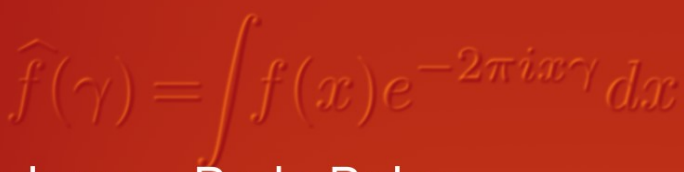
Applied and Numerical Harmonic Analysis



Travis D. Andrews · Radu Balan
John J. Benedetto · Wojciech Czaja
Kasso A. Okoudjou
Editors

Excursions in Harmonic Analysis, Volume 2

The February Fourier Talks at the Norbert Wiener Center

Birkhäuser

Applied and Numerical Harmonic Analysis

Travis D. Andrews • Radu Balan
John J. Benedetto • Wojciech Czaja
Kasso A. Okoudjou
Editors

Excursions in Harmonic Analysis, Volume 2

The February Fourier Talks at the Norbert Wiener Center

Birkhäuser

Editors
Travis D. Andrews
Norbert Wiener Center
Department of Mathematics
University of Maryland
College Park, MD, USA

Radu Balan
Norbert Wiener Center
Department of Mathematics
University of Maryland
College Park, MD, USA

John J. Benedetto
Norbert Wiener Center
Department of Mathematics
University of Maryland
College Park, MD, USA

Wojciech Czaja
Norbert Wiener Center
Department of Mathematics
University of Maryland
College Park, MD, USA

Kasso A. Okoudjou
Norbert Wiener Center
Department of Mathematics
University of Maryland
College Park, MD, USA

ISBN 978-0-8176-8378-8 ISBN 978-0-8176-8379-5 (eBook)
DOI 10.1007/978-0-8176-8379-5
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012951313

Mathematics Subject Classification (2010): 26-XX, 35-XX, 40-XX, 41-XX, 42-XX, 43-XX, 44-XX, 46-XX, 47-XX, 58-XX, 60-XX, 62-XX, 65-XX, 68-XX, 78-XX, 92-XX, 93-XX, 94-XX

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.birkhauser-science.com)

Dedicated to
Tom Grasso,
Friend and Editor Extraordinaire

ANHA Series Preface

The *Applied and Numerical Harmonic Analysis (ANHA)* book series aims to provide the engineering, mathematical, and scientific communities with significant developments in harmonic analysis, ranging from abstract harmonic analysis to basic applications. The title of the series reflects the importance of applications and numerical implementation, but richness and relevance of applications and implementation depend fundamentally on the structure and depth of theoretical underpinnings. Thus, from our point of view, the interleaving of theory and applications and their creative symbiotic evolution is axiomatic.

Harmonic analysis is a wellspring of ideas and applicability that has flourished, developed, and deepened over time within many disciplines and by means of creative cross-fertilization with diverse areas. The intricate and fundamental relationship between harmonic analysis and fields such as signal processing, partial differential equations (PDEs), and image processing is reflected in our state-of-theart *ANHA* series.

Our vision of modern harmonic analysis includes mathematical areas such as wavelet theory, Banach algebras, classical Fourier analysis, time-frequency analysis, and fractal geometry, as well as the diverse topics that impinge on them.

For example, wavelet theory can be considered an appropriate tool to deal with some basic problems in digital signal processing, speech and image processing, geophysics, pattern recognition, biomedical engineering, and turbulence. These areas implement the latest technology from sampling methods on surfaces to fast algorithms and computer vision methods. The underlying mathematics of wavelet theory depends not only on classical Fourier analysis, but also on ideas from abstract harmonic analysis, including von Neumann algebras and the affine group. This leads to a study of the Heisenberg group and its relationship to Gabor systems, and of the metaplectic group for a meaningful interaction of signal decomposition methods. The unifying influence of wavelet theory in the aforementioned topics illustrates the justification for providing a means for centralizing and disseminating information from the broader, but still focused, area of harmonic analysis. This will be a key role of *ANHA*. We intend to publish the scope and interaction that such a host of issues demands.

Along with our commitment to publish mathematically significant works at the frontiers of harmonic analysis, we have a comparably strong commitment to publish major advances in the following applicable topics in which harmonic analysis plays a substantial role:

Biomedical signal processing
Compressive sensing
Communications applications
Data mining/machine learning
Digital signal processing
Fast algorithms
Gabor theory and applications
Image processing
Numerical partial differential equations
Prediction theory
Radar applications
Sampling theory
Spectral estimation
Speech processing
Time-frequency and time-scale analysis
Wavelet theory

The above point of view for the *ANHA* book series is inspired by the history of Fourier analysis itself, whose tentacles reach into so many fields.

In the last two centuries, Fourier analysis has had a major impact on the development of mathematics, on the understanding of many engineering and scientific phenomena, and on the solution of some of the most important problems in mathematics and the sciences. Historically, Fourier series were developed in the analysis of some of the classical PDEs of mathematical physics; these series were used to solve such equations. In order to understand Fourier series and the kinds of solutions they could represent, some of the most basic notions of analysis were defined, e.g., the concept of "function". Since the coefficients of Fourier series are integrals, it is no surprise that Riemann integrals were conceived to deal with uniqueness properties of trigonometric series. Cantor's set theory was also developed because of such uniqueness questions.

A basic problem in Fourier analysis is to show how complicated phenomena, such as sound waves, can be described in terms of elementary harmonics. There are two aspects of this problem: first, to find, or even define properly, the harmonics or spectrum of a given phenomenon, e.g., the spectroscopy problem in optics; second, to determine which phenomena can be constructed from given classes of harmonics, as done, e.g., by the mechanical synthesizers in tidal analysis.

Fourier analysis is also the natural setting for many other problems in engineering, mathematics, and the sciences. For example, Wiener's Tauberian theorem in Fourier analysis not only characterizes the behavior of the prime numbers, but also provides the proper notion of spectrum for phenomena such as white light; this latter process leads to the Fourier analysis associated with correlation functions in filtering and prediction problems, and these problems, in turn, deal naturally with Hardy spaces in the theory of complex variables.

Nowadays, some of the theory of PDEs has given way to the study of Fourier integral operators. Problems in antenna theory are studied in terms of unimodular

trigonometric polynomials. Applications of Fourier analysis abound in signal processing, whether with the fast Fourier transform (FFT), or filter design, or the adaptive modeling inherent in time-frequency-scale methods such as wavelet theory. The coherent states of mathematical physics are translated and modulated Fourier transforms, and these are used, in conjunction with the uncertainty principle, for dealing with signal reconstruction in communications theory. We are back to the raison d'être of the *ANHA* series!

University of Maryland
College Park

John J. Benedetto
Series Editor

Preface

The chapters in these two volumes have at least one (co)author who spoke at the February Fourier Talks during the period 2006–2011.

The February Fourier Talks

The February Fourier Talks (*FFT*) were initiated in 2002 as a small meeting on harmonic analysis and applications, held at the University of Maryland, College Park. Since 2006, the *FFT* has been organized by the Norbert Wiener Center in the Department of Mathematics, and it has become a major annual conference. The *FFT* brings together applied and pure harmonic analysts along with scientists and engineers from industry and government for an intense and enriching two-day meeting. The goals of the *FFT* are the following:

- To offer a forum for applied and pure harmonic analysts to present their latest cutting-edge research to scientists working not only in the academic community but also in industry and government agencies,
- To give harmonic analysts the opportunity to hear from government and industry scientists about the latest problems in need of mathematical formulation and solution,
- To provide government and industry scientists with exposure to the latest research in harmonic analysis,
- To introduce young mathematicians and scientists to applied and pure harmonic analysis,
- To build bridges between pure harmonic analysis and applications thereof.

These goals stem from our belief that many of the problems arising in engineering today are directly related to the process of making pure mathematics applicable. The Norbert Wiener Center sees the *FFT* as the ideal venue to enhance this process in a constructive and creative way. Furthermore, we believe that our vision is shared

by the scientific community, as shown by the steady growth of the *FFT* over the years.

The *FFT* is formatted as a two-day single-track meeting consisting of thirty-minute talks as well as the following:

- Norbert Wiener Distinguished Lecturer series
- General interest keynote address
- Norbert Wiener Colloquium
- Graduate and postdoctoral poster session

The talks are given by experts in applied and pure harmonic analysis, including academic researchers and invited scientists from industry and government agencies.

The Norbert Wiener Distinguished Lecture caps the technical talks of the first day. It is given by a senior harmonic analyst, whose vision and depth through the years have had profound impact on our field. In contrast to the highly technical day sessions, the keynote address is aimed at a general public audience and highlights the role of mathematics, in general, and harmonic analysis, in particular. Furthermore, this address can be seen as an opportunity for practitioners in a specific area to present mathematical problems that they encounter in their work. The concluding lecture of each *FFT*, our Norbert Wiener Colloquium, features a mathematical talk by a renowned applied or pure harmonic analyst. The objective of the Norbert Wiener Colloquium is to give an overview of a particular problem or a new challenge in the field. We include here a list of speakers for these three lectures:

Distinguished lecturer

- Peter Lax
- Richard Kadison
- Elias Stein
- Ronald Coifman
- Gilbert Strang

Keynote address

- Frederick Williams
- Steven Schiff
- Peter Carr
- Barry Cipra
- William Noel
- James Coddington
- Mario Livio

Colloquium

- Christopher Heil
- Margaret Cheney
- Victor Wickerhauser
- Robert Fefferman
- Charles Fefferman
- Peter Jones

The Norbert Wiener Center

The Norbert Wiener Center for Harmonic Analysis and Applications provides a national focus for the broad area of mathematical engineering. Applied harmonic analysis and its theoretical underpinnings form the technological basis for this area. It can be confidently asserted that mathematical engineering will be to today's mathematics departments what mathematical physics was to those of a century ago. At that time, mathematical physics provided the impetus for tremendous advances within mathematics departments, with particular impact in fields such as differential

equations, operator theory, and numerical analysis. Tools developed in these fields were essential in the advances of applied physics, e.g., the development of the solid-state devices which now enable our information economy.

Mathematical engineering impels the study of fundamental harmonic analysis issues in the theories and applications of topics such as signal and image processing, machine learning, data mining, waveform design, and dimension reduction into mathematics departments. The results will advance the technologies of this millennium.

The golden age of mathematical engineering is upon us. The Norbert Wiener Center reflects the importance of integrating new mathematical technologies and algorithms in the context of current industrial and academic needs and problems. The Norbert Wiener Center has three goals:

- Research activities in harmonic analysis and applications
- Education—undergraduate to postdoctoral
- Interaction within the international harmonic analysis community

We believe that educating the next generation of harmonic analysts, with a strong understanding of the foundations of the field and a grasp of the problems arising in applications, is important for a high-level and productive industrial, government, and academic workforce.

The Norbert Wiener Center web site: www.norbertwiener.umd.edu

The Structure of the Volumes

To some extent the eight parts of these two volumes are artificial placeholders for all the diverse chapters. It is an organizational convenience that reflects major areas in harmonic analysis and its applications, and it is also a means to highlight significant modern thrusts in harmonic analysis. Each of the following parts includes an introduction that describes the chapters therein:

Volume 1	Volume 2
I Sampling Theory	V Measure Theory
II Remote Sensing	VI Filtering
III Mathematics of Data Processing	VII Operator Theory
IV Applications of Data Processing	VIII Biomathematics

Acknowledgments

The Norbert Wiener Center gratefully acknowledges the indispensable support of the following groups: Australian Academy of Science, Birkhäuser the IEEE Baltimore Section, MiMoCloud, Inc., Patton Electronics Co., Radyn, Inc., the SIAM Washington–Baltimore Section, and SR2 Group, LLC. One of the successes of the February Fourier Talks has been the dynamic participation of graduate student and postdoctoral engineers, mathematicians, and scientists. We have been fortunate to be able to provide travel and living expenses to this group due to continuing, significant grants from the National Science Foundation, which, along with the aforementioned organizations and companies, believes in and supports our vision of the FFT.

Contents

Part V
Measure Theory

The four chapters in this part treat old themes as well as some modern applications of measure theory. Measure theory was developed in the late nineteenth and early twentieth centuries, and its founders include E. Borel, H. Lebesgue, and J. Radon. It remains a central field in mathematics offering rigorous tools to tackle problems arising in mathematical analysis. Measure theory is also one of the foundational courses that many aspiring mathematicians are required to take. Even more, modern probability theory owes its very existence to measure theory, which also plays essential roles in areas such dynamical systems, harmonic analysis, and partial differential equations. The breadth and the depth of the chapters in this part partially illustrate how these two volumes are intended for various audiences ranging from graduate students, researchers, to practitioners in harmonic analysis and its applications.

The first chapter of this part is by ROBERT B. BURCKEL, who provides new proofs of two classical results about measures on the circle. The first result characterizing nonzero analytic measures on the unit circle is due to the Riesz brothers (1916). The second result, due to Szegö (1920), gives a condition under which the closed linear span of the monomials on the unit circle is dense in the Hilbert space associated with a measure on the circle. Burckel's beautiful proofs are based largely on Hilbert space arguments rather than being purely measure theoretical.

In the second chapter of this part, CARLOS CABRELLI, UDAYAN B. DARJI, AND URSULA MOLTER give a nontrivial extension of a construction of strongly invisible sets due to R. O. Davies to measures on Polish spaces (complete separable metric spaces). The setting of this type of result is measure theoretic dimension theory which offers methods to classify measurable sets. In this short but elegant chapter, Cabrelli, Darji, and Molter introduce tools that allow them to construct a large class of visible Cantor sets in Polish spaces. Visible Cantor sets are sets of finite positive measure for some Hausdorff measure or some translation invariant Borel measure. They also investigated strongly invisible sets in the setting of Polish groups. The new results they obtain include the density of certain of these classes of measurable sets within the set of all compact sets in the underlying space.

JEAN-PIERRE GABARDO'S chapter offers new insights to some applications of measure theory in time-frequency analysis. In particular, he establishes relationships between convolution inequalities for positive Borel measures in Euclidean space and the notion of upper and lower Beurling density for these measures. Such relationships arise in the study of the packing and tiling properties of translates of sets in Euclidean spaces, in which case the Borel measures considered are sums of Dirac masses. The connection to time-frequency analysis lies in the fact that the quantitative behavior of Beurling density is important in the theory of uniform and nonuniform Gabor systems. A Gabor system consists of the time-frequency shifts of a fixed window function along a discrete set of points in Euclidean space. If this system is a frame, then the upper Beurling density of the corresponding discrete set must be finite, while its lower Beurling density must be at least one. In this case, an appropriate choice of the window function can lead to the decomposition and reconstruction of any function from samples of its short-time Fourier transform.

In the final chapter of this part, BILL MORAN, STEPHEN HOWARD, AND DOUG COCHRAN use the parallel between frame theory and the theory of positive operator-valued measures (POVMs) on a Hilbert space, $\mathscr{H}$, to introduce the new concept of a *framed POVM*. Frames are redundant sets of vectors that provide stable and efficient algorithms for the decomposition and reconstruction of any vector in $\mathscr{H}$. Consequently, frames are a natural tool in many signal processing applications. On the other hand, the notion of POVM was introduced and developed in quantum mechanics as a tool to represent the most general form of quantum measurement of a system. After giving an overview of the similarities between frames for $\mathscr{H}$ and POVMs associated to $\mathscr{H}$, the authors show that the class of framed POVMs is large in a sense defined in their chapter. In fact, this class includes frames as well as more recent extensions such as fusion frames and generalized frames. One of the interesting aspects of framed POVMs is that tools from quantum mechanics can now be translated to obtain new results in frame theory.

Absolute Continuity and Singularity of Measures Without Measure Theory

R.B. Burckel

Abstract This chapter presents proofs, largely by Hilbert-space arguments, of two classical results about measures on the circle associated with the Riesz Brothers (1916) and Gabor Szegö (1920).

Keywords Absolutely continuous measures • Approximation theorems of Weierstrass and Féjer • Hilbert-space methods in measure theory • M. Riesz Theorem • Shifts in Hilbert space • Singular measures • Szegö's Theorem • Wold decomposition

This chapter presents proofs of two classical results about measures on the circle associated with the Riesz Brothers [3] and G. Szegö [5]. Of course, measure theory cannot be fully eschewed (so the title is something of a come-on), but the proofs are largely by Hilbert-space arguments, and some minimal notation is needed:

$\mathbb{Z}$ is the integers, $\mathbb{N} := \{n \in \mathbb{Z} : n \geq 1\}$, $\mathbb{N}_0 := \{n \in \mathbb{Z} : n \geq 0\}$, $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$, $\mathbb{T} := \partial\mathbb{D}$, $C(\mathbb{T})$ is the continuous functions on $\mathbb{T}$, λ is normalized Lebesgue measure on the Borel subsets of $\mathbb{T}$. For a complex-valued Borel measure ν on $\mathbb{T}$ the number

$$\hat{\nu}(k) := \int_{\mathbb{T}} z^{-k}\, d\nu(z)$$

is called the *k-th Fourier coefficient* of ν, for each $k \in \mathbb{Z}$. The absolute continuity relation is signalled, as is customary, by $\ll$. For a subset S of a vector space span(S) will denote its (algebraic) linear span. Classical theorems of Weierstrass and Fejér assert that

$$\text{span } \{z^k : k \in \mathbb{Z}\} \text{ is uniformly dense in } C(\mathbb{T}). \qquad \text{(WF)}$$

R.B. Burckel (✉)
Department of Mathematics, Kansas State University,
138 Cardwell Hall, Manhattan, KS-66506, USA
e-mail: burckel@math.ksu.edu

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_1,

A consequence of (WF) is

$$\text{span } \{z^k : k \in \mathbb{Z}\} \text{ is dense in every } L^2(\nu) \text{ space,} \qquad (\text{WF})_\nu$$

and so the Fourier coefficients uniquely determine ν; that is, $\hat{\nu}(k) = 0$ for all k if and only if $\nu = 0$. ν is called *analytic* if

$$\hat{\nu}(-n) := \int_{\mathbb{T}} z^n \, \mathrm{d}\nu(z) = 0 \quad \forall n \in \mathbb{N}. \qquad (\text{A})$$

This terminology derives from the fact (Cauchy's integral theorem) that if f is continuous on $\overline{\mathbb{D}}$ and analytic in $\mathbb{D}$, then the measure $f\,\mathrm{d}\lambda$ is analytic.

Theorem 1 (F. and M. Riesz Theorem). *Every non-zero analytic measure ν is mutually absolutely continuous with respect to λ.*

Proof. Let μ denote the total variation measure of ν and f the Radon–Nikodym derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$. That is, $\mathrm{d}\nu = f\,\mathrm{d}\mu$ and $|f| = 1$ μ-a.e. Then (A) can be written

$$\int_{\mathbb{T}} z^n f(z) \mathrm{d}\mu(z) = 0 \quad \forall n \in \mathbb{N}. \qquad (1)$$

Let $\langle\,,\,\rangle$ and $\|\,\|$ denote inner product and norm in $L^2(\mu)$ and U (suggesting *unilateral* shift) the operator, evidently unitary, of multiplication by z on this Hilbert space. According to (1) the constant function 1 is orthogonal to every $\mathrm{U}^n f$ $(n \in \mathbb{N})$, so the set

$$M := L^2(\mu)\text{-closure of span } \{\mathrm{U}^n f : n \in \mathbb{N}\} \qquad (2)$$

is a *proper* subspace of $L^2(\mu)$, evidently U-invariant. Since $|f| = 1$ μ-a.e., $(\text{WF})_\mu$ entails that

$$\text{span}\{z^k f : k \in \mathbb{Z}\} \text{ is dense in } L^2(\mu). \qquad (3)$$

Let us note that

$$\mathrm{U}M \subsetneq M. \qquad (4)$$

For if $\mathrm{U}M = M$, then $\mathrm{U}^* M = \mathrm{U}^* \mathrm{U} M = M$, so M would contain $(\mathrm{U}^*)^m \mathrm{U}^n f = (\bar{z})^m z^n f = z^{n-m} f$ for all n, $m \in \mathbb{N}$ and consequently $z^k f$ for all $k \in \mathbb{Z}$. From this and (3) would follow the contradiction $L^2(\mu) \subset M$. This confirms (4).

Form the orthogonal complement $M \ominus \mathrm{U}M$, which is not $\{0\}$ by (4), and note that the (closed) subspaces $\mathrm{U}^k(M \ominus \mathrm{U}M)$ are orthogonal, which is pretty clear when they are written as $\mathrm{U}^k M \ominus \mathrm{U}^{k+1} M : m_1, m_2 \in M \ominus \mathrm{U}M$, $p > k \Rightarrow \langle \mathrm{U}^p m_1, \mathrm{U}^k m_2 \rangle = \langle \mathrm{U}^{p-k} m_1, m_2 \rangle = 0$ since $\mathrm{U}^{p-k} m_1 \in \mathrm{U}M$ and $m_2 \perp \mathrm{U}M$.

As a special case of this

$$\{\mathrm{U}^k h\}_{k \in \mathbb{Z}} \text{ is an orthonormal sequence in } L^2(\mu) \qquad (5)$$

for every unit vector $h \in M \ominus \mathrm{U}M$. Note that

$$\text{subspace } \bigcap_{n\geq 0} \mathrm{U}^n M \text{ is orthogonal to } \mathrm{U}^k(M \ominus \mathrm{U}M) \quad \forall k \in \mathbb{Z}. \tag{6}$$

For if $m_1 \in M \ominus \mathrm{U}M$ and m_0 lies in this intersection, then $m_0 = \mathrm{U}^{|k|+1} m_2$ for some $m_2 \in M$, and so

$$\langle m_0, \mathrm{U}^k m_1\rangle = \langle \mathrm{U}^{|k|+1} m_2, \mathrm{U}^k m_1\rangle = \langle \mathrm{U}^{|k|-k+1} m_2, m_1\rangle = 0,$$

since $\mathrm{U}^{|k|-k+1} m_2 \in \mathrm{U}M$. The same argument shows that

$$\bigcap_{n\geq 0} \mathrm{U}^n M \text{ is orthogonal to } \mathrm{U}^k 1 = z^k \quad \forall k \in \mathbb{Z}. \tag{6}_1$$

For $\langle m_0, \mathrm{U}^k 1\rangle = \langle \mathrm{U}^{|k|+1} m_2, \mathrm{U}^k 1\rangle = \langle \mathrm{U}^{|k|-k+1} m_2, 1\rangle = 0$, since, as already noted, $1 \perp M$.

Now the well-known *Wold decomposition* (see Sz.-Nagy and Foiaş [6, p.3]) says that M is the orthogonal sum

$$M = \bigcap_{n\geq 0} \mathrm{U}^n M \oplus \bigoplus_{k\geq 0} \mathrm{U}^k(M \ominus \mathrm{U}M). \tag{7}$$

Again, this is pretty transparent when the right side is written out:

$$\bigcap_{n\geq 0} \mathrm{U}^n M \oplus (M \ominus \mathrm{U}M) \oplus (\mathrm{U}M \ominus \mathrm{U}^2 M) \oplus (\mathrm{U}^2 M \ominus \mathrm{U}^3 M) \oplus \cdots$$

From Eq. $(6)_1$ and $(\mathrm{WF})_\mu$ it follows that the space $\bigcap_{n\geq 0} \mathrm{U}^n M$ must be $\{0\}$ and (7) reads

$$M = \bigoplus_{k\geq 0} \mathrm{U}^k(M \ominus \mathrm{U}M) \tag{8}$$

Next we aim to show that

$$M \ominus \mathrm{U}M \text{ is 1-dimensional.} \tag{9}$$

To this end, consider a fixed $h \in M \ominus \mathrm{U}M$ of norm 1 [recalling (4)]. If (9) fails, there exists $g \in M \ominus \mathrm{U}M$ of norm 1 which is orthogonal to h. Then, by the familiar maneuvers, $\mathrm{U}^m h \perp \mathrm{U}^n g$ for all $m, n \in \mathbb{N}_0$.

$$0 = \langle \mathrm{U}^m h, \mathrm{U}^n g\rangle = \langle \mathrm{U}^{m-n} h, g\rangle \quad \forall m, n \geq 0 \Rightarrow$$
$$0 = \langle \mathrm{U}^k h, g\rangle = \int_{\mathbb{T}} z^k h\bar{g}\, \mathrm{d}\mu \quad \forall k \in \mathbb{Z}.$$

The function $h\bar{g}$ being in $L^1(\mu)$, these equations affirm that all Fourier coefficients of the (complex Borel) measure $h\bar{g}\,\mathrm{d}\mu$ vanish. Hence this measure is 0. That is,

$$|h\,\bar{g}| = |h||g| = 0 \quad \mu\text{-a.e.} \tag{10}$$

As noted in Eq. (5)

$$\int_{\mathbb{T}} z^k |h|^2\,\mathrm{d}\mu(z) = \langle \mathrm{U}^k h, h\rangle = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{if } k \in \mathbb{Z}\setminus\{0\}, \end{cases}$$

that is, the measure $|h|^2\mathrm{d}\mu$ has exactly the same Fourier coefficients as λ, so

$$|h|^2\mathrm{d}\mu = \mathrm{d}\lambda. \tag{11}$$

Similarly, $|g|^2\mathrm{d}\mu = \mathrm{d}\lambda$, which with Eq. (11) gives $|h|^3\mathrm{d}\mu = |h|\mathrm{d}\lambda = |h||g|^2\mathrm{d}\mu{=}0$, by (10). That is, $|h| = 0$ μ-a.e., contrary to h being of norm 1 in $L^2(\mu)$. This contradiction confirms (9). That is, $M \ominus \mathrm{U}M = \mathbb{C}h$ and then by Eq. (8)

$$L^2(\mu)\text{-closure of span } \{\mathrm{U}^n h : n \in \mathbb{N}_0\} = M.$$

In particular, since $\mathrm{U}f = zf \in M$, we see that zf lies in the $L^2(\mu)$-closure of span$\{\mathrm{U}^n h = z^n h : n \in \mathbb{N}_0\}$. It follows that for each $k \in \mathbb{Z}$, $z^k f$ lies in the $L^2(\mu)$-closure of span$\{z^n h : n \in \mathbb{Z}\}$. That is,

$$L^2(\mu)\text{-closure of span } \{z^k f : k \in \mathbb{Z}\} \subset L^2(\mu)\text{-closure of span } \{z^n h : n \in \mathbb{Z}\}. \tag{12}$$

Equality (3) forces equality in (12), which then clearly entails that

$$f\,\mathrm{d}\mu \ll h\,\mathrm{d}\mu \ll f\,\mathrm{d}\mu,$$

and thanks to Eq. (11)

$$\mathrm{d}\lambda \ll |h|\,\mathrm{d}\mu \ll \mathrm{d}\lambda.$$

Thus $\mathrm{d}\lambda$ and $f\,\mathrm{d}\mu$ (which is $\mathrm{d}\nu$) are mutually absolutely continuous.

Corollary 1 (Szegö). *Let σ be a Borel probability measure on $\mathbb{T}$ that annihilates some set of positive Lebesgue measure. Then the functions z^n, $n \in \mathbb{N}$, span $L^2(\sigma)$.*

Proof. Denote by M the $L^2(\sigma)$-closure of span$\{z^n : n \in \mathbb{N}\}$ and assume $M \neq L^2(\sigma)$. There is then a non-zero function $g \in L^2(\sigma)$ orthogonal to M:

$$0 = \langle g, z^n\rangle_{L^2(\sigma)} = \int_{\mathbb{T}} \overline{z^n} g\,\mathrm{d}\sigma \quad \forall n \in \mathbb{N}.$$

This says that $\bar{g}\mathrm{d}\sigma$ is a non-zero analytic measure, hence Lebesgue measure λ is absolutely continuous with respect to $\bar{g}\mathrm{d}\sigma$. By hypothesis, some Borel set B with

$\lambda(B) > 0$ has $\sigma(B) = 0$. The latter equality implies that $\bar{g}\mathrm{d}\sigma$ also gives measure 0 to B, contradicting $\mathrm{d}\lambda \ll \bar{g}\mathrm{d}\sigma$.

Remark 1. The usual formulation of Szegö's theorem is more quantitative and more general: If h denotes the Radon–Nikodym derivative with respect to λ of the absolutely continuous part of σ, then the distance in $L^2(\sigma)$ from 1 to span $\{z^n : n \in \mathbb{N}\}$ is $\exp(\int \log h \,\mathrm{d}\lambda)$. In the notation of the proof just given, h would be 0 on B, so $\int \log h \,\mathrm{d}\lambda = -\infty$ and the exponential of it is 0, which puts 1 into the $L^2(\sigma)$-closure of span$\{z^n : n \in \mathbb{N}\}$, from which the above corollary follows easily.

Remark 2. On the other hand, a weaker version of the corollary was proved by Holland [1]. He started with a Borel probability measure σ on $\mathbb{T}$ which is singular with respect to Lebesgue measure λ and by a very clever, explicit and elementary construction he manufactured a sequence of polynomials

$$P_n(z) = \sum_{k=1}^{n} A_k z^k \quad (n \in \mathbb{N})$$

such that

$$\sum_{k=1}^{\infty} |A_k|^2 = 1$$

and

$$\int_{\mathbb{T}} |1 - P_n|^2 \,\mathrm{d}\sigma = 1 - \sum_{k=1}^{n} |A_k|^2 \quad (n \in \mathbb{N}).$$

In fact, the A_k are the Taylor coefficients of the holomorphic function

$$\frac{F(z) - 1}{F(z) + 1}, \text{ where } F(z) := \int_{\mathbb{T}} \frac{u + z}{u - z} \,\mathrm{d}\sigma(u) \quad (z \in \mathbb{D}).$$

Mirabile dictu.

Remark 3. Also the half of the F. and M. Riesz theorem asserting that $\nu \ll \lambda$ was given a remarkable one-page function-theoretic proof by Øksendal [2]. A complex-valued Borel measure ν on $\mathbb{T}$ satisfying (A) is given and what has to be shown is that $\nu(K) = 0$ for every λ-null Borel set K. Because Borel measures on $\mathbb{T}$ are inner regular, it suffices to consider compact K. Clearly it suffices to show this for the modified measure $\nu_0 := \nu - \nu(\mathbb{T})\lambda$. This measure is also analytic but in addition annihilates 1. That is,

$$\widehat{\nu_0}(-n) = \int_{\mathbb{T}} z^n \,\mathrm{d}\nu_0(z) = 0 \quad \forall n \in \mathbb{N}_0. \tag{A}_0$$

For each $n \in \mathbb{N}$, an $N \in \mathbb{N}$, $z_j \in K$ and $\rho_j > 0$ are chosen appropriately and the rational functions

$$g_n(z) := 1 - \prod_{j=1}^{N} \frac{z - z_j}{z - (1 + \rho_j) z_j}$$

are introduced. They are bounded by 2 on $\mathbb{T}$ and are shown to converge there to the indicator function of K as $n \to \infty$. Since g_n is holomorphic in a neighborhood of $\overline{\mathbb{D}}$, the partial sums of its Taylor series at 0 approximate it uniformly on $\mathbb{T}$ and each sum has ν_0-integral 0 thanks to $(A)_0$. Consequently, $\int_{\mathbb{T}} g_n \, d\nu_0 = 0$. It follows then from the dominated convergence theorem that $\nu_0(K) = \lim \int g_n \, d\nu_0 = 0$, as wanted.

Acknowledgments In February of 2010 this proof was kindly communicated to me by Donald Sarason when I sought his insight on Holland [1]. He informed me that a version of it had been discovered by David Lowdenslager sometime between 1959 and 1963, and independently by himself while a graduate student of Paul Halmos' at the University of Michigan about 1962. Apparently neither was ever published, but some of Don's ideas appear in Sarason [4]. He has graciously agreed to my presenting his work here. It is an honor to be the purveyor of such elegant mathematics. I extend deepest thanks to John Benedetto for inviting me to speak to such a distinguished group.

References

1. Holland, F: Another proof of Szegö's theorem for a singular measure. Proc. Amer. Math. Soc. 45, 311–312 (1974) MR 50#2784
2. Øksendal, B.K.: A short proof of the F. & M. Riesz theorem. Proc. Amer. Math. Soc. 30 (1971), 204. MR 43#5039
3. Riesz, F., Riesz, M.: Über die Randwerte einer analytischen Funktion, Quatrième Congrès des Mathématiciens Scandinaves, Stockholm (1916), pp. 27–44. Almqvist and Wiksells, Uppsala (1920), JFM 47, p. 295
4. Sarason, D: Invariant subspaces. In: Pearcy, C. (eds.) pp. 1–47 of Topics in Operator Theory, Mathematical Surveys 13 (1974) A.M.S. MR 50#10862
5. Szegö, G.: Beiträge zur Theorie der Toeplitzschen Formen, erste Mitteilung, Math. Zeit. **6**, 167–202 (1920). JFM 48, p. 376
6. Sz.-Nagy, B., Foiaş, C.: Harmonic Analysis of Operators on Hilbert space, North-Holland Publishing Company, Amsterdam (1970). MR43# 947

Further Reading

Halmos, P: Shifts in Hilbert spaces. Jour. Reine und Angew. Math. **208**, 102–112 (1961). MR 27#2868
Wold, H: A Study in the Analysis of Stationary Time Series. Almqvist and Wiksells, Stockholm (1938). JFM 64, p. 1200

Visible and Invisible Cantor Sets

Carlos Cabrelli, Udayan B. Darji, and Ursula Molter

Abstract In this chapter we study for which Cantor sets there exists a *gauge*-function h, such that the $h-$Hausdorff measure—is positive and finite. We show that the collection of sets for which this is true is dense in the set of all compact subsets of a Polish space X. More general, any *generic* Cantor set satisfies that there exists a translation-invariant measure μ for which the set has positive and finite μ-measure. In contrast, we generalize an example of Davies of dimensionless Cantor sets (i.e., a Cantor set for which any translation invariant measure is either 0 or non-σ-finite) that enables us to show that the collection of these sets is also dense in the set of all compact subsets of a Polish space X.

Keywords Cantor set • Visible set • Hausdorff measure • Cantor space • Polish space • Dimensionless set • Strongly invisible set • Davies set • Comeager set • Tree • Cantor tree • Generic element

1 Introduction

Measure theoretic dimension theory provides a fundamental tool to classify sets. However, Hausdorff dimension as well as other notions of dimension such as packing dimension and Minkowski dimension are not completely satisfactory.

C. Cabrelli • U. Molter (✉)
Departamento de Matemática FCEyN, Universidad de Buenos Aires
C1428EGA C.A.B.A., Argentina IMAS - CONICET
e-mail: cabrelli@dm.uba.ar; umolter@dm.uba.ar

U.B. Darji
Department of Mathematics, University of Louisville,
Louisville, KY 40292, USA
e-mail: ubdarj01@louisville.edu

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_2,

For example, there are many examples of compact sets whose Hausdorff measure at its critical exponent is zero or infinite. This could also happen even if we consider the generalized Hausdorff h-measure where h is an appropriate gauge function in a well-defined class. See for example [2]. Furthermore, in 1930 Davies [5] produced a beautiful example of a Cantor set on the real line whose μ-measure is zero or no σ-finite for every translation invariant Borel measure μ. See [6] for more results in this direction.

In this chapter we study this phenomena. We try to estimate in some way the size of the class of visible sets, i.e., sets that have positive and finite measure for some h-Hausdorff measure or some translation invariant Borel measure.

We focus on Cantor sets in the context of Polish spaces. In [4] the authors proved that a large class of Cantor sets defined by monotone gap-sequences are visible. They explicitly construct the corresponding gauge function h. See also [3]. Here we extend this result to a larger class in general Polish spaces (Theorem 3.2). We also obtain density results for the class of visible sets and study *generic* visibility for subsets of the real line.

Then we focus on the concept of strong invisibility (see Definition 2.4). We were able to extend the ideas in the construction of Davies to a general abelian Polish group, obtaining a big class of strongly invisible compact sets in these groups. We also prove that the set of strongly invisible sets in the space of compact sets in the line with the Hausdorff distance is dense.

This chapter is organized as follows. We first introduce some notation and terminology in Sect. 2. A key ingredient will be the definition of visibility and strong invisibility and the analysis of the appropriate topology to be able to state density results. In Sect. 3 we show that a large class of Cantor sets is visible, and in Sect. 4 we show how to construct many strongly invisible sets.

2 Terminology and Notation

Throughout X will denote a *Polish space*, i.e., a separable space with a complete metric. We let $\mathscr{C}(X)$ denote the set of all compact subsets of X endowed with the Hausdroff metric d_H. We recall that for X Polish, $\mathscr{C}(X)$ is Polish, and for X compact, $\mathscr{C}(X)$ is compact. We let $\mathscr{B}(X)$ denote the set of Borel subsets of X.

A subset of a Polish space is a *Cantor space* (or *Cantor set*) if it is compact, has no isolated points, and has a basis of clopen sets, i.e., sets which are simultaneously open and closed. There is always a homeomorphism between two Cantor sets. We also consider several special types of Cantor sets subsets of the reals. We note that for a subset of the reals to be a Cantor set, it suffices to have the properties of being compact, perfect, and containing no interval.

We now describe a general way of describing any Cantor set subset of $[0, 1]$ of Lebesgue measure zero which contains $\{0, 1\}$ (see [1]). Let D be the set of dyadic rationals in $(0, 1)$.

$$D = \left\{ i2^{-k} | 1 \leq i \leq 2^k - 1, k \in \mathbb{N} \right\},$$

$$\mathscr{G} = \left\{ \varphi : D \to (0,1) | \sum_{d \in D} \varphi(d) = 1 \right\},$$

and for each $\varphi \in \mathscr{G}$ we associate the function,

$$\Phi(d) = \sum_{d' \in D, d' < d} \varphi(d').$$

The function φ can be thought of as a density function supported on D, and Φ is the associated cumulative distribution function. Associated to φ, we define the Cantor set K_φ as follows:

$$K_\varphi = [0,1] \setminus \bigcup_{d \in D} (\Phi(d), \Phi(d) + \varphi(d)).$$

We think of φ as the "gap function" of K_φ. We have the following basic facts.

Proposition 2.1. *For each $\varphi \in \mathscr{G}$, K_φ is Cantor set subset of $[0,1]$ containing $\{0,1\}$ with Lebesgue measure zero. Conversely, given any Cantor set K subset of $[0,1]$ with Lebesgue measure zero which also contains $\{0,1\}$, there is $\varphi \in \mathscr{G}$ such that $K = K_\varphi$.*

Proposition 2.2. *Suppose $\varphi_1, \varphi_2 \in \mathscr{G}$ are such that $K_{\varphi_1} = K_{\varphi_2}$. Then, there is a homeomorphism g on $[0,1]$, mapping D onto itself such that $\varphi_2 = \varphi_1 \circ g$.*

The following special subclass of K_φ's was studied in [4]. Let

$$\mathscr{D}S = \left\{ \alpha \in (\mathbb{R}^+)^{\mathbb{N}} : \alpha \text{ is decreasing and } \sum_{n=1}^{\infty} \alpha(i) = 1 \right\}.$$

For each $\alpha \in \mathscr{D}S$, we define $K_\alpha = K_\varphi$ where $\varphi(1/2) = \alpha(1)$, $\varphi(1/4) = \alpha(2)$, $\varphi(3/4) = \alpha(3), \ldots, \varphi(\frac{2s+1}{2^j}) = \alpha(2^{j-1} + s)$. We call the sequence α a "gap sequence."

We introduce necessary terminology and notation concerning measures. Let

$$\mathscr{H} = \{h : [0,\infty) \to [0,\infty) | h(0) = 0, h \text{ is continuous and nondecreasing}\},$$

and μ_h be the associated Hausdorff measure defined on the Borel subsets of X.

Definition 2.3. We call $M \in \mathscr{B}(X)$ **$\mathscr{H}$-visible** if there is $h \in \mathscr{H}$ such that $0 < \mu_h(M) < \infty$, i.e., M is an h-set for some $h \in \mathscr{H}$.

In [4] it was shown that for any $\alpha \in \mathscr{D}S$, K_α is $\mathscr{H}$-visible.

The main purpose of this chapter is to determine whether the previous result can be extended to other K_φ or if not, how big is the class of Cantor sets for which this is true.

In Polish groups, i.e., topological groups with Polish topology, Hausdorff measures are particular instances of general translation invariant Borel measures.

Definition 2.4. Let X be a Polish group. A set $M \in \mathscr{B}(X)$ is called **visible** if there exists a translation invariant Borel measure μ on $\mathscr{B}(X)$ such that $0 < \mu(M) < \infty$. A set $M \in \mathscr{B}(X)$ is called **strongly invisible** if for every translation invariant Borel measure μ on $\mathscr{B}(X)$ we have that $\mu(M) = 0$ or M is not $\mu\sigma$-finite.

Davies [5] showed that there is a compact subset of $\mathbb{R}$ which is strongly invisible. In a Polish group a **Davies set** is a compact set which is strongly invisible. Many natural examples of Borel sets which are strongly invisible were given in [6].

We would like to discuss visibility of a "randomly" chosen compact or Cantor set. Unfortunately, even in the case of the reals, there is no suitable natural measure on the set of compact or Cantor sets. Hence, we use the notion of genericity. Let X be a Polish space. A set $M \subseteq X$ is **meager** if it is the countable union of nowhere dense sets. The set of meager sets forms a σ-ideal, i.e., a subset of a meager set is meager, and the countable union of meager sets is meager. Moreover, as X is complete, the Baire category theorem holds, and hence no nonempty open set is meager. One thinks of meager sets as a collection of small sets and its complements as big sets. A set is **comeager** if its complement is meager. We say that a **generic element of** X **has property** P when the set of elements of X which has property P is comeager in X.

3 Visible Sets

In this section we study visible sets and $\mathscr{H}$-visible sets. In particular, Theorem 3.2 shows that a large class of Cantor sets are $\mathscr{H}$-visible. Proposition 3.3 shows that this includes the class of Cantor sets K_α, $\alpha \in \mathscr{D}S$, studied in [4]. Then, we discuss how big are the classes of visible, $\mathscr{H}$-visible, and strongly invisible sets. We show that the class of $\mathscr{H}$-visible and strongly invisible sets are dense in $\mathscr{C}([0,1])$ and, moreover, a generic compact subset of $[0,1]$ is visible. It remains open whether a generic compact subset of $[0,1]$ is $\mathscr{H}$-visible.

3.1 $\mathscr{H}$-Visibility

We now introduce the construction necessary for our main $\mathscr{H}$-visibility theorem.

Let

$$\mathbb{N}^{<\mathbb{N}} = \bigcup_{n\in\mathbb{N}} \mathbb{N}^n$$

be the set of all finite sequences from $\mathbb{N}$. For $\sigma \in \mathbb{N}^{<\mathbb{N}}$, $\sigma = \sigma_1 \dots \sigma_n$, we denote by $|\sigma| = n$ the length of σ, and for $k < |\sigma|$, $\sigma|k$ are the first k digits of σ. We say

that $\sigma \in \mathbb{N}^{<\mathbb{N}}$ is an extension of $\tau \in \mathbb{N}^{<\mathbb{N}}$ if $|\sigma| > |\tau|$ and $\sigma||\tau| = \tau$. Further, if $\sigma \in \mathbb{N}^{\mathbb{N}}$, $\sigma|n$ is the restriction of σ to the first n digits.

A **tree T** is simply a subset of $\mathbb{N}^{<\mathbb{N}}$ with the property that if $\sigma \in T$ and σ is an extension of τ, then $\tau \in T$. The **body** of T, denoted by $[T]$, is

$$[T] = \{\sigma \in \mathbb{N}^{\mathbb{N}} : \sigma|n \in T \text{ for all } n \in \mathbb{N}\}.$$

For a tree T and $\sigma \in T$, the valency of σ in T is the cardinality of $\{n \in \mathbb{N} : \sigma n \in T\}$. A tree T is a **Cantor tree** if for each $\sigma \in T$, the valency of σ is finite and at least 2. If T is a Cantor tree and each $\sigma \in T$ has valency n, then T is an **n-Cantor tree.**

Let X be a Polish space and T be a tree. A function f from T into the collection of all nonempty open subsets of X is called a **T-assignment** into X if the following conditions are satisfied:

1. For each $\sigma \in T$, $f(\sigma)$ is a nonempty open subset of X.
2. The diameter of $f(\sigma)$ is less than $1/|\sigma|$ for all $\sigma \in T$.
3. If $\sigma, \tau \in T$ with $\sigma \neq \tau$ and $|\sigma| = |\tau|$, then $f(\sigma) \cap f(\tau) = \emptyset$.
4. If $\sigma, \tau \in T$ with τ an extension of σ, then $\overline{f(\tau)} \subseteq f(\sigma)$.

If f is a T-assignment into X, then we let

$$[f] = \bigcup_{\sigma \in [T]} \bigcap_{n=1}^{\infty} \overline{f(\sigma|n)}.$$

The following proposition is obvious.

Proposition 3.1. *Let X be a Polish space, T be a Cantor tree, and f be a T-assignment into X. Then, $[f]$ is a Cantor set.*

For the following definitions assume that T is a Cantor tree and f is a T-assignment into X. We say that f is a **regular T-assignment** if for all $n \in \mathbb{N}$ the following holds:

$$\max\{\text{diam}(f(\sigma)) : \sigma \in T, |\sigma| = n+1\} < \min\{diam(f(\sigma)) : \sigma \in T, |\sigma| = n\}.$$

Let $2 \leq l < \infty$. We say that f satisfies the **l-intersection condition** if the following condition holds.

For each $n \in \mathbb{N}$ and each open ball B in X, if B intersects at least l elements of $\{f(\sigma) : \sigma \in T, |\sigma| = n\}$, then we must have that $\overline{f(\sigma)} \subseteq B$ for some $\sigma \in T$ with $|\sigma| = n$.

Theorem 3.2. *Let X be a Polish space, T be a n-Cantor tree, and f be a T-assignment which is regular and satisfies the l-intersection condition. Then, $[f]$ is $\mathscr{H}$-visible.*

Proof. For each $k \in \mathbb{N}$, let

$$m_k = \min\{\text{diam}(f(\sigma)) : \sigma \in T, |\sigma| = k\} \quad \text{and}$$
$$M_k = \max\{\text{diam}(f(\sigma)) : \sigma \in T, |\sigma| = k\}.$$

Let $h \in \mathscr{H}$ such that $h([m_k, M_k]) = n^{-k}, k \in \mathbb{N}$. This is possible since f is regular. We claim that $[f]$ is a h-set for μ^h.

It is clear from the construction that $\mu^h([f]) \leq 1$.

Let now $j_0 \in \mathbb{N}$ be such that $n^{j_0} > l$ and $c = n^{-j_0-1}$. We will show that $\mu^h([f]) \geq c$.

For this, let $\mathscr{C}$ be any collection of open balls in X which covers $[f]$. It suffices to show that $\sum_{B\in\mathscr{C}} h(\mathrm{diam}(B)) \geq c$. Let $C_1, C_2, \ldots C_t$ be distinct elements of $\mathscr{C}$ such that

$$[f] \subseteq \bigcup_{i=1}^{t} C_i \quad \text{and each} \quad C_i \cap [f] \neq \emptyset.$$

Let λ be the *Lebesgue number* associated with the covering $C_1, C_2, \ldots C_t$ and $[f]$ such that if B is any open set with $\mathrm{diam}(B) < \lambda$ and $B \cap [f] \neq \emptyset$, then $B \subseteq C_i$ for some $1 \leq i \leq t$.

Let $k_0 \in \mathbb{N}$ be such that $1/k_0 < \lambda$ and each of C_i, $1 \leq i \leq t$, contains more than n^{j_0} many elements from $\{f(\sigma) : \sigma \in T, |\sigma| = k_0\}$.

Now, for each $1 \leq i \leq t$, let c_i be the cardinality of

$$\left\{ f(\sigma) : \sigma \in T, |\sigma| = k_0, \overline{f(\sigma)} \subseteq C_i \right\}$$

and let $d_i \in \mathbb{N}$ be such that $n^{d_i} < c_i \leq n^{d_i+1}$. We note that each $d_i > j_0$ and

$$\sum_{i=1}^{t} c_i \geq n^{k_0} \text{ as } 1/k_0 < \lambda.$$

Fix $1 \leq i \leq t$. We next observe that C_i intersects at least l elements from the collection

$$\{f(\sigma) : \sigma \in T, |\sigma| = k_0 - (d_i - j_0)\}.$$

For otherwise, we would have that C_i intersects less than $l \cdot n^{d_i - j_0}$ elements from the collection $\{f(\sigma) : \sigma \in T, |\sigma| = k_0\}$. As $l \cdot n^{d_i-j_0} = l \cdot n^{-j_0} \cdot n^{d_i} < n^{d_i} < c_i$, this would lead to a contradiction.

Now, since f satisfies the l-intersection condition, we have that $\overline{f(\sigma_i)} \subseteq C_i$ for some $\sigma_i \in T$ with $|\sigma_i| = k_0 - (d_i - j_0)$.

Now,

$$\begin{aligned}
\sum_{i=1}^{t} h(\mathrm{diam}(C_i)) &\geq \sum_{i=1}^{t} h(\mathrm{diam}(f(\sigma_i))) \\
&\geq \sum_{i=1}^{t} n^{-(k_0-(d_i-j_0))} \\
&= n^{-k_0} n^{-j_0} \sum_{i=1}^{t} n^{d_i}
\end{aligned}$$

Since $n^{d_i} < c_i \leq n^{d_i+1}$ and $\sum_{i=1}^{t} c_i \geq n^{k_0}$, we have that $\sum_{i=1}^{t} n^{d_i} > n^{k_0}/n$. Hence, we have that $\sum_{i=1}^{t} h(\text{diam}(C_i)) > n^{-j_0}/n = n^{-j_0-1} = c$, as claimed.

We will now show that the Cantor sets K_α, considered in [4], fall under the hypothesis of the Theorem 3.2, and therefore the latter theorem extends the result obtained in the earlier paper.

Proposition 3.3. *Let $\alpha \in \mathscr{D}S$. There is a regular T-assignment f from a 2-Cantor tree, satisfying the 3-intersection condition, such that $[f] = K_\alpha$.*

Proof. Recall that $K_\alpha = K_\varphi$ where $\varphi \in \mathscr{G}$ is defined as $\varphi(1/2) = \alpha(1)$, $\varphi(1/4) = \alpha(2)$, $\varphi(3/4) = \alpha(3), \ldots, \varphi(\frac{2s+1}{2^j}) = \alpha(2^{j-1} + s)$. In addition, let $\varphi(0) = 0$, $\Phi(0) = 0$, and $\Phi(1) = 1$.

Let T be a 2-Cantor tree, i.e., $T = \bigcup_{k=1}^{\infty}\{0,1\}^k$.

To define f, let $d_\sigma = \sum_{i=1}^{|\sigma|} \frac{\sigma_i}{2^i}$ be the dyadic rational associated to σ. For $\sigma \in \{0,1\}^k$, define

$$f(\sigma) := \left(\Phi(d_\sigma) + \varphi(d_\sigma), \Phi(d_\sigma + \frac{1}{2^k})\right).$$

In this way, $\{\overline{f(\sigma)} : \sigma \in T, |\sigma| = k\}$ is the union of 2^k consecutive disjoint intervals; hence f clearly satisfies the 3-intersection property. Further,

$$K_\alpha = \bigcap_{k=1}^{\infty} \bigcup_{\{\sigma \in T, |\sigma|=k\}} \overline{f(\sigma)} = \bigcup_{\sigma \in [T]} \bigcap_{n=1}^{\infty} \overline{f(\sigma|n)}.$$

In addition, since α is decreasing, $\text{diam}(f(\sigma))$, $|\sigma| = k$ is at least the diameter of $f(\sigma)$ for any σ with $|\sigma| = k+1$, and therefore f is a regular T-assignment.

The previous theorem showed that there is in fact a very large class of Cantor sets that are $\mathscr{H}$-visible. We will now state a lemma that will allow us to conclude, that in fact, the set of Cantor sets that are $\mathscr{H}$-visible is dense in $\mathscr{C}(\mathbb{R})$.

Lemma 3.4. *Any collection $\mathscr{L}$ of compact subsets of $\mathbb{R}$ satisfying the following properties is dense in the set of all compact subsets of $\mathbb{R}$:*

1. *$\mathscr{L}$ contains sets with arbitrarily small diameters.*
2. *If $A \in \mathscr{L}$ and $A_1, \ldots, A_n$ are translates of A, then $\cup_{i=1}^{n} A_i \in \mathscr{L}$.*

Corollary 3.5. *The collection of $\mathscr{H}$-visible compact subset of $\mathbb{R}$ is dense in $\mathscr{C}(R)$.*

Proof. We first recall that the standard middle third Cantor set is $\mathscr{H}$-visible just as any clopen subset of it. Further, any translation of any of these sets is also $\mathscr{H}$-visible. Using Lemma 3.4, we have the result.

3.2 Generic Visibility

We next show that a generic compact subset of $\mathbb{R}$ is visible.

We need some lemmas in order to prove the theorem.

Lemma 3.6. *If $B \subset \mathbb{R}$ is a set that is linearly independent over the rationals and $\alpha \in \mathbb{R}, \alpha \neq 0$, then $(B+\alpha) \cap B$ contains at most one point.*

Proof. To obtain a contradiction, assume that $\alpha \neq 0$, $x_1 \neq x_2$ are points in B such that $y_1 = x_1 + \alpha$, $y_2 = x_2 + \alpha \in B$. Clearly, $x_1 \neq x_2$, $y_1 \neq y_2$, $y_1 \neq x_1$ and $y_2 \neq x_2$. If $x_1 = y_2$, then we have that $\alpha = x_1 - x_2$, which leads to $y_1 = x_1 + x_1 - x_2$, contradicting that B is linearly independent over the rationals. An analogous argument shows that $x_2 \neq y_1$. Hence, we have that all of x_1, x_2, y_1, y_2 are distinct. However, this implies that $y_1 - x_1 - y_2 + x_1 = 0$, contradicting that B is linearly independent over the rationals.

Lemma 3.7. *If K is an uncountable, compact set that is linearly independent over the rationals, then there exists a translation invariant Borel measure μ such that $0 < \mu(K) < +\infty$.*

Proof. Let ν be any nonatomic Borel measure such that $\nu(K) = 1$. Let $\mathscr{B}_K = \{B \subset K : B \text{ is open relative to } K\}$, and $\mathscr{C} := \{B + t,\ B \in \mathscr{B}_K,\ t \in \mathbb{R}\} \cup \{\mathbb{R}\}$ and let $P : \mathscr{C} \longrightarrow \mathbb{R}$ be a set function defined as

$$\begin{aligned} P(\emptyset) &= 0, \\ P(B+t) &= \nu(B) \quad B \in \mathscr{B}_K, t \in \mathbb{R}, \text{ and} \\ P(\mathbb{R}) &= +\infty. \end{aligned}$$

P is a premeasure and we use Method II [8] to construct the sought-after measure μ:

$$\mu(A) = \lim_{\delta \to 0} \mu_\delta(A) \quad \text{where} \quad \mu_\delta(A) = \inf\{\sum P(C_i) : \operatorname{diam}(C_i) \leq \delta,\ C_i \in \mathscr{C},\ \cup_i C_i \supset A\}.$$

Since we use Method II, we know that μ is Borel and metric. We need to show that

- μ is translation invariant.
- $0 < \mu(K) < +\infty$.

The first part is a direct consequence of the definition of P.

For the second, it is clear that $\mu(K) \leq 1$.

For the other inequality, let $\{C_i\} \subseteq \mathscr{C}$ be a covering of K with $\operatorname{diam}(C_i) \leq \delta, C_i = B_i + t_i$. Since K is compact,

$$K \subseteq (B_1 + \alpha_1) \cup (B_2 + \alpha_2) \cup \cdots \cup (B_n + \alpha_n), \text{ where } (B_i + \alpha_i) \cap K \neq \emptyset.$$

By the linear independence over the rationals of K, if $\alpha_i \neq 0$, $(B_i + \alpha_i) \cap K$ contains exactly one point and we call it x_i. Therefore, we can find $j_1, \ldots j_k$ for which $C_{j_i} = B_{j_i}$ and

$$K \setminus \{x_{i_1}, \ldots, x_{i_l}\} \subseteq B_{j_1} \cup B_{j_2} \cup \cdots \cup B_{j_k}.$$

Then

$$\sum_{s=1}^{k} P(B_{j_s}) = \sum_{s=1}^{k} \nu(B_{j_s}) \geq \nu(\cup_{s=1}^{k} B_{j_s})$$

since ν is a Borel measure. But

$$\nu(\cup_{s=1}^{k} B_{j_s}) \geq \nu(K \setminus \{x_{i_1}, \ldots, x_{i_l}\}) = \nu(K) = 1.$$

This yields the desired result.

We are now ready for the main theorem of this section.

Theorem 3.8. *A generic compact subset of* $\mathbb{R}$ *is visible.*

Proof. Now the proof follows from Lemma 3.7 and the fact that a generic compact subset of $\mathbb{R}$ is uncountable and it follows from Theorem 19.1 [7] that it is linearly independent over the rationals.

4 Davies Sets in Polish Abelian Groups

In this section we extract out key ideas from the construction of Davies [5] to give a general procedure for constructing compact strongly invisible sets in an abelian Polish group.

Let G be an abelian Polish group and $\{t_k\}$ be a sequence of distinct points in G such that $\sum_{k=1}^{\infty} d(0, t_k) < \infty$. For each $\sigma \in \{0,1\}^{\mathbb{N}}$ let $p(\sigma) = \sum_{k=1}^{\infty} \sigma(k) t_k$.

We say that sequence $\{t_k\}$ **is good** if

$$\left(p(\sigma) = p(\tau),\ \sigma, \tau \in \{0,1\}^{\mathbb{N}}\right) \implies (\sigma = \tau),$$

i.e., each element of $p(\{0,1\}^{\mathbb{N}})$ has a unique expansion in terms of $\{t_k\}$.

Lemma 4.1. *Let G be an uncountable, abelian Polish group. Then G has a good sequence.*

Proof. Let $\{t_k\}$ be any sequence in G such that for all $k \geq 1$ we have that

$$\sum_{j=k+1}^{\infty} d(0, t_j) < \frac{1}{2} d(0, t_k).$$

Then, this $\{t_k\}$ has the desired property.

Theorem 4.2. *Every uncountable abelian Polish group contains a strongly invisible compact set.*

Proof. Let G be an uncountable abelian Polish group, $\{t_k\}$ a good sequence in G constructed as in Lemma 4.1 and as before for each $\sigma \in \{0,1\}^{\mathbb{N}}$ let $p(\sigma) = \sum_{k=1}^{\infty} \sigma(k) t_k$.

Moreover, assume that $\{A_k : k = 0, 1, \ldots\}$ is a decreasing sequence of subsets of $\mathbb{N}$ such that $A_0 = \mathbb{N}$ and $A_k \setminus A_{k+1}$ is infinite for all $k = 0, 1, \ldots$.

Finally, define for $k \geq 1$,

$$B_k = \{\sigma \in \{0,1\}^{\mathbb{N}} : \sigma(i) = 0 \text{ if } i < k \text{ or } i \in A_k\}.$$

and

$$C_k = \{p(\sigma) : \sigma \in B_k\}.$$

The properties of the sequence $\{t_k\}$ imply that $\sum_{k=1}^{\infty} \operatorname{diam}(C_k) < \infty$.

We consider now a sequence $\{x_n\}$ of distinct points in G converging to zero, and a sequence of pairwise disjoint balls $\{B(x_n)\}$ centered at x_n.

Since the diameters of the sets C_k go to zero, it is possible to find an increasing sequence $\{n_k\}$ of positive integers, such that $C_{n_k} + l_k$ is included in the ball $B(x_k)$ for some translation $l_k \in G$. Define C to be the union $\bigcup_k (C_{n_k} + l_k)$ plus the element zero. So, C is compact.

We will see that the set C has the required properties.

For $k, l \in \mathbb{N}$, denote by $B^0_{k,l} = \{\sigma \in B_k : \sigma(k) = \cdots = \sigma(k + l - 1) = 0\}$ and by $C^0_{k,l} = \{p(\sigma) : \sigma \in B^0_{k,l}\}$.

Then C_k is a finite union of disjoint translates of $C^0_{k,l}$ since

$$C_k = \bigcup_u \left(C^0_{k,l} + u\right)$$

where $u \in \left\{\sum_{i=0}^{l-1} \alpha_i t_{k+i} : \alpha_i \in \{0,1\},\ \alpha_i = 0 \text{ if } k + i \in A_k\right\}$.

We want to see now that C_{k+l} is the disjoint union of uncountable translates of $C^0_{k,l}$. For this, define

$$D_{k,l} = \{\tau \in \{0,1\}^N : \tau(s) = 0 \text{ if } s < k + l \text{ or } s \notin A_k \setminus A_{k+l}\}$$

The set $D_{k,l}$ is uncountable and so is the set $\{p(\tau) : \tau \in D_{k,l}\}$ because the sequence $\{t_k\}$ is good.

It is easy to see now that the collection of translates $\{C^0_{k,l} + p(\tau) : \tau \in D_{k,l}\}$ is pairwise disjoint and that

$$C_{k+l} = \bigcup_{\tau \in D_{k,l}} \left(C^0_{k,l} + p(\tau)\right)$$

Now we are ready to see that C is strongly invisible.

Let μ be a translation invariant Borel measure. If $\mu(C)$ is not zero, then there exists $k \in \mathbb{N}$ such that $\mu(C_{n_k}) > 0$. Then $C^0_{n_k, n_{k+1}-n_k}$ has positive measure. It follows that $C_{n_{k+1}}$ is not $\sigma-$finite with respect to the measure μ. So, neither is C.

Using this construction, together with Lemma 3.4, we can show the following proposition.

Proposition 4.3. *The set of strongly invisible sets is dense in* $\mathscr{C}(\mathbb{R})$.

Proof. Following [5] we have that there are compact strongly invisible sets of arbitrary small diameters. Moreover, the finite union of translates of a fixed strongly invisible set is again strongly invisible. Using Lemma 3.4, we have the desired result.

Acknowledgments C. Cabrelli and U. Molter are partially supported by Grants UBACyT X638 and X502 (UBA) and PIP 112-200801-00398 (CONICET). U. Darji is partially supported by University of Louisville Project Initiation Grant.

References

1. Besicovitch, A.S., Taylor, S.J.: On the complementary intervals of a linear closed set of zero Lebesgue measure. J. London Math. Soc. **29**, 449–459 (1954)
2. Best, E.: A closed dimensionless linear set. Proc. Edinburgh Math. Soc. **2**(6), 105–108 (1939)
3. Cabrelli, C., Hare, K.E., Molter, U.M.: Classifying Cantor sets by their fractal dimensions. Proc. Amer. Math. Soc. **138**(11), 3965–3974 (2010)
4. Cabrelli, C., Mendivil, F., Molter, U.M., Shonkwiler, R.: On the h-Hausdorff measure of Cantor sets. Pac. J. Math. **217**, 29–43 (2004)
5. Davies, R.O.: Sets which are null or non-sigma-finite for every translation-invariant measure. Mathematika **18**, 161–162 (1971)
6. Elekes, M., Keleti, T.: Borel sets which are null or non-σ-finite for every translation invariant measure. Adv. Math. **201**(1), 102–115 (2006)
7. Kechris, A.S.: Descriptive Set Theory, Graduate Texts in Mathematics, vol. 156. Springer-Verlag New York (1995)
8. Rogers, C.A.: Hausdorff Measures, Cambridge Math Library. Cambridge University Press, Cambridge (1998)

Convolution Inequalities for Positive Borel Measures on $\mathbb{R}^d$ and Beurling Density

Jean-Pierre Gabardo

Abstract We consider certain convolution inequalities for positive Borel measures in Euclidean space and show how they are related to the notions of upper and lower-Beurling density for these measures. In particular, the upper-Beurling density of a measure μ is shown to be the infimum of the constants $C > 0$ such that $\mu * f \leq C$ a.e. on $\mathbb{R}^d$ for some nonnegative function f with $\int f(x)\,\mathrm{d}x = 1$, and a similar characterization is obtained for the lower-Beurling density of μ. We also consider convolution inequalities involving several measures and provide applications of these results to systems of windowed exponentials and Gabor systems.

Keywords Upper-Beurling density • Lower-Beurling density • Beurling density • Locally-finite measure • Translation-bounded measure • vector-valued measure • Convolution of positive measures • Dirac mass • Gabor frames

1 Introduction

Our main goal in this work is to establish a link between certain convolution inequalities for positive Borel measures in Euclidean space and corresponding notions of Beurling density associated with such measures. We show, for example, that if μ is a positive Borel measure on $\mathbb{R}^d$, if $f \geq 0$ is Lebesgue integrable on $\mathbb{R}^d$, and their convolution product satisfies $\mu * f \leq C$ almost everywhere on $\mathbb{R}^d$ for some constant $C > 0$, then we must have the inequality $\mathscr{D}^+(\mu) \int_{\mathbb{R}^d} f(x)\,dx \leq C$, where $\mathscr{D}^+(\mu)$ is the upper-Beurling density of μ and that the quantity $\mathscr{D}^+(\mu)$ is the best possible constant in this last inequality. Similarly, we will prove that if, in

J.-P. Gabardo (✉)
Department of Mathematics and Statistics, McMaster University,
Hamilton, ON, Canada, L8S 4K1,
e-mail: gabardo@mcmaster.ca

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_3,

addition, μ is translation-bounded and if $\mu * f \geq C$ almost everywhere on $\mathbb{R}^d$ for some constant $C > 0$, then $\mathscr{D}^-(\mu) \int_{\mathbb{R}^d} f(x)\,\mathrm{d}x \geq C$, where $\mathscr{D}^-(\mu)$, the lower-Beurling density of μ, is the best possible constant.

The fact that convolution inequalities of the type mentioned above give rise to some estimates on the corresponding Beurling densities of the associated measure is not new, and, in fact, particular cases of these results have been mentioned and used a number of times in the literature by various researchers. Examples of these situations can be found in the study of the packing and tiling properties of the translates of sets (or functions) in $\mathbb{R}^d$ (e.g., [2, 3]). In that case, the measure μ is simply a sum of Dirac masses $\mu = \sum_{\lambda\in\Lambda} \delta_\lambda$, where Λ corresponds to the set of translates involved in the packing or tiling. Other examples can be found in the theory of (weighted or unweighted) irregular Gabor systems (see e.g., [4, 5]). In that case, if such a system forms a Bessel collection or a frame for $\mathbb{R}^d$, one can deduce certain convolution inequalities for a related positive measure defined on the time-frequency space $\mathbb{R}^{2d}$. Using the results mentioned above, one can then obtain estimates on the Beurling density of the points associated with the time-frequency shifts of the window (or windows if there is more than one) involved in the system.

It is worth mentioning that in all the known results in the literature where these types of estimates appear, some restriction is made on the type of measure μ or the type of function f for which the convolution inequality is assumed to hold. The proof that the resulting inequality holds for the corresponding Beurling density is then dependent on these additional assumptions. This has the effect that the proofs are sometimes unnecessarily complicated and that the generality of the results is somewhat reduced. Hence, one of main tasks will be to establish the results mentioned above in full generality so that they can be applied in a systematic way to various situations.

This chapter is organized as follows. In Sect. 2, we define the convolution of (generally) unbounded positive Borel measures on $\mathbb{R}^d$ and prove that the result of the convolution of a positive Borel measure with a nonnegative locally integrable function is always given by a function (possibly taking the value ∞). In Sect. 3, we define the notion of translation-bounded measure and Beurling densities. We prove certain results relating these; in particular the fact that a measure μ is translation-bounded if and only if the upper-Beurling $\mathscr{D}^+(\mu)$ is finite. Sections 4 and 5 contain the main theorems of this chapter. Section 4 deals with the case of a convolution product with a single measure while Sect. 5 generalizes this situation to the case of a sum of convolution products involving several measures.

2 Convolution of Positive Borel Measures

The positive Borel measures on $\mathbb{R}^d$ that we will consider here will generally be unbounded, but they will be, unless otherwise specified, "locally finite" which means that the measure of any compact subset of $\mathbb{R}^d$ is finite. A major topic considered in this chapter will be the study of certain properties of the convolution

of such positive measures. The convolution of general unbounded complex-valued measures is not well defined in general, but this difficulty disappears when we deal with positive measures as we allow the measure of some sets to be infinite. We begin by recalling some known definitions and prove some elementary facts about the convolution of general positive Borel measures on $\mathbb{R}^d$. If μ and ν are (locally finite) positive Borel measures on $\mathbb{R}^d$, we can define their convolution product $\mu * \nu$ by the formula

$$(\mu * \nu)(E) = \int_{\mathbb{R}^{2d}} \chi_E(x+y)\, d(\mu \otimes \nu)(x,y),$$

for any Borel subset E of $\mathbb{R}^{\mathrm{d}}$, where $\mu \otimes \nu$ denotes the tensor product of the two measures. By the Fubini–Tonelli theorem, this last expression can be computed as

$$\begin{aligned}(\mu * \nu)(E) &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \chi_E(x+y)\, \mathrm{d}\nu(y) \right) \mathrm{d}\mu(x) \\ &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \chi_E(x+y)\ \mathrm{d}\mu(x) \right) \mathrm{d}\nu(y).\end{aligned}$$

Of course, the resulting measure, although still a Borel measure, might not be finite on compact sets even if μ and ν are both locally finite. It is easily checked that if μ, ν, and ρ are (locally finite) positive Borel measures on $\mathbb{R}^d$ and if $\mu * \nu$ and $\mu * \rho$ are locally finite as well, then $(\mu * \nu) * \rho = \mu * (\nu * \rho)$. The space of complex-valued (resp. locally) integrable functions on $\mathbb{R}^d$ will be denoted by $L^1(\mathbb{R}^d)$ (resp. $L^1_{\mathrm{loc}}(\mathbb{R}^d)$). Any Lebesgue measurable function f with $f \geq 0$ defines a Borel measure μ_f via the formula

$$\mu_f(E) = \int_E f(x)\, \mathrm{d}x, \quad E \text{ Borel subset of } \mathbb{R}^d.$$

We will use the notation $\mathrm{d}\mu_f = f(x)\mathrm{d}x$ to denote this measure. In the case where the convolution of two positive Borel measures μ, ν is of the form $d(\mu * \nu) = F(x)\, \mathrm{d}x$, for some nonnegative Lebesgue measurable function F, we write $\mu * \nu = F$, with a slight abuse of notation.

Lemma 1. *Let μ be a locally finite, positive Borel measure on $\mathbb{R}^d$ and let $f \in L^1_{loc}(\mathbb{R}^d)$ with $f \geq 0$. Then, there exists a Borel measurable function H on $\mathbb{R}^d$ with $0 \leq H(x) \leq \infty$ for $x \in \mathbb{R}^d$ such that $\mu * f = H$.*

Proof. Let B_N denote the closed ball centered at 0 with radius $N > 0$ in $\mathbb{R}^d$. Define $\mu_N = \mu\, \chi_{B_N}$ and $f_N = f\, \chi_{B_N}$ for any integer $N \geq 1$. Then, $\mu_N * f_N$ is a bounded Borel measure for any $N > 0$. If D is a Borel measurable subset of $\mathbb{R}^d$ with zero Lebesgue measure we have

$$(\mu_N * f_N)(D) = \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \chi_D(x+y)\, f_N(y)\, \mathrm{d}y \right) \mathrm{d}\mu_N(x) = 0.$$

By the Lebesgue–Radon–Nikodym theorem (see [7]) (applied to the situation where the sigma-finite measure in the theorem is the Lebesgue measure restricted to the Borel subsets of $\mathbb{R}^d$), there exists a Borel measurable (and integrable) function H_N on $\mathbb{R}^d$ with $H_N \geq 0$ such that

$$(\mu_N * f_N)(E) = \int_E H_N(x)\,\mathrm{d}x, \quad E \text{ Borel subset of } \mathbb{R}^d.$$

Clearly, the sequence $\{H_N\}_{N\geq 1}$ is pointwise increasing and $H(x) = \lim_{N\to\infty} H_N(x)$ is well defined as a Borel measurable function taking its values in $[0,\infty]$. Furthermore, the Lebesgue monotone convergence theorem shows that, for any Borel subset E of $\mathbb{R}^d$, we have

$$\lim_{N\to\infty} (\mu_N * f_N)(E) = \lim_{N\to\infty} \int_E H_N(x)\,\mathrm{d}x = \int_E H(x)\,\mathrm{d}x.$$

The same theorem shows also that

$$\begin{aligned}\lim_{N\to\infty} (\mu_N * f_N)(E) &= \lim_{N\to\infty} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \chi_E(x+y)\, f(y)\, \chi_{B_N}(x)\, \chi_{B_N}(y)\,\mathrm{d}(y) \right) \mathrm{d}\mu(x) \\ &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \chi_E(x+y)\, f(y)\,\mathrm{d}(y) \right) \mathrm{d}\mu(x) = (\mu * f)(E),\end{aligned}$$

which proves our claim. □

Note that, even in the case where μ is a bounded measure and f is integrable in the previous theorem, the convolution $\mu * f$ cannot, in general, be written in the form

$$(\mu * f)(x) = \int_{\mathbb{R}^d} f(x-y)\,\mathrm{d}\mu(y), \quad x \in \mathbb{R}^d, \tag{1}$$

as the integral does not even make sense if f is not Borel measurable. However, as the following lemma will show, the formula is correct when f is Borel measurable.

Lemma 2. *Let μ be a locally finite, positive Borel measure on $\mathbb{R}^d$. Suppose that $f \in L^1_{loc}(\mathbb{R}^d)$ is such that $f \geq 0$ and is Borel measurable. Then $\mu * f = H$ where H is given by (1).*

Proof. Since the function $(x,y) \mapsto \chi_E(x+y)\, f(y)$ is Borel measurable on $\mathbb{R}^{2d}$ for any Borel subset E of $\mathbb{R}^d$, the Fubini–Tonelli theorem yields

$$\begin{aligned}(\mu * \nu)(E) &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \chi_E(x+y)\, f(y)\,\mathrm{d}y \right) \mathrm{d}\mu(x) \\ &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \chi_E(y)\, f(y-x)\,\mathrm{d}y \right) \mathrm{d}\mu(x) = \int_E \left(\int_{\mathbb{R}^d} f(y-x)\,\mathrm{d}\mu(x) \right) \mathrm{d}y,\end{aligned}$$

which proves the lemma. □

Since every Lebesgue measurable function is equal almost everywhere to a Borel measurable function (see [1, Section 21]), the function H in Lemma 1 can be explicitly computed using formula (1) by simply choosing a Borel measurable representative of $f \in L^1_{\mathrm{loc}}(\mathbb{R}^d)$.

3 Translation-Bounded Measures and Beurling Densities

We will denote by $|E|$ the Lebesgue measure of a measurable subset E of $\mathbb{R}^d$ and by $E^{\complement}$, the complement of E, i.e., the set $\mathbb{R}^d \setminus E$. For $z = (z_1, \ldots, z_m) \in \mathbb{R}^d$, we define the d-dimensional box of side length $R > 0$ centered at z to be the set

$$I_R(z) = \{y = (y_1, \ldots, y_m) \in \mathbb{R}^d,\ |z_i - y_i| \leq R/2,\ i = 1, \ldots, m\}.$$

For simplicity, we will write I_R for $I_R(0)$. The notion of Beurling density plays an important role in many areas of modern Fourier analysis. One of the first use of the concept of Beurling density for discrete subsets of $\mathbb{R}^d$ appeared in the paper by H. Landau ([6]) studying sampling and interpolation in spaces of band-limited functions. More recently, many researchers have used this notion, particularly in the study of Gabor frames (e.g., [4, 5]). If μ is a positive Borel measure on $\mathbb{R}^d$, the quantities

$$\mathscr{D}^+(\mu) = \limsup_{R\to\infty} \sup_{z\in\mathbb{R}^d} \frac{\mu\,(I_R(z))}{R^d} \quad \text{and} \quad \mathscr{D}^-(\mu) = \liminf_{R\to\infty} \inf_{z\in\mathbb{R}^d} \frac{\mu\,(I_R(z))}{R^d}$$

are called the *upper and lower Beurling density* of the measure μ, respectively. If both these densities are equal and finite, we say that the Beurling density of the measure μ exists, and we define it to be the quantity $\mathscr{D}(\mu) := \mathscr{D}^+(\mu) = \mathscr{D}^-(\mu)$. If $\Lambda \subset \mathbb{R}^d$ is a discrete set, the corresponding Beurling densities $\mathscr{D}^+(\Lambda)$, $\mathscr{D}^-(\Lambda)$ and $\mathscr{D}(\Lambda)$ are defined as the corresponding Beurling density of the measure $\mu = \sum_{\lambda\in\Lambda} \delta_\lambda$, where δ_λ is the Dirac mass at λ. A positive Borel measure μ on $\mathbb{R}^d$ is called *translation-bounded* if, for every compact $K \subset \mathbb{R}^d$, there exists a constant $C_\mu(K) \geq 0$ such that

$$\mu\,(K + z) \leq C_\mu(K), \quad z \in \mathbb{R}^d,$$

where $K + z = \{z' + z,\ z' \in K\}$. Clearly, μ will be translation-bounded if the inequality above holds for just one compact set K with nonempty interior. We will need the following lemma.

Lemma 3. *Let μ be a translation-bounded Borel measure on $\mathbb{R}^d$. Then, for any rectangular box*

$$B = [a_1, b_1] \times \cdots \times [a_d, b_d]$$

contained in $\mathbb{R}^d$ *we have*

$$\mu(B) \leq C_\mu(I) \prod_{i=1}^{d} [(b_i - a_i) + 1]. \tag{2}$$

Proof. Let $I = [0,1]^d$ be the unit hypercube in $\mathbb{R}^d$. Fix a box B as above and, for each $i = 1, \ldots, d$, choose an integer k_i such that $a_i + k_i < b_i \leq a_i + k_i + 1$. Then,

$$B \subset [a_1, a_1+k_1+1] \times \cdots \times [a_d, a_d+k_d+1] = \bigcup_{r_1=0}^{k_1} \cdots \bigcup_{r_d=0}^{k_d} I + (a_1+r_1, \ldots, a_d+r_d).$$

Hence,

$$\mu(B) \leq C_\mu(I) \prod_{j=1}^{d} (k_j + 1) \leq C_\mu(I) \prod_{j=1}^{d} [(b_j - a_j) + 1],$$

which proves the lemma. □

Corollary 1. *Let* μ *be a translation-bounded, positive Borel measure on* $\mathbb{R}^d$. *Then, there exists a constant* C *depending only on* μ *such that*

$$\mu(I_R(z)) \leq C\,(1 + R^d), \quad R > 0,\ z \in \mathbb{R}^d. \tag{3}$$

Proof. Let $I = [0,1]^d$ be the unit hypercube in $\mathbb{R}^d$. Fix $z \in \mathbb{R}^d$. If $0 < R < 1$, we have the inclusion $I_R(z) \subset I + w$ where $w_i := z_i - 1/2$ for $i = 1, \ldots, d$. Hence, $\mu(I_R(z)) \leq C_\mu(I)$. If $R \geq 1$, the previous lemma shows that

$$\mu(I_R(z)) \leq C_\mu(I)\,(1 + R)^d \leq C_\mu(I)\,2^d\,(1 + R^d),$$

which proves our claim. □

Corollary 2. *Let* μ *be a positive Borel measure on* $\mathbb{R}^d$. *Then,* μ *is translation-bounded if and only if* $\mathscr{D}^+(\mu) < \infty$.

Proof. If μ is translation-bounded, then $\mathscr{D}^+(\mu) < \infty$ by Corollary 1. Conversely, given any compact set K, choose $R_0 > 0$ and $z_0 \in \mathbb{R}^d$ such that $K \subset I_{R_0}(z_0)$. By definition of $\mathscr{D}^+(\mu)$, we can find $R_1 > R_0$ such that $\mu(I_{R_1}(z)) \leq (\mathscr{D}^+(\mu) + 1)\,R_1^d$ for all $z \in \mathbb{R}^d$ which implies that

$$\mu(K + z) \leq (\mathscr{D}^+(\mu) + 1)\,R_1^d, \quad z \in \mathbb{R}^d.$$

Hence, μ is translation-bounded. □

The following lemma will play a crucial role to obtain certain estimates in the following section.

Lemma 4. *Let μ be a translation-bounded, positive Borel measure on $\mathbb{R}^d$. Suppose that $M > N \geq 0$. Then, for any $z \in \mathbb{R}^d$, we have*

$$|\mu(I_M(z) + y) - \mu(I_M(z))| \leq d\, C_\mu(I)\, (M+1)^{d-1}\, (N+2)/2, \quad \textit{for all } y \in I_N. \tag{4}$$

Proof. We have

$$\mu(I_M(z) + y) \geq \mu\left((I_M(z) + y) \cap I_M(z)\right) = \mu(I_M(z)) - \mu\left((I_M(z)) \cap (I_M(z) + y)^{\complement}\right)$$

and

$$(I_M(z) + y)^{\complement} = \bigcup_{j=1}^{d} R_j,$$

where

$$R_j = \{w = (w_1, \ldots, w_d) \in \mathbb{R}^d,\ z_j + y_j - w_j > M/2 \text{ or } z_j + y_j - w_j < -M/2\}.$$

Note that the set $I_M(z) \cap R_j$ is equal to

$$\{w \in \mathbb{R}^d,\ |z_i - w_i| \leq M/2,\ i = 1, \ldots, d,\ i \neq j \text{ and } M/2 - y_j < z_j - w_j \leq M/2\},$$

if $y_j \geq 0$, and to

$$\{w \in \mathbb{R}^d,\ |z_i - w_i| \leq M/2,\ i = 1, \ldots, d,\ i \neq j \\ \text{and } -M/2 < z_j - w_j \leq -M/2 - y_j\},$$

if $y_j \leq 0$. Using Lemma 3, we obtain thus that

$$\mu\left(I_M(z) \cap R_j\right) \leq C_\mu(I)\, (M+1)^{d-1}\, (|y_j| + 1) \leq C_\mu(I)\, (M+1)^{d-1}\, (N+2)/2.$$

Hence,

$$\mu\left(I_M(z) \cap (I_M(z) + y)^{\complement}\right) = \mu\left(\cup_{j=1}^{d} (I_M(z) \cap R_j)\right)$$

$$\leq \sum_{j=1}^{d} \mu\left(I_M(z) \cap R_j\right) \leq d\, C_\mu(I)\, (M+1)^{d-1}\, (N+2)/2$$

and

$$\mu(I_M(z) + y) - \mu(I_M(z)) \geq -d\, C_\mu(I)\, (M+1)^{d-1}\, (N+2)/2. \tag{5}$$

We have also,

$$\mu(I_M(z)) - \mu(I_M(z) + y) \geq -d\, C_\mu(I)\, (M+1)^{d-1}\, (N+2)/2,$$

using the previous estimate (5) with z replaced by $z+y$ and y by $-y$. This, together with Eq. (5), yields the required inequality. □

4 Convolution Inequalities and Beurling Densities

In this section, we consider the relationship between convolution inequalities involving a single positive Borel measure on $\mathbb{R}^d$ and the corresponding Beurling densities. We start by proving that the translation-bounded measures on $\mathbb{R}^d$ are exactly those measures whose convolution product with some nonnegative integrable function give rise to a bounded function on $\mathbb{R}^d$. To simplify the notations, we introduce the set

$$\mathscr{P}(\mathbb{R}^d) = \{f \in L^1(\mathbb{R}^d) \text{ with } f \geq 0 \text{ and } \int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = 1\}.$$

Proposition 1. *Let μ be a positive Borel measure on $\mathbb{R}^d$. Then, the following are equivalent:*

(a) μ is translation-bounded.
(b) There exist $f \in \mathscr{P}(\mathbb{R}^d)$ and a constant $C > 0$ such that $\mu * f \leq C$ a.e. on $\mathbb{R}^d$.

Proof. If μ is translation-bounded, let $f = \chi_{I_1}$. Then, $\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = 1$ and

$$\mu * f(x) = \int_{\mathbb{R}^d} \chi_{I_1}(x-y)\,\mathrm{d}\mu(y) = \mu(I_1 + x) \leq C_\mu(I_1), \quad x \in \mathbb{R}^d,$$

showing that (b) holds. Conversely, if there exists a function $f \in \mathscr{P}(\mathbb{R}^d)$ satisfying (b), choose a bounded measurable set E such that $0 < a \leq f < \infty$ on E where a is a positive constant. We have then $\mu * a\,\chi_E \leq C$ a.e. on $\mathbb{R}^d$. Letting $\check{E} = \{-x,\ x \in E\}$, we have thus

$$\mu * a\,\chi_E * \chi_{\check{E}} \leq C\,|E|, \quad \text{a.e. on } \mathbb{R}^d.$$

Since $\chi_E \in L^2(\mathbb{R}^d)$ and

$$\left(\chi_E * \chi_{\check{E}}\right)(x) = \int_{\mathbb{R}^d} \chi_E(x+y)\,\chi_E(y)\,\mathrm{d}y, \quad x \in \mathbb{R}^d,$$

the function $g := \chi_E * \chi_{\check{E}}$ is continuous and compactly supported. Furthermore $g(0) = |E|^2 > 0$. Therefore, $g > c > 0$ on some set I_r, for $r > 0$ and $c > 0$ small enough. Letting $C_1 = a^{-1}\,C\,|E|$, we have

$$\int_{\mathbb{R}^d} g(x-y)\,\mathrm{d}\mu(y) \leq C_1, \quad \text{for a.e. } x \in \mathbb{R}^d.$$

Since the left-hand side of the previous inequality is continuous, this inequality must hold for all $x \in \mathbb{R}^d$. Hence,

$$\begin{aligned}\mu(x + I_r) &\le c^{-1} \int_{x+I_r} g(x-y)\,\mathrm{d}\mu(y)\\ &\le c^{-1} \int_{\mathbb{R}^d} g(x-y)\,\mathrm{d}\mu(y) \le c^{-1}\, C_1, \quad x \in \mathbb{R}^d,\end{aligned}$$

which shows that μ is translation-bounded. □

Definition 1. Let $\mathscr{K}_d$ denotes the set of compact subsets of $\mathbb{R}^d$ with nonzero Lebesgue measure. Let μ be a positive Borel measure on $\mathbb{R}^d$. We define

$$\mathscr{E}^+(\mu) = \lim_{N\to\infty} \sup_{K\in\mathscr{K}_d} \inf_{x\in I_N} \frac{\mu(x+K)}{|K|} \quad \text{and}$$

$$\mathscr{E}^-(\mu) = \lim_{N\to\infty} \inf_{K\in\mathscr{K}_d} \sup_{x\in I_N} \frac{\mu(x+K)}{|K|}.$$

Proposition 2. *Let μ be a positive Borel measure on $\mathbb{R}^d$. Then, we have the inequalities*

$$\mathscr{E}^+(\mu) \le \mathscr{D}^+(\mu) \quad \textit{and} \quad \mathscr{E}^-(\mu) \ge \mathscr{D}^-(\mu). \tag{6}$$

Furthermore, if μ is translation-bounded, we have the equalities

$$\mathscr{E}^+(\mu) = \mathscr{D}^+(\mu), \quad \mathscr{E}^-(\mu) = \mathscr{D}^-(\mu), \tag{7}$$

and the collection of compact sets $\mathscr{K}_d$ in the definition of $\mathscr{E}^+(\mu)$ and $\mathscr{E}^-(\mu)$ can be replaced by the collection of translates of sets I_R, $R \ge 1$.

Proof. We will start by proving the inequalities in Eq. (6). Let us first assume that $\mathscr{E}^+(\mu) < \infty$. Fix $\epsilon > 0$ and let $\{K_N\}_{N\ge1}$ be a sequence in $\mathscr{K}_d$ such that

$$\mathscr{E}^+(\mu) - \epsilon \le \inf_{x\in I_N} \frac{\mu(x+K_N)}{|K_N|}, \quad N \ge 1.$$

We have then

$$\frac{1}{|I_N|} \int_{I_N} \frac{\mu(x+K_N)}{|K_N|}\,\mathrm{d}x \ge \mathscr{E}^+(\mu) - \epsilon, \quad N \ge 1.$$

Since, by Fubini's theorem,

$$\begin{aligned}\int_{I_N} \mu(x+K_N)\,\mathrm{d}x &= \int_{\mathbb{R}^d} \chi_{I_N}(x)\left(\int_{\mathbb{R}^d} \chi_{K_N}(y-x)\,\mathrm{d}\mu(y)\right)\mathrm{d}x\\ &= \int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} \chi_{K_N}(y-x)\,\chi_{I_N}(x)\,\mathrm{d}x\right)\mathrm{d}\mu(y)\\ &= \int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} \chi_{I_N}(y-x)\,\chi_{K_N}(x)\,\mathrm{d}x\right)\mathrm{d}\mu(y)\\ &= \int_{\mathbb{R}^d} \chi_{K_N}(x)\left(\int_{\mathbb{R}^d} \chi_{I_N}(y-x)\,\mathrm{d}\mu(y)\right)\mathrm{d}x = \int_{K_N} \mu(x+I_N)\,\mathrm{d}x,\end{aligned}$$

it follows that

$$\frac{1}{|K_N|}\int_{K_N} \frac{\mu(x+I_N)}{|I_N|}\,\mathrm{d}x \geq \mathscr{E}^+(\mu)-\epsilon, \quad N\geq 1.$$

In particular, for every $N \geq 1$, there must exist $x_N \in K_N$ such that

$$\frac{\mu(x_N+I_N)}{|I_N|} \geq \mathscr{E}^+(\mu)-\epsilon.$$

This shows that $\mathscr{D}^+(\mu) \geq \mathscr{E}^+(\mu)-\epsilon$ and, since $\epsilon > 0$ is arbitrary, that $\mathscr{D}^+(\mu) \geq \mathscr{E}^+(\mu)$. If $\mathscr{E}^+(\mu)=\infty$, we can find a sequence $\{K_N\}_{N\geq 1}$ in $\mathscr{K}_d$ such that

$$\inf_{x\in I_N} \frac{\mu(x+K_N)}{|K_N|} \to \infty, \quad N\to\infty,$$

and the same computation as above shows that $\mathscr{D}^+(\mu)=\infty=\mathscr{E}^+(\mu)$. Thus the first inequality in Eq. (6) holds. If $\mathscr{E}^-(\mu)=\infty$, the second inequality in Eq. (6) is obvious. We can thus assume that $\mathscr{E}^-(\mu)<\infty$. Fix $\epsilon>0$ and let $\{K_N\}_{N\geq 1}$ be a sequence in $\mathscr{K}_d$ such that

$$\sup_{x\in I_N} \frac{\mu(x+K_N)}{|K_N|} \leq \mathscr{E}^-(\mu)+\epsilon, \quad N\geq 1.$$

We have then

$$\frac{1}{|I_N|}\int_{I_N} \frac{\mu(x+K_N)}{|K_N|}\,\mathrm{d}x \leq \mathscr{E}^-(\mu)+\epsilon, \quad N\geq 1,$$

which implies, as before, that

$$\frac{1}{|K_N|}\int_{K_N} \frac{\mu(x+I_N)}{|I_N|}\,\mathrm{d}x \leq \mathscr{E}^-(\mu)+\epsilon, \quad N\geq 1.$$

and thus, since $\epsilon > 0$ is arbitrary, that $\mathscr{D}^-(\mu) \leq \mathscr{E}^-(\mu)$, proving thus the second inequality in Eq. (6). Let us now assume that μ is translation-bounded. Using Corollary 1, it follows that $\mathscr{D}^+(\mu) < \infty$. Fix $\epsilon > 0$. Using the definition of $\mathscr{D}^+(\mu)$, we can find an increasing sequence of positive numbers r_j with $\lim_{j\to\infty} r_j = \infty$ and corresponding points $x_j \in \mathbb{R}^d$ such that

$$\frac{\mu\left(x_j + I_{r_j}\right)}{r_j^d} \geq \mathscr{D}^+(\mu) - \epsilon/2.$$

Let $N \geq 1$ be an integer. We claim that we can find an integer J_0 such that,

$$\frac{\mu\left(x_j + I_{r_j} + x\right)}{r_j^d} \geq \frac{\mu\left(x_j + I_{r_j}\right)}{r_j^d} - \epsilon/2$$

for all $x \in I_N$ and all $j \geq J_0$. Indeed, using Lemma 4, we have, for $r_j > N$ and $x \in I_N$, that

$$\left|\frac{\mu\left(x_j + I_{r_j} + x\right) - \mu\left(x_j + I_{r_j}\right)}{r_j^d}\right| \leq \frac{d\, C_\mu(I)\,(r_j+1)^{d-1}\,(N+2)/2}{r_j^d} < \epsilon/2$$

if j is large enough. Hence, we have

$$\inf_{x\in I_N} \frac{\mu\left(x_j + I_{r_j} + x\right)}{r_j^d} \geq \mathscr{D}^+(\mu) - \epsilon,$$

which shows that

$$\sup_{K\in\mathscr{K}_d} \inf_{x\in I_N} \frac{\mu(x+K)}{|K|} \geq \mathscr{D}^+(\mu) - \epsilon,$$

and thus that $\mathscr{E}^+(\mu) \geq \mathscr{D}^+(\mu)$, since $\epsilon > 0$ is arbitrary. Using Eq. (6), this proves the first equality in Eq. (7). Since $\mathscr{E}^+(\mu) = \mathscr{D}^+(\mu)$, the previous computation also shows that the supremum over all compact sets in $\mathscr{K}_d$ in the definition of $\mathscr{E}^+(\mu)$ is the same as the one over the smaller collection of all translates of the sets I_R, $R \geq 1$. Similarly, using the definition of $\mathscr{D}^-(\mu)$ which is finite (since $\mathscr{D}^-(\mu) \leq \mathscr{D}^+(\mu)$ and $\mathscr{D}^+(\mu) < \infty$), we can find, for any $\epsilon > 0$, an increasing sequence of positive numbers s_j with $\lim_{j\to\infty} s_j = \infty$ and corresponding points $y_j \in \mathbb{R}^m$ such that

$$\frac{\mu\left(y_j + I_{s_j}\right)}{s_j^d} \leq \mathscr{D}^-(\mu) + \epsilon/2.$$

Let $N \geq 1$ be an integer. As before, we can find an integer J_0 such that,

$$\frac{\mu\left(y_j + I_{s_j} + x\right)}{s_j^d} \leq \frac{\mu\left(y_j + I_{s_j}\right)}{s_j^d} + \epsilon/2$$

for all $x \in I_N$ and all $j \geq J_0$. Hence, we have

$$\sup_{x \in I_N} \frac{\mu\left(y_j + I_{s_j} + x\right)}{s_j^d} \leq \mathscr{D}^-(\mu) + \epsilon,$$

which shows that

$$\inf_{K \in \mathscr{K}_d} \sup_{x \in I_N} \frac{\mu(x + K)}{|K|} \leq \mathscr{D}^-(\mu) + \epsilon,$$

and thus that $\mathscr{E}^-(\mu) \leq \mathscr{D}^-(\mu)$, since $\epsilon > 0$ is arbitrary. Using Eq. (6), this shows that the second equality in Eq. (7) holds. Since $\mathscr{E}^-(\mu) = \mathscr{D}^-(\mu)$, the previous computation also shows that the infimum over all compact sets in $\mathscr{K}_d$ in the definition of $\mathscr{E}^+(\mu)$ is the same as the one over the smaller collection of all translates of the sets I_R, $R \geq 1$. This concludes the proof. □

Let us introduce the following definitions. Let μ be a positive Borel measure on $\mathbb{R}^d$. We define

$$\mathscr{C}^+(\mu) = \inf\left\{C \geq 0,\ \mu * f \leq C \text{ a.e. for some } f \in \mathscr{P}(\mathbb{R}^d)\right\},$$

with the convention that $\mathscr{C}^+(\mu) = \infty$ if the set where the infimum above is taken happens to be empty. Similarly,

$$\mathscr{C}^-(\mu) = \sup\left\{D \geq 0,\ \mu * f \geq D \text{ a.e. for some } f \in \mathscr{P}(\mathbb{R}^d)\right\}.$$

Note that the fact that the inequalities appearing in the previous definitions must hold "almost everywhere" as opposed to "everywhere" is unimportant. For example, if $\mu * f \leq C$ a.e. on $\mathbb{R}^d$, we can replace f with $f * h$, where $h \geq 0$ is continuous, compactly supported, and with integral equal to 1, to obtain a corresponding inequality which now holds everywhere on $\mathbb{R}^d$. For the same reason, one can always restrict the functions f to be continuous or even smoother in our definition of $\mathscr{C}^+(\mu)$ or $\mathscr{C}^-(\mu)$ without affecting those values.

Proposition 3. *Let μ be a positive Borel measure on $\mathbb{R}^d$. Then,*

(a) $\mathscr{C}^+(\mu) < \infty$ if and only if μ is translation-bounded.
(b) $\mathscr{C}^-(\mu) \leq \mathscr{C}^+(\mu)$.

Proof. The statement in (a) follows directly from Proposition 1. In order to prove (b), we can assume that $\mathscr{C}^+(\mu) < \infty$. Given $\epsilon > 0$, choose $f, g \in \mathscr{P}(\mathbb{R}^d)$ with

$$\mu * f \leq \mathscr{C}^+(\mu) + \epsilon.$$

If $D \geq 0$ is a constant such that $\mu * g \geq D$ a.e. we have

$$D = D * f \leq \mu * g * f = \mu * f * g \leq \mathscr{C}^+(\mu) + \epsilon$$

which implies that $\mathscr{C}^-(\mu) \leq \mathscr{C}^+(\mu)$ since $\epsilon > 0$ is arbitrary. □

We can now state the main result in this section.

Theorem 1. *Let μ be a positive Borel measure on $\mathbb{R}^d$. Then,*

(a) $\mathscr{C}^+(\mu) = \mathscr{D}^+(\mu)$.
(b) $\mathscr{C}^-(\mu) \geq \mathscr{D}^-(\mu)$ *and if, in addition, μ is translation-bounded, then we have the equality* $\mathscr{C}^-(\mu) = \mathscr{D}^-(\mu)$.

Proof. We first prove (a). If μ is not translation-bounded, we have $\mathscr{D}^+(\mu) = \infty$ by Corollary 2 and $\mathscr{C}^+(\mu) = \infty$ by part (a) of Proposition 3. We can thus assume that μ is translation-bounded and thus that both $\mathscr{C}^+(\mu)$ and $\mathscr{D}^+(\mu)$ are finite. Given $\epsilon > 0$, we can find $R > 0$ such that

$$\frac{\mu(I_R(z))}{R^d} \leq \mathscr{D}^+(\mu) + \epsilon, \quad z \in \mathbb{R}^d. \tag{8}$$

Letting $f = R^{-d}\, \chi_{I_R}$, we have $f \in L^1(\mathbb{R}^d)$, $f \geq 0$, $\int_{\mathbb{R}^d} f(x)\, \mathrm{d}x = 1$ and

$$(\mu * f)(z) = \frac{\mu(I_R(z))}{R^d}, \quad z \in \mathbb{R}^d.$$

Using Eq. (8) and the definition of $\mathscr{C}^+(\mu)$, we deduce that $\mathscr{C}^+(\mu) \leq \mathscr{D}^+(\mu)$, since $\epsilon > 0$ is arbitrary. To prove the converse inequality, we use the fact that $\mathscr{D}^+(\mu) = \mathscr{E}^+(\mu)$ by Eq. (7) of Proposition 2. Given $\epsilon > 0$, we can find a sequence of compact set $\{K_N\}_{N \geq 1}$ with $|K_N| > 0$ such that

$$\frac{\mu(x + K_N)}{|K_N|} \geq \mathscr{E}^+(\mu) - \epsilon, \quad x \in I_N.$$

Suppose now that $C > 0$ is a constant such that $\mu * f \leq C$ a.e. for some function $f \in L^1(\mathbb{R}^d)$ with $f \geq 0$ and $\int_{\mathbb{R}^d} f(x)\, \mathrm{d}x = 1$. Define $h_N = |K_N|^{-1}\, \chi_{\check{K}_N} * \mu$, where $\check{K}_N = \{-x,\ x \in K_N\}$. Then,

$$h_N(x) = \frac{\mu(x + K_N)}{|K_N|} \geq \mathscr{E}^+(\mu) - \epsilon, \quad x \in I_N,$$

and

$$f * h_N = |K_N|^{-1}\, \chi_{\check{K}_N} * f * \mu \leq C \text{ a.e. on } \mathbb{R}^d.$$

Therefore,

$$\int_{I_1} \int_{\mathbb{R}^d} f(x - y)\, h_N(y)\, \mathrm{d}y\, \mathrm{d}x \leq C$$

which implies that

$$\left(\mathscr{E}^+(\mu)-\epsilon\right)\int_{I_1}\int_{I_N} f(x-y)\,\mathrm{d}y\,\mathrm{d}x \le C.$$

Since, for every $x \in I_1$, the sequence $\{\int_{I_N} f(x-y)\,dy\}_{N\ge 1}$ is increasing and converges to 1, it follows from the Lebesgue monotone convergence theorem that

$$\mathscr{E}^+(\mu)-\epsilon = \lim_{N\to\infty}\left(\mathscr{E}^+(\mu)-\epsilon\right)\int_{I_1}\int_{I_N} f(x-y)\,\mathrm{d}y\,\mathrm{d}x \le C.$$

By the definition of $\mathscr{C}^+(\mu)$, we obtain that $\mathscr{E}^+(\mu)-\epsilon \le \mathscr{C}^+(\mu)$ and thus that $\mathscr{D}^+(\mu) = \mathscr{E}^+(\mu) \le \mathscr{C}^+(\mu)$ since $\epsilon > 0$ is arbitrary. This proves the statement in (a).

For part (b), we will prove first that $\mathscr{D}^-(\mu) \le \mathscr{C}^-(\mu)$. If $\mathscr{D}^-(\mu) = \infty$, we can find for any $M > 0$, a number $R > 0$ such that

$$\frac{\mu(I_R(z))}{R^d} \ge M, \quad z \in \mathbb{R}^d. \tag{9}$$

As before, this implies that $\mathscr{C}^-(\mu) \ge M$ and thus $\mathscr{C}^-(\mu) = \infty$ since $M > 0$ is arbitrary. Similarly, if $\mathscr{D}^-(\mu) < \infty$, we can find for any $\epsilon > 0$ a number $R > 0$ such that

$$\frac{\mu(I_R(z))}{R^d} \ge \mathscr{D}^-(\mu)-\epsilon, \quad z \in \mathbb{R}^d \tag{10}$$

which implies that $\mathscr{C}^-(\mu) \ge \mathscr{D}^-(\mu)$, since $\epsilon > 0$ is arbitrary. Suppose now that μ is translation-bounded. To prove the converse inequality, we use the fact that $\mathscr{D}^-(\mu) = \mathscr{E}^-(\mu)$ by Eq. (7) of Proposition 2. Given $\epsilon > 0$, we can find a sequence of compact set $\{K_N\}_{N\ge 1}$ with $|K_N| > 0$ such that

$$\frac{\mu(x+K_N)}{|K_N|} \le \mathscr{E}^-(\mu)+\epsilon, \quad x \in I_N.$$

By Proposition 2, we can assume that each set K_N is equal to some translate of a set I_R, $R \ge 1$. Hence, using Corollary 1, we can also assume the existence of a constant $L > 0$ depending only on μ such that

$$\sup_{x\in\mathbb{R}^d}\frac{\mu(x+K_N)}{|K_N|} \le L.$$

Suppose now that $D > 0$ is a constant such that $\mu * f \ge D$ a.e. for some function $f \in L^1(\mathbb{R}^d)$ with $f \ge 0$ and $\int_{\mathbb{R}^d} f(x)\,dx = 1$. Define $h_N = |K_N|^{-1}\chi_{\check{K}_N} * \mu$, where $\check{K}_N = \{-x,\ x \in K_N\}$. Then,

$$h_N(x) = \frac{\mu(x+K_N)}{|K_N|} \le \mathscr{E}^-(\mu)+\epsilon, \quad x \in I_N,$$

$$0 \le h_N(x) \le L, \quad x \in \mathbb{R}^d,$$

and

$$f * h_N = |K_N|^{-1}\, \chi_{\check{K}_N} * f * \mu \geq D \text{ a.e. on } \mathbb{R}^d.$$

Therefore,

$$\int_{I_1} \int_{\mathbb{R}^d} f(x-y)\, h_N(y)\, \mathrm{d}y\, \mathrm{d}x \geq D$$

which implies that

$$(\mathscr{E}^-(\mu)+\epsilon) \int_{I_1} \int_{I_N} f(x-y)\, \mathrm{d}y\, \mathrm{d}x + \int_{I_1} \int_{\mathbb{R}^d \setminus I_N} f(x-y)\, h_N(y)\, \mathrm{d}y\, \mathrm{d}x \geq D.$$

Since, for every $x \in I_1$, the sequence $\{\int_{I_N} f(x-y)\, \mathrm{d}y\}_{N\geq 1}$ is increasing, it follows from the Lebesgue monotone convergence theorem that

$$\lim_{N\to\infty} (\mathscr{E}^-(\mu)+\epsilon) \int_{I_1} \int_{I_N} f(x-y)\, \mathrm{d}y\, \mathrm{d}x = (\mathscr{E}^-(\mu)+\epsilon) \int_{\mathbb{R}^d} f(t)\, \mathrm{d}t = \mathscr{E}^-(\mu)+\epsilon.$$

Furthermore, since, for each $x \in I_1$,

$$\left| \int_{\mathbb{R}^d \setminus I_N} f(x-y)\, h_N(y)\, \mathrm{d}y \right| = \int_{\mathbb{R}^d \setminus I_N} f(x-y)\, h_N(y)\, \mathrm{d}y \leq L \int_{\mathbb{R}^d} f(x-y)\, \mathrm{d}y$$

and

$$\int_{I_1} L \int_{\mathbb{R}^d} f(x-y)\, \mathrm{d}y\, \mathrm{d}x = L \int_{\mathbb{R}^d} f(y)\, \mathrm{d}y < \infty,$$

the Lebesgue dominated convergence theorem shows that

$$\lim_{N\to\infty} \int_{I_1} \int_{\mathbb{R}^d \setminus I_N} f(x-y)\, h_N(y)\, \mathrm{d}y\, \mathrm{d}x = 0.$$

We deduce thus that $\mathscr{E}^-(\mu)+\epsilon \geq D$ and, using the definition of $\mathscr{C}^-(\mu)$, we obtain that $\mathscr{E}^-(\mu)+\epsilon \geq \mathscr{C}^-(\mu)$. Hence, $\mathscr{D}^-(\mu) = \mathscr{E}^-(\mu) \geq \mathscr{C}^-(\mu)$ since $\epsilon > 0$ is arbitrary. This completes the proof. □

It is important to notice that the assumption that the measure μ be translation-bounded in part (b) of the previous theorem cannot be dropped. Indeed, consider the example where $f(x) = \mathrm{e}^{-|x|}$ and μ is the measure $\mu = \delta_0 + \sum_{j=1}^{\infty} \mathrm{e}^j\, (\delta_{j^2} + \delta_{-j^2})$. We have

$$\begin{aligned}(\mu * f)(x) &= \mathrm{e}^{-|x|} + \sum_{j=1}^{\infty} \mathrm{e}^j \left(\mathrm{e}^{-|x-j^2|} + \mathrm{e}^{-|x+j^2|}\right) \\ &= \left(1 + \sum_{1\leq j<\sqrt{|x|}} \mathrm{e}^j\, \mathrm{e}^{j^2} + \sum_{j=1}^{\infty} \mathrm{e}^j\, \mathrm{e}^{-j^2}\right) \mathrm{e}^{-|x|} + \left(\sum_{j\geq\sqrt{|x|}} \mathrm{e}^j\, \mathrm{e}^{-j^2}\right) \mathrm{e}^{|x|}.\end{aligned}$$

It is clear that $\mu * f > 0$ everywhere and is bounded on compact subsets of $\mathbb{R}$. Furthermore if $j \geq 0$ and $j^2 \leq x \leq (j+1)^2$, we have

$$(\mu * f)(x) \geq \mathrm{e}^j \, \mathrm{e}^{-|x-j^2|} + \mathrm{e}^{j+1} \, \mathrm{e}^{-|x-(j+1)^2|} = \mathrm{e}^{j+j^2} \, \mathrm{e}^{-x} + \mathrm{e}^{-j-j^2} \, \mathrm{e}^x := g(x).$$

If $j = 0$, we have $g(x) = \mathrm{e}^x + \mathrm{e}^{-x} \geq 2$ and if $j \geq 1$, we have

$$g'(x) = -\mathrm{e}^{j+j^2} \, \mathrm{e}^{-x} + \mathrm{e}^{-j-j^2} \, \mathrm{e}^x$$

and $g''(x) = g(x) > 0$. Hence, the minimum value of the function g on the interval $[j^2, (j+1)^2]$ is $g(j+j^2) = 2$. Thus it shows that $\mu * f \geq 2$ on $[0, \infty)$ and since $\mu * f$ is even, we have thus $\mu * f \geq 2$ on $\mathbb{R}$. Since $\int_{\mathbb{R}} f(x) \, \mathrm{d}x = 2$, it follows that $\mathscr{C}^-(\mu) \geq 1$. On the other hand, we have $\mathscr{D}^-(\mu) = 0$ since the support of μ contains gaps that are arbitrarily large. It is clear that μ is not translation-bounded in this example. Similar examples in higher dimensions can be obtained by considering tensor products.

We now state some straightforward consequences of Theorem 1 which are often used in applications.

Corollary 3. *Let μ be a positive Borel measure on $\mathbb{R}^d$ and let $h \in L^1(\mathbb{R}^d)$ with $h \geq 0$:*

(a) If there exists a constant $C > 0$ such that the inequality $\mu * h \leq C$ holds a.e. on $\mathbb{R}^d$, then $\mathscr{D}^+(\mu) \int_{\mathbb{R}^d} h(x) \, \mathrm{d}x \leq C$.

(b) If μ is translation-bounded and there exists a constant $C > 0$ such that the inequality $\mu * h \geq C$ holds a.e. on $\mathbb{R}^d$, then we have $\mathscr{D}^-(\mu) \leq \mathscr{D}^+(\mu) < \infty$ and $\mathscr{D}^-(\mu) \int_{\mathbb{R}^d} h(x) \, \mathrm{d}x \geq C$.

Corollary 4. *Let μ be a positive Borel measure on $\mathbb{R}^d$ and let $h \in L^1(\mathbb{R}^d)$ with $h \geq 0$. Suppose that there exists a constant $C > 0$ such that*

$$\mu * h = C \text{ a.e. on } \mathbb{R}^d.$$

Then, the Beurling density $\mathscr{D}(\mu)$ of μ exists, and we have the equality

$$\mathscr{D}(\mu) \int_{\mathbb{R}^d} h(x) \, \mathrm{d}x = C.$$

Conversely, if the Beurling density $\mathscr{D}(\mu)$ of μ exists, we can find, for every $\epsilon > 0$, a function $h \in \mathscr{P}(\mathbb{R}^d)$ such that

$$\mathscr{D}(\mu) - \epsilon \leq \mu * h \leq \mathscr{D}(\mu) + \epsilon \text{ a.e. on } \mathbb{R}^d.$$

Proof. The first statement of the corollary follows immediately from Theorem 1. To prove the second one, we note that, by the definition of $\mathscr{C}^+(\mu)$ and $\mathscr{C}^-(\mu)$ and Theorem 1 again, we can find function h_1 and h_2 in $\mathscr{P}(\mathbb{R}^d)$ such that

$$\mu * h_1 \leq \mathscr{D}(\mu) + \epsilon \quad \text{and} \quad \mu * h_2 \geq \mathscr{D}(\mu) - \epsilon \quad \text{a.e. on } \mathbb{R}^d.$$

The results follows by letting $h = h_1 * h_2$. □

We point out that, in general, if the Beurling density $\mathscr{D}(\mu)$ of a measure μ exists, it might not possible to find a function $h \in \mathscr{P}(\mathbb{R}^d)$ such that $\mu * h = \mathscr{D}(\mu)$. For example, if $\mu = 1 + \delta_0$, where δ_0 denotes the Dirac mass at the origin, we have $\mathscr{D}(\mu) = 1$, but if h is as above, we have $\mu * h = 1 + h$ which clearly cannot be equal to 1 almost everywhere.

Corollary 5. *Let μ be a positive Borel measure on $\mathbb{R}^d$ and let $h \in L^1(\mathbb{R}^d) \setminus \{0\}$ with $h \geq 0$. Then,*

(a) $\mathscr{D}^+(\mu * h) = \mathscr{D}^+(\mu) \int_{\mathbb{R}^d} h(x)\,\mathrm{d}x$.
(b) If μ is translation-bounded, $\mathscr{D}^-(\mu * h) = \mathscr{D}^-(\mu) \int_{\mathbb{R}^d} h(x)\,dx$.
(c) The Beurling density $\mathscr{D}(\mu * h)$ exists if and only if $\mathscr{D}(\mu)$ exists and $\mathscr{D}(\mu * h) = \mathscr{D}(\mu) \int_{\mathbb{R}^d} h(x)\,\mathrm{d}x$ *in that case.*

Proof. It is enough to prove the result in the case where $\int_{\mathbb{R}^d} h(x)\,\mathrm{d}x = 1$. If $f \in \mathscr{P}(\mathbb{R}^d)$ and $C \geq 0$ is a constant such that

$$\mu * f \leq C \quad \text{a.e. on } \mathbb{R}^d,$$

then we have also

$$\mu * h * f = \mu * f * h \leq C \quad \text{a.e. on } \mathbb{R}^d,$$

which shows that $\mathscr{C}^+(\mu * h) \leq \mathscr{C}^+(\mu)$. On the other hand, if

$$\mu * h * f \leq C \quad \text{a.e. on } \mathbb{R}^d,$$

then $\mathscr{C}^+(\mu) \leq C$ and thus $\mathscr{C}^+(\mu) \leq \mathscr{C}^+(\mu * h)$. It follows that $\mathscr{C}^+(\mu) = \mathscr{C}^+(\mu * h)$ and the result in part (a) follows from part (a) of Theorem 1. Similar arguments show that $\mathscr{C}^-(\mu) = \mathscr{C}^-(\mu * h)$. If μ is translation-bounded, so is $\mu * h$ and we have thus $\mathscr{D}^-(\mu) = \mathscr{D}^-(\mu * h)$ using part (b) of Theorem 1. This proves (b) and (c) is an immediate consequence of (a) and (b). □

The example given after Theorem 1 shows that the equality in part (b) of the previous corollary may fail if μ is not translation-bounded.

5 Vector-Valued Measures

In tiling or paving problems involving several tiles or in the theory of multi-windows Gabor frames, for example (see [5]), one is led to consider convolution inequalities as in the previous section involving more than one measure. Consider a finite

collection $\mu_1, \ldots, \mu_m$ of positive Borel measures on $\mathbb{R}^d$. We associate with it the vector-valued measure $\boldsymbol{\mu}$ defined by

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m).$$

Given a vector $\mathbf{a} = (a_1, \ldots, a_m) \in \mathbb{R}^m$ with $a_i > 0$ for each $i = 1, \ldots, m$, we define the collection

$$\mathscr{P}^m_{\mathbf{a}}(\mathbb{R}^d) = \left\{ (f_1, \ldots, f_m) \in \left(L^1(\mathbb{R}^d)\right)^m,\ f_i \geq 0, \int_{\mathbb{R}^d} f_i(x)\, \mathrm{d}x = a_i,\ i = 1, \ldots, m \right\}.$$

Our goal in this section is to compute the expressions

$$\mathscr{C}^+(\boldsymbol{\mu}, \mathbf{a}) = \inf \left\{ C \geq 0, \sum_{i=1}^m \mu_i * f_i \leq C \text{ a.e. for some } (f_1, \ldots, f_m) \in \mathscr{P}^m_{\mathbf{a}}(\mathbb{R}^d) \right\},$$

with the convention that $\mathscr{C}^+(\boldsymbol{\mu}, \mathbf{a}) = \infty$ if the set where the infimum above is taken turns out to be empty, and

$$\mathscr{C}^-(\boldsymbol{\mu}, \mathbf{a}) = \sup \left\{ D \geq 0, \sum_{i=1}^m \mu_i * f_i \geq D \text{ a.e. for some } (f_1, \ldots, f_m) \in \mathscr{P}^m_{\mathbf{a}}(\mathbb{R}^d) \right\}.$$

We let $\mathscr{P}^m(\mathbb{R}^d) = \mathscr{P}^m_{\mathbf{a_0}}(\mathbb{R}^d)$, $\mathscr{C}^+(\boldsymbol{\mu}) = \mathscr{C}^+(\boldsymbol{\mu}, \mathbf{a_0})$ and $\mathscr{C}^-(\boldsymbol{\mu}) = \mathscr{C}^-(\boldsymbol{\mu}, \mathbf{a_0})$, where $\mathbf{a_0} \in \mathbb{R}^m$ is defined by letting its components $a_i = 1$ for each $i = 1, \ldots, m$. It is clear that $\mathscr{C}^+(\boldsymbol{\mu}, \mathbf{a}) = \mathscr{C}^+(\tilde{\boldsymbol{\mu}})$ and $\mathscr{C}^-(\boldsymbol{\mu}, \mathbf{a}) = \mathscr{C}^-(\tilde{\boldsymbol{\mu}})$ where $\tilde{\boldsymbol{\mu}} = (a_1 \mu_1, \ldots, a_m \mu_m)$. We will need the following lemma.

Lemma 5. *Let $(f_1, \ldots, f_m) \in \mathscr{P}^m(\mathbb{R}^d)$ and let μ_i, $i = 1, \ldots, m$, be positive Borel measures on $\mathbb{R}^d$. Suppose that there exists a constant $C > 0$ such that*

$$\sum_{i=1}^m f_i * \mu_i \leq C, \quad \text{a.e. on } \mathbb{R}^d.$$

Suppose, furthermore, that there exists a constant C_0 such that, for every integer $N \geq 1$, we can find a compact set K_N with $|K_N| > 0$ and positive constants $C_{i,N}$, $i = 1, \ldots, m$ such that $\sum_{i=1}^m C_{i,N} \geq C_0$ and

$$\frac{\mu_i(x + K_N)}{|K_N|} \geq C_{i,N}, \quad x \in I_N.$$

Then, we have the inequality

$$C_0 \leq C.$$

Proof. Define $h_{i,N} = |K_N|^{-1} \chi_{\check{K}_N} * \mu_i$, $i = 1, \ldots, m$, where $\check{K}_N = \{-x,\ x \in K_N\}$. Then,

$$h_{i,N}(x) = \frac{\mu_i(x + K_N)}{|K_N|} \geq C_{i,N}, \quad x \in I_N,\ i = 1, \ldots, m.$$

and

$$\sum_{i=1}^{m} f_i * h_{i,N} = |K_N|^{-1} \chi_{\check{K}_N} * \left(\sum_{i=1}^{m} f_i * \mu_i \right) \leq C \text{ a.e. on } \mathbb{R}^d.$$

Therefore,

$$\sum_{i=1}^{m} \int_{I_1} \int_{\mathbb{R}^d} f_i(x - y)\, h_{i,N}(y)\, \mathrm{d}y\, \mathrm{d}x \leq C$$

which implies that

$$\sum_{i=1}^{m} C_{i,N} \int_{I_1} \int_{I_N} f_i(x - y)\, \mathrm{d}y\, \mathrm{d}x \leq C. \tag{11}$$

Since,

$$\int_{I_N} f_i(x - y)\, \mathrm{d}y \leq \int_{\mathbb{R}^d} f_i(t)\, \mathrm{d}t, \quad x \in I_1,\ N \geq 1,$$

the Lebesgue dominated convergence theorem shows that

$$\lim_{N \to \infty} \int_{I_1} \int_{I_N} f_i(x - y)\, \mathrm{d}y\, \mathrm{d}x = \int_{\mathbb{R}^d} f_i(t)\, \mathrm{d}t = 1.$$

Given $\epsilon > 0$, we can thus find an integer N_0 such that

$$\int_{I_1} \int_{I_N} f_i(x - y)\, \mathrm{d}y\, \mathrm{d}x \geq 1 - \epsilon, \quad N \geq N_0,\ i = 1, \ldots, m.$$

Using Eq. (11), we obtain that, for $N \geq N_0$,

$$C_0\,(1 - \epsilon) \leq \sum_{i=1}^{m} C_{i,N}\,(1 - \epsilon) \leq C.$$

Since $\epsilon > 0$ is arbitrary, the inequality $C_0 \leq C$ follows. □

Theorem 2. *Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ be a vector-valued measure where each μ_i, $i = 1, \ldots, m$ is a positive Borel measure on $\mathbb{R}^d$ and let $\mathbf{a} = (a_1, \ldots, a_m) \in \mathbb{R}^m$, with $a_i > 0$ for each $i = 1, \ldots, m$, be given. Then,*

$$\mathscr{C}^+(\boldsymbol{\mu}, \mathbf{a}) = \mathscr{D}^+(\mu), \text{ where } \mu = \sum_{i=1}^{m} a_i\, \mu_i.$$

Proof. We can assume that $\mathbf{a} = \mathbf{a_0}$ as above and thus that $\mu = \sum_{i=1}^{m} \mu_i$. If $\mathscr{C}^+(\boldsymbol{\mu})$ is finite, then there exists a constant $C > 0$ such that $\mu_i * f_i \leq C$ a.e. for each $i = 1, \ldots, m$ and each measure μ_i must be translation-bounded by Proposition 1. Hence, μ is translation-bounded and $\mathscr{D}^+(\mu) < \infty$ by Corollary 2. Conversely, if $\mathscr{D}^+(\mu) < \infty$, the measure μ is translation-bounded using the same result and thus so is each measure μ_i. By Proposition 1, there exist functions $h_i \in L^1(\mathbb{R}^d)$ with $h_i \geq 0$ and $\int_{\mathbb{R}^d} h_i(x)\,\mathrm{d}x = 1$ as well as positive constants C_i, $i = 1, \ldots, m$, such that $\mu_i * f_i \leq C_i$ a.e. on $\mathbb{R}^d$.
Hence,

$$\sum_{i=1}^{m} \mu_i * f_i \leq \sum_{i=1}^{m} C_i$$

and $\mathscr{C}^+(\boldsymbol{\mu}) < \infty$. It follows that $\mathscr{C}^+(\boldsymbol{\mu}) = \infty$ if and only if $\mathscr{D}^+(\mu) = \infty$. We can thus assume that μ is translation-bounded and that both $\mathscr{C}^+(\boldsymbol{\mu})$ and $\mathscr{D}^+(\mu)$ are finite. Given $\epsilon > 0$, we can find $R > 0$ such that

$$\frac{\mu(I_R(z))}{R^d} \leq \mathscr{D}^+(\mu) + \epsilon, \quad z \in \mathbb{R}^d. \tag{12}$$

Letting $f_i = R^{-d}\,\chi_{I_R}$, for $i = 1, \ldots, m$, we have $f_i \in L^1(\mathbb{R}^d)$, $f_i \geq 0$, $\int_{\mathbb{R}^d} f_i(x)\,\mathrm{d}x = 1$ and

$$\sum_{i=1}^{m} (\mu_i * f_i)(z) = \frac{\mu(I_R(z))}{R^d}, \quad z \in \mathbb{R}^d.$$

Using Eq. (12) and the definition of $\mathscr{C}^+(\boldsymbol{\mu})$, we deduce that $\mathscr{C}^+(\boldsymbol{\mu}) \leq \mathscr{D}^+(\mu)$, since $\epsilon > 0$ is arbitrary. Conversely, using the definition of $\mathscr{D}^+(\mu)$, we can find an increasing sequence of positive numbers r_j with $\lim_{j\to\infty} r_j = \infty$ and corresponding points $x_j \in \mathbb{R}^d$ such that

$$\frac{\mu\left(x_j + I_{r_j}\right)}{r_j^d} \geq \mathscr{D}^+(\mu) - \epsilon/2.$$

Let $N \geq 1$ be an integer. We claim that we can find an integer J_0 such that, for $1 \leq i \leq m$,

$$\frac{\mu_i\left(x_j + I_{r_j} + x\right)}{r_j^d} \geq \frac{\mu_i\left(x_j + I_{r_j}\right)}{r_j^d} - \epsilon/(2m)$$

for all $x \in I_N$ and all $j \geq J_0$. Indeed, using Lemma 4, we have, for $r_j > N$ and $x \in I_N$, that

$$\left|\frac{\mu_i\left(x_j + I_{r_j} + x\right) - \mu_i\left(x_j + I_{r_j}\right)}{r_j^d}\right| \leq \frac{d\, C_{\mu_i}(I)\,(r_j + 1)^{d-1}\,(N+2)/2}{r_j^d} < \epsilon/(2m)$$

if j is large enough.

Since

$$\sum_{i=1}^{m} \left\{ \frac{\mu_i\left(x_j + I_{r_j}\right)}{r_j^d} - \epsilon/(2m) \right\} = \frac{\mu\left(x_j + I_{r_j}\right)}{r_j^d} - \epsilon/2 \geq \mathscr{D}^+(\mu) - \epsilon,$$

we deduce from Lemma 5 with $K_N = x_j + I_{r_j}$ and

$$C_{i,N} = \frac{\mu_i\left(x_j + I_{r_j}\right)}{r_j^d} - \epsilon/(2m), \quad i = 1, \ldots, m,$$

for any $j \geq J_0$, that $\mathscr{D}^+(\mu) - \epsilon \leq \mathscr{C}^+(\boldsymbol{\mu})$. Since $\epsilon > 0$ is arbitrary, the inequality $\mathscr{D}^+(\mu) \leq \mathscr{C}^+(\boldsymbol{\mu})$ follows. This proves our claim. □

Corollary 6. *For $i = 1, \ldots, m$, let μ_i be positive Borel measures on $\mathbb{R}^d$ and let h_i be functions h_i in $L^1(\mathbb{R}^d)$ with $h_i \geq 0$. Suppose that there exists a constant $C > 0$ such that*

$$\sum_{i=1}^{m} \mu_i * h_i \leq C \text{ a.e. on } \mathbb{R}^d.$$

Then, we have the inequality

$$\mathscr{D}^+(\mu) \leq C$$

where $\mu = \sum_{i=1}^m \left(\int_{\mathbb{R}^d} h_i(x)\,\mathrm{d}x\right) \mu_i$.

Lemma 6. *For $i = 1, \ldots, m$), let μ_i be positive translation-bounded Borel measures on $\mathbb{R}^d$ and let $(f_1, \ldots, f_m) \in \mathscr{P}^m(\mathbb{R}^d)$. Suppose that there exists a constant $C > 0$ such that*

$$\sum_{i=1}^{m} f_i * \mu_i \geq C, \quad \text{a.e. on } \mathbb{R}^d.$$

Suppose, furthermore, that there exists a constant C_1 such that, for every integer $N \geq 1$, we can find a compact set K_N with $|K_N| > 0$ and positive constants $C_{i,N}$, $i = 1, \ldots, I$ such that $\sum_{i=1}^m C_{i,N} \leq C_1$ and

$$\frac{\mu_i(x + K_N)}{|K_N|} \leq C_{i,N}, \quad x \in I_N.$$

Suppose also that there exists a constant $L > 0$ such that

$$\frac{C_{\mu_i}(K_N)}{|K_N|} \leq L, \quad N \geq 1,\ i = 1, \ldots, m.$$

Then, we have the inequality

$$C_1 \geq C.$$

Proof. Define $h_{i,N} = |K_N|^{-1}\, \chi_{\check{K}_N} * \mu_i$, $i = 1, \ldots, m$, where $\check{K}_N = \{-x,\ x \in K_N\}$. Then,

$$h_{i,N}(x) = \frac{\mu_i(x + K_N)}{|K_N|} \leq C_{i,N}, \quad x \in I_N,\ i = 1, \ldots, m.$$

and

$$\sum_{i=1}^{m} f_i * h_{i,N} = |K_N|^{-1}\, \chi_{\check{K}_N} * \left(\sum_{i=1}^{m} f_i * \mu_i \right) \geq C \text{ a.e. on } \mathbb{R}^d.$$

We have also

$$h_{i,N}(x) \leq L, \quad x \in \mathbb{R}^d, \quad i = 1, \ldots, m.$$

Therefore,

$$\sum_{i=1}^{m} \int_{I_1} \int_{\mathbb{R}^d} f_i(x - y)\, h_{i,N}(y)\, \mathrm{d}y\, \mathrm{d}x \geq C$$

which implies that

$$\sum_{i=1}^{m} C_{i,N} \int_{I_1} \int_{I_N} f_i(x-y)\, \mathrm{d}y\, \mathrm{d}x + \sum_{i=1}^{m} \int_{I_1} \int_{\mathbb{R}^d \setminus I_N} f_i(x-y)\, h_{i,N}(y)\, \mathrm{d}y\, \mathrm{d}x \geq C. \tag{13}$$

Since

$$\int_{\mathbb{R}^d \setminus I_N} f_i(x - y)\, h_{i,N}(y)\, \mathrm{d}y \ \leq L \int_{\mathbb{R}^d} f_i(x - y)\, \mathrm{d}y, \quad x \in I_1$$

and

$$\int_{I_1} \int_{\mathbb{R}^d} f_i(x - y)\, \mathrm{d}y\, \mathrm{d}x = \int_{\mathbb{R}^d} f_i(y)\, \mathrm{d}y = 1 < \infty,$$

the Lebesgue dominated convergence theorem shows that

$$\lim_{N \to \infty} \int_{I_1} \int_{\mathbb{R}^d \setminus I_N} f_i(x - y)\, h_{i,N}(y)\, \mathrm{d}y\, \mathrm{d}x = 0, \quad i = 1, \ldots, m.$$

The Lebesgue monotone convergence theorem also shows that

$$\lim_{N \to \infty} \int_{I_1} \int_{I_N} f_i(x - y)\, \mathrm{d}y\, \mathrm{d}x = \int_{\mathbb{R}^d} f_i(t)\, \mathrm{d}t = 1, \quad i = 1, \ldots, m.$$

Given $\epsilon > 0$, we can thus find an integer N_0 such that for $N \geq N_0$ and $i = 1, \ldots, m$, we have

$$\int_{I_1} \int_{I_N} f_i(x - y)\, \mathrm{d}y\, \mathrm{d}x \leq 1 + \epsilon, \quad \sum_{i=1}^{m} \int_{I_1} \int_{\mathbb{R}^{\mathrm{d}} \setminus I_N} f_i(x - y)\, h_{i,N}(y)\, \mathrm{d}y\, \mathrm{d}x < \epsilon.$$

Using Eq. (13), we obtain that, for $N \geq N_0$,

$$C_1\,(1+\epsilon) \geq \sum_{i=1}^{m} C_{i,N}\,(1+\epsilon) \geq C-\epsilon.$$

Since $\epsilon > 0$ is arbitrary, the inequality $C_1 \geq C$ follows. □

Theorem 3. *Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ be a vector-valued measure where, for each $i = 1, \ldots, m$, μ_i is a positive Borel measure on $\mathbb{R}^d$ and let $\mathbf{a} \in \mathbb{R}^m$ be a vector whose components are all positive. Then,*

$$\mathscr{C}^-(\boldsymbol{\mu}, \mathbf{a}) \geq \mathscr{D}^-(\mu), \text{ where } \mu = \sum_{i=1}^{m} a_i\,\mu_i.$$

Furthermore, if each measure μ_i, $i = 1, \ldots, m$, is translation-bounded, we have the equality

$$\mathscr{C}^-(\boldsymbol{\mu}, \mathbf{a}) = \mathscr{D}^-(\mu). \tag{14}$$

Proof. As before, we can assume that $\mathbf{a} = \mathbf{a_0} = (1, \ldots, 1)$. The inequality $\mathscr{C}^-(\boldsymbol{\mu}) \geq \mathscr{D}^-(\mu)$ follows immediately from the inequality $\mathscr{C}^-(\mu) \geq \mathscr{D}^-(\mu)$ obtained in part (b) of Theorem 1 and the inequality $\mathscr{C}^-(\boldsymbol{\mu}) \geq \mathscr{C}^-(\mu)$, if $\mu = \sum_{i=1}^m \mu_i$, which follows immediately from the respective definitions of these quantities. Let us now prove (14) under the assumption that each measure μ_i, $i = 1, \ldots m$, and thus also μ, is translation-bounded. Using Corollary 1, it follows that $\mathscr{D}^+(\mu) < \infty$ and thus $\mathscr{D}^-(\mu) < \infty$. Fix $\epsilon > 0$. Using the definition of $\mathscr{D}^-(\mu)$, we can find an increasing sequence of positive numbers r_j with $\lim_{j\to\infty} r_j = \infty$ and corresponding points $x_j \in \mathbb{R}^d$ such that

$$\frac{\mu\left(x_j + I_{r_j}\right)}{r_j^d} \leq \mathscr{D}^-(\mu) + \frac{\epsilon}{2}.$$

Let $N \geq 1$ be an integer. We claim that we can find an integer J_0 such that, for $1 \leq i \leq m$ and $j \geq J_0$,

$$\frac{\mu_i\left(x_j + I_{r_j} + x\right)}{r_j^d} \leq \frac{\mu_i\left(x_j + I_{r_j}\right)}{r_j^d} + \frac{\epsilon}{2m}, \quad x \in I_N.$$

Indeed, using Lemma 4, we have, for $r_j > N$ and $x \in I_N$, that

$$\left|\frac{\mu_i\left(x_j + I_{r_j} + x\right) - \mu_i\left(x_j + I_{r_j}\right)}{r_j^d}\right| \leq \frac{d\,C_{\mu_i}(I)\,(r_j+1)^{d-1}\,(N+2)/2}{r_j^d} < \frac{\epsilon}{2m}$$

if j is large enough. Furthermore, since, by assumption, each measure μ_i, $i = 1, \ldots, m$, is translation-bounded, we can use Corollary 1 to show the existence of a constant $L > 0$ such that

$$\frac{C_{\mu_i}(z+I_R)}{|I_R|} \le L, \quad z \in \mathbb{R}^d, \ R \ge 1, \ 1 \le i \le m.$$

Since

$$\sum_{i=1}^{m} \left\{ \frac{\mu_i\left(x_j + I_{r_j}\right)}{r_j^d} + \frac{\epsilon}{2m} \right\} = \frac{\mu\left(x_j + I_{r_j}\right)}{r_j^d} + \frac{\epsilon}{2} \le \mathscr{D}^-(\mu) + \epsilon,$$

we deduce from Proposition 5 with $K_N = x_j + I_{r_j}$ and

$$C_{i,N} = \frac{\mu_i\left(x_j + I_{r_j}\right)}{r_j^d} + \frac{\epsilon}{2m}, \quad i = 1, \ldots, m,$$

for any $j \ge J_0$, that $\mathscr{D}^-(\mu) + \epsilon \ge \mathscr{C}^-(\boldsymbol{\mu})$. Since $\epsilon > 0$ is arbitrary, Eq. (14) follows. □

Corollary 7. *For $i = 1, \ldots, m$, let μ_i be translation-bounded, positive Borel measures on $\mathbb{R}^d$ and let h_i be functions h_i in $L^1(\mathbb{R}^d)$ with $h_i \ge 0$. Suppose that there exists a constant $C > 0$ such that*

$$\sum_{i=1}^{m} \mu_i * h_i \ge C \ \text{a.e. on } \mathbb{R}^d.$$

Then, we have the inequality

$$\mathscr{D}^-(\mu) \ge C.$$

where $\mu = \sum_{i=1}^m \left(\int_{\mathbb{R}^d} h_i(x)\,\mathrm{d}x\right) \mu_i$.

Combining Corollaries 6 and 7, we obtain the following result.

Corollary 8. *For $i = 1, \ldots, m$, let μ_i be positive Borel measures on $\mathbb{R}^d$ and let h_i be functions h_i in $L^1(\mathbb{R}^d)$ with $h_i \ge 0$. Suppose that there exists a constant $C > 0$ such that*

$$\sum_{i=1}^{m} \mu_i * h_i = C \ \text{a.e. on } \mathbb{R}^d.$$

Then, the Beurling density $\mathscr{D}(\mu)$ of the measure $\mu = \sum_{i=1}^m \left(\int_{\mathbb{R}^d} h_i(x)\,\mathrm{d}x\right) \mu_i$ exists and we have the equality

$$\mathscr{D}(\mu) = C.$$

Acknowledgments This work was supported by an NSERC grant.

References

1. Halmos, P.R.: Measure Theory. D. Van Nostrand Company, New York (1950)
2. Kolountzakis, M.N.: The study of translational tiling with Fourier analysis in Fourier analysis and convexity, pp. 131–187, Appl. Numer. Harmon. Anal. Birkhäuser, Boston (2004)
3. Kolountzakis, M.N., Lagarias, J.C.: Structure of tilings of the line by a function. Duke Math. J. **82**, 653–678 (1996)
4. Kutyniok, G.: Beurling density and shift-invariant weighted irregular Gabor systems. Sampl. Theory Signal Image Process. **5**, 163–181 (2006)
5. Kutyniok, G.: Affine Density in Wavelet Analysis. Lecture Notes in Mathematics, 1914. Springer, Berlin (2007)
6. Landau, H.J.: Necessary density conditions for sampling and interpolation of certain entire functions. Acta Math. **117**, 37–52 (1967)
7. Rudin, W.: Real and Complex Analysis, 3rd edn. McGraw-Hill, New York (1987)

Positive-Operator-Valued Measures: A General Setting for Frames

Bill Moran, Stephen Howard, and Doug Cochran

Abstract This chapter presents an overview of close parallels that exist between the theory of positive-operator-valued measures (POVMs) associated with a separable Hilbert space and the theory of frames on that space, including its most important generalizations. The concept of a framed POVM is introduced, and classical frames, fusion frames, generalized frames, and other variants of frames are all shown to arise as framed POVMs. This observation allows drawing on a rich existing theory of POVMs to provide new perspectives in the study of frames.

Keywords Frame • Generalized frame • Fusion frame • Positive operator-valued measure (POVM) • Framed POVM • Spectral measure • Naimark's Theorem • Stinespring's Theorem • Radon-Nikodym Theorem

1 Introduction

Frames have become a standard tool in signal processing, allowing uniform description of many linear but non-orthogonal transform techniques that underpin a wide variety of signal and image processing algorithms. Initially popularized in connection with wavelet applications, frames are now widely used in sampling,

B. Moran
Defence Science Institute, University of Melbourne, Parkville, VIC, Australia
e-mail: wmoran@unimelb.edu.au

S. Howard
Defence Science and Technology Organisation, Edinburgh, SA, Australia
e-mail: stephen.howard@dsto.defence.gov.au

D. Cochran (✉)
Arizona State University, Tempe, AZ, USA
e-mail: cochran@asu.edu

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_4,

compression, array processing, as well as in spectral and other transform methods for time series.

Frames were initially introduced in a 1952 paper of Duffin and Schaeffer [10], where they appeared as an abstraction of sampled Fourier transforms. Little interest was shown in them until the appearance of the 1986 paper [8] by Daubechies, Grossmann, and Meyer which coincided with the rise of wavelet methods in signal processing. Subsequently they were taken up by numerous authors. Several excellent sources, including [7, 11, 14, 15], are available for further details of both the theory and the many applications of frames.

The standard definition of a frame is as a collection $\mathscr{F} = \{\varphi_k : k \in K\}$ of elements of a separable Hilbert space $\mathscr{H}$. The index set K may be finite or infinite. In order for $\mathscr{F}$ to constitute a frame, there must exist constants $0 < A \leq B < \infty$ such that, for all $f \in \mathscr{H}$,

$$A\|f\|^2 \leq \sum_{k\in K} |\langle \varphi_k, f\rangle|^2 \leq B\|f\|^2. \tag{1}$$

Roughly speaking, a projection $f \mapsto \langle f, \varphi_k\rangle$ of a vector f representing the state of a system onto an individual element φ_k of a frame may be seen as a measurement of that system, and the aim is to reconstruct the state f from the collection of all individual measurements $\{\langle f, \varphi_k\rangle : k \in K\}$ in a robust way. The frame condition as stated in Eq. (1) expresses the ability to do that, and the frame bounds A and B provide a measure of robustness. If $A = B$, the frame is said to be *tight*. Orthonormal bases are special cases of tight frames, and for these $A = B = 1$.

Several generalizations of the basic concept of a frame have been proposed. These include, in particular, the possibility that the family $\{\varphi_k : k \in K\} \subset \mathscr{H}$ is indexed by a continuum rather than a discrete index set, resulting in what are called *generalized frames*. There are various formulations of generalized frames in the literature; see in particular [1]. From the perspective of this chapter, the infrastructure of a generalized frame is a measurable function from a measure space, which serves the role of the index set, to $\mathscr{H}$. Specifically, let $(\Omega, \mathscr{B}, \mu)$ be a measure space (e.g., $\Omega = \mathbb{R}$ with $\mathscr{B}$ its Borel sets and μ Lebesgue measure) and let $\Phi : \Omega \to \mathscr{H}$ be a μ-measurable function. The collection $\{\Phi(t) : t \in \Omega\} \subset \mathscr{H}$ is a generalized frame for $\mathscr{H}$ if it satisfies a condition analogous to the frame condition (1), i.e., for all $f \in \mathscr{H}$,

$$A\|f\|^2 \leq \int_\Omega |\langle \Phi(t), f\rangle|^2 \, \mathrm{d}\mu(t) \leq B\|f\|^2. \tag{2}$$

Define $\Pi_\varphi : \mathscr{H} \to \mathscr{H}$ to be orthogonal projection into the one-dimensional subspace spanned by the unit-norm element $\varphi \in \mathscr{H}$, i.e., $\Pi_\varphi(f) = \langle \varphi, f\rangle \varphi$. With this notation, Eq. (2) becomes

$$A\mathbb{I} \leq \int_\Omega \Pi_{\Phi(t)} \, \mathrm{d}\mu(t) \leq B\mathbb{I}, \tag{3}$$

where $\mathbb{I}$ denotes the identity operator on $\mathscr{H}$ and the inequalities mean that the differences are positive definite operators on $\mathscr{H}$. The integral in Eq. (3) is in the weak sense, i.e., for a suitable measurable family of operators $\{S(t) : t \in \Omega\}$ on $\mathscr{H}$, the integral $\int_\Omega S(t)\,\mathrm{d}\mu(t)$ is defined to be the operator D satisfying

$$\langle f, D\varphi\rangle = \int_\Omega \langle f, S(t)\varphi\rangle\,\mathrm{d}\mu(t)$$

for f and φ in $\mathscr{H}$.

Fusion frames generalize the concept of a frame in a different direction. They have received considerable recent attention in the signal processing literature; see, for example, [3,5,12,20]. In a fusion frame, the one-dimensional projections Π_{φ_k} are replaced by projections Π_k onto potentially higher-dimensional closed subspaces $W_k \subset \mathscr{H}$. Thus a fusion frame $\mathscr{F}$ is a family $\{(W_k, w_k) : k \in K\}$ of closed subspaces of $\mathscr{H}$ and a corresponding family of weights $w_k \geq 0$ satisfying the frame condition

$$A\|f\|^2 \leq \sum_{k\in K} w_k^2 \|\Pi_k(f)\|^2 \leq B\|f\|^2 \tag{4}$$

for all $f \in \mathscr{H}$. Some authors have promoted fusion frames as a means of representing the problem of fusion of multiple measurements in, for example, a sensor network. In this view, each projection corresponds to a node of the network, and the fusion frame itself, as its name suggests, provides the mechanism for fusion of these measurements centrally.

Not surprisingly, the ideas of generalized frames and fusion frames can be combined into a composite generalization. A *generalized fusion frame* $\mathscr{F}$ for $\mathscr{H}$ consists of a pair of measurable functions (Φ, w). In this setting, $w : \Omega \to \overline{\mathbb{R}_+}$ and $\Phi : \Omega \to \mathscr{P}(\mathscr{H})$ where $\mathscr{P}(\mathscr{H})$ denotes the space of orthogonal projections of any rank (including possibly ∞) on $\mathscr{H}$, endowed with the weak operator topology. Measurability of Φ is in the weak sense that $t \mapsto \langle\varphi, \Phi(t)\psi\rangle$ is μ-measurable for each φ and ψ in $\mathscr{H}$. As part of the definition, it is also required that the function $t \mapsto \Phi(t)f$ is in $L_2(\Omega, \mu)$ for each $f \in \mathscr{H}$. The frame condition in operator form, as in Eq. (3), becomes

$$A\mathbb{I} \leq \int_\Omega w(t)^2\Phi(t)\,\mathrm{d}\mu(t) \leq B\mathbb{I}.$$

As described in later sections of this chapter, this definition of a generalized fusion frame leads to a concept that is, in effect if not in formalism, remarkably similar to that of a positive-operator-valued measure (POVM) — a concept that has been prevalent in the quantum physics literature for many years. This is hardly unexpected from a signal processing viewpoint, as the concept of POVM was introduced and developed in quantum mechanics as a means to represent the most general form of quantum measurement of a system. Further, connections between POVMs and frames have been noted frequently in the physics literature

(e.g., [4, 18]), although these relationships seem to be unmentioned in mathematical work on frames.

The remainder of this chapter develops a generalization of the POVM concept as used in quantum mechanics, which encompasses the theory of frames — including all of the generalizations discussed above. Once generalized fusion frames are accepted, setting the discourse in terms of POVMs enables the importation of much theory from the quantum mechanics literature and also brings to light some decompositions that are not readily apparent from the frame formalism.

A key result used in what follows is the classical theorem of Naimark [16] which, long before frames became popular in signal processing or POVMs were used in quantum mechanics, formalized analysis and synthesis in this general context. When applied to the cases above, Naimark's perspective exactly reproduces those notions.

Subsequent sections describe positive-operator-valued measures, introduce the theorem of Naimark, and discuss how POVMs relate to frames and their generalizations. In this brief description of the relationship between POVMs and the generalizations of frames, it will only be possible to touch on the power of the POVM formalism.

2 Analysis and Synthesis

The various concepts of frame, fusion frame, and generalized frame all give rise to analysis and synthesis operations. In the case of a frame, a prevalent point of view is that an analysis operator F takes a "signal" in $\mathscr{H}$ to a set of complex "coefficients" in the space $\ell_2(K)$ of square-summable sequences on the index set K, i.e., F is the Bessel map given by $F(f) = \{\langle f, \varphi_k \rangle : k \in K\}$ where the finiteness of the upper frame bound B guarantees the square summability of this coefficient sequence. The synthesis operator is the adjoint map $F^* : \ell_2(K) \to \mathscr{H}$, given by

$$F^*(\{a_k\}) = \sum_{k \in K} a_k \varphi_k,$$

and corresponds to synthesis of a signal from a set of coefficients. It follows directly from Eq. (1) that the *frame operator* $\mathbb{F} = F^*F$ satisfies

$$A\mathbb{I} \leq \mathbb{F} \leq B\mathbb{I}. \tag{5}$$

To accommodate developments later in this chapter, it is useful to describe analysis and synthesis with frames in a slightly different way. With each φ_k in the frame $\mathscr{F}$, associate the one-dimensional orthogonal projection operator Π_k that takes $f \in \mathscr{H}$ to

$$\Pi_k(f) = \frac{\langle \varphi_k, f \rangle}{||\varphi_k||^2} \varphi_k$$

Note that $\Pi_k : \mathscr{H} \to W_k$ where W_k is the one-dimensional subspace of $\mathscr{H}$ spanned by φ_k. Also, $\|\Pi_k(f)\| = |\langle \varphi_k, f\rangle|/\|\varphi_k\|$. Thus the frame condition (1) is equivalent to

$$A\|f\|^2 \le \sum_{k\in K} w_k^2 \|\Pi_k(f)\|^2 \le B\|f\|^2$$

where $w_k = |\langle \varphi_k, f\rangle| \ge 0$. From a comparison of this expression with Eq. (4), it is clear that the weights w_k account for the possibility that the frame elements $\varphi_k \in \mathscr{F}$ are not of unit norm. Although it is typical to think of the analysis operator as producing a set of coefficients for each signal $f \in \mathscr{H}$ via the Bessel map, as described above, it is more suitable for generalization to regard it as a map from $\mathscr{H}$ to $\mathscr{H}$ that "channelizes" f into signals $w_k\Pi_k(f) \in W_k \subset \mathscr{H}$. The synthesis operator is then a linear rule for combining a set of signals from the channels W_k to form an aggregate signal in $\mathscr{H}$.

With this view, the analysis operator for a fusion frame is a natural generalization of its frame counterpart in which the subspaces W_k can be of dimension greater than one and the projection operators Π_k are from $\mathscr{H}$ to W_k. The analysis operator is $F : \mathscr{H} \to \bigoplus_{k\in K} W_k$ given by

$$F(f) = \{w_k\Pi_k(f) : k \in K\} \in \bigoplus_{k\in K} W_k.$$

The adjoint map $F^* : \bigoplus_{k\in K} W_k \to \mathscr{H}$ is given by

$$F^*(\{\xi_k\}) = \sum_{k\in K} w_k\xi_k \in \mathscr{H}, \quad \{\xi_k : k \in K\} \in \bigoplus_{k\in K} W_k.$$

The frame bound conditions guarantee that everything is well-defined. The corresponding fusion frame operator $\mathbb{F} = F^*F : \mathscr{H} \to \mathscr{H}$ is given by

$$\mathbb{F}(f) = \sum_{k\in K} w_k^2\Pi_k(f),$$

and the same kind of frame bound inequality as in Eq. (5) holds for fusion frames.

For the generalized frame described in Sect. 1, the analysis operator $F : \mathscr{H} \to L_2(\Omega, \mu)$ is given by

$$F(f)(t) = \langle f, \Phi(t)\rangle, \quad t \in \Omega,\ f \in \mathscr{H},$$

and its adjoint by

$$F^*(u) = \int_\Omega u(t)\Phi(t)\,\mathrm{d}\mu(t) \in \mathscr{H}, \quad u \in L_2(\Omega, \mu).$$

Again, the generalized frame operator $\mathbb{F} = F^*F$ satisfies inequalities (5).

For generalized fusion frames there is a corresponding definition of analysis and synthesis operators, but its description requires the definition of direct integrals of Hilbert spaces [9]. In any case the ideas will be subsumed under the more general development to follow.

It is immediately evident that, in each case discussed above, the synthesis operator does not reconstruct the analyzed signal, i.e., in general $F^*F \neq \mathbb{I}$. In the case of a frame, inversion of the analysis operator is performed by invoking a *dual frame*. There are various different usages of this terminology in the literature (see [5, 13, 15]). For the purposes here, given a frame $\{\varphi_k\}$ for the Hilbert space $\mathscr{H}$, a dual frame $\{\tilde{\varphi}_k\}$ satisfies

$$f = \sum_{k\in K} \langle \varphi_k, f\rangle \tilde{\varphi}_k = \sum_{k\in K} \langle \tilde{\varphi}_k, f\rangle \varphi_k. \tag{6}$$

In other words, the dual frame inverts the analysis and synthesis operations of the original frame to give perfect reconstruction. Such a dual frame always exists; indeed, it is easy to verify that

$$\tilde{\varphi}_k = \mathbb{F}^{-1}(\varphi_k) \tag{7}$$

has the appropriate property. Dual frames as defined in Eq. (6) are not in general unique; the one in Eq. (7) is called the *canonical dual frame*. In the case of a fusion frame $\{(W_k, w_k) : k \in K\}$, the concept of duality is more intricate. See [13] for discussion of dual frames in this context.

3 Positive-Operator-Valued Measures

The goal of this section is to define a *framed POVM* and give some examples of such objects. Consider a topological space Ω which, to avoid technicalities, will be assumed to be "nice," for example, a complete separable metric space or a locally compact second countable space. The crucial point is that Ω has sufficient structure to make the concept of regularity of measures meaningful and useful, though regularity will not be explicitly discussed in this chapter. Denote by $\mathscr{B}(\Omega)$ the σ-algebra of Borel sets on Ω and by $\mathscr{P}(\mathscr{H})$ the space of positive operators on a Hilbert space $\mathscr{H}$. A *framed POVM* is a function $M : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{H})$ satisfying the following two conditions:

(POVM-1) For all f in $\mathscr{H}$, $\omega \mapsto \langle f, M(\omega) f\rangle$ is a regular Borel measure on $\mathscr{B}(\Omega)$, denoted by μ_f.

(POVM-2) $A\mathbb{I} \leq M(\Omega) \leq B\mathbb{I}$ for some $0 < A \leq B < \infty$.

As in the case of frames, the numbers A and B are called the frame bounds for M. Without the condition POVM-2, the object is called a POVM, i.e., without the epithet "framed." Such a function is a measure on $\mathscr{B}(\Omega)$ that takes values in the

set of positive operators on $\mathscr{H}$, though the countable aspect of its additivity is only in a weak sense. In the quantum mechanics context, POVM-2 is replaced by the more strict requirement that $M(\Omega) = \mathbb{I}$. A framed POVM is *tight* if $A = B$, and if $A = B = 1$, M is a *probability POVM*. Probability POVMs are used in quantum mechanics as the most general form of quantum measurement.

As an example of a framed POVM, consider a fusion frame $\{(W_k, w_k) : k \in K\}$ in $\mathscr{H}$. Define $\Omega = K$ with the σ-field $\mathscr{B}(\Omega)$ taken to be the power set of Ω. Denoting, as above, projection onto W_k by Π_k,

$$M(\omega) = \sum_{k \in \omega} w_k \Pi_k. \tag{8}$$

It is straightforward to see that this satisfies both parts of the definition of a framed POVM, with the frame bounds being the bounds in the definition of the fusion frame. Thus, every fusion frame, and hence every frame, is trivially represented as a framed POVM.

If $\mathscr{F} = \{\Phi(t) : t \in \Omega\}$ is a generalized frame for $\mathscr{H}$, a POVM $M : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{H})$ can be defined by

$$M(\omega) = \int_{\omega} \Pi_{\Phi(t)} \, \mathrm{d}\mu(t), \tag{9}$$

where $\Pi_{\Phi(t)}$ denotes projection into the one-dimensional subspace of $\mathscr{H}$ spanned by $\Phi(t)$. M is a framed POVM with the same frame bounds as those of $\mathscr{F}$.

As will be discussed in Sect. 5, POVMs provide a rather general framework for analysis and reconstruction of signals. It will be seen that framed POVMs are only slightly more general than generalized fusion frames discussed briefly in Sect. 1. The impetus for studying POVMs in this context arises in part from the opportunity to draw on existing theory about POVMs in the physics literature for development and description of new constructs in signal processing. Some examples in this chapter illustrate this possibility, though much of the formalism is omitted from this overview.

4 Spectral Measures and the Naimark Theorem

A POVM S is a *spectral* POVM if

$$S(\omega_1 \cap \omega_2) = S(\omega_1)S(\omega_2), \qquad \omega_1, \omega_2 \in \mathscr{B}(\Omega).$$

Spectral POVMs arise, for example, in the spectral theorem for Hermitian operators on Hilbert space (see, e.g., [21]). If S is a spectral POVM, then $S(\Omega)$ is a projection, and every $S(\omega)$ with $\omega \in \mathscr{B}(\Omega)$ is a projection dominated by $S(\Omega)$, i.e.,

$$S(\omega)S(\Omega) = S(\Omega)S(\omega) = S(\omega).$$

Thus, for any $\omega \in \mathscr{B}(\Omega)$, $S(\omega)$ is completely specified by its behavior on the closed subspace $S(\Omega)\mathscr{H}$ of $\mathscr{H}$. Consequently, for most purposes, it suffices to assume $S(\Omega) = \mathbb{I}_{\mathscr{H}}$. In particular, if a spectral POVM is framed, then this condition must hold; conversely, imposing this condition on a spectral POVM ensures that it is framed. Since the focus here is on framed POVMs, it will be assumed that $S(\Omega) = \mathbb{I}_{\mathscr{H}}$ whenever a spectral POVM appears in subsequent discussion in this chapter. Note that, while a spectral POVM S need not be probability POVM in general, the condition that it is framed implies that S will be a probability POVM. Intuitively, spectral POVMs play an analogous role relative to framed POVMs to the one played by orthogonal bases relative to frames, i.e., spectral POVMs generalize orthogonal bases in a sense similar to that in which framed POVMs generalize frames.

With this machinery in place, it is possible to state the key theorem on POVMs due to Naimark [16], who formulated the result for POVMs without the framed condition. The following version is a relatively straightforward adaptation to framed POVMs.

Theorem 1. *Suppose $M : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{H})$ is a framed POVM with frame bounds A and B. Then there is an "auxiliary" Hilbert space $\mathscr{H}^\sharp$, a spectral POVM S with values in $\mathscr{P}(\mathscr{H}^\sharp)$, and a bounded linear map $V : \mathscr{H}^\sharp \to \mathscr{H}$ such that*

$$M(\omega) = VS(\omega)V^*, \qquad \omega \in \mathscr{B}(\Omega)$$

and $A\mathbb{I} \leq VV^* \leq B\mathbb{I}$.

For developments later in this chapter, it will be useful to have a sketch of the proof of this theorem. Given a POVM $M : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{H})$, consider the linear space $\mathscr{L}$ of $\mathscr{H}$-valued simple functions on Ω, i.e., finite linear combinations of functions of the form

$$\xi_\omega(t) = \begin{cases} \xi & \text{if } t \in \omega \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where $\omega \in \mathscr{B}(\Omega)$ and $\xi \in \mathscr{H}$. A pre-inner product on $\mathscr{L}$ is obtained by defining

$$\langle \xi_\omega, \xi'_{\omega'} \rangle_{\mathscr{L}} = \langle M(\omega)\xi, M(\omega')\xi' \rangle_{\mathscr{H}}. \tag{11}$$

Completion followed by factoring out zero-length vectors produces $\mathscr{H}^\sharp$, as a Hilbert space. The map from $\mathscr{H}$ to $\mathscr{L}$ taking ξ to ξ_Ω results in $V^* : \mathscr{H} \to \mathscr{H}^\sharp$ and V takes ξ_ω to $M(\Omega)^* M(\omega)\xi$. The spectral measure S arises first on $\mathscr{L}$ as

$$S(\omega')(\xi_\omega) = \xi_{\omega \cap \omega'} \qquad \xi \in \mathscr{H}, \quad \omega, \omega' \in \mathscr{B}(\Omega), \tag{12}$$

and then carries over to $\mathscr{H}^\sharp$.

The collection $(S, \mathscr{H}^\sharp, V)$ is known as a *Naimark representation* of the framed POVM $M : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{H})$. Further, a Naimark representation is *minimal* if the set

$$\{S(\omega)V^*\varphi : \varphi \in \mathscr{H}, \, \omega \in \mathscr{B}(\Omega)\}$$

is dense in $\mathscr{H}^\sharp$. Minimal Naimark representations are essentially unique in the sense that if $(S, \mathscr{H}_\sharp, V)$ and $(S', \mathscr{H}'_\sharp, V')$ are two such representations for the same M, then there is a surjective isometry $T : \mathscr{H}_\sharp \to \mathscr{H}'_\sharp$ such that $V'T = V$ and $T^{-1}S'(\omega)T = S(\omega)$ for all $\omega \in \mathscr{B}(\Omega)$. A fashionable way to handle the Naimark representation in recent literature (see [17]) is to convert POVMs to (completely) positive operators on commutative C^*-algebras via integration. In this setting, Naimark's theorem becomes a special case of Stinespring's theorem [19], which does not require commutativity of the C^*-algebra. A full description of this approach would be tangential to this chapter.

Example 1. Consider a generalized frame $\Phi : \Omega \to \mathscr{H}$ on the measure space (Ω, μ) with frame bounds $A \leq B$. Φ gives rise to a framed POVM M as in Eq. (9). To form a Naimark representation for M, define the Hilbert space $\mathscr{H}^\sharp$ to be $L_2(\Omega, \mu)$ and let the spectral measure S be the canonical one on this space, i.e.,

$$S(\omega) f(t) = \mathbb{1}_\omega(t) f(t), \qquad f \in L_2(\Omega, \mu).$$

S is clearly a spectral measure since the characteristic functions satisfy $\mathbb{1}_\omega \mathbb{1}_{\omega'} = \mathbb{1}_{\omega \cap \omega'}$. The map $V : L_2(\Omega, \mu) \to \mathscr{H}$ is defined by

$$V(f) = \int_\Omega f(t)\Phi(t)\, \mathrm{d}\mu(t),$$

where $f(t)\Phi(t)$ is the product of the scalar $f(t)$ and $\Phi(t) \in \mathscr{H}$. It can be verified that this is indeed a (the) minimal Naimark representation of M.

Example 2. Let $\mathscr{F} = \{(W_k, w_k) : k \in K\}$ be a fusion frame in $\mathscr{H}$. $\mathscr{F}$ corresponds to a framed POVM as in Eq. (8). In this case, $\mathscr{H}^\sharp$ may be taken to be the formal direct sum $\bigoplus_{k \in K} W_k$. The appropriate spectral measure S is defined on subsets J of $\Omega = K$ by

$$S(J) = \bigoplus_{k \in J} \Pi_k \tag{13}$$

where Π_k is the projection into W_k in $\mathscr{H}^\sharp$. Writing an element f of $\mathscr{H}^\sharp$ as $f = \{f_k \in W_k : k \in K\}$, the map $V : \mathscr{H}^\sharp \to \mathscr{H}$ is given by

$$V(f) = \sum_{k \in K} w_k f_k, \tag{14}$$

where the terms in the sum are considered as elements of $\mathscr{H}$. The square summability of the weights w_k guarantees that the sum in Eq. (14) converges in $\mathscr{H}$ because the Cauchy–Schwarz inequality gives

$$\sum_{k \in K} \|w_k \varphi_k\| \leq \Big(\sum_{k \in K} w_k^2\Big)^{1/2} \Big(\sum_{k \in K} \|\varphi_k\|^2\Big)^{1/2}. \tag{15}$$

Thus $V : \mathscr{H}^\sharp \to \mathscr{H}$ is a bounded linear map; in fact, by Eq. (15),

$$\|V\| \leq \Big(\sum_{k\in K} w_k^2\Big)^{1/2}.$$

Its adjoint $V^* : \mathscr{H} \to \mathscr{H}^\sharp$ is given by

$$V^*(\varphi) = \{w_k \Pi_k(\varphi) : k \in K\}.$$

Setting $\omega = \Omega = K$ gives $S(\Omega) = \mathbb{I}$ and

$$M(\Omega) = VS(\Omega)V^* = VV^*,$$

The frame bounds imply $A \leq VV^* \leq B$, and, if the fusion frame is tight, then $VV^* = A\mathbb{I}$.

From a comparison of the descriptions in Sect. 2 with the examples given here, it is evident that Naimark's theorem provides exactly the machinery for discussing analysis and synthesis operators in a general context. This is undertaken in the next section.

5 Analysis and Synthesis for General POVMs

The preceding examples indicate that the Naimark representation provides a mechanism for analysis and synthesis in POVMs that precisely extends the corresponding ideas for frames and fusion frames. To be specific, let $M : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{H})$ be a POVM and let $(S, \mathscr{H}^\sharp, V)$ be the corresponding minimal Naimark representation. In this context, $\mathscr{H}^\sharp$ will be called the *analysis space* and $V^* : \mathscr{H} \to \mathscr{H}^\sharp$ the *analysis operator*. Similarly, $V : \mathscr{H}^\sharp \to \mathscr{H}$ will be called the *synthesis operator*. The use of this terminology is directly analogous to the way it is used for frames and their generalizations in Sect. 2. Further, the Naimark representation also provides a means, via the spectral measure S, for keeping track of the labeling of the POVM.

Analysis of an element $f \in \mathscr{H}$ is the $\mathscr{H}^\sharp$-valued measure $\mathscr{A}$ on $\mathscr{B}(\Omega)$ defined by

$$\mathscr{A}(f)(\omega) = \hat{f}(\omega) = S(\omega)V^* f \in \mathscr{H}^\sharp. \tag{16}$$

In the case of a frame $\{\varphi_k : k \in K\}$, this measure on subsets of $\Omega = K$ associates the "coefficients" $\langle f, \varphi_k\rangle\, e_k \in \ell_2(K)$ with the signal f, where $\{e_k : k \in K\}$ is the standard basis of $\ell_2(K)$. Given a measure $\rho : \mathscr{B}(\Omega) \to \mathscr{H}^\sharp$ as in Eq. (16), the synthesis operator takes ρ to

$$\mathscr{S}(\rho) = V \int_\Omega \mathrm{d}\rho(t) \in \mathscr{H}. \tag{17}$$

As the examples in the preceding sections show, these analysis and synthesis operators correspond precisely to those of classical frames, fusion frames, and generalized fusion frames.

6 Isomorphism of POVMs

Two POVMs $(M_1, \Omega, \mathscr{H}_1)$ and $(M_2, \Omega, \mathscr{H}_2)$ are isomorphic if there is a surjective unitary transformation $U : \mathscr{H}_1 \to \mathscr{H}_2$ such that $UM_1(\omega)U^{-1} = M_2(\omega)$ for all $\omega \in \mathscr{B}(\Omega)$. The following result is a straightforward consequence of the proof of the Naimark theorem.

Theorem 2. *Suppose that POVMs $M_1 : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{H}_1)$ and $M_2 : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{H}_2)$ are isomorphic via the unitary transformation $U : \mathscr{H}_1 \to \mathscr{H}_2$. Let $(S_1, \mathscr{H}_1^\sharp, V_1)$ and $(S_2, \mathscr{H}_2^\sharp, V_2)$ be minimal Naimark representations of M_1 and M_2, respectively. Then there is a unitary transformation $U^\sharp : \mathscr{H}_1^\sharp \to \mathscr{H}_2^\sharp$ such that $U^\sharp S_1(\omega)(U^\sharp)^{-1} = S_2(\omega)$ for all $\omega \in \mathscr{B}(\Omega)$ and the following diagram commutes:*

$$\begin{array}{ccc} \mathscr{H}_1 & \overset{U}{\to} & \mathscr{H}_2 \\ V_1 \uparrow & & \uparrow V_2 \\ \mathscr{H}_1^\sharp & \underset{U^\sharp}{\to} & \mathscr{H}_2^\sharp \end{array} .$$

Although this result does not appear to be explicitly stated in the literature, it is implicit in many applications of the Naimark and Stinespring theorems. In particular, the paper of Arveson [2] discusses related ideas. The proof follows by consideration of the construction of the Naimark representation using Hilbert-space-valued functions as described in Sect. 4. Specifically, using the notation of the sketch proof of Naimark's theorem given in Sect. 4, observe that for the isomorphic POVMS M_1 and M_2, U gives rise to a map $\mathscr{L}_1 \to \mathscr{L}_2$ taking ξ_ω to $U(\xi)\omega$ which then produces $U^\sharp$. Moreover, it follows from the definition of the spectral measure in Eq. (12) that $U^\sharp S_1(\omega)(U^\sharp)^{-1} = S_2(\omega)$ for all $\omega \in \mathscr{B}(\Omega)$.

7 Canonical Representations and POVMs

Combining the Naimark theorem and Theorem 2 with the canonical representation of spectral POVMs (described in, e.g., [21]) yields a canonical representation of POVMs such that two isomorphic POVMs have the same canonical representation. This serves to illustrate the utility of the POVM formalism. The canonical representation decomposes $\mathscr{H}^\sharp$, the analysis space of a POVM $M : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{H})$ that arises in its Naimark representation, into a direct sum $\bigoplus_{n\in\mathbb{N}} \mathscr{G}_n$ such that:

1. Each of the spaces $\mathscr{G}_n$ is invariant under the spectral measure, i.e.,

$$S(\omega)\mathscr{G}_n \subset \mathscr{G}_n \qquad \omega \in \mathscr{B}(\Omega), n \in \mathbb{N}.$$

2. The restriction of S to $\mathscr{G}_n$ has uniform multiplicity, i.e., $\mathscr{G}_n \simeq \mathbb{C}^{u_n} \otimes L_2(\mu_n)$ if $\mathscr{G}_n$ has finite dimension u_n, and $\mathscr{G}_n \simeq \ell_2(\mathbb{N}) \otimes L_2(\mu_n)$ if $\mathscr{G}_n$ is infinite-dimensional.

This representation is essentially unique up to unitary equivalence and replacement of each of the measures μ_n by one having the same null sets. Denote by P_n the projection into $\mathscr{G}_n$, regarded as a subspace of $\mathscr{H}^\sharp$. Under the (minimal) Naimark representation, $V : \mathscr{H}^\sharp \to \mathscr{H}$ is such that $V^* S(\omega) V = M(\omega)$ for $\omega \in \mathscr{B}(\Omega)$. V can be decomposed as $V = \sum_n VP_n = \sum_n V_n$. The image of V_n^* is in $\mathscr{G}_n$, so that

$$M(\omega) = \sum_n V_n S(\omega) V_n^*, \qquad \omega \in \mathscr{B}(\Omega).$$

The map $M_n : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{G}_n)$, defined by $M_n(\omega) = V_n S(\omega) V_n^*$, is a POVM; more precisely, the values of M_n are positive operators on the closure of the image of V_n. The individual measures M_n are themselves POVMs, though they need not be framed even if M is framed. However, $M_n(\Omega) = V_n V_n^*$. Thus, if M is a framed POVM with frame bounds $A \leq B$,

$$A\mathbb{I}_{\mathscr{H}} \leq \sum M_n(\Omega) = \sum_n V_n V_n^* \leq B\mathbb{I}_{\mathscr{H}}.$$

Observe that $V_n^* V_n V_m^* V_m = 0$ for $n \neq m$, since the image of V_m^* lies in $\mathscr{G}_m$ which is in the kernel of V_n. So, in an obvious sense,

$$M = \sum_{n \in \mathbb{N}} M_n.$$

Thus every framed POVM is a sum of "uniform multiplicity" POVMs, though these need not be framed, and this composition is essentially unique. The canonical representation is characterized by the sequence of equivalence classes of measures $\{[\mu_n] : n \in \mathbb{N}\}$ and the linear map V.

Example 3. Consider a frame $\mathscr{F} = \{\varphi_k : k \in K\}$ in $\mathscr{H}$ with frame bounds $A \leq B$ and its corresponding framed POVM M. In this case $\mathscr{H}^\sharp$ is $\ell_2(K)$; $V : \mathscr{H}^\sharp \to \mathscr{H}$ is given by $V(e_k) = \varphi_k$. The spectral measure on the subsets of K is given by

$$S(J) = \sum_{k \in J} \Pi_k, \qquad J \subset K,$$

where Π_k denotes projection into the subspace of $\ell_2(K)$ spanned by the standard basis element e_k. Alternatively, this can be redefined by regarding members of

$\ell_2(K)$ as complex-valued functions on $\Omega = K$ and taking $S(J)(f) = \mathbb{1}_J f$ so that the spectral measure is uniform with multiplicity one.

Example 4. The case of a fusion frame $\{(W_k, w_k) : k \in K\}$ is more complicated than the frame case. The spectral measure S on subsets of $\Omega = K$ is given by Eq. (13). For each $j \in K$, denote

$$U_j = \{k \in K : \dim W_k = j\}.$$

Then

$$Y_j = \bigoplus_{k \in U_k} W_k \subset \mathscr{H}^\sharp$$

is isomorphic to $\mathbb{C}^j \otimes \ell_2(U_k)$ or, if $j = \infty$, $\ell_2(U_j) \otimes \ell_2(U_j)$. Evidently, Y_j has uniform multiplicity j, and the measure μ_j is counting measure on U_j, provided U_j is not empty. If all W_k for $k \in K$ have the same dimension, then the spectral measure S has uniform multiplicity.

8 Dual POVMs

As observed in Sect. 2, each of frame generalizations associates a "dual" object with the frame, and there is a canonical dual in each case. This is also possible for framed POVMs and indeed is relatively straightforward using the Naimark representation. Consider a POVM $M : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{H})$ and its minimal Naimark representation $(\Omega, S, \mathscr{H}^\sharp, V)$. The *canonical dual POVM* to M is the POVM $\widetilde{M} : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{H})$ having Naimark representation $(\Omega, S, \mathscr{H}^\sharp, (VV^*)^{-1}V)$, i.e.,

$$\widetilde{M}(\omega) = (VV^*)^{-1}VS(\omega)V^*(VV^*)^{-1}.$$

The frame condition on M guarantees $0 < A \leq V^*V \leq B < \infty$, which not only ensures the existence of $(VV^*)^{-1}$, but implies $\widetilde{M}$ is a framed POVM with bounds $B^{-1} \leq A^{-1}$ (see Theorem 1). Further,

$$M(\omega)\widetilde{M}(\omega) = \big(VS(\omega)V^*\big)\big((VV^*)^{-1}VS(\omega)V^*(VV^*)^{-1}\big)$$
$$\widetilde{M}(\omega)M(\omega) = \big((VV^*)^{-1}VS(\omega)V^*(VV^*)^{-1}\big)\big(VS(\omega)V^*\big)$$

In particular, invoking the assumption $S(\Omega) = \mathbb{I}_{\mathscr{H}^\sharp}$ gives

$$M(\Omega)\widetilde{M}(\Omega) = \widetilde{M}(\Omega)M(\Omega) = \mathbb{I}_{\mathscr{H}}.$$

From the point of view of analysis and synthesis, if $f \in \mathscr{H}$, its analysis with respect to M is the measure $\mathscr{A}(f)$ given in Eq. (16). Subsequently applying the synthesis operator $\widetilde{\mathscr{S}}$ associated with the canonical dual POVM $\widetilde{M}$ yields (17) gives

$$\widetilde{\mathscr{S}}\mathscr{A}(f)(\Omega) = (VV^*)^{-1}VS(\Omega)V^*f = f.$$

Similarly, analysis of f by $\widetilde{M}$ followed by synthesis with M is also the identity, i.e.,

$$\mathscr{S}\widetilde{\mathscr{A}}(f)(\Omega) = VS(\Omega)V^*(VV^*)^{-1}f = f.$$

9 Radon-Nikodym Theorem for POVMs

This section summarizes some results pertinent to framed POVMs on finite-dimensional Hilbert spaces. This setting is prevalent in signal processing applications, and it will be seen that the theory developed is valid in a number of infinite-dimensional examples as well. In this setting, the concept of a framed POVM is identical to that of a generalized fusion frame, as described in Sect. 1.

Let $M : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{H})$ be a framed POVM where $\dim \mathscr{H}$ is finite. The finite-dimensional assumption on $\mathscr{H}$ allows definition of a real-valued Borel measure $\mu(\omega) = \mathrm{Tr}(M(\omega))$ on the Borel sets of Ω. This positive regular Borel measure is a key element in the following *Radon-Nikodym theorem* for POVMs (see [6]).

Theorem 3. *Let $M : \mathscr{B}(\Omega) \to \mathscr{P}(\mathscr{H})$ be a POVM with $\mathscr{H}$ finite-dimensional. Then there exists a regular positive real-valued measure μ on $\mathscr{B}(\Omega)$ and an operator-valued bounded measurable function $r : \Omega \to \mathscr{P}(\mathscr{H})$ such that*

$$M(\omega) = \int_\omega r(t)\,\mathrm{d}\mu(t), \qquad \omega \in \mathscr{B}(\Omega).$$

The measure μ is called the *base measure* of the POVM and r the *Radon-Nikodym derivative* of the POVM M with respect to μ. This representation is useful in facilitating constructions of POVMs when $\mathscr{H}$ is finite-dimensional.

Corollary 1. *If M is a framed POVM with frame bounds $A \le B$, then*

$$A\mathbb{I}_{\mathscr{H}} \le \int_\Omega r(t)\,\mathrm{d}\mu(t) \le B\mathbb{I}_{\mathscr{H}}$$

It is instructive to observe how this Radon-Nikodym theorem manifests in the motivating examples. In particular, this result shows that, when $\mathscr{H}$ is finite-dimensional, framed POVMs correspond exactly to generalized fusion frames as introduced in Sect. 1.

Example 5. Let $\mathscr{F} = \{\varphi_k : k \in K\}$ be a frame in $\mathscr{H}$. The associated POVM is given by $M(J) = \sum_{k\in J} \Pi_k$ for subsets J of $\Omega = K$. In this case, the operator-valued function r is given by

$$r(k) = \Pi_k, \qquad k \in K.$$

In this special case, there is no need for the finite-dimensional restriction on $\mathscr{H}$. A POVM constructed from a frame in this way automatically possesses a Radon-Nikodym derivative with respect to counting measure on the subsets of K.

Example 6. In the case of a generalized frame $\Phi : \Omega \to \mathscr{H}$ for a Hilbert space $\mathscr{H}$, the associated POVM is given in Eq. (9). In this case, the operator-valued function is $r(t) = \Pi_{\Phi(t)}$. As in the previous case, a POVM constructed in this way satisfies a Radon-Nikodym theorem with respect to the given measure μ on Ω even when $\mathscr{H}$ is not finite-dimensional.

Example 7. For a fusion frame $\{(W_k, w_k) : k \in K\}$, $\Omega = K$ and μ is counting measure on subsets of K. The function $r : \mathscr{B}(K) \to \mathscr{P}(\mathscr{H})$ is given by $r(k) = w_k^2 \Pi_{W_k}$, which coincides with the previous observation that the POVM in this case is defined by

$$M(\omega) = \sum_{k \in \omega} w_k^2 \Pi_{W_k}, \qquad \omega \subset K.$$

Although the values of r are not projections, they are nonnegative multiples of projections. If the counting measure μ were replaced by $\nu(k) = w_k^2$, then the expression (9) for M would become

$$M(\omega) = \int_\omega \Pi_{W_k} \, \mathrm{d}\nu(k), \qquad \omega \subset K,$$

and the Radon-Nikodym derivative of M with respect to ν would have true projections as its values.

A POVM $M : \Omega \to \mathscr{P}(\mathscr{H})$ is *decomposable* if there is an essentially bounded measurable function $r : \Omega \to \mathscr{P}(\mathscr{H})$ and a measure μ on $\mathscr{B}(\Omega)$ such that

$$M(\omega) = \int_\omega r(t) \, \mathrm{d}\mu(t) \qquad \omega \in \mathscr{B}(\Omega).$$

As observed above, if $\dim \mathscr{H}$ is finite, the POVM is decomposable. Further, every POVM arising from a (generalized) frame is decomposable, even when $\mathscr{H}$ is not finite-dimensional. In effect, decomposable framed POVMs correspond to generalized fusion frames as described in Sect. 1, and thus this concept captures the simultaneous generalization of frames to fusion frames and generalized frames.

10 Conclusions

In this overview, we have set forth the concept of a framed positive-operator-valued measure and shown that classical frames, as well as several generalizations of frames, arise as special cases of framed POVMs. We have described how Naimark's theorem for POVMs leads to notions of analysis and synthesis for POVMs that subsume their frame counterparts. We have further discussed how

canonical representations of spectral POVMs lead to canonical descriptions of framed POVMs and that this leads to a notion of a canonical dual POVM analogous to that of the canonical dual of a frame.

Acknowledgments The authors are grateful to Somantika Datta and Benjamin Robinson for their helpful remarks on earlier drafts of this chapter.

References

1. Ali, S.T., Antoine, J.P., Gazeau, J.P.: Continuous frames in Hilbert space. Ann. Phys. **222**(1), 1–37 (1993)
2. Arveson, W.B.: Subalgebras of C*-algebras. Acta Mathematica **123**, 141–224 (1969)
3. Asgari, M.S., Khosravi, A.: Frames and bases of subspaces in Hilbert spaces. J. Math. Anal. Appl. **308**, 541–553 (2005)
4. Beukema, R.: Positive Operator-Valued Measures and Phase-Space Representations, Proefschrift, Technische Universiteit Eindhoven (2003)
5. Casazza, P.G., Kutyniok, G.: Frames of subspaces, In: Wavelets, Frames and Operator Theory. American Mathematical Society, vol. 345, pp. 87–113 (2004)
6. Chiribella, G., D'Ariano, G.M., Schlingemann, D.: Barycentric decomposition of quantum measurements in finite dimensions. J. Math. Phys. **51**, 022111-01–02111-16 (2010)
7. Christensen, O.: An Introduction to Frames and Riesz Bases. Birkhäuser, Boston (2003)
8. Daubechies, I., Grossmann, A., Meyer, Y.: Painless nonorthogonal expansions. J. Math. Phys. **27**, 1271–1283 (1986)
9. Dixmier, J.: Von Neumann Algebras. North-Holland, Amsterdam (1981)
10. Duffin, R.J., Schaeffer, A.C.: A class of nonharmonic Fourier series. Trans. Am. Math. Soc. **72**, 341–366 (1952)
11. Feichtinger, H.G., Strohmer, T.: Gabor Analysis and Algorithms: Theory and Applications. Birkhäuser, Boston (1998)
12. Fornasier, M.: Quasi-orthogonal decompositions of structured frames. J. Math. Anal. Appl. **289**, 180–199 (2004)
13. Gavruta, P.: On the duality of fusion frames. J. Math. Anal. Appl. **333**, 871–879 (2007)
14. Gröchenig, K.: Foundations of Time-Frequency Analysis. Birkhäuser, Boston (2001)
15. Heil, C., Walnut, D.:Continuous and discrete wavelet transforms. SIAM Rev. **31**, 628–666 (1989)
16. Naimark, M.A.: On a representation of additive operator set functions. Dokl. Acad. Sci. SSSR **41**, 373–375 (1943)
17. Parthasarathy, K.R.: Extremal decision rules in quantum hypothesis testing. In: Infinite Dimensional Analysis, Quantum Probability and Related Topics, vol. 2, pp. 557–568, World Scientific Singapore (1999)
18. Renes, J.M., Blume-Kohout, R., Scott, A.J., Caves, C.M.: Symmetric informationally complete quantum measurements. J. Math. Phys. **45**(6), 2171–2180 (2004)
19. Stinespring, W.F.: Positive Functions on C^*-algebras. Proc. Am. Math. Soc. **6**, 211–216 (1955)
20. Sun, W.: G-frames and G-Riesz bases. J. Math. Anal. Appl. **322**, 437–452 (2006)
21. Sunder, V.S.: Functional Analysis: Spectral Theory. Birkhäuser, Basel (1997)

Part VI
Filtering

The theory of filters, their design, and analysis take a prominent role in signal and image processing. The first examples of filtering appeared independently of and well before the introduction of automated computational devices. However, it was the applications enabled by the appearance of computers that transformed filtering from an aide to solve differential equations into a field of its own. Today, when we think of filters, we think, for example, of Wiener and Kalman filters and also about the impact that the modern theory of filters has on applications of harmonic analysis, for example, in terms of wavelet theory.

In the first chapter of this part, Daniel Alpay, Palle Jorgensen, and Izchak Lewkowicz discuss various aspects of wavelet filters. Nearly all practical implementations of wavelet transforms rely on finite impulse response (FIR) filters. In their work, the authors explore the extension of this classical understanding of wavelet filters, by considering them as matrix-valued meromorphic functions. This leads them to the introduction of generalized Schur functions, their associated reproducing kernel Pontryagin spaces, and the Cuntz relations. As such, this chapter provides a fascinating mathematical bridge between system theory and wavelet analysis. A broad introduction, aimed at both mathematicians and engineers, makes this work particularly useful.

Inspired by his work on the JPEG-2000 standard, Christopher M. Brislawn ventures into the world of group-theoretic characterizations of perfect reconstruction filter banks. This practical motivation yields a restriction of the broad class of FIR filters, studied in the previous chapter, to filter banks with well-behaved lifting factorizations: whole-sample (WS) symmetric and half-sample (HS) symmetric filter classes. Although the technique of lifting, introduced by Ingrid Daubechies and Wim Sweldens, found many applications in signal processing, here we are presented with a completely new approach. The notion of group lifting structure is defined, and its irreducible factorizations are analyzed. The use of group theory allows the author to find striking differences between WS and HS filters.

Shidong Li and Michael Hoffman explore the class of filter banks associated with biorthogonal wavelets through the notion of pseudoframes for subspaces (PFFS). PFFS were introduced by Li and Ogawa to overcome certain important limitations which arise in the theory of frames. This chapter starts with an overview of the state of the art in PFFS and proceeds to exploit the PFFS-based representations of perfect reconstruction biorthogonal filters. The authors show how the flexibility of PFFS aids in construction of biorthogonal filters with various desired properties, such as a required number of vanishing moments or a maximized coding gain.

FIR filters are also in the focus of the chapter written by Yang Wang and Zhengfang Zhou. Here, FIR filters are identified with banded Toeplitz matrices—diagonal-constant matrices studied by Otto Toeplitz and Gábor Szegö. Szegö limit theorems are among the most important results describing Toeplitz matrices, as they deal with the limit behavior of their eigenvalues. This, in turn, leads naturally to questions concerning convergence of iterations of banded Toeplitz operators. These types of questions arise independently in signal and data processing, through the so-called empirical mode decomposition (EMD). This chapter provides a mathematical

foundation underlying EMD theory through the introduction of FIR filters and the characterization of the eigenvectors of their banded Toeplitz operators.

Wei Zhu and M. Victor Wickerhauser study the ever-important question of finding most effective implementations of discrete wavelet transforms. Lifting analysis, discussed in the chapter by C. Brislawn, provides one way to achieve such effective algorithms. Among the most fundamental questions that need to be addressed are Daubechies and Sweldens listed analysis and exploitation of the non-uniqueness which arises in lifting through the Euclidean algorithm. In this chapter the authors provide important improvements to the lifting algorithm, by reducing the number of distant memory accesses and by showing that certain equivalent lifting factorizations possess much worse complexity and propagation-of-error properties. All this is achieved by exploiting the nonuniqueness of lifting factorizations and by building lifting steps that are optimized for these specific tasks, for example, minimizing the number of involved nearest neighbor array elements.

Extending Wavelet Filters: Infinite Dimensions, the Nonrational Case, and Indefinite Inner Product Spaces

Daniel Alpay, Palle Jorgensen, and Izchak Lewkowicz

Abstract In this chapter we are discussing various aspects of wavelet filters. While there are earlier studies of these filters as matrix-valued functions in wavelets, in signal processing, and in systems, we here expand the framework. Motivated by applications and by bringing to bear tools from reproducing kernel theory, we point out the role of non-positive definite Hermitian inner products (negative squares), for example, Krein spaces, in the study of stability questions. We focus on the nonrational case and establish new connections with the theory of generalized Schur functions and their associated reproducing kernel Pontryagin spaces and the Cuntz relations.

Keywords Cuntz relations • Schur analysis • Wavelet filters • Pontryagin spaces

MSC classes: 65T60, 46C20, 93B28

D. Alpay
Department of Mathematics, Ben Gurion University of the Negev,
P.O.B. 653, Be'er Sheva 84105, Israel
e-mail: daniel.alpay@gmail.com

P. Jorgensen (✉)
Department of Mathematics, 14 MLH, The University of Iowa City,
IA 52242-1419, USA,
e-mail: jorgen@math.uiowa.edu

I. Lewkowicz
Department of Electrical Engineering, Ben Gurion University of the Negev,
P.O.B. 653, Be'er Sheva 84105, Israel
e-mail: izchak@ee.bgu.ac.il

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_5,

1 Introduction

Roughly speaking, systems whose inputs and outputs may be viewed as *signals* are called *filters*. Mathematically, filters are often presented as operator-valued functions of a complex variable. In applications, filters are used in areas as (i) prediction, (ii) signal processing, (iii) systems theory, and (iv) Lax–Phillips scattering theory [55]. There, one is faced with spectral theoretic questions which can be formulated and answered with the use of a suitable choice of an operator-valued function defined on a domain in complex plane; in the case of scattering theory, the scattering operator and the scattering matrix; in the other areas, the names used include polyphase matrix; see, for example, [43, 49]. We also mention that more recently, filters are used in (iv) multiresolution analysis (MRA) in wavelets. We follow standard conventions regarding time-frequency duality, i.e., the correspondence between discrete time on one side and a complex frequency variable on the other. In the simplest cases, one passes from a time series to a generating function of a complex variable. These frequency response functions fall in various specific classes of functions of a complex variable; the particular function spaces in turn are dictated by applications. Again, motivated by applications, in our present study, we adopt a wider context for both sides of the duality divide. On the frequency side, we work with operator-valued functions. This framework is relevant to a host of applications, and we believe of independent interest in operator theory. From the literature, we mention [22, 57] (see also [21]) and the papers referenced below.

We here consider the set of $\mathbb{C}^{N\times N}$-valued functions meromorphic in the open unit disk $\mathbb{D}$[1] and define two subsets of it: We shall denote by $\mathscr{C}_N$ the family satisfying the symmetry

$$W(\epsilon_N z) = W(z)P_N, \tag{1}$$

where $\epsilon_N = \mathrm{e}^{\frac{2\pi i}{N}}$ and P_N denote the permutation matrix

$$P_N = \begin{pmatrix} 0_{1\times(N-1)} & 1 \\ I_{N-1} & 0_{(N-1)\times 1} \end{pmatrix}. \tag{2}$$

We shall also denote by $\mathscr{U}^{I_N}$ the set of $\mathbb{C}^{N\times N}$-valued functions which take unitary values[2] on the unit circle $\mathbb{T}$.

Classically *wavelet filters*, denoted by $\mathscr{W}_N$, are characterized by rational functions satisfying both symmetries, i.e.,

$$\mathscr{W}_N = \mathscr{U}^{I_N} \cap \mathscr{C}_N. \tag{3}$$

In a previous paper (see [9]), we have provided an easy-to-compute characterization of $\mathscr{W}_N$ as both a set of rational functions and in terms of state space realization.

[1] Classically, in the engineering literature, the functions are analytic, or more generally meromorphic, outside the closed unit disk. The map $z \mapsto 1/z$ relates the two settings.

[2] For rational functions, the term *para-unitary* is also used in the engineering literature.

The aim of this work is to explore the possibility of extending the notion of wavelet filters, described in Eq. (3). The functions considered still satisfy the symmetry in Eq. (1), but:

- The functions are not necessarily rational or finite dimensional.
- The functions are not necessarily unitary on the unit circle $\mathbb{T}$.
- The functions are meromorphic (rather than analytic) in $\mathbb{D}$.

To explain our strategy, first recall the following: If W is a $\mathbb{C}^{N\times N}$-valued function which is rational and takes unitary values on the unit circle, the kernel

$$K_W(z,w) = \frac{I_N - W(z)W(w)^*}{1 - zw^*}$$

is positive definite in the open unit disk $\mathbb{D}$ if W has no poles there or more generally has a finite number of negative squares in $\mathbb{D}$. See Definition 3.4 for the latter. In our approach, unitarity on the unit circle is replaced by the requirement that W is a generalized Schur function, in the sense that W is meromorphic in $\mathbb{D}$ and the associated kernel $K_W(z,w)$ has a finite number of negative squares there. *This family includes in particular the case of matrix-valued rational functions which take contractive values on the unit circle.* We will also consider the case where the values on the unit circle are, when defined, contractive with respect to indefinite metrics. These kernels are of the form

$$\frac{J_2 - W(z)J_1W(w)^*}{1 - zw^*} \tag{4}$$

when W is $\mathbb{C}^{p_2\times p_1}$-valued and analytic in a neighborhood of the origin, and where J_1 and J_2 are signature matrices, respectively, in $\mathbb{C}^{p_1\times p_1}$ and $\mathbb{C}^{p_2\times p_2}$, which have the same number of strictly negative eigenvalues:

$$\nu_-(J_1) = \nu_-(J_2), \tag{5}$$

and such that the kernel K_W has a finite number of negative squares. In [9] we studied the realization of wavelet filters in the $\mathbb{C}^{N\times M}$-valued (with $M \geq N$) rational case. The above approach allows us to extend these results to the case where the filter is not necessarily rational and M may be smaller than N. Furthermore, the conditions in [9] of the function being analytic in the open unit disk, and taking coisometric values on the unit circle, are both relaxed (in particular, in the previous case, in Eq. (5), we had $J_1 = I_M$ and $J_2 = I_N$).

This chapter is organized as follows. Since we address different audiences, Sects. 2–4 are of a review nature. In Sect. 2, we give background on the use of filters in mathematics. We note that the more traditional framework in the literature has so far been unnecessarily restricted by two kinds of technical assumptions: (i) restricting to rational operator-valued functions and (ii) restricting the range of the operator-valued functions considered. In Sect. 3 we address indefinite inner product spaces and survey the theory of Pontryagin and Krein spaces. This overview allows us in Sect. 4 to describe a setting that expands both the above-mentioned restrictions in (i) and (ii), namely the theory of generalized Schur functions. Our results in

Sects. 5 and 6 (Theorems 5.3, 5.4, and 6.4) deal with representations. We use these results in obtaining classifications and decomposition theorems. In Sect. 7, we employ these theorems in the framework of wavelets.

2 Some Background

2.1 Cuntz Relations

The Cuntz relations were realized by Cuntz in [24] as generators of a simple purely infinite C^*-algebra. Since then, they found many applications, and the related literature about Cuntz relations has flourished. Since Cuntz's paper [24], the study of their representations has mushroomed, and now makes up a big literature, see, for example, [13, 18–20, 25, 37, 39], and some of their applications [32, 38, 40–42], for example, to fractals [31].

In the initial framework, one is given a finite set $S_1, \ldots, S_N$ of isometries with orthogonal ranges adding up to the whole Hilbert space. Their representations play a role in a variety of applications, for example, wavelets, and more generally multi-scale phenomena. The study of what are called non-type I C^*-algebras was initiated in the pioneering work of Glimm [34, 35] and Dixmier [28]. This in turn was motivated by use of direct integrals in representation theory, both in the context of groups and C^*-algebras. Direct integrals of representations are done practically with the use of Borel cross sections. Glimm proved that there are purely infinite C^*-algebras which do not admit Borel cross sections as a parameter space for the set of equivalence classes of irreducible representations; the Cuntz algebra(s) O_N is the best-known examples [24]. Nonetheless, it was proved in [18] that there are families of equivalence classes of representations of O_N indexed by wavelet filters, the latter in turn being indexed by infinite-dimensional groups.

One illustration of the need for expanding the framework of O_N from Hilbert space to the case of Krein spaces is illustrated by applications to scattering theory for the automorphic wave equation [54]. The initial study was restricted to the case when the operators S_i act on Hilbert space and when they act isometrically. However, since then, there has been a need for generalizing the Cuntz relations. It was noted in [19] that the isometric case adapts well to the restricted framework of orthogonal wavelet families [26]. Nonetheless, applications to engineering dictate much wider families, such as wavelet frames.

In this work we extend what is known in the literature in a number of different directions, including to the case of Pontryagin spaces. We obtain Cuntz relations for isometries between certain reproducing kernel Pontryagin spaces of analytic functions.

2.2 *Wavelet Filters*

In electrical engineering terminology, systems whose inputs and outputs may be viewed as *signals* are called *filters*. By filter, we here mean functions $W(z)$ defined on the disk in the complex plane and taking operator values, i.e., linear operators mapping between suitable spaces.

While filters (in the sense of systems and signal processing) have already been used with success in analysis of wavelets, so far some powerful tools from systems theory have not yet been brought to bear on wavelet filters. The traditional restriction placed on these functions $W(z)$ is that they are rational and take values in the unitary group when z is restricted to have modulus 1. In models from systems theory, the complex variable z plays the role of complex frequency. A reason for the recent success of wavelet algorithms is a coming together of tools from engineering and harmonic analysis. While wavelets now enter into a multitude of applications from analysis and probability, it was the incorporation of ideas from signal processing that offered new and easy-to-use algorithms, and hence wavelets are now used in both discrete problems, as well as in harmonic analysis decompositions. It is our purpose to use tools from systems theory in wavelet problems and also show how ideas from wavelet decompositions shed light on factorizations used by engineers. Each of the various wavelet families demands a separate class of filters, for the case of compactly supported biorthogonal wavelets, see, for example, Resnikoff, Tian, Wells [60] and Sebert and Zou [63]. By now there is a substantial literature on the use of filters in wavelets (see, e.g., [18, 26, 37, 39]). For filters in wavelets, there are two pioneering papers [50, 51] and the book [56].

In a previous work [9] we characterized all rational wavelet filters attaining unitary values on the unit circle. It turned out that this family is quite small (and in particular the subset of finite impulse response filters, commonly used in engineering).

Thus, we here remove both restrictions on the filters, i.e., rational and unitary, and consider $W(z)$ which are generalized Schur functions and use reproducing kernel Pontryagin spaces associated with W. See [6] for background.

We hope that this message will be useful to practitioners in their use of these rigorous mathematics tools.

3 Pontryagin Spaces and Krein Spaces

For a number of problems in the study of signals and filters (e.g., stability considerations), it is necessary to work with Hermitian inner products that are not positive definite. This view changes the Hermitian quadratic forms, allowing for negative squares, as well as the associated linear spaces. But more importantly, this wider setting also necessitates changes in the analysis, for example, in the meaning of the notion of the adjoint operator, as well as the reproducing kernels. There are a number of subtle analytic points involved, as well as a new operator theory. We turn to these details below.

3.1 Krein Spaces

A *Krein space* is a pair $(V, [\cdot,\cdot])$, where V is a linear vector space on $\mathbb{C}$ endowed with an Hermitian form $[\cdot,\cdot]$, and with the following properties: V can be written as $V = V_+ + V_-$, where:

1. V_+ endowed with the Hermitian form $[\cdot,\cdot]$ is a Hilbert space.
2. V_- endowed with the Hermitian form $-[\cdot,\cdot]$ is a Hilbert space.
3. It holds that $V_+ \cap V_- = \{0\}$.
4. For all $v_\pm \in V_\pm$,

$$[v_+, v_-] = 0.$$

The representation $V = V_+ + V_-$ is called a *fundamental decomposition* and is highly nonunique as soon as $\dim V_- > 0$. Given such a decomposition, the map

$$\sigma(v_+ + v_-) = v_+ - v_-$$

is called a *fundamental symmetry*. Note that the space V endowed with the Hermitian form (where $w = w_+ + w_-$ is also an element of V, with $w_\pm \in V_\pm$)

$$\langle v, w\rangle = [v, \sigma w] = [v_+, w_+] - [v_-, w_-]$$

is a Hilbert space. These norms are called *natural norms*, and they are all equivalent. The Hilbert space topologies associated to any two such decompositions are equivalent, and V is endowed with any of them; see [15, p. 102]. When V_- is finite dimensional, V is called a Pontryagin space and the dimension of V_- is called the *negative index* (or the *index* for short) of the Pontryagin space. We refer to the books [6, 12, 15, 36] for more information on Krein and Pontryagin spaces. Note that in [36] it is the space V_+ rather than V_- which is assumed finite dimensional in the definition of a Pontryagin space. Surveys may be found in for instance in [7,29,30]. It is interesting to note that Laurent Schwartz introduced independently the notion of Krein and Pontryagin spaces (he used the terminology Hermitian spaces for Krein and Pontryagin spaces) in his paper [62]. For applications of Krein spaces to the study of boundary conditions for hyperbolic PDE, including wave equations, and exterior domains, see, for example, [23, 52, 53, 58]. We now give two examples, which will be important in the sequel.

Example 3.1. Let $J \in \mathbb{C}^{p\times p}$ be an Hermitian involution, i.e.,

$$J = J^{-1} = J^*.$$

Such a matrix is called a signature matrix. We denote by $\mathbb{C}_J$ the space $\mathbb{C}^p$ endowed with the associated indefinite inner product

$$[x, y]_J = y^* J x, \quad x, y \in \mathbb{C}^p.$$

It is a finite-dimensional Pontryagin space.

Example 3.2. Let J be a signature matrix. We consider the space $\mathbf{H}_2(\mathbb{D})^p$ of functions analytic in $\mathbb{D}$ and with values in $\mathbb{C}^p$:

$$f(z) = \sum_{n=0}^{\infty} a_n z^n, \quad a_n \in \mathbb{C}^p,$$

such that

$$\sum_{n=0}^{\infty} a_n^* a_n < \infty.$$

Then, $\mathbf{H}_2(\mathbb{D})^p$ endowed with the Hermitian form

$$[f, g]_J = \sum_{n=0}^{\infty} b_n^* J a_n \quad (with\ g(z) = \sum_{n=0}^{\infty} b_n z^n)$$

is a Krein space, which we denote by $\mathbf{H}_{2,J}(\mathbb{D})$.

In the above example, if $p = 1$ and $J = 1$ (as opposed to $J = -1$), the space $\mathbf{H}_{2,J}(\mathbb{D})$ is equal to the classical Hardy space $\mathbf{H}_2(\mathbb{D})$ of the unit disk.

3.2 Operators in Krein and Pontryagin Spaces

When one considers a bounded operator A between two Krein spaces $(\mathscr{K}_1, [\cdot,\cdot]_1)$ and $(\mathscr{K}_2, [\cdot,\cdot]_2)$ (in this chapter, it will be most of the time between two Pontryagin spaces) the adjoint can be computed in two different ways, with respect to the Hilbert spaces inner products (and then we use the notation A^*) and with respect to the Krein spaces inner products (and then we use the notation $A^{[*]}$). More precisely, if σ_1 and σ_2 are fundamental symmetries in $\mathscr{K}_1$ and $\mathscr{K}_2$ which define the Hilbert spaces inner products

$$\langle f_1, g_1\rangle_1 = [\sigma_1 f_1, g_1]_1 \quad \text{and} \quad \langle f_2, g_2\rangle_2 = [\sigma_2 f_2, g_2]_2,$$

(with $f_1, g_1 \in \mathscr{K}_1$ and $f_2, g_2 \in \mathscr{K}_2$), we have for $f_1 \in \mathscr{K}_1$ and $f_2 \in \mathscr{K}_2$

$$\begin{aligned}[A f_1, f_2]_2 &= \langle \sigma_2 A f_1, f_2\rangle_2 \\ &= \langle f_1, A^* \sigma_2 f_2\rangle_1 \\ &= [f_1, A^{[*]} f_2]_1,\end{aligned}$$

with

$$A^{[*]} = \sigma_1 A^* \sigma_2. \tag{6}$$

In the case of $\mathbb{C}_J$ (see Example 3.1) we have

$$A^{[*]} = JA^*J. \tag{7}$$

The operator A from $\mathscr{D}(A) \subset \mathscr{K}_1$, where $(\mathscr{K}_1, [\cdot,\cdot]_1)$ is a Krein space, into the Krein space $(\mathscr{K}_2, [\cdot,\cdot]_2)$ is a contraction if

$$[Ak_1, Ak_1]_2 \leq [k_1, k_1]_1, \quad \forall k_1 \in \mathscr{D}(A).$$

A densely defined contraction, or even isometry, operator A between Krein spaces need not be continuous, let alone have a continuous extension. See for instance [29, Theorem 1.1.7]. In the case of Pontryagin spaces with same negative index, A has a continuous extension to all of $\mathscr{K}_1$; see [6, Theorem 1.4.1, p. 27] and Theorem 3.3. Even when it is continuous and has a well-defined adjoint, this adjoint need not be a contraction. The operator is called a bicontraction if both it and its adjoint are contractions. When the Krein spaces are Pontryagin spaces with same negative index, a contraction is automatically continuous and its adjoint is also a contraction. An important notion in the theory of Pontryagin spaces is that of *relation*. Given two Pontryagin spaces $\mathscr{P}_1$ and $\mathscr{P}_2$, a relation is a linear subspace of $\mathscr{P}_1 \times \mathscr{P}_2$. For instance the graph of an operator is a relation. The domain of the relation $\mathscr{R}$ is the set of $f \in \mathscr{P}_1$ such that there is a $g \in \mathscr{P}_2$ for which $(f, g) \in \mathscr{R}$. A relation $\mathscr{R}$ is called *contractive* if

$$[g, g]_2 \leq [f, f]_1 \quad \forall (f, g) \in \mathscr{R}.$$

A key result is the following theorem of Shmulyan (see [6, Theorem 1.4.1, p. 27]).

Theorem 3.3. *A densely defined contractive relation between Pontryagin spaces with same negative index extends to the graph of a uniquely defined contraction operator from $\mathscr{P}_1$ into $\mathscr{P}_2$.*

3.3 Kernels

Recall that a (say, matrix-valued) function $K(z, w)$ of two variables, defined for z and w in a set Ω, is called a positive definite kernel if it is Hermitian: $K(z, w)^* = K(w, z)$ for all $z, w \in \Omega$ and if for every choice of $M \in \mathbb{N}$ and $w_1, \ldots, w_M \in \Omega$ the $M \times M$ Hermitian block matrix with (ℓ, j) block entry $K(w_\ell, w_j)$ is nonnegative. For instance, if b is a finite Blaschke product,

$$b(z) = \prod_{n=1}^{m} \frac{z - a_n}{1 - za_n^*}$$

for some $a_1, \ldots, a_m$ in the open unit disk, the kernel

$$k_b(z,w) = \frac{1 - b(z)b(w)^*}{1 - zw^*}$$

is positive definite, as can be seen from the formula

$$k_b(z,w) = \langle k_b(\cdot,w), k_b(\cdot,z)\rangle_{\mathbf{H}_2(\mathbb{D})}.$$

When b is replaced with a function s analytic and contractive in the open unit disk, the corresponding kernel $k_s(z,w) = \frac{1-s(z)s(w)^*}{1-zw^*}$ is still positive definite in $\mathbb{D}$; see [16, 17]. This follows, for instance, from the fact that the operator of multiplication by s is a contraction from $\mathbf{H}_2(\mathbb{D})$ into itself. In the special case of a finite Blaschke product (or more generally, of an inner function), this multiplication operator is an isometry. This makes the underlying computations much easier. More generally, the kernels which appear in the following section can be seen as far reaching generalizations of the kernels $k_b(z,w)$.

The notion of positive definite kernel has been extended by Krein as follows:

Definition 3.4. 1 Let $\kappa \in \mathbb{N}_0$. A (say, matrix-valued) function $K(z,w)$ defined on a set Ω has κ negative squares if it is Hermitian and if for every choice of $M \in \mathbb{N}$ and $w_1, \ldots, w_M \in \Omega$ the $M \times M$ Hermitian block matrix with (ℓ, j) block entry $K(w_\ell, w_j)$ has at most κ strictly negative eigenvalues and exactly κ strictly negative eigenvalues for some choice of $M, w_1, \ldots, w_M$. When $\kappa = 0$, the function is positive definite.

The one-to-one correspondence between positive definite kernels and reproducing kernel Hilbert spaces was first extended to the indefinite case by Schwartz; see [62]: *There is a one-to-one correspondence between reproducing kernel Pontryagin spaces and kernels with a finite number of negative squares*. For completeness, we mention that such a result fails if the number of negative squares is not finite. *A necessary and sufficient condition for a function to be the reproducing kernel of a Krein space is that this function is the difference of two positive kernels, but the associated Krein space need not be unique*. Here too we refer to Schwartz [62] and also to the paper [1]. Realization of operator-valued analytic functions (without assumptions on an associated kernel but with some symmetry hypothesis) has also been considered. See for instance [27]. The $\mathbb{C}^{p\times p}$-valued function $K(z,w)$ defined for z, w in an open set Ω of the complex plane will be called an *analytic kernel* if it is Hermitian and if it is analytic in z and w^*. If it has moreover a finite number of negative squares, the elements of the associated reproducing kernel Pontryagin space are analytic in Ω. See [6, Theorem 1.1.2, p. 7].

There are two important classes of operators between reproducing kernel spaces, namely multiplication and composition operators. We conclude this section with three results on these operators.

Theorem 3.5. *Let $(\mathscr{K}_1, [\cdot,\cdot]_1)$ and $(\mathscr{K}_2, [\cdot,\cdot]_2)$ be two reproducing kernel Krein spaces of vector-valued functions, defined in Ω, and with reproducing kernels $K_1(z,w)$ and $K_2(z,w)$, respectively, $\mathbb{C}^{p_1\times p_1}$- and $\mathbb{C}^{p_2\times p_2}$-valued. Let m be a $\mathbb{C}^{p_2\times p_1}$-valued function and let φ be a map from Ω into itself. Assume that the map*

$$(T_{m,\varphi} f)(z) = m(z) f(\varphi(z)) \tag{8}$$

defines a bounded operator from $(\mathscr{K}_1, [\cdot,\cdot]_1)$ into $(\mathscr{K}_2, [\cdot,\cdot]_2)$. Then, for every $z, w \in \Omega$, and $\xi_2 \in \mathbb{C}^{p_2}$,

$$\left(T^{[*]}_{m,\varphi} K_2(\cdot, w)\xi_2\right)(z) = K_1(z, \varphi(w))m(w)^*\xi_2. \tag{9}$$

Proof. Let $z, w \in \Omega$, $\xi_2 \in \mathbb{C}^{p_2}$, and $\xi_1 \in \mathbb{C}^{p_1}$. We have

$$\begin{aligned}
\xi_1^* \left(T^{[*]}_{m,\varphi} K_2(\cdot, w)\xi_2\right)(z) &= [T^{[*]}_{m,\varphi} K_2(\cdot, w)\xi_2\,,\ K_1(\cdot, z)\xi_1]_1 \\
&= [K_2(\cdot, w)\xi_2\,,\ T_{m,\varphi}(K_1(\cdot, z)\xi_1)]_2 \\
&= [K_2(\cdot, w)\xi_2\,,\ m(\cdot)K_1(\varphi(\cdot), z)\xi_1]_2 \\
&= [m(\cdot)K_1(\varphi(\cdot), z)\xi_1\,,\ K_2(\cdot, w)\xi_2]_2^* \\
&= \left(\xi_2^* m(w)K_1(\varphi(w)\,,\ z)\xi_1\right)^* \\
&= \xi_1^* K_1(z, \varphi(w))m(w)^*\xi_2.
\end{aligned}$$

□

As a corollary we have the following result.

Theorem 3.6. *Assume in the preceding theorem that $\mathscr{K}_1$ and $\mathscr{K}_2$ are Pontryagin spaces with same negative index. Then, $T_{m,\varphi}$ is a contraction if and only if the kernel*

$$K_2(z, w) - m(z)K_1(\varphi(z), \varphi(w))m(w)^* \tag{10}$$

is positive definite in Ω.

Proof. Assume that T is a contraction. Then, its adjoint is also a contraction since the Pontryagin spaces have the same negative index. Let $g \in \mathscr{K}_2$ be of the form

$$g(z) = \sum_{k=1}^{N} K_2(z, w_k)\xi_k,$$

where $N \in \mathbb{N}$, $w_1, \ldots, w_N \in \omega$ and $\xi_1, \ldots, \xi_N \in \mathbb{C}^{p_2}$. By Eq. (9) we have

$$\begin{aligned}
&\sum_{\ell,k=1}^{N} \xi_\ell^* m(w_\ell) K_1(\varphi(w_\ell), \varphi(w_k)) m(w_k)^* \xi_k \\
&\quad = \left[\sum_{k=1}^{N} K_1(z, \varphi(w_k)) m(w_k)^* \xi_k , \sum_{\ell=1}^{N} K_1(z, \varphi(w_\ell)) m(w_\ell)^* \xi_\ell \right]_1 \\
&\quad = \left[T_{m,\varphi}^{[*]} g, T_{m,\varphi}^{[*]} g \right]_1 \\
&\quad \leq [g, g]_2 \\
&\quad = \sum_{\ell,k=1}^{N} \xi_\ell^* K_2(w_\ell, w_k) \xi_k ,
\end{aligned}$$

and hence the kernel Eq. (10) is positive definite. Conversely, assume that the kernel Eq. (10) is positive definite. Then the linear span of the pairs of functions

$$(K_2(\cdot, w)\xi \,,\; K_1(\cdot, \varphi(w)) m(w)^* \xi), \quad w \in \Omega, \;\; \xi \in \mathbb{C}^{p_2},$$

defines a linear densely defined contractive relation in $\mathscr{H}_1 \times \mathscr{H}_2$. By Shmulyan's theorem (see Theorem 3.3), this relation has an everywhere defined extension which is the graph of a bounded contraction: There is a unique contraction X from $\mathscr{H}_2$ into $\mathscr{H}_1$ such that

$$X(K_2(\cdot, w)\xi) = K_1(\cdot, \varphi(w)) m(w)^* \xi, \quad w \in \Omega, \;\; \xi \in \mathbb{C}^{p_2}.$$

By Eq. (9), we have $X^{[*]} = T_{m,\varphi}$, and this concludes the proof. □

We will consider in the sequel special cases of this result, in particular, when

$$m(z) = \left(1 \; z \cdots z^{N-1}\right);$$

see Theorem 5.3, or more generally when

$$m(z) = \left(m_0(z) \; m_1(z) \cdots m_{N-1}(z)\right),$$

see Theorem 5.4. The operator $T_{m,\varphi}$ defined by Eq. (8) is then a block operator, and its components satisfy, under appropriate supplementary hypothesis, the Cuntz relations formally defined in Eqs. (18)–(19).

We conclude this section with a result on composition operators in reproducing kernel Pontryagin spaces.

Theorem 3.7. *Let $K(z, w)$ be a $\mathbb{C}^{p \times p}$-valued function which has κ negative squares in the set Ω. The associated reproducing kernel Pontryagin space will be denoted by $\mathscr{P}(K)$. Let φ be a map from Ω into itself, and assume that*

$$f(\varphi(z)) \equiv 0 \Longrightarrow f \equiv 0$$

for $f \in \mathscr{P}(K)$. Then:
(a) The function $K_\varphi(z,w) = K(\varphi(z),\varphi(w))$ has at most κ negative squares in Ω, and its associated reproducing Pontryagin space is the set of functions of the form $F(z) = f(\varphi(z))$, with $f \in \mathscr{P}(K)$ and Hermitian form

$$[F,G]_{\mathscr{P}(K_\varphi)} = [f,g]_{\mathscr{P}(K)}. \tag{11}$$

(b) The map $f \mapsto f(\varphi)$ is unitary from $\mathscr{P}(K)$ into itself if and only if

$$K(z,w) = K(\varphi(z),\varphi(w)), \quad \forall z,w \in \Omega. \tag{12}$$

Proof. Set

$$\mathscr{M}_\varphi = \{f(\varphi(z)),\ f \in \mathscr{P}(K)\}.$$

By hypothesis, we have $f(\varphi(z)) \equiv 0$ if and only if $f \equiv 0$, and so the Hermitian form (11) is well defined and induces a Pontryagin structure on $\mathscr{M}_\varphi$. Furthermore, with $c \in \mathbb{C}^p$ and $F(z) = f(\varphi(z)) \in \mathscr{M}_\varphi$, we have

$$\begin{aligned} {}_{\mathscr{P}(K_\varphi)} &= [f(\cdot), K(\cdot,\varphi(w))c]_{\mathscr{P}(K)} \\ &= c^* f(\varphi(z)) \\ &= F(w), \end{aligned}$$

and hence the reproducing kernel property is in force. To prove (b) we use the uniqueness of the kernel for a given reproducing kernel Pontryagin space. □

To fine-tune the previous result, note that for $\varphi(z) = z^N$, the composition map is an isometry from $\mathbf{H}_2(\mathbb{D})$ into itself but is not unitary (unless $N = 1$). We also note that the preceding theorem holds also for reproducing kernel Krein spaces. Indeed, the correspondence between functions which are difference of positive functions on a given set and reproducing kernel Krein spaces is not one-to-one, but a given reproducing kernel Krein space has a unique reproducing kernel.

4 Generalized Schur Functions and Associated Spaces

In this section we review the main aspects of the realization theory of generalized Schur functions and of their associated reproducing kernel Pontryagin spaces.

4.1 Generalized Schur Functions

In the positive definite case, this theory originates with the works of de Branges and Rovnyak; see [16,17]. In earlier work on models involving operators in Hilbert space and matrix factorization, de Branges spaces have served as a surprisingly powerful tool. The theory was developed in the indefinite case in a fundamental series of papers by Krein and Langer, see for instance [44–48], and using reproducing kernel methods in [7] and in the book [6]. It was later used in [6, p. 119] and in the paper [3] to study generalized Schur functions with some given symmetry. In this chapter we use this setting to present nonrational and non unitary wavelet filters. In [6] the case of operator-valued functions is studied, but we here consider the case of $\mathbb{C}^{p\times p}$-valued functions. We now recall the definition of a generalized Schur function. A (say $\mathbb{C}^{p\times p}$-valued) function W is called a Schur function if it is analytic and contractive in the open unit disk, or, equivalently, if the associated kernel

$$K_W(z,w) = \frac{I_p - W(z)W(w)^*}{1 - zw^*} \tag{13}$$

is positive definite in a neighborhood of the origin. Then, it has a unique analytic extension to the open unit disk, and this extension is such that the kernel K_W is still positive definite in $\mathbb{D}$. There are two other kernels associated to W, namely the kernel $K_{\widetilde{W}}(z,w)$ (with $\widetilde{W}(z) \stackrel{\text{def.}}{=} W(z^*)^*$) and the kernel

$$D_W(z,w) = \begin{pmatrix} K_W(z,w) & \frac{W(z)-W(w^*)}{z-w^*} \\ \frac{\widetilde{W}(z)-\widetilde{W}(w^*)}{z-w^*} & K_{\widetilde{W}}(z,w) \end{pmatrix}.$$

These three kernels are simultaneously positive definite in the open unit disk. The first is the state space for a unique coisometric realization of W, the second is the state space for a unique isometric realization of W, and the reproducing kernel Hilbert space with reproducing kernel D_W is the state space for a unique unitary realization of W. In these three cases, uniqueness is up to an invertible similarity operator.

Let $J \in \mathbb{C}^{p\times p}$ be a signature matrix. We now consider functions with values in $\mathbb{C}_J$ defined in Example 3.1, denoted by Θ (rather than W). A $\mathbb{C}^{p\times p}$-valued function Θ analytic in a neighborhood of the origin is called J-contractive if the associated kernel

$$K_\Theta(z,w) = \frac{J - \Theta(z)J\Theta(w)^*}{1 - zw^*} \tag{14}$$

is positive definite. It has a unique meromorphic extension to the open unit disk, and this extension is such that the kernel K_Θ is still positive definite in the domain of analyticity of Θ in $\mathbb{D}$. Here too, besides the kernel K_Θ we have the kernel $K_{\widetilde{\Theta}}(z,w)$ and the kernel

$$D_\Theta(z,w) = \begin{pmatrix} K_\Theta(z,w) & \frac{J\Theta(z)-J\Theta(w^*)}{z-w^*} \\ \frac{\widetilde{\Theta}(z)J-\widetilde{\Theta}(w^*)J}{z-w^*} & K_{\widetilde{\Theta}}(z,w) \end{pmatrix}.$$

We note that the kernel K_Θ can be written as

$$K_\Theta(z,w) = \frac{I_p - \Theta(z)\Theta(w)^{[*]}}{1 - zw^*},$$

where $[*]$ denotes the adjoint in $\mathbb{C}_J$. This conforms with the way these kernels and the two other related kernels are written down in [6].

As we already mentioned, Krein and Langer developed in [44–48], the theory of operator-valued functions such that the corresponding kernels K_Θ (with a signature operator rather than a signature matrix) have a finite number of negative squares in some open subset of the open unit disk. Then, Θ has a unique meromorphic extension to the open unit disk, and this extension is such that K_Θ has the same number of negative squares in $\Omega(\Theta)$, the domain of analyticity Θ in $\mathbb{D}$. The three kernels have simultaneously the same number of negative squares and, as in the positive definite case, are respectively state spaces for coisometric, isometric, and unitary realizations of Θ.

In the special case $J = I$ (we return to the notation W rather than Θ for the function), Krein and Langer proved (see [44]) that W can be written as $W_0 B_0^{-1}$, where W_0 is analytic and contractive in the open unit disk and where B_0 is a finite matrix-valued Blaschke product. It follows that W has a finite number of poles in the open unit disk. In the rational case and when W takes unitary values on the unit circle, W is a quotient of two matrix-valued rational Blaschke product. Note however that when J has mixed inertia, W may have an infinite number of poles, even when $\kappa = 0$. For example, take

$$J = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad W(z) = \begin{pmatrix} 1 & 0 \\ 0 & b(z)^{-1} \end{pmatrix},$$

where b is a convergent Blaschke product with an infinite number of zeros. Such examples originate with the work of Potapov [59].

Definition 4.1. We denote by $\mathcal{S}_\kappa^{p\times p}(\mathbb{D})$ the family of $\mathbb{C}^{p\times p}$-valued functions W meromorphic in the open unit disk and such that the kernel K_W (defined by Eq. (13)) has κ negative squares in the domain of analyticity of W in $\mathbb{D}$.

Given a signature matrix J, we denote by $\mathcal{S}_\kappa^J(\mathbb{D})$ the family of $\mathbb{C}^{p\times p}$-valued functions Θ meromorphic in the open unit disk and such that the kernel K_Θ (defined by Eq. (14)) has κ negative squares in the domain of analyticity of Θ in $\mathbb{D}$.

We denote by $\mathscr{P}(W)$ and $\mathscr{P}(\Theta)$, respectively, the associated reproducing kernel Pontryagin spaces.

Since the kernels K_W and K_Θ are analytic in z and w^*, the elements of the associated reproducing kernel Pontryagin spaces are analytic in the domain of definition of W or Θ, respectively. See [6, Theorem 1.1.3, p. 7].

More generally, it is useful to consider non square generalized Schur functions. We consider $J_1 \in \mathbb{C}^{p_1\times p_1}$ and $J_2 \in \mathbb{C}^{p_1\times p_1}$ two signature matrices, of possibly different sizes, such that Eq. (5) is in form denoted by ν_-:

$$\nu_-(J_1) = \nu_-(J_2).$$

Reproducing kernel Pontryagin spaces with reproducing kernel of the form (4):

$$\frac{J_2 - \Theta(z) J_1 \Theta(w)^*}{1 - z w^*}$$

when Θ is $\mathbb{C}^{p_2 \times p_1}$-valued and analytic in a neighborhood of the origin, have been characterized in [6, Theorem 3.1.2, p. 85] (in fact, the result there is more general and considers operator-valued functions). In the statement below R_0 denotes the backward-shift operator

$$R_0 f(z) = \frac{f(z) - f(0)}{z}.$$

Theorem 4.2. *Let $(\mathscr{P}, [\cdot,\cdot]_{\mathscr{P}})$ be a reproducing kernel Pontryagin space of $\mathbb{C}^{p_2}$-valued functions. It has a reproducing kernel of the form (4) if and only if it is invariant under the backward-shift operator R_0 and*

$$[R_0 f, R_0 f]_{\mathscr{P}} \leq [f, f]_{\mathscr{P}} - f(0)^* J_2 f(0), \quad \forall f \in \mathscr{P}.$$

An example of such non square Θ appears in Sect. 6.2. See Eq. (37).

4.2 State Spaces and Realizations

We begin with recalling the following definition. Let W be an operator-valued function analytic in a neighborhood of the origin. A realization of W is an expression of the form

$$W(z) = D + zC(I - zA)^{-1}B, \tag{15}$$

where $D = W(0)$ and A, B, C are operators between appropriate spaces. It is an important problem to connect the properties of W and of the operator matrix

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}. \tag{16}$$

When the values of W are linear bounded operators between two Krein spaces, Azizov proved that a realization exists and that M can be chosen unitary. See [11] and [27] for further discussion and additional references. When W is a matrix-valued rational function without a pole at the origin, the spaces may be chosen finite dimensional, when no special structure is forced on the operator matrix M.

In Sect. 4.1, we have studied the correspondence between kernels and operator-valued Schur functions. Here we then pass to the realizations of Schur functions. The introduction of Schur functions offers many advantages, relevant to algorithms

and to computation. Case in point: In the next section, we give explicit formulas for realizations, i.e., for the computation of the four block operator entries A through D making up admissible realizations of a given Schur function and therefore of a kernel. As we show, there are several such choices, the coisometric realization (Theorem 4.3) and the unitary realization of de Branges and Rovnyak (Theorem 4.4), among others. There is in turn a rich literature on Schur algorithms in various special cases; see, for example, [2] for an overview. In preparation of Sect. 4.3 we need some definitions. Let $\mathscr{P}$ denote the space where A acts in Eq. (15). We say that the realization is closely inner connected if the span of the functions

$$(I - zA)^{-1}B\xi,$$

where ξ runs through $\mathbb{C}^p$ (recall that $J \in \mathbb{C}^{p\times p}$) and z runs through a neighborhood of the origin, is dense in $\mathscr{P}$. With the same choices of ξ and z, it will be called closely outer connected if the span of the functions

$$(I - zA^{[*]})^{-1}C^{[*]}\xi$$

is dense in $\mathscr{P}$, and connected if the span of the functions

$$(I - zA)^{-1}B\xi, \quad \text{and} \quad (I - wA^{[*]})^{-1}C^{[*]}\eta$$

is dense in $\mathscr{P}$ (η running through $\mathbb{C}^p$ and w through the same neighborhood of the origin as z). Here the adjoints are between Pontryagin spaces. We note that the terminology is different from that of classical system theory. In the finite-dimensional case, what is called here closely inner connected corresponds to observability, and what is called outer connected corresponds to controlability. The notion of being closely connected is specific to this domain and is, in general, different from minimality.

4.3 Coisometric and Unitary Realizations

Let $\Theta \in \mathscr{S}_\kappa^J$ be a generalized Schur function, assumed analytic in a neighborhood of the origin. In this section we review how the spaces $\mathscr{P}(\Theta)$ and $\mathscr{D}(\Theta)$ are the state spaces for coisometric and unitary realizations, respectively. For the following theorems, see [6, Theorem 2.2.1, p. 49] and [6, Theorem 2.1.3], respectively. In Theorems 4.3 and 4.4 the notions of coisometry and unitarity mean that M in Eq. (16) is an operator coisometric (resp. unitary) from the Pontryagin $\mathscr{P}(\Theta) \oplus \mathbb{C}_J$ into itself (resp. from $\mathscr{D}(\Theta) \oplus \mathbb{C}_J$ into itself).

Theorem 4.3. *Let $J \in \mathbb{C}^{p\times p}$ be a signature matrix and $\Theta \in \mathscr{S}_\kappa^J$ be analytic in a neighborhood of the origin. Then the formulas*

$$
\begin{aligned}
Af(z) &= \frac{f(z) - f(0)}{z}, \\
(B\xi)(z) &= \frac{\Theta(z) - \Theta(0)}{z}\xi, \\
Cf &= f(0), \\
D\xi &= \Theta(0)\xi,
\end{aligned}
$$

with $f \in \mathscr{P}(\Theta)$ and $\xi \in \mathbb{C}^p$, define a closely outer connected realization of Θ which is coisometric. This realization is unique up to a continuous and continuously invertible similarity operator.

This coisometric realization was introduced by de Branges and Rovnyak in [16] for scalar Schur functions and extended to the operator-valued case in [17]. We note that the coisometric realization is also known as the *backward-shift realization*; see, for example, [33].

L. de Branges and J. Rovnyak also formulated the unitary realization below.

Theorem 4.4. *Let $J \in \mathbb{C}^{p\times p}$ be a signature matrix and $\Theta \in \mathscr{S}_\kappa^J$ be analytic in a neighborhood of the origin. The formulas*

$$
\begin{aligned}
A\begin{pmatrix} f \\ g \end{pmatrix} &= \begin{pmatrix} \dfrac{f(z) - f(0)}{z} \\ zg(z) - \widetilde{\Theta}(z)Jf(0) \end{pmatrix}, \\
(B\xi)(z) &= \begin{pmatrix} \dfrac{\Theta(z) - \Theta(0)}{z}\xi \\ (J - \widetilde{\Theta}(z)J\widetilde{\Theta}(0)^*)\xi \end{pmatrix}, \\
C\begin{pmatrix} f \\ g \end{pmatrix} &= f(0), \\
D\xi &= \Theta(0)\xi,
\end{aligned}
$$

with $f \in \mathscr{D}(\Theta)$ and $\xi \in \mathbb{C}^p$, define a closely connected realization of Θ which is unitary. This realization is unique up to a continuous and continuously invertible similarity operator.

It is important to note that, in some cases, all three realizations are unitary. This is in particular the case when Θ is rational and J-unitary on the unit circle. See Sect. 4.4.

4.4 Finite-Dimensional de Branges Spaces

The finite-dimensional case is of special importance, and the case $J = I$ was considered in details in our previous work [9]. Then the three realizations are unitary, and it is easier to focus on the $\mathscr{P}(\Theta)$ spaces. As proved in [7, Corollary

p. 111] for the case $J = I$ and in [7, Theorem 5.5, p. 112] for the general case, given $\Theta \in \mathscr{S}_\kappa^J$, the associated space $\mathscr{P}(\Theta)$ is finite dimensional if and only if Θ is rational and J-unitary on the unit circle:

$$\Theta(e^{it})^* J \Theta(e^{it}) = J,$$

at all points e^{it} ($t \in [0, 2\pi]$) where it is defined. If moreover Θ is analytic in a neighborhood of the closed unit disk, we have

$$\mathscr{P}(\Theta) = \mathbf{H}_{2,J} \ominus \Theta \mathbf{H}_{2,J}.$$

Rationality is not enough to insure that $\mathscr{P}(\Theta)$ is finite dimensional, as illustrated by the case $J = 1$ and $\Theta = 0$. Then, $\mathscr{P}(\Theta) = \mathbf{H}_2(\mathbb{D})$.

Definition 4.5. We will denote by $\mathscr{U}_\kappa^J$ the multiplicative group of rational $\mathbb{C}^{p\times p}$-valued functions Θ which take J-unitary values on the unit circle and for which the corresponding kernel K_Θ has κ negative squares. We set

$$\mathscr{U}^J = \bigcup_{\kappa=0}^{\infty} \mathscr{U}_\kappa^J.$$

The results and realizations presented in the previous section take now an easier form. The various operators can be seen as matrices. Unitarity above is with respect to the indefinite metric of $\mathscr{P}(\Theta) \oplus \mathbb{C}_J$, and we can rephrase Theorem 4.4 as follows:

Theorem 4.6. *Let W be a rational $\mathbb{C}^{p\times p}$-valued function analytic at the origin, and let*

$$W(z) = D + zC(I - zA)^{-1}B$$

be a minimal realization of W. Then, W is J-unitary on the unit circle if and only if there exists an invertible Hermitian matrix H (which is uniquely determined from the given realization) such that

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^* \begin{pmatrix} H & 0 \\ 0 & J \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} H & 0 \\ 0 & J \end{pmatrix}. \tag{17}$$

The change of variable $z \mapsto 1/z$ yields

Theorem 4.7. *Let W be analytic at infinity, and let*

$$W(z) = D + C(zI - A)^{-1}B$$

be a minimal realization of W. Then, W is J-unitary on the unit circle if and only if there exists an invertible Hermitian matrix H (which is uniquely determined from the given realization) and such that Eq. (17) holds.

The matrix H is called the *associated Hermitian matrix* (to the given minimal realization). This result was proved in [8, Theorem 3.10] for the case where A is non-singular. For the approach using reproducing kernel Hilbert spaces, see [4,6,7].

5 Cuntz Relations

5.1 Cuntz Relations and the de Branges–Rovnyak Spaces

The results of this section are related to [10, 22]. In that last paper, the functions $1, \ldots, z^{N-1}$ below are replaced by the span of a finite-dimensional backward-shift invariant subspace, but the discussion is restricted to the Hilbert space case and scalar-valued functions.

Normally by Cuntz relations we refer to a finite system of isometries $S_1, \ldots, S_N$ in a Hilbert space $\mathscr{H}$ satisfying two conditions:

(a) Different isometries in the system must have orthogonal ranges

$$S_j^* S_k = 0, \quad j \neq k, \tag{18}$$

(b) The sum of the ranges equals $\mathscr{H}$:

$$\sum_{j=1}^{N} S_j S_j^* = I_{\mathscr{H}}. \tag{19}$$

Note that (a) already forces $\mathscr{H}$ to be infinite dimensional. Indeed, if $\mathscr{H}$ is finite dimensional, an isometry is unitary and the orthogonality of the ranges is not possible; see the discussion below and Sect. 5.3. If we allow the isometries to operate between two finite-dimensional spaces of different dimensions, then one can find isometries which satisfy the Cuntz relations. It is the set of three conditions: Each S_i is isometric in a Hilbert space $\mathscr{H}$, and (a) and (b) together imply that every realization is a representation of a simple, purely infinite C^*-algebra, called O_N. In applications to filters, the N individual subspaces represent frequency bands. This allows for versatile computational algorithms tailored to multiscale problems such as wavelet decompositions and analysis on fractals. In our present paper, we relax some of the original very restrictive axioms, while maintaining the computational favorable properties. Our more general framework still allows for algorithms based on iteration of the operator family $S_1, \ldots, S_N$ in a particular representation.

If one allows isometries between two Hilbert spaces, then the finite-dimensional case may occur, as illustrated by the following example:

$$\mathscr{H}_1 = \mathbb{C}, \quad \mathscr{H}_2 = \mathbb{C}^2,$$

and

$$S_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad S_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

We have

$$S_1^* S_1 = S_2^* S_2 = 1, \quad S_1^* S_2 = S_2^* S_1 = 0,$$

and

$$S_1 S_1^* + S_2 S_2^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We go beyond the setting of Hilbert space and relax the conditions (a) and (b) imposed in the original framework from C^*-algebra theory, allowing here isometric operators between two Pontryagin spaces. We still preserve the features of the representations of use in iterative algorithms.

It is not surprising that in Sect. 5.3 we have finite-dimensional spaces. Now for the generalized theory, we must allow for de Branges and Rovnyak spaces and for negative squares and signature matrix. The resulting modifications in the form of the Cuntz relations, in the case of Hilbert space, entail some nontrivial modifications addressed in the next two sections. Our main results for this are proved in the present section and in Sect. 5.3 for the finite-dimensional case.

The main result of this section is that one can associate in a natural way to an element $\Theta \in \mathscr{S}_k^J(\mathbb{D})$ a family of operators which satisfy the Cuntz relations. We begin with a preliminary result, which is a corollary of Theorem 3.7 with $\varphi(z) = z^N$.

Proposition 5.1. *Let $\Theta \in \mathscr{S}_\kappa^J(\mathbb{D})$, and let $\mathscr{P}(\Theta)$ be the associated Pontryagin space, with reproducing kernel*

$$K_\Theta(z, w) = \frac{J - \Theta(z) J \Theta(w)^*}{1 - z w^*}.$$

The function

$$K_\Theta(z^N, w^N) = \frac{J - \Theta(z^N) J \Theta(w^N)^*}{1 - z^N w^{*N}}$$

has also κ negative squares in its domain of definition in $\mathbb{D}$. The associated reproducing kernel Pontryagin space $\mathscr{M}_N$ is equal to the space of functions of the form $F(z) = f(z^N)$, where $f \in \mathscr{P}(\Theta)$, with the following indefinite inner product

$$[F, G]_{\mathscr{M}_N} = [f, g]_{\mathscr{P}(\Theta)}, \tag{20}$$

where $g \in \mathscr{P}(\Theta)$ and $G(z) = g(z^N)$.

We have:

Theorem 5.2. *Let $\Theta \in \mathscr{S}_\kappa^J(\mathbb{D})$, and let $\mathscr{P}(\Theta)$ be the associated Pontryagin space with reproducing kernel*

$$K_\Theta(z,w) = \frac{J - \Theta(z)J\Theta(w)^*}{1 - zw^*}.$$

Then, for $N \in \mathbb{N}$, the function Θ_N defined by $\Theta_N(z) = \Theta(z^N)$ belongs to $\mathcal{S}^J_{N\kappa}$. Furthermore, $\mathcal{P}(\Theta_N)$ consists of all the functions of the form

$$f(z) = \sum_{j=0}^{N-1} z^j f_j(z^N), \quad f_j \in \mathcal{P}(\Theta).$$

Any such representation is unique, and the inner product in $\mathcal{P}(\Theta_N)$ is given by

$$[f,g]_{\mathcal{P}(\Theta_N)} = \sum_{j=0}^{N-1} [f_j, g_j]_{\mathcal{P}(\Theta)},$$

where $g(z) = \sum_{j=0}^{N-1} z^j g_j(z^N)$ for some $g_0, \ldots, g_{N-1} \in \mathcal{P}(\Theta)$.

Proof. We proceed in a number of steps.

Step 1. It holds that $\nu_-(\Theta_N) \le N \cdot \kappa$.

Indeed,

$$\begin{aligned}\frac{J - \Theta(z^N)J\Theta(w^N)^*}{1 - zw^*} &= \frac{J - \Theta(z^N)J\Theta(w^N)^*}{1 - z^N w^{N*}} \cdot \frac{1 - z^N w^{N*}}{1 - zw^*} \\ &= \frac{J - \Theta(z^N)J\Theta(w^N)^*}{1 - z^N w^{N*}} \cdot \left(\sum_{k=0}^{N-1} z^k w^{*k} \right).\end{aligned}$$

This expresses the kernel K_{Θ_N} as the sum of N kernels, each with κ negative squares. Thus, $\nu_-(\Theta_N) \le N\kappa$. To show that there is equality, we need to show that the associated spaces have pairwise intersections which all reduce to the zero function.

Step 2. Let $k, \ell \in \{0, \ldots, N-1\}$, such that $k \ne \ell$. Then, with $\mathcal{M}_N$ as in the previous theorem:

$$z^k \mathcal{M}_N \cap z^\ell \mathcal{M}_N = \{0\}.$$

Indeed, assume that $k > \ell$ and let $f, g \in \mathcal{M}_N$ be such that

$$z^k f(z^N) = z^\ell g(z^N).$$

Then, f and g will simultaneously be identically equal to $0_{p\times 1}$. Assume $f \not\equiv 0_{p\times 1}$. One of its components, say the first, with $f = \begin{pmatrix} x_1(z) \cdots x_p(z) \end{pmatrix}^t$ is not identically equal to zero (p is the size of the signature matrix J). Then we obtain

$$z^{k-\ell} = \frac{y_1(z^N)}{x_1(z^N)},$$

where y_1 denotes the first component of g. Since f and g are meromorphic in $\mathbb{D}$, the function y_1/x_1 has a Laurent expansion at the origin. Moreover the Laurent expansion of $\frac{y_1}{x_1}(z^N)$ contains only powers which are multiple of N. By the uniqueness of the Laurent expansion, this contradicts the fact that it is equal to $z^{k-\ell}$, with $|k-\ell| < N$.

Step 3. It holds that

$$\mathscr{P}(\Theta_N) = \oplus_{j=0}^{N-1} z^j \mathscr{M}_N,$$

and it holds that $\nu_{\Theta_N} = N\kappa$.

This is because the spaces $z^j \mathscr{M}_N$ have pairwise intersections which reduce to the zero functions in view of Step 2.

□

Theorem 5.3. *In the notation above, set*

$$(S_j f)(z) = z^j f(z^N) \quad \mathscr{P}(\Theta) \longrightarrow \mathscr{P}(\Theta_N).$$

Then,

$$S_j^{[*]} f = f_j \quad \mathscr{P}(\Theta_N) \longrightarrow \mathscr{P}(\Theta), \tag{21}$$

and

$$\begin{aligned} S_j^{[*]} S_k &= \delta_{j,k} I_{\mathscr{P}(\Theta)} \\ \sum_{j=0}^{N-1} S_j S_j^{[*]} &= I_{\mathscr{P}(\Theta_N)}, \end{aligned} \tag{22}$$

where the $[*]$ *denotes adjoint between Pontryagin spaces.*

Proof. We proceed in a number of steps.

Step 1. The operators S_j are continuous.

The operators S_j are between Pontryagin spaces of different indices, and some care is required to check continuity. To this end, fix $j \in \{0, \ldots, N-1\}$ and note that S_j is everywhere defined. Furthermore we claim that it is a closed operator. Indeed, let $f_1, f_2 \ldots$ be a sequence of elements in $\mathscr{P}(\Theta)$ converging strongly to $f \in \mathscr{P}(\Theta)$ and such that the sequence $S_j f_1, S_j f_2, \ldots$ converges strongly to $g \in \mathscr{P}(\Theta_N)$. Strong convergence in a Pontryagin space implies weak convergence, and in a reproducing kernel Pontryagin space, weak convergence implies pointwise convergence. Therefore, for every w where Θ is defined,

$$\lim_{k\to\infty} f_k(w) = f(w),$$

and

$$\lim_{k\to\infty}(S_j f_k)(w) = g(w).$$

Since $(S_j f_k)(w) = w^j f_k(w)$, and thus $g(w) = w^j f(w)$. Therefore $g = S_j f$, and the operator S_j is closed and hence continuous.

Step 2. Equation (21) is in force.

Let $g(z) = \sum_{k=0}^{N-1} z^k g_k(z^N) \in \mathscr{P}(\Theta_N)$ where the $g_k \in \mathscr{P}(\Theta)$, and let $u \in \mathscr{P}(\Theta)$. Then,

$$\begin{aligned}\mathscr{P}(\Theta_N) &= \left[z^j u(z^N)\,,\ \sum_{k=0}^{N-1} z^k g_k(z^N) \right]_{\mathscr{P}(\Theta_N)} \\ &= [u\,,\ g_j]_{\mathscr{P}(\Theta)} \\ &= [u\,,\ S_j^{[*]} g]_{\mathscr{P}(\Theta)},\end{aligned}$$

where $[\,,\,]_{\mathscr{P}(\Theta)}$ and $[\,,\,]_{\mathscr{P}(\Theta_N)}$ denote the indefinite inner products in the corresponding spaces. Hence, we have $S_j^{[*]} g = g_j$.

Step 3. The Cuntz relations hold.

From Eq. (21) we have for $u \in \mathscr{P}(\Theta)$

$$S_j^{[*]} S_k u = S_j^{[*]}(z^k u(z^N)) = \begin{cases} 0 & \text{if} \quad j \neq k, \\ u & \text{if} \quad j = k. \end{cases}$$

Furthermore, for $f(z) = \sum_{j=0}^{N-1} z^j f_j(z^N) \in \mathscr{P}(\Theta_N)$ (where the $f_j \in \mathscr{P}(\Theta)$), we have

$$S_k S_k^{[*]} f = S_k(f_k) = z^k f_k(z^N),$$

and thus

$$\sum_{k=0}^{N-1} S_k S_k^{[*]} = I_{\mathscr{P}(\Theta_N)}.$$

□

We note that, with

$$S = \begin{pmatrix} S_0 & S_1 & \cdots & S_{N-1} \end{pmatrix} \quad \mathscr{P}(\Theta)^N \longrightarrow \mathscr{P}(\Theta_N),$$

the Cuntz relations (21) can be rewritten as

$$S S^{[*]} = I_{\mathscr{P}(\Theta_N)} \quad \text{and} \quad S^{[*]} S = I_{\mathscr{P}(\Theta)^N}.$$

At this stage, let us introduce some more notation. We set

$$\Theta_{N^k}(z) = \Theta(z^{N^k}),$$

and $S_i^{(0)} = S_i$ for $i = 0, \ldots, N-1$. We can reiterate the preceding analysis with Θ_N instead of Θ. We then obtain N isometries $S_0^{(1)}, \ldots, S_{N-1}^{(1)}$ from $\mathscr{P}(\Theta_N)$ into $\mathscr{P}(\Theta_{N^2})$ satisfying the Cuntz relations. Iterating k times, one obtains k sets of isometries

$$S_0^{(j-1)}, \ldots, S_{N-1}^{(j-1)}, \quad j = 1, \ldots, k,$$

from $\mathscr{P}(\Theta_{N^{j-1}})$ into $\mathscr{P}(\Theta_{N^j})$, which also satisfy the Cuntz relations. This gives us N^k isometries

$$S_{i_1}^{(0)} S_{i_2}^{(1)} \cdots S_{i_k}^{(k-1)},$$

with $(i_1, i_2, \ldots, i_k) \in \{0, \ldots, N-1\}^k$, from $\mathscr{P}(\Theta)$ into $\mathscr{P}(\Theta_{N^k})$, all satisfying the Cuntz relations.

5.2 Cuntz Relation: The General Case

We now wish to extend the results of Sect. 5.1 and in particular Theorem 5.2 to the case where the N functions $1, z, \ldots, z^{N-1}$ are replaced by prescribed functions $m_0(z), m_1(z), \ldots, m_{N-1}(z)$, whose finite-dimensional linear span we denote by $\mathscr{L}$, and the kernel $K_\Theta(z, w)$ is replaced by a given analytic $\mathbb{C}^{N\times N}$-valued kernel $K(z, w)$ and the kernel $K_{\Theta_N}(z, w)$ is replaced by a kernel $\widetilde{K}(z, w)$. Let as in Sect. 5.1, $K_N(z, w) = K(z^N, w^N)$. We address the following problem: Given K and $\widetilde{K}$ two Hermitian kernels defined on a set Ω and with a finite number of negative squares there, when can one find decompositions of the form

$$f(z) = \sum_{n=0}^{N-1} m_n(z) g_n(z^N), \tag{23}$$

where the functions $g_0, \ldots, g_{N-1}$ belong to $\mathscr{P}(K)$ for some, or all, elements in $\mathscr{P}(\widetilde{K})$. We have:

Theorem 5.4. *Let $K(z, w)$ and $\widetilde{K}(z, w)$ be two kernels defined on a set Ω, and assume that*

$$\nu_-(\widetilde{K}) = N\nu_-(K). \tag{24}$$

Let $m_0, \ldots, m_{N-1}$ be N functions on Ω. Assume that the kernel

$$\widetilde{K}(z, w) - \left(\sum_{n=0}^{N-1} m_n(z) m_n(w)^* \right) K(z, w)$$

is positive definite in Ω. Then, with $\varphi(z) = z^N$, the choice $g_n = T_{m_n,\varphi}^{[*]} f_n$, $n = 0, \ldots, N-1$ solves Eq. (23).

Proof. We use Theorem 3.6 with $K_2(z, w) = \widetilde{K}(z, w)$ and

$$K_1(z,w) = \begin{pmatrix} K(z,w) & 0 & 0 & \cdots & 0 \\ 0 & K(z,w) & 0 & \cdots & 0 \\ & & & & \\ 0 & 0 & \cdots & 0 & K(z,w) \end{pmatrix}.$$

Then

$$\left(\sum_{n=0}^{N-1} m_n(z) m_n(w)^*\right) K(z,w) = m(z) K_1(z) m(w)^*,$$

and Theorem 3.6 with K_1 and K_2 as above and

$$m(z) = \left(m_0(z)\ m_1(z) \cdots m_{N-1}(z)\right), \quad \text{and} \quad \varphi(z) = z^N,$$

leads to the fact that the map

$$f \mapsto m(z) f(z^N)$$

is a contraction from $(\mathscr{P}(K))^N$ into $\mathscr{P}(\widetilde{K})$. □

In applications, one uses the kernel $\widetilde{K}(z,w) = K_N(z,w)$ in the above result.

Proposition 5.5. *A sufficient condition for Eq. (24) to hold is that*

$$m_j \mathscr{P}(K) \cap m_k \mathscr{P}(K) = \{0\}, \tag{25}$$

for all $j,k \in \{0,\ldots,N-1\}$ *such that* $j \neq k$.

Proof. Indeed, when this condition is in force, we have that the Pontryagin space with reproducing kernel $m(z)K_1(z,w)m(w)^*$ is the direct sum of the Pontryagin spaces with reproducing kernels $m_j(z)K(z,w)m_j(w)^*$, $j = 0,\ldots,N-1$. □

We note that there is similarity between Eq. (23) and the solution of Gleason's problem. Gleason's problem is the following: Given a linear space of functions $\mathscr{M}$ of functions analytic in a set $\Omega \subset \mathbb{C}^N$ and given $a \in \Omega$, Gleason's problem is the following: when can we find functions $g_1,\ldots,g_N \in \mathscr{M}$ (which depend on a) such that

$$f(z) - f(a) = \sum_{n=1}^{N} (z_n - a_n) g_n(z,a).$$

5.3 Cuntz Relations: Realizations in the Rational Case

Recall that for a given generalized Schur function Θ, we presented in Theorems 4.3 and 4.4 coisometric and unitary realizations, respectively. The unitary realization turns to be more involved than the coisometric backward-shift realization. In some

cases, these two realizations are unitarily equivalent, in particular, when Θ is rational and J-unitary on the unit circle. As we already discussed in Sect. 4.4, this is equivalent to having the space $\mathscr{P}(\Theta)$ finite dimensional. In this section we adopt this simplifying assumption and study the realization of $\Theta_N(z) = \Theta(z^N)$ in terms of the realization of Θ.

We take the signature matrix J to belong to $\mathbb{C}^{L\times L}$. We know (see [6, 7] and Theorem 4.3) that

$$\Theta(z^N) = \mathscr{D} + z\mathscr{C}(I - z\mathscr{A})^{-1}\mathscr{B},$$

where $\mathscr{D} = \Theta_N(0) = \Theta(0)$ and where $\mathscr{A}, \mathscr{B}$, and $\mathscr{C}$ are defined as follows: $\mathscr{C}$ is the evaluation at the origin:

$$\mathscr{C} f = f(0).$$

$\mathscr{B}$ is defined by

$$\mathscr{B}\xi = \frac{\Theta_N(z) - \Theta_N(0)}{z}\xi, \quad \xi \in \mathbb{C}^L,$$

and $\mathscr{A}$ is the backward shift in $\mathscr{P}(\Theta_N)$. The matrix (see [7])

$$\begin{pmatrix} \mathscr{A} & \mathscr{B} \\ \mathscr{C} & \mathscr{D} \end{pmatrix}$$

is unitary in the $\mathscr{P}(\Theta_N)$ metric. We know from Theorem 5.2 that $\mathscr{P}(\Theta_N)$ is equal to the space of functions of the form

$$f(z) = \sum_{k=0}^{N-1} z^k f_k(z^N), \tag{26}$$

where the $f_k \in \mathscr{P}(\Theta)$ are uniquely defined. We will denote by U the map

$$f \hookrightarrow \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{N-1} \end{pmatrix}$$

from $\mathscr{P}(\Theta_N)$ onto $(\mathscr{P}(\Theta))^N$. In view of Eq. (22), U is a unitary map (between Pontryagin spaces).

Let T denote the following map from $(\mathscr{P}(\Theta))^N$ into itself defined by

$$TUf = \begin{pmatrix} 0 & I & 0 & \cdots & 0 \\ 0 & 0 & I & \cdots & 0 \\ & & & & \\ & & & & I \\ R_0 & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ \vdots \\ f_{N-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ R_0 f_0 \end{pmatrix}.$$

Proposition 5.6. *Let $f \in \mathscr{P}(\Theta_N)$, with representation (26). It holds that*

$$U\mathscr{A}f = (TUf)(z^N), \tag{27}$$

and it holds that

$$\langle \mathscr{A}f, \mathscr{A}g\rangle_{\mathscr{P}(\Theta_N)} = \langle TUf, TUg\rangle_{(\mathscr{P}(\Theta))^N}. \tag{28}$$

Proof. Indeed, with f of the form (26), we have

$$\mathscr{A}f(z) = R_0 f(z) = \frac{f(z) - f(0)}{z} = \sum_{k=1}^{N-1} z^{k-1} f_k(z^N) + z^{N-1}\frac{f_0(z^N) - f_0(0)}{z},$$

so that $U\mathscr{A}U^* f$ is equal to

$$\begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{N-1} \end{pmatrix} \mapsto \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ R_0 f_0 \end{pmatrix},$$

i.e., Eq. (27) is in force. Finally Eq. (28) follows from the formula for the inner product in $\mathscr{P}(\Theta_N)$. □

Proposition 5.7. *Let $f \in \mathscr{P}(\Theta_N)$ with representation (26). Then,*

$$\mathscr{C}f = C\begin{pmatrix} I_L & 0 \cdots 0 \end{pmatrix} Uf, \tag{29}$$

where C is the evaluation at the origin in $\mathscr{P}(\Theta)$.

Proof. This is clear from

$$\mathscr{C}f = f_0(0) = \begin{pmatrix} C & 0 & 0 \cdots 0 \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ \vdots \\ f_{N-1} \end{pmatrix}.$$

□

Proposition 5.8. *We have*

$$\mathscr{B}\xi = z^{N-1}(B\xi)(z^N),$$

where B is the operator from $\mathbb{C}^L$ into $\mathscr{P}(\Theta)$:

$$B\xi = R_0\Theta\xi,$$

and we have

$$\langle \mathscr{B}\xi, \mathscr{B}\eta\rangle_{\mathscr{P}(\Theta_N)} = \langle B\xi, B\eta\rangle_{\mathscr{P}(\Theta)}, \quad \eta, \xi \in \mathbb{C}^L. \tag{30}$$

Proof. We have

$$\mathscr{B}\xi(z) = R_0\Theta_N\xi(z) = \frac{\Theta(z^N) - \Theta(0)}{z} = z^{N-1}(B\xi)(z^N).$$

Equality (30) follows form the definition of the inner product in $\mathscr{P}(\Theta_N)$. □

These various formulas allow to show directly that the realization is indeed unitary and to compute the associated Hermitian matrix in the finite-dimensional case.

6 Decompositions

6.1 Generalized Down-Sampling and an Hermitian Form

In the preceding section we considered decompositions of a function in the form (23). Here we consider different kind of decompositions. We consider matrices $P \in \mathbb{C}^{N\times N}$ satisfying

$$\det(I_N - \epsilon_N^\ell P^\ell) \neq 0, \quad \ell = 1, \dots, N-1, \quad \text{and} \quad P^N = I_N. \tag{31}$$

We do not assume that $P^{N-1} \neq I_N$, and in particular the choice $P = I_N$ is allowed. The special case $P = \epsilon_N P_N$ plays also an important role.

Theorem 6.1. *Let W be a $\mathbb{C}^{N\times M}$-valued function defined in the open unit disk (typically, $M = 1$ or $M = N$). Let $P \in \mathbb{C}^{N\times N}$ satisfying Eq. (31), and let, for $k = 0, \dots, N-1$,*

$$W_k(z) = \frac{1}{N}\sum_{\ell=0}^{N-1} (\epsilon_N P)^{k\ell}\, W\left(\epsilon_N^\ell z\right). \tag{32}$$

Then,

$$W_k(\epsilon_N z) = (\epsilon_N P)^{-k}(W_k(z)), \quad k = 0, \dots, N-1, \tag{33}$$

$$W(z) = \sum_{k=0}^{N-1} W_k(z). \tag{34}$$

Proof. We have

$$
\begin{aligned}
W_k(\epsilon_N z) &= \frac{1}{N}\sum_{\ell=0}^{N-1}(\epsilon_N P)^{k\ell} W(\epsilon_N^{\ell}\epsilon_N z) \\
&= (\epsilon_N P)^{-k}\left(\frac{1}{N}\sum_{\ell=0}^{N-1}(\epsilon_N P)^{k(\ell+1)} W(\epsilon_N^{\ell+1} z)\right) \\
&= (\epsilon_N P)^{-k}(W_k(z)),
\end{aligned}
$$

since $(\epsilon_N P)^{kN} = I_N$, and this proves Eq. (33). To prove Eq. (34) we write

$$
\begin{aligned}
\sum_{k=0}^{N-1} W_k(z) &= \sum_{k=0}^{N-1}\left(\frac{1}{N}\sum_{\ell=0}^{N-1}(\epsilon_N P)^{k\ell} W(\epsilon_N^{\ell} z)\right) \\
&= \frac{1}{N}\left(\sum_{\ell=0}^{N-1}\left(\sum_{k=0}^{N-1}(\epsilon_N P)^{k\ell}\right) W(\epsilon_N^{\ell} z)\right) \\
&= W(z),
\end{aligned}
$$

since, in view of Eq. (31),

$$
\sum_{k=0}^{N-1}(\epsilon_N P)^{k\ell} = \begin{cases} N, & \text{if } \ell = 0, \\ (I_N - (\epsilon_N P)^{N\ell})(I - (\epsilon_N P)^{\ell})^{-1} = 0 & \text{if } \ell = 1, 2, \ldots N-1. \end{cases}
$$

□

When $P = I_N$, the index $k = 1$ corresponds to the down-sampling operator.

6.2 *Orthogonal Decompositions in Krein Spaces*

In some cases the decomposition (34) is orthogonal for the underlying Krein space (or Pontryagin space) structure. We will assume that the Krein space $(\mathscr{K}, [\cdot,\cdot]_{\mathscr{K}})$ consists of $\mathbb{C}^N$-valued functions and satisfies the following property:

Hypothesis 6.2. *Let P be a matrix satisfying Eq. (31), and let $\varphi(z) = \epsilon_N z$. We assume that:*

1. *The composition operator $f \mapsto f(\varphi)$ is continuous and unitary from $\mathscr{K}$ into itself.*
2. *The operator of multiplication by P on the left is continuous and unitary from $\mathscr{K}$ into itself.*

We note that, in particular, the operator $T_{P,\varphi}$ defined by Eq. (8),

$$T_{P,\varphi} f(z) = P f(\epsilon_N z),$$

is continuous and unitary from $\mathscr{K}$ into itself. Note also that

$$T_{P,\varphi}^N = I_{\mathscr{K}}.$$

Hypothesis 6.2 holds in particular for the spaces $\mathbf{H}_{2,J}$ when P is J-unitary, i.e., satisfies

$$P^* J P = J.$$

Theorem 6.3. *Let $(\mathscr{K}, [\cdot\,,\cdot])$ be a Krein space of $\mathbb{C}^N$-valued functions, satisfying Hypothesis 6.2. Let $W \in \mathscr{K}$ and let*

$$W_k(z) = \frac{1}{N} \sum_{\ell=0}^{N-1} (\epsilon_N P)^{k\ell} W(\epsilon_N^{\ell} z). \tag{35}$$

Then,

$$[W_\ell, W_k] = 0, \quad \ell \neq k,$$
$$W(z) = W_0(z) + \cdots + W_{N-1}(z),$$

and

$$W_k(\epsilon_N z) = (\epsilon_N P)^{-k} W(z).$$

Proof. The last two claims are proved in Theorem 6.1. The first claim takes into account the hypothesis on $\mathscr{K}$ and is proved as follows: We take k_1 and k_2 in $\{0, \ldots, N-1\}$ and assume that $k_2 < k_1$. Taking into account the definition of W_k, we see that the inner product $[W_{k_1}, W_{k_2}]_{\mathscr{K}}$ is a sum of N^2 inner products, namely

$$[(\epsilon_N P)^{k_1 \ell_1} W(\epsilon_N^{\ell_1} z), (\epsilon_N P)^{k_2 \ell_2} W(\epsilon_N^{\ell_2} z)]_{\mathscr{K}}, \quad \ell_1, \ell_2 \in \{0, \ldots, N-1\}.$$

These N^2 inner products can be rearranged as N sums of inner product, each sum being equal to 0. Indeed, consider first the inner products corresponding to $\ell_1 = \ell_2$. In view of the unitary of the operator $T_{P,\varphi}$, we have

$$\sum_{\ell_1=0}^{N-1} \left[(\epsilon_N P)^{k_1 \ell_1} W(\epsilon_N^{\ell_1} z), (\epsilon_N P)^{k_2 \ell_1} W(\epsilon_N^{\ell_1} z) \right]_{\mathscr{K}} = \left[\left(\sum_{\ell_1=0}^{N-1} (\epsilon_N^{k_1 - k_2} P)^{\ell_1} \right) W, W \right]_{\mathscr{K}}$$
$$= 0.$$

Indeed, using $0 < k_1 - k_2 \leq N - 1$ and so, by hypothesis on P, we have

$$\det(I_N - (\epsilon_N P)^{k_1 - k_2}) \neq 1,$$

and the sum

$$\sum_{\ell_1=0}^{N-1}((\epsilon_N P)^{k_1-k_2})^{\ell_1} = 0.$$

Let us now regroup the factors of $[W(z), W(\epsilon_N z)]_{\mathscr{K}}$. Taking into account that

$$[P^{k_1(N-1)}W(\epsilon_N^{N-1}z), W(z)]_{\mathscr{K}} = [P^{k_1(N-1)}W(z), W(\epsilon_N z)]_{\mathscr{K}},$$

we have

$$\begin{aligned}
&\sum_{\ell=0}^{N-2}\left[(\epsilon_N P)^{k_1\ell}W(\epsilon_N^{\ell}z), (\epsilon_N P)^{k_2(\ell+1)}W(\epsilon_N^{\ell+1}z)\right]_{\mathscr{K}} \\
&\quad + \left[(\epsilon_N P)^{k_1(N-1)}\, W\left(\epsilon_N^{N-1}z\right), W(z)\right]_{\mathscr{K}} \\
&= \left[\left(\sum_{\ell=0}^{N-2}(\epsilon_N P)^{\ell k_1-(\ell+1)k_2} + (\epsilon_N P)^{k_1(N-1)}\right) W(z), W(\epsilon_N z)\right]_{\mathscr{K}} \\
&= \left[(\epsilon_N P)^{-k_2}\left(\sum_{\ell=0}^{N-1}((\epsilon_N P)^{k_1-k_2})^{\ell}\right) W(z), W(\epsilon_N z)\right]_{\mathscr{K}} \\
&= 0.
\end{aligned}$$

The remaining terms are summed up to 0 in the same way. □

6.3 *Decompositions in Reproducing Kernel Spaces*

We begin with a result in the setting of Schur functions, as opposed to generalized Schur functions.

Theorem 6.4. *Let W be a $\mathbb{C}^{p\times q}$-valued Schur function and let $\varphi(z) = \epsilon_N z$. Then the operator of composition by φ is a contraction from $\mathscr{H}(W)$ into itself if and only if there exists a $\mathbb{C}^{q\times q}$-valued Schur function $X(z)$ such that*

$$W(z) = W(\epsilon_N z)X(z). \tag{36}$$

Proof. By Theorem 3.6, the map T_φ is a contraction if and only if the kernel

$$K_W(z,w) - K_W(\epsilon_N z, \epsilon_N w) = \frac{W(\epsilon_N z)W(\epsilon_N w)^* - W(z)W(w)^*}{1 - zw^*}$$

is positive definite in the open unit disk. By Leech's factorization theorem (see [61, p. 107]), the above kernel is positive definite if and only if there is a Schur function $X(z)$ such that Eq. (36) is in force. □

As an example, take any Schur function s and build

$$W(z) = \frac{1}{\sqrt{N}} \left(s(z)\ s(\epsilon_N z) \cdots s(\epsilon_N^{N-1} z) \right). \tag{37}$$

Then

$$W(z) = W(\epsilon_N z) P_N,$$

where P_N is defined by Eq. (2).

7 The Family $\mathscr{C}_N$

An effective approach to generating wavelet bases is the use of MRA; see, for example, [13, 18, 26]. Traditionally one looks for a finite family of functions in $\mathbf{L}_2(\mathbb{R}, dx)$ or $\mathbf{L}_2(\mathbb{R}^d, dx)$ for some dimension d. If $d = 1$, one chooses a scale number, say N. If $d > 1$, instead one scales with a $d \times d$ matrix A over the integers. We assume that A is expansive, i.e., with eigenvalues bigger than 1 in modulus. If A is given, let N be the absolute value of its determinant. To create MRA wavelets we need an initial finite family $\mathscr{F}$ of N functions in $\mathbf{L}_2(\mathbb{R})$ or $\mathbf{L}_2(\mathbb{R}^d)$. One of the functions is called the scaling function (ϕ in the discussion below). For the moment, we will set $d=1$, but the outline below easily generalizes to $d > 1$. An MRA wavelet basis is a basis for $\mathbf{L}_2(\mathbb{R})$ or $\mathbf{L}_2(\mathbb{R}^d)$ which is generated from the initial family $\mathscr{F}$ and two operations: one operation is scaling by the number N (or the matrix A if $d>1$) and the other is action by integer translates of functions. The special property for the finite family of functions $\mathscr{F}$ is that if the N-scaling is applied for each function ψ in F the result is in the closed span of the integer translates of the scaling function ϕ. The corresponding coefficients are called masking coefficients. The reason for this is that the scaled functions represent refinements, and they are computed from masking points in a refinement. The role of the functions $m_0, m_1, \ldots, m_{N-1}$ are the frequency response functions corresponding to the system of masking coefficients. From these functions we then build a matrix-valued function $W(z)$ as in Eq. (38). The question we address here is the characterization of the matrix-valued function which arises this way. Now the wavelet filters we consider here go beyond those studied earlier in that we allow for wider families of multiresolution analyses (MRAs). This includes more general wavelet families, allowing, for example, for wavelet frame bases, see, e.g., [13,41,42], multi-scale systems in dynamics, and analysis of fractals; see [31].

7.1 The Family $\mathscr{C}_N$: Characterization

The filters we consider are matrix-valued (or operator-valued) functions of a complex variable. In general if a positive integer N is given, and if a matrix function $W(z)$ is designed to take values in $\mathbb{C}^{N\times N}$, then of course, there are N^2 scalar-valued functions occurring as matrix entries. However, in the case of filters arising in applications involving N distinct frequency bands, for example, in wavelet constructions with scale number N, then we can take advantage of an additional symmetry for the given matrix function $W(z)$; see, for example, Eq. (1) in the Introduction. Here we point out that this N-symmetry condition (or N-periodicity) means that $W(z)$ is then in fact determined by only N scalar-valued functions; see Eq. (38). These functions play three distinct roles as follows: they are (1) the scalar-valued filter functions, $\hat{s}_i$, for $i = 0, 1, \ldots, N-1$, in generalized quadrature-mirror filter systems (the quadrature case corresponds to $N = 2$); (2) scaling filters for scale number N with each of the N scalar functions $\hat{s}_i$ generating an element in a wavelet system of functions on the real line and corresponding to scale number N; and (3) the system of scalar functions $(\hat{s}_i)_{i=0,\ldots N-1}$ generates an operator family $(S_i)_{i=0,\ldots,N-1}$ constituting a representation of the Cuntz relations, thus generalizing Theorem 5.3. The results presented in this section are related to [9].

Recall that $\epsilon_N = \mathrm{e}^{\frac{2\pi i}{N}}$. We shall say that a $\mathbb{C}^{N\times N}$-valued ($N \geq 2$) function W meromorphic in the open unit disk $\mathbb{D}$ belongs to $\mathscr{C}_N$ if it is of the form

$$W(z) = \frac{1}{\sqrt{N}} \begin{pmatrix} \hat{s}_0(z) & \hat{s}_0(\epsilon_N z) & \cdots & \hat{s}_0(\epsilon_N^{N-1} z) \\ \hat{s}_1(z) & \hat{s}_1(\epsilon_N z) & \cdots & \hat{s}_1(\epsilon_N^{N-1} z) \\ \vdots & \vdots & & \vdots \\ \hat{s}_{N-1}(z) & \hat{s}_{N-1}(\epsilon_N z) & \cdots & \hat{s}_{N-1}(\epsilon_N^{N-1} z) \end{pmatrix}, \tag{38}$$

where $\hat{s}_0, \ldots, \hat{s}_{N-1}$ are complex-valued functions meromorphic in $\mathbb{D}$. Note that such a function, when analytic at the origin, will never be invertible there. A special case of this analyticity restriction of course is when $W(z)$ has polynomial entries. Under the filter-to-wavelet[3] correspondence [18], polynomial filters are the compactly supported wavelets. In the sequel, it will turn out that we shall concentrate on the opposite cases. Namely, not only $W(z)$ will have a pole at the origin, in fact, we shall have $W(z)^{-1}_{|_{z=0}} = 0_{N\times N}$.

Recall that we have denoted by P_N the permutation matrix

$$P_N = \begin{pmatrix} 0_{1\times(N-1)} & 1 \\ I_{N-1} & 0_{(N-1)\times 1} \end{pmatrix}$$

(see Eq. (2)).

[3]This correspondence: *polynomial filter* to *compactly supported wavelet* even works if $d > 1$.

Lemma 7.1. *A $\mathbb{C}^{N\times N}$-valued function meromorphic in the open unit disk is of the form (38) if and only if it satisfies Eq. (1):*

$$W(\epsilon_N z) = W(z)P_N.$$

Proof. Let W be a $\mathbb{C}^{N\times N}$-valued function meromorphic in $\mathbb{D}$ and satisfying Eq. (1), and let $s_1, \dots s_N$ denote its columns, i.e.,

$$W(z) = \begin{pmatrix} s_1(z) & s_2(z) & \dots & s_N(z) \end{pmatrix}. \tag{39}$$

Namely, from Eq. (38),

$$s_j(z) := \frac{1}{\sqrt{N}} \begin{pmatrix} \hat{s}_0(\epsilon_N^{j-1} z) \\ \hat{s}_1(\epsilon_N^{j-1} z) \\ \vdots \\ \hat{s}_{N-1}(\epsilon_N^{j-1} z) \end{pmatrix}, \qquad j = 1, \cdots, N.$$

Multiplying W by P_N from the right makes a cyclic shift of the columns to the left, namely

$$W(z)P_N = \begin{pmatrix} s_2(z) & s_3(z) & \cdots & s_N(z) & s_1(z) \end{pmatrix}.$$

Equation (1) then leads to

$$\begin{pmatrix} s_1(\epsilon_N z) & s_2(\epsilon_N z) & \cdots & s_{N-1}(\epsilon_N z) & s_N(\epsilon_N z) \end{pmatrix} = \\ = \begin{pmatrix} s_2(z) & s_3(z) & \cdots & s_N(z) & s_1(z) \end{pmatrix}.$$

Thus

$$s_2(z) = s_1(\epsilon_N z), \quad s_3(z) = s_1(\epsilon_N^2 z), \dots, s_N(z) = s_1(\epsilon_N^{N-1} z),$$

and so W is of the asserted form. The converse is clear. □

Note that in contrast to Lemma 7.1, in Eq. (37), we did not assume that W is square.

When one assumes that the function W in the previous lemma is a generalized Schur function, the symmetry condition (1) can be translated into the realization. We present the result for the closely outer connected coisometric realization, but similar results hold for the closely inner connected isometric realization and connected unitary realizations as well (see Sect. 4.2 for these notions). In the statement, recall that the state space $\mathscr{P}$ will in general be infinite dimensional and endowed with a Pontryagin space structure.

Theorem 7.2. *Let W be a generalized Schur function, and let*

$$W(z) = D + zC(I - zA)^{-1}B$$

be a closely inner coisometric realization of W, with state space $\mathscr{P}$. Then, W satisfies Eq. (1) if and only if there is a bounded invertible operator T from $\mathscr{H}$ into itself such that

$$\begin{pmatrix} \epsilon_N A & B \\ \epsilon_N C & D \end{pmatrix} \begin{pmatrix} T & 0 \\ 0 & I_N \end{pmatrix} = \begin{pmatrix} T & 0 \\ 0 & I_N \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix}. \tag{40}$$

Furthermore, the operator T satisfies

$$T^N = I. \tag{41}$$

Proof. The first equation follows from the uniqueness of the closely connected coisometric realization. Iterating Eq. (40) and taking into account that $\epsilon_N^N = 1$ we get

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} T^N & 0 \\ 0 & I_N \end{pmatrix} = \begin{pmatrix} T^N & 0 \\ 0 & I_N \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

By uniqueness of the similarity operator we have $T^N = I$. □

Proposition 7.3. *Let W_1 and W_2 be in $\mathscr{C}_N$. Then the functions*

$$W_1(z)W_2(\overline{z})^* \quad and \quad W_1(z)W_2(1/\overline{z})^*$$

are meromorphic functions of z^N.

Proof. Let $W(z) = W_1(z)W_2(\overline{z})^*$. Since $P_N P_N^* = I_N$, we have

$$\begin{aligned} W(\epsilon_N z) &= W_1(\epsilon_N z)W_2(\overline{\epsilon_N z})^* \\ &= W_1(z)P_N P_N^* W_2(\overline{z})^* \\ &= W_1(z)W_2(\overline{z})^* \\ &= W(z), \end{aligned}$$

i.e.,

$$W(\epsilon_N z) = W(z). \tag{42}$$

The functions W_1 and W_2 are meromorphic in the open unit disk and so is the function W. We denote by Λ the set of poles of W and by Λ_N the set of points w in the open unit disk such that $w^N \in \Lambda$. Let, for $z = re^{i\theta}$ with $r > 0$ and $\theta \in (-\pi, \pi]$,

$$R(z) = W(\sqrt[N]{r}e^{i\frac{\theta}{N}}).$$

The function R is analytic in $\mathbb{D}\setminus\{\Lambda_N \cup (-1, 0]\}$. Thanks to Eq. (42), it is continuous across the negative axis at those points in $(-1, 0)$ which are not in $\mathbb{D}\setminus\Lambda_N$. It follows that R is analytic in $\mathbb{D}\setminus\Lambda_N \cup \{0\}$. Furthermore, $W(z) = R(z^N)$. Any singular point

of R is a pole (otherwise its roots of order N would be essential singularities of W), and so R is meromorphic in $\mathbb{D}$. □

In the rational case, the previous result has an easier and more precise proof. Indeed consider the Laurent expansion at the origin of W:

$$W(z) = \sum_{-m_0}^{\infty} W_k z^k.$$

It converges in a punctured disk $0 < |z| < r$ for some $r > 0$. Equation (42) implies that

$$\sum_{-m_0}^{\infty} W_k z^k = \sum_{-m_0}^{\infty} W_k \epsilon_N^k z^k.$$

By uniqueness of the Laurent expansion we get that

$$W_k = 0, \quad \text{for} \quad k \notin N\mathbb{Z}.$$

Thus, if $m > 0$, we may assume without loss of generality that $m_0 = Nn_0$ for some $n_0 \in \mathbb{N}$. The function

$$W_-(z) = \sum_{k=-m_0}^{-1} W_k z^k$$

is rational and so is the function

$$W_+(z) = \sum_{k=0}^{\infty} W_k z^k.$$

We see that

$$W_-(z) = \sum_{-m_0 \le nN \le -N} W_{nN} z^{nN},$$

and so $W_-(z) = R_-(z^N)$, where the function

$$R_-(z) = \sum_{-m_0 \le nN \le -N} W_{nN} z^n$$

is rational and analytic at infinity. The function W_+ is analytic at the origin and thus can be written in realized form as

$$W_+(z) = D + zC(I_p - zA)^{-1}B.$$

Comparing with

$$W_+(z) = \sum_{n=0}^{\infty} W_{nN} z^{nN},$$

we have that

$$CA^p B = \begin{cases} 0 & \text{if } p+1 \notin N\mathbb{N}, \\ W_{nN} & \text{if } p+1 = nN, \quad n \in \mathbb{N}. \end{cases}$$

It follows that $W_+(z) = R_+(z^N)$, where R_+ is the rational function defined by

$$\begin{aligned} R_+(z) &= D + \sum_{n=1}^{\infty} z^n C A^{nN-1} B \\ &= D + \sum_{n=1}^{\infty} z^n C A^{(n-1)N} A^{N-1} B \\ &= D + zC(I_p - zA^N)^{-1} A^{N-1} B. \end{aligned}$$

The function

$$R(z) = R_-(z) + R_+(z)$$

is rational.

The proof of the preceding proposition can be mimicked to obtain the following result:

Proposition 7.4. *Let W_1 and W_2 be in $\mathscr{C}_N$, with nonidentically vanishing determinant. Then there exists a meromorphic function R such that*

$$W_1(z) W_2(z)^{-1} = R(z^N). \tag{43}$$

To this end, recall that the unitary matrix F_N:

$$F_N := \frac{1}{\sqrt{N}} \begin{pmatrix} \epsilon_N^{-(0\cdot 0)} & \epsilon_N^{-(0\cdot 1)} & \epsilon_N^{-(0\cdot 2)} & \cdots & e^{-(0\cdot(N-1))} \\ \epsilon_N^{-(1\cdot 0)} & \epsilon_N^{-(1\cdot 1)} & \epsilon_N^{-(1\cdot 2)} & \cdots & e^{-(1\cdot(N-1)} \\ \epsilon_N^{-(2\cdot 0)} & \epsilon_N^{-(2\cdot 1)} & \epsilon_N^{-(2\cdot 2)} & \cdots & e^{-(2\cdot(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \epsilon_N^{-((N-1)\cdot 0)} & \epsilon_N^{-((N-1)\cdot 1)} & \epsilon_N^{-((N-1)\cdot 2)} & \cdots & e^{-((N-1)\cdot(N-1)} \end{pmatrix}$$

generates the discrete Fourier transform. Namely, the discrete Fourier transform of $x \in \mathbb{C}^N$ is given by $X = F_N x$, and the inverse discrete Fourier transform is given by $x = F_N^* X$. Let furthermore

$$\hat{W}_N(z) := \operatorname{diag}\{1, z^{-1}, \ldots, z^{1-N}\} F_N. \tag{44}$$

With this special choice of W_2 the previous proposition becomes

Proposition 7.5. *$W \in \mathscr{C}_N$ and $\det W \not\equiv 0$ if and only if it can be written as*

$$W(z) = R(z^N)\hat{W}_N(z),$$

where R and $\hat{W}_N$ are as in Eqs. (43) and (44), respectively.

7.2 A Connection with Periodic Systems

Let

$$D_N(z) = \operatorname{diag}\,(z^N, z^{N-1}\epsilon_N^{N-1}, z^{N-2}\epsilon_N^{k-2}, \ldots, z\epsilon_N),$$

so that

$$D_N(1) = \operatorname{diag}\,(1, \epsilon_N^{N-1}, \epsilon_N^{k-2}, \ldots, \epsilon_N).$$

Functions which satisfy the related symmetry

$$W(\epsilon_N z) = D_N(1)^{-1} W(z) P_N \tag{45}$$

appear in the theory of periodic systems. A function W satisfies Eq. (45) if and only if it is of the form

$$W(z) = \frac{1}{\sqrt{N}} \begin{pmatrix} \hat{s}_0(z) & \hat{s}_0(\epsilon_N z) & \cdots & \hat{s}_0(\epsilon_N^{N-1} z) \\ \hat{s}_1(z) & \frac{1}{\epsilon_N}\hat{s}_1(\epsilon_N z) & \cdots & \frac{1}{\epsilon_N^{N-1}}\hat{s}_1(\epsilon_N^{N-1} z) \\ \vdots & & & \\ \hat{s}_{N-1}(z) & \frac{1}{\epsilon_N^{N-1}}\hat{s}_{N-1}(\epsilon_N z) & \cdots & \frac{1}{\epsilon_N^{(N-1)^2}}\hat{s}_{N-1}(\epsilon_N^{N-1} z). \end{pmatrix} \tag{46}$$

See [5, Theorem 4.1, p. 381]. We note that the corresponding general bitangential interpolation problem (see [14] for references) was solved in [5] for functions analytic and contractive in the open unit disk (i.e., for Schur functions). Let us denote by $\mathscr{P}er_N$ the family of functions meromorphic in the open unit disk and which satisfy Eq. (45).

Proposition 7.6. *The map $W \mapsto D_N W$ is one-to-one from $\mathscr{P}er_N$ onto $\mathscr{C}_N$. If W is analytic and contractive in the open unit disk, so is $D_N W$.*

Proof. We first note that

$$D_N(\epsilon_N z) = D_N(z) D_N(1). \tag{47}$$

Let now $W \in \mathscr{P}er_N$. In view of Eqs. (47) and (45) we have

$$\begin{aligned} D_N(\epsilon_N z)W(\epsilon_N z) &= D_N(z)D_N(1)D_N(1)^{-1}W(z)P_N \\ &= D_N(z)W(z)P_N, \end{aligned}$$

and so $D_N W \in \mathscr{C}_N$. □

Epilogue: A reason for the recent success of wavelet algorithms is a coming together of tools from engineering and harmonic analysis. While wavelets now enter into a multitude of applications from analysis and probability, it was the incorporation of ideas from signal processing that offered new and easy-to-use algorithms, and hence wavelets are now used in both discrete problems, as well as in harmonic analysis decompositions. Following this philosophy we here employed tools from system theory to wavelet problems and tried to show how ideas from wavelet decompositions throw light on factorizations used by engineers.

Since workers in wavelet theory often are not familiar with filterers in general, and FIR filters (short for finite impulse response) in particular, widely used in the engineering literature, we have taken the opportunity to include a section for mathematicians about filters. Conversely (in the other direction), engineers are often not familiar with wavelet analysis, and we have included a brief exposition of wavelet facts addressed to engineers . We showed that there are explicit actions of infinite-dimensional Lie groups which account for all the wavelet filters, as well as for other classes of filters used in systems theory. Moreover, we described these groups and explained how they arise in systems. The corresponding algorithms, including the discrete wavelet algorithms, are used in a variety of multi-scale problems, as used, for example, in data mining. These are the discrete algorithms, and we described their counterparts in harmonic analysis in standard $\mathbf{L}_2$ Lebesgue spaces, as well as in reproducing kernels Hilbert spaces. We also outlined the role of Pontryagin spaces in the study of stability questions.

In the engineering literature the study of filters is mostly confined to FIR filters. Recall that FIR filters correspond to having the spectrum at the origin. In our previous work [9] we have explained that the set of FIR wavelet filters is small in a sense we made precise. This suggests two possible conclusions:

1. It is unrealistic to offer optimization schemes over all FIR wavelet filters as part of the design procedure.
2. It calls upon using, at least in some circumstances, also stable IIR (short for infinite impulse response) wavelet filters, i.e., the spectrum is confined to the open unit disk.

The above extension to $\mathscr{U}^{I_N}$ allows us to consider filters whose spectrum is in $\mathbb{C}\setminus\mathbb{T}$. The generalization to $\mathscr{U}^{J}$ permits the spectrum to be everywhere in the complex plane.

Roughly, we hope that this message will be useful to practitioners in their use of these rigorous mathematics tools. We offer algorithms hopefully improving on those used before.

Acknowledgments D. Alpay thanks the Earl Katz family for endowing the chair which supported his research. The work was done in part while the second named author visited Department of Mathematics, Ben Gurion University of the Negev, supported by a BGU distinguished visiting scientist program. Support and hospitality are much appreciated. We acknowledge discussions with colleagues there and in the USA, Dorin Dutkay, Myung–Sin Song, and Erin Pearse.

References

1. Alpay, D.: Some remarks on reproducing kernel Kreĭn spaces. Rocky Mountain J. Math. **21**, 1189–1205 (1991)
2. Alpay, D.: The Schur algorithm, reproducing kernel spaces and system theory. American Mathematical Society, Providence, RI (2001) Translated from the 1998 French original by Stephen S. Wilson, Panoramas et Synthèses [Panoramas and Syntheses]
3. Alpay, D., Azizov, T.Ya., Dijksma, A., Rovnyak, J.: Colligations in Pontryagin spaces with a symmetric characteristic function. In: Linear operators and matrices, vol. 130 of Oper. Theory Adv. Appl., pp. 55–82. Birkhäuser, Basel (2002)
4. Alpay, D., Bolotnikov, V., Dijksma, A., De Snoo, H.: On some operator colligations and associated reproducing kernel Pontryagin spaces. J. Func. Anal. **136**, 39–80 (1996)
5. Alpay, D., Bolotnikov, V., Loubaton, Ph.: Dissipative periodic systems and symmetric interpolation in Schur classes. Arch. Math. (Basel) **68**, 371–387 (1997)
6. Alpay, D., Dijksma, A., Rovnyak, J., de Snoo, H.: Schur functions, operator colligations, and reproducing kernel Pontryagin spaces. vol. 96 of Operator theory: Advances and Applications, Birkhäuser, Basel (1997)
7. Alpay, D., Dym, H.: On applications of reproducing kernel spaces to the Schur algorithm and rational J-unitary factorization. In: Gohberg, I. (ed.) I. Schur methods in operator theory and signal processing, vol. 18 of Operator Theory: Advances and Applications, pp. 89–159. Birkhäuser, Basel (1986)
8. Alpay, D., Gohberg, I.: Unitary rational matrix functions. In: Gohberg, I. (ed.) Topics in interpolation theory of rational matrix-valued functions, vol. 33 of Operator Theory: Advances and Applications, pp. 175–222. Birkhäuser, Basel (1988)
9. Alpay, D., Jorgensen, P., Lewkowicz, I.: An easy-to-compute parameterizations of all wavelet filters: input-output and state-space, Preprint (2011) Avalaible at Arxiv at http://arxiv.org/abs/1105.0256
10. Alpay, D., Jorgensen, P., Lewkowicz, I., Marziano, I.: Representation formulas for Hardy space functions through the Cuntz relations and new interpolation problems. In: Shen, X., Zayed, A. (eds.) Multiscale signal analysis and modeling, Lecture Notes in Electrical Engineering, pp. 161–182. Springer (2013)
11. Azizov, T.Ya.: On the theory of extensions of J-isometric and J-symmetric operators. Funktsional. Anal. i Prilozhen. **18**(1), 57–58 (1984) English translation: Functional Analysis and Appl. **18**, 46–48 (1984)
12. Azizov, T.Ya., Iohvidov, I.S.: Foundations of the theory of linear operators in spaces with indefinite metric. Nauka, Moscow (1986) (Russian). English translation: Linear operators in spaces with an indefinite metric. Wiley, New York (1989)
13. Baggett, L., Jorgensen, P., Merrill, K., Packer, J.: A non-MRA C^r frame wavelet with rapid decay. Acta Appl. Math. **89**(1–3), 251–270 (2006) (2005)
14. Ball, J., Gohberg, I., Rodman, L.: Interpolation of rational matrix functions. vol. 45 of Operator Theory: Advances and Applications. Birkhäuser, Basel (1990)
15. Bognár, J.: Indefinite inner product spaces. Springer, Berlin (1974)
16. Branges, L.de., Rovnyak, J.: Canonical models in quantum scattering theory. In: Wilcox, C. (ed.) Perturbation theory and its applications in quantum mechanics, pp. 295–392. Wiley, New York (1966)

17. Branges, L.de., Rovnyak, J.: Square summable power series. Holt, Rinehart and Winston, New York (1966)
18. Bratteli, O., Jorgensen, P.: Wavelets through a looking glass. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2002) The world of the spectrum
19. Bratteli, O., Jorgensen, P.E.T.: Isometries, shifts, Cuntz algebras and multiresolution wavelet analysis of scale N. Integral Equations Operator Theory **28**(4), 382–443 (1997)
20. Bratteli, O., Jorgensen, P.E.T.: Wavelet filters and infinite-dimensional unitary groups. In: Wavelet analysis and applications (Guangzhou, 1999), vol. 25 of AMS/IP Stud. Adv. Math., pp. 35–65. Amer. Math. Soc., Providence, RI (2002)
21. Courtney, D., Muhly, P.S., Schmidt, S.W.: Composition Operators and Endomorphisms. ArXiv e-prints, March 2010
22. Courtney, D., Muhly, P.S., Schmidt, S.W.: Composition operators and endomorphisms, Complex Analysis and Operator Theory **6**(1), 163–188 (2012)
23. Crandall, M.G., Phillips, R.S.: On the extension problem for dissipative operators. J. Funct. Anal. **2**, 147–176 (1968)
24. Cuntz, J.: Simple C^*-algebras generated by isometries. Comm. Math. Phys. **57**(2), 173–185 (1977)
25. D'Andrea, J., Merrill, K.D., Packer, J.: Fractal wavelets of Dutkay-Jorgensen type for the Sierpinski gasket space. In: Frames and operator theory in analysis and signal processing, vol. 451 of Contemp. Math., pp. 69–88. Amer. Math. Soc., Providence, RI (2008)
26. Daubechies, I.: Ten lectures on wavelets, vol. 61 of CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1992)
27. Dijksma, A., Langer, H., de Snoo, H.S.V.: Representations of holomorphic operator functions by means of resolvents of unitary or selfadjoint operators in Kreĭn spaces. In: Operators in indefinite metric spaces, scattering theory and other topics (Bucharest, 1985), vol. 24 of Oper. Theory Adv. Appl., pp. 123–143. Birkhäuser, Basel (1987)
28. Dixmier, J.: C^*-algebras. North-Holland, Amsterdam (1977) Translated from the French by Francis Jellett, vol. 15. North-Holland Mathematical Library
29. Dritschel, M., Rovnyak, J.: Extensions theorems for contractions on Kreĭn spaces, vol. 47 of Operator theory: Advances and Applications, pp. 221–305. Birkhäuser, Basel (1990)
30. Dritschel, M., Rovnyak, J.: Operators on indefinite product spaces. In: Lancaster, P. (ed.) Lectures on operator theory and its applications, vol. 3 of Fields Institute Monographs, pp. 143–232. American Mathematical Society (1996)
31. Dutkay, D.E., Jorgensen, P.E.T.: Wavelets on fractals. Rev. Mat. Iberoam. **22**(1), 131–180 (2006)
32. Dutkay, D.E., Jorgensen, P.E.T.: Methods from multiscale theory and wavelets applied to nonlinear dynamics. In: Wavelets, multiscale systems and hypercomplex analysis, vol. 167 of Oper. Theory Adv. Appl., pp. 87–126. Birkhäuser, Basel (2006)
33. Fuhrmann, P.A.: A Polynomial Approach to Linear Algebra. Universitext. Springer, New York (1996)
34. Glimm, J.: Locally compact transformation groups. Trans. Amer. Math. Soc. **101**, 124–138 (1961)
35. Glimm, J.: Type I C^*-algebras. Ann. Math. **73**(2), 572–612 (1961)
36. Iohvidov, I.S., Kreĭn, M.G., Langer, H.: Introduction to the Spectral Theory of Operators in Spaces with an Indefinite Metric. Akademie, Berlin (1982)
37. Jorgensen, P.E.T.: Matrix factorizations, algorithms, wavelets. Notices Amer. Math. Soc. **50**(8), 880–894 (2003)
38. Jorgensen, P.E.T.: Analysis and probability: wavelets, signals, fractals. vol. 234 of Graduate Texts in Mathematics. Springer, New York (2006)
39. Jorgensen, P.E.T.: Certain representations of the Cuntz relations, and a question on wavelets decompositions. In: Operator theory, operator algebras, and applications, vol. 414 of Contemp. Math., pp. 165–188. Amer. Math. Soc., Providence (2006)

40. Jorgensen, P.E.T.: Use of operator algebras in the analysis of measures from wavelets and iterated function systems. In: Operator theory, operator algebras, and applications, vol. 414 of Contemp. Math., pp. 13–26. Amer. Math. Soc., Providence (2006)
41. Jorgensen, P.E.T.: Frame analysis and approximation in reproducing kernel Hilbert spaces. In: Frames and operator theory in analysis and signal processing, vol. 451 of Contemp. Math., pp. 151–169. Amer. Math. Soc., Providence (2008)
42. Jorgensen, P.E.T., Song, M.-S.: Analysis of fractals, image compression, entropy encoding, Karhunen-Loève transforms. Acta Appl. Math. **108**(3), 489–508 (2009)
43. Keinert, F.: Wavelets and multiwavelets. Studies in Advanced Mathematics. Chapman & Hall/CRC, Boca Raton (2004)
44. Kreĭn, M.G., Langer, H.: Über die verallgemeinerten Resolventen und die charakteristische Funktion eines isometrischen Operators im Raume Π_k. In: Hilbert space operators and operator algebras (Proc. Int. Conf. Tihany, 1970), pp. 353–399. North–Holland, Amsterdam (1972) Colloquia Math. Soc. János Bolyai
45. Kreĭn, M.G., Langer, H.: Über einige Fortsetzungsprobleme, die eng mit der Theorie hermitescher Operatoren im Raume π_k zusammenhangen. I. Einige Funktionenklassen und ihre Darstellungen. Math. Nachrichten **77**, 187–236 (1977)
46. Kreĭn, M.G., Langer, H.: Über einige Fortsetzungsprobleme, die eng mit der Theorie hermitescher Operatoren im Raume Π_κ zusammenhängen. II. Verallgemeinerte Resolventen, u-Resolventen und ganze Operatoren. J. Funct. Anal. **30**(3), 390–447 (1978)
47. Kreĭn, M.G., Langer, H.: On some extension problems which are closely connected with the theory of Hermitian operators in a space Π_κ. III. Indefinite analogues of the Hamburger and Stieltjes moment problems. Part II. Beiträge Anal. **15**, 27–45 (1981) (1980)
48. Kreĭn, M.G., Langer, H.: Some propositions on analytic matrix functions related to the theory of operators in the space π_k. Acta Sci. Math. **43**, 181–205 (1981)
49. Krommweh, J.: Tight frame characterization of multiwavelet vector functions in terms of the polyphase matrix. Int. J. Wavelets Multiresolut. Inf. Process. **7**(1), 9–21 (2009)
50. Lawton, W.M.: Multiresolution properties of the wavelet Galerkin operator. J. Math. Phys. **32**(6), 1440–1443 (1991)
51. Lawton, W.M.: Necessary and sufficient conditions for constructing orthonormal wavelet bases. J. Math. Phys. **32**(1), 57–61 (1991)
52. Lax, P.D., Phillips, R.S.: Purely decaying modes for the wave equation in the exterior of an obstacle. In: Proc. Internat. Conf. on Functional Analysis and Related Topics (Tokyo, 1969), pp. 11–20. Univ. of Tokyo Press, Tokyo (1970)
53. Lax, P.D., Phillips, R.S.: A logarithmic bound on the location of the poles of the scattering matrix. Arch. Rational Mech. Anal. **40**, 268–280 (1971)
54. Lax, P.D., Phillips, R.S.: Scattering theory for automorphic functions. Princeton Univ. Press, Princeton (1976) Annals of Mathematics Studies, No. 87
55. Lax, P.D., Phillips, R.S.: Scattering theory. vol. 26 of Pure and Applied Mathematics, 2nd edn. Academic, Boston (1989) With appendices by Cathleen S. Morawetz and Georg Schmidt
56. Mallat, S.: A wavelet tour of signal processing, 3rd edn. Elsevier/Academic Press, Amsterdam (2009) The sparse way, With contributions from Gabriel Peyré
57. Muhly, P.S., Solel, B.: Schur class operator functions and automorphisms of Hardy algebras. Doc. Math. **13**, 365–411 (2008)
58. Phillips, R.S.: The extension of dual subspaces invariant under an algebra. In: Proc. Internat. Sympos. Linear Spaces (Jerusalem, 1960), pp. 366–398. Jerusalem Academic, Jerusalem (1961)
59. Potapov, V.P.: The multiplicative structure of J-contractive matrix–functions. Trudy Moskow. Mat. Obs. **4**, 125–236 (1955) English translation in: American mathematical society translations **15**(2), 131–243 (1960)
60. Resnikoff, H.L., Tian, J., Wells, Jr., R.O.: Biorthogonal wavelet space: parametrization and factorization. SIAM J. Math. Anal. **33**(1), 194–215 (electronic) (2001)
61. Rosenblum, M., Rovnyak, J.: Hardy Classes and Operator Theory. Birkhäuser, Basel (1985)

62. Schwartz, L.: Sous espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). J. Analyse Math. **13**, 115–256 (1964)
63. Sebert, F.M., Zou, Y.M.: Factoring Pseudoidentity Matrix Pairs. SIAM J. Math. Anal. **43**, 565–576 (2011)

On the Group-Theoretic Structure of Lifted Filter Banks

Christopher M. Brislawn

Abstract The polyphase-with-advance matrix representations of whole-sample symmetric (WS) unimodular filter banks form a multiplicative matrix Laurent polynomial group. Elements of this group can always be factored into lifting matrices with half-sample symmetric (HS) off-diagonal lifting filters; such linear phase lifting factorizations are specified in the ISO/IEC JPEG 2000 image coding standard. Half-sample symmetric unimodular filter banks do not form a group, but such filter banks can be *partially* factored into a cascade of whole-sample *antisymmetric* (WA) lifting matrices starting from a concentric, equal-length HS base filter bank. An algebraic framework called a *group lifting structure* has been introduced to formalize the group-theoretic aspects of matrix lifting factorizations. Despite their pronounced differences, it has been shown that the group lifting structures for both the WS and HS classes satisfy a *polyphase order-increasing property* that implies uniqueness ("modulo rescaling") of irreducible group lifting factorizations in both group lifting structures. These unique factorization results can in turn be used to characterize the group-theoretic structure of the groups generated by the WS and HS group lifting structures.

Keywords Lifting • Filter bank • Linear phase filter • Group theory • Group lifting structure • JPEG 2000 • Wavelet • Polyphase matrix • Unique factorization • Matrix polynomial

1 Introduction

Lifting [9, 22, 23] is a general technique for factoring the polyphase matrix representation of a perfect reconstruction multirate filter bank into elementary matrices over the Laurent polynomials. As one might expect of a technique as

C.M. Brislawn (✉)
Los Alamos National Laboratory, Los Alamos, NM, 87545, USA
e-mail: brislawn@lanl.gov

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_6,

universal as elementary matrix factorization, lifting has proven extremely useful for both theoretical investigations and practical applications. For instance, lifting forms the basis for specifying discrete wavelet transforms in the ISO/IEC JPEG 2000 standards [12, 13].

In addition to providing a completely general mathematical framework for standardizing discrete wavelet transforms, lifting also provides a cascade structure for *reversible* filter banks—nonlinear implementations of linear filter banks that furnish bit-perfect invertibility in fixed-precision arithmetic [5, 6, 19, 26]. Reversibility allows digital communication systems to realize the efficiency and scalability of subband coding while also providing the option of lossless transmission, a key feature that made lifting a particularly attractive choice for the JPEG 2000 standard.

The author became acquainted with lifting while serving on the JPEG 2000 standard, and he was struck by the group-theoretic flavor of the subject. After completing his standards committee work, he began studying the lifting structure of two-channel linear phase FIR filter banks in depth, leading to the publications outlined in the present chapter. In spite of its universality, lifting is not particularly well suited for analyzing *paraunitary* filter banks because, as discussed in [1, Sect. IV], lifting matrices are never paraunitary. This means lifting factorization takes place *outside* of the paraunitary group, whereas we shall show that lifting factorization can be defined to take place entirely *within* the group of whole-sample symmetric (WS, or odd-length linear phase) filter banks by decomposing WS filter banks into linear phase lifting steps. This allows us to prove both existence and (rather surprisingly) uniqueness of "irreducible" WS group lifting factorizations. One consequence of this unique factorization theory is that we can characterize the group-theoretic structure of the unimodular WS filter bank group up to isomorphism using standard group-theoretic constructs.

Besides WS filter banks, there is also a class of half-sample symmetric (HS, or even-length linear phase) filter banks. The differences between the group-theoretic structure of WS and HS filter banks are striking. For instance, HS filter banks do *not* form a matrix group, but linear phase "partial" lifting factorizations partition the class of unimodular HS filter banks into *cosets* of a particular matrix group generated by whole-sample *antisymmetric* (WA) lifting filters. The complete group-theoretic classification of unimodular HS filter banks is still incomplete as of this writing but comprises an extremely active area of research by the author.

This chapter is an expository overview of recent research [1–4]. It is targeted at a mathematical audience that has at least a passing familiarity with elementary group theory and with the connections between wavelet transforms and multirate filter banks.

1.1 Perfect Reconstruction Filter Banks

This chapter studies two-channel multirate digital filter banks of the form shown in Fig. 1 [7, 8, 15, 21, 24, 25]. We only consider systems in which both the analysis

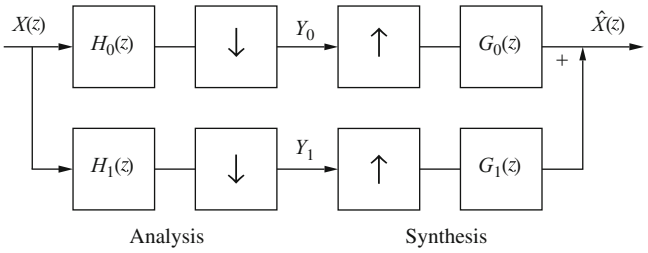


Fig. 1 Two-channel perfect reconstruction multirate filter bank

filters $\{H_0(z),\ H_1(z)\}$ and the synthesis filters $\{G_0(z),\ G_1(z)\}$ are linear translation-invariant (or time-invariant) finite impulse response (FIR) filters. A system like that in Fig. 1 is called a *perfect reconstruction multirate filter bank* (frequently abbreviated to just "filter bank" in this chapter) if it is a linear translation-invariant system with a transfer function satisfying

$$\frac{\hat{X}(z)}{X(z)} = az^{-d} \tag{1}$$

for some integer $d \in \mathbb{Z}$ and some constant $a \neq 0$.

FIR filters are written in the transform domain as Laurent polynomials,

$$F(z) \equiv \sum_{n=a}^{b} f(n)\, z^{-n} \in \mathbb{C}\left[z, z^{-1}\right],$$

with impulse response $f(n)$. The *support interval* of an FIR filter, denoted

$$\text{supp\_int}(F) \equiv \text{supp\_int}(f) \equiv [a, b] \subset \mathbb{Z}, \tag{2}$$

is the smallest closed interval of integers containing the support of the filter's impulse response or, equivalently, the largest closed interval for which $f(a) \neq 0$ and $f(b) \neq 0$. If $\text{supp\_int}(f) = [a, b]$ then the *order* of the filter is

$$\text{order}(F) \equiv b - a. \tag{3}$$

1.2 The Polyphase-with-Advance Representation

It is more efficient to compute the decimated output of a filter bank like the one in Fig. 1 by splitting the signal into even- and odd-indexed subsequences,

$$x_i(n) \equiv x(2n + i),\ i = 0,\ 1; \quad X(z) = X_0(z^2) + z^{-1} X_1(z^2). \tag{4}$$

The *polyphase vector form* of a discrete-time signal is defined to be

$$\boldsymbol{x}(n) \equiv \begin{bmatrix} x_0(n) \\ x_1(n) \end{bmatrix}; \quad \boldsymbol{X}(z) \equiv \begin{bmatrix} X_0(z) \\ X_1(z) \end{bmatrix}. \tag{5}$$

The *analysis polyphase-with-advance representation* of a filter [4, Eq. (9)] is

$$f_j(n) \equiv f(2n-j),\ j=0,\ 1; \quad F(z) = F_0(z^2) + zF_1(z^2).$$

Its *analysis polyphase vector representation* is

$$\boldsymbol{F}(z) \equiv \begin{bmatrix} F_0(z) \\ F_1(z) \end{bmatrix} = \sum_{n=c}^{d} \boldsymbol{f}(n)\, z^{-n}, \tag{6}$$

$$\boldsymbol{f}(n) \equiv \begin{bmatrix} f_0(n) \\ f_1(n) \end{bmatrix} \quad \text{with } \boldsymbol{f}(c),\ \boldsymbol{f}(d) \neq \mathbf{0}. \tag{7}$$

Since we generally work with analysis filter bank representations, "polyphase" will mean "analysis polyphase-with-advance." The polyphase filter (6), (7) has the *polyphase support interval*

$$\text{supp\_int}(\boldsymbol{f}) \equiv [c, d], \tag{8}$$

which differs from the scalar support interval (2) for the same filter. The *polyphase order* of (6) is

$$\text{order}(\boldsymbol{F}) \equiv d - c\ . \tag{9}$$

These definitions generalize for FIR filter banks, $\{H_0(z),\ H_1(z)\}$. Decompose each filter $H_i(z)$ into its polyphase vector representation $\boldsymbol{H}_i(z)$ as in (6) and form the *polyphase matrix*

$$\mathbf{H}(z) \equiv \begin{bmatrix} \boldsymbol{H}_0^T(z) \\ \boldsymbol{H}_1^T(z) \end{bmatrix} = \sum_{n=c}^{d} \mathbf{h}(n)\, z^{-n}, \tag{10}$$

$$\mathbf{h}(n) \equiv \begin{bmatrix} \boldsymbol{h}_0^T(n) \\ \boldsymbol{h}_1^T(n) \end{bmatrix} \quad \text{with } \mathbf{h}(c),\ \mathbf{h}(d) \neq \mathbf{0}. \tag{11}$$

Bold italics denote column vectors and bold roman (upright) fonts denote matrices.

The polyphase support interval of the filter bank in (10), (11) is defined to be

$$\text{supp\_int}(\mathbf{h}) \equiv [c, d], \tag{12}$$

and the polyphase order is defined to be

$$\text{order}(\mathbf{H}) \equiv d - c. \tag{13}$$

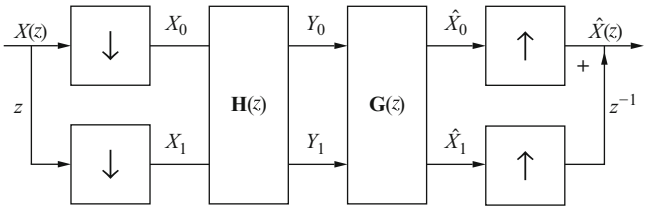


Fig. 2 The polyphase-with-advance representation of a two-channel multirate filter bank

With this notation, the output of the analysis bank in Fig. 1 can be written

$$\boldsymbol{Y}(z) = \mathbf{H}(z)\boldsymbol{X}(z).$$

An analogous synthesis polyphase matrix representation, $\mathbf{G}(z)$, can be defined for the synthesis filter bank $\{G_0(z),\ G_1(z)\}$; see [4, Sect. II-A].

The block diagram for this matrix-vector filter bank representation, which we call the *polyphase-with-advance representation* [4], is shown in Fig. 2. The polyphase representation transforms the non-translation-invariant analysis bank of Fig. 1 into a demultiplex operation, $x(k) \mapsto \boldsymbol{x}(n)$, followed by a linear translation-invariant operator acting on vector-valued signals. The polyphase representation therefore reduces the study of multirate filter banks to the study of invertible transfer matrices over the Laurent polynomials.

Since Laurent *mono*mials are units, invertibility of $\mathbf{H}(z)$ over $\mathbb{C}[z, z^{-1}]$ is equivalent to

$$|\mathbf{H}(z)| \equiv \det \mathbf{H}(z) = \breve{a} z^{-\breve{d}}; \quad \breve{a} \neq 0,\ \breve{d} \in \mathbb{Z}. \tag{14}$$

$\breve{d}$ is called the *determinantal delay* of $\mathbf{H}(z)$ and $\breve{a}$ is called the *determinantal amplitude*. A filter bank satisfying (14) is called an *FIR perfect reconstruction (PR) filter bank* [24]. It was noted in [4, Theorem 1] that the family $\mathcal{F}$ of all FIR PR filter banks forms a nonabelian matrix group, called the *FIR filter bank group*. The *unimodular group*, $\mathcal{N}$, is the normal subgroup of $\mathcal{F}$ consisting of all matrices of determinant 1,

$$|\mathbf{H}(z)| = 1\,. \tag{15}$$

The unimodular group can also be regarded as $SL(2, \mathbb{C}\left[z, z^{-1}\right])$.

1.3 Linear Phase Filter Banks

It is easily shown [4, Eq. (20)] that a discrete-time signal is symmetric about one of its samples, $x(i_0)$, if and only if its polyphase vector representation (5) satisfies

$$\boldsymbol{X}(z^{-1}) = z^{i_0}\boldsymbol{\Lambda}(z)\boldsymbol{X}(z), \quad \text{where} \quad \boldsymbol{\Lambda}(z) \equiv \operatorname{diag}(1, z^{-1}). \tag{16}$$

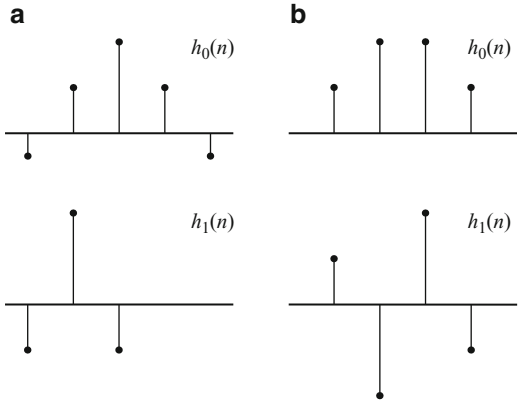


Fig. 3 (**a**) Whole-sample symmetric filter bank (**b**) Half-sample symmetric filter bank

We say a signal satisfying (16) is *whole-sample symmetric (WS)* about $i_0 \in \mathbb{Z}$. Similarly, a discrete-time signal is *half-sample symmetric (HS)* about an odd multiple of 1/2 (indexed by $i_0 \in \mathbb{Z}/2$) if and only if

$$\boldsymbol{X}(z^{-1}) = z^{(2i_0-1)/2}\mathbf{J}\boldsymbol{X}(z), \quad \text{where} \quad \mathbf{J} \equiv \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{17}$$

Analogous characterizations of whole- and half-sample *antisymmetry* (abbreviated WA and HA, respectively) are obtained by putting minus signs in (16) and (17). Real-valued discrete-time signals (or filters) possessing any of these symmetry properties are called *linear phase* signals (filters).

It was proven in [16] that the only *nontrivial* classes (classes with at least one nontrivial real degree of freedom) of two-channel FIR PR linear phase filter banks are the whole- and half-sample symmetric classes shown in Fig. 3. Arbitrary combinations of symmetry are not necessarily compatible with invertibility; e.g., if both filters have odd lengths then both must be symmetric (WS). In an even-length filter bank, one filter must be symmetric (HS) while the other must be antisymmetric (HA). It was also proven in [16] that the sum of the impulse response lengths must be a multiple of 4, so it is possible for HS (but not WS) filter banks to have filters of *equal lengths*, as shown in Fig. 3.

Linear phase properties of filter banks are also straightforward to characterize in the polyphase domain [4, Sect. III]. The *group delay* [17] of a linear phase FIR filter is equal to the midpoint (or axis of symmetry) of the filter's impulse response. Let d_i denote the group delay of h_i for $i = 0, 1$.

Lemma 1 ([4], Lemma 2). *A real-coefficient FIR transfer matrix* $\mathbf{H}(z)$ *is a WS analysis filter bank with group delays* d_0 *and* d_1 *if and only if*

$$\mathbf{H}(z^{-1}) = \operatorname{diag}(z^{d_0}, z^{d_1})\mathbf{H}(z)\boldsymbol{\Lambda}(z^{-1}). \tag{18}$$

If $\mathbf{H}(z)$ satisfies (14) then the *delay-minimized WS filter bank* normalization

$$d_0 = 0,\ d_1 = -1 \tag{19}$$

ensures that the determinantal delay, $\check{d} = (d_0 + d_1 + 1)/2$, is zero and (18) becomes

$$\mathbf{H}(z^{-1}) = \mathbf{\Lambda}(z)\mathbf{H}(z)\mathbf{\Lambda}(z^{-1}). \tag{20}$$

The analogous delay-minimized HS filter bank normalization is

$$d_0 = -1/2 = d_1. \tag{21}$$

Both filters have the same axis of symmetry, as in Fig. 3b; we call such filter banks *concentric*. Delay-minimized HS filter banks are characterized by the relation

$$\mathbf{H}(z^{-1}) = \mathbf{L}\mathbf{H}(z)\mathbf{J} \quad \text{where} \quad \mathbf{L} \equiv \text{diag}(1, -1). \tag{22}$$

We now see a striking difference between the algebraic properties of WS and HS filter banks. Since $\mathbf{\Lambda}(z^{-1}) = \mathbf{\Lambda}^{-1}(z)$, (20) says that $\mathbf{\Lambda}(z)$ *intertwines* $\mathbf{H}(z)$ and $\mathbf{H}(z^{-1})$, so the set of all filter banks satisfying (20) (i.e., the set of all delay-minimized WS filter banks) forms a multiplicative group. In sharp contrast, filter banks satisfying (22) do *not* form a group.

Definition 1 ([1], Definition 8). The *unimodular WS group*, $\mathcal{W}$, is the group of all real FIR transfer matrices that satisfy both (15) and (20).

Definition 2 ([1], Definition 9). The *unimodular HS class*, $\mathfrak{H}$, is the set of all real FIR transfer matrices that satisfy both (15) and (22).

2 Lifting Factorization of Linear Phase Filter Banks

We now define lifting and apply it to linear phase filter banks, focusing on the problem of factoring linear phase filter banks into linear phase lifting steps.

2.1 Lifting Factorizations

Daubechies and Sweldens [9] used the Euclidean algorithm for $\mathbb{C}[z, z^{-1}]$ to prove that any unimodular FIR transfer matrix can be decomposed into a *lifting factorization* (or *lifting cascade*) of the form

$$\mathbf{H}(z) = \mathbf{D}_K\, \mathbf{S}_{N-1}(z) \cdots \mathbf{S}_1(z)\, \mathbf{S}_0(z)\ . \tag{23}$$

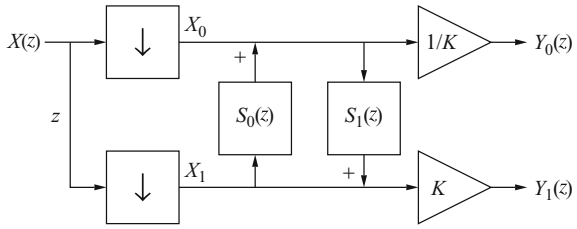


Fig. 4 Two-step lifting representation of a unimodular filter bank

The diagonal matrix $\mathbf{D}_K \equiv \mathrm{diag}(1/K, K)$ is a *unimodular gain-scaling matrix* with *scaling factor* $K \neq 0$. The lifting matrices $\mathbf{S}_i(z)$ are upper- or lower-triangular with ones on the diagonal and a lifting filter, $S_i(z)$, in the off-diagonal position.

In the factorization corresponding to Fig. 4, the lifting matrix for the step $S_0(z)$ (which is a lowpass update) is upper-triangular and the matrix for the second step (a highpass update) is lower-triangular. For example, the Haar filter bank

$$H_0(z) = (z+1)/2, \quad H_1(z) = z - 1, \tag{24}$$

has a unimodular polyphase representation with two different lifting factorizations,

$$\mathbf{H}_{haar}(z) \equiv \begin{bmatrix} 1/2 & 1/2 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1/2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \tag{25}$$

$$= \begin{bmatrix} 1 & 1/2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}. \tag{26}$$

Factorization (25) fits the ladder structure of Fig. 4 with $S_0(z) = 1$, $S_1(z) = -1/2$, and $K = 2$. Factorization (26), on the other hand, begins with a *highpass* lifting update and does not require a gain-scaling operation.

Definition 3 ([13], Annex G). The *update characteristic* of a lifting step (or lifting matrix) is a binary flag, $m = 0$ or 1, indicating which polyphase channel is being updated by the lifting step.

For instance, the update characteristic, m_0, of the first lifting step in Fig. 4 is "lowpass," coded with a zero ($m_0 = 0$), while the update characteristic of the second step is "highpass" ($m_1 = 1$). The update characteristic m_i is defined similarly for each matrix $\mathbf{S}_i(z)$ in a lifting cascade (23).

Next, we generalize (23) slightly to accommodate factorizations that lift one filter bank to another. A *partially factored lifting cascade*,

$$\mathbf{H}(z) = \mathbf{D}_K \, \mathbf{S}_{N-1}(z) \cdots \mathbf{S}_0(z) \, \mathbf{B}(z), \tag{27}$$

is an expansion relative to some *base* filter bank, $\mathbf{B}(z)$, with scalar filters $B_0(z)$ and $B_1(z)$. We sometimes write such factorizations in recursive form:

$$\begin{aligned} \mathbf{H}(z) &= \mathbf{D}_K\, \mathbf{E}^{(N-1)}(z), \\ \mathbf{E}^{(n)}(z) &= \mathbf{S}_n(z)\, \mathbf{E}^{(n-1)}(z), \quad 0 \le n < N, \\ \mathbf{E}^{(-1)}(z) &\equiv \mathbf{B}(z). \end{aligned} \tag{28}$$

2.2 Whole-Sample Symmetric Filter Banks

The fact that delay-minimized WS filter banks form a group makes it easy to characterize the lifting matrices that lift one delay-minimized WS filter bank to another,

$$\mathbf{F}(z) = \mathbf{S}(z)\, \mathbf{H}(z). \tag{29}$$

Lemma 2 ([4], Lemma 8). *A lifting matrix, $\mathbf{S}(z)$, lifts a filter bank satisfying (20) to another filter bank satisfying (20) if and only if $\mathbf{S}(z)$ also satisfies (20). An upper-triangular lifting matrix satisfies (20) if and only if its lifting filter is half-sample symmetric about $1/2$. A lower-triangular lifting matrix satisfies (20) if and only if its lifting filter is HS about $-1/2$.*

Note that HS lifting *filters* with appropriate group delays form lifting *matrices* that are WS filter banks. It is easy to show that the lifting filters symmetric about $1/2$ form an additive group, $\mathcal{P}_0$, of Laurent polynomials and that the upper-triangular lifting matrices with lifting filters in $\mathcal{P}_0$ form a multiplicative group, $\mathcal{U}$. Similarly, the lifting filters symmetric about $-1/2$ form an additive group, $\mathcal{P}_1$, and the lower-triangular lifting matrices with lifting filters in $\mathcal{P}_1$ form a multiplicative group, $\mathcal{L}$.

Given Lemma 2, it is natural to ask whether every filter bank in $\mathcal{W}$ has a lifting factorization of the form (23) in which every lifting matrix $\mathbf{S}_i(z)$ satisfies (20). The answer is yes, and the proof is a constructive, order-reducing recursion that does not rely on the Euclidean algorithm.

Theorem 1 ([4], Theorem 9). *A unimodular filter bank, $\mathbf{H}(z)$, satisfies the delay-minimized WS condition (20) if and only if it can be factored as*

$$\mathbf{H}(z) = \mathbf{D}_K\, \mathbf{S}_{N-1}(z) \cdots \mathbf{S}_1(z)\, \mathbf{S}_0(z), \tag{30}$$

where each lifting matrix, $\mathbf{S}_i(z)$, satisfies (20).

We refer to such decompositions as *WS group lifting factorizations*. This is the form of lifting factorizations specified in [13, Annex G] for user-defined WS filter banks.

Definition 1 of the unimodular WS group, $\mathcal{W}$, is independent of lifting, but we need lifting to define *reversible* WS filter banks. Let $\mathcal{U}_r$ and $\mathcal{L}_r$ be the subgroups of $\mathcal{U}$ and $\mathcal{L}$ with matrices whose lifting filters have *dyadic* coefficients of the form

$k \cdot 2^n$, $k, n \in \mathbb{Z}$. Since gain-scaling operations are not generally invertible in fixed-precision arithmetic, gain scaling is not used in reversible implementations.

Definition 4 ([1], Example 3). The group $\mathcal{W}_r$ of reversible unimodular WS filter banks is defined to be the group of all transfer matrices $\mathbf{H}(z)$ generated by lifting factorizations (30) where $\mathbf{S}_i(z) \in \mathcal{U}_r \cup \mathcal{L}_r$ and $\mathbf{D}_K = \mathbf{I}$.

2.3 Half-Sample Symmetric Filter Banks

Lifting factorization of HS filter banks is harder (i.e., more interesting) than lifting factorization of WS filter banks, in part "because" HS filter banks do not form a group. For instance, the characterization in Lemma 2 of lifting matrices that lift one WS filter bank to another is equally valid for *left* lifts, as in (29), and *right* lifts in which $\mathbf{S}(z)$ acts on the right. This fails badly for HS filter banks.

Theorem 2 ([4], Theorem 12). *Suppose that* $\mathbf{H}(z)$ *is an HS filter bank satisfying the concentric delay-minimized condition (22). If* $\mathbf{F}(z)$ *is right-lifted from* $\mathbf{H}(z)$,

$$\mathbf{F}(z) = \mathbf{H}(z)\,\mathbf{S}(z),$$

then $\mathbf{F}(z)$ *can only satisfy (22) if* $\mathbf{S}(z) = \mathbf{I}$ *and* $\mathbf{F}(z) = \mathbf{H}(z)$.

Fortunately, half-sample symmetry can be preserved by left-lifting operations.

Lemma 3 ([4], Lemma 10). *If either* $\mathbf{H}(z)$ *or* $\mathbf{F}(z)$ *in (29) is an HS filter bank satisfying the concentric delay-minimized condition (22), then the other filter bank also satisfies (22) if and only if* $\mathbf{S}(z)$ *satisfies*

$$\mathbf{S}(z^{-1}) = \mathbf{L}\,\mathbf{S}(z)\,\mathbf{L} = \mathbf{S}^{-1}(z), \tag{31}$$

which says that the lifting filter is whole-sample antisymmetric (WA) about 0.

WA lifting filters form an additive group, $\mathcal{P}_a$, and the upper-triangular (resp., lower-triangular) lifting matrices with lifting filters in $\mathcal{P}_a$ form a group, $\mathcal{U}$ (resp., $\mathcal{L}$). In contrast to WS group lifting factorizations, concentric delay-minimized HS filter banks *never* factor completely into WA lifting steps [4, Theorem 13]. The obstruction, which does not exist for WS filter banks, is the possibility that a reduced-order intermediate HS filter bank in the factorization process will correspond to filters $H_0(z)$ and $H_1(z)$ of *equal lengths.* Given a concentric equal-length HS filter bank, it is *never* possible to reduce its order by factoring off a WA lifting step. This leaves us with an incomplete lifting theory for unimodular HS filter banks.

Theorem 3 ([4], Theorem 14). *A unimodular filter bank,* $\mathbf{H}(z)$, *satisfies the concentric delay-minimized HS convention (22) if and only if it can be decomposed into a partially factored lifting cascade of WA lifting steps satisfying (31) and a concentric equal-length HS base filter bank* $\mathbf{B}(z)$ *satisfying (22):*

$$\mathbf{H}(z) = \mathbf{S}_{N-1}(z) \cdots \mathbf{S}_0(z)\, \mathbf{B}(z). \tag{32}$$

There is no gain-scaling matrix, $\mathbf{D}_K$, in (32) since $\mathbf{B}(z)$ has been left unfactored.

One popular choice for the equal-length base filter bank in HS lifting constructions is the Haar filter bank, which has a particularly simple lifting factorization (26). The 2-tap/10-tap HS filter bank specified in JPEG 2000 Part 2 [13, Annex H.4.1.1.3] is lifted from the Haar via a lower-triangular 4th-order WA lifting step. Another important example is the 6-tap/10-tap HS filter bank in [13, Annex H.4.1.2.1]. This filter bank was originally constructed by spectral factorization and has a lifting factorization of the form $\mathbf{H}(z) = \mathbf{S}(z)\mathbf{B}(z)$, where $S(z)$ is a second-order WA filter and $\mathbf{B}(z)$ is an equal-length (6-tap/6-tap) HS filter bank.

Defining a class $\mathfrak{H}_r$ of *reversible* HS filter banks is awkward; see [1, Example 5].

3 Uniqueness of Linear Phase Lifting Factorizations

In the last section we saw that every filter bank in the unimodular WS and HS classes factors into linear phase lifting steps of an appropriate form. Lifting factorizations, like other elementary matrix decompositions, are highly nonunique, and although linear phase factorizations are more specialized than general lifting decompositions, there seems little reason *a priori* to expect them to be unique. There are, however, a few trivial causes of nonuniqueness that we can exclude in an *ad hoc* fashion.

Definition 5 ([1], Definition 3). A lifting cascade (27) is *irreducible* if all lifting steps are nontrivial ($\mathbf{S}_i(z) \neq \mathbf{I}$) and there are no consecutive lifting matrices with the same update characteristic, i.e., the lifting matrices strictly alternate between lower- and upper-triangular.

Every lifting cascade can be simplified to an irreducible cascade using matrix multiplication. Merely restricting attention to irreducible lifting cascades is far from sufficient to ensure unique factorizations, as the two irreducible lifting factorizations of the Haar filter bank (25)–(26) show. To view nonuniqueness in a different light, move the lifting steps from (26) over to the right end of (25) and use [9, Sect. 7.3] to factor $\mathrm{diag}(1/2, 2)$ into lifting steps. This results in an irreducible lifting factorization of the identity,

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1/2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1/2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1/2 \\ 0 & 1 \end{bmatrix}. \tag{33}$$

In a similar manner, *any* transfer matrix with two distinct irreducible lifting factorizations gives rise to an irreducible factorization of the identity; cf. [1, Example 1], which presents an irreducible, reversible lifting factorization of the

identity using linear phase (HS and HA) lifting filters. By constructing irreducible lifting factorizations of the identity, it is possible to sharpen the universal lifting factorization result of [9] into the following universal *nonunique* factorization result.

Proposition 1 ([1], Proposition 1). *If* $\mathbf{G}(z)$ *and* $\mathbf{H}(z)$ *are any FIR perfect reconstruction filter banks then* $\mathbf{G}(z)$ *can be irreducibly lifted from* $\mathbf{H}(z)$ *in infinitely many different ways.*

3.1 Group Lifting Structures

In light of the rich supply of elementary matrices, this plethora of irreducible lifting factorizations (almost all of which are useless for applications) results from our failure to specify precisely which liftings we regard as *useful*. The JPEG committee restricted the scope of [13, Annex G] to linear phase lifting factorizations of WS filter banks because these were considered to be the most useful liftings for conventional image coding, while [13, Annex H] was written to accommodate arbitrary lifted filter banks for niche applications. Taking a cue from the JPEG committee, we formalize a framework for specifying *restricted universes* of lifting factorizations. Group theory turns out to be a convenient tool for this task.

3.1.1 Lifting Matrix Groups

As mentioned above, upper-triangular (resp., lower-triangular) lifting matrices form multiplicative groups, $\mathcal{U}$ (resp., $\mathcal{L}$), as do lifting matrices whose lifting filters are restricted to additive groups of Laurent polynomials. This includes groups of filters whose symmetry and group delay are given, such as the groups $\mathcal{P}_0$ and $\mathcal{P}_1$ of HS lifting filters associated with Lemma 2. Define abelian group isomorphisms

$$\upsilon,\ \lambda : \mathbb{C}[z, z^{-1}] \to \mathcal{N}$$

that map a lifting filter $S(z) \in \mathbb{C}[z, z^{-1}]$ to lifting matrices,

$$\upsilon(S) \equiv \begin{bmatrix} 1 & S(z) \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \lambda(S) \equiv \begin{bmatrix} 1 & 0 \\ S(z) & 1 \end{bmatrix}. \tag{34}$$

Definition 6 ([1], Definition 4). Given two additive groups of Laurent polynomials, $\mathcal{P}_i < \mathbb{C}[z, z^{-1}]$, $i = 0, 1$, the groups $\mathcal{U} \equiv \upsilon(\mathcal{P}_0)$ and $\mathcal{L} \equiv \lambda(\mathcal{P}_1)$ are called the *lifting matrix groups* generated by $\mathcal{P}_0$ and $\mathcal{P}_1$.

3.1.2 Gain-Scaling Automorphisms

The unimodular gain-scaling matrices $\mathbf{D}_K \equiv \mathrm{diag}(1/K, K)$ also form an abelian group with the product $\mathbf{D}_K\mathbf{D}_J = \mathbf{D}_{KJ}$, which says that we have an isomorphism

$$\mathbf{D}\colon \mathbb{R}^* \equiv \mathbb{R}\backslash\{0\} \xrightarrow{\cong} \mathcal{D} < \mathcal{N}. \tag{35}$$

$\mathcal{D}$ acts on $\mathcal{N}$ via inner automorphisms,

$$\gamma_K\mathbf{A}(z) \equiv \mathbf{D}_K\,\mathbf{A}(z)\,\mathbf{D}_K^{-1}, \qquad \gamma_K\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & K^{-2}b \\ K^2c & d \end{bmatrix}. \tag{36}$$

This is equivalent to the intertwining relation

$$\mathbf{D}_K\,\mathbf{A}(z) = (\gamma_K\mathbf{A}(z))\,\mathbf{D}_K \tag{37}$$

and makes $\gamma\colon \mathbf{D}_K \mapsto \gamma_K$ a homomorphism of $\mathcal{D}$ onto a subgroup $\gamma(\mathcal{D}) < \mathrm{Aut}(\mathcal{N})$.

Definition 7 ([1], Definition 5). A group $\mathcal{G} < \mathcal{N}$ is $\mathcal{D}$-*invariant* if all of the inner automorphisms $\gamma_K \in \gamma(\mathcal{D})$ fix the group $\mathcal{G}$; i.e., $\gamma_K\mathcal{G} = \mathcal{G}$, so that $\gamma_K|_{\mathcal{G}} \in \mathrm{Aut}(\mathcal{G})$. This is equivalent to saying that $\mathcal{D}$ lies in the normalizer of $\mathcal{G}$ in $\mathcal{N}$:

$$\mathcal{D} \;<\; N_{\mathcal{N}}(\mathcal{G}) \;\equiv\; \left\{\mathbf{A} \in \mathcal{N} : \mathbf{A}\mathcal{G}\mathbf{A}^{-1} = \mathcal{G}\right\}.$$

For instance, when the lifting filter groups $\mathcal{P}_0$ and $\mathcal{P}_1$ are *vector spaces*, it follows easily from (36) that $\mathcal{U} \equiv \upsilon(\mathcal{P}_0)$ and $\mathcal{L} \equiv \lambda(\mathcal{P}_1)$ are $\mathcal{D}$-invariant matrix groups.

3.1.3 Definition of Group Lifting Structures

We now have the machinery needed to define a "universe" of lifting factorizations. In the following, $\mathfrak{B}$ denotes a set (not necessarily a group) of base filter banks from which other filter banks are lifted in partially factored lifting cascades (27).

Definition 8 ([1], Definitions 6 and 7). A *group lifting structure* is an ordered four-tuple,

$$\mathfrak{S} \equiv (\mathcal{D},\ \mathcal{U},\ \mathcal{L},\ \mathfrak{B}),$$

where $\mathcal{D}$ is a gain-scaling group, $\mathcal{U}$ and $\mathcal{L}$ are upper- and lower-triangular lifting matrix groups, and $\mathfrak{B} \subset \mathcal{N}$. The *lifting cascade group*, $\mathcal{C}$, generated by $\mathfrak{S}$ is the subgroup of $\mathcal{N}$ generated by $\mathcal{U}$ and $\mathcal{L}$:

$$\mathcal{C} \equiv \langle \mathcal{U} \cup \mathcal{L} \rangle = \{\mathbf{S}_1 \cdots \mathbf{S}_k : k \geq 1,\ \mathbf{S}_i \in \mathcal{U} \cup \mathcal{L}\}. \tag{38}$$

The *scaled lifting group*, $\mathcal{S}$, generated by $\mathfrak{S}$ is the subgroup generated by $\mathcal{D}$ and $\mathcal{C}$:

$$\mathcal{S} \equiv \langle \mathcal{D} \cup \mathcal{C} \rangle = \{ \mathbf{A}_1 \cdots \mathbf{A}_k : k \geq 1,\ \mathbf{A}_i \in \mathcal{D} \cup \mathcal{U} \cup \mathcal{L} \}\,. \tag{39}$$

We say $\mathfrak{S}$ is a $\mathcal{D}$-*invariant group lifting structure* if $\mathcal{U}$ and $\mathcal{L}$, and therefore $\mathcal{C}$, are $\mathcal{D}$-invariant groups.

Given a group lifting structure, the universe of all filter banks generated by $\mathfrak{S}$ is

$$\mathcal{DCB} \equiv \{ \mathbf{DCB} : \mathbf{D} \in \mathcal{D},\ \mathbf{C} \in \mathcal{C}\,,\ \mathbf{B} \in \mathfrak{B} \}\,.$$

The statement "**H** has a (group) lifting factorization in $\mathfrak{S}$" means $\mathbf{H} \in \mathcal{DCB}$. **H** has a lifting factorization in $\mathfrak{S}$ if and only if it has an *irreducible* factorization in $\mathfrak{S}$.

The group lifting structure that characterizes the universe of WS group lifting factorizations is defined as follows. The lifting matrix groups $\mathcal{U} \equiv \upsilon(\mathcal{P}_0)$ and $\mathcal{L} \equiv \lambda(\mathcal{P}_1)$ are determined by the groups $\mathcal{P}_0$ and $\mathcal{P}_1$ of HS lifting filters defined in Sect. 2.2. By Theorem 1 unimodular WS filter banks factor completely over $\mathcal{U}$ and $\mathcal{L}$, so we set $\mathfrak{B} \equiv \{\mathbf{I}\}$. Since $\mathcal{P}_0$ and $\mathcal{P}_1$ are vector spaces, setting $\mathcal{D} \equiv \mathbf{D}(\mathbb{R}^*)$ results in a $\mathcal{D}$-invariant group lifting structure, $\mathfrak{S}_{\mathcal{W}} \equiv (\mathcal{D}, \mathcal{U}, \mathcal{L}, \mathfrak{B})$. The conclusion of Theorem 1 can be stated succinctly in terms of $\mathcal{C}_{\mathcal{W}} \equiv \langle \mathcal{U} \cup \mathcal{L} \rangle$ as

$$\mathcal{W} = \mathcal{D}\mathcal{C}_{\mathcal{W}}\mathfrak{B} = \mathcal{D}\mathcal{C}_{\mathcal{W}}\,. \tag{40}$$

The group lifting structure for delay-minimized HS lifting factorizations is more complicated. The lifting matrix groups $\mathcal{U} \equiv \upsilon(\mathcal{P}_a)$ and $\mathcal{L} \equiv \lambda(\mathcal{P}_a)$ are determined by the group $\mathcal{P}_a$ of WA lifting filters defined in Sect. 2.3. Per Theorem 3, we define $\mathfrak{B}_{\mathfrak{H}}$ to be the set of all concentric equal-length HS filter banks. Defining $\mathcal{D} \equiv \mathbf{D}(\mathbb{R}^*)$ results in a $\mathcal{D}$-invariant group lifting structure, $\mathfrak{S}_{\mathfrak{H}} \equiv (\mathcal{D}, \mathcal{U}, \mathcal{L}, \mathfrak{B}_{\mathfrak{H}})$. With $\mathcal{C}_{\mathfrak{H}} \equiv \langle U \cup L \rangle$ the conclusion of Theorem 3 can be stated as

$$\mathfrak{H} = \mathcal{D}\mathcal{C}_{\mathfrak{H}}\mathfrak{B}_{\mathfrak{H}}\,. \tag{41}$$

Group lifting structures $\mathfrak{S}_{\mathcal{W}_r}$ and $\mathfrak{S}_{\mathfrak{H}_r}$ for *reversible* WS and HS filter banks are defined in [1, Sect. IV].

3.2 *Unique Irreducible Group Lifting Factorizations*

We need one more hypothesis in addition to irreducibility to infer uniqueness of group lifting factorizations within a given group lifting structure. The key is found in the fact that nonunique lifting factorizations can be rewritten as irreducible lifting factorizations of the identity, such as (33). Given a (nonconstant) lifting of the identity like [1, Eq. (21)], if some partial product $\mathbf{E}^{(n)}(z)$ of lifting steps (28) has positive polyphase order then the order of subsequent partial products must

eventually *decrease* because the final product, $\mathbf{I}$, has order zero. This suggests that lifting structures that only generate "order-increasing" cascades will generate *unique* factorizations, an idea that will be made rigorous in Theorem 4.

Definition 9 ([1], Definition 10). A lifting cascade (27) is *strictly polyphase order-increasing* (usually shortened to *order-increasing*) if the order (13) of each intermediate polyphase matrix (28) is strictly greater than that of its predecessor:

$$\text{order}\left(\mathbf{E}^{(n)}\right) > \text{order}\left(\mathbf{E}^{(n-1)}\right) \quad \text{for } 0 \le n < N.$$

A group lifting structure, $\mathfrak{S}$, is called order-increasing if every irreducible cascade in $\mathcal{CB}$ is order-increasing.

3.2.1 An Abstract Uniqueness Theorem

Theorem 4 ([1], Theorem 1). *Suppose that $\mathfrak{S}$ is a $\mathcal{D}$-invariant, order-increasing group lifting structure. Let $\mathbf{H}(z)$ be a transfer matrix generated by $\mathfrak{S}$, and suppose we are given two irreducible group lifting factorizations of $\mathbf{H}(z)$ in $\mathcal{DCB}$:*

$$\mathbf{H}(z) = \mathbf{D}_K\, \mathbf{S}_{N-1}(z) \cdots \mathbf{S}_0(z)\, \mathbf{B}(z) \tag{42}$$

$$= \mathbf{D}_{K'}\, \mathbf{S}'_{N'-1}(z) \cdots \mathbf{S}'_0(z)\, \mathbf{B}'(z)\ . \tag{43}$$

Then (42) and (43) satisfy the following three properties:

$$N' = N\,, \tag{44}$$

$$\mathbf{B}'(z) = \mathbf{D}_\alpha\, \mathbf{B}(z) \quad \textit{where } \alpha \equiv K/K', \tag{45}$$

$$\mathbf{S}'_i(z) = \gamma_\alpha \mathbf{S}_i(z) \quad \textit{for } i = 0, \ldots, N-1. \tag{46}$$

If, in addition, $\mathbf{B}(z)$ and $\mathbf{B}'(z)$ share a nonzero matrix entry at some point z_0, then the factorizations (42) and (43) are identical; *i.e., $K' = K$, $\mathbf{B}'(z) = \mathbf{B}(z)$, and*

$$\mathbf{S}'_i(z) = \mathbf{S}_i(z) \quad \textit{for } i = 0, \ldots, N-1. \tag{47}$$

It also follows that $K' = K$ if either of the scalar base filters, $B_0(z)$ or $B_1(z)$, shares a nonzero value with its primed counterpart; e.g., if the base filter banks have equal lowpass DC responses.

The relationship described by (44)–(46) leads to the following definition.

Definition 10 ([1], Definition 11). Two factorizations of $\mathbf{H}(z)$ that satisfy (44)–(46) are said to be *equivalent modulo rescaling*. If *all* irreducible group lifting factorizations of $\mathbf{H}(z)$ are equivalent modulo rescaling for *every* $\mathbf{H}(z)$ generated by $\mathfrak{S}$, we say that irreducible factorizations in $\mathfrak{S}$ are *unique modulo rescaling*.

3.2.2 Application to WS and HS Group Lifting Structures

Applying Theorem 4 is nontrivial, and verifying the order-increasing property is the hardest aspect of the whole theory. The key lemma for proving the order-increasing property for the WS and HS group lifting structures is the following result.

Lemma 4 ([2], Lemma 2). *Let $\mathfrak{S}$ be a group lifting structure satisfying the following two polyphase vector conditions.*

1. *For all $\mathbf{B}(z) \in \mathfrak{B}$, the polyphase support intervals (8) for the base polyphase filter vectors are equal:*

$$\text{supp\_int}(\boldsymbol{b}_0) = \text{supp\_int}(\boldsymbol{b}_1). \tag{48}$$

2. *For all irreducible lifting cascades in $\mathcal{C}\mathfrak{B}$, the polyphase support intervals (8) for the intermediate polyphase filter vectors satisfy the proper inclusions*

$$\text{supp\_int}\left(\boldsymbol{e}_{1-m_n}^{(n)}\right) \subsetneq \text{supp\_int}\left(\boldsymbol{e}_{m_n}^{(n)}\right) \quad \text{for } n \geq 0. \tag{49}$$

It then follows that $\mathfrak{S}$ is strictly polyphase order-increasing.

Hypothesis (48) is the correct answer to the ill-posed question, "What do all concentric equal-length HS base filter banks have in common with the lazy wavelet filter bank, **I**?" This was one of the last pieces of the uniqueness puzzle to be solved and unified the uniqueness proofs for the WS and HS cases.

Theorem 5 ([2], Theorem 1). *Let $\mathfrak{S}_{\mathcal{W}}$ and $\mathfrak{S}_{\mathcal{W}_r}$ be the group lifting structures defined in [1, Sect. IV-A]. Every filter bank in $\mathcal{W}$ has a unique irreducible lifting factorization in $\mathfrak{S}_{\mathcal{W}}$, and every filter bank in $\mathcal{W}_r$ has a unique irreducible lifting factorization in $\mathfrak{S}_{\mathcal{W}_r}$.*

Corollary 1 ([2], Corollary 1). *A delay-minimized unimodular WS filter bank can be specified in JPEG 2000 Part 2 Annex G syntax in one and only one way.*

The proof of Theorem 5 involves deriving the support-interval covering property (49) needed to invoke Lemma 4 and Theorem 4. The support-interval covering property results from the following tedious lemma based on the recursive formulation of lifting (28). The update characteristic of $\mathbf{S}_n(z)$ (Definition 3) is m_n, and the *support radius* of a filter is the radius of its support interval,

$$\text{supp\_rad}(f) \equiv \left\lfloor \frac{b-a+1}{2} \right\rfloor, \quad \text{where} \quad [a,b] = \text{supp\_int}(f). \tag{50}$$

Lemma 5 ([2], Lemma 5). *Let $\mathbf{S}_{N-1}(z)\cdots\mathbf{S}_0(z) \in \mathcal{C}_{\mathcal{W}}$ be an irreducible cascade with intermediate scalar filters $E_i^{(n)}(z), i = 0, 1$. Let $r_i^{(n)}$ be the support radius of $e_i^{(n)}$, and let $t^{(n)} \geq 1$ be the support radius of the HS lifting filter $S_n(z)$.*

Then $\text{supp\_int}\left(e_i^{(n)}\right)$ *is centered at* $-i$*,*

$$\text{supp\_int}\left(e_i^{(n)}\right) = \left[-r_i^{(n)} - i,\, r_i^{(n)} - i\right], \quad i = 0,\, 1,$$

where

$$r_{m_n}^{(n)} = r_{1-m_n}^{(n)} + 2t^{(n)} - 1 \quad \textit{for } n \geq 0, \tag{51}$$

$$r_{1-m_n}^{(n)} = r_{m_n}^{(n-1)} + 2t^{(n-1)} - 1 \quad \textit{for } n \geq 1, \tag{52}$$

with $r_{1-m_0}^{(0)} = r_{1-m_0}^{(-1)} = 0$.

There is a similar unique factorization result for unimodular HS filter banks.

Theorem 6 ([2], Theorem 2). *Let* $\mathfrak{S}_{\mathfrak{H}}$ *and* $\mathfrak{S}_{\mathfrak{H}_r}$ *be the group lifting structures defined in [1, Sect. IV-B]. Every filter bank in* $\mathfrak{H}$ *has an irreducible group lifting factorization in* $\mathfrak{S}_{\mathfrak{H}}$ *that is unique modulo rescaling. Every filter bank in* $\mathfrak{H}_r$ *has a unique irreducible group lifting factorization in* $\mathfrak{S}_{\mathfrak{H}_r}$.

4 Group-Theoretic Structure of Linear Phase Filter Banks

We can now characterize the group-theoretic structure of the groups generated by a $\mathcal{D}$-invariant, order-increasing group lifting structure. First we consider the lifting cascade group, $\mathcal{C}$, which only depends on $\mathcal{U}$ and $\mathcal{L}$, after which we consider the structure generated by scaling operations in the scaled lifting group, $\mathcal{S}$.

4.1 Free Product Structure of Lifting Cascade Groups

Recall the definition of free products in the category of groups.

Definition 11 ([11, 18]). Let $\{\mathcal{G}_i : i \in I\}$ be an indexed family of groups, and let $\mathcal{P}$ be a group with homomorphisms $\jmath_i : \mathcal{G}_i \to \mathcal{P}$. Then $\mathcal{P}$ is called a *free product of the groups* $\mathcal{G}_i$ if and only if, for every group $\mathcal{H}$ and family of homomorphisms $f_i : \mathcal{G}_i \to \mathcal{H}$, there exists a unique homomorphism $\phi : \mathcal{P} \to \mathcal{H}$ such that $\phi \circ \jmath_i = f_i$ for all $i \in I$. This is equivalent to saying that there exists a unique homomorphism ϕ such that the diagram in Fig. 5 commutes for all $i \in I$.

Defining free products via the universal mapping property in Fig. 5 means free products are *coproducts* in the category of groups and are therefore uniquely determined (up to isomorphism) by their generators $\mathcal{G}_i$ [11, Theorem I.7.5], [18, Theorem 11.50]. There is a constructive procedure (the "reduced word construction" [11, 18]) that generates a canonical realization of the free product of an arbitrary family of groups. Standard notation for free products is $\mathcal{P} = \mathcal{G}_1 * \mathcal{G}_2 * \cdots$.

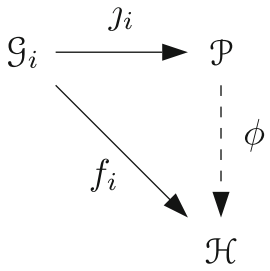


Fig. 5 Commutative diagram defining a free product of the groups $\mathcal{G}_i$

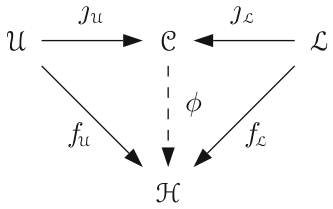


Fig. 6 Universal mapping property for the coproduct $\mathcal{C} \cong \mathcal{U} * \mathcal{L}$

The intuition behind Theorem 7 is the identification of irreducible group lifting factorizations over $\mathcal{U}$ and $\mathcal{L}$ with the group of reduced words over the alphabet $\mathcal{U} \cup \mathcal{L}$, which is the canonical realization of $\mathcal{U} * \mathcal{L}$. The reduced word construction of $\mathcal{U} * \mathcal{L}$ is a somewhat technical chore when done rigorously, and it would be a messy affair at best to write down and verify an isomorphism between the group of reduced words over $\mathcal{U} \cup \mathcal{L}$ and a lifting cascade group in one-to-one correspondence with a collection of irreducible group lifting factorizations. For this reason the proof presented in [3] avoids the details of the reduced word construction and instead uses uniqueness of irreducible group lifting factorizations to show that $\mathcal{C}$ satisfies the categorical definition of a coproduct.

4.1.1 Lifting Cascade Groups Are Free Products of $\mathcal{U}$ and $\mathcal{L}$

An easy lemma is needed to deal with group lifting structures whose irreducible group lifting factorizations are only unique modulo rescaling.

Lemma 6 ([3], Lemma 1). *If $(\mathcal{D}, \mathcal{U}, \mathcal{L}, \mathfrak{B})$ is a $\mathcal{D}$-invariant, order-increasing group lifting structure with lifting cascade group $\mathcal{C} \equiv \langle \mathcal{U} \cup \mathcal{L} \rangle$ then irreducible group lifting factorizations in $\mathcal{C}$ are unique, even if irreducible group lifting factorizations of filter banks in $\mathcal{DCB}$ are only unique modulo rescaling.*

Lemma 6 ensures that all $\mathcal{D}$-invariant, order-increasing group lifting structures satisfy the hypotheses of the following theorem, whose proof consists of showing that $\mathcal{C}$ satisfies the universal mapping property in Fig. 6.

Theorem 7 ([3], Theorem 1). *Let $\mathcal{U}$ and $\mathcal{L}$ be upper- and lower-triangular lifting matrix groups with lifting cascade group $\mathcal{C} \equiv \langle \mathcal{U} \cup \mathcal{L} \rangle$. If every element of $\mathcal{C}$ has a unique irreducible group lifting factorization over $\mathcal{U} \cup \mathcal{L}$ then $\mathcal{C}$ is isomorphic to the free product of $\mathcal{U}$ and $\mathcal{L}$:*

$$\mathcal{C} \cong \mathcal{U} * \mathcal{L} .$$

This free product structure, $\mathcal{C} \cong \mathcal{U} * \mathcal{L}$, is one of the conditions that are required for $\mathcal{C}$ to be a free group.

Theorem 8 ([3], Theorem 2). *Let $\mathcal{C} \equiv \langle \mathcal{U} \cup \mathcal{L} \rangle$ be a lifting cascade group over nontrivial lifting matrix groups $\mathcal{U}$ and $\mathcal{L}$. $\mathcal{C}$ is a free group (necessarily on two generators) if and only if $\mathcal{U}$ and $\mathcal{L}$ are infinite cyclic groups and $\mathcal{C} \cong \mathcal{U} * \mathcal{L}$.*

4.2 Semidirect Product Structure of Scaled Lifting Groups

Consider the interaction between the gain-scaling group $\mathcal{D}$ and the lifting cascade group $\mathcal{C}$ in a scaled lifting group, $\mathcal{S} \equiv \langle \mathcal{D} \cup \mathcal{C} \rangle$. As we have seen, $\mathcal{D}$ acts on $\mathcal{C}$ via inner automorphisms so it is not surprising that, under suitable hypotheses, $\mathcal{S}$ has the structure of a semidirect product, whose definition we now review.

Definition 12 ([11,14,18]). Let $\mathcal{G}$ be a (multiplicative) group with identity element $1_{\mathcal{G}}$ and subgroups $\mathcal{K}$ and $\mathcal{Q}$. $\mathcal{G}$ is an *(internal) semidirect product of $\mathcal{K}$ by $\mathcal{Q}$*, denoted $\mathcal{G} = \mathcal{Q} \ltimes \mathcal{K}$, if the following three axioms are satisfied:

$$\mathcal{G} = \langle \mathcal{K} \cup \mathcal{Q} \rangle \quad (\mathcal{K} \text{ and } \mathcal{Q} \text{ generate } \mathcal{G}), \tag{53}$$

$$\mathcal{K} \lhd \mathcal{G} \quad (\mathcal{K} \text{ is a normal subgroup of } \mathcal{G}), \tag{54}$$

$$\mathcal{K} \cap \mathcal{Q} = 1_{\mathcal{G}} \quad \text{(the trivial group)}. \tag{55}$$

If $\mathcal{G} = \mathcal{Q} \ltimes \mathcal{K}$ then $\langle \mathcal{K} \cup \mathcal{Q} \rangle = \mathcal{Q}\mathcal{K}$ and such product representations, $g = qk$ for $g \in \mathcal{G} = \mathcal{Q}\mathcal{K}$, are unique.

For groups $\mathcal{K}$ and $\mathcal{Q}$ that are not subgroups of a common parent, a similar construction called an *external semidirect product*, denoted $\mathcal{G} = \mathcal{Q} \ltimes_\theta \mathcal{K}$, can be performed whenever we have an automorphic group action $\theta: \mathcal{Q} \to \mathrm{Aut}(\mathcal{K})$.

4.2.1 Scaled Lifting Groups Are Semidirect Products of $\mathcal{C}$ by $\mathcal{D}$

Let $\mathfrak{S} = (\mathcal{D}, \mathcal{U}, \mathcal{L}, \mathfrak{B})$ be a group lifting structure with lifting cascade group $\mathcal{C}$ and scaled lifting group $\mathcal{S}$. The following theorem has the same hypotheses as those of Theorem 4, but rather than invoking the unique factorization theorem, the argument in [3] proves Theorem 9 directly from the hypotheses.

Theorem 9 ([3], Theorem 3). *If $\mathfrak{S}$ is a $\mathcal{D}$-invariant, order-increasing group lifting structure then $\mathcal{S}$ is the internal semidirect product of $\mathcal{C}$ by $\mathcal{D}$:*

$$\mathcal{S} = \mathcal{D} \ltimes \mathcal{C}.$$

This result can be combined with Theorem 7 to yield a complete group-theoretic description of the group of unimodular WS filter banks,

$$\mathcal{W} = \mathcal{S}_{\mathcal{W}} = \mathcal{D}\mathcal{C}_{\mathcal{W}}.$$

Corollary 2 ([3], Corollary 2). *Let $\mathfrak{S}_{\mathcal{W}} \equiv (\mathcal{D}, \mathcal{U}, \mathcal{L}, \mathbf{I})$ be the group lifting structure for the unimodular WS group, $\mathcal{W}$, defined in [1, Sect. IV]. The group-theoretic structure of $\mathcal{W}$ is*

$$\mathcal{W} \cong \mathcal{D} \ltimes_\theta (\mathcal{U} * \mathcal{L}).$$

A similar characterization is possible for HS filter banks. While $\mathfrak{H}$ is not a group, the product representation

$$\mathfrak{H} = \mathcal{D}\mathcal{C}_{\mathfrak{H}}\mathfrak{B}_{\mathfrak{H}} = \mathcal{S}_{\mathfrak{H}}\mathfrak{B}_{\mathfrak{H}}, \tag{56}$$

$$\mathfrak{B}_{\mathfrak{H}} \equiv \{\mathbf{B} \in \mathfrak{H} : \operatorname{order}(B_0) = \operatorname{order}(B_1)\} \tag{57}$$

exhibits $\mathfrak{H}$ as a collection of *right cosets*, $\mathcal{S}_{\mathfrak{H}}\mathbf{B}$, of $\mathcal{S}_{\mathfrak{H}}$ by elements of $\mathfrak{B}_{\mathfrak{H}}$. These cosets do not *partition* $\mathfrak{H}$, however, since they are not disjoint: $\mathbf{B}' \equiv \mathbf{D}_a\mathbf{B} \in \mathfrak{B}_{\mathfrak{H}}$ implies $\mathcal{S}_{\mathfrak{H}}\mathbf{B} = \mathcal{S}_{\mathfrak{H}}\mathbf{B}'$. To obtain a nonredundant partition of $\mathfrak{H}$ into cosets, we can either eliminate scaling matrices (i.e., form cosets of $\mathcal{C}_{\mathfrak{H}}$ rather than of $\mathcal{S}_{\mathfrak{H}}$) or else normalize the elements of $\mathfrak{B}_{\mathfrak{H}}$.

Corollary 3 ([3], Corollary 3). *Let $\mathfrak{S}_{\mathfrak{H}} \equiv (\mathcal{D}, \mathcal{U}, \mathcal{L}, \mathfrak{B}_{\mathfrak{H}})$ be the group lifting structure for the unimodular HS class, $\mathfrak{H}$, defined in [1, Sect. IV]. The group-theoretic structure of $\mathcal{S}_{\mathfrak{H}}$ is*

$$\mathcal{S}_{\mathfrak{H}} \cong \mathcal{D} \ltimes_\theta (\mathcal{U} * \mathcal{L}),$$

and $\mathfrak{H}$ can be partitioned into disjoint right cosets (but not left cosets) of either $\mathcal{C}_{\mathfrak{H}}$ or $\mathcal{S}_{\mathfrak{H}}$:

$$\mathfrak{H} = \bigcup \{\mathcal{C}_{\mathfrak{H}}\mathbf{B} : \mathbf{B} \in \mathfrak{B}_{\mathfrak{H}}\} \tag{58}$$

$$= \bigcup \{\mathcal{S}_{\mathfrak{H}}\mathbf{B} : \mathbf{B} \in \mathfrak{B}'_{\mathfrak{H}}\}, \tag{59}$$

where $\mathfrak{B}'_{\mathfrak{H}}$ is given by, e.g.,

$$\mathfrak{B}'_{\mathfrak{H}} \equiv \{\mathbf{B} \in \mathfrak{B}_{\mathfrak{H}} : B_0(1) = 1\}. \tag{60}$$

Scaled lifting groups with the structure $\mathcal{S} \cong \mathcal{D} \ltimes_\theta (\mathcal{U} * \mathcal{L})$ have formal similarities [3, Sect. IV] to other examples in the mathematical literature of continuous groups

with dilations, such as *homogeneous groups* [10, 20]. Unlike homogeneous groups, however, scaled lifting groups are neither nilpotent nor finite dimensional, so scaled lifting groups at present appear to be a new addition to the realm of continuous groups with scaling automorphisms.

5 Conclusions

We have surveyed recent results characterizing the group-theoretic structure of the two principal classes of two-channel linear phase perfect reconstruction unimodular filter banks, the whole-sample symmetric and the half-sample symmetric classes. WS filter banks presented in the polyphase-with-advance representation naturally form a multiplicative subgroup, $\mathcal{W}$, of the group of all unimodular matrix Laurent polynomials. Although the class $\mathfrak{H}$ of unimodular HS filter banks does not form a group, lifting factorization theory shows that HS filter banks form cosets of a particular group generated by unimodular diagonal gain-scaling matrices and lifting matrices with whole-sample antisymmetric lifting filters. An algebraic framework known as a group lifting structure has been introduced for formalizing the group-theoretic structure of lifting factorizations, and it has been shown that the group lifting structures for WS (respectively, HS) filter banks satisfy a nontrivial polyphase order-increasing property that implies uniqueness of irreducible group lifting factorizations.

These unique factorization results have in turn been used to characterize the structure (up to isomorphism) of the lifting cascade group and the scaled lifting group associated with each of these classes of linear phase filter banks. Specifically, in both cases the lifting cascade group generated by the linear phase lifting matrices is the free product of the upper- and lower-triangular lifting matrix groups, $\mathcal{C} \cong \mathcal{U} * \mathcal{L}$. Also in both cases, the scaled lifting group generated by the lifting cascade group and the diagonal gain-scaling matrix group has the structure of a semidirect product, $\mathcal{S} = \mathcal{D}\mathcal{C} \cong \mathcal{D} \ltimes_\theta (\mathcal{U} * \mathcal{L})$. In the case of WS filter banks this directly furnishes the structure of the unimodular WS group, $\mathcal{W}$, since $\mathcal{W} = \mathcal{S}_{\mathcal{W}}$. In the case of HS filter banks, $\mathfrak{H}$ is partitioned by the family of all right cosets of $\mathcal{C}_{\mathfrak{H}}$ by concentric equal-length base HS filter banks. Alternatively, $\mathfrak{H}$ is also partitioned by the family of all right cosets of $\mathcal{S}_{\mathfrak{H}}$ by concentric equal-length base HS filter banks with unit lowpass DC response.

Acknowledgements The original research papers [1–4] described in this chapter were supported by the Los Alamos Laboratory-Directed Research & Development Program. Preparation of this chapter was supported by the DOE Office of Science and Kristi D. and Reilly R. Brislawn. The author also thanks the producers of the Ipe drawing editor (http://ipe7.sourceforge.net) and the TEXLive/MacTEX distribution (http://www.tug.org/mactex).

References

1. Brislawn, C.M.: Group lifting structures for multirate filter banks I: Uniqueness of lifting factorizations. IEEE Trans. Signal Process. **58**(4), 2068–2077 (2010). URL http://dx.doi.org/10.1109/TSP.2009.2039816
2. Brislawn, C.M.: Group lifting structures for multirate filter banks II: Linear phase filter banks. IEEE Trans. Signal Process. **58**(4), 2078–2087 (2010). DOI 10.1109/TSP.2009.2039818. URL http://dx.doi.org/10.1109/TSP.2009.2039818
3. Brislawn, C.M.: Group-theoretic structure of linear phase multirate filter banks. Tech. Rep. LA-UR-12-20858, Los Alamos National Laboratory (2012). Submitted for publication. URL http://viz.lanl.gov/paper.html
4. Brislawn, C.M., Wohlberg, B.: The polyphase-with-advance representation and linear phase lifting factorizations. IEEE Trans. Signal Process. **54**(6), 2022–2034 (2006). DOI 10.1109/TSP.2006.872582. URL http://dx.doi.org/10.1109/TSP.2006.872582
5. Bruekers, F.A.M.L., van den Enden, A.W.M.: New networks for perfect inversion and perfect reconstruction. IEEE J. Selected Areas Commun. **10**(1), 129–137 (1992)
6. Calderbank, A.R., Daubechies, I., Sweldens, W., Yeo, B.L.: Wavelet transforms that map integers to integers. Appl. Comput. Harmon. Anal. **5**(3), 332–369 (1998). DOI 10.1006/acha.1997.0238. URL http://www.sciencedirect.com/science/article/B6WB3-45KKTPF-H/2/489d06bc0099f6a398eb534b2a8a8481
7. Crochiere, R.E., Rabiner, L.R.: Multirate Digital Signal Processing. Prentice Hall, Englewood Cliffs (1983)
8. Daubechies, I.C.: Ten Lectures on Wavelets. No. 61 in CBMS-NSF Regional Conf. Series in Appl. Math., (Univ. Mass.—Lowell, June 1990). Soc. Indust. Appl. Math., Philadelphia (1992)
9. Daubechies, I.C., Sweldens, W.: Factoring wavelet transforms into lifting steps. J. Fourier Anal. Appl. **4**(3), 245–267 (1998)
10. Folland, G.B., Stein, E.M.: Hardy Spaces on Homogeneous Groups. No. 28 In: Princeton Mathematical Notes. Princeton Univ. Press, Princeton (1982)
11. Hungerford, T.W.: Algebra. Springer, New York (1974)
12. Information technology—JPEG 2000 image coding system, Part 1, ISO/IEC Int'l. Standard 15444-1, ITU-T Rec. T.800. Int'l. Org. Standardization (2000)
13. Information technology—JPEG 2000 image coding system, Part 2: Extensions, ISO/IEC Int'l. Standard 15444-2, ITU-T Rec. T.801. Int'l. Org. Standardization (2004)
14. MacLane, S., Birkhoff, G.: Algebra. Macmillan, New York (1967)
15. Mallat, S.: A Wavelet Tour of Signal Processing, 2nd edn. Academic Press, San Diego (1999)
16. Nguyen, T.Q., Vaidyanathan, P.P.: Two-channel perfect-reconstruction FIR QMF structures which yield linear-phase analysis and synthesis filters. IEEE Transactions on Acoustics, Speech and Signal Processing **37**(5), 676–690 (1989)
17. Oppenheim, A.V., Schafer, R.W., Buck, J.R.: Discrete-Time Signal Processing, 2nd edn. Prentice Hall, Upper Saddle River (1998)
18. Rotman, J.J.: An Introduction to the Theory of Groups, 4th edn. Springer, New York (1995)
19. Said, A., Pearlman, W.A.: Reversible image compression via multiresolution representation and predictive coding. In: Visual Commun. & Image Proc., Proc. SPIE, vol. 2094, pp. 664–674. SPIE, Cambridge (1993)
20. Stein, E.M.: Harmonic Analysis. Princeton Univ. Press, Princeton (1993)
21. Strang, G., Nguyen, T.: Wavelets and Filter Banks. Wellesley-Cambridge, Wellesley (1996)
22. Sweldens, W.: The lifting scheme: a custom-design construction of biorthogonal wavelets. Appl. Comput. Harmonic Anal. **3**(2), 186–200 (1996)
23. Sweldens, W.: The lifting scheme: a construction of second generation wavelets. SIAM J. Math. Anal. **29**(2), 511–546 (1998)

24. Vaidyanathan, P.P.: Multirate Systems and Filter Banks. Prentice Hall, Englewood Cliffs (1993)
25. Vetterli, M., Kovačević, J.: Wavelets and Subband Coding. Prentice Hall, Englewood Cliffs (1995)
26. Zandi, A., Allen, J.D., Schwartz, E.L., Boliek, M.: Compression with reversible embedded wavelets. In: Proc. Data Compress. Conf., pp. 212–221. IEEE Computer Soc., Snowbird (1995)

Parametric Optimization of Biorthogonal Wavelets and Filterbanks via Pseudoframes for Subspaces

Shidong Li and Michael Hoffman

Abstract We present parametric optimizations of biorthogonal wavelets and associated filter banks using *pseudoframes for subspaces* (PFFS). PFFS extends the theory of frames in that pseudoframe sequences need not reside within the subspace of interest. In particular, when PFFS is applied to biorthogonal wavelets, the underlying flexibility presents opportunities to incorporate optimality, regularity, as well as perfect reconstruction into one parametric design approach. This approach reduces certain filter optimization problems to optimization over a free parameter. While past constructions can be reproduced, results with additional optimality are also obtained and presented here with numerical examples. Tables of filter coefficients along with graphs are provided.

Keywords Pseudoframes • Frames • Biorthogonal wavelets • Filter banks • Compression • Filter design

1 Introduction: Pseudoframes for Subspaces and Biorthogonal Wavelets

Frames and frame variations have generated vast interest and applications, see, e.g., [1–3, 6–8, 12–17, 19, 20, 23, 24] with pseudoframes first appearing in [25, 26].

S. Li (✉)
Department of Mathematics, San Francisco State University, San Francisco, CA, USA
e-mail: shidong@sfsu.edu

M. Hoffman
Department of Mathematics, Cañada College, Redwood City, CA, USA
e-mail: hoffmanm@smccd.edu

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_7,

Pseudoframes for subspaces (PFFS) is an extension of frames. It is a notion of frame-like expansions for a subspace $\mathscr{X}$ of a separable Hilbert space [27].

Let $\mathscr{X}$ be a closed subspace of a separable Hilbert space $\mathscr{H}$. Let $\{x_n\} \subseteq \mathscr{H}$ be a Bessel sequence w.r.t. $\mathscr{X}$, and let $\{x_n^*\}$ be a Bessel sequence in $\mathscr{H}$. We say $\{x_n\}$ is a $\mathscr{X}$ PFFS w.r.t. $\{x_n^*\}$ if

$$\forall f \in \mathscr{X}, \quad f = \sum_n \langle f, x_n \rangle x_n^*. \tag{1}$$

The important distinction between PFFS and frames is that none of the sequences $\{x_n\}$ and $\{x_n^*\}$ are necessarily required to be in $\mathscr{X}$. Consequently, $\{x_n\}$ and $\{x_n^*\}$ are not generally in the same subspace either. The resulting flexibility is the key point in PFFS.

The purpose of this chapter is to elaborate on the construction of new biorthogonal wavelets and filter banks utilizing PFFS as a method to obtain certain optimal design criteria.

To this end we will restate the most fundamental characterization of PFFS [27]. Let $\{x_n\} \subseteq \mathscr{H}$ and $\{x_n^*\} \subseteq \mathscr{H}$. Assume that $\{x_n\}$ is a Bessel sequence with respect to (w.r.t.) the subspace $\mathscr{X}$. Assume also that $\{x_n^*\}$ is a Bessel sequence in $\mathscr{H}$. Define $U : \mathscr{X} \to l^2(\mathbf{Z})$ by

$$\forall f \in \mathscr{X}, \quad Uf = \{\langle f, x_n \rangle\}, \tag{2}$$

and define $V : l^2(\mathbf{Z}) \to \mathscr{H}$ such that

$$\forall c \equiv \{c(n)\} \in l^2(\mathbf{Z}), \quad Vc = \sum_n c(n) x_n^*. \tag{3}$$

Then the following characterization of PFFS holds.

Theorem 1 ([27]). *Let $\{x_n\}$ and $\{x_n^*\}$ be two sequences in $\mathscr{H}$ (not necessarily in $\mathscr{X}$). Assume that $\{x_n\}$ is a Bessel sequence w.r.t. the subspace $\mathscr{X}$, and $\{x_n^*\}$ is a Bessel sequence in $\mathscr{H}$. Let U be defined by Eq. (2), and V be defined by Eq. (3). Suppose that $\mathscr{P}$ is any projection from $\mathscr{H}$ onto $\mathscr{X}$. Then $\{x_n\}$ is a pseudoframe for $\mathscr{X}$ w.r.t. $\{x_n^*\}$ if and only if*

$$VU\mathscr{P} = \mathscr{P}. \tag{4}$$

Therefore, the constructions of PFFS all start from Eq. (4) which basically requires the pseudoframes to preserve projections onto the subspace in question. The construction of PFFS has typically two (nonsymmetrical) directions. One corresponds to finding the "left inverse" V from a given U; the other relates to finding the U from a given V, all according to Eq. (4). We refer to [27] for details of the constructions.

Let $\{x_n^*\}$ be a Bessel sequence in $\mathscr{H}$ such that $\overline{sp}\{x_n^*\} \supseteq \mathscr{X}$. All PFFS-dual sequences $\{x_n\}$ can be constructed as follows. Let $\{x_n^\dagger\} \subseteq \mathscr{H}$ be such that $x_n^\dagger = (V^*)^\dagger e_n$, where $(V^*)^\dagger$ denotes the pseudoinverse of (V^*), and

$$\forall c \in l^2, \quad (V^*)^\dagger c = \sum_n c(n) x_n^\dagger, \tag{5}$$

and let $\{y_n\}$ be an arbitrary Bessel sequence in $\mathscr{H}$. We have then the following:

Corollary 1 ([27]). *Let $\{x_n^*\}$ be a Bessel sequence in $\mathscr{H}$ such that $\overline{sp}\{x*_n\} \supseteq \mathscr{X}$. Assume further that $R(V)$ is closed. Let $\{x_n^\dagger\}$ be defined in Eq. (5) and $\{y_n\}$ be a Bessel sequences in $\mathscr{H}$. All dual PFFS sequences $\{x_n\}$ for $\mathscr{X}$ w.r.t. $\{x_n^*\}$ are given by*

$$x_n = \mathscr{P}^* x_n^\dagger + y_n - \sum_m \langle x_n^\dagger, x_m^* \rangle \mathscr{P}^* y_m, \quad \forall n \in \mathbf{Z}, \tag{6}$$

where $\mathscr{P}^*$ is the adjoint operator of the projection $\mathscr{P}$.

Applying this construction to a shift-invariant subspace $\mathscr{X}$ such that $\overline{sp}\{\tau_n \phi\} \supseteq \mathscr{X}$ and assuming $\{\tau_n \phi\} \in L^2(\mathbf{R})$ is a Bessel sequence, we have seen in [27] that PFFS allows for the characterization of the entire class of duals of translates as follows:

$$\tilde{\phi}_n = \tau_n \tilde{\phi} \quad \forall n \in \mathbf{Z},$$

where

$$\tilde{\phi} \equiv P\phi^\dagger + y - P \sum_m \langle \phi^\dagger, \tau_m \phi \rangle \tau_m y \quad \text{in } L^2(\mathbf{R}), \tag{7}$$

and $\phi^\dagger \equiv V^{*\dagger} e_n$ is the dual corresponding to the pseudoinverse of V in $\overline{sp}\{x_n\}$, and $y \in L^2(\mathbf{R})$ is such that $\{\tau_n y\}$ is a Bessel sequence. Here, we have chosen P as an orthogonal projection.

PFFS Applied to Biorthogonal Wavelets

In a special case, let us assume that $\{\tau_n \phi\}$ is an exact frame of $\overline{sp}\{\tau_n \phi\} \equiv \mathscr{X}$, and translate the result of *Corollary 1* into this special setting. With the given assumption that $\{\tau_n \phi^\dagger\}$ is the unique biorthogonal dual frame to $\{\tau_n \phi\}$, $\langle \tau_n \phi^\dagger, \tau_m \phi \rangle = \delta_{nm}$. Therefore, we may arrive at the following expected result:

$$\tau_n \tilde{\phi} = P\tau_n \phi^\dagger + \tau_n y - P\tau_n y = \tau_n \phi^\dagger + (I - P)\tau_n y. \tag{8}$$

As long as $y \notin \overline{sp}\{\tau_n \phi\} = \mathscr{X}$, Eq. (8) would yield nonunique biorthogonal dual sequences $\{\tau_n \tilde{\phi}\}$.

We shall show that biorthogonal wavelet construction via PFFS opens up opportunities for design optimization without disrupting any important features of the original pair such as symmetry, compact support, improved regularity, or vanishing moments. Here, we will show that maximum attenuation and desired filter response can be incorporated with the design of FIR bi-filter banks while maintaining the vanishing moments, symmetry, etc.

2 Construction of Biorthogonal Wavelets and Filters via PFFS

Assume that $\phi \in L^2(\mathbf{R})$ and that $\{\tau_n\phi\}$ forms a biorthogonal basis of $V_0 \equiv \overline{sp}\{\tau_n\phi\}$. Assume also that $\{\phi, V_j\}$ generates a (biorthogonal) MRA of $L^2(\mathbf{R})$. As we analyzed, all biorthogonal PFFS-dual scaling functions $\{\tilde{\phi}_n = \tau_n\tilde{\phi}\}$ are given by

$$\tilde{\phi} = \phi^0 + \triangle\phi, \tag{9}$$

where $\phi^0 \in V_0$ is the standard dual function of ϕ and $\triangle\phi \in V_0^{\perp}$. In general, it is evident that if ϕ^* is any biorthogonal PFFS-dual function of ϕ, then so is $\tilde{\phi} = \phi^* + \triangle\phi$ for any $\triangle\phi \in V_0^{\perp}$. With a slight abuse of notation, we shall be considering equation (9) with ϕ^0 being any biorthogonal dual.

Assume that we are only interested in sufficiently regular and refinable $\tilde{\phi}$ such that $\tilde{\phi} = \sum_n \tilde{h}_n\tilde{\phi}_{1n}$. Considerable studies on the conditions for sequences such as $\tilde{h}$ can be found in [10]. Then the following relationship holds:

$$\tilde{h}_n = \langle\tilde{\phi}, \phi_{1n}\rangle = \langle\phi^0 + \triangle\phi, \phi_{1n}\rangle \equiv h_n^0 + \triangle h_n,$$

where we have assumed $\triangle\phi = \triangle\phi_0 + \triangle\phi_1 + \cdots$ with $\triangle\phi_j \in W_j \equiv V_{j+1} \setminus V_j$, and

$$\triangle h_n \equiv \langle\triangle\phi_0, \phi_{1n}\rangle.$$

What Does It Mean to Have $\{\triangle h_n\} \sim \triangle\phi_0 \in W_0$?

In the context of biorthogonal wavelets and multiresolution analysis, the add-on filter sequence component $\{\triangle h_n\}$ is solely relevant to information in the subspace W_0. In the case of B-spline wavelets, since there is no compactly supported biorthogonal wavelets and the corresponding linear phase FIR filters in the conventional biorthogonal sense within V_0 [9], we have thus observed that the regularity, vanishing moments, and compact support properties are the consequences of "add-on" components from information in the complement $W_0 = V_1 \setminus V_0$.

From the sub-band processing point of view, since these nice properties are demanded in practical applications [4, 5, 11, 18, 22, 28–31], we now understand the need to bring some information in the high-pass band (W_0) back to the low-pass band to offset some of the drawbacks in the conventional local structure.

All of the above have to be done in such a way that the perfect reconstruction principle is not violated. Recall that a set of four filter sequences $\{h_n\}$, $\{g_n\}$, $\{\tilde{h}_n\}$, and $\{\tilde{g}_n\}$ are said to form a biorthogonal sub-band system (or perfect reconstruction filter bank (PRFB)) if

$$\sum_n \left(h_{2n-k}\tilde{h}_{2n-l} + g_{2n-k}\tilde{g}_{2n-l} \right) = \delta_{kl}, \tag{10}$$

where $\{\tilde{h}_n\}$ is often termed the *dual* filter sequence to $\{h_n\}$ and $\{\tilde{g}_n\}$ the dual filter to $\{g_n\}$.

The following are the basic parametric construction of FIR biorthogonal filters from the "add-on" component point of view.

Theorem 2. *Let $\{h_n\}$ and $\{g_n\}$ be a set of filter bank filters, and let $\{h_n^0\}$ and $\{g_n^0\}$ be the corresponding biorthogonal dual filters satisfying (10). Let $\triangle H(\gamma)$ be the Fourier series of $\{\triangle h_n\}$, and $H(\gamma)$ be the Fourier series of $\{h_n\}$. Then $\{\tilde{h}_n = h_n^0 + \triangle h_n\}$ is a biorthogonal PFFS-dual filter if and only if*

$$\triangle H(\gamma) = \overline{H(\gamma + \frac{1}{2})} \cdot \hat{q}(\gamma), \tag{11}$$

where $\hat{q}(\gamma)$ is trigonometric polynomial satisfying

$$\overline{H(\gamma)} \overline{H\left(\gamma + \frac{1}{2}\right)} \left(\hat{q}(\gamma) + \hat{q}(\gamma + \frac{1}{2}) \right) = 0. \tag{12}$$

Moreover, the other two corresponding biorthogonal filters g and $\tilde{g}$ are given by

$$G(\gamma) = \mathrm{e}^{-2\pi i\gamma} \left(\overline{H^0(\gamma + \frac{1}{2})} + \overline{H(\gamma)\hat{q}(\gamma + \frac{1}{2})} \right), \tag{13}$$

$$\tilde{G}(\gamma) = \mathrm{e}^{-2\pi i\gamma} \overline{H\left(\gamma + \frac{1}{2}\right)}. \tag{14}$$

Proof. Following a similar proof in [10], the Fourier transform of Eq. (10) shows

$$H(\gamma)\overline{\tilde{H}(\gamma)} + H\left(\gamma + \frac{1}{2}\right) \overline{\tilde{H}\left(\gamma + \frac{1}{2}\right)} = 2. \tag{15}$$

Since $\tilde{H}(\gamma) = H^0(\gamma) + \triangle H(\gamma)$, Eq. (15) implies that

$$H(\gamma)\overline{\triangle H(\gamma)} + H\left(\gamma + \frac{1}{2}\right) \overline{\triangle H\left(\gamma + \frac{1}{2}\right)} = 0. \tag{16}$$

This in turn implies that

$$\triangle H(\gamma) = \overline{H\left(\gamma + \frac{1}{2}\right)} \cdot \hat{q}(\gamma)$$

for some 1-periodic trig polynomial $\hat{q}$ such that

$$\overline{H(\gamma)}\overline{H\left(\gamma+\frac{1}{2}\right)}\left(\hat{q}(\gamma)+\hat{q}\left(\gamma+\frac{1}{2}\right)\right)=0, \quad a.e.$$

This finishes the proof of the first half of the assertion. The proof of the filter relationships (13) and (14) for the corresponding high-pass filters G and $\tilde{G}$ is similar. □

We comment that, with the filter relationships as in Eqs. (13) and (14), the conditions we have just derived are essentially to require that

$$\sum_n h_n\overline{\tilde{h}_{n-2k}}=\delta_{k0}. \quad \forall k \tag{17}$$

Bring in the fact that $\tilde{h}_n = h_n^0 + \Delta h_n$ and that $\{h_n^0\}$ is biorthogonal to $\{h_n\}$, we essentially have

$$\sum_n h_n\overline{\Delta h_{n-2k}}=0, \quad \forall k, \tag{18}$$

with $\{\Delta h_n\} \sim \Delta\phi \in W_0 \subseteq V_0^{\perp}$.

We demonstrate how to construct new linear phase and symmetric FIR biorthogonal filters that maintain a given number of vanishing moments etc. in their wavelets while introducing the flexibility of other optimization opportunities.

2.1 Basic Relationships for Symmetric Cases

In the following presentations, we recall that the number of vanishing moments of a biorthogonal wavelets equals to the number of zeros of the corresponding low-pass filters at $\gamma = \frac{1}{2}$.

Theorem 3. *Let H be a symmetric biorthogonal FIR filter such that $H(\gamma) \neq 0$ for all γ except for $\gamma = \frac{1}{2}$ and perhaps a set of measure zero. Assume that a dual filter H^0 is symmetric with $2l$ zeros at $\gamma = \frac{1}{2}$. Let $\hat{q}(\gamma)$ be the trig polynomial in Theorem 2 satisfying $\hat{q}(0) = \hat{q}\left(\frac{1}{2}\right) = 0$. Then a set of symmetric $\hat{q}$ function with $2l$ zeros at $\gamma = \frac{1}{2}$ is given by*

$$\hat{q}(\gamma) = (1-\cos 4\pi\gamma)^l\hat{q}_1(\gamma)\cos 2\pi N\gamma, \quad N = odd, \tag{19}$$

where $\hat{q}_1(\gamma) = \hat{p}(\cos 4\pi\gamma)$ is a trig polynomial s.t. $\hat{q}_1(\frac{1}{2}) \neq 0$. In particular, let $\hat{q}_1 = 2$, the corresponding sequence $\{q_n\}$ of $\hat{q}(\gamma)$ is given by

$$q_n = \tau_{N-2l}b_n + \tau_{-N-2l}b_n,$$

where

$$b_n = \begin{cases} (-1)^m \binom{2l}{m} \Big/ (-2)^l, & n = 2m \\ 0, & n = 2m+1. \end{cases} \tag{20}$$

For such choices of $\hat{q}$, new biorthogonal PFFS-duals $\tilde{H}$ remain symmetric, and $\tilde{H}$ has at least the same number of zeros at $\frac{1}{2}$ as that of H^0. Here $\tilde{H} = H^0 + \Delta H$ and, with $N = 1$,

$$\Delta H(\gamma) = \overline{H(\gamma + \frac{1}{2})}\hat{q} = 2\overline{H(\gamma + \frac{1}{2})}(1 - \cos 4\pi\gamma)^l \cos 2\pi\gamma. \tag{21}$$

Notice that if $H = \chi_{[-\frac{1}{2},\frac{1}{2})}$, the choice of $\hat{q}$ would be free for any trigonometric polynomial. However, this is not of interest here because we require that H corresponds to FIR filters.

Proof. By the assumption of the theorem, Eq. (12) holds if and only if

$$\hat{q}(\gamma) = -\hat{q}\left(\gamma + \frac{1}{2}\right), \tag{22}$$

implying that $|\hat{q}|$ is $\frac{1}{2}$-periodic. Viewing the symmetry requirement and the number of zeros needed at $\frac{1}{2}$, we see that $\hat{q}$ can be a trig polynomial of $\cos(4\pi\gamma)$ modulated by a factor of $\cos 2\pi N\gamma$ for some odd integer N. Hence,

$$\hat{q}(\gamma) = (1 - \cos 4\pi\gamma)^l \hat{q}_1(\cos 4\pi\gamma) \cos 2\pi N\gamma,$$

where $\hat{q}_1$ should have no zeros at $\frac{1}{2}$. A simple trig identity simplification will show that such a $\hat{q}$ will indeed have at least $2l$ zeros at $\frac{1}{2}$. Let us find the filter sequence associated with $\hat{q}$. For $\hat{q}_1 = 2$,

$$\begin{aligned} \hat{q}(\gamma) &= 2^l \sin^{2l} 2\pi\gamma \cdot \hat{q}_1 \cdot \cos 2\pi N\gamma \\ &= (-1)^l 2^{-l} \left(\mathrm{e}^{-2\pi i\gamma} - \mathrm{e}^{2\pi i\gamma}\right)^{2l} 2\cos 2\pi N\gamma \\ &= (-1)^l 2^{-l} (2\cos 2\pi N\gamma)\, \mathrm{e}^{2\pi i(2l)\gamma} \sum_{k=0}^{2l} (-1)^k \binom{2l}{k} e^{-2\pi i(2k)\gamma} \\ &= 2\cos \pi N\gamma \mathrm{e}^{2\pi i(2l)\gamma} \sum_{n=0}^{4l+1} b_n \mathrm{e}^{-2\pi i n\gamma}, \end{aligned}$$

where

$$b_n \equiv \begin{cases} (-1)^m \binom{2l}{m} \Big/ (-2)^l, & n = 2m, \\ 0, & n = 2m+1. \end{cases}$$

Therefore,

$$\hat{q}(\gamma) = \left(\mathrm{e}^{-2\pi i N\gamma} + \mathrm{e}^{2\pi i N g a}\right) \mathrm{e}^{2\pi i (2l)\gamma} \sum_{n=0}^{4l+1} b_n \mathrm{e}^{-2\pi i n\gamma}$$

$$= \sum_{n=0}^{4l+1} \left(\tau_{N-2l} b_n + \tau_{-N-2l} b_n\right) \mathrm{e}^{-2\pi i n\gamma}.$$

Here $\tau_k b_n \equiv b_{n-k}$. We have therefore proved the assertion.

Letting $\hat{q}_1$ Be a Scalar Parameter λ

In *Theorem 3* the choice was made to let $\hat{q}_1 = 2$. This is valid since the only condition placed on the trig polynomial $\hat{q}_1$ is that this function must have no zeros at $\frac{1}{2}$. In the implementations discussed below and in *Proposition 1*, we have instead set $\hat{q}_1 = 2\lambda$. Here the scalar parameter λ is very useful for optimization purposes (without increasing the filter length than that of the choice of $\hat{q}_1 = 2$). While $\hat{q}_1$ could be *a*ny polynomial in $cos 4\pi\gamma$ (see discussions at the end of this article), the choice of the scalar parameterization λ minimizes the length of PFFS bi-filters.

From this point and on, the λ parameter would always be the consequence of such choices, and eventually become the optimization parameter in various optimal design procedures.

Combining the result of *Theorem 3* with Eq. (11), we have obtained the following parametric PFFS-dual bi-filters equation.

Proposition 1. *Let $\{h_n\}$ be a pair of symmetric biorthogonal filters with $2l$ zeros at $\gamma = \frac{1}{2}$. Let $\{\tilde{h}_n\}$ be any symmetric bi-dual filter with $2l$ zeros at $\gamma = \frac{1}{2}$, then, with $\{b_n\}$ being given by Eq. (20),*

$$h_n' = \tilde{h}_n + \lambda \left(\sum_k (-1)^k \overline{h_k} \left(\tau_{1-2l} b_{n+k} + \tau_{-1-2l} b_{n+k}\right) \right) \tag{23}$$

is a biorthogonal PFFS-dual with at least $2l$ zeros at $\frac{1}{2}$. Here we have set $N = 1$.

The proof of this result amounts to a deconvolution, plus the fact Eq. (18) is always satisfied whenever Δh is equal to the second term of Eq. (23).

2.2 Basic Relationships for Shift-Symmetric cases

As seen in the constructions in [10], among symmetric biorthogonal wavelets, there are also ones that are (without loss of generality) symmetric after shifting by a time index $t = \frac{1}{2}$. These, in relevance to linear phase filters, translate into the fact that

$$\tilde{H}(-\gamma) = \mathrm{e}^{2\pi i\gamma}\tilde{H}(\gamma). \tag{24}$$

We shall term those filters shift-symmetric. For this class of PFFS-duals, the $\hat{q}$ as in Eq. (11) will be slightly different. We have the following construction.

Theorem 4. *Let H and H^0 be a pair of biorthogonal filters, both satisfying (24). Assume that $H(\gamma) \neq 0$ for all γ except for $\gamma = \frac{1}{2}$ and perhaps a set of measure zero. Assume further that the dual filter H^0 has $2l+1$ zeros at $\gamma = \frac{1}{2}$ ($l = 0, 1, \cdots$). Then a class of $\triangle H(\gamma)$ satisfying the shifted-symmetry property (24) with $2l+1$ zeros at $\frac{1}{2}$ is given by*

$$\triangle H(\gamma) = (\cos\pi\gamma)^{2l+1}(\sin\pi\gamma)^{2l+1}\cos 2\pi\gamma\,\hat{q}_1(\gamma)\overline{H(\gamma+\frac{1}{2})}\mathrm{e}^{-2\pi i N\gamma}, \tag{25}$$

where $\hat{q}_1(\gamma) = \hat{q}_1(\cos 4\pi\gamma)|_{\gamma=\frac{1}{2}} \neq 0$, and N is an odd integer. Without particular specification, $\hat{q}_1 = 1$ and $N = 1$ whenever ΔH appears for shift-symmetric cases.

The proof of *Theorem 4* is similar to that of *Theorem 3*.

We can derive the expression of the filter sequence $\{\triangle h_n\}$ with a straightforward calculation.

Proposition 2. *Let $\triangle H$ and N be given in Theorem 4, and let $\tilde{h}_n$ be a bi-dual filter whose corresponding $\tilde{H}(\gamma)$ has $2l+1$ zeros at $\gamma = \frac{1}{2}$. Then*

$$h_n' = \tilde{h}_n + \lambda\left(\sum_k (-1)^k\overline{h_k}\left(\tau_{1-2l}c_{n+k} + \tau_{-1-2l}c_{n+k}\right)\right) \tag{26}$$

is a PFFS biorthogonal dual whose Fourier series H' has at least $2l+1$ zeros at $\frac{1}{2}$, and $\{c_n\}$ is given by

$$c_n = \begin{cases} (-1)^m\binom{2l+1}{m}\Big/(-2)^l, & n = 2m+1 \\ 0, & n = 2m. \end{cases}$$

The proof of this proposition is very similar to that of *Theorem 3* and *Proposition 1*.

3 Design Opportunities Provided by the Parametric Construction

In this section we provide examples of construction of bi-filters optimized to three different criteria:

1. Targeting a desired filter response in order to keep the filter length short, while keeping a given number of vanishing moments
2. Maximizing stopband attenuation, while keeping a given number of vanishing moments
3. Adding integer vanishing moments to a given bi-filter, which was the traditional focus of known constructions

In each case we demonstrate that the problem is reduced to the optimization over the free scalar parameter λ.

3.1 Targeting a Desired Filter

In filter design it is common to attempt to emulate the frequency response characteristics of some target filter H^t. Here we characterize this problem as finding $\min_{\tilde{H}}\{||\tilde{H} - H^t||_2\}$ subject to the condition that $\tilde{H}$ is dual to a B-spline filter H.

In the PFFS context, one such problem is reduced to an unconstrained optimization over a free parameter while maintaining perfect reconstruction and a given number of vanishing moments, without greatly lengthening the filter length. Namely, we consider problems such as

$$\min_{\lambda\in\mathbf{R}}\{||H^t - (H^0 + \lambda\Delta H)||_2\} \tag{27}$$

Proposition 3. *Let H be the filter response corresponding to a given B-spline wavelet. Let H^0 be a bi-filter having p zeros at $\gamma = \frac{1}{2}$. Let H^t be the targeting filter response with desirable characteristics, and let ΔH be given by Eq. (21) or (25) with also p zeros at $\gamma = \frac{1}{2}$. If*

$$\lambda^t = \frac{\int_0^1 Re\left(H^t\overline{\Delta H}\right)\,\mathrm{d}\gamma + \int_0^1 Re\left(H^0\overline{\Delta H}\right)\mathrm{d}\gamma}{||\Delta H||^2} \tag{28}$$

then

$$\|H^t - (H^0 + \lambda^t\Delta H)\|_2 = \min_{\lambda\in\mathbf{R}}\{\|H^t - (H^0 + \lambda\Delta H)\|_2\}$$

and $\tilde{H} = H^0 + \lambda^t\Delta H$ will have at least p zeros at $\gamma = \frac{1}{2}$.

Proof. With the given assumptions we can expand the square of the norm from Eq. (27),

$$\|\tilde{H} - H^t\|_2^2 = \|H^t - (H^0 + \lambda \Delta H)\|_2^2.$$

So our objective function is

$$\Gamma(\lambda) = \int_0^1 |H^t - (H^0 + \lambda \Delta H)|^2 \mathrm{d}\gamma.$$

Expanding the integrand, the objective function becomes quadratic in λ, and the optimal solution to λ^t arrives at Eq. (28).

3.2 Similar Performance with Shorter Filters

We can now apply the result of *Proposition 3* to design shorter bi-dual filters of spline wavelets that have frequency characteristics similar to those of longer length, while still keeping a given number of vanishing moments.

We begin by writing out one of the B-spline bi-filters from [10] as a sum of appropriately weighted "out of subspace" components (which is shown in *Proposition 5*). Let H^0 be a spline dual filter as constructed in [10] with p zeros at $\gamma = \frac{1}{2}$. It is true that a bi-filter with an additional four zeros at $\gamma = \frac{1}{2}$ will be given by

$$H_2 = H^0 + \lambda_1^* \Delta H_1 + \lambda_2^* \Delta H_2, \tag{29}$$

where ΔH_1 and ΔH_2 are given by Eq. (21) or (25) with p and $p+2$ zeros at $\gamma = \frac{1}{2}$, respectively. λ_1^* and λ_2^* are the corresponding parameter values from *Proposition 5* so that H_2 will then correspond to one of the filters in [10] with $p + 4$ zeros at $\gamma = \frac{1}{2}$.

Applying *Proposition 3* we can now construct a significantly *s*horter bi-filter targeting H_2 while still keeping p zeros at $\gamma = \frac{1}{2}$. This is done by adding $\lambda^t \Delta H_1$ to H^0 where λ^t is from Eq. (28). The following corollary states this result in detail and can be easily verified by substituting Eq. (29) for H^t into *Proposition (3)* with $\tilde{H} = H^0 + \lambda \Delta H_1$.

Corollary 2. *Let H^0 be a spline bi-filter with p zeros at $\gamma = \frac{1}{2}$ and let ΔH_1, ΔH_2, λ_1^*, λ_2^*, and the corresponding H_2 be as in Eq. (29). For $\tilde{H} = H^0 + \lambda \Delta H_1$, then the choice of*

$$\lambda = \lambda_1^* + \lambda_2^* \frac{\int Re\left(\Delta H_1 \overline{\Delta H_2}\right) \mathrm{d}\gamma}{\int |\Delta H_1|^2 \, g\gamma}$$

will minimize $\|H_2 - \tilde{H}\|_2$ while $\tilde{H}$ will maintain at least p vanishing moments and the time sequence $\tilde{h}$ will have four fewer taps than that of h_2.

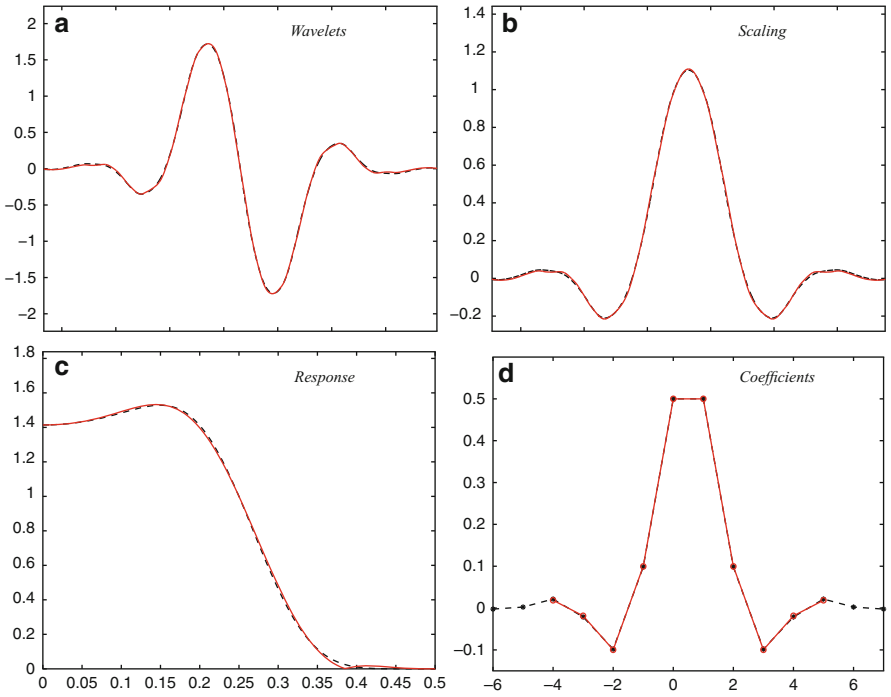


Fig. 1 *10-tap* PFFS-dual filters (*solid*) versus the *14-tap* filter (*dashed*) derived in [10]. Shown are the (**a**) wavelets, (**b**) Sscaling functions, (**c**) frequency response, and (**d**) filter coefficients, all dual to those of the first-order B-spline

Figures 1 and 2 show two examples of this construction. Notice that the frequency responses as well as the wavelets and scaling functions are nearly identical, whereas the filter sequences are shorter by four-taps.

3.3 Maximum Stopband Attenuation

In certain applications it is desirable that the filter response has a sharp decay in a given stopband. In this section we show how the parametric PFFS construction can produce dual filters with optimal attenuation for any given stopband frequency γ_s. We also show that this implementation reproduces the filters of the construction in [10] as $\gamma_s \to \frac{1}{2}$.

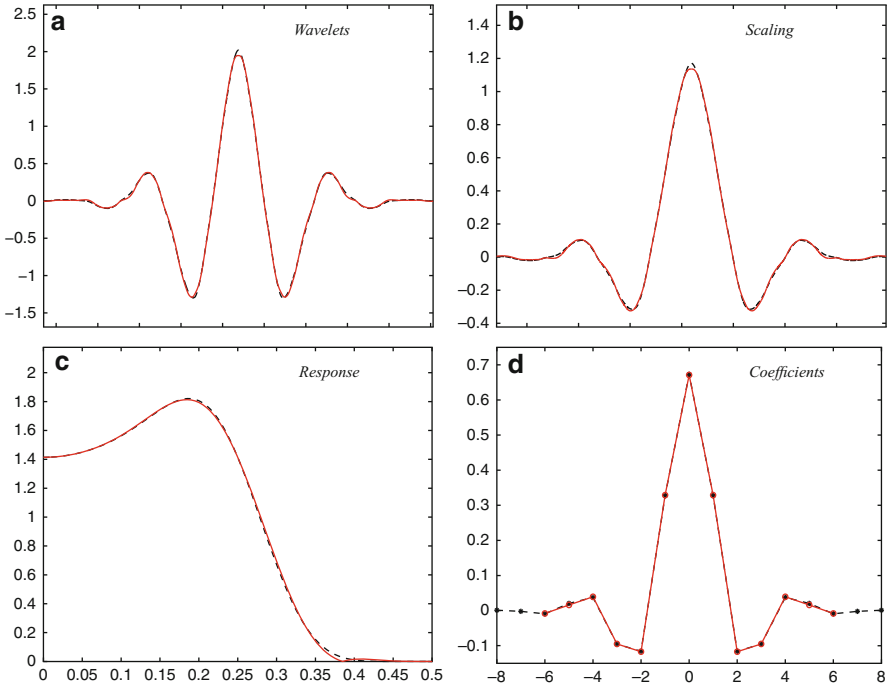


Fig. 2 *13-tap* PFFS-dual filters (*solid*) versus the *17-tap* filter (*dashed*) derived in [10]. Shown are the (**a**) wavelets, (**b**) scaling functions, (**c**) frequency response, and (**d**) filter coefficients dual to the second-order B-spline

3.3.1 The Parametric Formula

One stopband optimization problem (also seen in [28]) can be formulated as follows. Let H^0 be a given bi-dual filter with p zeros at $\frac{1}{2}$. Consider a set of all bi-dual filters $\tilde{H}$ with at least p zeros at $\gamma = \frac{1}{2}$. We may then set an objective function J be the energy of the frequency response $\tilde{H}$ between a chosen stopband γ_s and $\frac{1}{2}$ and minimize J over a given set of such dual filters $\tilde{H}$. Among all such dual $\tilde{H}$, we shall consider those of the same length and parameterized by a nonzero scaler λ, namely, consider only those $\tilde{H}$ that are given by $\tilde{H} = H^0 + \lambda \Delta H$ where ΔH, as given by Eq. (21) or (25), has p zeros at $\frac{1}{2}$. We then see that the minimization of J over these $\tilde{H}$ can be carried out by simply manipulating the λ parameter.

$$\begin{aligned} \min_{\tilde{H}} J &\equiv \min_{\tilde{H}} \int_{\gamma_s}^{\frac{1}{2}} \left|\tilde{H}(\gamma)\right|^2 \mathrm{d}\gamma \\ &= \min_{\lambda} \int_{\gamma_s}^{\frac{1}{2}} \left|H^0(\gamma) + \lambda \Delta H(\gamma)\right|^2 \mathrm{d}\gamma, \end{aligned} \tag{30}$$

where we minimize only over real λ to keep the sequence $\{\tilde{h}_n\}$ real valued.

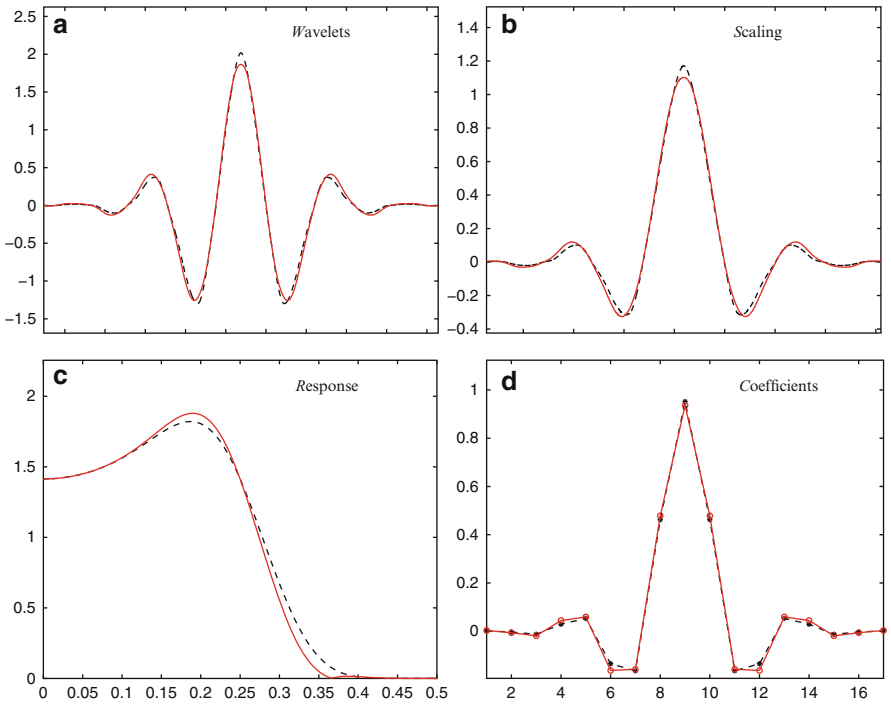


Fig. 3 PFFS duals with maximum stopband attenuation to a given stopband frequency (*solid*) are compared with those in [10] (*dashed*). Shown are the (**a**) wavelets, (**b**) scaling functions, (**c**) frequency response, and (**d**) filter coefficients associated with the length 17 filter, dual to the second-order B-spline

Again J is quadratic in λ, one can easily derive that the minimum is achieved at

$$\lambda^d = -\frac{\int_{\gamma_s}^{\frac{1}{2}} Re\left(H^0(\gamma)\overline{\triangle H(\gamma)}\right) \mathrm{d}\gamma}{\int_{\gamma_s}^{\frac{1}{2}} |\triangle H(\gamma)|^2 \mathrm{d}\gamma}. \tag{31}$$

This formula thus provides a means to maximize the stopband attenuation of bi-dual filters $\tilde{H}$ to any $\gamma_s \in (0, \frac{1}{2}]$.

Examples with maximum stopband attenuation for *given* stopbands are shown in Figs. 3 and 4. We see that the stopband attenuation of the new bi-filter frequency responses is indeed greater than those examples in [10] of the same length. Meanwhile, the smoothness of the scaling and wavelet functions has also been seemingly improved.

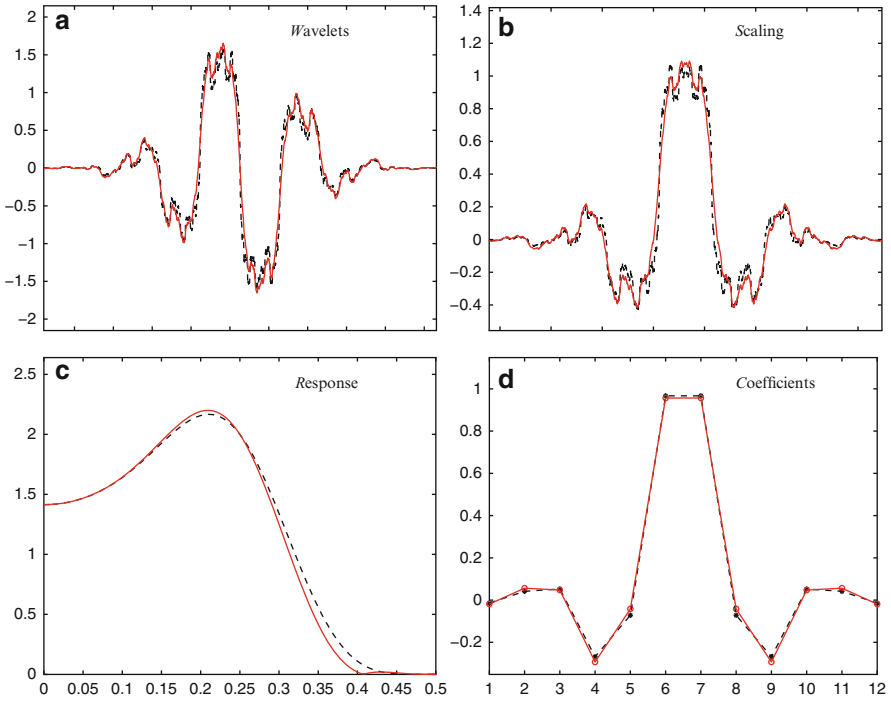


Fig. 4 PFFS duals with maximum stopband attenuation to a given stopband frequency (*solid*) are compared to those in [10] (*dashed*) with the same length filter. Shown are the (**a**) wavelets, (**b**) scaling functions, (**c**) frequency response, and (**d**) filter coefficients associated with the length 12 filter, dual to the third-order B-spline

3.3.2 Limiting Behavior of $\lambda^d(\gamma_s)$

Here is an observation worth of mentioning. By letting $\gamma_s \to \frac{1}{2}$ the value of λ^d from Eq. (31) converges to λ^* as in *Proposition 5* which adds precisely two zeros at $\gamma = \frac{1}{2}$ to the original bi-dual H^0.

Proposition 4. *Let H^0 be a biorthogonal dual of a spline function with p vanishing moments as constructed in [10], and let ΔH be given by Eq. (21) or (25). Then as $\gamma_s \to \frac{1}{2}$*

$$\lambda^d \to \lambda^*$$

where λ^* is such that $\widetilde{H} = H^0 + \lambda^* \Delta H$ has two additional integer vanishing moments as seen in Proposition 5.

Proof. For the symmetric case, we start by simply expanding the general form of the filters as given in *Proposition 1* into Eq. (31). Since the functions in the integrands are bounded and integrable, as $\gamma_s \to \frac{1}{2}$ the integrals must go to zero, and we are justified in using L'Hopital's rule to evaluate the limit. The fundamental theorem of

calculus then allows us to evaluate the limit:

$$\lim_{\gamma_s \to \frac{1}{2}} \lambda^d = \lim_{\gamma_s \to \frac{1}{2}} -\frac{\sum_{n=0}^{k-1} \binom{k-1+n}{n} (\cos \pi\gamma_s)^{4\tilde{l}} (\sin \pi\gamma_s)^{2n} (\sin \pi\gamma_s)^{2k} \cos 2\pi\gamma_s}{2^{3\tilde{l}+1} (\sin \pi\gamma_s)^{4k} (\cos \pi\gamma_s)^{4\tilde{l}} (\cos 2\pi\gamma_s)^2}$$

$$= \frac{\sum_{n=0}^{k-1} \binom{k-1+n}{n}}{2^{3\tilde{l}+1}}.$$

Plugging this value into Eq. (1) and setting $\gamma_s = \frac{1}{2}$ will confirm the result.

The proof of the shift-symmetric case is completely similar.

This observation indicates that the bi-filters designed in [10] could not have the stopband attenuation considered since the stopband frequency γ_s is equivalently set at $\gamma_s = \frac{1}{2}$ in [10].

3.4 Ability to Increase Wavelet Vanishing Moments

The number of vanishing moments of a wavelet is directly related to the regularity of the bi-scaling and bi-wavelet functions, which is also known to be equal to the number of zeros of the filter H and $\tilde{H}$ at $\gamma = \frac{1}{2}$. We shall demonstrate how zeros at $\gamma = \frac{1}{2}$ of a PFFS-dual filter can be easily increased by the parametric approach in the general setting, then show how results for the setting of [10] can be reproduced using the PFFS approach.

Assume that H and H^0 are a pair of dual biorthogonal filters. Then according to our earlier discussions, any new PFFS-dual biorthogonal filter can be written as

$$\widetilde{H}(\gamma) = H^0(\gamma) + \overline{H(\gamma + \frac{1}{2})}\,\hat{q}(\gamma)$$

where $\hat{q}$ is a trig polynomial satisfying $\hat{q}(\gamma) + \hat{q}(\gamma + \frac{1}{2}) = 0$. One can verify that to keep at least the same number of zeros at $\frac{1}{2}$ (same number of vanishing moment in ψ), $\hat{q}$ can be of the following form:

$$\hat{q}(\gamma) = H^0(\gamma) H^0(\gamma + \frac{1}{2}) \hat{q}_1(\cos 4\pi\gamma) \cos 2\pi N\gamma$$

for some odd integer N. Hence,

$$\widetilde{H}(\gamma) = H^0(\gamma) \left(1 + H^0(\gamma + \frac{1}{2}) \overline{H(\gamma + \frac{1}{2})} \hat{q}_1(\cos 4\pi\gamma) \cos 2\pi N\gamma\right)$$

$$= H^0(\gamma) F(\gamma)$$

where $F(\gamma) \equiv 1 + H^0(\gamma + \frac{1}{2})\overline{H(\gamma + \frac{1}{2})}\hat{q}_1(\cos 4\pi\gamma)\cos 2\pi N\gamma$. Evidently, the new PFFS-dual biorthogonal filter has at least the same number of zeros at $\gamma = \frac{1}{2}$. Observe however,

$$F(\frac{1}{2}) = 1 - H^0(0)\overline{H(0)}\,\hat{q}_1(1),$$

which can easily yield another zero at $\gamma = \frac{1}{2}$ for $\hat{q}_1(1) = 1/H^0(0)\overline{H(0)}$ since $H^0(0)\overline{H(0)} \neq 0$.

In the context of the spline bi-wavelet system constructed in [10] the following proposition can easily be verified and shows that the PFFS approach reproduces the results of [10].

Proposition 5. *Suppose H is the filter corresponding to a spline function with $2l$ zeros at $\frac{1}{2}$ for the symmetric case (or $2l+1$ zeros at $\frac{1}{2}$ for the shift-symmetric case) and that the biorthogonal dual filter H^0 has $2\tilde{l}$ zeros at $\frac{1}{2}$ for the symmetric case (or $2\tilde{l}+1$ zeros at $\frac{1}{2}$ for the shift-symmetric case). If ΔH is given by Eq. (21) (or (25) for the shift-symmetric case), and*

$$\lambda^* = \frac{\sum_{n=0}^{k-1}\binom{k-1+n}{n}}{2^{3\tilde{l}+1}}$$

with $k = l + \tilde{l}$, or for the shift-symmetric case,

$$\lambda^* = \frac{\sum_{n=0}^{k-1}\binom{k-1+n}{n}}{2^{3\tilde{l}+3}}$$

with $k = l + \tilde{l} + 1$, the PFFS dual given by $\tilde{H} = H^0 + \lambda^*\Delta H$ will have an additional two zeros at $\frac{1}{2}$.

3.5 Other Optimization Potentials

We demonstrated in the previous sections a few possibilities for optimizing the bi-filter construction over a scalar λ. This is quite effective and the major rationale for working with a scalar is to keep the bi-filter length as small as possible. We also mentioned only three criteria that are all signal *independent*. So, we mention here two other possible optimization problems that could be carried out with our PFFS approach.

3.5.1 Nonscalar $\hat{q}_1$ Polynomials

We can choose the trig polynomial $\hat{q}_1(\cos 4\pi\gamma)$ as in *Theorem 3* to have more than one parameter to work with so as to enhance the optimization potentials. Take the symmetric case, for instance, let

$$\hat{q}_1 = \lambda + \xi(1 - \cos 4\pi\gamma). \tag{32}$$

λ and ξ could be used to tune the filter to meet two different criteria at the same time. For example, one would be able to increase the number of zeros at $\gamma = \frac{1}{2}$ by a choice of λ since the second term is zero at $\frac{1}{2}$, and ξ could be adjusted to enhance uniform Lipschitz-α regularity. The only sacrifice is that such "out of subspace" components add more number of nonzero coefficients to the filters.

Choices such as Eq. (32) with two or more parameters open up a vast degree of flexibility in filter design and are believed to have further applications to signal-dependent methods not discussed here.

3.5.2 Ability to Maximize the Coding Gain

Maximum coding gain is often a design goal to maximize the energy compaction after the sub-band decomposition and to enhance the coding/compression efficiency. In [21] the PFFS construction methodology is used to construct filters with maximum coding gain for common signal models. This is a signal-dependent design approach and it is thus not discussed further here.

Matlab programs for the PFFS optimal filter designs that can be installed and added into the Wavelet Toolbox are available upon request to the authors.

Acknowledgements Shidong Li is partially supported by NSF grants DMS-0406979 and DMS-0709384.

References

1. Aldroubi, A.: Portraits of frames. Proc. Amer. Math. Soc. **123**, 1661–1668 (1995)
2. Benedetto, J., Walnut, D.F.: Chapter 3, Gabor frames for L^2 and related spaces. In: Benedetto, J.J., Frazier, M.W. (eds.) Wavelets: Mathematics and Applications. CRC Press, Boca Raton (1994)
3. Benedetto, J.J.: Chapter 7, Frame decompositions, sampling, and uncertainty principle inequalities. In: Benedetto, J.J. Frazier, M.W. (eds.) Wavelets: Mathematics and Applications. CRC Press, Boca Raton (1994)
4. Benedetto, J.J., Li, S.: The theory of multiresolution analysis frames and applications to filter banks. Appl. Comput. Harmon. Anal. **5**, 389–427 (1998)
5. Bolcskei, H., Hlawatsch, F.: Oversampled filter banks: Optimal noise shaping, design freedom, and noise analysis. In: ICASSP '97: Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)-Volume 3, pp. 2453. IEEE Computer Society, Washington, DC, USA (1997)

6. Casazza, P.G., Christensen, O.: Hilbert space frames containing a Riesz basis and Banach spaces which have no subspace isomorphic to C_0. J. Math. Anal. Appl. **202**(3), 940 (1996)
7. Christensen, O.: Frames and pseudo-inverses. J. Math. Anal. Appl. **195**, 401–414 (1995)
8. Christensen, O., Heil, C.: Perturbations of Banach frames and atomatic decompositions. Math. Nach. **185**, 33–47 (1997)
9. Chui, C.K.: Wavelets: A Tutorial in Theory and Applications. Academic, Boston (1992)
10. Cohen, A., Daubechies, I., Feauveau, J.C.: Biorthogonal bases of compactly supported wavelets. Comm. Pure and Appl. Math. **XLV**, 485–560 (1992)
11. Cvetkovic, Z., Vetterli, M.: Oversampled filter banks. IEEE Trans. Signal Process **46**, 1245–1255 (1998)
12. Daubechies, I.: The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Inform. Theor. **36**(5), 961–1005 (1990)
13. Daubechies, I.: Ten lectures on wavelets. SIAM **61** (1992)
14. Daubechies, I., Grossmann, A., Meyer, Y.: Painless nonorthogonal expansions. J. Math. Phys. **27**, 1271–1283 (1986)
15. Duffin, R., Schaeffer, A.: A class of nonharmonic Fourier series. Trans. Amer. Math. Soc. **72**, 341–366 (1952)
16. Feichtinger, H.G., Grochenig, K.: Gabor wavelets and the Heisenberg group: Gabor expansions and short time Fourier transform from the group theoretical point of view. In: Chui, C.K. (ed.) Wavelets: A Tutorial in Theory and Applications, pp. 359–398. Academic, Boston **2**, (1992)
17. Feichtinger, H.G., Zimmermann, G.: A Banach space of test functions for Gabor analysis. In: Feichtinger, H.G., Strohmer, T. (eds.) in Gabor Analysis and Algorithms: Theory and Applications. Birkhäuser, Basel (1997)
18. Gopinath, R.A., Burrus, C.S.: Wavelet transforms and filter banks. In: Chui, C.K. (ed.) Wavelets: A Tutorial in Theory and Applications, pp. 603-654. Academic, Boston (1992)
19. Gröchenig, K.: Describing functions: atomic decompositions versus frames. Monatsh. Math. **112**, 87–104 (1991)
20. Heil, C., Walnut, D.: Continuous and discrete wavelet transforms. SIAM Rev. **31**, 628–666 (1989)
21. Herman, M., Li, S.: Biorthogonal wavelets of maximum coding gain through pseudoframes for subspaces. In: Mathematics of Data/Image Pattern Recognition, Compression, and Encryption with Applications IX., vol. 6315, pp. 631501 (2006)
22. Jayant, N.S., Noll, P.: Digital Coding of Waveforms. Prentice Hall, Englewood Cliffs, 07632 (1984)
23. Larson, D.: Frames and wavelets from an operator-theoretical point of view. Contemp. Math. **228**, 201–218 (1998)
24. Larson, D., Han, D.: Frames, bases and group representations. No. 697, American Mathematical Society **147** (2000).
25. Li, S., Ogawa, H.: Pseudoframes for subspaces with applications. Proc. SPIE Conf. on Wavelet Applications and Mathematical Imaging, VI, San Diego, July 22 (1998)
26. Li, S., Ogawa, H.: Pseudoframes in separable Hilbert spaces. Tech. Report, T.I.T. (2000)
27. Li, S., Ogawa, H.: Pseudoframes for subspaces with applications. J. Fourier Anal. Appl. June **10**(4), 409–431 (2004)
28. Strang, G., Nguyen, T.: Wavelets and Filter Banks. Welesley-Cambridge Press, Wellesley (1996)
29. Vaidyanathan, P.P.: Multirate Systems and Filter Banks. Prentice Hall, Englewood Cliffs, New Jersey (1993)
30. Vetterli, M., Herley, C.: Wavelets and filter banks: theory and design. IEEE ASSP **40**(9), 2207–2232 (1992)
31. Vitterli, M.: Filter banks allowing perfect reconstruction. Signal Process. **10**, 219–244 (1986)

On the Convergence of Iterative Filtering Empirical Mode Decomposition

Yang Wang and Zhengfang Zhou

Abstract Empirical mode decomposition (EMD), an adaptive technique for data and signal decomposition, is a valuable tool for many applications in data and signal processing. One approach to EMD is the iterative filtering EMD, which iterates certain banded Toeplitz operators in $l^\infty(\mathbb{Z})$. The convergence of iterative filtering is a challenging mathematical problem. In this chapter we study this problem, namely for a banded Toeplitz operator T and $\mathbf{x} \in l^\infty(\mathbb{Z})$ we study the convergence of $T^n(\mathbf{x})$. We also study some related spectral properties of these operators. Even though these operators don't have any eigenvalue in Hilbert space $l^2(\mathbb{Z})$, all eigenvalues and their associated eigenvectors are identified in $l^\infty(\mathbb{Z})$ by using the Fourier transform on tempered distributions. The convergence of $T^n(\mathbf{x})$ relies on a careful localization of the generating function for T around their maximal points and detailed estimates on the contribution from the tails of $\mathbf{x}$.

Keywords Finite impulse response filter • Toeplitz operator • Empirical mode decomposition • Intrinsic mode functions • Iterative filtering

1 Introduction

Let $\mathbf{a} = (a_k) \in l^1(\mathbb{Z})$. We consider the operator
$T_\mathbf{a} : l^\infty(\mathbb{Z}) \longrightarrow l^\infty(\mathbb{Z})$ associated with $\mathbf{a}$, given by

$$T_\mathbf{a}(\mathbf{x}) = \Big(\sum_{j\in\mathbb{Z}} a_j x_{k+j}\Big)_{k\in\mathbb{Z}}$$

Y. Wang (✉) • Z. Zhou
Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA,
e-mail: ywang@msu.math.edu; zfzhou@msu.math.edu

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_8,

where $\mathbf{x} = (x_k) \in l^\infty(\mathbb{Z})$. In the signal processing literature $T_\mathbf{a}$ is called a *filter*, and it is a *finite impulse response (FIR)* filter if $a_k \neq 0$ for only finitely many $k \in \mathbb{Z}$. Note that $T_\mathbf{a}$ is in fact a Toeplitz operator and an FIR filter simply means the Toeplitz operator $T_\mathbf{a}$ is *banded*. In this chapter we shall use the terms filter and Toeplitz operator interchangeably, and only FIR filters and banded Toeplitz operators will be considered. Toeplitz operators are classical operators that have been studied extensively, see [2] and the references therein. There is an even larger literature on filters, which we shall not divulge into. In this chapter our main focus is on the iteration of certain type of banded Toeplitz operators. More precisely, we consider the following question: Let $T_\mathbf{a}$ be banded and $\mathbf{x} \in l^\infty(\mathbb{Z})$. When will $T_\mathbf{a}^n(\mathbf{x})$ converge (in the sense that every entry converges) as $n \to \infty$? This question arises from signal and data processing using *empirical mode decomposition (EMD)*, which is an important tool for analyzing digital signals and data sets [8, 12]. Our study is motivated primarily by the desire to provide a mathematical framework for EMD.

Signal and data analysis is an important and necessary part in both research and practical applications. Understanding large data set is particularly important and challenging given the explosion of data and numerous ways they are being collected today. Often the challenge is to find hidden information and structures in data and signals. To do so one might encounter several difficulties with the data: The data represent a nonlinear process and is nonstationary; the essential information in the data is often mingled together with noise or other irrelevant information, and others. Historically, Fourier spectral analysis has provided a general method for analyzing signals and data. The term "spectrum" is synonymous with the Fourier transform of the data. Another popular technique is wavelet transform. These techniques are often effective but are known to have their limitations. To begin with, none of these techniques is data adaptive. This can be a disadvantage in some applications. There are other limitations. For example, Fourier transform may not work well for nonstationary data or data from nonlinear systems. It also does not offer spatial and temporal localization to be useful for some applications in signal processing. Wavelet transform captures discontinuities very successfully. But it too has many limitations; see [8] for a more detailed discussion. These limitations have led Huang et al. [8] to propose the *EMD* as a highly adaptive technique for analyzing data. EMD has turned out to be a powerful complementary tool to Fourier and wavelet transforms. The goal of EMD is to decompose a signal into a finite number of *intrinsic mode functions (IMF)*, from which hidden information and structures can often be captured by analyzing their Hilbert transformations. We shall not discuss the details of IMF and EMD in this chapter. They can be found in [3,5,8,9,12,13,15] and the references therein.

The original EMD is obtained through an algorithm called the *sifting algorithm*. The local maxima and minima of a function (signal) are respectively connected via cubic splines to form the so-called upper and lower envelopes. The average of the two envelopes is then subtracted from the original data. This process is iterated to obtain the first IMF in the EMD. The other IMFs are obtained by the same process on the residual signal. The sifting algorithm is highly adaptive. A small perturbation, however, can alter the envelopes dramatically, raising some questions

about its stability. Another drawback is that there is no natural way to generalize EMD to higher dimensions, which severely limits the scope of its applications. As powerful as EMD is in many applications, a mathematical foundation is virtually nonexistent. Many fundamental mathematical issues such as the convergence of the sifting algorithm have never been established.

To address these concerns, a new approach, the *iterative filtering EMD*, is proposed in [12]. Instead of the average of the upper and lower envelopes, the iterative filtering EMD replaces them by certain FIR filters, usually low-pass filters that yield a "moving average" similar to the mean of the envelopes in the original sifting algorithm. It is shown in [12] that iterative filtering approach often leads to comparable EMD as the classical EMD, and in general it serves as a useful alternative or complement. Furthermore iterative filtering EMD has some advantages over the classic EMD, making it well suited for certain applications [10, 14, 16].

The iterative filtering EMD proposed in [12] has the following set up: let $\mathbf{a} = (a_k)_{k\in\mathbb{Z}}$ be finitely supported, i.e., only finitely many $a_k \neq 0$, which we choose so that $T_{\mathbf{a}}(\mathbf{x})$ represents a "moving average" of $\mathbf{x}$. Now let

$$\mathscr{L}(\mathbf{x}) = \mathbf{x} - T_{\mathbf{a}}(\mathbf{x}). \tag{1}$$

The first IMF in the EMD is given by $\mathbf{I}_1 = \lim_{n\to\infty} \mathscr{L}^n(\mathbf{x})$, and subsequent IMF's are obtained recursively via $\mathbf{I}_k = \lim_{n\to\infty} \mathscr{L}_k^n(\mathbf{x} - \mathbf{I}_1 - \cdots - \mathbf{I}_{k-1})$. In practical applications the process stops when some stopping criterion is met. For a periodic $\mathbf{x}$ the convergence of $\mathscr{L}^n(\mathbf{x})$ is completely characterized in [12]. However, the convergence for $\mathbf{x} \in l^\infty(\mathbb{Z})$ in general is a much more difficult problem. The main purpose of this chapter is to study this question.

The rest of this chapter is organized as follows: In Sect. 2 we introduced the notations and state the main theorem. In Sect. 3 we prove a result on sum of Dirac measures, which is closely related to the Poisson summation formula as well as a classical result of Cordoba [4]. We use it to characterize all eigenvectors of banded Toeplitz operators on $l^p(\mathbb{Z})$ for $1 \leq p \leq \infty$. The proof of the main theorem, which is quite tedious, is given in Sect. 4.

2 Main Result and Notations

For any $\mathbf{x} = (x_k)_{k\in\mathbb{Z}} \in l^\infty(\mathbb{Z})$ we shall use $\mathbf{x}_N$ to denote the cutoff of $\mathbf{x}$ from $k = -N$ to N, i.e., $\mathbf{x}_N = (y_k)$ such that $y_k = x_k$ for $-N \leq x \leq N$ and $y_k = 0$ otherwise. We shall often also view $\mathbf{x} = (x_k)_{k\in\mathbb{Z}} \in l^\infty(\mathbb{Z})$ as a function $\mathbf{x} : \mathbb{Z}\longrightarrow\mathbb{C}$ with $\mathbf{x}(k) = x_k$. We say $\mathbf{x} = (x_k) \in l^\infty(\mathbb{Z})$ is *symmetric* if $x_k = x_{-k}$ for all $k \in \mathbb{Z}$, and it is *finitely supported* if $\text{supp}(\mathbf{x}) := \{k \in \mathbb{Z} : x_k \neq 0\}$ is a finite set.

Throughout this chapter the Fourier transform of a function $f(x)$ is defined as

$$\mathscr{F}(f)(\xi) = \widehat{f}(\xi) := \int_{\mathbb{R}} f(x)\mathrm{e}^{2\pi i x\xi}\,\mathrm{d}x.$$

The inverse Fourier transform of $g(\xi)$ is

$$\mathscr{F}^{-1}(g)(x) := \int_{\mathbb{R}} g(\xi)\mathrm{e}^{-2\pi i\xi x}\,\mathrm{d}\xi.$$

For each $\mathbf{x} \in l^\infty(\mathbb{Z})$ there is an associated complex measure $\mu_{\mathbf{x}} := \sum_{k\in\mathbb{Z}} x_k\delta_k$, where δ_b is the Dirac measure supported at b for any $b \in \mathbb{R}$, i.e., $\delta_b(x) = \delta(x-b)$. It is well known that $\mu_{\mathbf{x}}$ is a tempered distribution. Thus $\widehat{\mu_{\mathbf{x}}}$ is also well defined as a tempered distribution. We shall often use $\widehat{\mathbf{x}}$ to denote $\widehat{\mu_{\mathbf{x}}}$ for simplicity, especially when $\mathbf{x}$ is finitely supported; in such case $\widehat{\mathbf{x}}(\xi)$ is a trigonometric polynomial.

Going back to Toeplitz operators, it is easy to check that for any $\mathbf{a} \in l^1(\mathbb{Z})$ we have

$$\widehat{T_{\mathbf{a}}(\mathbf{x})}(\xi) = \widehat{\mathbf{a}}(-\xi)\widehat{\mu_{\mathbf{x}}}(\xi) = \widehat{\mathbf{a}}(-\xi)\widehat{\mathbf{x}}(\xi).$$

For any finitely supported $\mathbf{a}$ the spectrum of $T_{\mathbf{a}}$ is precisely $\{\widehat{\mathbf{a}}(\xi) : \xi \in [0,1)\}$. Let $Z_{\mathbf{a},\lambda} = \{\theta \in [0,1) : \widehat{\mathbf{a}}(\theta) = \lambda\}$. This set will occur very frequently in this chapter.

Before stating our main theorem we introduce a few more notations. For any $\theta \in \mathbb{R}$ let $\mathbf{v}_\theta := (\mathrm{e}^{2\pi i k\theta})_{k\in\mathbb{Z}}$. If $\theta \in Z_{\mathbf{a},\lambda}$ then $T_{\mathbf{a}}(\mathbf{v}_\theta) = \lambda\mathbf{v}_\theta$. For any $\mathbf{x} = (x_k)$ and $\mathbf{y} = (y_k)$ in $l^\infty(\mathbb{Z})$ define

$$[\mathbf{x},\mathbf{y}] = \lim_{n\to\infty} \frac{1}{2n+1}\sum_{k=-n}^{n} x_k\bar{y}_k$$

if it exists. One can view this as a form of "inner product."

One of the main objectives of this chapter is to study the convergence of the new sifting algorithm from which we obtain the IMFs by $\mathbf{I}_k = \lim_{n\to\infty}(I - T_{\mathbf{a}_k})^n(\mathbf{x} - \mathbf{I}_1 - \cdots - \mathbf{I}_{k-1})$. Since $I - T_{\mathbf{a}}$ is simply the Toeplitz operator $T_{\delta-\mathbf{a}}$ where $\delta = (\delta_{k0})$ with $\delta_{00} = 1$ and $\delta_{k0} = 0$ for $k \neq 0$. So we shall focus on iterations of $T_{\mathbf{a}}$ for general finitely supported $\mathbf{a}$. Our main theorem of this chapter is:

Theorem 2.1. *Let* $\mathbf{a} = (a_k)$ *be finitely supported and symmetric such that* $-1 < \widehat{\mathbf{a}}(\xi) \leq 1$ *and* $\widehat{\mathbf{a}}(\xi) \not\equiv 1$. *For any* $\mathbf{x} \in l^\infty(\mathbb{Z})$, *if* $[\mathbf{x},\mathbf{v}_\theta]$ *exists for all* $\theta \in Z_{\mathbf{a},1}$ *then*

$$\lim_{n\to\infty} T_{\mathbf{a}}^n(\mathbf{x}) = \sum_{\theta\in Z_{\mathbf{a},1}} [\mathbf{x},\mathbf{v}_\theta]\mathbf{v}_\theta \qquad \textit{pointwise}. \tag{2}$$

Here pointwise convergence means the kth entry $T_{\mathbf{a}}^n(\mathbf{x})(k)$ of $T_{\mathbf{a}}^n(\mathbf{x})$ converges for each $k \in \mathbb{Z}$. Informally speaking, $T_{\mathbf{a}}^n(\mathbf{x})$ converges pointwise to the "projection" of $\mathbf{x}$ onto the 1-eigenspace of $T_{\mathbf{a}}$. Note that the eigenvalues of $T_{\mathbf{a}}$ are precisely $\{\widehat{\mathbf{a}}(\xi) : \xi \in [0,1)\}$ (see Sect. 3), so the condition $-1 < \widehat{\mathbf{a}}(\xi) \leq 1$ is natural. It is not clear whether the condition $[\mathbf{x},\mathbf{v}_\theta]$ exists for each $\theta \in Z_{\mathbf{a},1}$ is a necessary condition. The following example shows that $\lim_{n\to\infty} T_{\mathbf{a}}^n(\mathbf{x})$ does not exist for a $x \in l^\infty(\mathbb{Z})$.

Example 2.1. Let $\mathbf{a} = (a_k)$ with $a_0 = \frac{1}{2}$, $a_1 = a_{-1} = \frac{1}{4}$, and $a_k = 0$ for all other k. $\widehat{\mathbf{a}}(\xi) = \sin^2 \frac{\xi}{2}$ satisfies the hypothesis of Theorem 2.1. Let $\mathbf{x} = (x_k)$ where $x_k = 0$ for all $k \leq 0$ and $x_k = (-1)^{n-1}$ for $2^{n!} \leq K < 2^{(n+1)!}$. Then it is easy to show that $T^n_{\mathbf{a}}(\mathbf{x})(0)$ does not converge. In fact every point in $[-\frac{1}{2}, \frac{1}{2}]$ is a limit point of the sequence.

3 Eigenvectors of Banded Toeplitz Operators

To study the iterations of banded Toeplitz operators it is natural to ask about their eigenvalues and eigenvectors in $l^\infty(\mathbb{Z})$. We state some results here. While these results may not be new (although we have not found them in the literature), our approach appears to be.

For any $\mathbf{x} = (x_k) \in l^\infty(\mathbb{Z})$ the associated measure $\mu_{\mathbf{x}}$ is a tempered distribution [7, 11]. Hence its Fourier transform, given by

$$\langle \widehat{\mu_{\mathbf{x}}}, \phi \rangle := \langle \mu_{\mathbf{x}}, \widehat{\phi} \rangle = \sum_{k \in \mathbb{Z}} x_k \widehat{\phi}(k) \tag{3}$$

for any ϕ in the Schwartz class, is also a tempered distribution.

Lemma 3.1. *Let* $\mathbf{x} \in l^\infty(\mathbb{Z})$ *such that* $\operatorname{supp}(\widehat{\mu_{\mathbf{x}}}) = \Lambda$ *is a uniformly discrete set in* $\mathbb{R}$. *Then*

$$\widehat{\mu_{\mathbf{x}}} = \sum_{\beta \in \Lambda} c_\beta \delta_\beta \tag{4}$$

for some bounded sequence $(c_\beta)_{\beta \in \Lambda}$ *in* $\mathbb{C}$.

Proof. Since Λ is uniformly discrete we may find $\psi_\beta \in C_0^\infty(\mathbb{R})$ for each $\beta \in \Lambda$ such that $\sum_{\beta \in \Lambda} \psi_\beta = 1$ and $\operatorname{supp}(\psi_\beta) \cap \Lambda = \{\beta\}$. Now for any $\beta \in \Lambda$, $\psi_\beta \widehat{\mu_{\mathbf{x}}}$ is a tempered distribution supported on a single-point $\{\beta\}$. It follows that

$$\psi_\beta \widehat{\mu_{\mathbf{x}}} = \sum_{j=0}^{N} a_j \delta_\beta^{(j)},$$

where $\delta_\beta^{(j)}$ denotes the jth derivative of δ_β (see, e.g., Folland [6]). We only need to show that $a_j = 0$ for $j > 0$ for all $\beta \in \Lambda$. If not there exists some $\alpha^* \in \Lambda$ such that $\psi_{\alpha^*} \widehat{\mu_{\mathbf{x}}} = \sum_{j=0}^{N} a_j \delta_{\alpha^*}^{(j)}$ with $a_N \neq 0$, $N > 0$. □

Without loss of generality we assume that $\alpha^* = 0$. Now take a test function $\phi \in C_0^\infty(\mathbb{R})$ such that $\operatorname{supp}(\phi) = [-\varepsilon, \varepsilon]$, $\operatorname{supp}(\phi) \cap \operatorname{supp}(\psi_\beta) = \emptyset$ for all $\beta \neq 0$, and $\phi^{(j)}(0) = 0$ for $j < N$ but $\phi^{(N)}(0) \neq 0$. Set $\phi_\lambda(x) = \phi(\lambda x)$. Then

$$\langle \widehat{\mu_{\mathbf{x}}}, \phi_\lambda \rangle = \left\langle \sum_{\beta \in \Lambda} \psi_\beta \widehat{\mu_{\mathbf{x}}}, \phi_\lambda \right\rangle = \langle \psi_0 \widehat{\mu_{\mathbf{x}}}, \phi_\lambda \rangle = \left\langle \sum_{j=0}^{N} a_j \delta_0^{(j)}, \phi_\lambda \right\rangle = a_N \lambda^N \phi^{(N)}(0),$$

which goes to ∞ as $\lambda\to\infty$. On the other hand,

$$|\langle\widehat{\mu_{\mathbf{x}}},\phi_\lambda\rangle| \;=\; |\langle\mu_{\mathbf{x}},\widehat{\phi_\lambda}\rangle| \;=\Big|\;\frac{1}{\lambda}\sum_{k\in\mathbb{Z}} x_k\widehat{\phi}\Big(\frac{k}{\lambda}\Big)\;\Big|\le\;\frac{C}{\lambda}\sum_{k\in\mathbb{Z}}\Big|\;\widehat{\phi}\Big(\frac{k}{\lambda}\Big)\;\Big|,$$

which goes to $C\int_{\mathbb{R}}|\widehat{\phi}(\xi)|\,d\xi$ as $\lambda\to\infty$. This is a contradiction. Thus $\widehat{\mu_{\mathbf{x}}} = \sum_{\beta\in\Lambda} c_\beta\delta_\beta$.

It remains to show c_β are bounded. Take a test function $\phi\in C^\infty(\mathbb{R})$ such that $\phi(0)=1$ and $\operatorname{supp}(\phi)=[-\varepsilon,\varepsilon]$, where $\varepsilon<\inf\{|\alpha_1-\alpha_2|:\alpha_1,\alpha_2\in\Lambda,\alpha_1\neq\alpha_2\}$. Applying $\langle\widehat{\mu_{\mathbf{x}}},\varphi\rangle=\langle\mu_{\mathbf{x}},\widehat{\varphi}\rangle$ to $\varphi(x)=\phi(x-\beta)$ for each $\beta\in\Lambda$ we obtain

$$|c_\beta| = |\langle\mu_{\mathbf{x}},\widehat{\varphi}\rangle| \;=\Big|\sum_{k\in\mathbb{Z}} x_k\mathrm{e}^{-2\pi ik\beta}\widehat{\phi}(k)\;\Big|\le\;\|\mathbf{x}\|_\infty\sum_{k\in\mathbb{Z}}\Big|\;\widehat{\phi}(k)\;\Big|\,.$$

This proves the lemma.

The following theorem is closely related to a well-known result of Cordoba [4], which is a classic result in the study of quasicrystals.

Theorem 3.2. *Let Λ be a uniformly discrete set in $\mathbb{R}$ and $\mu=\sum_{\beta\in\Lambda}x_\beta\delta_\beta$ where (x_β) is bounded. Assume that* $\operatorname{supp}(\widehat{\mu})\subset\mathbb{Z}$. *Then*

(A) There exist $\alpha_1,\dots,\alpha_m\in[0,1)$ such that $\Lambda=\bigcup_{j=1}^m(\alpha_j+\mathbb{Z})$.
(B) There exist $c_1,\dots,c_m$ such that $x_\beta=c_j$ for all $\beta\in\alpha_j+\mathbb{Z}$. Thus,

$$\mu=\sum_{j=1}^m c_j\sum_{k\in\mathbb{Z}}\delta_{\alpha_j+k}.$$

(C) $\widehat{\mu}=\sum_{k\in\mathbb{Z}}\Big(\sum_{j=1}^m c_j\mathrm{e}^{2\pi i\alpha_j k}\Big)\delta_k$.

Proof. By Lemma 3.1 we have $\widehat{\mu}=\sum_{k\in\mathbb{Z}}p_k\delta_k$. For $\phi\in C_0^\infty(\mathbb{R})$ denote $\phi_{\lambda,t}(x):=\phi(\lambda x)\mathrm{e}^{2\pi itx}$. Then $\widehat{\phi_{\lambda,t}}(\xi)=\lambda^{-1}\phi\big(\lambda^{-1}(\xi-t)\big)$.

It follows from $\langle\mu,\widehat{\phi_{\lambda,t}}\rangle=\langle\widehat{\mu},\phi_{\lambda,t}\rangle$ that

$$\sum_{k\in\mathbb{Z}}p_k\phi_{\lambda,t}(k)=\sum_{\alpha\in\Lambda}x_\alpha\widehat{\phi_{\lambda,t}}(\alpha).$$

This yields

$$\sum_{k\in\mathbb{Z}}p_k\phi(\lambda k)\mathrm{e}^{2\pi ikt}=\frac{1}{\lambda}\sum_{\alpha\in\Lambda}x_\alpha\widehat{\phi}\Big(\frac{\alpha-t}{\lambda}\Big). \tag{5}$$

Substituting $1/\lambda$ for λ we can rewrite the equation as

$$\lambda^{-1}\sum_{k\in\mathbb{Z}}p_k\phi(\lambda^{-1}k)\mathrm{e}^{2\pi ikt}=\sum_{\alpha\in\mathbb{R}}x_\alpha\widehat{\phi}\,(\lambda(\alpha-t))\,, \tag{6}$$

where $x_\alpha = 0$ for $\alpha \notin \Lambda$. Observe that because all x_α are bounded and Λ is uniformly discrete we have

$$\lim_{\lambda\to\infty} \sum_{\alpha\in\mathbb{R}} x_\alpha \widehat{\phi}\,(\lambda(\alpha - t)) = x_t \widehat{\phi}(0).$$

However, the right-hand side of Eq. (6) has

$$\lambda^{-1} \sum_{k\in\mathbb{Z}} p_k \phi(\lambda^{-1}k) \mathrm{e}^{2\pi i k t_1} = \lambda^{-1} \sum_{k\in\mathbb{Z}} p_k \phi(\lambda^{-1}k) \mathrm{e}^{2\pi i k t_2}$$

for any t_1, t_2 with $t_1 - t_2 \in \mathbb{Z}$. By choosing ϕ such that $\widehat{\phi}(0) = \int_{\mathbb{R}} \phi \neq 0$ it follows that $x_{t_1} = x_{t_2}$ whenever $t_1 - t_2 \in \mathbb{Z}$. Thus Λ must be the union of equivalent classes modulo $\mathbb{Z}$, i.e., cosets of $\mathbb{Z}$. Being uniformly discrete Λ can only be a finitely union of cosets of $\mathbb{Z}$. Hence there exist $\alpha_1, \ldots, \alpha_m \in [0, 1)$ such that $\Lambda = \bigcup_{j=1}^{m} (\alpha_j + \mathbb{Z})$. Furthermore, $x_\beta = c_j$ for all $\beta \in \alpha_j + \mathbb{Z}$. Finally (C) follows directly from taking the Fourier transform of μ and the Poisson summation formula

$$\widehat{\sum_{k\in\mathbb{Z}} \delta_{k+\alpha}} = \sum_{k\in\mathbb{Z}} \mathrm{e}^{2\pi i \alpha k} \delta_k. \qquad \square$$

Remark 1. The condition $\mathrm{supp}(\widehat{\mu}) \subset \mathbb{Z}$ in the theorem can be replaced with $\mathrm{supp}(\widehat{\mu}) \subset \Gamma$ for some lattice Γ. In this setting the theorem still holds if the set $\mathbb{Z}$ in (A) and (B) is replaced by the dual lattice Γ^* of Γ, and the $\mathbb{Z}$ in (C) is replaced by Γ.

Remark 2. A theorem of Cordoba [4] draws the same conclusions under the hypotheses that $\mathrm{supp}(\widehat{\mu})$ is a uniformly discrete set but requires that the set $\{x_\beta : \beta \in \Lambda\}$ is finite.

Theorem 3.3. *Let $\mathbf{a} = (a_k)$ be finitely supported. Suppose $T_{\mathbf{a}} \neq c\, I$ where I is the identity map. Then λ is an eigenvalue of $T_{\mathbf{a}}$ if and only if $\lambda \in \{\widehat{\mathbf{a}}(\xi) : \xi \in [0, 1)\}$. Furthermore $\mathbf{x} \in l^\infty(\mathbb{Z})$ is an eigenvector of $T_{\mathbf{a}}$ for the eigenvalue λ if and only if*

$$\mathbf{x} = \sum_{\theta\in Z_{\mathbf{a},\lambda}} c_\theta \mathbf{v}_\theta \tag{7}$$

for some constants c_θ, where $Z_{\mathbf{a},\lambda} := \{t \in [0, 1) : \widehat{\mathbf{a}}(t) = \lambda\}$.

Proof. For any $\lambda = \widehat{\mathbf{a}}(t)$ it is easy to check that $\mathbf{v}_t$ is an eigenvector of $T_{\mathbf{a}}$. Let λ be an eigenvalue of $T_{\mathbf{a}}$ with $T_{\mathbf{a}}(\mathbf{x}) = \lambda\mathbf{x}$ for some nonzero $\mathbf{x} \in l^\infty(\mathbb{Z})$. Observe that $\mathscr{F}^{-1}(\mu_{T_{\mathbf{a}}(\mathbf{x})}) = \widehat{\mathbf{a}}\mathscr{F}^{-1}(\mu_{\mathbf{x}})$. Thus

$$\widehat{\mathbf{a}}\mathscr{F}^{-1}(\mu_{\mathbf{x}}) = \lambda\mathscr{F}^{-1}(\mu_{\mathbf{x}}), \qquad \text{and} \qquad (\widehat{\mathbf{a}} - \lambda)\mathscr{F}^{-1}(\mu_{\mathbf{x}}) = 0.$$

It follows that $\mathrm{supp}(\mathscr{F}^{-1}(\mu_{\mathbf{x}})) \subseteq Z_{\mathbf{a},\lambda} + \mathbb{Z}$. Thus $\lambda \in \{\widehat{\mathbf{a}}(\xi) : \xi \in [0,1)\}$, and because $T_{\mathbf{a}} \neq c\,I$ the set $Z_{\mathbf{a},\lambda}$ is finite. Hence $Z_{\mathbf{a},\lambda} + \mathbb{Z}$ is uniformly discrete. Lemma 3.1 implies that

$$\mathscr{F}^{-1}(\mu_{\mathbf{x}}) = \sum_{\alpha \in Z_{\mathbf{a},\lambda}+\mathbb{Z}} b_\alpha \delta_\alpha$$

for some bounded sequence(b_α). Theorem 3.2 now applies to $\mathscr{F}^{-1}(\mu_{\mathbf{x}})$ to show that

$$\mathscr{F}^{-1}(\mu_{\mathbf{x}}) = \sum_{\theta \in Z_{\mathbf{a},\lambda}} c_\theta \sum_{k \in \mathbb{Z}} \delta_{\theta+k}.$$

The structure of $\mu_{\mathbf{x}}$ now follows from part (C) of Theorem 3.2, which yields

$$\mathbf{x} = \sum_{\theta \in Z_{\mathbf{a},\lambda}} c_\theta \mathbf{v}_\theta.$$

□

Corollary 3.4. *For any finitely supported* $\mathbf{a} = (a_k)$ *the operator* $T_{\mathbf{a}}$ *has no point spectrum in* $l^p(\mathbb{Z})$ *for any* $1 \leq p < \infty$ *unless* $T_{\mathbf{a}} = c\,I$.

Proof. Clearly any eigenvector for $T_{\mathbf{a}}$ in $l^p(\mathbb{Z})$ is also an eigenvector in $l^\infty(\mathbb{Z})$ for the same eigenvalue. If $T_{\mathbf{a}} \neq c\,I$ then by Theorem 3.3, all eigenvectors of $T_{\mathbf{a}}$ in $l^\infty(\mathbb{Z})$ are of the form (7), which do not belong to $l^p(\mathbb{Z})$ for an $1 \leq p < \infty$. This is easily seen from the fact that such $\mathbf{x}$ are almost periodic so the entries do not tend to 0. Thus $T_{\mathbf{a}}$ has no point spectrum in $l^p(\mathbb{Z})$ for $1 \leq p < \infty$. □

4 Proof of Main Theorem

In this section we assume the hypotheses of Theorem 2.1 and prove the theorem by breaking it down into a series of lemmas and estimates. Without loss of generality we assume that $\mathbf{a} = (a_k)$ is symmetric and $a_k = 0$ for $k > q$ or $k < -q$, i.e., $\mathrm{supp}(\mathbf{a}) \subset [-q, q]$. To prove the theorem it suffices to prove that $\lim_{n\to\infty} T_{\mathbf{a}}^n(\mathbf{x})(0) = \sum_{\theta \in Z_{\mathbf{a},1}} [\mathbf{x}, \mathbf{v}_\theta]$.

Lemma 4.1. *Let* $\mathbb{T} = \mathbb{R}/\mathbb{Z}$. *Then*

$$\lim_{n\to\infty} T_{\mathbf{a}}^n(\mathbf{x})(0) = \int_{\mathbb{T}} \widehat{\mathbf{a}}^n(\xi)\widehat{\mathbf{x}_{qn}}(\xi)\,\mathrm{d}\xi \tag{8}$$

and

$$\lim_{n\to\infty} \left(T_{\mathbf{a}}^n(\mathbf{x})(0) - \sum_{\theta \in Z_{\mathbf{a},1}} \int_{|\xi-\theta|<\delta} \widehat{\mathbf{a}}^n(\xi)\widehat{\mathbf{x}_{qn}}(\xi)\,\mathrm{d}\xi \right) = 0 \tag{9}$$

for any $\delta > 0$ such that the intervals $\{(\theta - \delta, \theta + \delta) : \theta \in Z_{\mathbf{a},1}\}$ in $\mathbb{T}$ are disjoint.
Proof. Note that $\widehat{\mathbf{a}}^n$ is a trigonometric polynomial of degree qn, $T^n_{\mathbf{a}}(\mathbf{x})(0)$ is the constant term of $\widehat{\mathbf{a}}^n(-\xi)\widehat{\mathbf{x}}_{qn}(\xi)$. Integrating it over $\mathbb{T}$ yields $T^n_{\mathbf{a}}(\mathbf{x})(0)$. Equation (8) follows from the fact that $\widehat{\mathbf{a}^n}(-\xi) = \widehat{\mathbf{a}^n}(\xi)$.

To prove Eq. (9) we observe that

$$(T^n_{\mathbf{a}}(\mathbf{x})(0) - \sum_{\theta\in Z_{\mathbf{a},1}} \int_{|\xi-\theta|<\delta} \widehat{\mathbf{a}^n}(\xi)\widehat{\mathbf{x}_{qn}}(\xi)\,\mathrm{d}\xi = \int_E \widehat{\mathbf{a}^n}(\xi)\widehat{\mathbf{x}_{qn}}(\xi)\,\mathrm{d}\xi,$$

where $|\xi - \theta| \geq \delta$ on E for any $\theta \in Z_{\mathbf{a},1}$. Thus there exists an $\varepsilon > 0$ such that $|\widehat{\mathbf{a}}(\xi)| \leq 1 - \varepsilon$ on E. Also $|\widehat{\mathbf{x}}_{qn}(\xi)| \leq \|\mathbf{x}\|_\infty qn$, so

$$\lim_{n\to\infty} \left| \int_E \widehat{\mathbf{a}^n}(\xi)\widehat{\mathbf{x}_{qn}}(\xi)\,\mathrm{d}\xi \right| \leq \lim_{n\to\infty} (1-\varepsilon)^n \|\mathbf{x}\|_\infty qn = 0. \tag{10}$$

□

Throughout this section we shall assume that $\delta > 0$ is small enough so that $\{(\theta - \delta, \theta + \delta) : \theta \in Z_{\mathbf{a},1}\}$ in $\mathbb{T}$ are disjoint. Our next step shows that with small enough $\delta > 0$, for any $\varepsilon > 0$, the estimate

$$\left| \int_{|\xi-\theta|<\delta} \widehat{\mathbf{a}^n}(\xi)\widehat{\mathbf{x}_{qn}}(\xi)\,\mathrm{d}\xi - [\mathbf{x}, \mathbf{v}_\theta] \right| < \varepsilon \tag{11}$$

holds for sufficiently large n. This is achieved by performing a series of delicate estimates. Obviously Theorem 2.1 follows readily from Eq. (11).

We now fix any $\theta \in Z_{\mathbf{a},1}$. Note that $\widehat{\mathbf{a}} \leq 1$ so $\widehat{\mathbf{a}}'(\theta) = 0$ and

$$\widehat{\mathbf{a}}(\xi) = 1 - c_\theta(\xi - \theta)^{2m} + O((\xi - \theta)^{2m+1})$$

near θ, where $c_\theta > 0$. It follows that $\widehat{\mathbf{a}}(\theta + t) = 1 - c_\theta t^{2m} + h_\theta(t)$ where $h_\theta(t) = O(t^{2m+1})$ is bounded and

$$\int_{|\xi-\theta|<\delta} \widehat{\mathbf{a}^n}(\xi)\widehat{\mathbf{x}_{qn}}(\xi)\,\mathrm{d}\xi = \int_{-\delta}^{\delta} \widehat{\mathbf{a}^n}(\theta + t)\widehat{\mathbf{x}_{qn}}(\theta + t)\,\mathrm{d}t = A(n, \theta, \delta) + B(n, \theta, \delta),$$

where

$$A(n, \delta, \theta) = \int_{-\delta}^{\delta} (1 - c_\theta t^{2m})^n \widehat{\mathbf{x}_{qn}}(\theta + t)\,\mathrm{d}t, \tag{12}$$

$$B(n, \delta, \theta) = \int_{-\delta}^{\delta} \left(\widehat{\mathbf{a}}(\theta + t)^n - (1 - c_\theta t^{2m})^n\right) \widehat{\mathbf{x}_{qn}}(\theta + t)\,\mathrm{d}t. \tag{13}$$

We first prove that $\lim_{n\to\infty} B(n, \delta, \theta) = 0$. To do so we study the term $\widehat{\mathbf{a}}(\theta + t)^n - (1 - c_\theta t^{2m})^n$ on $[0, \delta]$.

Lemma 4.2. *Let $F_n(t) = (1 - t^k + h(t))^n - (1 - t^k)^n$ where $k \geq 2$ and $h(t) = o(t^k)$ is analytic and nonzero. Let $\delta > 0$ be sufficiently small. Then for sufficiently large n the function F_n has only one critical point $t_n \in (0, \delta]$, at which $|F(t_n)| = \max_{t\in[0,\delta]} |F_n(t)| \leq Cn^{-\frac{1}{k}}$.*

Proof. Let $f(t) = 1-t^k+h(t)$ and $g(t) = 1-t^k$. Since h is analytic and nonzero, we have $h(t) = Kt^m + O(t^{m+1})$ where $m > k$ and $K \neq 0$. Note that $F_n(0) = 0$ and $F_n(\delta)\to 0$ exponentially. But $F_n(n^{-\frac{1}{k}})$ does not go to 0 exponentially. Hence F_n has at least one critical point in $(0, \delta]$ for sufficiently large n. Let t_n be a critical point, i.e., $F_n'(t_n) = 0$. It follows that $f^{n-1}(t_n)f'(t_n) - g^{n-1}(t_n)g'(t_n) = 0$, and thus

$$\frac{g^{n-1}(t_n)}{f^{n-1}(t_n)} = \frac{f'(t_n)}{g'(t_n)} = 1 - \frac{Km}{k}t_n^{m-k} + O(t_n^{m-k+1}). \tag{14}$$

□

Claim. $\lim_{n\to\infty}(n-1)t_n^k = m/k$.

It is clear that if $t_{n_j}\to\varepsilon$ where $\varepsilon > 0$ then $\frac{g^{n_j-1}(t_{n_j})}{f^{n_j-1}(t_{n_j})}\to 0$ when $K > 0$, the limit is $+\infty$ when $K < 0$. This is a contradiction. Hence $t_n\to 0$. Now observe that $g(t_n)/f(t_n) = 1 - Kt_n^m + O(t_n^{m+1})$. By taking logarithm on both sides of Eq. (14) we obtain

$$-(n-1)Kt_n^m + O((n-1)t_n^{m+1}) = -\frac{Km}{k}t_n^{m-k} + O(t_n^{m-k+1}).$$

The claim $\lim_{n\to\infty}(n-1)t_n^k = m/k$ now follows.

We next show that this t_n is unique for sufficiently large n. This is done by the sign of $F''(t_n)$.

$$F_n'' = n\left(f''f^{n-1} - g''g^{n-1}\right) + n(n-1)\left(f'^2f^{n-2} - g'^2g^{n-2}\right).$$

Thus

$$\frac{F_n''}{nf^{n-2}} = \left(f''f - g''\frac{g^{n-1}}{f^{n-2}}\right) + (n-1)\left(f'^2 - g'^2\frac{g^{n-2}}{f^{n-2}}\right).$$

At $t = t_n$ we have $\frac{g^{n-1}}{f^{n-1}} = \frac{f'}{g'}$. It is not hard to verify that this yields

$$\frac{F_n''}{nf^{n-2}} = fg'\left(\frac{f'}{g'}\right)' + (n-1)f'g\left(\frac{f}{g}\right)' \qquad \text{at} \quad t = t_n.$$

One can also check easily that at $t = t_n$,

$$fg'(f'/g')' = Km(m-k)t_n^{m-2} + O(t_n^{m-1}),$$
$$f'g(f/g)' = -Km(m-k)t_n^{m+k-2} + O(t_n^{m+k-1}).$$

Thus by the claim,

$$\lim_{n\to\infty} \frac{F_n''(t_n)}{n\, t_n^{l-2} g^{n-2}(t_n)} = Km(m-k) - \lim_{n\to\infty} Kmk(n-1)t_n^k = -Kmk.$$

If $K > 0$, we have $F_n(t) \geq 0$ on $[0, \delta]$ and $F_n''(t_n) < 0$, which implies that any critical point t_n is a local maximum for sufficiently large n. But any two local maximum must sandwich a local minimum. Thus there can only be one critical point, at which F_n must achieve its maximum. If $K < 0$, $F_n(t) \leq 0$ on $[0, \delta]$ and $F_n''(t_n) > 0$. By the same reason t_n is the only critical point of F_n and it is the minimum. Thus in either case we must have $|F(t_n)| = \max_{t\in[0,\delta]} |F_n(t)|$. Finally,

$$|F_n(t)| \leq (1 - t^k)^n \left((1 + K_1 t^m)^n - 1\right)$$

for some $K_1 > 0$ on $[0, \delta]$. It follows from $\lim_{n\to\infty} nt_n^k = m/k$ that $(1 - t_n^m)^n \to \mathrm{e}^{\frac{m}{k}}$ and $(1 + K_1 t_n^m)^n - 1 = O(t_n^{m-k}) = O(n^{-\frac{m-k}{k}})$. This proves the lemma.

Lemma 4.3. $\lim_{n\to\infty} B(n, \delta, \theta) = 0$.

Proof. Let $F_n(t) = \widehat{\mathbf{a}}(\theta + t)^n - (1 - c_\theta t^{2m})^n$. Without loss of generality we may assume that $c_\theta = 1$. Then F_n satisfies the hypothesis of Lemma 4.2. Let t_n be the unique critical point of F_n on $(0, \delta]$. So F_n is monotone on $[0, t_n]$ and $[t_n, \delta]$. A special mean value theorem for integration (see, e.g., Bartle [1], Theorem 30.11) now implies that for some $\eta \in [0, \delta]$,

$$\int_0^{t_n} F_n(t)\widehat{\mathbf{x}_{qn}}(\theta + t)\,\mathrm{d}t = F_n(0)\int_0^{\eta} \widehat{\mathbf{x}_{qn}}(\theta + t)\,\mathrm{d}t + F_n(t_n)\int_{\eta}^{t_n} \widehat{\mathbf{x}_{qn}}(\theta + t)\,\mathrm{d}t.$$

Note that $F_n(0) = 0$ and

$$\left|\int_{\eta}^{\delta} \widehat{\mathbf{x}_{qn}}(\theta + t)\,\mathrm{d}t\right| \leq C_1 \sum_{j=1}^{qn} \frac{1}{j} = O(\ln n).$$

It follows that

$$\left|\int_0^{t_n} F_n(t)\widehat{\mathbf{x}_{qn}}(\theta + t)\,\mathrm{d}t\right| = O(n^{-\frac{1}{2m}} \ln n).$$

By the same token,

$$\left|\int_{t_n}^{\delta} F_n(t)\widehat{\mathbf{x}_{qn}}(\theta + t)\,\mathrm{d}t\right| = O(n^{-\frac{1}{2m}} \ln n).$$

Thus $\int_0^{\delta} F_n(t)\widehat{\mathbf{x}_{qn}}(\theta + t)\,\mathrm{d}t \longrightarrow 0$. Similarly $\int_{-\delta}^{0} F_n(t)\widehat{\mathbf{x}_{qn}}(\theta + t)\,\mathrm{d}t \longrightarrow 0$. These combine to yield $\lim_{n\to\infty} B(n, \delta, \theta) = 0$. □

Lemma 4.4. *Assume that $b, k > 0$ and $p \geq 0$. For any $\epsilon > 0$ such that $b\epsilon^k < 1$ we have*

$$\int_0^\epsilon (1 - bt^k)^n t^p \, \mathrm{d}t \leq \min\Big\{\frac{1}{p+1}\epsilon^{p+1}, C\, n^{-\frac{p+1}{k}}\Big\},$$

where $C = \int_0^\infty \mathrm{e}^{-bs^k} s^p \, \mathrm{d}s$.

Proof. Using the fact that $1 - x \leq \mathrm{e}^{-x}$ for all $x \geq 0$, we have $(1 - bt^k)^n \leq \mathrm{e}^{-nbt^k}$. Making the substitution $s = \sqrt[k]{n}t$ we have

$$\int_0^\epsilon (1 - bt^k)^n t^p \, \mathrm{d}t \leq n^{-\frac{p+1}{k}} \int_0^{\sqrt[k]{n}\epsilon} \mathrm{e}^{-bs^k} s^p \, \mathrm{d}s.$$

The lemma follows from two estimates. First, the integral $\int_0^{\sqrt[k]{n}\epsilon} \mathrm{e}^{-bs^k} s^p \, \mathrm{d}s$ is bounded by $C_2 = \int_0^\infty \mathrm{e}^{-bs^k} s^p \, \mathrm{d}s$. Second, it is also bounded by $\frac{1}{p+1}(\sqrt[k]{n}\epsilon)^{p+1}$. □

Next we concentrate on estimating $A(n, \delta, \theta)$. To achieve this, for each $\varepsilon > 0$, we break $\widehat{\mathbf{x}_{qn}}(\theta + t)$ up into three parts:

$$\widehat{\mathbf{x}_{qn}}(\theta + t) = \Big(\sum_{|k| \leq \varepsilon n^\sigma} + \sum_{\varepsilon n^\sigma < |k| \leq \varepsilon^{-1} n^\sigma} + \sum_{\varepsilon^{-1} n^\sigma < |k| \leq qn}\Big) x_k \mathrm{e}^{2\pi i k(\theta + t)} = J_1 + J_2 + J_3,$$

where $\sigma = \frac{1}{2m}$. As a result we write $A(n, \delta, \theta) = A_1(n, \delta, \theta, \varepsilon) + A_2(n, \delta, \theta, \varepsilon) + A_3(n, \delta, \theta, \varepsilon)$, where

$$A_j(n, \delta, \theta, \varepsilon) = \int_{-\delta}^{\delta} (1 - c_\theta t^{2m})^n J_j(\theta + t) \, \mathrm{d}t, \quad j = 1, 2, 3. \tag{15}$$

Note that here all $J_j(\theta + t)$ depend on n, ε, but for simplicity of notations, we keep the dependence in the background.

Lemma 4.5. *Let $\varepsilon > 0$. Then $|A_1(n, \delta, \theta, \varepsilon)| \leq C_1\sqrt{\varepsilon}$ for sufficiently large n, where $C_1 > 0$ is independent of n.*

Proof. By the Cauchy–Schwartz inequality we have

$$|A_1(n, \delta, \theta, \varepsilon)|^2 \leq \int_{-\delta}^{\delta} (1 - c_\theta t^{2m})^{2n} \, \mathrm{d}t \int_{-\delta}^{\delta} |J_1(\theta + t)|^2 \, \mathrm{d}t.$$

Using the orthogonality of $\mathrm{e}^{2\pi i k t}$ on $\mathbb{T}$, we have

$$\int_{-\delta}^{\delta} |J_1(\theta + t)|^2 \, \mathrm{d}t \leq \int_{-\frac{1}{2}}^{\frac{1}{2}} |J_1(\theta + t)|^2 \, \mathrm{d}t \leq 4\varepsilon n^\sigma \|\mathbf{x}\|_\infty^2.$$

Also by Lemma 4.4, $\int_{-\delta}^{\delta} (1 - c_\theta t^{2m})^{2n} \, \mathrm{d}t \leq C n^{-\sigma}$. The lemma now follows. □

Lemma 4.6. *Let $\varepsilon > 0$. Then $|A_3(n,\delta,\theta,\varepsilon)| \le C_3\varepsilon$ for sufficiently large n, where $C_3 > 0$ is independent of n.*
Proof. We first establish the inequality

$$\left| \int_{-\delta}^{\delta} (1 - c_\theta t^{2m})^n \mathrm{e}^{2\pi i k t}\, \mathrm{d}t \right| = 2 \left| \int_0^{\delta} (1 - c_\theta t^{2m})^n \cos(2\pi k t)\, \mathrm{d}t \right| \le \frac{C_3' n^\sigma}{k^2} \quad (16)$$

for all $|k| > \varepsilon^{-1} n^\sigma$, where again $\sigma = \frac{1}{2m}$. The substitution $s = n^\sigma t$ yields

$$\int_0^{\delta} (1 - c_\theta t^{2m})^n \cos(2\pi k t)\, \mathrm{d}t = \frac{1}{n^\sigma} \int_0^{\delta n^\sigma} g_n(s) \cos(Ls)\, \mathrm{d}s \quad (17)$$

where $L = \frac{2\pi k}{n^\sigma}$ and $g_n(s) = (1 - \frac{c_\theta s^{2m}}{n})^n$. Again, $1 - \frac{c_\theta s^{2m}}{n} \le \mathrm{e}^{-\frac{c_\theta s^{2m}}{n}}$ so $g_n(s) \le \mathrm{e}^{-c_\theta s^{2m}}$. Observe that we have $g_n(\delta n^\sigma) = O(\mathrm{e}^{-c_\theta \delta^{2m} n})$, $g_n'(0) = 0$, and $g_n'(\delta n^\sigma) = O(n\mathrm{e}^{-c_\theta \delta^{2m} n})$. Combining these with integration by parts twice on Eq. (17) we obtain

$$\int_0^{\delta n^\sigma} g_n(s) \cos(Ls)\, \mathrm{d}s = O(n\mathrm{e}^{-c_\theta \delta^2 n}) - \frac{1}{L^2} \int_0^{\delta n^\sigma} g_n''(s) \cos(Ls)\, \mathrm{d}s.$$

It is easy to check that $|g_n''(s)\cos(Ls)| \le (a_1 s^{2m-2} + a_2 s^{4m-2})\mathrm{e}^{-c_\theta s^{2m}}$ for some constants $a_1, a_2 > 0$. Thus $\int_0^{\delta n^\sigma} g_n''(s)\cos(Ls)\,\mathrm{d}s$ is bounded by $\int_0^\infty (a_1 s^{2m-2} + a_2 s^{2m-2})\mathrm{e}^{-c_\theta s^{2m}}\,\mathrm{d}s$, which is finite. Hence there exists $C_3'' > 0$ such that

$$\left| \int_0^{\delta n^\sigma} g_n(s) \cos(Ls)\, \mathrm{d}s \right| \le \frac{C_3''}{L^2} = \frac{C_3'' n^{2\sigma}}{4\pi^2 k^2},$$

which yields Eq. (16). Finally by Eq. (16),

$$|A_3(n,\delta,\theta,\varepsilon)| \le C_3' \|\mathbf{x}\|_\infty \sum_{\varepsilon^{-1} n^\sigma < |k| \le qn} \frac{n^\sigma}{k^2} \le C_3 \varepsilon. \qquad \square$$

Lemma 4.7. *Assume that $[\mathbf{x}, \mathbf{v}_{-\theta}] = 0$. Let $\varepsilon > 0$. Then $|A_2(n,\delta,\theta,\varepsilon)| \le C_2\varepsilon$ for sufficiently large n, where $C_2 > 0$ is independent of n.*
Proof. Set $\mathbf{y} = (y_k) := (x_k \mathrm{e}^{2\pi i k\theta})_{k\in\mathbb{Z}}$. Then $\widehat{\mathbf{y}_{qn}}(t) = \widehat{\mathbf{x}_{qn}}(\theta + t)$. By the fact that $\int_{-\delta}^{\delta} (1 - c_\theta t^{2m})^n \sin(2\pi k t)\, \mathrm{d}t = 0$,

$$\begin{aligned} A_2(n,\delta,\theta,\varepsilon) &= \sum_{\varepsilon n^\sigma < |k| \le \varepsilon^{-1} n^\sigma} y_k \int_{-\delta}^{\delta} (1 - c_\theta t^{2m})^n \mathrm{e}^{2\pi i k t}\, \mathrm{d}t \\ &= 2 \sum_{\varepsilon n^\sigma < k \le \varepsilon^{-1} n^\sigma} (y_k + y_{-k}) \int_0^{\delta} (1 - c_\theta t^{2m})^n \cos(2\pi k t)\, \mathrm{d}t. \end{aligned}$$

Now denote $S_k := \sum_{j=-k}^{k} y_j$ and $U_k = \int_0^\delta (1 - c_\theta t^{2m})^n \cos(2\pi k t)\,\mathrm{d}t$. Then $y_k + y_{-k} = S_k - S_{k-1}$. Using summation by parts

$$A_2(n, \delta, \theta, \varepsilon) = \sum_{k=M}^{N} (S_k - S_{k-1}) U_k = \sum_{k=M}^{N-1} S_k\,(U_k - U_{k+1}) - S_{M-1}U_M + S_N U_N$$

where $M = \lfloor \varepsilon n^\sigma \rfloor + 1$ and $N = \lfloor \varepsilon^{-1} n^\sigma \rfloor$. Using the fact $|S_k| \le a_1 k$ for some constant a_1 and Lemma 4.4 we have

$$|S_{M-1}U_M| \le a_1 \varepsilon n^\sigma \int_0^\delta (1 - c_\theta t^{2m})^n\,\mathrm{d}t \le a_1' \varepsilon.$$

By Eq. (16) there exists some $a_2 > 0$ such that

$$|S_N U_N| \le a_1 C_3' N \frac{n^\sigma}{N^2} \le a_2 \varepsilon^{-1} n^\sigma \frac{n^\sigma}{(\varepsilon^{-1} n^\sigma)^2} = a_2 \varepsilon.$$

It remains to estimate $T := \sum_{k=M}^{N-1} S_k\,(U_k - U_{k+1})$. The hypothesis $[\mathbf{x}, \mathbf{v}_{-\theta}] = 0$ implies that $\lim_{k\to\infty} S_k/k = 0$. Thus for $n > N_0$ we have $\sup_{k \ge \varepsilon n^\sigma} |S_k/k| \le \varepsilon^3$. It follows from the Cauchy–Schwartz inequality that

$$|T|^2 = \Big| \sum_{k=M}^{N-1} \frac{S_k}{k}\, k\,(U_k - U_{k+1}) \Big|^2 \le \varepsilon^6 \Big(\sum_{k=M}^{N-1} k^2 \Big) \Big(\sum_{k=M}^{N-1} (U_k - U_{k+1})^2 \Big).$$

Now $U_k - U_{k+1} = \int_0^\delta (1 - c_\theta t^{2m})^n \sin(\pi t) \sin(\pi(2k+1)t)\,\mathrm{d}t$. Observe that the functions $\{\sqrt{2}\sin(\pi(2k+1)t)\}$ are orthonormal on $[0, 1]$. Parseval's inequality yields

$$\sum_{k=M}^{N-1} (U_k - U_{k+1})^2 \le \frac{1}{2} \int_0^\delta (1 - c_\theta t^{2m})^{2n} \sin^2(\pi t)\,\mathrm{d}t \le \frac{\pi^2}{2} \int_0^\delta (1 - c_\theta t^{2m})^{2n} t^2\,\mathrm{d}t. \tag{18}$$

By Lemma 4.4

$$\int_0^\delta (1 - c_\theta t^{2m})^{2n} t^2\,\mathrm{d}t \le C n^{-\frac{3}{2m}} = C n^{-3\sigma}.$$

Thus

$$|T|^2 \le a_3 \varepsilon^6 (\varepsilon^{-1} n^\sigma)^3 n^{-3\sigma} = a_3 \varepsilon^3.$$

These estimates show that for sufficiently large n we have

$$|A_2(n, \delta, \theta, \varepsilon)| \le C_2 \varepsilon. \qquad \square$$

We can now complete the proof of Theorem 2.1. Let $\tilde{\mathbf{x}} := \mathbf{x} - \sum_{\theta \in Z_{\mathbf{a},1}} [\mathbf{x}, \mathbf{v}_\theta]\mathbf{v}_\theta$. Then $T^n_{\mathbf{a}}(\mathbf{x}) = T^n_{\mathbf{a}}(\tilde{\mathbf{x}}) + \sum_{\theta \in Z_{\mathbf{a},1}} [\mathbf{x}, \mathbf{v}_\theta]\mathbf{v}_\theta$. Note that $\tilde{\mathbf{x}}$ satisfies the hypothesis of Lemma 4.7. Combining Lemma 4.3 and Lemmas 4.5–4.7 yields $T^n_{\mathbf{a}}(\tilde{\mathbf{x}})(0) \to 0$. Thus

$$\lim_{n\to 0} T^n_{\mathbf{a}}(\mathbf{x})(0) = \sum_{\theta \in Z_{\mathbf{a},1}} [\mathbf{x}, \mathbf{v}_\theta]\mathbf{v}_\theta(0) = \sum_{\theta \in Z_{\mathbf{a},1}} [\mathbf{x}, \mathbf{v}_\theta].$$

Finally, let τ be the left shift operator on $l^\infty(\mathbb{Z})$, i.e., $\tau((x_k)) = (x_{k+1})$. Then $T_{\mathbf{a}} \circ \tau = \tau \circ T_{\mathbf{a}}$. It follows that

$$T^n_{\mathbf{a}}(\mathbf{x})(k) = \tau^k \circ T^n_{\mathbf{a}}(\mathbf{x})(0) = T^n_{\mathbf{a}}\left(\tau^k(\mathbf{x})\right)(0).$$

But $[\tau^k(\mathbf{x}), \mathbf{v}_\theta] = [\mathbf{x}, \mathbf{v}_\theta]\mathrm{e}^{2\pi i k\theta}$ for $\theta \in Z_{\mathbf{a},1}$. Thus

$$T^n_{\mathbf{a}}(\mathbf{x})(k) = T^n_{\mathbf{a}}\left(\tau^k(\mathbf{x})\right)(0) = \sum_{\theta \in Z_{\mathbf{a},1}} [\mathbf{x}, \mathbf{v}_\theta]\mathrm{e}^{2\pi i k\theta}.$$

This completes the proof of Theorem 2.1.

Remark. Lemma 4.7 is the only place where the condition $[\mathbf{x}, \mathbf{v}_\theta]$ exists for all $\theta \in Z_{\mathbf{a},1}$ is being used. With this condition we may apply summation by parts and the convergence of S_k/k to obtain the necessary final estimates. It is also clear from the proof that we can apply summation by parts again to show the following: Let $S_k(\theta) := \sum_{j=-k}^{k} x_j \mathrm{e}^{-2\pi j\theta}$ and $S'_k(\theta) := \sum_{1\le|j|\le k} S_j(\theta)/j$. Assume that $\lim_{k\to\infty} S'_k(\theta)/k$ exists for every $\theta \in Z_{\mathbf{a},1}$ then the conclusion of the theorem still holds. Unfortunately the convergence of $S'_k(\theta)/k$ is equivalent to the convergence of $S_k(\theta)/k$. We shall omit the proof here.

References

1. Bartle, R.: The Elements of Real Analysis, 2nd edn. Wiley, New York (1976)
2. Böttcher, A., Grudsky, S.: Toeplitz Matrices, Asymptotic Linear Algebra, and Functional Analysis. Birkhäuser (2000)
3. Chen, Q., Huang, N., Riemenschneider, S., Xu, Y.: B-spline approach for empirical mode decomposition, Adv. Comput. Math. **24**, 171–195 (2006)
4. Cordoba, A.: Dirac combs. Lett. Math. Phys. **17**, 191–196 (1989)
5. Echeverria, J.C., Crowe, J.A., Woolfson, M.S., Hayes-Gill, B.R.: Application of empirical mode decomposition to heart rate variability analysis. Med. Biol. Eng. Comput. **39**, 471–479 (2001)
6. Folland, G.: Real Analysis. Modern techniques and their application. 2nd ed. John Wiley and Sons Inc., New York, (1999)
7. Hörmander, L.: The Analysis of Linear Partial Differential Operators I. Springer (1983)
8. Huang, N., et al.: The empirical mode decomposition and the Hilbert spectrum for nonlinear nonstationary time series analysis. Proceedings of Royal Society of London A **454**, 903–995 (1998)

9. Huang, N., Shen, Z., Long, S.: A new view of nonlinear water waves: the Hilbert spectrum. Annu. Rev. Fluid Mech. **31**, 417–457 (1999)
10. Hughes, J., Mao, D., Rockmore, D., Wang, Y., Wu, Q.: Empirical mode decomposition analysis of visual stylometry, preprint
11. Lagarias, J.: Mathematical quasicrystals and the problem of diffraction. In: Baake, M., Moody, R.V. (eds.) Directions in Mathematical Quasicrystals, CRM Monograph Series, Amer. Math. Soc., vol. 13, pp. 61–93. Providence, RI (2000)
12. Lin, L., Wang, Y., Zhou, H.: Iterative filtering as an alternative algorithm for empirical mode decomposition. Adv. Adapt. Data Anal. **1**(4), 543–560 (2009)
13. Liu, B., Riemenschneider, S., Xu, Y.: Gearbox fault diagnosis using empirical mode decomposition and hilbert spectrum, preprint
14. Mao, D., Wang, Y., Wu, Q.: A new approach for analyzing physiological time series, preprint
15. Pines, D., Salvino, L.: Health monitoring of one dimensional structures using empirical mode decomposition and the Hilbert-Huang Transform. In: Proceedings of SPIE 4701, pp. 127–143 (2002)
16. Yu, Z.-G., Anh, V., Wang, Y., Mao, D.: Modeling and simulation of the horizontal component of the magnetic field by fractional stochastic differential equation in conjunction with epirical mode decomposition. J. Geophys. Res. Space Phys. to appear

Wavelet Transforms by Nearest Neighbor Lifting

Wei Zhu and M. Victor Wickerhauser

Abstract We show that any discrete wavelet transform (DWT) using finite impulse response (FIR) filters may be factored into lifting steps that use only nearest neighbor array elements. We then discuss the advantages and disadvantages of imposing this additional requirement.

Keywords Condition number • Euclid's algorithm • Laurent polynomial • Partial division • Polyphase matrix • Symmetric extension lifting step • Shift matrix • Symmetric division • Z transform

1 Introduction

Our goal is to implement discrete wavelet transforms (DWT) efficiently. The recursive algorithms of Daubechies [3] and Mallat [5] offer an $O(n)$ algorithm for n-point time series. The *lifting* implementation of Daubechies and Sweldens [4] offers an alternative which is also $O(n)$ complex but which only requires about half as many arithmetic operations in the most common cases. Additionally, it acts on the input in such a way that requires just $O(1)$ auxiliary memory.

For DWT on an interval, artifacts may arise at the boundary if the input's periodization from that interval is discontinuous. *Symmetric extension* before periodization, thoroughly described in [1], reduces these artifacts and is easy to include within a lifting implementation.

In this chapter we consider two further enhancements to the lifting method, with or without symmetric extension and periodization:

- **Nearest neighbor lifting** to reduce the number of distant memory accesses.
- **Lifting sequence choice** allowing some utility to be maximized.

W. Zhu • M.V. Wickerhauser (✉)
Washington University in St. Louis, St. Louis, MO, Missouri, USA,
e-mail: zhuwei@math.wustl.edu; victor@math.wustl.edu

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_9,

Nearest neighbors in an input array are elements whose indices differ by one. The corresponding memory locations are thus close enough to ensure that both are very likely to reside in the same physical cache and thus are quick to access. Different but equivalent *lifting sequences* all give the same filter transformation but may have different arithmetic complexity or propagation-of-error properties. In this chapter, we will show how the existence and construction of these enhancements improve the efficiency of DWT.

2 Review of Discrete Wavelet Transforms

Recall that DWT consists of:

- **Signal:** $u \in \ell^2$, in practice finitely supported or periodic.
- **Analysis filters:** linear maps $\tilde{H}, \tilde{G} : \ell^2 \to \ell^2$, composed of convolution and downsampling.
- **DWT:** for integer $J > 0$ levels, filter the signal into a collection of wavelet components

$$u \mapsto \{\tilde{H}^J u;\ \tilde{G}\tilde{H}^{J-1}u,\ \tilde{G}\tilde{H}^{J-2}u,\ \ldots,\ \tilde{G}\tilde{H}u,\ \tilde{G}u\}.$$

- **Synthesis filters:** linear maps $H, G : \ell^2 \to \ell^2$, composed of convolution and resampling and related to $\tilde{H}, \tilde{G}$.
- **Wavelet reconstruction:**

$$\begin{aligned} u &= G\tilde{G}u + H\tilde{H}u \\ &= G\tilde{G}u + H\left(G\tilde{G}\tilde{H}u + H\tilde{H}^2u\right) \\ &= \cdots \\ &= G\tilde{G}u + H\left(G\tilde{G}\tilde{H}u + H\left(\cdots + H\left(G\tilde{G}\tilde{H}^{J-1}u + H\tilde{H}^J u\right)\right)\cdots\right). \end{aligned}$$

An example of DWT to depth $J = 4$ is depicted in Fig. 1. In it the analysis filters are determined by sequences $h \leftrightarrow \tilde{H}$ and $g \leftrightarrow \tilde{G}$, while the synthesis filters are adjoints $H = \tilde{H}^* \leftrightarrow h^*$ and $G = \tilde{G}^* \leftrightarrow g^*$. Reconstruction of u is depicted as moving up and adding.

A *filter* $F : \ell^2 \to \ell^2$ is a linear transformation determined by an absolutely summable sequence $f = \{f_n : n \in \mathbf{Z}\}$:

$$Fx_m = \sum_n f_{2m-n}x_n, \qquad m \in \mathbf{Z}.$$

The *adjoint filter* F^* determined by the same sequence f is

$$F^*x_n = \sum_m \bar{f}_{2m-n}x_m, \qquad n \in \mathbf{Z}.$$

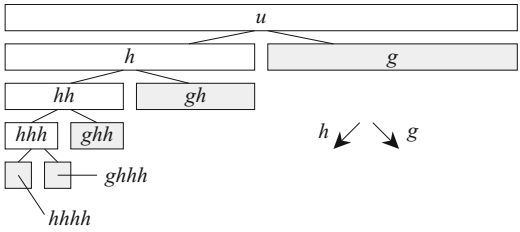


Fig. 1 Four-level discrete wavelet transform with filters h and g

Thus $\langle Fx, y\rangle = \langle x, F^*y\rangle$ for all $x, y \in \ell^2$, using the Hermitian inner product in ℓ^2.

The *conjugate filter* $\dot{F}$ of F has sequence $\dot{f} = \{\dot{f}_n : n \in \mathbf{Z}\}$ defined by

$$\dot{f}_n = (-1)^n f_{1-n} \qquad \Rightarrow \qquad f_n = (-1)^{1-n}\dot{f}_{1-n}, \quad \Rightarrow \ddot{F} = -F.$$

Filter F is called *finite*, equivalently *finite impulse response (FIR)*, if its sequence f is finitely supported. Such filters have a *support interval* $I = [\min S, \max S]$ of finite length $|I|$, where $S = \{n \in \mathbf{Z} : f_n \neq 0\}$. If F is finite then $\dot{F}$ is also finite, with the same support length.

Filter H is called *orthogonal* if it and its conjugate filter $G = \dot{H}$ satisfy the *orthogonality conditions*:

$$HH^* = Id; \quad GG^* = Id; \quad GH^* = HG^* = 0; \quad H^*H + G^*G = Id.$$

Filters H, G form a *perfect reconstruction pair* if they and their conjugates $\tilde{H} = \dot{G}$ and $\tilde{G} = \dot{H}$ satisfy the weaker *biorthogonality conditions*:

$$\tilde{H}H^* = Id; \quad \tilde{G}G^* = Id; \quad \tilde{G}H^* + \tilde{H}G^* = 0; \quad H^*\tilde{H} + G^*\tilde{G} = Id.$$

These may be satisfied for some $G \neq \dot{H}$. However, since $\ddot{H} = -H$ and $\ddot{G} = -G$, any perfect reconstruction pair H, G also satisfies

$$H\tilde{H}^* = Id; \quad G\tilde{G}^* = Id; \quad G\tilde{H}^* + H\tilde{G}^* = 0; \quad \tilde{H}^*H + \tilde{G}^*G = Id.$$

Thus $(\tilde{H}, \tilde{G}) = (\dot{G}, \dot{H})$ form a perfect reconstruction pair whenever (H, G) are a perfect reconstruction pair.

Call H a *perfect reconstruction filter* if there exists a *complement* filter G such that (H, G) is a perfect reconstruction pair. Any orthogonal filter H is evidently a perfect reconstruction filter: we get a perfect reconstruction pair using $G = \dot{H}$ as its complement.

Equivalent perfect reconstruction conditions may be stated for filter sequences. For example, with $H \leftrightarrow h$ and its conjugate $\dot{H} = G \leftrightarrow g$, the orthogonality conditions become:

$$\sum_k h(k)\bar{h}(k+2n) = \mathbf{1}(n) = \sum_k g(k)\bar{g}(k+2n);$$

$$\sum_k g(k)\bar{h}(k+2n) = \quad 0 \quad = \sum_k h(k)\bar{g}(k+2n);$$

and

$$\sum_k h(2k+m)\bar{h}(2k+n) + \sum_k g(2k+m)\bar{g}(2k+n) = \mathbf{1}(n-m),$$

for all $n, m \in \mathbf{Z}$. Here

$$\mathbf{1}(n) = \begin{cases} 1, & \text{if } n = 0, \\ 0, & \text{otherwise.} \end{cases}$$

It is a straightforward exercise to rewrite the remaining conditions for biorthogonal perfect reconstruction pairs in terms of filter sequences.

3 Review of Lifting

Recall the definition of the Z-*transform* of a sequence $x = \{x_n \in \mathbf{C} : n \in \mathbf{Z}\} \in \ell^2$:

$$x(z) = \sum_n x_n z^{-n}, \qquad \text{with} \quad \begin{cases} \textit{even part } x_e(z) \stackrel{\text{def}}{=} \sum_n x_{2n} z^{-n}; \\ \textit{odd part } \ x_o(z) \stackrel{\text{def}}{=} \sum_n x_{2n+1} z^{-n}. \end{cases}$$

We recover the Z-transform of x from the even and odd parts x_e, x_o:

$$x(z) = x_e(z^2) + z^{-1} x_o(z^2).$$

Likewise, we get the even and odd parts from the Z-transform:

$$x_e(z^2) = \frac{x(z) + x(-z)}{2}, \qquad x_o(z^2) = \frac{x(z) - x(-z)}{2z^{-1}}.$$

Any filter F determined by a sequence $\{f_n : n \in \mathbf{Z}\}$ is likewise determined by the Z-transform $f(z) = \sum_n f_n z^{-n}$. We may denote the even and odd parts by $f_e(z)$ and $f_o(z)$, respectively, and write the actions of F and its adjoint F^* on x as pointwise multiplication of Z-transforms:

$$
\begin{aligned}
Fx(z) &= \sum_m Fx_m z^{-m} \;=\; \sum_m \sum_n f_{2m-n} x_n z^{-m} \\
&= \sum_m \sum_n f_{2m-2n} x_{2n} z^{-m} + \sum_m \sum_n f_{2m-2n-1} x_{2n+1} z^{-m} \\
&= \left(\sum_m f_{2m} z^{-m}\right)\left(\sum_n x_{2n} z^{-n}\right) + \left(\sum_m f_{2m-1} z^{-m}\right)\left(\sum_n x_{2n+1} z^{-n}\right) \\
&= \left(\sum_m f_{2m} z^{-m}\right)\left(\sum_n x_{2n} z^{-n}\right) + z^{-1}\left(\sum_m f_{2m+1} z^{-m}\right)\left(\sum_n x_{2n+1} z^{-n}\right) \\
&= f_e(z) x_e(z) + z^{-1} f_o(z) x_o(z);
\end{aligned}
$$

$$
\begin{aligned}
F^* x(z) &= \sum_n F^* x_n z^{-n} \;=\; \sum_n \sum_m \bar{f}_{2m-n} x_m z^{-n} \\
&= \sum_m \sum_n \bar{f}_n x_m z^{-n-2m} \;=\; \left(\sum_n \bar{f}_n z^{-n}\right)\left(\sum_m x_m z^{-2m}\right) \;=\; \bar{f}(z) x(z^2).
\end{aligned}
$$

Remark 1. There is also a "correlation and downsampling" definition of filter and adjoint:

$$
Fx_m = \sum_n f_{2m+n} x_n, \quad m \in \mathbf{Z}; \qquad F^* x_n = \sum_m \bar{f}_{2m+n} x_m, \quad n \in \mathbf{Z}.
$$

Writing this in terms of Z-transforms is straightforward and left to the reader.

There are algebraic relations between the Z-transforms of a filter F and its conjugate $\dot{F}$, respectively denoted by f and $\dot{f}$. Namely,

$$
\dot{f}(z) = -z^{-1} f(-z^{-1}), \qquad \begin{cases} \dot{f}_e(z) = f_o(z^{-1}), \\ \dot{f}_o(z) = -f_e(z^{-1}). \end{cases}
$$

Remark 2. It is possible to generalize to the *M-conjugate* for fixed $M \in \mathbf{Z}$:

$$
\dot{f}_n = (-1)^n f_{2M+1-n} \qquad \Rightarrow \qquad f_n = (-1)^{1-n} \dot{f}_{2M+1-n}, \quad \Rightarrow \ddot{F} = -F
$$

For the M-conjugate of filter F, compute

$$\dot{f}(z) = -z^{-2M-1} f(-z^{-1}), \qquad \begin{cases} \dot{f}_e(z) = z^{-2M} f_o(z^{-1}), \\ \dot{f}_o(z) = -z^{-2M} f_e(z^{-1}). \end{cases}$$

Using the relations just stated, perfect reconstruction conditions for filters may be written in terms of Z-transforms. For filters $H, G, \tilde{H} = \dot{G}, \tilde{G} = \dot{H}$, these become:

$$h(z)\tilde{h}(z^{-1}) + g(z)\tilde{g}(z^{-1}) = 1; \qquad h(z)\tilde{h}(-z^{-1}) + g(z)\tilde{g}(-z^{-1}) = 0.$$

In terms of the even and odd parts:

$$h_e(z)\tilde{h}_e(z^{-1}) + g_e(z)\tilde{g}_e(z^{-1}) = 1; \;\; h_e(z)\tilde{h}_o(z^{-1}) + g_e(z)\tilde{g}_o(z^{-1}) = 0;$$
$$h_o(z)\tilde{h}_o(z^{-1}) + g_o(z)\tilde{g}_o(z^{-1}) = 1; \;\; h_o(z)\tilde{h}_e(z^{-1}) + g_o(z)\tilde{g}_e(z^{-1}) = 0.$$

We now turn our attention to finite filters. If $p \in \ell^2$ is finitely supported, then its Z-transform $p(z)$ is a *Laurent polynomial*:

$$p(z) = \sum_{n=a}^{b} p_n z^{-n}, \qquad a \le b, \quad a, b \in \mathbf{Z}.$$

If $p \not\equiv 0$, then $S = \{n : p_n \neq 0\}$ is a finite nonempty set and the *degree* may be defined by $\deg p = \max S - \min S$, a nonnegative integer one less than the support length of the sequence p.

Laurent polynomials form the commutative ring $\mathbf{C}[z, z^{-1}]$ with multiplicative identity 1. Element $p \not\equiv 0$ is called a *unit*, if and only if p has a multiplicative inverse, if and only if p is a *monomial* $p(z) = Kz^n$ for some constants $K \neq 0$ and $n \in \mathbf{Z}$, and if and only if $\deg p = 0$. Then $p^{-1}(z) = K^{-1}z^{-n}$.

We may also form matrices over the Laurent polynomials. The case we will use is the matrix ring $\mathbf{Mat}\left(2 \times 2, \mathbf{C}[z, z^{-1}]\right)$, with elements:

$$M(z) = \begin{bmatrix} a(z) & b(z) \\ c(z) & d(z) \end{bmatrix}, \qquad a, b, c, d \in \mathbf{C}[z, z^{-1}].$$

M is invertible if and only if $\det M = a(z)d(z) - b(z)c(z)$ is invertible in $\mathbf{C}[z, z^{-1}]$, namely is a nonzero monomial Kz^n. Then

$$M^{-1}(z) = K^{-1}z^{-n} \begin{bmatrix} d(z) & -b(z) \\ -c(z) & a(z) \end{bmatrix}.$$

Now, any pair H, G of finite filters determines a *polyphase matrix*:

$$P(z) = \begin{bmatrix} h_e(z) & g_e(z) \\ h_o(z) & g_o(z) \end{bmatrix}.$$

Likewise, their conjugates $\tilde{H} = \dot{G}, \tilde{G} = \dot{H}$ determine the related polyphase matrix:

$$\tilde{P}(z) = \begin{bmatrix} \tilde{h}_e(z) & \tilde{g}_e(z) \\ \tilde{h}_o(z) & \tilde{g}_o(z) \end{bmatrix} = \begin{bmatrix} \dot{g}_e(z) & \dot{h}_e(z) \\ \dot{g}_o(z) & \dot{h}_o(z) \end{bmatrix}.$$

Both P and $\tilde{P}$ belong to $\mathbf{Mat}(2 \times 2, \mathbf{C}[z, z^{-1}])$. A straightforward calculation now show that the perfect reconstruction condition for (H, G) is equivalent to:

$$P(z)\tilde{P}(z^{-1})^t = Id.$$

Remark 3. In practice, Id may be replaced by any invertible diagonal matrix in $\mathbf{Mat}(2\times 2, \mathbf{C}[z, z^{-1}])$. The two units of $\mathbf{C}[z, z^{-1}]$ appearing on the diagonal will then be monomials Kz^n or multiples by nonzero K of shifts by n indices. The original sequence x is easily reconstructed from such a shifted and multiplied version.

Say that a Laurent polynomial $h(z)$ has a *complement* $g = g(z)$ if the polyphase matrix determined by h, g is invertible. It follows immediately that finite filter H is part of a perfect reconstruction pair, if and only if H has a complement, if and only if its Z-transform has a complement. This reduces part of filter design to algebra.

Now, $\mathbf{C}[z, z^{-1}]$ is a Euclidean domain, so the division lemma holds:

Lemma 1. *Suppose* $a, b \in \mathbf{C}[z, z^{-1}]$ *with* $\deg a \geq \deg b \geq 0$. *Then there exists a quotient* q *and a remainder* r *with* $\deg r < \deg b$ *so that*

$$a(z) = q(z)b(z) + r(z).$$

Note that $\deg q = \deg a - \deg b$.

Write $q = a/b$ and $r = a\%b$, as in the C programming language, but note that neither q nor r is unique.

Lemma 2. *There are at most* $2^{1+\deg a-\deg b}$ *different ways to divide* $a(z)/b(z)$, *among which at most* $2 + \deg a - \deg b$ *quotients are different.*

Proof. First note that division is a generalization of Gaussian elimination. The vector of b's coefficients is shifted, scaled, and added to the vector of a's coefficients to eliminate either the highest or lowest power of z, namely the leftmost or rightmost term. After at most $1 + \deg a - \deg b$ such eliminations, the remainder will have degree less than $\deg b$. Each sequence of eliminations is determined by its sequence of "left" and "right" directives, making at most $2^{1+\deg a-\deg b}$ different ways to find the quotient.

However, left and right eliminations commute as long as $\deg a > \deg b$. Hence, two quotients will be the same if their elimination sequences contain the same number of left and right eliminations, regardless of order. Thus, there can be no more distinct quotients than the number of sequences of length $1 + \deg a - \deg b$ with distinct numbers of "left" directives, which is $2 + \deg a - \deg b$. □

Example 1. Let $a(z) = 2z^{-1} + 4 + z$ and $b(z) = 1 + z$. Then $\deg a = 2$ and $\deg b = 1$, so there are three distinct quotients:

$$a(z) = (2z^{-1} + 2)b(z) + (-z) \qquad \text{(left, left)}.$$
$$a(z) = (3z^{-1} + 1)b(z) + (-z^{-1}) \qquad \text{(right, right)}.$$
$$a(z) = (2z^{-1} + 1)b(z) + 1 \qquad \text{(left, right) or (right, left)}.$$

We may say "left division" to mean always eliminating the leftmost term and "right division" to mean always eliminating the rightmost term. When $\deg a - \deg b$ is even, there will be an even number of terms to eliminate so we may say "symmetric division" to mean an equal number of left and right eliminations.

A Laurent polynomial $p = p(z)$ is said to be *symmetric* if it is unchanged by reversing the order of its coefficients. This is equivalent to the property

$$(\exists M)(\forall z)\ z^M p(z^{-1}) = p(z).$$

For symmetric p not identically zero, the *reflection index* M is unique. Monomials z^k are evidently symmetric with $M = 2k$. We may further distinguish whole or half index symmetry, depending upon whether M is even or odd. The parity of M will be the same as that of $\deg p$.

Lemma 3. *If $a, b \in \mathbf{C}[z, z^{-1}]$ are symmetric Laurent polynomials, then symmetric division results in a symmetric quotient a/b and symmetric remainder $a\%b$.*

Proof. The result holds if $\deg a < \deg b$, since then $a/b = 0$ and $a\%b = a$.

For all other cases, use induction on $n = \deg a - \deg b$.

If $n = 0$, left elimination and right elimination produce identical monomial quotients, which are trivially symmetric. The remainders are evidently identical and symmetric as well.

If $n = 1$, the quotient will have degree 1 with two identical coefficients, hence it will be symmetric. The remainder will be the difference between two symmetric polynomials with the same reflection index M, hence it will itself be symmetric with that same M.

The induction step follows from the observation that a symmetric (left, right) pair of elimination steps reduces the degree of the dividend by 2 while preserving its symmetry. This reduces n by 2 and contributes to the quotient a symmetric polynomial of the same reflection index M as a. □

We now recall some basic notions useful in Euclidean domains:

- Write $b|a$ (b *divides* a) if $a = qb + 0$ for some q. Thus $b|a \Rightarrow \deg b \leq \deg a$.
- Say that d is a *common divisor* of a and b if $d|a$ and $d|b$.
- Say that a common divisor d is a *greatest common divisor* of a and b if every common divisor c of a and b also divides d.

Lemma 4. *If d_1 and d_2 are greatest common divisors for a and b, then $d_1 = ud_2$ for some unit $u \in \mathbf{C}[z, z^{-1}]$.*

Theorem 1. *Every pair $a, b \in \mathbf{C}[z, z^{-1}]$, not both zero, has a greatest common divisor that is unique up to multiplication by a unit.*

Denote this set of greatest common divisors by $\gcd(a, b)$. Say that a, b are *coprime* if $\gcd(a, b)$ is contained in the set of units.

Assume a, b are Laurent polynomials with $\deg a \geq \deg b \geq 0$. Their greatest common divisor may be found by the Euclidean algorithm for Laurent polynomials. Put $a_0 \stackrel{\text{def}}{=} a$ and $b_0 \stackrel{\text{def}}{=} b$, and define a_k, b_k recursively:

$$a_{k+1} = b_k; \qquad b_{k+1} = a_k - q_k b_k, \qquad k = 0, 1, 2, \ldots,$$

where q_k is one of the possible quotients a_k/b_k. It thus determines b_{k+1} as the corresponding one of the possible remainders $a_k \% b_k$.

Lemma 5. *Let n be the smallest positive integer for which $b_n = 0$. Then $a_n \in \gcd(a, b)$.*

By Lemma 4, finding any representative in $\gcd(a, b)$ determines all the others.

Example 2. Consider $a(z) = a_0(z) = 2z^{-1} + 4 + z$ and $b(z) = b_0(z) = 1 + z$. Using symmetric division, the first step is

$$a_1(z) = 1 + z, \qquad b_1(z) = 1, \qquad q_1(z) = 2z^{-1} + 1.$$

The second step is

$$a_2(z) = 1, \qquad b_2(z) = 0, \qquad q_2(z) = 1 + z.$$

Therefore

$$\begin{bmatrix} 2z^{-1} + 4 + z \\ 1 + z \end{bmatrix} = \begin{bmatrix} 2z^{-1} + 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 + z & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

so $\gcd(a, b) = 1$.

Note that a_n is determined only up to a unit, defined by the sequence of quotients $q_0, \ldots, q_{n-1}$.

Theorem 2. *Laurent polynomial h has a complement g if and only if h_e and h_o are coprime.*

Proof. Apply the Euclidean algorithm to find the polyphase matrix. Write the recursion in matrix form:

$$\begin{bmatrix} a_{k+1}(z) \\ b_{k+1}(z) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & -q_k(z) \end{bmatrix} \begin{bmatrix} a_k(z) \\ b_k(z) \end{bmatrix}, \quad \Rightarrow \begin{bmatrix} a_n(z) \\ 0 \end{bmatrix} = \prod_{k=1}^{n} \begin{bmatrix} 0 & 1 \\ 1 & -q_{n-k}(z) \end{bmatrix} \begin{bmatrix} a(z) \\ b(z) \end{bmatrix}.$$

Inverting the product of matrices gives

$$\begin{bmatrix} a(z) \\ b(z) \end{bmatrix} = (-1)^n \prod_{k=0}^{n-1} \begin{bmatrix} q_k(z) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a_n(z) \\ 0 \end{bmatrix}.$$

If n is odd, absorb the unit $(-1)^n$ term into a_n.

Put $a = h_e$ and $b = h_o$ and assume $\gcd(h_e, h_o) = Kz^m$, $K \neq 0$. Define g_e, g_o by

$$P(z) \stackrel{\text{def}}{=} \begin{bmatrix} h_e(z) & g_e(z) \\ h_o(z) & g_o(z) \end{bmatrix} = \prod_{k=0}^{n-1} \begin{bmatrix} q_k(z) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} Kz^m & 0 \\ 0 & K^{-1}z^{-m} \end{bmatrix}.$$

Then $P(z)$ is evidently invertible. Get $\tilde{h}, \tilde{g}$ from $\tilde{P}(z^{-1})^t = P(z)^{-1}$. □

This leads immediately to general implementations of DWT by lifting:

Theorem 3 (Daubechies and Sweldens). *For every perfect reconstruction finite filter pair (H, G) with polyphase matrix P, there exist finitely many Laurent polynomials $s_i(z)$ and $t_i(z)$, $1 \leq i \leq m < \infty$, and a nonzero constant K such that*

$$P(z) = \prod_{i=1}^{m} \begin{bmatrix} 1 & s_i(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ t_i(z) & 1 \end{bmatrix} \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix}.$$

Remark 4. The matrix factors correspond to the following operations on sequences:

- *Prediction*: Unit upper triangular factor; $u_e \leftarrow u_e + Su_o$.
- *Updating*: Unit lower triangular factor, $u_o \leftarrow u_o + Tu_e$.
- *Scaling*: Last diagonal matrix, $u_e \leftarrow Ku_e$, $u_o \leftarrow K^{-1}u_o$.

Since u_e, u_e may be stored as disjoint arrays, this transform can be performed in place, without extra memory for temporary results.

Proof. Observe first:

$$\begin{bmatrix} q_k(z) & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & q_k(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ q_k(z) & 1 \end{bmatrix}.$$

The *flip* matrices $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ cancel if predict and update steps alternate.

Second, note that a leftover flip matrix may be factored into lifting steps in a number of ways:

$$\begin{aligned}
\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.
\end{aligned}$$

Third, note that diagonal shift matrices may be factored into lifting steps:

$$\begin{aligned}
\begin{bmatrix} z & 0 \\ 0 & z^{-1} \end{bmatrix} &= \begin{bmatrix} 1 & -z \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ z^{-1} & 1 \end{bmatrix} \begin{bmatrix} 1 & 1-z \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1-z & 1 \end{bmatrix} \begin{bmatrix} 1 & z^{-1} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -z & 1 \end{bmatrix}
\end{aligned}$$

$$\begin{bmatrix} z^{-1} & 0 \\ 0 & z \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ z & 1 \end{bmatrix} \begin{bmatrix} 1 & -z^{-1} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1+z & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}.$$

Other factorizations exist, but at most five lifting steps are needed per shift. Thus, $\begin{bmatrix} z^m & 0 \\ 0 & z^{-m} \end{bmatrix}$ or $\begin{bmatrix} z^{-m} & 0 \\ 0 & z^m \end{bmatrix}$ factor into at most $5m$ lifting steps. Afterward, only the constant diagonal matrix $\begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix}$ remains. □

4 Nearest Neighbor Factorization

Assume that a smooth sampled signal $u \in \ell^2$ is finitely supported in the index interval $[0, N-1]$. Big endpoint values $|u(0)|$ or $|u(N-1)|$ may result in misleading large DWT coefficients. Similarly, a big difference $|u(N-1) - u(0)|$ may result in large periodized DWT coefficients. These undesirable effects are mitigated through the use of *symmetric extension*, as described in [1]. It requires symmetric H, G, defining

$$u(n) = \begin{cases} u(-n), & if\text{-N¡n¡0;} \\ u(n) = u(2N-1-n), & if\,\text{N¡n¡2N,} \end{cases}$$

and then treating u as $2N-2$-periodic. Several other extensions are possible, depending on the symmetry type of H, G.

It is easy to implement symmetric extension for certain special implementations. The lifting factorization

$$P(z) = \prod_{k=1}^{n} \begin{bmatrix} 1 & s_k(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ t_k(z) & 1 \end{bmatrix} \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix}$$

uses only *nearest neighbors* if it satisfies the following conditions:

$$s_k(z) = \alpha_k + \beta_k z^{-1},$$
$$t_k(z) = \gamma_k z + \delta_k,$$

with $\alpha_k, \beta_k, \gamma_k, \delta_k \in \mathbf{C}$. Nearest neighbor action on sequences has the explicit forms:

- *Nearest neighbor predict:* $u_{2k} \leftarrow u_{2k} + \alpha_k u_{2k-1} + \beta_k u_{2k+1}$.
- *Nearest neighbor update:* $u_{2k+1} \leftarrow u_{2k+1} + \gamma_k u_{2k} + \delta_k u_{2k+2}$.

For nearest neighbor factorizations, symmetric extension becomes:

- *Symmetric extension nearest neighbor predict step:*

$$u_{2k} \leftarrow u_{2k} + \begin{cases} \alpha(u_{2k-1} + u_{2k+1}), & if\,2k \neq 0; \\ 2\alpha\, u_{2k+1}, & if\,2k = 0. \end{cases}$$

- *Symmetric extension nearest neighbor update step:*

$$u_{2k+1} \leftarrow u_{2k+1} + \begin{cases} \gamma(u_{2k} + u_{2k+2}), & if\,2k+1 \neq N-1; \\ 2\gamma\, u_{N-2}, & if\,2k+1 = N-1. \end{cases}$$

Hence the endpoints get almost the same treatment as the interior points.

5 All Lifting Can Be Nearest Neighbor Lifting

Unfortunately, not every perfect reconstruction filters factor into nearest neighbor lifting steps directly, even allowing for any choice of quotients in Euclid's algorithm. For example, let

$$h(z) = \frac{1}{\sqrt{2}}(1 + z^{-9}) \qquad g(z) = \frac{1}{\sqrt{2}}(-z^{8} + z^{-1}).$$

This is similar to the Haar orthogonal filter pair. Then

$$h_e(z) = \frac{1}{\sqrt{2}}; \quad h_o(z) = \frac{1}{\sqrt{2}}z^{-4}; \quad g_e(z) = \frac{1}{\sqrt{2}}z^{4}; \quad g_o(z) = \frac{1}{\sqrt{2}}.$$

There are no division steps in Euclid's algorithm, so the (empty) sequence of quotients is unique. The ordinary lifting factorization gives:

$$P(z) = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & z^{4} \\ z^{-4} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ z^{-4} & 1 \end{bmatrix}\begin{bmatrix} 1 & -\frac{1}{2}z^{4} \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \sqrt{2} \end{bmatrix}.$$

This is not a nearest neighbor factorization because the off-diagonal nonconstant terms have powers other than z and z^{-1}. However, in common with nearest neighbor lifting steps, the off-diagonal terms $s(z), t(z)$ have $\deg s \leq 1$ and $\deg t \leq 1$, and further factorization is possible (see Lemma 7) to obtain nearest neighbor steps.

We may obtain quotients of constrained degree through a modification of the division lemma:

Lemma 6 (Partial division). *Suppose $a, b \in \mathbf{C}[z, z^{-1}]$ with $\deg a \geq \deg b \geq 0$. Then there exists a partial quotient q and a partial remainder r with $\deg q \leq 1$ and $\deg r < \deg a$ so that*

$$a(z) = q(z)b(z) + r(z).$$

Proof. We limit ourselves to eliminating just one or two terms from a with a partial quotient of the form $q(z) = z^{m}(\gamma + \delta z)$. Then $\deg q \leq 1$, but not both $\gamma = 0$ and $\delta = 0$, so $\deg r = \deg(a - qb) < \deg a$. □

It is clear that any common divisor of $a(z)$ and $b(z)$ also divides the partial remainder $r(z) = a(z) - q(z)b(z)$, so we obtain lifting factors of degree one or less with Euclid's algorithm expanded to use partial division:

Theorem 4. *Assume a and b are two coprime nonzero Laurent polynomials. Then there exist Laurent polynomials $q_1, \ldots, q_n$ with $\deg q_k \leq 1$ for all $k = 1, \ldots, n$, such that*

$$\begin{bmatrix} a(z) \\ b(z) \end{bmatrix} = \prod_{k=1}^{n}\begin{bmatrix} q_k(z) & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} K \\ 0 \end{bmatrix},$$

where $n \leq 2(\deg a + \deg b + 1)$.

We may now slightly strengthen Theorem 3:

Corollary 1. *For every perfect reconstruction finite filter pair (H, G) with polyphase matrix P, there is a lifting factorization*

$$P(z) = \prod_{i=1}^{m} \begin{bmatrix} 1 & s_i(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ t_i(z) & 1 \end{bmatrix} \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix},$$

where the Laurent polynomials $s_i(z)$ and $t_i(z)$, $1 \le i \le m < \infty$, each have degree one or less, and K is a nonzero constant.

As shown earlier, the degree condition is not enough to guarantee that the factorization gives a nearest neighbor filter transform. To get from degree one or less to nearest neighbor polynomials requires additional factorization:

Lemma 7.

$$\begin{bmatrix} 1 & z^{2m}(\alpha z^{-1} + \beta) \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} z^m & 0 \\ 0 & z^{-m} \end{bmatrix} \begin{bmatrix} 1 & \alpha z^{-1} + \beta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z^{-m} & 0 \\ 0 & z^m \end{bmatrix};$$

$$\begin{bmatrix} 1 & 0 \\ z^{2m}(\gamma + \delta z) & 1 \end{bmatrix} = \begin{bmatrix} z^{-m} & 0 \\ 0 & z^m \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \gamma + \delta z & 1 \end{bmatrix} \begin{bmatrix} z^m & 0 \\ 0 & z^{-m} \end{bmatrix},$$

where m is any integer and α, β, γ, and δ are constants.

Factoring the z^m shifts into at most $5m$ nearest neighbor lifting steps, each yields:

Corollary 2. *Any degree-one predict or update matrix factors further into a finite number of nearest neighbor lifting steps.*

Remark 5. The conditions $\deg s_k \le 1$ and $\deg t_k \le 1$ may also be obtained from an ordinary lifting factorization by further decomposition:

$$\begin{bmatrix} 1 & s_1(z) + s_2(z) \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & s_1(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & s_2(z) \\ 0 & 1 \end{bmatrix},$$

$$\begin{bmatrix} 1 & 0 \\ t_1(z) + t_2(z) & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ t_1(z) & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ t_2(z) & 1 \end{bmatrix}.$$

However, this may create multiple successive predict factors and multiple successive update factors and ultimately requires more matrices.

6 Backward Error Analysis for Lifting Factorizations

We now turn consider how various lifting factorizations affect the conditioning of DWT. In terms of the polyphase matrix, this may be computed as follows:

Lemma 8. *If $P(z)$ is the polyphase matrix of a perfect reconstruction filter pair, then*

$$\operatorname{cond}(P) \stackrel{\text{def}}{=} \frac{\sup\left\{\sqrt{\lambda_{\max}(P(z)^* P(z))} : |z| = 1\right\}}{\inf\left\{\sqrt{\lambda_{\min}(P(z)^* P(z))} : |z| = 1\right\}}.$$

Furthermore, if $P = P_1 \cdots P_n$, then $\operatorname{cond}(P) \le \operatorname{cond}(P_1) \cdots \operatorname{cond}(P_n)$.

For a proof of this fact and extensive discussion of polyphase matrices, see [6].

We may use the special form of lifting step matrices to estimate the condition number of the factorization of P in another, simpler way. For definiteness, consider an "update" step in floating-point arithmetic with absolute truncation error ϵ. Its polyphase matrix G may be represented within the computer by something that differs by as much as δG, where

$$G(z) = \begin{bmatrix} 1 & 0 \\ t(z) & 1 \end{bmatrix} \qquad \Rightarrow \qquad \delta G(z) = \begin{bmatrix} \epsilon & 0 \\ \delta t(z) & \epsilon \end{bmatrix},$$

and Laurent polynomial $t(z) = \sum_k t_k z^{-k}$ has error $\delta t(z) = \sum_k \delta t_k z^{-k}$. Each polynomial coefficient δt_k is computed from the filter h by elimination and therefore satisfies

$$|\delta t_k| \le (1 + \deg h)\epsilon + O(\epsilon^2).$$

For such G, define

$$|G|(z) \stackrel{\text{def}}{=} \begin{bmatrix} 1 & 0 \\ |t|(z) & 1 \end{bmatrix}, \qquad \text{where} \quad |t|(z) \stackrel{\text{def}}{=} \sum_k |t_k| z^{-k}.$$

Then the maximum matrix infinity norm of $G(z)$ may be computed as follows:

$$\|G\|_\infty \stackrel{\text{def}}{=} \sup_{|z|=1} \|G(z)\|_\infty \le \sup_{|z|=1} \| \, |G|(z) \, \|_\infty = 1 + \sup_{|z|=1} |t|(z) = 1 + \sum_k |t_k|.$$

The same estimate applies to "predict" steps as well.

Now assume that $P(z)$ is a polyphase matrix with lifting factorization

$$P(z) = \prod_i \begin{bmatrix} 1 & s_i(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ t_i(z) & 1 \end{bmatrix} \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix} \stackrel{\text{def}}{=} \prod_j G_j(z),$$

where each G_j is a lifting step. Taking truncation errors into account, the computed results using floating-point arithmetic therefore satisfy

$$P(z) + \delta P(z) = \prod_j (G_j(z) + \delta G_j(z)).$$

Table 1 Condition number bounds for nearest neighbor factorization versus ordinary lifting versus the polyphase matrix

Filter	Cond of P(z)	Cond of lifting	Cond of N-N
9-7	1.32	205	205
D4	1	77	77
D6	1	76	76
Cubic B-spline	4	56	56
CDF-1-1	1	8.59	8.59
CDF-1-3	1.28	8.72	3,100
CDF-1-5	1.42	6.25	1,200
CDF-2-2	2	8.59	8.59
CDF-2-4	2	99	1,900
CDF-3-1	4	643	643
CDF-3-3	4	723	3,200
CDF-4-2	8	111	111
CDF-4-4	8	113	2,800

By expanding the product and using the submultiplicativity of the matrix infinity norm, we get

$$\|\delta P\|_\infty \le \epsilon\,(1+\deg h)\sum_j ||G_j||_\infty + O(\epsilon^2),$$

indicating that to obtain the smallest condition number, we should use a factorization with small lifting coefficients and not too many low-degree lifting steps.

Assuming the worst case, equality, we can estimate the condition number for three implementations of the filter bank:

1. The original polyphase matrix P
2. The usual (shortest) lifting factorization of P
3. The nearest neighbor lifting factorization of P

The results are displayed in Table 1, for a number of symmetric and nonsymmetric-orthogonal filters.

7 Applications and Examples

Because of nonuniqueness in the quotients in Euclid's algorithm for matrices over Laurent polynomials, we may choose a lifting factorization optimized for a minimal number of nearest neighbors.

In some cases, the original lifting steps yield a nearest neighbor algorithm. The filter indexing may need to be adjusted to eliminate or at least minimize the number of z^m shift matrices. In addition, the sequence of quotients $\{q_k(z) : k = 0, 1, \ldots, n-1\}$ may be chosen to minimize the condition number bound.

Example 3. To show how re-indexing may result in a nearest neighbor factorization, consider Daubechies' D4 filters, defined as follows:

$$h(z) = h_3 z^{-3} + h_2 z^{-2} + h_1 z^{-1} + h_0,$$
$$g(z) = h_0 z^{-1} - h_1 + h_2 z - h_3 z^2,$$

with

$$h_0 = \frac{1+\sqrt{3}}{4\sqrt{2}}, \qquad h_1 = \frac{3+\sqrt{3}}{4\sqrt{2}}, \qquad h_2 = \frac{3-\sqrt{3}}{4\sqrt{2}}, \qquad h_3 = \frac{1-\sqrt{3}}{4\sqrt{2}}.$$

Then the polyphase matrix is

$$P(z) = \tilde{P}(z) = \begin{bmatrix} h_e(z) & g_e(z) \\ h_o(z) & g_o(z) \end{bmatrix} = \begin{bmatrix} h_2 z^{-1} + h_0 & -h_1 - h_3 z \\ h_3 z^{-1} + h_1 & h_0 + h_2 z \end{bmatrix}.$$

It has the following two factorizations, the first obtained by left division in Euclid's algorithm and the second by right division:

$$P(z) = \begin{bmatrix} 1 & -\sqrt{3} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{\sqrt{3}}{4} + \frac{\sqrt{3}-2}{4} z^{-1} & 1 \end{bmatrix} \begin{bmatrix} 1 & z \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{3}+1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}-1}{\sqrt{2}} \end{bmatrix}$$
$$= \begin{bmatrix} 1 & \frac{\sqrt{3}}{3} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{-\sqrt{3}}{4} + \frac{3\sqrt{3}+6}{4} z & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{-z^{-1}}{3} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{3-\sqrt{3}}{3\sqrt{2}} z^{-1} & 0 \\ 0 & \frac{3+\sqrt{3}}{\sqrt{2}} z \end{bmatrix}.$$

The forward transform corresponding to the first (left-division) factorization is not nearest neighbor:

$$\begin{aligned} x_{2m+1} &\leftarrow x_{2m+1} - \sqrt{3} x_{2m}; \\ x_{2m} &\leftarrow x_{2m} + \frac{\sqrt{3}}{4} x^{(1)}_{2m+1} + \frac{\sqrt{3}-2}{4} x^{(1)}_{2m+3}; \\ x_{2m+1} &\leftarrow x_{2m} + x_{2m-2}; \\ x_{2m} &\leftarrow \frac{\sqrt{3}+1}{\sqrt{2}} x_{2m}; \\ x_{2m+1} &\leftarrow \frac{\sqrt{3}-1}{\sqrt{2}} x_{2m+1}. \end{aligned}$$

The inverse transform for the left-division factorization into lifting steps is similarly not nearest neighbor, as may be seen from its polyphase matrix:

$$\tilde{P}(z^{-1})^t = \begin{bmatrix} \frac{\sqrt{3}+1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ z^{-1} & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{\sqrt{3}}{4} + \frac{\sqrt{3}-2}{4} z \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\sqrt{3} & 1 \end{bmatrix}.$$

It is left as an exercise to derive the predict and update steps from these matrices.

The second (right-division) factorization can be made nearest neighbor by factoring the leftover diagonal shift matrices into lifting steps. However, if the coefficients of h are first shifted so that $h(z) = \sum_{-2}^{1} h_{k+2} z^{-k}$, then right division in Euclid's algorithm yields a nearest neighbor transform directly:

$$P(z) = \begin{bmatrix} 1 & \frac{\sqrt{3}}{3} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{-\sqrt{3}}{4} + \frac{3\sqrt{3}+6}{4} z & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{-z^{-1}}{3} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{3-\sqrt{3}}{3\sqrt{2}} & 0 \\ 0 & \frac{3+\sqrt{3}}{\sqrt{2}} \end{bmatrix}.$$

It is again left as an exercise to find the corresponding predict and update steps.

Example 4. Not all orthogonal filters will give nearest neighbor factorization directly after a simple index shift. Consider the orthogonal filters defined earlier:

$$h(z) = \frac{\sqrt{2}}{2}(1 + z^{-9}) \qquad g(z) = \frac{\sqrt{2}}{2}(-z^8 + z^{-1}).$$

We cannot get a nearest neighbor factorization simply by using an index shift. For this filter, it is necessary to use Lemma 7 and pay the price of additional lifting steps.

Example 5. Not all filters offer a choice of lifting factorizations. The biorthogonal perfect reconstruction filters `CDF-2-4h` and `CDF-2g` have the following coefficients:

$$\begin{aligned} h(z) = \sqrt{2}\Big(& \frac{3}{128} z^{-4} - \frac{3}{64} z^{-3} - \frac{1}{8} z^{-2} + \frac{19}{64} z^{-1} + \\ & + \frac{45}{64} + \frac{19}{64} z - \frac{1}{8} z^2 - \frac{3}{64} z^3 + \frac{3}{128} z^4 \Big) \\ g(z) = \sqrt{2}\Big(& \frac{1}{4} z^{-2} - \frac{1}{2} z^{-1} + \frac{1}{4} \Big). \end{aligned}$$

CDF biorthogonal filters [2] are symmetric, so there is a unique lifting factorization using the Euclidean algorithm. But since the degrees of $h(z)$ and $g(z)$ are very different, in most cases it does not yield a nearest neighbor factorization directly.

$$\begin{aligned} P(z) &= \begin{bmatrix} -\frac{1}{2} z^{-1} - \frac{1}{2} & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -\frac{3}{64} z^{-1} + \frac{19}{64} + \frac{19}{64} z - \frac{3}{64} z^2 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & -\frac{\sqrt{2}}{2} \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{2} z^{-1} - \frac{1}{2} & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -\frac{3}{64} z^{-1} + \frac{19}{64} & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{aligned}$$

$$\times \begin{bmatrix} \frac{19}{64}z - \frac{3}{64}z^2 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & -\frac{\sqrt{2}}{2} \end{bmatrix}$$
$$= \begin{bmatrix} 1 & -\frac{1}{2}z^{-1} - \frac{1}{2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & -\frac{3}{64}z^{-1} + \frac{19}{64} \\ 0 & 1 \end{bmatrix}$$
$$\begin{bmatrix} z & 0 \\ 0 & z^{-1} \end{bmatrix} \begin{bmatrix} 1 & \frac{19}{64}z^{-1} - \frac{3}{64} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z^{-1} & 0 \\ 0 & z \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & -\frac{\sqrt{2}}{2} \end{bmatrix}.$$

An application of Lemma 7, Corollary 2, and Theorem 3 may be used to convert this into a nearest neighbor factorization.

Example 6. If the filter h has a perfect reconstruction complement, then its coefficients may be re-indexed and the proper quotients chosen in the division algorithm so that $\gcd(h_e, h_o)$ is a constant. However, this is not guaranteed to produce nearest neighbor lifting.

Consider Daubechies' orthogonal `D6` filter; it may have its indices shifted so that $h(z) = \sum_{k=-2}^{3} h_k z^{-k}$. The filter thus indexed, together with its complement $g = \dot{h}$, has the following polyphase matrix:

$$h_e(z) = h_2 z^{-1} + h_0 + h_{-2} z \qquad g_e(z) = -h_{-1} z^{-1} - h_1 - h_3 z$$
$$h_o(z) = h_3 z^{-1} + h_1 + h_{-1} z \qquad g_o(z) = h_{-2} z^{-1} + h_0 + h_2 z.$$

Then the lifting factorization by symmetric division yields a constant $\gcd(h_e, h_o) \approx 1.918$, entirely through nearest neighbor predict and update steps:

$$P(z) = \begin{bmatrix} 1 & 0 \\ -0.412 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1.565z^{-1} + 0.352 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0.028 + 0.492z & 1 \end{bmatrix}$$
$$\begin{bmatrix} 1 & -0.390 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1.918 & 0 \\ 0 & 0.521 \end{bmatrix}.$$

Alternatively, with the indexing $h(z) = \sum_{k=0}^{5} h_k z^{-k}$, we get a different polyphase matrix:

$$h_e(z) = h_4 z^{-2} + h_2 z^{-1} + h_0 \qquad g_e(z) = -h_1 z^{-1} - h_3 z^1 - h_5 z^2$$
$$h_o(z) = h_5 z^{-2} + h_3 z^{-1} + h_1 \qquad g_o(z) = h_0 + h_2 z + h_4 z^2.$$

Using the same division as for `D4` now gives nonconstant $\gcd(h_e, h_o) = 1.918 z^{-1}$, but with nearest neighbor predict and update steps:

$$P(z) = \begin{bmatrix} 1 & 0 \\ -0.412 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1.565z^{-1} + 0.352 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0.028 + 0.492z & 1 \end{bmatrix}$$
$$\begin{bmatrix} 1 & -0.390 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1.918z^{-1} & 0 \\ 0 & 0.521z \end{bmatrix}.$$

However, choosing right division so that the gcd comes out constant gives

$$P(z) = \begin{bmatrix} 1 & 0 \\ -0.412 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1.565z^{-1} + 0.355 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0.001645z^{-1} - 0.028 & 1 \end{bmatrix}$$
$$\begin{bmatrix} 1 & 607.65z - 116.5z^2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -33.172 & 0 \\ 0 & 0.03 \end{bmatrix},$$

which is evidently not a nearest neighbor factorization.

References

1. Brislawn, C.: Classification of nonexpansive symmetric extension transforms for multirate filter banks. Appl. Comput. Harmon. Anal. **3**(4), 337–357 (1996)
2. Cohen, A., Daubechies, I., Feauveau, J.-C.: Biorthogonal bases of compactly supported wavelets. Comm. Pure. Appl. Math. **45**, 485–500 (1992)
3. Daubechies, I.: Orthonormal bases of compactly supported wavelets. Comm. Pure. Appl. Math. **41**, 909–996 (1988)
4. Daubechies, I., Sweldens, W.: Factoring wavelet transforms into lifting steps. J. Fourier Anal. Appl. **4**(3), 245–267 (1998)
5. Mallat, S.G.: Multiresolution approximation and wavelet orthonormal bases of $L^2(\mathbf{R})$. Trans. Am. Math. Soc. **315**, 69–87 (1989)
6. Strang, G., Nguyen, T.: Wavelets and Filter Banks. Wellesley–Cambridge Press, Wellesley, Massachusetts (1996)

Part VII
Operator Theory

Operator theory is at the crossroads of several areas of mathematics, physics, and engineering. For instance, linear and non linear operator theory is an essential tool in partial differential equations and mathematical physics, while operator algebras are fundamental in noncommutative analysis. In engineering, operators are used to model effectively the actions of certain systems on signals. The seven chapters in this part give an overview of the modern role of operator theory in analysis, mathematical physics, time-frequency analysis, multi-scale harmonic analysis, and medical imaging reconstructions.

The first chapter of this part by OVIDIU CALIN, DER-CHEN CHANG, and YUTIAN LI offers an elegant construction of the heat kernel for certain operators on a Lie group. In particular, they construct a heat kernel associated to operators consisting of the finite sum of squares of left invariant vector fields on a finite dimensional noncommutative Lie group that satisfies an involutivity condition. This construction is achieved through a novel approach they developed, and it relies on replacing the usual bracket-generating condition by a new one they call the *total involutivity property*. This condition together with a geometric method involving Hamiltonian formalism are the main tools in their construction.

In their chapter JAVIER DUOANDIKOETXEA and VIRGINIA NAIBO give an excellent overview of mixed-norm inequalities for the k-plane transform and some related potential-type operators supported on k-planes. The history of these types of operators goes back to the work of J. Radon on the reconstruction of sufficiently smooth two-dimensional functions from their line integrals: this is the Radon transform. The applications of such techniques include x-rays in medical imaging. The chapter focuses on a review of recent results on the boundedness of k-plane transforms with special attention paid to their actions on radial functions. One of the main tools used by the authors is the maximal operator associated to these operators. The boundedness of this maximal operator leads to interesting consequences for the bounds of the Kakeya maximal operator.

PETER C. GIBSON, MICHAEL P. LAMOUREUX, and GARY F. MARGRAVE address some fundamental questions on Gabor multipliers. Gabor multipliers are operators obtained by composing the short-time Fourier transform (STFT), multiplication by a distribution on phase space, and the inverse STFT. The STFT, also known as the Gabor transform, appeared in D. Gabor's seminal work on the theory of communication. In particular, Gabor proposed to decompose and reconstruct finite energy signals using the STFT with a Gaussian window. Gabor's proposal can also be viewed as a resolution of the identity operator in terms of time-frequency shifts of the Gaussian. This is in fact an example of a Gabor multiplier in which the distribution on phase space is identically equal to one. Today, other choices of window functions that allow similar analysis are known, but the design of windows in Gabor analysis is still not completely resolved. In this chapter, the authors give a characterization of all linear operators that can be written as Gabor multipliers. Moreover, given one such linear operator, they analyze the choice of window needed to write it as a Gabor multiplier. Though theoretical, these results have applications in seismic image data processing. In this context, the authors have

developed an approach that uses Gabor multipliers to represent nonstationary filters and wave field extrapolators.

The fourth chapter of this part is by JOHN R. KLAUDER and BO-STURE K. SKAGERSTAM. While their contribution is rooted in quantum mechanics, Klauder and Skagerstam point out its applicability in time-frequency analysis. The specific context for this chapter is the Berezin-Lieb inequalities which dictate upper and lower bounds for partition functions based on phase space integrals involving certain representations. These inequalities were first established by F. A. Berezin and E. H. Lieb using the well-known Husimi and Glauder-Sudarshan symbol classes. In an earlier work, Klauder and Skagerstam defined larger classes of these phase space symbols. In the present chapter, they use these new classes of space phase symbols to extend the Berezin-Lieb inequalities as a consequence of more general integral inequalities.

In his chapter, DIEGO MALDONADO gives a survey of boundedness properties and the symbolic calculus for certain bilinear pseudo-differential operators. After a brief review of some of the classical results due to Hörmander, Kohn, Nirenberg, Calderón, and Vaillancourt, Maldonado considers the bilinear pseudo-differential operators defined by R. R. Coifman and Y. Meyer in their seminal work. The remaining part of the chapter consists of a review of recent deep results by him and his collaborators. These include novel results on paraproducts, as well as results on a special class of bilinear pseudo-differential operators, viz., the bilinear Calderón-Zygmund operators.

Part VII ends with MARÍA C. PEREYRA'S exposition of the recent solution of the A_2 conjecture. This asserts that all Calderón-Zygmund singular integral operators are bounded on weighted L^2 spaces with a bound that depends linearly on the A_2 characteristic of the weight as well as corresponding results for the associated commutators. Pereyra guides us through the history of the conjecture culminating in its recent proof by Tuomas Hytönen. The special case of the boundedness on weighted L^2 spaces of the Hilbert transform as well as its commutators with certain BMO functions is considered and shown to lead to sharp bounds on the dyadic paraproduct on corresponding weighted L^p space. Some of the main tools that appear in her exposition are dyadic harmonic analysis, Bellman function techniques, the martingale transform, and dyadic Haar shift operators.

On the Heat Kernel of a Left Invariant Elliptic Operator

Ovidiu Calin, Der-Chen Chang, and Yutian Li

Dedicate to Professor John J. Benedetto

Abstract The goal of this chapter is to find the heat kernel of a left invariant operator on a Lie group, by using a geometric method involving Hamiltonian formalism.

Keywords Heat kernel • Lie group • Curvature • Geodesic

1 Introduction

There is a vast literature dealing with properties of the sum of squares $L = X_1^2 + \cdots X_k^2$, with $k \leq n$ of left invariant vector fields X_i on a noncommutative Lie group G of dimension n (see [1, 2, 5, 7, 9]). Most papers assume that the bracket generating condition holds, i.e., the vector fields X_i together with their iterated Lie brackets generate the Lie algebra $\mathcal{L}(G)$. This condition implies the operator L

O. Calin
Department of Mathematics, Eastern Michigan University,
Ypsilanti, MI 48197, USA
e-mail: ocalin@emich.edu

D.-C. Chang (✉)
Department of Mathematics and Department of Computer Science, Georgetown University,
Washington D.C. 20057, USA

Department of Mathematics, Fu Jen Catholic University, Taipei 242, Taiwan, ROC
e-mail: chang@georgetown.edu

Y. Li
Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong
e-mail: yutianli@hkbu.edu.hk

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_10,

hypoelliptic (see Hörmander [8]) and also global connectivity by piecewise smooth curves tangent to the distribution $\mathcal{H} = \text{span}\{X_1, \ldots, X_k\}$ (see Chow and Rashevskii [6, 12], see also [4, Chap. 3]). The distribution $\mathcal{H}$ in this case is not involutive.

We are in a completely different situation if for any two vector fields $X, Y \in \mathcal{L}(G)$ then the Lie bracket $[X, Y]$ is a linear combination of X and Y:

$$\forall X, Y \in \mathcal{L}(G) \Longrightarrow [X, Y] = \alpha X + \beta Y, \qquad \alpha, \beta \in \mathbb{R}. \tag{1}$$

The vector field $[X, Y]$ always belongs to $\mathcal{L}(G)$; however, the aforementioned property states that $[X, Y]$ belongs to the plane spanned by the vector fields X and Y. This will be regarded in the future as the *total involutivity property* of the group G.

This can be stated equivalently in terms of constants of structure as follows. Let $\{E_1, \ldots, E_n\}$ be a basis of the Lie algebra $\mathcal{L}(G)$. The constants of structure C_{ij}^k are defined by

$$[E_i, E_j] = \sum_{k=1}^{n} C_{ij}^k E_k.$$

Then the total involutivity property can be written as

$$C_{ij}^k = 0 \text{ for } k \notin \{i, j\}.$$

This property does not always hold for any noncommutative Lie group. A counterexample on hand is the Heisenberg group $\mathbb{H}$ (see [7]). If $\{X_1, X_2, X_3\}$ is a basis of its Lie algebra $\mathfrak{h}$, then

$$[X_1, X_2] = X_3, \ [X_1, X_3] = [X_2, X_3] = 0,$$

so $[X_1, X_2]$ is not a linear combination of X_1 and X_2.

Milnor [10] proved that if a noncommutative Lie group G of dimension $n \geq 2$ has the property (1), then there is a one-form ω such that

$$[X, Y] = \omega(X)Y - \omega(Y)X.$$

Furthermore, he showed that any left invariant metric g on the Lie group G has constant negative sectional curvature. This means that for any 2-plane π we have $K_\pi = -\|\omega\|^2 < 0$, where

$$\|\omega\|^2 = \omega_1^2 + \cdots + \omega_n^2,$$

with $\omega = \sum \omega_i \, dx_i$. In the case $n \geq 3$, by Schur's theorem, it follows that the Riemannian space (G, g) has a constant negative curvature.

The result was extended by Nomizu [11] for left invariant Lorentz metrics on G; however, in this case the sectional curvature is not necessary negative.

In the following we shall consider an example of a Lie group that satisfies the total involutivity property (1).

Let $G = (\mathbb{R}^{2k}, \circ)$ be endowed with the noncommutative group law

$$\begin{aligned}(x_1, x_2, \ldots, x_{2k-1}, x_{2k}) \circ (x_1', x_2', \ldots, x_{2k-1}', x_{2k}') \\ = (x_1 + \mathrm{e}^{-x_2} x_1',\ x_2 + x_2', \ldots, x_{2k-1} + \mathrm{e}^{-x_{2k}} x_{2k-1}',\ x_{2k} + x_{2k}').\end{aligned} \tag{2}$$

One can check that

$$E_1 = \mathrm{e}^{-x_2} \partial_{x_1},\ E_2 = \partial_{x_2}, \ldots, E_{2k-1} = \mathrm{e}^{-x_{2k}} \partial_{x_{2k-1}},\ E_{2k} = \partial_{x_{2k}} \tag{3}$$

are left invariant vector fields that span the Lie algebra $\mathcal{L}(G)$. The commutator relations are

$$[E_1, E_2] = E_1, \ldots, [E_{2k-1}, E_{2k}] = E_{2k-1},$$

and hence G satisfies the property (1).

The authors are not aware of any general result regarding the heat kernel of an operator defined as sum of squares of vector fields $L = X_1^2 + \cdots X_k^2$, with $k \leq n$ of left invariant vector fields X_i on a noncommutative Lie group G of dimension n that satisfies the total involutivity property (1).

In the case when $n = k$, then L is an elliptic operator on the connected, simple connected, and constant curvature space G. Hence it must be isomorphic to the hyperbolic space $\mathcal{H}_n$. Then our problem reduces to the hope of being able to express the heat kernel of the elliptic operator L using the heat kernel of the elliptic operator

$$\widetilde{L} = Y_1^2 + \cdots + Y_n^2,$$

where $Y_k = F_* X_k$ are vector fields induced by the isomorphism $F : G \to \mathcal{H}_n$.

However, in this chapter, we shall consider a direct method of a geometric flavor. The purpose of this chapter is to find the heat kernel of the sum of squares of the vector fields given by (3). It suffices to solve the problem in the case $k = 1$, since the heat kernel in the general case is a product of heat kernels in the lower dimension. This shall be done using a similar analysis to the one done in the case of the Heisenberg group for the heat kernel of the Heisenberg Laplacian (see [1,2]). In this case the Lie group is 2-dimensional, and it resembles the Grushin case, replacing the polynomial coefficient by an exponential.

2 Introducing the Geometry of a Lie Group

Consider the Lie group $G = (\mathbb{R}^2, \circ)$ with the noncommutative group law

$$(x_1, x_2) \circ (x_1', x_2') = (x_1 + \mathrm{e}^{-x_2} x_1',\ x_2 + x_2'). \tag{4}$$

One can check that

$$E_1 = \mathrm{e}^{-x_2} \partial_{x_1}, \qquad E_2 = \partial_{x_2} \tag{5}$$

are left invariant vector fields which span the Lie algebra of G. The commutator relation is $[E_1, E_2] = E_1$, so the constants of structure are zero with the exception of $C_{12}^1 = -C_{21}^1 = 1$. Consider the Riemannian metric on $\mathbb{R}^2$ in which the vector fields $\{E_1, E_2\}$ form an orthonormal basis

$$g_{ij} = g(\partial_{x_i}, \partial_{x_j}) = \begin{pmatrix} \mathrm{e}^{2x_2} & 0 \\ 0 & 1 \end{pmatrix}.$$

A computation shows that all Christoffel symbols are zero with the exception of

$$\Gamma_{12,1} = \Gamma_{21,1} = -\Gamma_{11,2} = \mathrm{e}^{2x_2},$$
$$\Gamma_{12}^1 = \Gamma_{21}^1 = 1, \quad \Gamma_{11}^2 = -\mathrm{e}^{2x_2}.$$

By a straightforward computation or just by using a MAPLE package we get $R_{212}^1 = -1$, and hence

$$R_{1212} = g_{11} R^{212} = -\mathrm{e}^{2x_2}.$$

Since $g_{11}g_{22} - (g_{12})^2 = \mathrm{e}^{2x_2}$, it follows that the relation

$$R_{1212} = K\big(g_{11}g_{22} - (g_{12})^2\big)$$

holds for $K = -1$, i.e., the Riemannian manifold (G, g) is a space with negative constant curvature. By Hadamard's theorem it follows that the space is geodesically complete, i.e., any two points can be joined by a unique geodesic. This is an important feature which will be used when computing the heat kernel of the associated operator.

3 The Associated Elliptic Operator

The geometry of the Riemannian space (G, g) is associated with the sum of squares elliptic operator[1]

$$\mathbb{L} = \frac{1}{2}(E_1^2 + E_2^2) = \frac{1}{2}(\mathrm{e}^{-2x_2}\partial_{x_1}^2 + \partial_{x_2}^2). \tag{6}$$

We are interested in finding an explicit formula for the heat kernel of the operator (6). The principal symbol

[1]This operator resembles the Grushin operator $\frac{1}{2}(x_2^2\partial_{x_1}^2 + \partial_{x_2}^2)$, but in this case it is left invariant with respect to a group law.

$$H(x,p) = \frac{1}{2}(\mathrm{e}^{-2x_2}p_1^2 + p_2^2) = \frac{1}{2}\sum g^{ij}(x)\,p_i\,p_j$$

is considered as a Hamiltonian function. It is known that the geodesics of the space (G,g) can be obtained as the x-projection of the bicharacteristics associated with the Hamiltonian H (see for instance [3], Chap. 6.) In order to find the equations of the geodesics, we shall solve the bicharacteristics system

$$\dot{x}_1 = H_{p_1} = \mathrm{e}^{-2x_2}p_1, \tag{7}$$

$$\dot{x}_2 = H_{p_2} = p_2, \tag{8}$$

$$\dot{p}_1 = -H_{x_1} = 0 \implies p_1 = c \text{ (constant)}, \tag{9}$$

$$\dot{p}_2 = -H_{x_2} = \mathrm{e}^{-2x_2}p_1^2, \tag{10}$$

with the boundary conditions

$$x_1(0) = x_2(0) = 0, \quad x_1(\tau) = x_1, \quad x_2(\tau) = x_2. \tag{11}$$

3.1 Finding Geodesics

Since we are working on a group it suffices to study only the geodesics $\big(x_1(t), x_2(t)\big)$ starting at the origin. All the other geodesics are obtained by left translations with respect to the law (4).

3.1.1 Finding the x_2-Component

Differentiating in Eq. (8) and using Eqs. (9) and (10) yields the ODE

$$\ddot{x}_2 = c^2\mathrm{e}^{-2x_2}. \tag{12}$$

It can be checked by direct differentiation that a first integral of motion for (12) is the total energy

$$\frac{1}{2}\dot{x}_2^2 + \frac{1}{2}c^2\mathrm{e}^{-2x_2} = E \text{ (constant)}. \tag{13}$$

Consider that parametrization of the bicharacteristics for which $E = \frac{1}{2}$. This is obtained for the parameter $s = \sqrt{2}\times$(arc length). This way the parameters are the momentum $p_1 = c$ and the geodesic length $\tau/\sqrt{2}$. Since there is a unique geodesic between the origin and any given point in $\mathbb{R}^2$, then the parameters c and τ are uniquely determined by the boundary conditions (11).

Substituting $E = \frac{1}{2}$ in (13), then separating and integrating between 0 and s, yields

$$\int_0^{x_2(s)} \frac{\mathrm{d}u}{\sqrt{1 - c^2 \mathrm{e}^{-2u}}} = \pm s.$$

Making the substitution $v = \mathrm{e}^u$ transforms the previous relation into

$$\int_1^{\mathrm{e}^{x_2(s)}} \frac{\mathrm{d}v}{\sqrt{v^2 - c^2}} = \pm s.$$

Since the left side is always positive and $s > 0$, only the positive sign on the right side will provide a solution. Integrating and solving for $x_2(s)$ yields

$$\begin{aligned}
\cosh^{-1}(\mathrm{e}^{x_2(s)}/c) &= \cosh^{-1}(1/c) + s \iff \\
\mathrm{e}^{x_2(s)} &= c \cosh\left(\cosh^{-1}(1/c) + s\right) \iff \\
\mathrm{e}^{x_2(s)} &= \cosh s + c \sinh\left(\cosh^{-1}(1/c)\right) \sinh s \iff \\
\mathrm{e}^{x_2(s)} &= \cosh s + \sqrt{1 - c^2} \sinh s, \qquad \forall s \in [0, \tau].
\end{aligned}$$

This can be also written as

$$\mathrm{e}^{x_2(s)} = \frac{\cosh(K + s)}{\cosh K}, \tag{14}$$

where $K = \cosh^{-1}(1/c)$.

3.1.2 Finding the x_1-Component

Substituting (14) in (7) yields

$$\dot{x}_1 = c\mathrm{e}^{-2x_2} = \frac{1}{c \cosh^2(K + s)}. \tag{15}$$

Integrating between 0 and s yields

$$\begin{aligned}
x_1(s) &= \frac{1}{c}\big[\tanh(K + s) - \tanh K\big] \\
&= \cosh K \tanh(K + s) - \sinh K.
\end{aligned} \tag{16}$$

Proposition 1. *The geodesics starting at the origin and parameterized by*

$$s = \sqrt{2} \times (arc\ length)$$

are given by

$$x_1(s) = \cosh K \, \tanh(K+s) - \sinh K,$$
$$x_2(s) = \log\cosh(K+s) - \log\cosh K,$$

with $K = \cosh^{-1}(1/c)$*, where* c *is the constant value of the momentum* p_1.

4 The Riemannian Distance

We shall express the Riemannian distance from the origin in terms of the boundary conditions

$$x_1(\tau) = x_1, \quad x_2(\tau) = x_2.$$

Making $s = \tau$ in (14) and (16) yields

$$\mathrm{e}^{x_2} = \frac{\cosh(K+\tau)}{\cosh K}, \tag{17}$$
$$x_1 = \cosh K \, \tanh(K+\tau) - \sinh K \tag{18}$$

We shall solve the system (17) and (18) for τ and K in terms of the boundary conditions x_1 and x_2. It will be useful to denote $W = K + \tau$ and solve the system for the unknowns W and K. Then $\tau = W - K$ will provide the distance from the origin to x.

Using (17) Eq. (18) becomes

$$\begin{aligned} x_1 &= \frac{\cosh K}{\cosh W} \sinh W - \sinh K \\ &= \mathrm{e}^{-x_2} \sinh W - \sinh K. \end{aligned} \tag{19}$$

Equation (17) can be also written as

$$\mathrm{e}^{-x_2} \cosh W - \cosh K = 0. \tag{20}$$

Taking the square and converting into sinh we have

$$\mathrm{e}^{-2x_2} \cosh^2 W - \cosh^2 K = 0 \iff$$
$$\mathrm{e}^{-2x_2} \sinh^2 W - \sinh^2 K = 1 - \mathrm{e}^{-2x_2}.$$

Factoring and using (19) yields

$$x_1(\mathrm{e}^{-x_2} \sinh W + \sinh K) = 1 - \mathrm{e}^{-2x_2}.$$

Assuming $x_1 \neq 0$ yields

$$e^{-x_2} \sinh W + \sinh K = \frac{1 - e^{-2x_2}}{x_1}. \tag{21}$$

Adding and subtracting Eqs. (19) and (21) yields

$$\sinh W = \frac{e^{x_2}(1 - e^{-2x_2} + x_1^2)}{2x_1} = \frac{\sinh x_2}{x_1} + e^{x_2}\frac{x_1}{2}, \tag{22}$$

$$\sinh K = \frac{1 - e^{-2x_2} - x_1^2}{2x_1} = e^{-x_2}\frac{\sinh x_2}{x_1} - \frac{x_1}{2}. \tag{23}$$

Then we can compute the value of the parameter τ in terms of x_1 and x_2

$$\begin{aligned}
\tau &= W - K \\
&= \sinh^{-1}\left(\frac{\sinh x_2}{x_1} + \frac{x_1 e^{x_2}}{2}\right) - \sinh^{-1}\left(e^{-x_2}\frac{\sinh x_2}{x_1} - \frac{x_1}{2}\right) \\
&= \log \frac{\sinh x_2 + e^{x_2}\frac{x_1^2}{2} + \sqrt{x_1^2 + \left(\sinh x_2 + e^{x_2}\frac{x_1^2}{2}\right)^2}}{e^{-x_2}\sinh x_2 - \frac{x_1^2}{2} + \sqrt{x_1^2 + \left(e^{-x_2}\sinh x_2 - \frac{x_1^2}{2}\right)^2}},
\end{aligned} \tag{24}$$

where we used the formula $\sinh^{-1} u = \log(u + \sqrt{1 + u^2})$.

Remark 1. If let $x_1 = 0$ in (18) then $\tanh(K + \tau) = \tanh K$, which implies $\tau = 0$. Substituting in (17) yields $e^{x_2} = 1$ and hence $x_2 = 0$. In this case the geodesic starts and ends at the origin and has the length zero and hence it reduces to a point.

5 Heat Kernels

Given two points x_0 and x in space $\mathbb{R}^n$ and a time $t > 0$ the propagator from x_0 to x within time t, is given by the path integral

$$\mathcal{P}(x_0, x; t) = \int_{\Phi_{x_0,x;t}} e^{-S(\phi;t)}\, dm(\phi), \tag{25}$$

where

$$\Phi_{x_0,x;t} = \{\phi : [0,t] \to \mathbb{R}^n : \phi(0) = x_0,\ \phi(t) = x\}$$

is the space of continuous paths ϕ from x_0 to x in time t. This means that $\mathcal{P}$ depends on all continuous paths joining x_0 and x parametrized by $[0, t]$. Among all the possible paths between the aforementioned points, the *classical path* plays a

distinguished role. This is the path on which a classical particle would travel and is given by the solution of the Euler–Lagrange system of equations. It is a remarkable fact that for a *classical Lagrangian*, i.e., a Lagrangian which is at most quadratic in $\dot{x}_j$ and x_j, the path integral (25) depends only on the classical action. This is the famous van Vleck's formula which expresses the path integral in the following closed form (see [13]):

$$\begin{aligned}\mathcal{P}(x_0, x; t) &= \int_{\Phi_{x_0,x;t}} \mathrm{e}^{-S(\phi;t)} \mathrm{d}m(\phi) \\ &= \sqrt{\det\left(-\frac{1}{2\pi}\frac{\partial^2 S(x_0, x; t)}{\partial x_0 \partial x}\right)}\, \mathrm{e}^{-S(x_0,x;t)},\end{aligned} \tag{26}$$

where $S(x_0, x; t)$ is the classical action obtained integrating the Lagrangian along the solution $x(t)$ of the Euler–Lagrange equation. It is known that the action function $S(x_0, x; t)$ satisfies the Hamilton–Jacobi equation:

$$\frac{\partial S}{\partial t} + H(x, y; \nabla S) = 0,$$

where $\nabla S = \left(\frac{\partial S}{\partial x}, \frac{\partial S}{\partial y}\right)$. The factor in (26)

$$V(t) = \sqrt{\det\left(-\frac{1}{2\pi}\frac{\partial^2 S(x_0, x; t)}{\partial x_0 \partial x}\right)}$$

is called the *van Vleck determinant*, and it plays the role of the *volume element* in the geometric method described in many articles (see, for example [2, 3, 5]). Now we just need to construct the action function and the volume element to obtain the heat kernels.

5.1 The Action Function

Since the underlying geometry is Riemannian, the action function between the origin and the point $x = (x_1, x_2)$ in time t is given by

$$S(0, x; t) = \frac{\mathrm{dist}(0, x)^2}{2t} = \frac{\mathrm{dist}(0, x_1, x_2)^2}{2t};$$

see for instance [3], Chap. 7. Therefore it suffices to find the distance $\mathrm{dist}(0, x)$ in the g-metric. According to our previous parametrization, the Riemannian distance is the length of the geodesic, and it is given by $\mathrm{dist}(0, x) = \tau/\sqrt{2}$, with τ given by formula (24).

5.2 The Volume Element

As we mentioned in the beginning of this section, we need to compute the volume element as the van Vleck determinant; see [5, Sect. 7.9]. This method works only for the case where the Lagrangian is at most quadratic in the positions and momenta. In order to achieve this goal, we make the change of variables

$$u = x_1, \qquad v = \mathrm{e}^{-x_2}.$$

Then in the new coordinates (u, v),

$$\partial_{x_1}^2 = \partial_u^2, \qquad \partial_{x_2}^2 = v\partial_v + v^2\partial_v^2 \tag{27}$$

and

$$\mathbb{L} = \frac{1}{2}(\mathrm{e}^{-2x_2}\partial_{x_1}^2 + \partial_{x_2}^2) = \frac{1}{2}(v^2\partial_u^2 + v\partial_v + v^2\partial_v^2). \tag{28}$$

Now the Lagrangian $\mathcal{L}(u, v, \mu, \nu)$ and Hamiltonian $H(u, v, \mu, \nu)$ of the operator (28) are at most quadratic in u, v, μ, ν. Then the action function between $O(0, 1)$ and (u, v) is

$$\begin{aligned} S(0, 1, u, v; t) &= \frac{\mathrm{dist}(0, 1; u, v)^2}{2t} \\ &= \frac{1}{4t}\log^2 \frac{\frac{(1+u^2)v^{-1}-v}{2} + \sqrt{u^2 + \left(\frac{(1+u^2)v^{-1}-v}{2}\right)^2}}{\frac{1-u^2-v^2}{2} + \sqrt{u^2 + \left(\frac{1-u^2-v^2}{2}\right)^2}}, \end{aligned}$$

and

$$S(u_0, v_0, u, v; t) = S(0, 1, u - u_0, v/v_0; t).$$

Hence,

$$\frac{\partial S}{\partial u_0} = -\frac{\partial S}{\partial u}, \qquad \frac{\partial S}{\partial v_0} = -\frac{v}{v_0}\frac{\partial S}{\partial v},$$

and

$$\frac{\partial^2 S}{\partial v \partial v_0} = -\frac{1}{v_0}\left(\frac{\partial S}{\partial v} + v\frac{\partial^2 S}{\partial v^2}\right).$$

Now the van Vleck determinant is

$$V(t) = \sqrt{\det\left(-\frac{1}{2\pi}\frac{\partial^2 S}{\partial \mathbf{u}\partial \mathbf{u}_0}\right)}$$

$$= \frac{1}{2\pi}\sqrt{\frac{\partial^2 S}{\partial u \partial u_0}\frac{\partial^2 S}{\partial v \partial v_0} - \frac{\partial^2 S}{\partial u \partial v_0}\frac{\partial^2 S}{\partial v \partial u_0}}$$

$$= \frac{1}{2\pi}\sqrt{\frac{1}{vv_0}\left[\frac{\partial^2 S}{\partial u^2}\left(v\frac{\partial S}{\partial v} + v^2\frac{\partial^2 S}{\partial v^2}\right) - \left(v\frac{\partial^2 S}{\partial u \partial v}\right)^2\right]}.$$

Noting that $S = \frac{\tau^2}{4t}$ and using (27), we have

$$V(t) = \frac{1}{2\pi}\sqrt{e^{x_2+x_2^0}\left[\frac{\partial^2 S}{\partial x_1^2}\frac{\partial^2 S}{\partial x_2^2} - \left(\frac{\partial^2 S}{\partial x_1 \partial x_2}\right)^2\right]}$$

$$= \frac{e^{\frac{x_2+x_2^0}{2}}}{4\pi t}\left\{\left[\left(\frac{\partial \tau}{\partial x_1}\right)^2 + \tau\frac{\partial^2 \tau}{\partial x_1^2}\right]\left[\left(\frac{\partial \tau}{\partial x_2}\right)^2 + \tau\frac{\partial^2 \tau}{\partial x_2^2}\right]\right.$$

$$\left. - \left[\frac{\partial \tau}{\partial x_1}\frac{\partial \tau}{\partial x_2} + \tau\frac{\partial^2 \tau}{\partial x_1 \partial x_2}\right]^2\right\}^{1/2}, \tag{29}$$

where the derivatives of τ with respect to x_1 and x_2 can be obtained from (24) though it is in a complicated form. Finally, combining the calculations in this section, we obtain the heat kernel for the operator

$$\frac{\partial}{\partial t} - \mathbb{L} = \frac{\partial}{\partial t} - \frac{1}{2}(e^{-2x_2}\partial_{x_1}^2 + \partial_{x_2}^2).$$

In fact, we may consider the heat kernel for the operator

$$\frac{\partial}{\partial t} - \sum_{j=1}^{k}\mathbb{L}_j = \frac{\partial}{\partial t} - \sum_{j=1}^{m}\frac{1}{2}\left(e^{-2x_{2j}}\partial_{x_{2j-1}}^2 + \partial_{x_{2j}}^2\right)$$

defined on $\mathbb{R}_+ \times G$. Since

$$[\mathbb{L}_i, \mathbb{L}_j] = 0 \qquad \text{for} \qquad i \neq j,$$

then the heat kernel will be the product of heat kernels for each operator $\mathbb{L}_j$. One has the following theorem.

Theorem 1. *The heat kernel for the operator* $\frac{\partial}{\partial t} - \sum_{j=1}^{m}\mathbb{L}_j$ *defined on* $\mathbb{R}_+ \times G$ *has the following form:*

$$\mathcal{P}(0, \mathbf{x}, t) = \mathcal{P}(0, x_1, x_2, \ldots, x_{2k-1}, x_{2k}, t) = \prod_{j=1}^{k} V_j(t)e^{-\frac{\tau_j^2}{4t}},$$

where

$$\tau_j \;=\; \log \frac{\sinh x_{2j} \,+\, \mathrm{e}^{x_{2j}} \frac{x_{2j-1}^2}{2} + \sqrt{x_{2j-1}^2 + \left(\sinh x_{2j} + \mathrm{e}^{x_{2j}} \frac{x_{2j-1}^2}{2}\right)^2}}{\mathrm{e}^{-x_{2j}} \sinh x_{2j} - \frac{x_{2j-1}^2}{2} + \sqrt{x_{2j-1}^2 + \left(\mathrm{e}^{-x_{2j}} \sinh x_{2j} - \frac{x_{2j-1}^2}{2}\right)^2}},$$

$$V_j(t) = \frac{\mathrm{e}^{\frac{x_{2j}}{2}}}{4\pi t} \left\{ \left[\left(\frac{\partial \tau}{\partial x_{2j-1}} \right)^2 + \tau \frac{\partial^2 \tau}{\partial x_{2j-1}^2} \right] \left[\left(\frac{\partial \tau}{\partial x_{2j}} \right)^2 + \tau \frac{\partial^2 \tau}{\partial x_{2j}^2} \right] - \left[\frac{\partial \tau}{\partial x_{2j-1}} \frac{\partial \tau}{\partial x_{2j}} + \tau \frac{\partial^2 \tau}{\partial x_{2j-1} \partial x_{2j}} \right]^2 \right\}^{1/2}.$$

Moreover,

$$\mathcal{P}(\mathbf{x}_0, \mathbf{x}, t) \;=\; \mathcal{P}(\mathbf{0}, \mathbf{x} \circ \mathbf{x}_0^{-1}, t),$$

with $\circ$ *being the group law which is defined by (2) and* $\mathbf{x}_0^{-1}$ *being the inverse of* $\mathbf{x}_0$.

So far, all the calculations are based on the hypothesis that the constant c in (9) is not zero. When $c = 0$, then $p_1 = 0$ (from (9)). Then (7) tells us that $\dot{x}_1 = 0$ which implies that $x_1(s) =$ constant. Therefore, along any vertical line, the operator becomes $\partial_{x_2}^2$ (which is the one-dimensional Laplacian) and the corresponding geodesics are line segments (along the vertical line $x_1 =$ constant.) In this case, we may assume that $x_1 = 0$. From the formula (24), we know that

$$\begin{aligned} \tau \;&=\; \log \frac{\sinh x_2 + \sqrt{\sinh^2 x_2}}{\mathrm{e}^{-x^2} \sinh x_2 + \sqrt{\mathrm{e}^{-2x_2} \sinh^2 x_2}} \\ &= \log \frac{2 \sinh x_2}{2\mathrm{e}^{-x_2} \sinh x_2} \;=\; \log \mathrm{e}^{x_2} \;=\; x_2. \end{aligned}$$

It follows that the action function is

$$S(0, x_2, t) = \frac{\mathrm{dist}(0, 0, x_2)^2}{2t} = \frac{\tau^2}{4t} = \frac{x_2^2}{4t}$$

and the volume element is

$$V(t) = \frac{\mathrm{e}^{x_2}}{4\pi t} \sqrt{\frac{x_2}{\sinh x_2}}.$$

The heat kernel in this case takes the form

$$\mathcal{P}(0, x_2, t) \;=\; \frac{1}{4\pi t} \sqrt{\frac{x_2}{\sinh x_2}} \mathrm{e}^{x_2 - \frac{x_2^2}{4t}}.$$

Acknowledgment This research project is partially supported by an NSF grant DMS-1203845 and Hong Kong RGC competitive earmarked research grants #600607, #601410.

References

1. Beals, R., Gaveau, B., Greiner, P.C.: On a geometric formula for the fundamental solution of subelliptic Laplacians. Math. Nachr. **181**, 81–163 (1996)
2. Beals, R., Gaveau, B., Greiner, P.C.: Hamilton–Jacobi theory and the heat kernel on Heisenberg groups. J. Math. Pures Appl. **79**(7), 633–689 (2000)
3. Calin, O., Chang, D.C.: Geometric mechanics on Riemannian manifolds: applications to partial differential equations. Applied and Numerical Analysis. Birhäuser, Boston (2004)
4. Calin, O., Chang, D.C.: Sub-Riemannian geometry, general theory and examples. Encyclopedia of Mathematics and Its Applications, vol. 126, Cambridge University Press, Cambridge (2009)
5. Calin, O., Chang, D.C., Furutani, K., Iwasaki, K.: Heat Kernels, Methods and Techniques, Applied and Numerical Analysis. Birhäuser, Boston (2010)
6. Chow, W.L.: Über systeme von linearen partiellen differentialgleichungen erster ordnung. Math. Ann. **117**, 98–105 (1939)
7. Gaveau, B.: Systémes dynamiques associés à certains opérateurs hypoelliptiques. Bull. Sci. Math. **102**, 203–229 (1978)
8. Hörmander, L.: Hypo-elliptic second order differential equations. Acta Math. **119**, 147–171 (1978)
9. Hulanicki, A.: The distribution of energy in the Brownian motion in the Gausssian field and analytic hypoellipticity of certain subelliptic operators on the Heisenberg group. Studia Math. **56**, 165–173 (1976)
10. Milnor, J.: Curvatures of left invariant metrics on Lie groups. Adv. Math. **21**, 293–329 (1976)
11. Nomizu, K.: Left invariant Lorentz metrics on Lie groups. Osaka J. math. **16**, 143–150 (1979)
12. Rashevskii, P.K.: About connecting two points of complete nonholonomic space by admissible curve. Uch. Zapiski Ped. Inst. Libknexta **2**, 83–94 (1938) [Russian]
13. van Vleck, J.H.: The correspondence principle in the statistical interpretation of quantum mechanics. Proc. Natl. Acad. Sci. USA **14**(176), 178–188 (1928)

Mixed-Norm Estimates for the k-Plane Transform

Javier Duoandikoetxea and Virginia Naibo

Abstract The Radon transform constitutes a fundamental concept for X-rays in medical imaging, and more generally, in image reconstruction problems from diverse fields. The Radon transform in Euclidean spaces assigns to functions their integrals over affine hyperplanes. This can be extended so that the integration is performed on affine k-dimensional subspaces; the corresponding transform is called k-plane transform.

An overview of mixed-norm inequalities for the k-plane transform and related potential-type operators supported on k-planes is presented. Particular attention is given to the action of these operators on classes of radial functions, and applications to bounds for the Kakeya maximal operator are discussed.

Keywords Radon transform • X-ray transform • k-plane transform • Grassmannian manifold • Mixed-norm estimates • Directional operators • Homogeneous singular integrals • Potential operators • Kakeya maximal operator

1 Introduction

The reconstruction of an object from a series of projections has its roots in the work of Johann Radon [28, 29], who in 1917 showed that a sufficiently smooth function of two variables is uniquely determined by its integrals along

J. Duoandikoetxea
Departamento de Matemáticas, Universidad del País Vasco-Euskal Herriko Unibertsitatea, Apartado 644, 48080 Bilbao, Spain
e-mail: javier.duoandikoetxea@ehu.es

V. Naibo (✉)
Department of Mathematics, Kansas State University, 138 Cardwell Hall, Manhattan, KS-66506, USA
e-mail: vnaibo@math.ksu.edu

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_11,

arbitrary lines. A two-dimensional object can be thought of in terms of its density function $f(x_1, x_2)$, $(x_1, x_2) \in \mathbb{R}^2$, and the integral

$$P_1 f((x_1, x_2), (u_1, u_2)) = \int_{-\infty}^{\infty} f(x_1 - t\, u_1, x_2 - t\, u_2)\, \mathrm{d}t,$$

where (u_1, u_2) is a unit vector in $\mathbb{R}^2$, corresponds to the total mass along the line through the point (x_1, x_2) in the direction of (u_1, u_2). The operator $P_1 f$ is called the X-ray transform or the Radon transform of f.

More generally, one can consider smooth functions on $\mathbb{R}^n$ and integration over affine subspaces of dimension other than one. Given $f \in \mathscr{S}(\mathbb{R}^n)$, the Schwartz class of rapidly decreasing functions on $\mathbb{R}^n$ and an integer $k \in (0, n)$, the k-plane transform of f is defined by

$$P_k f(x, \pi) = \int_{\pi} f(x - y)\, \mathrm{d}\lambda_k(y),$$

where $x \in \mathbb{R}^n$, π is a k-dimensional subspace of $\mathbb{R}^n$, and λ_k denotes Lebesgue measure on π. The operators P_1 and P_{n-1} correspond to the X-ray transform and the Radon transform, respectively, and they coincide in the case $n = 2$. For an extensive overview of results and applications concerning these transformations consult the book by Markoe [26] and references therein.

We denote by $G(n, k)$ the Grassmannian manifold of all k-dimensional subspaces of $\mathbb{R}^n$ endowed with a finite measure $\gamma_{n,k}$, unique up to a constant factor, that is invariant under orthogonal transformations. We use the notation $\mathscr{G}_{n,k}$ for the affine Grassmannian, $\mathscr{G}_{n,k} := \{(\pi, x) : \pi \in G(n, k) \text{ and } x \in \pi^{\perp}\}$, with the measure induced by the linear functional $h \to \int_{\pi \in G(n,k)} \left(\int_{x \in \pi^{\perp}} h(\pi, x)\, \mathrm{d}\lambda_{n-k}(x) \right) \mathrm{d}\gamma_{n,k}(\pi)$.

If $f \in L^1(\mathbb{R}^n)$ then Fubini's theorem yields

$$\int_{\pi^{\perp}} P_k(|f|)(x, \pi)\, \mathrm{d}\lambda_{n-k}(x) = \|f\|_{L^1(\mathbb{R}^n)}, \quad \pi \in G(k, n),$$

and therefore $P_k f(x, \pi)$ exists almost everywhere for $x \in \pi^{\perp}$. In fact, the above equality gives that $P_k f$ is integrable on $\mathscr{G}_{k,n}$ and

$$\|P_k f\|_{L^1(\mathscr{G}_{n,k})} \leq \gamma_{n,k}(G_{n,k})\, \|f\|_{L^1(\mathbb{R}^n)}.$$

As a consequence, $P_k f(\pi, x)$ exists almost everywhere in $(\pi, x) \in \mathscr{G}_{n,k}$ when $f \in L^1(\mathbb{R}^n)$. Even more, if $1 \leq p < \frac{n}{k}$ then every $f \in L^p(\mathbb{R}^n)$ is integrable over almost every k-plane and therefore P_k is well defined on all of $L^p(\mathbb{R}^n)$. On the other hand, if $p \geq \frac{n}{k}$ then the k-plane transform is not well defined on $L^p(\mathbb{R}^n)$ since there exist nonnegative functions $f \in L^p(\mathbb{R}^n)$ that are not integrable over any k-plane of $\mathbb{R}^n$; for instance, $f(x) = \chi_{(|x|>e)}(x)|x|^{-\frac{n}{p}} \frac{1}{\log|x|}$. These results on lower integrability of L^p functions are due to Solmon–Smith [31] and Solmon [33]; see also Oberlin–Stein [27], Calderón [3], and Rubin [30].

In this chapter we present a survey of results concerning boundedness properties of the k-plane transform and other related operators on Lebesgue spaces in the form of mixed-norm estimates. More precisely, we are interested in the following inequalities:

$$\left(\int_{\mathbb{R}^{\kappa}}\left(\int_{G(n,k)}|P_k f(x,\pi)|^r\,\mathrm{d}\gamma_{n,k}(\pi)\right)^{q/r}\mathrm{d}x\right)^{1/q}\leq C_{p,q,r}\|f\|_{L^p(\mathbb{R}^n)},\qquad(1)$$

and

$$\left(\int_{G(n,k)}\left(\int_{\pi^{\perp}}|P_k f(x,\pi)|^q\,\mathrm{d}\lambda_{n-k}(x)\right)^{r/q}\mathrm{d}\gamma_{n,k}(\pi)\right)^{1/r}\leq C_{p,q,r}\|f\|_{L^p(\mathbb{R}^n)}.\quad(2)$$

Inequalities of type (1) are motivated by the study of boundedness properties of integral operators with variable kernel as explained in Sect. 2. Since $G(n,1)$ can be identified with half S^{n-1} and $\gamma_{n,1}$ with the surface measure on it, inequality (4) of Sect. 2 coincides with (1) for $k=1$. When $r=q$, inequalities of type (2), when true, describe boundedness properties of P_k from $L^p(\mathbb{R}^n)$ into $L^q(\mathscr{G}_{n,k})$ such as the one for $p=q=1$ in the previous page.

The organization of the chapter is as follows. In Sect. 2 we motivate the study of mixed-norm inequalities for directional operators that arise when applying the method of rotations introduced by Calderón–Zygmund in [4] to study boundedness properties of homogenous singular integrals with variable kernel. Section 3 contains an overview of results on mixed-norm inequalities (1) and (2) for the k-plane transform, and Sect. 4 emphasizes these inequalities when restricting the operators to classes of radial functions. Section 5 is devoted to mixed-norm estimates of a family of potential-type operators defined on k-planes of which the k-plane transform is a particular case. The proofs of the theorems appearing in Sect. 5 comprise the study of a maximal operator associated to k-planes, which we address in Sect. 6. The boundedness properties of this last operator on classes of radial functions have fine consequences on the bounds for the Kakeya maximal operator when restricted to such classes of functions and we present these in Sect. 7.

We conclude this section with a few words on notation. We will frequently use the Lorentz spaces $L^{p,1}(\mathbb{R}^n)$ and $L^{p,\infty}(\mathbb{R}^n)$. We say that the operator T satisfies a weak (p,q) inequality if T is bounded from $L^p(\mathbb{R}^n)$ into $L^{q,\infty}(\mathbb{R}^n)$; that is, there exists a constant C such that

$$|\{x\in\mathbb{R}^n:|Tf(x)|>\lambda\}|\leq C\left(\frac{\|f\|_{L^p(\mathbb{R}^n)}}{\lambda}\right)^q,$$

for all $f\in L^p(\mathbb{R}^n)$ and all $\lambda>0$. If the above inequality only holds for characteristic functions of measurable sets we say that the operator T satisfies

a restricted weak (p,q) inequality; this is equivalent to T being bounded from $L^{p,1}(\mathbb{R}^n)$ into $L^{q,\infty}(\mathbb{R}^n)$. If $1 \le p \le \infty$ then p' will denote the conjugate index of p, $\frac{1}{p} + \frac{1}{p'} = 1$. For a measurable set $E \subset \mathbb{R}^n$ we use the notation $|E|$ for its Lebesgue measure and χ_E for its characteristic function.

2 Integral Operators with Variable Kernel and Directional Operators

Suppose that T is a bounded linear operator on $L^p(\mathbb{R})$. Given $u \in S^{n-1}$, we define the directional operator T_u on the space $\mathscr{S}(\mathbb{R}^n)$ of Schwartz functions as follows: If $x \in \mathbb{R}^n$, then $x = t_x u + y_x$ for unique $t_x \in \mathbb{R}$ and $y_x \in \mathbb{R}^n$ perpendicular to u, and $T_u f(x) := T(f_{u,x})(t_x)$, where $f \in \mathscr{S}(\mathbb{R}^n)$ and $f_{u,x}(s) = f(su + y_x)$. Note that Fubini's theorem implies that

$$\|T_u f\|_{L^p(\mathbb{R}^n)} \le C \, \|f\|_{L^p(\mathbb{R}^n)}, \quad f \in \mathscr{S}(\mathbb{R}^n),$$

where C is independent of u, since T is bounded on $L^p(\mathbb{R}^n)$.

Typical examples of directional operators are the directional Hilbert transform, the directional maximal operator associated to the Hardy–Littlewood maximal operator, and the X-ray transform. All of these appear when applying the method of rotations introduced by Calderón and Zygmund in [4] in the study of homogenous singular integrals of variable kernel. In the next two sections we explain how mixed-norm inequalities for directional operators are sufficient to prove these boundedness properties. See also the survey on directional operators and mixed norms by Duoandikoetxea [15] and references therein.

2.1 The Directional Hilbert Transform and the Directional Maximal Operator

Consider the singular integral with variable kernel

$$T_\Omega f(x) = \text{p.v.} \int_{\mathbb{R}^n} \frac{\Omega(x, y')}{|y|^n} \, f(x-y) \, \mathrm{d}y, \quad x \in \mathbb{R}^n,$$

where $y' = y/|y|$ and we assume that $\Omega(x,\cdot)$ is integrable and has mean value zero so that T_Ω is well defined on $\mathscr{S}(\mathbb{R}^n)$. If Ω is odd in its second variable, the method of rotations gives

$$T_\Omega f(x) = \frac{1}{2} \int_{S^{n-1}} \Omega(x,u) \, H_u f(x) \, \mathrm{d}\sigma(u),$$

where H_u is the directional Hilbert transform,

$$H_u f(x) = \text{p.v} \int_{-\infty}^{\infty} \frac{f(x-tu)}{t}\, \mathrm{d}t, \quad x \in \mathbb{R}^n,\ u \in S^{n-1}.$$

Assuming $\sup_{x\in\mathbb{R}^n} \int_{S^{n-1}} |\Omega(x,u)|^{r'}\, \mathrm{d}\sigma(u) < \infty$ for some $r \geq 1$, it then follows that T_Ω is bounded from $L^p(\mathbb{R}^n)$ into $L^p(\mathbb{R}^n)$, $1 \leq p \leq \infty$, if

$$\left(\int_{\mathbb{R}^n} \left(\int_{S^{n-1}} |H_u f(x)|^r\, \mathrm{d}\sigma(u) \right)^{p/r} \mathrm{d}x \right)^{1/p} \leq C_{p,r} \|f\|_{L^p(\mathbb{R}^n)}. \tag{3}$$

Inequality (3) is easily proved to be true for $p \geq r$ since the directional Hilbert transform is uniformly bounded on $L^p(\mathbb{R}^n)$; on the other hand, only partial results for the case $p < r$ are known. It is conjectured that if $1 < p < \infty$ then (3) holds if and only if $1 \leq r < \infty$ and $\frac{n}{p} < \frac{n-1}{r} + 1$. This last condition is shown to be necessary by considering $f(x) = \chi_{(|x|\leq 1)}(x)$ in (3), while the characteristic function of a Besicovitch-type set rules out the index $r = \infty$. The conjecture was completely proved in the two-dimensional case and for $1 \leq p \leq 2$ in higher dimensions by Calderón and Zygmund in [5]. These results were improved by Christ, Duoandikoetxea, and Rubio de Francia in [8] to the range $1 < p \leq \max\left(2, \frac{n+1}{2}\right)$.

When Ω is even in its second variable, the study of boundedness properties on Lebesgue spaces of the associated maximal singular integral,

$$T_\Omega^* f(x) = \sup_{0<\varepsilon<N<\infty} \int_{\varepsilon<|y|<N} \frac{\Omega(x,y')}{|y|^n} f(x-y)\, \mathrm{d}y, \quad x \in \mathbb{R}^n,$$

relies on mixed-norm inequalities for the directional maximal operator

$$M_u f(x) = \sup_{h>0} \frac{1}{h} \int_{-h}^{h} |f(x-tu)|\, \mathrm{d}t, \qquad u \in S^{n-1},\ x \in \mathbb{R}^n.$$

Mixed-norm inequalities for M_u analogous to those in (3) have been proved for indices in the same ranges as for the directional Hilbert transform. See Fefferman [21], Cowling–Mauceri [11], and Christ–Duoandikoetxea–Rubio de Francia [8].

2.2 The X-ray Transform

The X-ray transform appears when considering integral operators with variable kernel of the form

$$I_{1,\Omega} f(x) = \int_{\mathbb{R}^n} \frac{\Omega(x, y')}{|y|^{n-1}} f(x-y)\, \mathrm{d}y.$$

Again, by the method of rotations,

$$I_{1,\Omega} f(x) = \frac{1}{2} \int_{S^{n-1}} \Omega(x, u)\, P_1 f(x, u)\, \mathrm{d}\sigma(u),$$

and therefore, under certain integrability conditions on Ω, boundedness properties of $I_{1,\Omega}$ from $L^p(\mathbb{R}^n)$ into $L^q(\mathbb{R}^n)$ follow if the X-ray transform satisfies mixed-norm inequalities of the type

$$\left(\int_{\mathbb{R}^n} \left(\int_{S^{n-1}} |P_1 f(x, u)|^r \, \mathrm{d}\sigma(u) \right)^{q/r} \mathrm{d}x \right)^{1/q} \leq C_{p,q,r} \| f \|_{L^p(\mathbb{R}^n)}. \tag{4}$$

3 Mixed-Norm Inequalities for the k-Plane Transform

In this section we tour results concerning inequalities (1) and (2) for the k-plane transform.

Necessary conditions on the indices p, q, and r for (1) and (2) follow from scaling arguments and by checking the inequalities with appropriate functions. More specifically, regarding inequality (1):

- A scaling argument replacing f by $f_{\lambda(x)} = f(\lambda x)$, $\lambda > 0$, in (1), yields $\frac{n}{p} - \frac{n}{q} = k$.
- The condition $1 \leq p < \frac{n}{k}$ is implied by the function $f(x) = \chi_{(|x|>e)}(x)|x|^{-\frac{n}{p}} \frac{1}{\log|x|}$, which is in $L^p(\mathbb{R}^n)$ for all $p > 1$ but is not integrable on any k-plane with $k \geq \frac{n}{p}$, as mentioned in the introduction. Note that the condition from the previous argument implies $1 \leq p \leq \frac{n}{k}$.
- The function $f = \chi_{B(0,1)}$ forces $\frac{n-k}{r} > \frac{n}{p} - k$; indeed, it can be seen that $P_k f(x, \pi) \sim 1$ for large x and π in a subset of $G(n, k)$ of measure proportional to $|x|^{k-n}$, from which it follows that

$$\int_{\mathbb{R}^k} \left(\int_{G(n,k)} |P_k f(x, \pi)|^r \, \mathrm{d}\gamma_{n,k}(\pi) \right)^{q/r} \mathrm{d}x \geq C \int_{|x|>c} |x|^{(k-n)q/r}.$$

  The integral on the right-hand side is finite if and only if $(k-n)q/r < -n$ and since $\frac{n}{q} = \frac{n}{p} - k$ we obtain the above-mentioned condition.
- The characteristic function of an appropriate parallelepiped implies $\frac{1}{r} \geq \frac{1}{n}$.

As for inequality (2):

- The same scaling argument as above now requires $\frac{n}{p} - \frac{n-k}{q} = k$.
- The condition $1 \le p < \frac{n}{k}$ is again forced by $f(x) = \chi_{(|x|>e)}(x)|x|^{-\frac{n}{p}} \frac{1}{\log|x|}$.
- The characteristic function of a parallelepiped of side lengths $1 \times \delta \times \cdots \times \delta$ for a small positive number δ gives $r \le (n-k)p'$.

For inequality (1), the above conditions are also sufficient for the X-ray transform ($k = 1$) and the Radon transform ($k = n-1$) as proved by Duoandikoetxea–Oruetxebarria in [16, 17], respectively. As for inequality (2), it is conjectured that the above necessary conditions on the indices are also sufficient. The conjecture has been completely proved for $k \ge n/2$ and partially proved for the case $k < n/2$; see Solmon [32, 33], Oberlin–Stein [27] ($k = n-1$), Drury [12–14], Christ [7], Wolff [35] ($k = 1$). We next state the best-known results on mixed-norm inequalities of type (2).

Theorem 1 (Christ [7]). *Assuming* $\frac{n}{p} - \frac{n-k}{q} = k$, *inequality (2) holds for all* $f \in L^p(\mathbb{R}^n)$ *if*

$$1 \le p \le \frac{n+1}{k+1} \quad \text{and} \quad r \le (n-k)p',$$

or if

$$1 \le p \le 2, \quad 1 \le p < \frac{n}{k} \quad \text{and} \quad r \le (n-k)p'.$$

4 Mixed-Norm Inequalities for the k-plane Transform on Radial Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is radial if $f(x) = f_0(|x|)$ for some $f_0 : [0,\infty) \to \mathbb{R}$. In Duoandikoetxea–Naibo–Oruetxebarria [20], the authors considered inequalities (1) and (2) when restricted to radial functions. In this context, those necessary conditions on the indices stated in the previous section that follow from the scaling arguments or that use radial functions to force the inequalities are also sufficient.

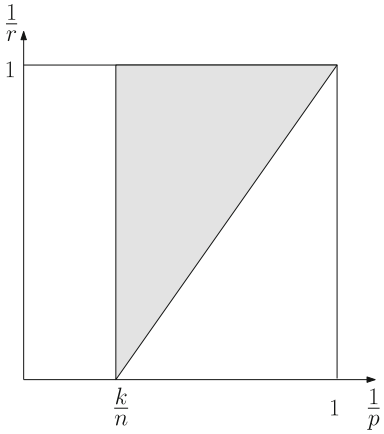
Theorem 2 (Duoandikoetxea–Naibo–Oruetxebarria [20]). *For radial functions inequality (1) holds if and only if*

$$1 \le r < \infty, \quad 1 < p < \frac{n}{k}, \quad \frac{n}{p} - \frac{n}{q} = k, \quad \frac{n-k}{r} > \frac{n}{p} - k,$$

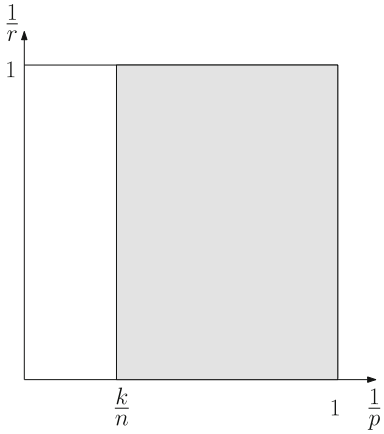
and inequality (2) holds if and only if

$$1 \le r \le \infty, \quad 1 \le p < \frac{n}{k}, \quad \frac{n}{p} - \frac{n-k}{q} = k\,.$$

The following figures illustrate the range of indices p and r indicated in Theorem 2, with q given by the scaling relation outside of the picture.



Inequality (1) - radial functions



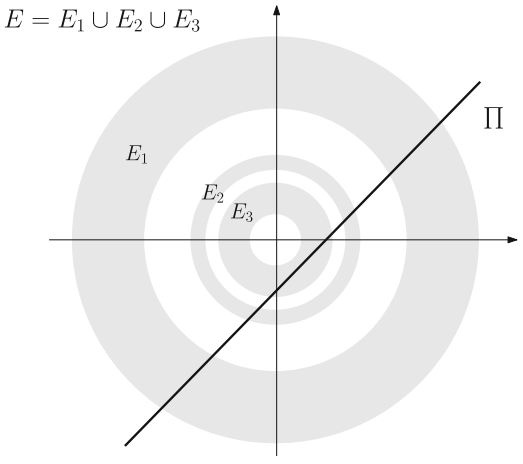
Inequality (2) - radial functions

The key point in the proof of Theorem 2 is based on an endpoint estimate corresponding to the indices $p = n/k$, $q = \infty$, and $r = \infty$, as stated in the following lemma.

Lemma 1 (Duoandikoetxea–Naibo–Oruetxebarria [20]). *If $E \subset \mathbb{R}^n$ is radially symmetric and Π is a translate of a k-plane in $\mathbb{R}^n$ then*

$$\lambda_k(E \cap \Pi) \leq C_{k,n}|E|^{k/n}.$$

The figure below illustrates Lemma 1 when $n = 2$ (and therefore $k = 1$): the length of the intersection of the line Π with the set E given by the union of the annuli is controlled by the square root of the area of E.



The proof of Theorem 2 supplies the following estimates for the pairs $(\frac{1}{p}, \frac{1}{r})$ on the borders of the regions depicted above for inequalities (1) and (2):

- Noting that the left-hand side of the inequality in Lemma 1 corresponds to the k-plane transform of χ_E, we can rewrite it as

$$\sup_{x\in\mathbb{R}^n,\, \pi\in G(n,k)} P_k\chi_E(x,\pi) \le C_{k,n}\|\chi_E\|_{L^{\frac{n}{k},1}(\mathbb{R}^n)}.$$

  This, in turn, implies that P_k is bounded, when restricted to radial functions, from $L^{n/k,1}(\mathbb{R}^n)$ into $L^\infty(L^r)$ (in the spirit of inequality (1) and from $L^{n/k,1}(\mathbb{R}^n)$ into $L^r(L^\infty)$ (in the spirit if inequality (2), $1 \le r \le \infty$.
- The identity

$$\int_{\mathbb{R}^n} h(x)\,\mathrm{d}x = \int_{G(n,n-k)}\int_{\pi^\perp} |y|^{n-k} h(y)\,\mathrm{d}\lambda_k(y)\mathrm{d}\gamma_{n,n-k}(\pi)$$

  (see Solmon [32]) and the identification of the manifolds $G(n,k)$ and $G(n,n-k)$ through the correspondence $\pi \to \pi^\perp$ yield

$$\int_{G(n,k)} P_k f(x,\pi)\,\mathrm{d}\gamma_{n,k}(\pi) = c\; I_k f(x),$$

  where I_k is the Riesz potential of order k. The boundedness properties of I_k then show (1) for $r = 1$, $1 < p < \frac{n}{k}$, $\frac{n}{p} - \frac{n}{q} = k$, and give a weak-type estimate for $r = 1$, $p = 1$, $q = \frac{n}{n-k}$.
- In relation to inequality (1), P_k is bounded from the class of radial functions in $L^{p,1}(\mathbb{R}^n)$ into $L^{q,\infty}(L^r)$ for $\frac{n-k}{r} = \frac{n}{p} - k$ and $\frac{n}{p} - \frac{n}{q} = k$.

We refer the reader to Kumar–Ray [25] for weighted versions of Theorem 2 with power weights, and to Kumar–Ray [24] for weighted mixed norm inequalities for the Radon transform on general functions.

5 Mixed-Norm Inequalities for Potential-Type Operators Associated to k-planes

The X-ray transform is an element in a scale of potential-type directional operators which appear when the method of rotations is applied to integral operators with variable kernels that have the homogeneity of the Riesz potentials. Indeed, for $0 < \alpha < n$, we have

$$I_{\alpha,\Omega} f(x) = \int_{\mathbb{R}^n} \frac{\Omega(x,y')}{|y|^{n-\alpha}}\, f(x-y)\,\mathrm{d}y = \frac{1}{2}\int_{S^{n-1}} \Omega(x,u)\, P_\alpha f(x,u)\,\mathrm{d}u,$$

where

$$P_\alpha f(x,u) = \int_{-\infty}^{\infty} f(x-tu)|t|^{\alpha-1}\,\mathrm{d}t.$$

Then the X-ray transform corresponds to $\alpha = 1$. Analogously to what was explained in Sect. 2.2, boundedness properties of $I_{\alpha,\Omega}$ on Lebesgue spaces follow from mixed-norm estimates of the type (1) for $P_\alpha f(x,u)$. More generally, Duoandikoetxea–Naibo–Oruetxebarria [20] considered mixed-norm inequalities for potential-like operators supported on k-planes of the form

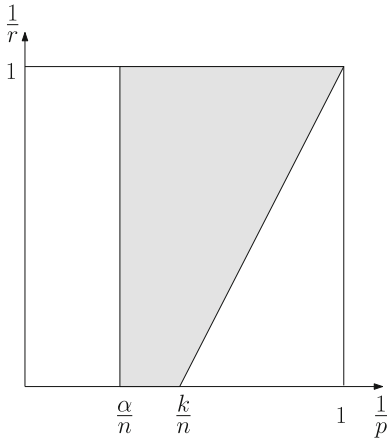
$$P_{k,\alpha} f(x,\pi) = \int_\pi f(x-y)|y|^{\alpha-k}\,\mathrm{d}\lambda_k(y), \qquad 0 < \alpha \le n,$$

(then $P_\alpha = P_{1,\alpha}$) and proved the following theorem.

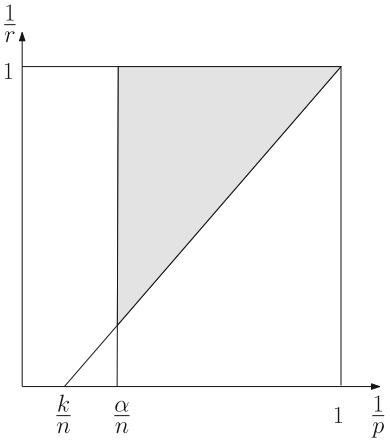
Theorem 3 (Duoandikoetxea–Naibo–Oruetxebarria [20]). *For radial functions inequality (1) holds for $P_{k,\alpha}$ if and only if*

$$1 < p < \frac{n}{\alpha}, \quad \frac{n}{p} - \frac{n}{q} = \alpha, \quad \frac{n-k}{r} > \frac{n}{p} - k\,.$$

The following figures illustrate the regions given by Theorem 3 for the indices p and r according to the cases $\alpha < k$ and $\alpha \ge k$.



Theorem 3, $\alpha < k$



Theorem 3, $\alpha \ge k$

When $\alpha < k$, the proof of Theorem 3 is based on Lemma 1 and boundedness properties of the Hardy–Littlewood maximal operator associated to k-planes. This maximal operator is defined by

$$M_k f(x,\pi) = \sup_{R>0} \frac{1}{R^k} \int_{\{y\in\pi:|y|<R\}} |f(x-y)|\,\mathrm{d}\lambda_k(y), \quad x \in \mathbb{R}^n,\ \pi \in G(n,k),$$

which coincides with the directional maximal operator of Sect. 2.1 when $k = 1$. Indeed, Lemma 1 allows to get an endpoint estimate corresponding to $p = \frac{n}{\alpha}$, $r = \infty, and\ q = \infty$, mainly,

$$\sup_{x\in\mathbb{R}^n,\,\pi\in G(n,k)} P_{k,\alpha}\chi_E(x,\pi) \le C\,|E|^{\alpha/n}, \quad E \text{ radially symmetric},$$

and the following version of Hedberg's inequality [22] for potential operators associated to k-planes,

$$P_{k,\alpha}\chi_E(x,\pi) \le C\, M_k\chi_E(x,\pi)^{1-\alpha/k}\,|E|^{\alpha/n}.$$

This last inequality and boundedness properties of M_k when restricted to radial functions (see part (i) of Theorem 4) give an endpoint estimate for $P_{k,\alpha}$ corresponding to $p = \frac{n}{k}, r = \infty, and\ q = n/(k-\alpha)$. More precisely, $P_{k,\alpha}$ restricted to radial functions is bounded from $L^{n/k,1}(\mathbb{R}^n)$ into $L^{n/(k-\alpha),\infty}(L^\infty)$.

All of the above allows to get appropriate weak estimates for indices p, r, q such that $(\frac{1}{p}, \frac{1}{r})$ is on the segments with endpoints $(\frac{\alpha}{n}, 0)$ and $(\frac{\alpha}{n}, 1)$ or endpoints $(\frac{k}{n}, 0)$ and $(1, 1)$, respectively, and q is given by the scaling relation $\frac{n}{p} - \frac{n}{q} = \alpha$. Finally, for each fixed r real interpolation between Lorentz spaces yields Theorem 3 when $\alpha < k$.

The case $\alpha \ge k$ uses Lemma 1 and the pointwise inequality for nonnegative f, $x \in \mathbb{R}^n$ and $\pi \in G(n,k)$ given by

$$P_{k,\alpha}f(x,\pi) \le P_{k,\beta}f(x,\pi)^{1-s}\,P_{k,\gamma}f(x,\pi)^s, \quad \alpha = (1-s)\beta + s\gamma,$$
$$0 < \beta < \alpha < \gamma \le n,$$

applied with $\beta = k$ and $\gamma = n$ to get that $P_{k,\alpha}$ is bounded, when restricted to radial functions, from $L^{n/\alpha,1}(\mathbb{R}^n)$ into $L^\infty(L^{(n-k)/(\alpha-k)})$. The indices $p = \frac{n}{\alpha}$ and $r = (n-k)/(\alpha-k)$ correspond to the point of intersection of the segments with endpoints $(\frac{\alpha}{n}, 0)$ and $(\frac{\alpha}{n}, 1)$ and endpoints $(\frac{k}{n}, 0)$ and $(1, 1)$. The rest of the proof follows similarly to that of Theorem 2.

Almost sharp versions (in terms of the conditions on p, q, and r being necessary and sufficient) of Theorem 3 for $k = 1$ and $k = n-1$ in the non-radial case were proved by Duoandikoetxea–Oruetxebarria in [16] and [17], respectively.

6 The Hardy–Littlewood Maximal Operator Associated to k-planes

In this section, we study boundedness properties of the maximal operator

$$\mathscr{M}_k f(x) := \sup_{\pi\in G(n,k)} M_k f(x,\pi),$$

with M_k as introduced in Sect. 5. The operator, $\mathscr{M}_1$ is known as the universal maximal operator and it follows that

$$\mathscr{M}_1 f(x) \sim \sup_{x \in R} \frac{1}{|R|} \int_R |f(y)| \, \mathrm{d}y, \quad x \in \mathbb{R}^n,$$

where R is a parallelepiped in $\mathbb{R}^n$. A construction using Besicovitch-type sets allows to show that $\mathscr{M}_k$ is unbounded on $L^p(\mathbb{R}^n)$, $1 \leq p < \infty$. However, this situation changes for $p > n/k$ when restricting to radial functions as stated in the following theorem.

Theorem 4 (Duoandikoetxea–Naibo–Oruetxebarria [20]). *(i) The operator $\mathscr{M}_k$ is of restricted weak-type $(\frac{n}{k}, \frac{n}{k})$ when restricted to radial functions; this is, there exists a constant C such that*

$$|\{x \in \mathbb{R}^n : \mathscr{M}_k f(x) > \lambda\}| \leq C \left(\frac{\|f\|_{L^{\frac{n}{k},1}(\mathbb{R}^n)}}{\lambda} \right)^{\frac{n}{k}},$$

for all radial functions $f \in L^{\frac{n}{k},1}(\mathbb{R}^n)$ and $\lambda > 0$.

(ii) If $p > n/k$ then $\mathscr{M}_k$ is bounded on $L^p(\mathbb{R}^n)$ when restricted to radial functions; this is, there exists a constant C such that

$$\|\mathscr{M}_k f\|_{L^p(\mathbb{R}^n)} \leq C \, \|f\|_{L^p(\mathbb{R}^n)}$$

for all radial functions $f \in L^p(\mathbb{R}^n)$.

The case $k = 1$ of Theorem 4 was treated in Carbery–Hernández–Soria [6]. However, the methods of proofs in [20] are different and unify well for all $k \geq 1$. We recall that part (i) is an essential estimate in our proof of Theorem 3. The proof of Theorem 4 is a consequence of the boundedness properties of the Hardy–Littlewood maximal operator and the following pointwise inequality:

Theorem 5 (Duoandikoetxea–Naibo–Oruetxebarria [20]). *If $E \subset \mathbb{R}^n$ is a radially symmetric set of finite measure, then*

$$M_k \chi_E(x, \pi) \leq C \, M_{hl} \chi_E(x)^{k/n}, \quad x \in \mathbb{R}^n, \pi \in G(n,k), \tag{5}$$

where M_{hl} denotes the usual Hardy–Littlewood maximal operator in $\mathbb{R}^n$ and the constant C depends only on n and k.

An elementary argument was used in [20] to prove that there is a constant C depending only on k and n such that

$$\frac{\lambda_k(B \cap E)}{\lambda_k(B)} \leq C \left(\frac{|A[B] \cap E|}{|A[B]|} \right)^{k/n} \tag{6}$$

for all radially symmetric sets E in $\mathbb{R}^n$, all k-balls B lying on translates of k-planes and where

$$A[B] := \{x \in \mathbb{R}^n : |x| = |y| \text{ for some } y \in B\}.$$

Inequality (6) then says that

$$M_k \chi_E(x,\pi) \le C \ (\mathscr{A}\chi_E(x))^{k/n}, \quad x \in \mathbb{R}^n, \pi \in G(n,k),$$

where $\mathscr{A}$ is the maximal operator over annuli on $\mathbb{R}^n$, this is,

$$\mathscr{A} f(x) = \sup_{x \in A_{a,b},\, 0<a<b<\infty} \frac{1}{|A_{a,b}|} \int_{A_{a,b}} |f(y)|\, \mathrm{d}y,$$

with $A_{a,b} := \{x \in \mathbb{R}^n : a < |x| < b\}$. Theorem 5 then follows after observing that

$$\mathscr{A} f(x) \sim M_{hl} f(x), \quad x \in \mathbb{R}^n, \ f \text{ radial}.$$

A by-product of the pointwise inequality (5) is a weighted version of Theorem 4. We recall that if $1 < p < \infty$, the weight w is in the Muckenhoupt class A_p if there exists a constant C such that

$$\left(\frac{1}{|B|}\int_B w(x)\,\mathrm{d}x\right)\left(\frac{1}{|B|}\int_B w(x)^{1-p'}\,\mathrm{d}x\right)^{p-1} \le C,$$

for all balls $B \subset \mathbb{R}^n$. The class A_1 is defined as the class of weights w such that

$$M_{hl} w(x) \sim w(x), \quad x \in \mathbb{R}^n.$$

Let $L^p_w(\mathbb{R}^n)$ be the L^p spaces based on $\mathbb{R}^n$ with respect to the measure $w(x)\mathrm{d}x$, and $L^{p,q}_w(\mathbb{R}^n)$ be the corresponding Lorentz spaces. Then M_{hl} is bounded on $L^p_w(\mathbb{R}^n)$, $1 < p < \infty$, if and only if $w \in A_p$, and from $L^1_w(\mathbb{R}^n)$ into $L^{1,\infty}_w(\mathbb{R}^n)$ if and only if $w \in A_1$. As a consequence we obtain:

Corollary 1 (Weighted version of Theorem 4). *(i) If $w \in A_1$, the operator $\mathscr{M}_k$ is of restricted weak-type $(\frac{n}{k}, \frac{n}{k})$ with respect to the measure $w(x)dx$ when restricted to radial functions; that is, there exists a constant C such that*

$$\int_{\{x \in \mathbb{R}^n :\, \mathscr{M}_k f(x) > \lambda\}} w(x)\,\mathrm{d}x \le C \left(\frac{\|f\|_{L^{\frac{n}{k},1}_w(\mathbb{R}^n)}}{\lambda}\right)^{\frac{n}{k}},$$

for all radial functions $f \in L^{\frac{n}{k},1}_w(\mathbb{R}^n)$ and $\lambda > 0$.

(ii) If $p > n/k$ and $w \in A_{pk/n}$ then $\mathscr{M}_k$ is bounded on $L^p_w(\mathbb{R}^n)$ when restricted to radial functions; that is, there exists a constant C such that

$$\|\mathscr{M}_k f\|_{L^p_w(\mathbb{R}^n)} \le C\, \|f\|_{L^p_w(\mathbb{R}^n)}$$

for all radial functions $f \in L^p_w(\mathbb{R}^n)$.

In Duoandikoetxea–Naibo [18] inequality (5) was extended to other classes of functions in the case of the universal maximal operator ($k = 1$)

$$\mathscr{M}_1 f(x) = \sup_{u\in S^{n-1}} \sup_{R>0} \frac{1}{R} \int_{-R}^{R} |f(x-tu)|\, \mathrm{d}t.$$

We say that a function f defined on $\mathbb{R}^n$ is l_q-radial if $f(x) = f_0(|x|_q)$, where f_0 is a function defined on $(0,\infty)$ and $|x|_q = (\sum_{j=1}^n |x_j|^q)^{1/q}$, $x = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$.

Theorem 6 (Duoandikoetxea–Naibo [18]). *Let $1 < q \le n$ and let $E \subset \mathbb{R}^n$ be an l^q-radial set of finite measure. Then*

$$\mathscr{M}_1 \chi_E(x) \le C\, M_{hl}\chi_E(x)^{1/n}, \quad x \in \mathbb{R}^n, \tag{7}$$

where the constant C depends only on n and q.

As in the case $q = 2$, the proof of (7) combines an inequality with the maximal operator over l_q-annuli and the pointwise equivalence of this last operator with the Hardy–Littlewood maximal operator when restricted to l^q-radial functions. Inequality (7) does not hold when $q = 1$ or $q > n$. Indeed, let δ be a small positive number, E be the l_q-annulus with inner radius $1-\delta$ and outer radius $1+\delta$, and $x = (2, 1, 0, \cdots, 0)$; then $\mathscr{M}_1\chi_E(x) \sim \delta^{1/q}$ for $q > n$, $\mathscr{M}_1\chi_E(x) \sim 1$ for $q = 1$, and $M_{hl}\chi_E(x) \sim \delta$.

Corollary 1. *Let $1 < q \le n$. When restricted to l^q-radial functions, $\mathscr{M}_1$ is bounded on $L^p(\mathbb{R}^n)$ for $p > n$ and unbounded for $p \le n$. For $p = n$ it is of restricted weak type.*

Unboundedness for $p \le n$ follows when considering χ_E where E is the unit l^q-ball; indeed, $\mathscr{M}_1\chi_E(x) \sim |x|_q^{-1}$ for big x, and therefore $\mathscr{M}_1\chi_E \notin L^p(\mathbb{R}^n)$ if $p \le n$. In addition, $\mathscr{M}_1$ is not of weak-type (n,n) when restricted to l_q-radial functions. This follows from previous mentioned examples adapted to the l_q-norm setting; for instance, $f(x) = |x|_q^{-1}(\log|x|_q)^{-1}\chi_{\{|x|_q>e\}}(x)$ is in $L^n(\mathbb{R}^n)$, but $\mathscr{M}_1 f(x) = \infty$.

A substitute of Theorem 6 for the cases $q = 1$ and $q = \infty$ using a larger maximal operator was also proved in Duoandikoetxea–Naibo [18]. This result also implies sharp L^p estimates, $p > n$, for $\mathscr{M}_1$ when restricted to l^∞- and l^1-radial functions.

Duoandikoetxea–Naibo [18] extended Lemma 1 to l^q-radial functions for $1 < q \le n$ and showed that it is not true for $q = 1$ or $q > n$. As a consequence, Theorem 3 also holds for $P_{1,\alpha}$ when restricted to the class of l^q-radial functions, $1 < q \le n$, since its proof is based on Corollary 1 and this extension of Lemma 1.

We refer the reader to Duoandikoetxea–Moyua–Oruetxebarria [19] for a sharper version of inequality (7), which allows for general radial functions (for $q = 2$) and uses the two-dimensional Hardy–Littlewood maximal operator as an upper bound in all dimensions.

7 Application: Bounds for the Kakeya Maximal Operator

Given $N > 0$, let $\mathscr{R}_N$ be the set of all parallelepipeds in $\mathbb{R}^n$ with $n-1$ sides of length a and the remaining side of length Na, for some $a > 0$. The Kakeya maximal operator is defined by

$$\mathscr{K}_N f(x) = \sup_{x \in R \in \mathscr{R}_N} \frac{1}{|R|} \int_R |f(y)|\, \mathrm{d}y, \quad x \in \mathbb{R}^n.$$

Since every parallelepiped with $n-1$ sides of length a and the remaining side of length Na, for some $a > 0$, is contained in a ball of radius comparable to Na, then $\mathscr{K}_N$ is pointwise dominated by the Hardy–Littlewood maximal operator; indeed, there exists a constant C that depends only on n such that

$$\mathscr{K}_N f(x) \le C\, N^{n-1} M_{hl} f(x), \quad x \in \mathbb{R}^n.$$

It then follows that $\mathscr{K}_N$ is bounded on $L^p(\mathbb{R}^n)$ for $1 < p \le \infty$ and is of weak-type $(1,1)$.

A very important problem in harmonic analysis consists in estimating the norm of the operator $\mathscr{K}_N$, $\|\mathscr{K}_N\|_{L^p \to L^p}$, as a bounded operator on $L^p(\mathbb{R}^n)$. The above pointwise inequality gives a bound for the weak-type $(1,1)$ operator norm proportional to N^{n-1}. Interpolation with the $L^\infty(\mathbb{R}^n)$ operator norm gives that

$$\|\mathscr{K}_N\|_{L^p \to L^p} \lesssim N^{\frac{n-1}{p}}, \quad 1 < p < \infty.$$

It is conjectured that $\mathscr{K}_N$ is bounded on $L^p(\mathbb{R}^n)$ with norm majorized as

$$\|\mathscr{K}_N\|_{L^p \to L^p} \le \begin{cases} C(p)(\log N)^{a(p)}, & \text{for some } a(p) > 0 \text{ if } p \ge n, \\ C(p) N^{n/p-1} (\log N)^{a(p)}, & \text{for some } a(p) \ge 0 \text{ if } 1 < p < n. \end{cases}$$

This conjecture is related to the problem of determining the Hausdorff dimension of the Kakeya set and the boundedness properties of Bochner–Riesz multipliers. The conjecture has been proved when $n = 2$ in Córdoba [9] and when $n \ge 3$ and $1 < p \le (n+2)/2$ (Córdoba [10], Christ–Duoandikoetxea–Rubio de Francia [8], Bourgain [1, 2], Wolff [34], Katz–Tao [23]). The conjecture is also related to sharp mixed-norm estimates for the directional maximal operator M_u introduced in Sect. 2.1 (see Duoandikoetxea [15]).

Let $S \subset \mathbb{R}^n$ be a star-shaped set with respect to the origin (i.e., if $x \in S$, then the segment joining x with the origin is contained in S). Then S can be described in polar coordinates as $S \setminus \{0\} = \{(\rho, u) \in (0,\infty) \times S^{n-1} : 0 < \rho < F_S(u)\}$, $|S| = \frac{1}{n}\int_{S^{n-1}} F_S(u)^n \, \mathrm{d}u$, and

$$\begin{aligned}\int_S |f(x-y)| \, \mathrm{d}y &= \int_{S^{n-1}} \int_0^{F_S(u)} |f(x-\rho u)| \rho^{n-1} \, \mathrm{d}\rho \, \mathrm{d}u \\ &\leq \int_{S^{n-1}} F_S(u)^n \, M_1 f(x,u) \, \mathrm{d}u \leq n \, |S| \, \mathscr{M}_1 f(x).\end{aligned}$$

This inequality implies that the Kakeya maximal operator is pointwise controlled by the universal maximal operator,

$$\mathscr{K}_N f(x) \leq n \, \mathscr{M}_1 f(x), \quad x \in \mathbb{R}^n.$$

As a consequence of the boundedness properties of $\mathscr{M}_1$ given in Corollary 1 on classes of radial functions we obtain the conjecture on the bounds of the Kakeya maximal operator when restricted to l^q-radial functions.

Theorem 7 (Duoandikoetxea–Naibo [18]). *Fix $1 < q \leq n$.*

- *If $p > n$ there exists a constant C independent of N such that*

$$\|\mathscr{K}_N f\|_{L^p(\mathbb{R}^n)} \leq C \, \|f\|_{L^p(\mathbb{R}^n)},$$

  for all l^q-radial functions $f \in L^p(\mathbb{R}^n)$.
- *There exists a constant C independent of N such that*

$$\sup_{\lambda>0} \lambda |\{x : |\mathscr{K}_N f(x)| > \lambda\}|^{\frac{1}{n}} \leq C \, \|f\|_{L^{n,1}(\mathbb{R}^n)},$$

  for all l^q-radial functions $f \in L^{n,1}(\mathbb{R}^n)$, (restricted weak-type (n,n)).
- *There exists a constant C independent of N such that*

$$\sup_{\lambda>0} \lambda |\{x : |\mathscr{K}_N f(x)| > \lambda\}|^{\frac{1}{n}} \leq C \, (\log N)^{1-\frac{1}{n}} \, \|f\|_{L^n(\mathbb{R}^n)},$$

  for all l^q-radial functions $f \in L^n(\mathbb{R}^n)$, (weak-type (n,n)).
- *There exists a constant C independent of N such that*

$$\|\mathscr{K}_N f\|_{L^n(\mathbb{R}^n)} \leq C \, \log N \, \|f\|_{L^n(\mathbb{R}^n)},$$

  for all l^q-radial functions $f \in L^n(\mathbb{R}^n)$.

The case $q = 2$ (radial functions) of this theorem was already treated in Carbery–Hernandez–Soria [6]. An analogous result for l^1-radial and l^∞-radial functions was

also proved in Duoandikoetxea–Naibo [18] but with worse exponents for $\log N$ in the weak-type and strong-type estimates corresponding to $p = n$.

Acknowledgments Javier Duoandikoetxea's research is supported in part by grant MTM2007-62186 of MEC (Spain) and FEDER. Virginia Naibo's research is supported in part by the National Science Foundation under grant DMS 1101327. Virginia Naibo thanks the members of the Organizing Committee (Profs. Radu Balan, John J. Benedetto, Wojtek Czaja, and Kasso Okoudjou) for the invitation to speak in the February Fourier Talks 2009.

References

1. Bourgain, J.: Besicovitch type maximal operators and applications to Fourier analysis. Geom. Funct. Anal. **1**(2), 147–187 (1991)
2. Bourgain, J.: On the dimension of Kakeya sets and related maximal inequalities. Geom. Funct. Anal. **9**(2), 256–282 (1999)
3. Calderón, A.P.: On the Radon transform and some of its generalizations. In: Conference on Harmonic Analysis in Honor of Antoni Zygmund, Vol. I, II, Chicago, Illinois, 1981, pp. 673–689. Wadsworth Mathematical Series, Wadsworth, Belmont, CA (1983)
4. Calderón, A.P., Zygmund, A.: On singular integrals. Amer. J. Math. **78**, 289–309 (1956)
5. Calderón, A.P., Zygmund, A.: On singular integrals with variable kernels. Appl. Anal. **7**, 221–238 (1978)
6. Carbery, A., Hernández, E., Soria, F.: Estimates for the Kakeya maximal operator on radial functions in $\mathbb{R}^n$. Harmonic Analysis, Sendai, 1990, pp. 41–50. ICM-90 Satellite Conference Proceedings. Springer, Tokyo (1991)
7. Christ, M.: Estimates for the k-plane transform. Indiana Univ. Math. J. **33**, 891–910 (1984)
8. Christ, M., Duoandikoetxea, J., Rubio de Francia, J.L.: Maximal operators related to the Radon transform and the Calderón–Zygmund method of rotations. Duke Math. J. **53**, 189–209 (1986)
9. Córdoba, A.: The Kakeya maximal function and the spherical summation multipliers. Amer. J. Math. **99**(1), 1–22 (1977)
10. Córdoba, A.: A note on Bochner–Riesz operators. Duke Math. J. **46**(3), 505–511 (1979)
11. Cowling, M., Mauceri, G.: Inequalities for some maximal functions I. Trans. Amer. Math. Soc. **287**(2), 431–455 (1985)
12. Drury, S.W.: L^p estimates for the X-ray transform. Illinois J. Math. **27**(1), 125–129 (1983)
13. Drury, S.W.: Generalizations of Riesz potentials and L^p estimates for certain k-plane transforms. Illinois J. Math. **28**(3), 495–512 (1984)
14. Drury, S.W.: A survey of k-plane transform estimates. Commutative Harmonic Analysis, Canton, NY, 1987, pp. 43–55. Contemporary Mathematics, vol. 91. American Mathematical Society, Providence (1989)
15. Duoandikoetxea, J.: Directional operators and mixed norms. Proceedings of the 6th International Conference on Harmonic Analysis and Partial Differential Equations, El Escorial, 2000. Publications Mathematics, vol. Extra, pp. 39–56 (2002)
16. Duoandikoetxea, J., Oruetxebarria, O.: Mixed norm inequalities for directional operators associated to potentials. Potential Anal. **15**, 273–283 (2001)
17. Duoandikoetxea, J., Oruetxebarria, O.: Mixed norm estimates for potential operators related to the Radon transform. J. Aust. Math. Soc. **84**(2), 181–191 (2008)
18. Duoandikoetxea, J., Naibo, V.: The universal maximal operator on special classes of functions. Indiana Univ. Math. J. **54**(5), 1351–1369 (2005)
19. Duoandikoetxea, J., Moyua, A., Oruetxebarria, O.: The spherical maximal operator on radial functions. J. Math. Anal. Appl. **387**(2), 655–666 (2012)

20. Duoandikoetxea, J., Naibo, V., Oruetxebarria, O.: k-plane transforms and related operators on radial functions. Michigan Math. J. **49**, 265–276 (2001)
21. Fefferman, R.: On an operator arising in the Calderón-Zygmund method of rotations and the Bramble–Hilbert lemma. Proc. Nat. Acad. Sci. U.S.A. **80**(12), Phys. Sci., 3877–3878 (1983)
22. Hedberg, L.: On certain convolution inequalities. Proc. Amer. Math. Soc. **36**, 505–510 (1972)
23. Katz, N., Tao, T.: New bounds for Kakeya problems. J. Anal. Math. **87**, 231–263 (2002)
24. Kumar, A., Ray, S.K.: Mixed norm estimate for Radon transform on L^p spaces. Proc. Indian Acad. Sci. (Math. Sci.) **120**(4), 441–456 (2010)
25. Kumar, A., Ray, S.K.: Weighted estimates for the k-plane transform of radial functions on Euclidean spaces. Israel J. Math. DOI: 10.1007/s11856-011-0091-8
26. Markoe, A.: Analytic tomography. Encyclopedia of Mathematics and its Applications, vol. 106. Cambridge University Press, Cambridge (2006)
27. Oberlin, D.M., Stein, E.M.: Mapping properties of the Radon transform. Indiana Univ. Math. J. **31**(5), 641–650 (1982)
28. Radon, J.: Über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten. Ber. Verh. Sächs. Akad. Wiss. Leipzig, Math. Nat. Kl. **69**, 262–277 (1917)
29. Radon, J.: On the determination of functions from their integral values along certain manifolds. IEEE Trans. Med. Imaging **MI-5**(4), 170–176 (1986). Translated by P.C. Parks from the original German text, with corrections
30. Rubin, B.: Reconstruction of functions from their integrals over k-planes. Israel J. Math. **141**, 93–117 (2004)
31. Smith, K., Solmon, D.C.: Lower dimensional integrability of L^2 functions. J. Math. Anal. Appl. **51**(3), 539–549 (1975)
32. Solmon, D.C.: The X-ray transform. J. Math. Anal. Appl. **56**, 61–83 (1976)
33. Solmon, D.C.: A note on k-plane integral transforms. J. Math. Anal. Appl. **71**(2), 351–358 (1979)
34. Wolff, T.: An improved bound for Kakeya type maximal functions. Rev. Mat. Iberoamericana **11**(3), 651–674 (1995)
35. Wolff, T.: A mixed norm estimate for the X-ray transform. Rev. Mat. Iberoamericana **14**(3), 561–600 (1998)

Representation of Linear Operators by Gabor Multipliers

Peter C. Gibson, Michael P. Lamoureux, and Gary F. Margrave

Abstract We consider a continuous version of Gabor multipliers: operators consisting of a short-time Fourier transform, followed by multiplication by a distribution on phase space (called the Gabor symbol), followed by an inverse short-time Fourier transform, allowing different localizing windows for the forward and inverse transforms. This chapter focuses on the following broad questions. Firstly, for a given pair of forward and inverse windows, which linear operators can be represented as a Gabor multiplier, and what is the relationship between the Kohn–Nirenberg symbol of such an operator and the corresponding Gabor symbol? We answer this question completely. Secondly, for a linear operator of a given type, can windows be specially chosen, or "tuned", to suit the operator so that the Gabor symbol reflects the operator's type? In studying this latter question for product-convolution operators, we derive a new class of "extreme value" windows that, with respect to the representation of linear operators, are more general than standard Gaussian windows while sharing many of Gaussian windows' desirable properties. The results in this chapter help to justify techniques developed for seismic imaging that use Gabor multipliers to represent nonstationary filters and wavefield extrapolators.

P.C. Gibson (✉)
Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, ON, Canada, M3J1P3
e-mail: pcgibson@yorku.ca

M.P. Lamoureux
Department of Mathematics and Statistics, University of Calgary, 2500 University Drive NW, Calgary, AB, Canada, T2N1N4
e-mail: mikel@ucalgary.ca

G.F. Margrave
Department of Geoscience, University of Calgary, 2500 University Drive NW, Calgary, AB, Canada, T2N1N4
e-mail: margrave@ucalgary.ca

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_12,

Keywords Gabor multiplier • Gabor symbol • Kohn–Nirenberg operator • Schwartz kernel • Short-time Fourier transform • Analysis window • Synthesis window • Window pair • Compatible window pair • Extreme value window • Spreading function • Underspread operator • Symplectic Fourier transform

1 Introduction

1.1 Overview

We are interested in exploiting the short-time Fourier transform, also known as the Gabor transform, to establish a general framework for evaluating linear operators. A Gabor multiplier is an operator consisting of a short-time Fourier transform, followed by multiplication by a distribution on phase space—called the Gabor symbol—followed by an inverse short-time Fourier transform. Such operators are also known as localization operators, or anti-Wick operators, depending on the context. (See, for instance, [2], [5, Chap. 2], and the references therein.) We envisage a two-step scheme for numerical evaluation of a linear operator L: (1) represent L as a Gabor multiplier and (2) numerically evaluate the Gabor multiplier on given data. The scheme is predicated on the existence of fast algorithms to evaluate forward and inverse Gabor transforms, which we discuss in detail elsewhere [9, 11]. The present chapter is devoted to step (1). Thus we concentrate entirely on the theoretical issue of precisely which linear operators may be represented as Gabor multipliers based on given windows. In addition we study the subsidiary question of how to adapt windows to suit the class of operators at hand. As a testing ground for the latter analysis, we focus on product-convolution operators, which are closely related to partial differential operators. This work was motivated by two particular applications in seismic imaging, nonstationary filtering and wavefield extrapolation. In both cases Gabor methods have yielded major improvements, even before the underlying mathematics had been fully understood; details of the particular applications appear in [10–13]. The present chapter supplies the mathematical analysis needed to properly justify some of these newly developed techniques.

This chapter is organized as follows. In Sects. 1.2 and 1.3, we fix our notation and introduce the well-established Kohn–Nirenberg formalism, which provides a convenient framework in which to discuss general linear operators. In Sect. 1.4 we discuss a particular class, product-convolution operators, which provides some of the rationale for Sect. 4.

Our principal object of study, Gabor multipliers, is defined precisely in Sect. 2.1. In Sect. 2.2 we derive their Schwartz kernels, and then in Sect. 2.3 we derive the key equation, Theorem 2, that relates the Gabor symbol of a linear operator to its Kohn–Nirenberg symbol.

In Sect. 3 we use Theorem 2 to describe precisely the class of operators that may be represented as Gabor multipliers based on a given pair of windows. We develop these results further for the special case of Gaussian windows and then compactly supported windows in Sects. 3.2 and 3.3, respectively.

Based on the results established in Sect. 3 for Gaussian windows, in Sect. 4 we take up the problem of studying a wider class windows that are suited to Gabor multiplier representation of product-convolution operators. More precisely, we investigate window pairs that, like Gaussians, respect the separation of time and frequency inherent in product-convolution operators. This leads, via a classification discussed in Sect. 4.2, to the unexpected emergence of what we term "extreme value" windows. These are analysed and compared to Gaussians in Sect. 4.3. Our final result, Theorem 4, shows that extreme value windows, while sharing some of the desirable properties of Gaussians, are in certain respects superior. We anticipate that they will be useful windows in applications of Gabor analysis.

Finally, in Sect. 5, we give a brief summary.

1.2 Notation and Conventions

Our notation is mostly standard, with the exceptions that: (1) we use the version of the Fourier transform that has a factor of 2π in the exponent, and (2) we take tempered distributions to be continuous, *conjugate* linear, rather than linear, functionals on the space of Schwartz class functions. Generally speaking, we deal with functions and distributions on $\mathbb{R}^n$, where the value of n is fixed within a given context, and $\mathbb{R}^n$ is always the domain of integration, which we omit. We indicate a point in $\mathbb{R}^{2n}$ by a pair (x, y) of points $x, y \in \mathbb{R}^n$, and integration over $\mathbb{R}^{2n}$ is indicated by a pair of integral signs. For convenient reference we have compiled in Table 1 a list of some of the function spaces and operators that we use repeatedly.

We work within the basic framework of $\mathscr{L}_n$, the space of continuous, linear operators

$$L : \mathscr{S}_n \to \mathscr{S}'_n,$$

making frequent use of the correspondence between $\mathscr{L}_n$ and $\mathscr{S}'_{2n}$ [14]. More precisely, given a linear operator $L \in \mathscr{L}_n$, there is a unique distribution $K = K(L) \in \mathscr{S}'_{2n}$, its Schwartz kernel, that satisfies the equation

$$\langle K, \theta \otimes \bar{\varphi} \rangle = \langle L\varphi, \theta \rangle \qquad \forall \varphi, \theta \in \mathscr{S}_n. \tag{1}$$

And conversely, given $K \in \mathscr{S}'_{2n}$, the Eq. (1) evidently determines a unique $L \in \mathscr{L}_n$. (Here $\otimes$ denotes the tensor product: $f \otimes g(x, y) = f(x)g(y)$.) The adjoint of a linear operator $L \in \mathscr{L}_n$ is the linear operator $L^* \in \mathscr{L}_n$ defined by the equation

$$\langle L^*\varphi, \theta \rangle = \overline{\langle L\theta, \varphi \rangle} \qquad \forall \varphi, \theta \in \mathscr{S}_n. \tag{2}$$

Table 1 Notation

$\mathscr{S}_n$	Function and distribution spaces Schwartz class functions $\varphi : \mathbb{R}^n \to \mathbb{C}$
$\mathscr{S}'_n$	(conjugate linear) tempered distributions $u : \mathscr{S}_n \to \mathbb{C}$
$\mathscr{P}_n$	C^∞ functions $\varphi : \mathbb{R}^n \to \mathbb{C}$ such that $\lvert\partial^\alpha \varphi\rvert$ is bounded by a polynomial p_α for every multi-index α

Operators

Description	Symbol	Action on functions	Adjoint
Composition with a change of variables $\psi : \mathbb{R}^m \to \mathbb{R}^m$	$\mathscr{T}_\psi$	$\varphi \mapsto \varphi \circ \psi$	$\mathscr{T}_\psi^* = \lvert\det J_{\psi^{-1}}\rvert\, \mathscr{T}_{\psi^{-1}}$
Fourier transform	$\mathscr{F} : \mathscr{S}_m \to \mathscr{S}_m$ $(\mathscr{S}'_m \to \mathscr{S}'_m)$	$\varphi(x) \mapsto \widehat{\varphi}(\xi) = \int \mathrm{e}^{-2\pi \mathrm{i}\xi\cdot x}\varphi(x)\,\mathrm{d}x$	$\mathscr{F}^* = \mathscr{F}^{-1}$
Partial Fourier transform	$\mathscr{F}_1 : \mathscr{S}_{2n} \to \mathscr{S}_{2n}$ $(\mathscr{S}'_{2n} \to \mathscr{S}'_{2n})$	$\varphi(x,y) \mapsto \mathscr{F}_1\varphi(\xi,y) = \int \mathrm{e}^{-2\pi \mathrm{i}\xi\cdot x}\varphi(x,y)\,\mathrm{d}x$	$\mathscr{F}_1^* = \mathscr{F}_1^{-1}$
Partial Fourier transform	$\mathscr{F}_2 : \mathscr{S}_{2n} \to \mathscr{S}_{2n}$ $(\mathscr{S}'_{2n} \to \mathscr{S}'_{2n})$	$\varphi(x,y) \mapsto \mathscr{F}_1\varphi(x,\eta) = \int \mathrm{e}^{-2\pi \mathrm{i}\eta\cdot y}\varphi(x,y)\,\mathrm{d}y$	$\mathscr{F}_2^* = \mathscr{F}_2^{-1}$
Symplectic Fourier transform	$\mathscr{F}_s : \mathscr{S}_{2n} \to \mathscr{S}_{2n}$ $(\mathscr{S}'_{2n} \to \mathscr{S}'_{2n})$	$\varphi(x,\xi) \mapsto \widehat{\varphi^s}(u,\eta) = \iint \mathrm{e}^{2\pi \mathrm{i}(\xi\cdot u - x\cdot\eta)}\varphi(x,\xi)\,\mathrm{d}x\mathrm{d}\xi$	${\mathscr{F}_s}^* = \mathscr{F}_s$
Modulation	$M_\xi : \mathscr{S}_n \to \mathscr{S}_n$ $(\mathscr{S}'_n \to \mathscr{S}'_n)$	$\varphi(x) \mapsto \mathrm{e}^{2\pi \mathrm{i}\xi\cdot x}\varphi(x)$	$M_\xi^* = M_{-\xi}$
Translation	$T_x : \mathscr{S}_n \to \mathscr{S}_n$ $(\mathscr{S}'_n \to \mathscr{S}'_n)$	$\varphi(t) \mapsto \varphi(t-x)$	$T_x^* = T_{-x}$
Multiplication by $\lambda \in \mathscr{S}'_m$	$P_\lambda : \mathscr{P}_m \to \mathscr{S}'_m$	$\varphi \mapsto \lambda\varphi$	–

In order for the operator $\mathscr{T}_\psi$, listed in Table 1, to be well behaved, some restrictions have to be placed on ψ; in this regard we introduce the notion of a "tempered change of variables", as follows.

Definition 1. We say that a smooth, invertible map

$$\psi : \mathbb{R}^m \to \mathbb{R}^m$$

is a *tempered change of variables* if each of the operators $\mathscr{T}_\psi$, $\mathscr{T}_\psi^*$, $\mathscr{T}_{\psi^{-1}}$, *and* $\mathscr{T}_{\psi^{-1}}^*$ maps $\mathscr{S}_m$ into $\mathscr{S}_m$.

There are two tempered changes of variables on $\mathbb{R}^{2n}$ whose corresponding operators we assign special notations:

$$\mathscr{T}_- = \mathscr{T}_{\psi_-}, \text{ where } \psi_-(x,y) = (x, y-x); \tag{3}$$

$$\mathscr{T}_s = \mathscr{T}_{\psi_s}, \text{ where } \psi_s(x,y) = (y, -x). \tag{4}$$

In terms of this notation, the symplectic Fourier transform is simply the usual Fourier transform composed with the operator $\mathscr{T}_s$, i.e., $\mathscr{F}_s = \mathscr{T}_s \mathscr{F}$.

1.3 The Kohn–Nirenberg Formalism

Given an arbitrary tempered distribution $\sigma(x, \xi) \in \mathscr{S}'_{2n}$, we write $\sigma(X, D)$ for the operator defined by the formula

$$\sigma(X, D) : \mathscr{S}_n \to \mathscr{S}'_n; \qquad \sigma(X, D)\varphi(x) = \int \mathrm{e}^{2\pi \mathrm{i} x \cdot \xi} \sigma(x, \xi) \mathscr{F}\varphi(\xi)\, \mathrm{d}\xi. \quad (5)$$

(Here $X : \mathbb{R}^n \to \mathbb{R}^n$ denotes the identity map, so that, for example, $X^\alpha(x) = x^\alpha$. D stands for the differential operator $D = \frac{1}{2\pi \mathrm{i}} \partial$.) We refer to $\sigma(X, D)$ as a *Kohn–Nirenberg pseudodifferential operator*; the distribution $\sigma(x, \xi)$ is its *Kohn–Nirenberg symbol*.

In the context of classical pseudodifferential operators, where the symbol $\sigma(x, \xi)$ is required to be smooth and of bounded growth, the integral on the right-hand side of (5) is inherently well defined. A simple way to give the integral an unambiguous interpretation in the present much more general setting is to define $\sigma(X, D)$ in terms of its Schwartz kernel:

$$K(\sigma(X, D)) = \mathscr{T}_- \mathscr{F}_2 \sigma. \quad (6)$$

Since each of the operators $\mathscr{T}_-$ and $\mathscr{F}_2$ carries $\mathscr{S}'_{2n}$ bijectively onto itself, it is evident from the representation (6) that the class of Kohn–Nirenberg pseudodifferential operators on $\mathbb{R}^n$ is identical with $\mathscr{L}_n$ itself. Among the various ways to represent a linear operator, however, the Kohn–Nirenberg symbol and accompanying formal representation (5) are of particular interest since, from the physical point of view, they are natural both for partial differential operators and for nonstationary filters [7, § 14.2]. In other words, in applications one is sometimes given the Kohn–Nirenberg symbol of a linear operator directly.

Note that the Kohn–Nirenberg symbol σ_L and the Schwartz kernel $K(L)$ of a linear operator $L : \mathscr{S}_n \to \mathscr{S}'_n$ both belong to $\mathscr{S}'_{2n}$. But there is a sense in which these are distinct versions of $\mathscr{S}'_{2n}$, in that σ_L is a distribution on phase space, $\mathbb{R}^n \times \widehat{\mathbb{R}^n}$, while $K(L)$ is a distribution on the cross product $\mathbb{R}^n \times \mathbb{R}^n$ of the underlying space with itself. Indeed, since it is sometimes useful to make this distinction between space and frequency, we reserve ξ and η for frequency variables, using other letters (such as x, y, τ, t, v, w) for spatial variables. Of course here we are using the terms "space" and "frequency" in a generic sense to indicate that the variables in question are Fourier dual to one another. Depending on the particular context, "space" could represent any of space, time, or space-time, with "frequency" representing wavenumber, frequency, or wavenumber-frequency.

One last notion pertaining to the Kohn–Nirenberg formalism plays a central role in the present chapter. The *spreading function* of a linear operator $L \in \mathscr{L}_n$ is defined

to be the symplectic Fourier transform of its Kohn–Nirenberg symbol. Hence the spreading function $\widehat{\sigma}_L^s$ is a distribution on phase space that represents the operator L, just as the original symbol σ_L does.

1.4 Operators with Distinct Characteristics in Space and Frequency

The Kohn–Nirenberg formalism gives a precise notion of the "spatial" and the "frequency" structure of an operator $L \in \mathscr{L}_n$: the behaviour of the symbol $\sigma_L(x, \xi)$ in the x variables corresponds to spatial characteristics; the behaviour of $\sigma_L(x, \xi)$ in ξ corresponds to frequency. We are especially interested in operators that have distinct characteristics in space and frequency. The prototypical class exhibiting such a distinction is product-convolution operators, which are operators of the form $P_g C_f$, where C_f denotes convolution with f and P_g denotes multiplication by g [1]. The Kohn–Nirenberg symbol of $P_g C_f$ is easily seen to be

$$\sigma(x, \xi) = g(x)\widehat{f}(\xi). \tag{7}$$

Thus the Kohn–Nirenberg symbol of a product-convolution operator is a tensor product, $\sigma = g \otimes \widehat{f}$. In general the nature of g and $\widehat{f}$ may be completely different from one another, in which case the corresponding operator has completely different spatial and frequency characteristics. The fact that every linear partial differential operator is a finite sum of product-convolution operators, as can be seen from the general form of the symbol

$$\sigma(x, \xi) = \sum_{\alpha} g_\alpha(x)\xi^\alpha,$$

underlies the importance of the latter class. Moreover it is in the nature of partial differential operators for the symbol to be polynomial in frequency, while the spatial structure, corresponding to the functions g_α, typically represents physical parameters that are of a very different nature—possibly not even smooth.

Based on these considerations we tailor our analysis in this chapter to operators that have distinct characteristics in space and frequency, and we focus on representations of such operators that preserve this distinction. More precisely, we study the representation of operators as Gabor multipliers, described in Sect. 2. A Gabor multiplier has an associated Gabor symbol, analogous to the Kohn–Nirenberg symbol. In Sect. 4 we introduce the notion of compatible window pairs, which lead to Gabor multiplier representations that respect the distinction between space and frequency in the following sense: up to a phase factor, the resulting Gabor symbol is a tensor product if and only if the Kohn–Nirenberg symbol is.

2 The Relation Between Gabor and Kohn–Nirenberg Symbols

2.1 Gabor Multipliers

If $g \in \mathscr{S}'_n$ is a tempered distribution, then the formula

$$V_g\varphi(x,\xi) = \overline{\langle M_\xi T_x g, \varphi\rangle}$$

defines a *short-time Fourier transform* $V_g : \mathscr{S}_n \to \mathscr{P}_{2n}$ with *analysis window* g. The basic theory of short-time Fourier transforms is described in [7, § 3]. For $\gamma \in \mathscr{S}_n$, the range of V_γ lies in $\mathscr{S}_{2n}$; the adjoint V_γ^* of V_γ is the map

$$V_\gamma^* : \mathscr{S}'_{2n} \to \mathscr{S}'_n; \qquad \langle V_\gamma^* u, \varphi\rangle = \langle u, V_\gamma\varphi\rangle.$$

If the distribution $u \in \mathscr{S}'_{2n}$ happens to be an L^2 function, then $V_\gamma^* u$ is given by the formula

$$V_\gamma^* u(t) = \iint u(x,\xi) M_\xi T_x \gamma(t)\,\mathrm{d}x\mathrm{d}\xi.$$

For $r > 0$, let φ_r denote the scaled Gaussian on $\mathbb{R}^n$, $\varphi_r(x) = r^n \mathrm{e}^{-\pi r^2 x\cdot x}$. Given a pair of distributions $(g,\gamma) \in \mathscr{S}'_n \times \mathscr{S}'_n$, we write $\langle g,\gamma\rangle$ for the limit

$$\begin{aligned}\langle g,\gamma\rangle &= \lim_{r\to\infty} \int \overline{V_g\varphi_r}(x,0) V_\gamma\varphi_r(x,0)\,\mathrm{d}x \\ &= \lim_{r\to\infty} \int g * \varphi_r(x)\overline{\gamma * \varphi_r}(x)\,\mathrm{d}x \end{aligned} \tag{8}$$

whenever the latter exists and is finite. Here the symbol $*$ denotes convolution.

Definition 2. A *window pair* on $\mathbb{R}^n$ is a pair of distributions $(g,\gamma) \in \mathscr{S}'_n \times \mathscr{S}'_n$ such that (i) for every pair of Schwartz class functions $(\varphi,\theta) \in \mathscr{S}_n \times \mathscr{S}_n$, $\overline{V_g\varphi} V_\gamma\theta \in \mathscr{S}_{2n}$, and (ii) in the sense of (8), $\langle g,\gamma\rangle$ is well defined and non-zero. We write WP_n for the set of all window pairs in $\mathscr{S}'_n \times \mathscr{S}'_n$.

It is evident from the definition that $(g,\gamma) \in \mathrm{WP}_n$ if and only if $(\gamma, g) \in \mathrm{WP}_n$. Examples of window pairs include:

- Any pair (g,γ) of Gaussians $g(x) = \mathrm{e}^{-mx\cdot x}, \gamma(x) = \mathrm{e}^{-\mu x\cdot x}$, where m,μ are positive scalars
- Any non-orthogonal pair of Schwartz class functions.
- (g,γ) when g is bounded by a polynomial and locally integrable, $\gamma \in \mathscr{S}_n$ and $\langle g,\gamma\rangle \neq 0$
- The pair $(1,\delta)$ (i.e., the constant function 1 and Dirac's delta function)

It is a basic fact about the short-time Fourier transform that for any window pair (g, γ), the map

$$\frac{1}{\langle g, \gamma \rangle} V_\gamma^* V_g : \mathscr{S}_n \to \mathscr{S}_n \tag{9}$$

is the identity. Recall from Table 1 that we use the symbol P_λ to denote multiplication by λ. Inserting P_λ between the forward and inverse transforms of (9) leads to the following.

Definition 3. Given a window pair (g, γ) on $\mathbb{R}^n$ and a distribution $\lambda \in \mathscr{S}'_{2n}$, we call

$$\mathscr{M}_\lambda^{g,\gamma} = \frac{1}{\langle g, \gamma \rangle} V_\gamma^* P_\lambda V_g$$

a *Gabor multiplier*; we refer to the distribution λ as its *Gabor symbol*. g and γ are the *analysis* window and *synthesis* window, respectively.

Definition 2 ensures that $\mathscr{M}_\lambda^{g,\gamma}$ is well defined. More precisely, $\mathscr{M}_\lambda^{g,\gamma}$ determines a well-defined sesquilinear functional $(\varphi, \theta) \mapsto \langle \mathscr{M}_\lambda^{g,\gamma} \varphi, \theta \rangle$, with the prescribed interpretation

$$\langle \mathscr{M}_\lambda^{g,\gamma} \varphi, \theta \rangle = \frac{1}{\langle g, \gamma \rangle} \langle \lambda, \overline{V_g \varphi} V_\gamma \theta \rangle.$$

(Note that Feichtinger and Nowak [3] use the term "short-time Fourier transform multiplier" for a Gabor multiplier based on identical windows (g, g), while in [3] "Gabor multiplier" refers to a more general object than we have defined.) A Gabor multiplier, in the sense of Definition 3, is a linear operator belonging to $\mathscr{L}_n$. Its adjoint is also a Gabor multiplier, and the precise connection between the two works out as follows.

Proposition 4. *For any window pair (g, γ) and any distribution $\lambda \in \mathscr{S}'_{2n}$, the adjoint of the Gabor multiplier $\mathscr{M}_\lambda^{g,\gamma}$ is $\left(\mathscr{M}_\lambda^{g,\gamma}\right)^* = \mathscr{M}_{\bar{\lambda}}^{\gamma,g}$.*

The structure of a Gabor multiplier is such that it carries an implicit diagonalization on phase space. From the theoretical point of view, this fact makes it desirable to express a given linear operator $L \in \mathscr{L}_n$ as a Gabor multiplier, the structure of the operator then being encoded in its Gabor symbol. Note that the Kohn–Nirenberg form is itself a Gabor multiplier, as the following easily verified formula attests:

$$\forall \sigma \in \mathscr{S}'_{2n}, \quad \mathscr{M}_\sigma^{1,\delta} = \sigma(X, D). \tag{10}$$

This is just one extreme of a whole range of such representations, each of which has its own characteristics. Since there exist fast computational methods to evaluate discretized Gabor multipliers [9] it is also desirable to express an operator as a Gabor multiplier from the point of view of applications. However, speed of computation is contingent on localization of the windows, so from this point of view the Kohn–Nirenberg form (10) is not particularly advantageous, as the analysis window 1 has no localization whatsoever.

2.2 The Schwartz Kernel of a Gabor Multiplier

One problem that we are concerned with in the present chapter is to express a given Kohn–Nirenberg pseudodifferential operator as a Gabor multiplier based on prescribed windows. Before considering the issue in detail, we deal briefly with the converse problem of expressing a given Gabor multiplier as a pseudodifferential operator. In light of the expression (6) for the Schwartz kernel of a pseudodifferential operator, the latter problem is equivalent to computing the Schwartz kernel of a Gabor multiplier. This turns out to be relatively straightforward and can be carried out in full generality. In stating the basic result we make use of the following notation. Let $E : \mathscr{S}_{4n} \to \mathscr{S}_{2n}$ denote the map defined by

$$E\rho(x,\xi) = \rho(x,x,\xi,-\xi). \tag{11}$$

The corresponding adjoint is the map

$$E^* : \mathscr{S}'_{2n} \to \mathscr{S}'_{4n}; \qquad \langle E^*u, \varphi\rangle = \langle u, E\varphi\rangle.$$

Theorem 1. *An arbitrary Gabor multiplier $\mathscr{M}_\lambda^{g,\gamma}$ has Schwartz kernel*

$$K(\mathscr{M}_\lambda^{g,\gamma}) = \frac{1}{\langle g,\gamma\rangle} V^*_{\gamma\otimes\bar{g}} E^*\lambda. \tag{12}$$

Proof. By Definition 3,

$$\langle \mathscr{M}_\lambda^{g,\gamma}\varphi, \theta\rangle = \frac{1}{\langle g,\gamma\rangle}\langle \lambda, \overline{V_g\varphi} V_\gamma\theta\rangle,$$

and we have

$$\begin{aligned}
\overline{V_g\varphi}V_\gamma\theta(x,\xi) &= \overline{\int \varphi(t)\mathrm{e}^{-2\pi \mathrm{i} t\cdot\xi}\bar{g}(t-x)\,\mathrm{d}t \int \theta(\tau)\mathrm{e}^{-2\pi\mathrm{i}\tau\cdot\xi}\bar{\gamma}(\tau-x)\,\mathrm{d}\tau} \\
&= \iint \bar{\varphi}(t)\mathrm{e}^{2\pi\mathrm{i}t\cdot\xi} g(t-x)\theta(\tau)\mathrm{e}^{-2\pi\mathrm{i}\tau\cdot\xi}\bar{\gamma}(\tau-x)\,\mathrm{d}\tau\mathrm{d}t \\
&= \iint \mathrm{e}^{2\pi\mathrm{i}(t-\tau)\cdot\xi}\bar{\gamma}(\tau-x)g(t-x)\theta\otimes\bar{\varphi}(\tau,t)\,\mathrm{d}\tau\mathrm{d}t \qquad (13) \\
&= \iint \mathrm{e}^{-2\pi\mathrm{i}(\tau,t)\cdot(\xi,-\xi)} T_{(x,x)}\bar{\gamma}\otimes g(\tau,t)\theta\otimes\bar{\varphi}(\tau,t)\,\mathrm{d}\tau\mathrm{d}t \\
&= V_{\gamma\otimes\bar{g}}\theta\otimes\bar{\varphi}(x,x,\xi,-\xi).
\end{aligned}$$

Thus,

$$\begin{aligned}\langle \mathscr{M}_{\lambda}^{g,\gamma}\varphi, \theta\rangle &= \frac{1}{\langle g,\gamma\rangle}\langle \lambda, EV_{\gamma\otimes\bar{g}}\theta\otimes\bar{\varphi}\rangle \\ &= \frac{1}{\langle g,\gamma\rangle}\langle V^*_{\gamma\otimes\bar{g}}E^*\lambda, \theta\otimes\bar{\varphi}\rangle.\end{aligned}$$

□

2.3 The Key Equation Relating Symbols

Since generalized Kohn–Nirenberg operators encompass all of $\mathscr{L}_n$ (by (6)), given an arbitrary Gabor multiplier $\mathscr{M}_{\lambda}^{g,\gamma}$ on $\mathbb{R}^n$, there exists a distribution $\sigma \in \mathscr{S}_{2n}$ such that $\sigma(X, D) = \mathscr{M}_{\lambda}^{g,\gamma}$. By Theorem 1, this is equivalent, in terms of Schwartz kernels, to the equation

$$\mathscr{T}_{-}\mathscr{F}_2\sigma = \frac{1}{\langle g,\gamma\rangle}V^*_{\gamma\otimes\bar{g}}E^*\lambda, \tag{14}$$

which can be solved for σ in terms of λ to yield

$$\sigma = \frac{1}{\langle g,\gamma\rangle}\mathscr{F}_2^{-1}\mathscr{T}_{-}^{-1}V^*_{\gamma\otimes\bar{g}}E^*\lambda. \tag{15}$$

On the other hand, it is not possible in general to solve Eq. (14) for λ in terms of σ. In order to characterize precisely when a solution does exist, it is simpler to compare the symplectic Fourier transforms of σ and λ rather than the distributions themselves. Recalling the operators $\mathscr{T}_{-}$, $\mathscr{T}_s$, and E defined earlier on lines (3), (4), and (11) and letting $\mathscr{T}_{+}$ denote the inverse of $\mathscr{T}_{-}$, we begin with a preliminary calculation.

Lemma 5. *As operators on* $\mathscr{S}'_{2n}$,

$$\mathscr{T}_s\mathscr{F}_1\mathscr{T}_{+}V^*_{\gamma\otimes\bar{g}}E^* = V_g\gamma\mathscr{F}_s.$$

Proof. It suffices to compare the action of the respective adjoint operators

$$EV_{\gamma\otimes\bar{g}}\mathscr{T}_{-}\mathscr{F}_1^{-1}\mathscr{T}_s^* \text{ and } \mathscr{F}_s P_{\overline{V_g\gamma}},$$

where $P_{\overline{V_g\gamma}}$ denotes multiplication by $\overline{V_g\gamma}$, on the space $\mathscr{S}_{2n}$ of test functions. Thus we verify that for every test function $\varphi \in \mathscr{S}_{2n}$,

$$EV_{\gamma\otimes\bar{g}}\mathscr{T}_{-}\mathscr{F}_1^{-1}\mathscr{T}_s^*\varphi = \mathscr{F}_s\left(\overline{V_g\gamma}\varphi\right).$$

This is a matter of direct calculation:

$$
\begin{aligned}
E V_{\gamma\otimes\bar{g}} \mathscr{T}_- \mathscr{F}_1^{-1} \mathscr{T}_s^* \varphi(x,\xi) &= V_{\gamma\otimes\bar{g}} \mathscr{T}_- \mathscr{F}_1^{-1} \mathscr{T}_s^* \varphi(x,x,\xi,-\xi) \qquad (16)\\
&= \iiint \mathrm{e}^{-2\pi \mathrm{i}(y\cdot\xi - t\cdot\xi)} \mathrm{e}^{2\pi i y\cdot\eta} \mathscr{T}_s^* \varphi(\eta, t-y)\\
&\quad \bar{\gamma}(y-x) g(t-x) \mathrm{d}\eta \, \mathrm{d}y \, \mathrm{d}t\\
&= \iiint \mathrm{e}^{2\pi \mathrm{i}(t-y)\cdot\xi} \mathrm{e}^{2\pi i y\cdot\eta} \mathscr{T}_s^* \varphi(\eta, t-y) \bar{\gamma}(y-x)\\
&\quad g(t-x) \mathrm{d}\eta \, \mathrm{d}y \, \mathrm{d}t. \qquad (17)
\end{aligned}
$$

Applying the change of variables $(y,t) \mapsto (\tau,u) = (y-x, t-y)$, the integral (17) transforms to

$$
\begin{aligned}
&\iiint \mathrm{e}^{2\pi i u\cdot\xi} \mathrm{e}^{2\pi i(\tau+x)\cdot\eta} \mathscr{T}_s^* \varphi(\eta,u) \bar{\gamma}(\tau) g(\tau+u) \mathrm{d}\eta \, \mathrm{d}\tau \, du\\
&= \iiint \mathrm{e}^{2\pi \mathrm{i}(x\cdot\eta + \xi\cdot u)} \mathrm{e}^{2\pi i \tau\cdot\eta} \bar{\gamma}(\tau) g(\tau+u) \varphi(-u,\eta) \mathrm{d}\tau \, \mathrm{d}\eta \, du\\
&= \iiint \mathrm{e}^{2\pi \mathrm{i}(x\cdot\eta - \xi\cdot u)} \mathrm{e}^{2\pi i \tau\cdot\eta} \bar{\gamma}(\tau) g(\tau-u) \varphi(u,\eta) \mathrm{d}\tau \, \mathrm{d}\eta \, du\\
&= \mathscr{F}_s \left(\overline{V_g \gamma} \varphi \right)(x,\xi). \qquad (18)
\end{aligned}
$$

Comparing (16) and (18) yields the desired result. □

The relation between the symplectic Fourier transforms of σ and λ is as follows.

Theorem 2. *The equation $\sigma(X,D) = \mathscr{M}_\lambda^{g,\gamma}$ holds if and only if $\widehat{\sigma}^s = \frac{1}{\langle g,\gamma\rangle} V_g\gamma \, \widehat{\lambda}^s$.*

Proof. Suppose $\sigma(X,D) = \mathscr{M}_\lambda^{g,\gamma}$. Then

$$
\begin{aligned}
\widehat{\sigma}^s &= \mathscr{T}_s \mathscr{F}_1 \mathscr{T}_+ \mathscr{T}_- \mathscr{F}_2 \sigma\\
&= \mathscr{T}_s \mathscr{F}_1 \mathscr{T}_+ \left(\frac{1}{\langle g,\gamma\rangle} V_{\gamma\otimes\bar{g}}^* E^* \lambda \right) \text{ (by (14))}\\
&= \frac{1}{\langle g,\gamma\rangle} \mathscr{T}_s \mathscr{F}_1 \mathscr{T}_+ V_{\gamma\otimes\bar{g}}^* E^* \lambda\\
&= \frac{1}{\langle g,\gamma\rangle} V_g\gamma \, \widehat{\lambda}^s \text{ (by Lemma 5).}
\end{aligned}
$$

Conversely, if $\widehat{\sigma}^s = \frac{1}{\langle g,\gamma\rangle} V_g\gamma \, \widehat{\lambda}^s$, then, again by Lemma 5, Eq. (14) holds, which says that $\sigma(X,D)$ and $\mathscr{M}_\lambda^{g,\gamma}$ have the same Schwartz kernel and hence are equal. □

Theorem 2 is not a new result. The special case where $g = \gamma$ are identical, normalized Gaussians appears in [5, p. 141] and more recently as Theorem 17.1 in [15]. In the one-dimensional setting, with the additional assumption that the operator in question be traceclass, the special case corresponding to identical, but not necessarily Gaussian, windows appears as Equation (10.3.19) in [8]. In any case our main interest in this result is as a means to characterize the class of operators that may be represented as Gabor multipliers based on a fixed window pair.

3 The Class of Operators Based on a Fixed Window Pair

Although every Gabor multiplier has a corresponding Kohn–Nirenberg symbol, Theorem 2 implies that for a fixed window pair (g, γ) there need not exist a Gabor symbol corresponding to every Kohn–Nirenberg operator. The general situation is that given a window pair (g, γ), there is a corresponding class of operators that may be expressed as Gabor multipliers based on (g, γ):

$$\mathrm{Op}(g, \gamma) = \left\{ \mathscr{M}_{\lambda}^{g,\gamma} \mid \lambda \in \mathscr{S}'_{2n} \right\}.$$

Theorem 2 facilitates a simple characterization of $\mathrm{Op}(g, \gamma)$, illustrated in the present section. Our ultimate objective, which motivates Sect. 4, is to "tune" $\mathrm{Op}(g, \gamma)$ to include operators of a desired type by judicious choice of g and γ.

3.1 Characterization of the Spreading Function

The basic relation between Gabor and K-N symbols provides a representation of the set $\mathrm{Op}(g, \gamma)$ of all Gabor multipliers based on (g, γ) in terms of the corresponding set of spreading functions. We use the following notation: given $\varphi \in \mathscr{P}_{2n}$ we write $\varphi \mathscr{S}'_{2n}$ for the set of distributions

$$\varphi \mathscr{S}'_{2n} = \left\{ \varphi\rho \;\middle|\; \rho \in \mathscr{S}'_{2n} \right\}.$$

Theorem 3. *Given a window pair (g, γ) on $\mathbb{R}^n$,*

$$Op(g, \gamma) = \left\{ \sigma(X, D) \mid \widehat{\sigma}^s \in V_g\gamma \mathscr{S}'_{2n} \right\}.$$

Proof. Given $\sigma \in \mathscr{S}'_{2n}$, there exists a distribution $\lambda \in \mathscr{S}'_{2n}$ such that $\widehat{\sigma}^s = \frac{1}{\langle g,\gamma \rangle} V_g\gamma \widehat{\lambda}^s$ if and only if $\widehat{\sigma}^s \in \frac{1}{\langle g,\gamma \rangle} V_g\gamma \mathscr{S}'_{2n} = V_g\gamma \mathscr{S}'_{2n}$. □

The problem of describing $\mathrm{Op}(g, \gamma)$ is thus tied to the description of the set of distributions $V_g\gamma \mathscr{S}'_{2n}$. Let us now specialize to the case of Gaussian windows.

3.2 Gabor Multipliers Based on Gaussian Windows

We study the dependence of $\mathrm{Op}(g,\gamma)$ on (g,γ) in the case where g and γ are Gaussians, from the following perspective. If L is a given product-convolution operator, say, to what extent can we choose g and γ to ensure that $L \in \mathrm{Op}(g,\gamma)$? The starting point in this regard is a calculation based on Theorem 3.

Proposition 6. *Let $g(x) = \mathrm{e}^{-mx\cdot x}$ and $\gamma(x) = \mathrm{e}^{-\mu x\cdot x}$ be Gaussians on $\mathbb{R}^n$, where $m,\mu > 0$. Then*

$$Op(g,\gamma) = \{\sigma(X,D)\,|\,\widehat{\sigma}^s \in (h\otimes k)\,\mathscr{S}'_{2n}\},$$

where $h(u) = \mathrm{e}^{-\frac{m\mu}{m+\mu}u\cdot u}$ and $k(\eta) = \mathrm{e}^{-\frac{\pi^2}{m+\mu}\eta\cdot\eta}$.

Proof. By a straightforward calculation,

$$V_g\gamma(u,\eta) = \sqrt{\frac{\pi}{m+\mu}}^{\,n}\,\mathrm{e}^{-2\pi\mathrm{i}\frac{m}{m+\mu}u\cdot\eta}\mathrm{e}^{-\frac{m\mu}{m+\mu}u\cdot u}\mathrm{e}^{-\frac{\pi^2}{m+\mu}\eta\cdot\eta}.$$

Since $\sqrt{\frac{\pi}{m+\mu}}^{\,n}\,\mathrm{e}^{-2\pi\mathrm{i}\frac{m}{m+\mu}u\cdot\eta}\mathscr{S}'_{2n} = \mathscr{S}'_{2n}$, it follows that $V_g\gamma\,\mathscr{S}'_{2n} = (h\otimes k)\,\mathscr{S}'_{2n}$ for the given h and k. The proposition then follows from Theorem 3. □

Roughly speaking the broad implication of Proposition 6 is that if $L \in \mathrm{Op}(g,\gamma)$, then the spreading function of L must be rapidly decaying, the rate of decay being dictated by the Gaussian $h\otimes k$. For example, every *underspread operator*, defined as an operator having compactly supported spreading function (hence decaying more rapidly than $h\otimes k$), belongs to $\mathrm{Op}(g,\gamma)$. To analyse the situation in more detail, let us consider product-convolution operators L of a special type, namely those having a Gaussian Kohn–Nirenberg symbol of the form

$$\sigma_L(x,\xi) = \mathrm{e}^{-ax\cdot x}\mathrm{e}^{-b\xi\cdot\xi}, \tag{19}$$

where $a,b > 0$. The spreading function for such an operator is

$$\widehat{\sigma}^s_L(u,\eta) = \left(\frac{\pi}{\sqrt{ab}}\right)^n \mathrm{e}^{-\frac{\pi^2}{b}u\cdot u}\mathrm{e}^{-\frac{\pi^2}{a}\eta\cdot\eta}.$$

For h,k as in Proposition 6,

$$\begin{aligned}
&\widehat{\sigma}^s_L \in (h\otimes k)\mathscr{S}'_{2n}\\
&\quad\Longleftrightarrow\ (h\otimes k)^{-1}\widehat{\sigma}^s_L \in \mathscr{S}'_{2n}\\
&\quad\Longleftrightarrow\ \frac{\pi^2}{b} - \frac{m\mu}{m+\mu} \geq 0 \text{ and } \frac{\pi^2}{a} - \frac{\pi^2}{m+\mu} \geq 0.
\end{aligned} \tag{20}$$

The quadratic inequalities (20) easily yield the following result.

Proposition 7. *Let g, γ be as in Proposition 6, and let $L \in \mathscr{L}_n$ have symbol (19). In the case where $g = \gamma$ (i.e. $m = \mu$), $L \in Op(g, \gamma)$ if and only if $ab \leq 4\pi^2$ and $a/2 \leq m \leq 2\pi^2/b$. In the general case, for any a, b, the inequalities (20) determine an unbounded region in the (m, μ) plane for which $L \in Op(g, \gamma)$.*

The foregoing proposition illustrates several facts. Firstly, it is restrictive to use identical analysis and synthesis windows: an operator L of the given type need not be representable as a Gabor multiplier based on identical Gaussian windows. Secondly, from the point of view of the operator L, the smaller the value of the product ab (corresponding to slower decay of the symbol σ_L), the larger the class of window pairs (g, γ) for which $L \in \mathrm{Op}(g, \gamma)$. Finally, even if ab is large, L can be represented as a Gabor multiplier based on windows (g, γ) provided that g, γ are chosen to have sufficiently different widths (i.e. with $m\mu$ sufficiently small and $m + \mu$ sufficiently large).

In dealing with product-convolution operators it is very convenient to be able to analyze the separate behaviour of the functions h and k arising in Proposition 6, facilitating results such as Proposition 7. The fact that $\mathrm{Op}(g, \gamma)$ can be described in terms of a tensor product $h \otimes k$ is a special feature of Gaussian windows, but it is not unique to them. In Sect. 4 we introduce a broader class of window pairs that share this same advantage.

3.3 Multipliers Based on Compactly Supported Windows

We have mentioned already that rapidly decreasing windows are desirable for numerical computation. In this respect one cannot do better than to use windows that have compact support, which of course decay even faster than Gaussian windows. In this section we examine briefly the representation of a very simple class, namely translation operators, as Gabor multipliers based on compactly supported windows. Translation operators are underspread, so the results above in Sect. 3.2 show that using arbitrary Gaussian windows, any translation operator can be represented as a Gabor multiplier. This is not the case for compactly supported windows, as we now show.

The Kohn–Nirenberg symbol of translation by $\tau \in \mathbb{R}^n$, denoted T_τ, is

$$\sigma(x, \xi) = \mathrm{e}^{-2\pi \mathrm{i} \tau \cdot \xi},$$

which has symplectic Fourier transform

$$\widehat{\sigma}^s(y, \eta) = \delta(y + \tau)\delta(\eta).$$

Thus the support of $\widehat{\sigma}^s$ is the singleton $\{(-\tau, 0)\}$.

Now, suppose that we wish to express the operator $T_\tau = \sigma(X, D)$ as a Gabor multiplier based on compactly supported windows g, γ. Theorem 2 implies that

if this can be achieved using Gabor symbol λ, then $\operatorname{supp}\widehat{\sigma^s} = \operatorname{supp}(V_g\gamma\widehat{\lambda^s})$. In particular, since

$$\operatorname{supp}(V_g\gamma\widehat{\lambda^s}) \subseteq \operatorname{supp} V_g\gamma \cap \operatorname{supp}\widehat{\lambda^s},$$

it is necessary that $\operatorname{supp}\widehat{\sigma^s} \subseteq \operatorname{supp} V_g\gamma$. But the support of $V_g\gamma$ is constrained by that of g and γ, as follows.

Lemma 8. *Let (g, γ) be a window pair on $\mathbb{R}^n$ where at least one of supp g and supp γ is bounded. Then the support of $V_g\gamma$ is contained in*

$$(\mathit{supp}\,\gamma - \mathit{supp}\,g) \times \mathbb{R}^n,$$

where the minus sign denotes Minkowski difference:

$$\mathit{supp}\,\gamma - \mathit{supp}\,g = \{x - y \mid x \in \mathit{supp}\,\gamma \text{ and } y \in \mathit{supp}\,g\}.$$

Proof. Note that boundedness of supp g or supp γ implies supp $\gamma -$ supp g is a closed set. The integrand of the expression defining $V_g\gamma$, $\mathrm{e}^{-2\pi\mathrm{i}\tau\cdot\eta}\gamma(\tau)\bar{g}(\tau - u)$, is different from zero only if $\tau \in \operatorname{supp}\gamma$ and $\tau - u \in \operatorname{supp} g$, or, equivalently, if $\tau \in \operatorname{supp}\gamma$ and $u \in \tau - \operatorname{supp} g$, which implies that

$$u \in \operatorname{supp}\gamma - \operatorname{supp} g. \tag{21}$$

So if (21) fails to hold, then $V_g\gamma(u, \eta) = 0$; since $\operatorname{supp}\gamma - \operatorname{supp} g$ is closed, this implies that $(u, \eta) \notin \operatorname{supp} V_g\gamma$. □

Thus the possibility of realizing a translation operator as a Gabor multiplier is contingent upon the windows used, as follows.

Proposition 9. *Given windows (g, γ) where at least one of supp g and supp γ is bounded, the equality $T_\tau = \mathscr{M}_\lambda^{g,\gamma}$ is possible only if $\tau \in \mathit{supp}\,g - \mathit{supp}\,\gamma$.*

Proof. If $T_\tau = \sigma(X, D)$ can be expressed as a Gabor multiplier $\mathscr{M}_\lambda^{g,\gamma}$, then, by Theorem 2,

$$\operatorname{supp} V_g\gamma \supseteq \operatorname{supp}\widehat{\sigma^s} = \{(-\tau, 0)\}.$$

Lemma 8 therefore requires that

$$(-\tau, 0) \in (\operatorname{supp}\gamma - \operatorname{supp} g) \times \mathbb{R}^n$$
$$\iff \tau \in \operatorname{supp} g - \operatorname{supp}\gamma.$$

□

Concerning the converse question of when a translation *can* be expressed as a Gabor multiplier, we give a simple example. Let $h \in C_0^\infty(\mathbb{R}^n)$ be a function that is strictly positive on the open unit cube, $(0, 1)^n$, and zero elsewhere, and consider windows $g = \gamma = h$. In this case, $\operatorname{supp} g - \operatorname{supp}\gamma = [-1, 1]^n$, and one can easily

verify that $V_g\gamma(-\tau,0) > 0$ if and only if $\tau \in (-1,1)^n$. For $\tau \in (-1,1)^n$, the choice of Gabor symbol

$$\lambda(x,\xi) = \frac{\langle g,\gamma\rangle}{V_g\gamma(-\tau,0)}\mathrm{e}^{-2\pi\mathrm{i}\tau\cdot\xi}$$

results in $\mathscr{M}_\lambda^{g,\gamma} = T_\tau$. Thus almost every translation not excluded by Proposition 9 *is* realizable as a Gabor multiplier with the given windows, the only exceptions being boundary points of the cube $[-1,1]^n$.

4 A General Class of Gaussian-like Window Pairs

In this section we focus on a particular class of window pairs for which $V_g\gamma$ has an explicit form that generalizes the Gaussian case and which facilitates analysis of $\mathrm{Op}(g,\gamma)$ by separating space and frequency components.

4.1 Definition of Compatible Window Pairs

According to Proposition 6, $V_g\gamma\mathscr{S}'_{2n}$ has the form $(h\otimes k)\mathscr{S}'_{2n}$ when g and γ are Gaussian. The tensor product form $h\otimes k$ is a great convenience from the point of view of trying to ensure that a given product-convolution operator L belongs to $\mathrm{Op}(g,\gamma)$, since the spreading function of any product-convolution operator is itself a tensor product. On the other hand, Proposition 6 also shows that for Gaussian windows g and γ, the class $\mathrm{Op}(g,\gamma)$ is limited to operators whose spreading functions decay very rapidly—what one might call near-underspread operators. Do there exist rapidly decaying windows for which $V_g\gamma\mathscr{S}'_{2n}$ has the form $(h\otimes k)\mathscr{S}'_{2n}$, as for Gaussians, but for which $\mathrm{Op}(g,\gamma)$ is less restrictive? We show in this section that the answer is yes. To do so, we study window pairs having the following property.

Definition 10. A window pair is said to be *compatible*, or, for emphasis, strongly compatible, if $V_g\gamma$ can be expressed in the form

$$V_g\gamma(u,\eta) = \mathrm{e}^{2\pi\mathrm{i}\eta\cdot\zeta(u)}h(u)k(\eta),$$

for some smooth function $\zeta:\mathbb{R}^n\to\mathbb{R}^n$.

The class of compatible window pairs includes all Gaussian pairs, but it includes many other functions as well—see Sect. 4.2. The following generalization of compatibility serves a purely technical role in the present discussion.

Definition 11. A window pair (g,γ) is *weakly compatible* if the spectrogram $|V_g\gamma|^2$ splits as a tensor product:

$$|V_g\gamma(x,\xi)|^2 = H(x)K(\xi),$$

for some $H, K : \mathbb{R}^n \to \mathbb{R}$. Equivalently, (g, γ) is weakly compatible if $V_g\gamma$ splits as a tensor product up to a phase factor:

$$V_g\gamma(u, \eta) = \mathrm{e}^{2\pi \mathrm{i}\rho(u,\eta)} h(u)k(\eta), \tag{22}$$

for some $h, k : \mathbb{R}^n \to \mathbb{C}$ and $\rho : \mathbb{R}^{2n} \to \mathbb{R}$.

For a weakly compatible window pair (g, γ) the function $V_g\gamma$ has a special form, given in the next proposition. Here the notation $\widetilde{f}$ is used to denote reflection in the argument: $\widetilde{f}(x) = f(-x)$. (We omit the proposition's proof, which is straightforward.)

Proposition 12. *A window pair (g, γ) is weakly compatible if and only if there exists a phase function $\rho : \mathbb{R}^{2n} \to \mathbb{R}$ satisfying $\rho(u, 0) = \rho(0, \eta) = 0$ such that*

$$\frac{1}{\langle g, \gamma\rangle} V_g\gamma(u, \eta) = \frac{\mathrm{e}^{2\pi \mathrm{i}\rho(u,\eta)}}{|\langle g, \gamma\rangle|^2} \left(\gamma * \widetilde{\overline{g}}\right)(u) \left(\widehat{\gamma} * \widehat{\overline{g}}\right)(\eta). \tag{23}$$

The main thrust of the above proposition is that if (g, γ) is weakly compatible, then the functions h, k in Definition 11 may be assumed to have the form $h = \gamma * \widetilde{\overline{g}}$ and $k = \widehat{\gamma} * \widehat{\overline{g}}$, up to multiplication by a scalar. We return now to the implications of the original Definition 10.

Proposition 13. *A window pair (g, γ) is (strongly) compatible if and only if there exists a smooth function $\psi : \mathbb{R}^n \to \mathbb{R}^n$ such that for every $u \in \mathbb{R}^n$,*

$$\gamma\, T_u\overline{g} = \frac{\gamma * \widetilde{\overline{g}}(u)}{\langle \gamma, g\rangle} T_{\psi(u)}(\gamma\overline{g}). \tag{24}$$

Proof. If the pair (g, γ) is compatible then Eq. (23) holds, where ρ has the specific form $\rho(u, \eta) = \eta \cdot \zeta(u)$ for some smooth $\zeta : \mathbb{R}^n \to \mathbb{R}^n$. Taking the inverse Fourier transform with respect to η of this equation yields (24), with $\psi = -\zeta$. □

The formulation (24) can be viewed as product-translation invariance of $\overline{g}, \gamma$: the product of γ with a translate of $\overline{g}$ is itself a rescaled translation of the product of γ with $\overline{g}$. The advantage of this formulation is that it can be used to obtain an explicit classification of compatible windows.

4.2 Classification of Compatible Windows

It turns out that there is one key example of a compatible window pair that is essentially distinct from Gaussians. The class of all Schwartz class compatible windows can be generated from this key example, together with Gaussians, by means of a simple family of transformations and constructions, which we now describe. The following two propositions may be easily verified directly from Definition 10.

Proposition 14. *Let A be a non-singular linear transformation on $\mathbb{R}^n$; let ω_0, ω_1, b_0, b_1 be vectors in $\mathbb{R}^n$; and let $\alpha_0, \alpha_1 \in \mathbb{C}$ be non-zero scalars. If (g, γ) is a compatible window pair on $\mathbb{R}^n$, then so is the pair (g', γ'), where*

$$g'(t) = \alpha_0 \mathrm{e}^{2\pi \mathrm{i}\omega_0 \cdot t} g(At + b_0) \quad \text{and} \quad \gamma'(t) = \alpha_1 \mathrm{e}^{2\pi \mathrm{i}\omega_1 \cdot t} \gamma(At + b_1).$$

In other words, compatibility is invariant under rescaling, translation, modulation, and linear changes of variables, provided the same linear transformation is applied to both windows.

Proposition 15. *If (g_0, γ_0) and (g_1, γ_1) are compatible window pairs on $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, then the window pair $(g_0 \otimes g_1, \gamma_0 \otimes \gamma_1)$ is compatible on $\mathbb{R}^{m+n}$.*

A Gaussian on $\mathbb{R}^n$ is a tensor product of Gaussians on $\mathbb{R}^1$, so in light of Propositions 14 and 15, all pairs of Gaussian windows can be generated starting with a pair of Gaussians on $\mathbb{R}^1$.

As mentioned earlier, the formulation (24) facilitates a classification of compatible windows. More precisely, by taking repeated derivatives of the logarithm of (24), one can set up a differential equation that must be satisfied by each component of a compatible window pair and thereby explicitly compute all possibilities. The details of this analysis are rather involved and will be presented elsewhere in [6]. The key example that arises as a solution to the aforementioned differential equation is the following.

Proposition 16. *Let m, μ and α be positive scalars, and set $g(t) = \mathrm{e}^{mt - \mathrm{e}^{\alpha t}}$ and $\gamma(t) = \mathrm{e}^{\mu t - \mathrm{e}^{\alpha t}}$. Then the window pair (g, γ) is compatible on $\mathbb{R}^1$, and the associated function ζ required by Definition 10 is $\zeta(u) = \frac{1}{\alpha} \log\left(\frac{1+\mathrm{e}^{-\alpha u}}{2}\right)$.*

Like Gaussians, these windows have a fundamental role in probability theory and arise in connection to a variant of the central limit theorem. More precisely, let E denote the function g of Proposition 16 in the special case $m = \alpha = 1$:

$$E(t) = \mathrm{e}^{t - \mathrm{e}^t}.$$

Properly scaled affine transformations of E, of the form

$$\frac{1}{b} E\left(\frac{a - t}{b}\right),$$

are the density functions of so-called "extreme value" probability distributions, described in [4]. Based on this, we refer to the windows of Proposition 16 as *extreme value windows*—see Fig. 1.

The Fourier transform of an extreme value window is proportional to the gamma function, restricted to a line of constant real part:

$$\mathscr{F}\left(\mathrm{e}^{mt - \mathrm{e}^{\alpha t}}\right)(\xi) = \frac{1}{\alpha} \Gamma\left(\frac{m}{\alpha} - \frac{2\pi i}{\alpha} \xi\right).$$

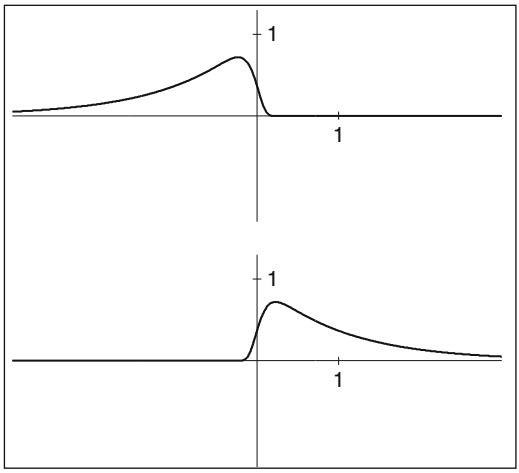


Fig. 1 Extreme value windows. The function $f(t) = \mathrm{e}^{t-\mathrm{e}^{\alpha t}}$, *above*, and its reversal $f(-t)$, *below*, are plotted for $\alpha = 10$. Note that $f(-t)$ is close to being *causal*, i.e. zero for negative values of t. This is because the double exponential $\mathrm{e}^{-\mathrm{e}^{-\alpha t}}$ tends to the Heaviside function with increasing α

Thus the Fourier transform of an extreme value distribution is analytic, never zero, and rapidly decreasing. Of course, since extreme value windows are compatible, they are by design suited to the representation of product-convolution operators—or other operators that have distinct space and frequency structure. The analogue of Proposition 6 for extreme value windows is the following.

Proposition 17. *Let* $g(t) = \mathrm{e}^{mt-\mathrm{e}^{\alpha t}}$, $\gamma(t) = \mathrm{e}^{\mu t-\mathrm{e}^{\alpha t}}$ *be extreme value windows on* $\mathbb{R}^1$, *where* $m, \mu, \alpha > 0$. *Then*

$$Op(g,\gamma) = \{\sigma(X,D) \,|\, \widehat{\sigma}^s \in (h \otimes k)\, \mathscr{S}_2'\},$$

where

$$h(u) = \left(\mathrm{e}^{\frac{m\alpha}{m+\mu}u} + \mathrm{e}^{-\frac{\mu\alpha}{m+\mu}u}\right)^{-\frac{m+\mu}{\alpha}} \qquad \textit{and} \qquad k(\eta) = \Gamma\left(\frac{m+\mu}{\alpha} - \frac{2\pi\mathrm{i}}{\alpha}\eta\right).$$

Proof. By Theorem 3 it suffices to show that $V_g\gamma\mathscr{S}_2' = (h \otimes k)\mathscr{S}_2'$ for the given h, k. By direct calculation,

$$V_g\gamma(u,\eta) = \frac{1}{\alpha}\,(1+\mathrm{e}^{-\alpha u})^{\frac{2\pi\mathrm{i}}{\alpha}\eta}\left(\mathrm{e}^{\frac{m\alpha}{m+\mu}u} + \mathrm{e}^{-\frac{\mu\alpha}{m+\mu}u}\right)^{-\frac{m+\mu}{\alpha}} \Gamma\left(\frac{m+\mu}{\alpha} - \frac{2\pi\mathrm{i}}{\alpha}\eta\right).$$

The proposition then follows from the observation that $\frac{1}{\alpha}\,(1+\mathrm{e}^{-\alpha u})^{\frac{2\pi\mathrm{i}}{\alpha}\eta}\,\mathscr{S}_2' = \mathscr{S}_2'$. □

In Sect. 4.3 we use Proposition 17 to compare extreme value windows with Gaussians in terms of representation of linear operators as Gabor multipliers. As a final remark in the present section, we reiterate that the class of *all* Schwartz class compatible window pairs can be generated by means of Propositions 14 and 15, starting with pairs of Gaussians and pairs of extreme value windows on $\mathbb{R}^1$ (see [6]).

4.3 Extreme Value Versus Gaussian Windows

Extreme value windows are attractive from the point of view of numerical computation since they are rapidly decreasing. From the perspective of representing operators as Gabor multipliers, extreme value windows have a decided advantage over Gaussians in that the former are more general: a strictly wider class of operators can be represented as Gabor multipliers based on extreme value windows than can be represented using Gaussians.

Theorem 4. *Let $\alpha, m, \mu, n, \nu > 0$ be arbitrary positive scalars, and set*

$$f(t) = \mathrm{e}^{-mt^2}, \quad \varphi(t) = \mathrm{e}^{-\mu t^2}, \quad g(t) = \mathrm{e}^{nt - \mathrm{e}^{\alpha t}}, \quad \gamma(t) = \mathrm{e}^{\nu t - \mathrm{e}^{\alpha t}}.$$

Then (g, γ) is a more general window pair on $\mathbb{R}^1$ than (f, φ), in the sense that

$$Op(f, \varphi) \subset Op(g, \gamma).$$

Proof. By Propositions 6 and 17, the conclusion of the theorem is equivalent to the inclusion

$$(h \otimes k)\mathscr{S}_2' \subset (H \otimes K)\mathscr{S}_2', \tag{25}$$

where

$$h(u) = \mathrm{e}^{-\frac{m\mu}{m+\mu}u \cdot u}, \qquad k(\eta) = \mathrm{e}^{-\frac{\pi^2}{m+\mu}\eta \cdot \eta},$$

$$\text{and } H(u) = \left(\mathrm{e}^{\frac{n\alpha}{n+\nu}u} + \mathrm{e}^{-\frac{\nu\alpha}{n+\nu}u}\right)^{-\frac{n+\nu}{\alpha}}, \qquad K(\eta) = \Gamma\left(\frac{n+\nu}{\alpha} - \frac{2\pi \mathrm{i}}{\alpha}\eta\right).$$

Note that $H(u)$ behaves like $(\mathrm{e}^{nu} - \mathrm{e}^{-\nu u})^{-1}$ for large $|u|$, while Stirling's formula shows that $|K(\eta)|$ behaves like $|\eta|^{\frac{n+\nu}{\alpha} - \frac{1}{2}} \mathrm{e}^{-\frac{\pi^2}{\alpha}\eta}$ for large $|\eta|$. Therefore both h/H and k/K are rapidly decreasing. Moreover, h/H and k/K are both in $\mathscr{P}_1$ (and in fact they are Schwartz class functions). It follows that

$$\left(\frac{h}{H} \otimes \frac{k}{K}\right)\mathscr{S}_2' \subset \mathscr{S}_2',$$

which implies (25). □

Of course the theorem holds for the reversals $(f(-t), \varphi(-t))$ of the given extreme value windows as well. The difference in generality between the two classes of windows is significant. For example, Theorem 4 immediately implies that for any fixed extreme value window f,

$$\bigcup_{(g,\gamma)\ \text{Gaussian}} \mathrm{Op}(g,\gamma) \subset \mathrm{Op}(f,f).$$

That is, even without using distinct analysis and synthesis windows, a fixed extreme value window pair is more general than the totality of Gaussian window pairs.

5 Summary

Given the Kohn–Nirenberg symbol σ_L of a linear operator L and a window pair (g,γ), one may apply Theorem 3 to decide whether L can be represented as a Gabor multiplier based on g and γ. And if it can, Theorem 2 shows how to compute the Gabor symbol λ in terms of σ_L. Of course if L does not belong to $\mathrm{Op}(g,\gamma)$, the problem is to find new windows that allow representation of L. This is potentially difficult since in general one wants to use windows that are "nice" (highly localized and smooth, say), but the results we have obtained in the present chapter nevertheless provide some guiding principles as well as certain concrete results. For example, if L is near-underspread, then one may determine appropriate Gaussian windows using Proposition 6. The general principle at play is that the window pair (g,γ) is more general (i.e., $\mathrm{Op}(g,\gamma)$ is larger), the greater the difference in localization, or width, between g and γ. Thus, from the perspective of representing a wide class of operators as Gabor multipliers, it pays to use distinct analysis and synthesis windows.

For operators that have very different characteristics in space (or time) versus in frequency, compatible windows can be used to construct a Gabor multiplier representation whose symbol exhibits this same distinction. Extreme value windows seem a natural window to use in this context, or even generally. However, Proposition 17 shows that, although any extreme value windows g,γ are more general than Gaussians, still not every linear operator belongs to $\mathrm{Op}(g,\gamma)$. Thus, for instance, a partial differential operator whose spreading function does not decay rapidly at infinity could not be represented as a Gabor multiplier in a non-trivial way using the windows considered in this chapter; this is a direction for further research. Concerning the somewhat technical notion of compatibility, it is remarkable that it leads to extreme value windows, which in retrospect seem very natural. Note that we do not have a complete classification of *weakly* compatible window pairs, and so there may be some interesting windows of this type which are not strongly compatible. Indeed it can be shown that a pair (g,γ) of gamma functions $g(t) = \Gamma(a - ibt)$, $\gamma(t) = \Gamma(c - ibt)$ falls into this category.

Lastly, we reiterate that, coupled with algorithms for fast computation of discretized Gabor multipliers, the results presented in the present chapter help to establish a general framework for practical evaluation of a wide class of linear operators, including nonstationary filters.

References

1. Busby, R.C., Smith, H.A.: Product-convolution operators and mixed-norm spaces. Trans. Amer. Math. Soc., **263**(2), 309–341 (1981)
2. Cordero, E., Tabacco, A.: Localization operators via time-frequency analysis. In: Ashino, R., Boggiatto, P., Wong, M.W. (eds.) Advances in pseudo-differential operators. Operator Theory: Advances and Applications, vol. 155. Birkhäuser, Boston (2004)
3. Feichtinger, H.G., Nowak, K.: A first survey of Gabor multipliers. In: Feichtinger, H.G., Strohmer, T. (eds.) Advances in Gabor Analysis, Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2003)
4. Fisher, R.A., Tippett, L.H.C.: Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc. Camb. Phil. Soc. **XXIV**(Pt. 2), 180–190 (1928)
5. Folland, G.B.: Harmonic analysis in phase space. Annals of Mathematics Studies. Princeton University Press, Princeton (1989)
6. Gibson, P.C., Zizler, P.: Compatible windows in Gabor analysis. Unpublished manuscript (2002)
7. Gröchenig, K.: Foundations of time-frequency analysis. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2001)
8. Kozek, W.: Adaptation of Weyl–Heisenberg frames to underspread environments. In: Feichtinger, H.G., Strohmer, T. (eds.) Gabor Analysis and Algorithms: Theory and Applications. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (1998)
9. Lamoureux, M.P., Gibson, P.C., Grossman, J.P., Margrave, G.F.: A fast, discrete Gabor transform via a partition of unity, CREWES Technical Report Volume 15, University of Calgary – Consortium for Research in Elastic Wave Exploration Seismology (2003)
10. Margrave, G.F., Ferguson, R.J.: Wavefield extrapolation by nonstationary phase shift. Geophys. **64**, 1067–1078 (1999)
11. Margrave, G.F., Ferguson, R.J., Lamoureux, M.P.: Approximate Fourier integral wavefield extrapolators for heterogeneous, anisotropic media. Canad. Appl. Math. Quarterly **10**(2), 331–343 (2002)
12. Margrave, G.F., Lamoureux, M.P., Grossman, J.P., Iliescu, V.: Gabor deconvolution of seismic data for source waveform and Q correction. In: 72nd Annual International Meeting, Society of Exploration Geophysicists, Expanded Abstract Volume, pp. 2190–2193 (2002)
13. Margrave, G.F., Gibson, P.C., Grossman, J.P., Henley, D.C., Iliescu, V., Lamoureux, M.P.: The Gabor transform, pseudodifferential operators, and seismic deconvolution. Integrated Comput. Aided Eng. **12**(1), 43–56 (2005)
14. Schwartz, L.: Théorie des noyaux. Proc. Internat. Congress Math. **1**, 220–230 (1950)
15. Wong, M.W.: Weyl Transforms. Universitext. Springer, New York (1998)

Extension of Berezin–Lieb Inequalities

John R. Klauder and Bo-Sture K. Skagerstam

Abstract The Berezin–Lieb inequalities provide upper and lower bounds for a partition function based on phase-space integrals that involve the Glauber–Sudarshan and Husimi representations, respectively. Generalizations of these representations have recently been introduced by the present authors, and in this article, we extend the use of these new representations to develop numerous analogs of the Berezin–Lieb inequalities that may offer improved bounds. Several examples illustrate the use of the new inequalities. Although motivated by problems in quantum mechanics, these results may also find applications in time-frequency analysis, a valuable cross-fertilization that has been profitably used at various times in the past.

Keywords Coherent states • Berezin–Lieb inequality • Quantum partition function • Partition function • Classical bounds • Husimi representation • Upper symbol • Lower symbol

1 Introduction

The Berezin–Lieb inequalities offer upper and lower bounds for partition functions of elementary quantum systems. In particular, for a system composed of a single canonical degree of freedom, let P and Q denote canonical Heisenberg variables, fulfilling the commutation relation $[Q, P] = iI$, in units where $\hbar = 1$. Let $|0\rangle$

J.R. Klauder (✉)
Department of Physics and Department of Mathematics, University of Florida, Gainesville, FL 32611, USA
e-mail: klauder@phys.ufl.edu

B.-S.K. Skagerstam
Department of Physics, The Norwegian University of Science and Technology, N-7491 Trondheim, Norway
e-mail: bo-sture.skagerstam@ntnu.no

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_13,

denote the normalized ground state of an elementary oscillator for which $(Q + iP)|0\rangle = 0$. Canonical coherent states for this system are taken to be states of the form (see, e.g., [1–3])

$$|p,q\rangle \equiv U[p,q]|0\rangle, \qquad U[p,q] \equiv \mathrm{e}^{\mathrm{i}(p\,Q - qP)} \tag{1}$$

for all $(p,q) \in \mathbb{R}^2$, where $U[p,q]$ denotes the unitary Weyl operator. Let $\mathcal{H} = \mathcal{H}(P,Q)$ denote the Hamiltonian for the system in question. The corresponding classical Hamiltonian is denoted by $H_{\mathrm{cl}}(p,q)$. We introduce two well-known symbols associated with $\mathcal{H}$, namely, the Husimi [4] symbol $H_H(p,q)$ defined by

$$H_H(p,q) \equiv \langle p,q|\mathcal{H}(P,Q)|p,q\rangle = \langle 0|\mathcal{H}(P+p,Q+q)|0\rangle, \tag{2}$$

and the Glauber–Sudarshan [5, 6] symbol $H_{\mathrm{G-S}}(p,q)$ implicitly defined by the operator representation

$$\mathcal{H}(P,Q) = \int H_{\mathrm{G-S}}(p,q)\,|p,q\rangle\,\langle p,q|\,\mathrm{d}p\,\mathrm{d}q/2\pi. \tag{3}$$

It follows from Eq. (2) that these two symbols are related by the integral equation

$$\begin{aligned} H_H(p',q') &= \int |\langle p',q'|p,q\rangle|^2\, H_{\mathrm{G-S}}(p,q)\,\mathrm{d}p\,\mathrm{d}q/2\pi \\ &= \int \mathrm{e}^{-[(p'-p)^2+(q'-q)^2]/2}\, H_{\mathrm{G-S}}(p,q)\,\mathrm{d}p\,\mathrm{d}q/2\pi. \end{aligned} \tag{4}$$

Armed with these definitions, the Berezin–Lieb inequalities [7, 8] read

$$\int \mathrm{e}^{-\beta\, H_H(p,q)}\,\mathrm{d}p\,\mathrm{d}q/2\pi \le \mathrm{Tr}[\mathrm{e}^{-\beta\,\mathcal{H}(P,Q)}] \le \int \mathrm{e}^{-\beta\, H_{\mathrm{G-S}}(p,q)}\,\mathrm{d}p\,\mathrm{d}q/2\pi\,. \tag{5}$$

In what follows we will implicitly rederive this inequality as a special example of our generalizations.

The purpose of this chapter is to extend such inequalities by offering infinitely many additional symbol pairs that can stand in place of the Husimi and Glauber–Sudarshan symbols in Eq. (5), thereby generalizing the original Berezin–Lieb inequalities.

2 Multiple Phase-Space Symbols

In a recent paper [9], the authors have introduced a wide class of phase-space symbols that are analogues of the Husimi and Glauber–Sudarshan dual pair. Let us first recall the principal elements of that study specialized to the discussion at hand.

We first introduce a nonnegative, trace-class operator $\sigma = \sigma^\dagger \geq 0$ which we normalize so that $\mathrm{Tr}(\sigma) = 1$. Such operators have the generic form given by

$$\sigma = \sum_{l=1}^{\infty} c_l \, |b_l\rangle \langle b_l|, \tag{6}$$

where $\{|b_l\rangle\}_{l=1}^{\infty}$ denotes a complete orthonormal sets of vectors, and the coefficients $\{c_l\}_{l=1}^{\infty}$ satisfy the conditions $c_l \geq 0$ and $\sum_{l=1}^{\infty} c_l = 1$. In short, σ enjoys all the properties to be a density matrix.

We shall make use of the function $\mathrm{Tr}(U[k,x]\sigma)$ defined for all (k,x) in phase space, and we restrict σ so that the expression

$$\mathrm{Tr}(U[k,x]\,\sigma) \neq 0 \tag{7}$$

for all $(k,x) \in \mathbb{R}^2$.

We next recall the Weyl representation of operators given by

$$A = \int \tilde{A}(k,x)\, U[k,x]\, \mathrm{d}k\, \mathrm{d}x/2\pi, \tag{8}$$

where

$$\tilde{A}(k,x) \equiv \mathrm{Tr}(U[k,x]^\dagger A). \tag{9}$$

Given two such operators A and B, it follows that

$$\mathrm{Tr}(A^\dagger B) = \int \tilde{A}(k,x)^*\, \tilde{B}(k,x)\, \mathrm{d}k\, \mathrm{d}x/2\pi. \tag{10}$$

In terms of the double Fourier transformation, given by

$$A(p,q) = \int \mathrm{e}^{\mathrm{i}(qk-px)}\, \tilde{A}(k,x)\, \mathrm{d}k\, \mathrm{d}x/2\pi, \tag{11}$$

and likewise for $B(p,q)$, it also follows that

$$\mathrm{Tr}(A^\dagger B) = \int A(p,q)^*\, B(p,q)\, \mathrm{d}p\, \mathrm{d}q/2\pi. \tag{12}$$

We next modify the symmetric expression for $\mathrm{Tr}(A^\dagger B)$ given by Eq. (10) so that

$$\mathrm{Tr}(A^\dagger B) = \int \left\{ \frac{\tilde{A}(k,x)^*}{\mathrm{Tr}(U[k,x]\sigma)} \right\} \{\mathrm{Tr}(U[k,x]\sigma)\tilde{B}(k,x)\}\, \mathrm{d}k\, \mathrm{d}x/2\pi$$

$$= \int \left\{ \frac{\tilde{A}(k,x)}{\mathrm{Tr}(U[k,x]^\dagger \sigma)} \right\}^* \{\mathrm{Tr}(U[k,x]\sigma)\, \tilde{B}(k,x)\}\, \mathrm{d}k\, dx/2\pi$$

$$\equiv \int \tilde{A}_{-\sigma}(k,x)^*\, \tilde{B}_\sigma(k,x)\, \mathrm{d}k\, \mathrm{d}x/2\pi$$

$$\equiv \int A_{-\sigma}(p,q)^*\, B_\sigma(p,q)\, \mathrm{d}p\, \mathrm{d}q/2\pi. \tag{13}$$

In the final line we have introduced the Fourier transform of the symbols in the line above. We next show that there are alternative expressions involving the symbols $A_{-\sigma}(p,q)$ and $B_\sigma(p,q)$ directly in their own space of definition rather than implicitly through a Fourier transformation.

We begin first with the symbol $B_\sigma(p,q)$. In particular, we note that

$$B_\sigma(p,q) = \int \mathrm{e}^{\mathrm{i}(kq-xp)}\, \mathrm{Tr}(U[k,x]\sigma)\, \tilde{B}(k,x)\, \mathrm{d}k\, \mathrm{d}x/2\pi$$

$$= \int \mathrm{Tr}(U[p,q]^\dagger\, U[k,x]\, U[p,q]\sigma)\, \mathrm{Tr}(U[k,x]^\dagger\, B)\, \mathrm{d}k\, \mathrm{d}x/2\pi$$

$$= \int \mathrm{Tr}(U[k,x]\, U[p,q]\sigma\, U[p,q]^\dagger)\, \mathrm{Tr}(U[k,x]^\dagger\, B)\, \mathrm{d}k\, \mathrm{d}x/2\pi$$

$$= \mathrm{Tr}(U[p,q]\sigma\, U[p,q]^\dagger\, B), \tag{14}$$

where in the second line we have used the Weyl form of the commutation relations, and in the last line we have used the Weyl representation Eq. (10), which leads us to the desired expression for $B_\sigma(p,q)$. This expression is the sought-for generalization of the Husimi representation; indeed, if $\sigma = |0\rangle\langle 0|$, it follows immediately that

$$B_\sigma(p,q) = \mathrm{Tr}(U[p,q]\,|0\rangle\langle 0|\, U[p,q]^\dagger\, B)$$

$$= \langle p,q|\, B\, |p,q\rangle = B_H(p,q). \tag{15}$$

For general σ, to find the expression for $A_{-\sigma}(p,q)$, we appeal to the relation

$$\mathrm{Tr}(A^\dagger\, B) = \int A_{-\sigma}(p,q)^*\, B_\sigma(p,q)\, \mathrm{d}p\, \mathrm{d}q/2\pi$$

$$= \int A_{-\sigma}(p,q)^* \mathrm{Tr}(U[p,q]\sigma\, U[p,q])^\dagger\, B)\, \mathrm{d}p\, \mathrm{d}q/2\pi, \tag{16}$$

an equation, which, thanks to its validity for all suitable operators B, carries the important implication that

$$A \equiv \int A_{-\sigma}(p,q)\, U[p,q]\sigma\, U[p,q]^\dagger\, \mathrm{d}p\, \mathrm{d}q/2\pi. \tag{17}$$

Observe that this equation implies a very general operator representation as a linear superposition of basic operators given by $U[p,q]\sigma\, U[p,q]^\dagger$, for a general choice of σ.

Equation (17) for A is the sought-for generalization of the Glauber–Sudarshan representation; indeed, if $\sigma = |0\rangle\langle 0|$, it follows immediately that

$$\begin{aligned} A &= \int A_{-\sigma}(p,q)\, U[p,q]\,|0\rangle\langle 0|U[p,q]^\dagger \,\mathrm{d}p\,\mathrm{d}q/2\pi \\ &= \int A_{-\sigma}(p,q)\; |p,q\rangle\langle p,q|\,\mathrm{d}p\,\mathrm{d}q/2\pi \\ &= \int A_{\mathrm{G-S}}(p,q)\; |p,q\rangle\langle p,q|\,\mathrm{d}p\,\mathrm{d}q/2\pi. \end{aligned} \tag{18}$$

Once again there is a direct connection between the generalization of the Husimi representation, $A_\sigma(p,q)$, and the generalization of the Glauber–Sudarshan representation, $A_{-\sigma}(p,q)$. In particular, it follows that

$$\begin{aligned} A_\sigma(r,s) &= \int A_{-\sigma}(p,q)\,\mathrm{Tr}(U[r,s]\sigma U[r,s]^\dagger U[p,q]\sigma U[p,q]^\dagger)\,\mathrm{d}p\,\mathrm{d}q/2\pi \\ &= \int A_{-\sigma}(p,q)\,\mathrm{Tr}(U[r-p,q-s] \\ &\quad \sigma U[r-p,q-s]^\dagger\sigma)\,\mathrm{d}p\,\mathrm{d}q/2\pi. \end{aligned} \tag{19}$$

This equation is a convolution, which just reflects the multiplicative connection between these two symbols in Fourier space.

3 Derivation of Inequalities

Let $\{|r\rangle\}_{r=1}^{\infty}$ denote an arbitrary, complete, orthonormal basis. Consider the expression [cf., Eq. (6)]

$$\begin{aligned} f(p,q|r) &\equiv \langle r|U[p,q]\sigma\, U[p,q)]^\dagger|r\rangle \\ &= \sum_{l=1}^{\infty} c_l\, |\langle r|U[p,q]|b_l\rangle|^2. \end{aligned} \tag{20}$$

It follows that

$$\int f(p,q|r)\,\mathrm{d}p\,\mathrm{d}q/2\pi = 1, \tag{21}$$

and also that

$$\sum_{r=1}^{\infty} f(p,q|r) = 1. \tag{22}$$

We can interpret these results in two different ways: On the one hand, $f(p,q|r)$ is a *probability density* on $\mathbb{R}^2$ for each value of r; on the other hand, $f(p,q|r)$ forms a *discrete probability* on $\{1,2,3,\ldots\}$ for each phase-space point (p,q).

3.1 Jensen's Inequality

The Jensen inequality [10, 11] applies to convex functions $\phi(x)$—such as $\mathrm{e}^{-\beta x}$—and arbitrary probability distributions on $x \in \mathbb{R}$. If $\langle(\cdot)\rangle$ denotes an average over that probability distribution, then the Jensen inequality reads

$$\phi(\langle x\rangle) \le \langle\phi(x)\rangle, \tag{23}$$

or, in particular,

$$\mathrm{e}^{-\beta\langle x\rangle} \le \langle \mathrm{e}^{-\beta x}\rangle. \tag{24}$$

This equation will be important in what follows.

Let $\mathcal{H}$ denote the Hamiltonian with a discrete spectrum $\{\mu_r\}_{r=1}^{\infty}$ and an associated set of eigenvectors $\{|r\rangle\}_{r=1}^{\infty}$ such that

$$\mathcal{H}|r\rangle = \mu_r|r\rangle. \tag{25}$$

It also follows that

$$\mathcal{H} = \sum_{r=1}^{\infty} \mu_r\,|r\rangle\langle r|. \tag{26}$$

Following Lieb [8], we first observe that

$$\begin{aligned}\langle r|\mathrm{e}^{-\beta\mathcal{H}}|r\rangle &= \exp[-\beta\langle r|\mathcal{H}|r\rangle] \\ &= \exp\left[-\beta\int H_{-\sigma}(p,q)\,f(p,q|r)\,\mathrm{d}p\,\mathrm{d}q/2\pi\right] \\ &\le \int \mathrm{e}^{-\beta H_{-\sigma}(p,q)}\,f(p,q|r)\,\mathrm{d}p\,\mathrm{d}q/2\pi. \end{aligned}\tag{27}$$

Summing on r leads to

$$\mathrm{Tr}(\mathrm{e}^{-\beta\mathcal{H}}) \leq \int \mathrm{e}^{-\beta H_{-\sigma}(p,q)}\,\mathrm{d}p\,\mathrm{d}q/2\pi. \tag{28}$$

Second, we learn that

$$\begin{aligned}\exp[-\beta H_\sigma(p,q)] &= \exp[-\beta\Sigma_r\,\mu_r\,f(p,q|r)] \\ &\leq \sum_{r=1}^{\infty}\exp[-\beta\mu_r]\,f(p,q|r).\end{aligned} \tag{29}$$

Integrating over $\mathbb{R}^2$ leads to

$$\int \mathrm{e}^{-\beta H_\sigma(p,q)}\,\mathrm{d}p\,\mathrm{d}q/2\pi \leq \mathrm{Tr}(\mathrm{e}^{-\beta\mathcal{H}}). \tag{30}$$

Above we have two separate inequalities, one an upper bound, the other a lower bound. These bounds apply for *any* choice of σ that fits our requirements, and so we can decouple the choice of σ and assert that σ can be chosen independently in the two cases. In summary, therefore, we have established the inequalities

$$\int \mathrm{e}^{-\beta H_{\sigma'}(p,q)}\,\mathrm{d}p\,\mathrm{d}q/2\pi \leq \mathrm{Tr}(\mathrm{e}^{-\beta\mathcal{H}}) \leq \int \mathrm{e}^{-\beta H_{-\sigma}(p,q)}\,\mathrm{d}p\,\mathrm{d}q/2\pi, \tag{31}$$

where σ' and σ may be chosen independently of each other. This possibility permits optimizing both bounds by taking the supremum over the lower bound and taking the infimum over the upper bound. The bounds as given by Eq. (31) now lead to upper and lower bounds, respectively, of the free energy $F(\beta) \equiv -\ln Z(\beta)/\beta$, where $Z(\beta)$ denotes the partition function, as well as bounds on the ground-state energy E_0 since $E_0 = \lim_{\beta\to\infty} F(\beta)$.

4 Symbols for the Lower Bound

We focus on the symbol

$$\begin{aligned}H_\sigma(p,q) &= \mathrm{Tr}(U[p,q]\sigma U[p,q]^\dagger\mathcal{H}(P,Q)) \\ &= \mathrm{Tr}(\mathcal{H}(P+p,Q+q)\sigma).\end{aligned} \tag{32}$$

For simplicity, we introduce the shorthand notation that

$$\overline{(\cdot)} \equiv \mathrm{Tr}((\cdot)\sigma). \tag{33}$$

In that case we find, e.g., that

$$(q)_\sigma \equiv \mathrm{Tr}((Q+q)\sigma) \equiv q + \overline{Q}, \tag{34}$$

where the notation $(q)_\sigma$ is the symbol $H_\sigma(p,q)$ when the operator $\mathcal{H}$ is simply Q. Below we list a table of symbols needed for our present purposes:

$$\begin{aligned}
(q)_\sigma &= q + \overline{Q}\ ,\\
(p)_\sigma &= p + \overline{P}\ ,\\
(q^2)_\sigma &= q^2 + 2q\overline{Q} + \overline{Q^2}\ ,\\
(p^2)_\sigma &= p^2 + 2p\overline{P} + \overline{P^2}\ ,\\
(qp)_\sigma &= qp + q\overline{P} + p\overline{Q} + \overline{QP}\ ,\\
(pq)_\sigma &= pq + p\overline{Q} + q\overline{P} + \overline{PQ}\ ,\\
(q^4)_\sigma &= q^4 + 4q^3\overline{Q} + 6q^2\overline{Q^2} + 4q\overline{Q^3} + \overline{Q^4}\ ,\\
(p^4)_\sigma &= p^4 + 4p^3\overline{P} + 6p^2\overline{P^2} + 4p\overline{P^3} + \overline{P^4}\ ,\\
(q^2p^2)_\sigma &= q^2p^2 + 2pq^2\overline{P} + 2qp^2\overline{Q} + q^2\overline{P^2} + p^2\overline{Q^2} + 4qp\overline{QP}\\
&\quad +2q\overline{QP^2} + 2p\overline{Q^2P} + \overline{Q^2P^2}\ ,\\
(p^2q^2)_\sigma &= p^2q^2 + 2pq^2\overline{P} + 2qp^2\overline{Q} + q^2\overline{P^2} + p^2\overline{Q^2} + 4pq\overline{PQ}\\
&\quad +2q\overline{P^2Q} + 2p\overline{PQ^2} + \overline{P^2Q^2}\ .
\end{aligned} \tag{35}$$

Note that on the left-hand side the order matters, i.e., $(qp)_\sigma \neq (pq)_\sigma$, etc. We also notice that for the quadratic symbols

$$\begin{aligned}
(q^2)_\sigma &= (q+\overline{Q})^2 + \Delta(Q)\ ,\\
(p^2)_\sigma &= (p+\overline{P})^2 + \Delta(P)\ ,\\
\frac{1}{2}[(qp)_\sigma + (pq)_\sigma] &= (q+\overline{Q})(p+\overline{P}) + \Delta(Q,P)\ ,
\end{aligned} \tag{36}$$

in terms of the variances $\Delta(\mathcal{O}) \equiv \overline{\mathcal{O}^2} - \overline{\mathcal{O}}^2$ and $\Delta(\mathcal{O}_1,\mathcal{O}_2) \equiv (\overline{\mathcal{O}_1\mathcal{O}_2 + \mathcal{O}_2\mathcal{O}_1})/2 - \overline{\mathcal{O}_1}\,\overline{\mathcal{O}_2}$. For a conventional minimal uncertainty state, e.g., $\Delta(Q)\Delta(P) = 1/4$ and $\Delta(Q,P) = 0$.

We also introduce a special-case table based on a symmetry we shall impose on σ and to be made use of below, namely, that all odd-order averages vanish, i.e., $\overline{Q} = \overline{Q^3} = \overline{P} = \overline{P^3} = \overline{Q^2P} = 0$. This special-case table reads

$$\begin{aligned}
(q)_\sigma &= q,\\
(p)_\sigma &= p,
\end{aligned}$$

$$
\begin{aligned}
(q^2)_\sigma &= q^2 + \overline{Q^2}, \\
(p^2)_\sigma &= p^2 + \overline{P^2}\,, \\
(qp)_\sigma &= qp + \overline{QP}\,, \\
(pq)_\sigma &= pq + \overline{PQ}\,, \\
(q^4)_\sigma &= q^4 + 6q^2\overline{Q^2} + \overline{Q^4}\,, \\
(p^4)_\sigma &= p^4 + 6p^2\overline{P^2} + \overline{P^4}\,, \\
(q^2p^2)_\sigma &= q^2p^2 + q^2\overline{P^2} + p^2\overline{Q^2} + 4qp\overline{QP} + \overline{Q^2P^2}\,, \\
(p^2q^2)_\sigma &= p^2q^2 + q^2\overline{P^2} + p^2\overline{Q^2} + 4pq\overline{PQ} + \overline{P^2Q^2}\,.
\end{aligned} \tag{37}
$$

5 Symbols for the Upper Bound

The construction of the upper limit is somewhat more involved than that for the lower limit. We start with Eq. (17), which is

$$
A = \int A_{-\sigma}(p,q)\, U[p,q]\sigma U[p,q]^\dagger \,\mathrm{d}p\,\mathrm{d}q/2\pi. \tag{38}
$$

For reasons of clarity we limit ourselves to a number-operator diagonal form for σ, i.e., $\sigma = \sum_{n=0}^{\infty} c_n\,|n\rangle\langle n|$, $c_n \geq 0$, and $\sum_{n=0}^{\infty} c_n = 1$, where $N\,|n\rangle = n\,|n\rangle$ for the number eigenstates $\{\,|n\rangle\,\}$. We learn that in general

$$
\begin{aligned}
A &= \sum_{n=0}^{\infty} c_n \int A_{-\sigma}(p,q)\, U[p,q]|n\rangle\langle n|U[p,q]^\dagger \,\mathrm{d}p\,\mathrm{d}q/2\pi \\
&\equiv \sum_{n=0}^{\infty} c_n \int A_{-\sigma}(p,q)\,|p,q;n\rangle\langle p,q;n|\,\mathrm{d}p\,\mathrm{d}q/2\pi,
\end{aligned} \tag{39}
$$

in terms of the so-called semi-coherent states or displaced coherent states $|p,q;n\rangle \equiv U[p,q]|n\rangle$ (see, e.g., [12–17]). To see what this means, let us take a simple example with $A = P^2 + Q^2$. Since an operator is determined by its expectation value in canonical coherent states, it is sufficient to consider the Husimi symbol $A_H(p,q)$ as given by Eq. (2), i.e.,

$$
\begin{aligned}
\langle r,s;0|\,(P^2+Q^2)\,|r,s;0\rangle &= \langle 0|\,[(P+r)^2+(Q+s)^2]\,|0\rangle = (r^2+s^2)+1 \\
&\equiv \sum_{n=0}^{\infty} c_n \int [(p+r)^2 + (q+s)^2 + k_2]\; p_n(p,q)\,\mathrm{d}p\,\mathrm{d}q/2\pi\,,
\end{aligned} \tag{40}
$$

where

$$p_n(p,q) \equiv |\langle 0|p,q;n\rangle|^2 = \mathrm{e}^{-(p^2+q^2)/2}(p^2+q^2)^n/2^n n!\,. \tag{41}$$

Here, we have made use of the Ansatz

$$(P^2+Q^2)_{-\sigma}(p,q) \equiv p^2+q^2+k_2, \tag{42}$$

where k_2 is a constant to be determined, and we immediately learn that

$$k_2 = -1-2\sum_{n=0}^{\infty} c_n n \equiv -1-2\overline{n}, \tag{43}$$

where we have defined mean values $\overline{f(n)} \equiv \sum_{n=0}^{\infty} c_n f(n)$. It now, e.g., follows that the right-hand side of Eq. (17), with the upper symbol as given by Eqs. (42) and (43), has $|n\rangle$ as an eigenvector with eigenvalue $2n+1$. It is not entirely trivial to verify this explicitly, but it follows using the properties of displaced coherent states as well as properties of the conventional associated Laguerre polynomials L_n^m:

$$L_n^m(x) = \sum_{k=0}^{n}(-1)^k \frac{(n+m)!x^k}{(n-k)!k!(m+k)!}\,. \tag{44}$$

In like fashion, it follows for $A = (P^2+Q^2)^2$ and the corresponding Husimi symbol that

$$\begin{aligned}\langle r,s;0|(P^2+Q^2)^2|r,s;0\rangle &= \langle 0|[(P+r)^2+(Q+s)^2]^2|0\rangle = (r^2+s^2+1)^2\\ &\equiv \sum_{n=0}^{\infty} c_n \int [((p+r)^2+(q+s)^2)^2 + k_4((p+r)^2+(q+s)^2)+k_6]\\ &\quad \times p_n(p,q)\,\mathrm{d}p\,\mathrm{d}q/2\pi,\end{aligned} \tag{45}$$

expressed in terms of the (assumed) symbol

$$((P^2+Q^2)^2)_{-\sigma}(p,q) = (p^2+q^2)^2 + k_4(p^2+q^2)+k_6. \tag{46}$$

One now finds, making use of Eq. (41), that

$$k_4 = 2-8\sum_{n=0}^{\infty} c_n(n+1) = -6-8\overline{n}\,, \tag{47}$$

and

$$k_6 = 1 - 4\sum_{n=0}^{\infty} c_n\,(n+1)(n+2) - 2k_4 \sum_{n=0}^{\infty} c_n\,(n+1)$$
$$= 5 + 16\overline{n} + 16\overline{n}^2 - 4\overline{n^2}. \tag{48}$$

In a similar manner and for $A = Q^4$, we can write

$$(Q^4)_{-\sigma}(p,q) = q^4 + a_2 q^2 + a_4\ , \tag{49}$$

where

$$a_2 = -3(1 + 2\overline{n})\ , \tag{50}$$

and

$$a_4 = 3\left(\frac{1}{4} + \frac{3}{2}\overline{n} + 2\overline{n}^2 - \frac{1}{2}\overline{n^2}\right)\ . \tag{51}$$

The expressions above now relate the standard symbols to the generalized symbols. Extension of these expressions to other polynomials in P and Q is straightforward.

6 Examples

With the special choice for σ considered above, i.e., $\sigma = \sum_{n=0}^{\infty} c_n |n\rangle\langle n|$, $c_n \geq 0$, and $\sum_{n=0}^{\infty} c_n = 1$, we will now consider some specific examples in order to illustrate the use of the generalized upper and lower symbols. We first remark that in the trivial case of a harmonic oscillator with $H = (P^2 + Q^2)/2$, such that $Z(\beta) = 1/[2\sinh(\beta/2)]$, the lower symbol Eq. (36) and the upper symbol Eq. (42), together with the bounds in Eq. (31), lead to the expression

$$\mathrm{e}^{-\beta(\Delta(P)+\Delta(Q))/2}/\beta \leq Z(\beta) \leq \mathrm{e}^{\beta(1/2+\overline{n})/2}/\beta, \tag{52}$$

which, obviously, is true. We can optimize this expression in the form

$$\mathrm{e}^{-\beta/2}/\beta \leq Z(\beta) \leq \mathrm{e}^{\beta/2}/\beta. \tag{53}$$

From the corresponding lower bound we then obtain an upper bound on the ground-state energy $E_0 \leq 1/2$ since $E_0 = -\lim_{\beta\to\infty} \ln Z(\beta)/\beta$. In the high-temperature limit, i.e., $\beta \to 0$, the bounds in Eq. (53) exactly reproduce the classical Gibbs

partition function $Z_{\rm cl}(\beta)/2\pi = 1/\beta$ taking the fundamental phase-space volume 2π into account and making use of

$$Z_{\rm cl}(\beta) = \int {\rm e}^{-\beta H_{\rm cl}(p,q)}\, {\rm d}p\, {\rm d}q\ , \tag{54}$$

with, of course, $H_{\rm cl}(p,q) = (p^2+q^2)/2$.

6.1 A Nonlinear Oscillator

Here we consider Hamiltonians of the form $\mathcal{H} = \mathcal{H}(N)$, where N is the usual number operator. We study this example more for its ease of analysis and pedagogical value. We choose as our example $\mathcal{H} = (N-a)(N-b)$. Such a form of a Hamiltonian has its roots in, e.g., the description of a single-mode nonlinear Kerr medium in quantum optics or a single vibrational mode beyond the harmonic approximation. We make the choice $a = 1$ and $b = 5$. We observe that the partition function $Z(\beta) = \sum_{n=0}^{\infty} \exp[-\beta(n-1)(n-5)]$ then has the form $Z(\beta) \simeq \exp(4\beta)$ for large values of β. A straightforward application of Poisson resummation techniques also leads to the behavior $Z(\beta) \simeq \sqrt{\pi/\beta}/2$ for small values of β, which corresponds to the high-temperature limit of the classical partition function $Z_{\rm cl}/2\pi$ using Eq. (54) with $H_{\rm cl} = (p^2+q^2)^2/4 - 7(p^2+q^2)/2 + 33/4$.

We may then combine these factors for $\mathcal{H} = (N-1)(N-5)$ at hand by noting that

$$\begin{aligned}(N-1)(N-5) &= \frac{1}{4}(P^2+Q^2-1)^2 - 6\frac{1}{2}(P^2+Q^2-1)+5\\ &= \frac{1}{4}(P^4+Q^4+P^2Q^2+Q^2P^2) - 7\frac{1}{2}(P^2+Q^2)+33/4.\end{aligned} \tag{55}$$

Consequently,

$$\begin{aligned}H_\sigma(p,q) &= \frac{1}{4}[(p^4)_\sigma + (q^4)_\sigma + (p^2q^2)_\sigma + (q^2p^2)_\sigma] - 7\frac{1}{2}[(p^2)_\sigma + (q^2)_\sigma] + 33/4\\ &= \frac{1}{4}[\,p^4 + 6p^2\overline{P^2} + \overline{P^4} + q^4 + 6q^2\overline{Q^2} + \overline{Q^4} + q^2p^2 + q^2\overline{P^2} + p^2\overline{Q^2}\\ &\quad +4qp\overline{QP} + \overline{Q^2P^2} + p^2q^2 + q^2\overline{P^2} + p^2\overline{Q^2} + 4pq\overline{PQ} + \overline{P^2Q^2}\,]\\ &\quad -7\frac{1}{2}[\,p^2 + \overline{P^2} + q^2 + \overline{Q^2}\,] + 33/4.\end{aligned} \tag{56}$$

Since we have restricted our choice of σ so that it is only a function of N, i.e., $\sigma = \sigma(N)$, σ has now a symmetry that makes $\overline{Q^2} = \overline{P^2} \equiv C_2$, $\overline{P^4} = \overline{Q^4} \equiv C_4$, $\overline{Q^2P^2} = \overline{P^2Q^2} \equiv C_{22}$, and importantly that $\overline{QP} + \overline{PQ} = 0$. The three constants C_2, $C_4, and\ C_{22}$ are the only remnants of σ in $H_\sigma(p,q)$, and of necessity, they satisfy $C_2 \geq 1/2$, $C_4 \geq C_2^2$, and $C_4 \geq C_{22}$. With the restriction $\sigma = \sigma(N)$ we can actually be more precise and write

$$C_2 = \frac{1}{2} + \bar{n},\ C_{22} = \frac{1}{2}\left(\overline{n^2} + \bar{n} + \frac{1}{2}\right),\ C_4 = \frac{3}{2}\left(\overline{n^2} + \bar{n} + \frac{1}{2}\right). \tag{57}$$

Putting this information together, we find that

$$H_\sigma(p,q) = \frac{1}{4}(p^2+q^2)^2 + K_1(p^2+q^2) + K_2, \tag{58}$$

where

$$\begin{aligned} K_1 &\equiv \frac{7}{4}(C_2 - 2) = \frac{7}{4}\left(\bar{n} - \frac{3}{2}\right), \\ K_2 &\equiv C_4 + \frac{1}{2}C_{22} - 7C_2 + \frac{33}{4} = \overline{(n-3)^2} - 4. \end{aligned} \tag{59}$$

We note the fact that $H_\sigma(p,q)$ is a function only of the combination (p^2+q^2) on the basis of our restriction that $\sigma = \sigma(N)$. It follows, therefore, that the lower bound of interest is given by

$$\int \exp\{-\beta H_\sigma(p,q)\}\,\mathrm{d}p\,\mathrm{d}q/2\pi = \frac{1}{2}\int_0^\infty \exp\left\{-\beta\left[\frac{1}{4}s^2 + K_1 s + K_2\right]\right\}\mathrm{d}s, \tag{60}$$

where we have passed to polar coordinates and set $s \equiv (p^2+q^2)$. The upper-bound integral is a function of β as well as the σ-parameters, C_2, C_4, and C_{22}, i.e., the independent mean value $\bar{n}$ and dispersion $\overline{(n-\bar{n})^2}$ parameters.

The lower bound of Eq. (31) together with Eq. (60) now leads to the lower bound $\sqrt{\pi/\beta}/2 \leq Z(\beta)$ as $\beta \to 0$. This lower-bound again corresponds to the high-temperature limit for the classical partition function $Z_{\mathrm{cl}}(\beta)/2\pi$. By making use of $E_0 = -\lim_{\beta\to\infty} \ln Z(\beta)/\beta$, Eq. (60) leads to the upper limit $E_0 \leq -4$ using the state $\sigma = |3\rangle\langle 3|$. We observe that such a state will not strictly satisfy the restriction imposed by Eq. (7) since $\mathrm{Tr}(U[k,x]\sigma)$ then will be zero at isolated points away from the origin $k = x = 0$. But, in fact, the restriction Eq. (7) is then not required if A is a polynomial in P and Q since the symbol $\tilde{A}(k,x)$ as defined in Eq. (9) will involve derivatives of delta functions with support at the origin [9, 18].

The upper bound of Eq. (31), using Eqs. (42) and (46), now leads to

$$Z(\beta) \leq \frac{1}{2}\int_0^\infty \exp\left\{-\beta\left[\frac{1}{4}(s^2+k_4 s+k_6)-\frac{7}{2}(s+k_2)+\frac{33}{4}\right]\right\}\mathrm{d}s, \quad (61)$$

where the parameters k_2, k_4, and k_6 are given by Eqs. (43), (47), and (48), respectively. It is now evident again that Eq. (61) reproduces the high-temperature limit of the classical partition function $Z_{\mathrm{cl}}(\beta) \simeq \sqrt{\pi/\beta}/2$. The upper bound of Eq. (31) gives unfortunately now a rather poor lower bound on the ground-state energy $E_0 \geq -12 - 9\bar{n} - \overline{n^2}$, i.e., $E_0 \geq -12$.

6.2 An Anharmonic Oscillator

We next consider the Hamiltonian $H = (P^2+Q^2)/2+\lambda Q^4/2 \geq 0, \lambda > 0$, to define the partition function. With the lower and upper symbols as given by Eqs. (37), (42), and (49), we now find that

$$Z(\beta) \leq \frac{1}{\beta\sqrt{2\pi}}\mathrm{e}^{-\beta(k_2+\lambda a_4)/2}\int \mathrm{e}^{-(x^2+\lambda(x^4/\beta+a_2x^2))/2}\mathrm{d}x\,, \quad (62)$$

and

$$Z(\beta) \geq \frac{1}{\beta\sqrt{2\pi}}\mathrm{e}^{-\beta(\Delta(P)+\Delta(Q)+\lambda\overline{Q^4})/2}\int \mathrm{e}^{-(x^2+\lambda(x^4/\beta+6x^2\overline{Q^2}))/2}\mathrm{d}x\,. \quad (63)$$

In the limit of large β, the lower bound on $Z(\beta)$ and the fact that $H \geq 0$ then lead to $0 \leq E_0 \leq (1+\lambda\overline{Q^4})/2$. With $\sigma = |0\rangle\langle 0|$ one finds the upper-bound $E_0 \leq (1+3\lambda/4)/2$ which, e.g., can be compared to the "exact" numerical value of $2E_0 = 1.392351641530\ldots$ for $\lambda = 1$ [19]. We expect that this upper bound could be improved with a different choice of σ. A consequence of the upper and lower bounds in Eqs. (62) and (63) now is that for sufficiently small β the upper and lower bounds converge to the well-studied (see, e.g., Refs. [20–22]) classical and asymptotic form

$$\begin{aligned} Z(\beta) &= \frac{1}{\beta\sqrt{2\pi}}\int \mathrm{e}^{-x^2/2-\lambda x^4/2\beta}\mathrm{d}x \equiv Z_{\mathrm{cl}}(\beta)/2\pi \\ &= \frac{1}{2\lambda\sqrt{2\pi}}\sqrt{\frac{\lambda}{\beta}}\mathrm{e}^{\beta/16\lambda}K_{1/4}(\beta/16\lambda) \end{aligned} \quad (64)$$

using Eq. (54) with $H_{\mathrm{cl}}(p,q) = (p^2+q^2)/2+\lambda q^4/2$. The expression in Eq. (64) involves all the energy states of the anharmonic oscillator in a highly non-trivial manner. In our case we are specifically interested in the limit $\beta \to 0$, i.e., $Z(\beta) \simeq \Gamma(1/4)(2\beta/\lambda)^{1/4}/2\beta\sqrt{2\pi}$.

7 Comments

For clarity, we have mainly focused on matrices $\sigma = \sigma(N)$ which meant that $\sigma = \sum_{n=0}^{\infty} c_n |n\rangle\langle n|$. More general matrices of course would involve expansions of the form

$$\sigma = \sum_{n,n'=0}^{\infty} c_{n,n'} |n\rangle\langle n'| \tag{65}$$

expressed in terms of a general matrix $\{c_{n,n'}\}$ that still ensures that σ has all the properties of a partition function. The use of such more general choices for σ will inevitably lead to expressions involving the matrix elements [12–17]

$$\langle n|U[p,q]|n'\rangle$$

$$= \sqrt{\frac{2^{n'}n!}{2^n n'!}} \exp\left[-\frac{1}{4}(p^2+q^2)\right](q+ip)^{n-n'} L_n^{n-n'}\left(\frac{1}{2}(p^2+q^2)\right), \tag{66}$$

for $n \geq n'$ expressed in terms of the associated Laguerre polynomials Eq. (44); instead, when $n < n'$, use $\langle n|U[p,q]|n'\rangle = \langle n'|U[-p,-q]|n\rangle^*$. The simple example where $\mathcal{H} = P^2 + \omega^2 Q^2$, $\omega \neq 1$, shows that the optimal choice of σ is not always given by $|0\rangle\langle 0|$, where $(Q+iP)|0\rangle = 0$, but in the present case by $\sigma = |0;\omega\rangle\langle 0;\omega|$, where $(\omega Q + iP)|0;\omega\rangle = 0$. This remark serves to confirm that the generalized representations have the possibility to make better bounds. It may be true that choices for σ of the form $|\psi\rangle\langle\psi|$ (analogues of pure states) may be optimal and that perhaps choosing $|\psi\rangle$ as the ground state of the Hamiltonian under examination may lead to optimal bounds. Those are interesting questions for the future.

8 Conclusion

We have developed new, classical, phase-space bounds to deal with specialized (i.e., the partition function) questions that arise in quantum mechanics, and which, by their very nature, are technically easier to deal with than in their original form. It is quite likely that the generalized phase-space symbols we have introduced may have additional applications both in quantum mechanics and in time-frequency analysis.

References

1. Klauder, J.R., Skagerstam, B.-S.: Coherent States: Applications in Physics and Mathematical Physics. World Scientific, Singapore, (1985) and Beijing (1986); Perelomov, A.M.: Generalized Coherent States and Their Applications. Springer, Berlin/Heidelberg (1986); Zhang, W.M., Feng, D.H., Gilmore, R.: Coherent states: theory and some applications. Rev. Mod. Phys. **62**, 927 (1990)
2. Skagerstam, B.-S.: Coherent states—some applications in quantum field theory and particle physics. In: Feng, D.H., Klauder, J.R., Strayer, M.R. (eds.) Coherent States: Past, Present, and Future, World Scientific, Singapore (1994)
3. Klauder, J.R.: The current state of coherent states. Contribution to the 7th ICSSUR Conference, Boston (2001) (arXiv:quant-ph/0110108v1)
4. Husimi, K.: Some formal properties of the density matrix. Proc. Math. Phys. Soc. Japan **22**, 264 (1940)
5. Sudarshan, E.C.G.: Equivalence of semiclassical and quantum mechanical descriptions of statistical light beams. Phys. Rev. Lett. **10**, 277 (1963)
6. Glauber, R.J.: Coherent and incoherent states of the radiation field. Phys. Rev. **131**, 2766 (1963)
7. Berezin, F.A.: The general concept of quantization. Commun. Math. Phys. **40**, 153 (1975)
8. Lieb, E.H.: The classical limit of quantum spin systems. Commun. Math. Phys. **31**, 327 (1973); and Coherent states as a tool for obtaining rigorous bounds. In: Feng, D.H., Klauder, J.R., Strayer, M.R. (eds.) Coherent States: Past, Present, and Future. World Scientific, Singapore (1994)
9. Klauder, J.R., Skagerstam, B.-S.: Generalized phase-space representation of operators. J. Phys. A: Math. Theor. **40**, 2093 (2007)
10. Hardy, G.H., Littlewood, J.E., Polya, G.: Some theorems concerning monotonic functions. In: Inequalities. Cambridge University Press, Cambridge (1988)
11. Rudin, W.: Real and Complex Analysis. McGraw-Hill, New York (1970), Chapter 3
12. Carruthers, P., Nieto, M.N.: Coherent states and the harmonic oscillator. Am. J. Phys. **33**, 537 (1965)
13. Bagrov, V.G., Gitman, D.M., Kuchin, V.A.: External field in quantum electrodynamics and coherent states. In: Actual Problems of Theoretical Problems. Collection of papers to D. D. Ivanenko, MGU, Moscow (1976)
14. Fradkin, E.S., Gitman, D.M., Shvartsman, S.M.: Quantum Electrodynamics with Unstable Vacuum. Springer, Berlin (1991)
15. de Olivera, F.A.M., Kim, M.S., Knight, P.L.: Properties of displaced number states. Phys. Rev. A **41**, 2645 (1990)
16. Möller, K.B., Jörgensen, T.G., Dahl, J.P.: Displaced squeezed numbers states: position space representations, inner product and some other applications. Phys. Rev. A **54**, 5378 (1996)
17. Nieto, M.M.: Displaced and squeezed number states. Phys. Lett. A **229**, 135 (1997)
18. Klauder, J.R.: Continuous-representation theory III. On functional quantization of classical systems. J. Math. Phys. **5**, 177 (1964)
19. Banerjee, K., Bhatnagar, S.P., Choudhry, V., Kanwal, S.S.: The anharmonic oscillator. Proc. Roy. Soc. A **360**, 575 (1978)
20. Borel, E.: Mémoire sur les séries divergentes. Ann. Sci. École Norm. Sup. (3) **16**, 9–131 (1899)
21. Hardy, G.H.: Divergent Series, Chelsea, New York (1992). ISBN 978-0-8218-2649-2
22. Zinn-Justin, J.: Perturbation series at large order in quantum mechanics and field theories: application to the problem of resummation. Phys. Rep. **70**, 109 (1981)

Bilinear Calderón–Zygmund Operators

Diego Maldonado

Abstract This chapter is based on the presentation "*Generalized bilinear Calderón –Zygmund operators and applications*" delivered by the author during the 2008 February Fourier Talks at the Norbert Wiener Center for Harmonic Analysis and Applications, Department of Mathematics, University of Maryland, College Park, on February 21st. In turn, that presentation was based on material from the article "*Weighted norm inequalities for paraproducts and bilinear pseudodifferential operators with mild regularity*," J. Fourier Anal. Appl. **15** (2), (2009), 218–261, by Virginia Naibo and the author. This chapter also surveys some more recent results concerning the symbolic calculus and mapping properties of bilinear pseudodifferential operators.

The author would like to thank Professor John Benedetto for his inspiration and constant support as well as the organizers of the FFT for their kind invitation and hospitality.

Keywords Hörmander class • Bilinear pseudodifferential operator • Bilinear Hörmander class • Molecular paraproduct • Bilinear Calderón–Zygmund kernel • Bilinear Calderón–Zygmund operator

1 Introduction

Let $\mathscr{S}(\mathbb{R}^n)$ denote the Schwartz class in $\mathbb{R}^n$. Given $f \in \mathscr{S}(\mathbb{R}^n)$ its Fourier transform $\hat{f}$ is

$$\hat{f}(\xi) := \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}x\cdot\xi} f(x)\,\mathrm{d}x, \qquad \xi \in \mathbb{R}^n.$$

D. Maldonado (✉)
Kansas State University, Department of Mathematics, 138 Cardwell Hall, Manhattan, KS-66506, USA,
e-mail: dmaldona@math.ksu.edu

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_14,

A classical theorem of Mikhlin ([34], [38, p. 263]) establishes that for a function $m \in C^\infty(\mathbb{R}^n \setminus \{0\})$, satisfying the estimates

$$|\partial_\xi^\alpha m(\xi)| \leq C_\alpha |\xi|^{-|\alpha|}, \quad \xi \in \mathbb{R}^n \setminus \{0\}, \tag{1}$$

for all multi-index $\alpha \in \mathbb{N}_0^n$ with $|\alpha| \leq [n/2] + 1$, the operator T_m defined by

$$T_m f(x) := \int_{\mathbb{R}^n} m(\xi) \hat{f}(\xi) \mathrm{e}^{\mathrm{i}\xi \cdot x} \, \mathrm{d}\xi$$

is bounded from $L^p(\mathbb{R}^n)$ into $L^p(\mathbb{R}^n)$ whenever $1 < p < \infty$. $T_m f$ is the result of multiplying by m the spectrum of f, and when this action preserves the space $L^p(\mathbb{R}^n)$, T_m (as well as m) is called a *multiplier* for $L^p(\mathbb{R}^n)$. When the recipe for multiplication of the spectrum depends on the point x, we are looking at *pseudo-differential operators*, and, instead of multipliers, we speak about *symbols*. Formally, a symbol σ defines a pseudo-differential operator T_σ given by

$$T_\sigma f(x) := \int_{\mathbb{R}^n} \sigma(x, \xi) \hat{f}(\xi) \mathrm{e}^{\mathrm{i}\xi \cdot x} \, \mathrm{d}\xi.$$

Among the most useful classes of symbols is the Hörmander class $S^m_{\rho,\delta}$. More precisely, let $m \in \mathbb{R}$ and $0 \leq \delta, \rho \leq 1$. A function $\sigma \in C^\infty(\mathbb{R}^n \times \mathbb{R}^n)$ belongs to *Hörmander's class* $S^m_{\rho,\delta}$ if

$$|\partial_x^\alpha \partial_\xi^\beta \sigma(x, \xi)| \leq C_{\alpha,\beta} (1 + |\xi|)^{m + \delta|\alpha| - \rho|\beta|}, \quad x, \xi \in \mathbb{R}^n, \tag{2}$$

for all $\alpha, \beta \in \mathbb{N}_0^n$. Notice that now m denotes real number, called the *order* of σ. Here are some classical theorems involving Hörmander classes that will be of relevance in the later discussion.

Theorem 1 (Kohn–Nirenberg [29]). *For $\sigma \in S^0_{1,0}$, we have that $T_\sigma : L^2 \to L^2$ is bounded, and the classes $S^0_{1,0}$ possess a symbolic calculus for transposition and composition.*

Theorem 2 (Hörmander [27]). *For $0 \leq \delta < \rho < 1$ and $\sigma \in S^0_{\rho,\delta}$, we have that $T_\sigma : L^2 \to L^2$ is bounded, and the classes $S^0_{\rho,\delta}$ possess a symbolic calculus for transposition and composition.*

Theorem 3 (Calderón–Vaillancourt [9]). *For $0 \leq \rho < 1$ and $\sigma \in S^0_{\rho,\rho}$, we have that $T_\sigma : L^2 \to L^2$ is bounded.*

Around 1969, L. Nirenberg asked the following question: Suppose that the symbol σ verifies

$$|\partial_\xi^\beta \sigma(x, \xi)| \leq C_\beta (1 + |\xi|)^{-|\beta|}, \quad x, \xi \in \mathbb{R}^n, \tag{3}$$

for all $\beta \in \mathbb{N}_0^n$, with no a priori regularity in the x-variable (compare (3) to (2) with $\rho = 1$ and $m = \delta = 0$), does it follow that $T_\sigma : L^2 \to L^2$? In 1972, this question was answered in the negative by Ching, one of Nirenberg's students, who constructed a counterexample; see [12]. Actually, Ching's counterexample is smooth in the x-variable, but its x-derivatives lack a pointwise control as in (2). The answer to the question, however, becomes positive if, in addition to (3), σ is assumed to be homogeneous of degree 0 in the ξ-variable (see [38, p. 268]), and this goes back to the pioneering contributions of Calderón and Zygmund.

Another question is: How about a pointwise control as in (2), but adapted to milder regularity conditions in the x-variable? In this direction, let ω be a modulus of continuity and let Σ_ω be the class of all $\sigma(x,\xi)$ satisfying, for $x, \xi \in \mathbb{R}^n$ and $\beta \in \mathbb{N}_0^n$, the inequalities

$$|\partial_\xi^\beta \sigma(x,\xi)| \le C_\beta \, (1+|\xi|)^{-|\beta|}$$

and

$$|\partial_\xi^\beta \sigma(x+h,\xi) - \partial_\xi^\beta \sigma(x,\xi)| \le C_\beta \, \omega(|h|) \, (1+|\xi|)^{-|\beta|}.$$

Then we have

Theorem 4 (Coifman–Meyer, [13, p. 38]). *The following statements are equivalent:*

(i) $\displaystyle\int_0^1 \omega^2(t)\,\frac{\mathrm{d}t}{t} < \infty.$

(ii) For all $\sigma \in \Sigma_\omega$, T_σ *is bounded on* $L^2(\mathbb{R}^n)$.

(iii) For all $\sigma \in \Sigma_\omega$ *and all* $p \in (1,\infty)$, T_σ *is bounded on* $L^p(\mathbb{R}^n)$.

(iv) For all $\sigma \in \Sigma_\omega$, T_σ *is bounded from* $H^1(\mathbb{R}^n)$ *into* $L^1(\mathbb{R}^n)$.

Now we move on to the bilinear setting. A smooth function $\sigma(x,\xi,\eta)$ defined on $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$ has an associated bilinear pseudodifferential operator T_σ (formally) defined by

$$T_\sigma(f,g)(x) := \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}x\cdot(\xi+\eta)}\sigma(x,\xi,\eta)\,\hat{f}(\xi)\hat{g}(\eta)\,\mathrm{d}\xi\mathrm{d}\eta, \quad x \in \mathbb{R}^n,\ f,g \in \mathscr{S}(\mathbb{R}^n).$$

We say that the *bilinear symbol* $\sigma(x,\xi,\eta)$ belongs to the *bilinear Hörmander class* $BS^m_{\rho,\delta}$ if

$$|\partial_x^\alpha \partial_\xi^\beta \partial_\eta^\gamma \sigma(x,\xi,\eta)| \le C_{\alpha,\beta,\gamma}(1+|\xi|+|\eta|)^{m+\delta|\alpha|-\rho(|\beta|+|\gamma|)}, \quad x,\xi,\eta \in \mathbb{R}^n, \quad (4)$$

for all $\alpha,\beta,\gamma \in \mathbb{N}_0^n$. The study of bilinear pseudo-differential operators grew from the seminal work of Coifman and Meyer [13–15], who used them as models to represent Calderón–Zygmund commutators. Further applications now include the study of compensated compactness (see [16, 17, 44]), and, as bilinear pseudo-differential operators also model expressions of the type $\sum_{\alpha,\beta} c_{\alpha,\beta}\,\partial_x^\alpha f\,\partial_x^\beta g$, they

are useful in generalizing Leibniz's rule in the spirit of the Kato-Ponce inequality, see [3, 36]. Motivated by the study of certain bilinear operators including bilinear pseudo-differential operators and paraproducts with mild regularity, we will next describe how some of the linear results above have a bilinear counterpart and how some others fail to have a natural bilinearization.

A counterpart to Theorem 1 regarding the mapping properties of bilinear pseudo-differential operators is due to Coifman–Meyer [14], Grafakos–Torres [25, 26], and Kenig–Stein [28]. Namely,

Theorem 5. *Given $\sigma \in BS^0_{1,0}$ and $1 < p, q < \infty$ with $1/p + 1/q = 1/r$, the mapping property*

$$T_\sigma : L^p(\mathbb{R}^n) \times L^q(\mathbb{R}^n) \to L^r(\mathbb{R}^n).$$

holds true.

Notice that the Hölder scaling $1/p + 1/q = 1/r$ in Theorem 5 is to be expected as the symbol $\sigma_0 \equiv 1$ belongs to $BS^0_{1,0}$ and T_{σ_0} renders the product of two functions.

In contrast, the natural counterpart to Theorem 3 fails to hold true. Indeed, as Bényi and Torres showed in [4], the class $BS^0_{0,0}$ does not produce the expected mapping behavior in the bilinear setting. Namely,

Theorem 6 (Bényi–Torres, [4]). *There exists a bilinear symbol $\sigma \in BS^0_{0,0}$ such that T_σ does not map $L^p(\mathbb{R}^n) \times L^q(\mathbb{R}^n)$ into $L^r(\mathbb{R}^n)$ for any choice of indices $1 < p, q < \infty$ with $1/p + 1/q = 1/r$.*

On the other hand, a key feature in the theory of linear pseudo-differential operators, namely, the symbolic calculus, does possess a bilinear counterpart. Theorems 7 and 8 , proved in [7], establish the invariance under transposition of the bilinear Hörmander classes. Recall that the two transposes of a bilinear operator T, denoted by T^{*1} and T^{*2}, are defined by the duality relations

$$\langle T(f, g), h\rangle = \langle T^{*1}(h, g), f\rangle = \langle T^{*2}(f, h), g\rangle.$$

Theorem 7 (Invariance under transposition, [7]). *Assume that $m \in \mathbb{R}$, $0 \le \delta \le \rho \le 1$, $\delta < 1$, and $\sigma \in BS^m_{\rho,\delta}$. Then, for $j = 1, 2$, $T^{*j}_\sigma = T_{\sigma^{*j}}$, where $\sigma^{*j} \in BS^m_{\rho,\delta}$.*

For the next result, we write

$$\sigma \sim \sum_{j=0}^{\infty} \sigma_j$$

if there is a non-increasing sequence $m_N \searrow -\infty$ such that

$$\sigma - \sum_{j=0}^{N-1} \sigma_j \in BS^{m_N}_{\rho,\delta},$$

for all $N > 0$.

Theorem 8 (Asymptotic expansion, [7]). *If $m \in \mathbb{R}$, $0 \le \delta < \rho \le 1$ and $\sigma \in BS^m_{\rho,\delta}$, then σ^{*1} and σ^{*2} have the asymptotic expansions*

$$\sigma^{*1} \sim \sum_{\alpha} \frac{i^{|\alpha|}}{\alpha!} \partial_x^\alpha \partial_\xi^\alpha (\sigma(x, -\xi - \eta, \eta))$$

and

$$\sigma^{*2} \sim \sum_{\alpha} \frac{i^{|\alpha|}}{\alpha!} \partial_x^\alpha \partial_\eta^\alpha (\sigma(x, \xi, -\xi - \eta)).$$

More precisely, if $N \in \mathbb{N}$, then

$$\sigma^{*1} - \sum_{|\alpha|<N} \frac{i^{|\alpha|}}{\alpha!} \partial_x^\alpha \partial_\xi^\alpha (\sigma(x, -\xi - \eta, \eta)) \in BS^{m+(\delta-\rho)N}_{\rho,\delta} \tag{5}$$

and

$$\sigma^{*2} - \sum_{|\alpha|<N} \frac{i^{|\alpha|}}{\alpha!} \partial_x^\alpha \partial_\eta^\alpha (\sigma(x, \xi, -\xi - \eta)) \in BS^{m+(\delta-\rho)N}_{\rho,\delta}. \tag{6}$$

In addition, regarding mild regularity of the bilinear symbols, Theorem 4 has the following bilinear counterpart.

Theorem 9 (Coifman–Meyer, [13, p. 55]). *Let ω be a modulus of continuity. If*

$$\int_0^1 \omega^2(t) \, \frac{\mathrm{d}t}{t} < \infty$$

and $\sigma(x, \xi, \eta)$ satisfies

$$|\partial_\xi^\alpha \partial_\eta^\beta \sigma(x, \xi, \eta)| \le C_{\alpha,\beta} (1 + |\xi| + |\eta|)^{-(|\alpha|+|\beta|)}$$

and

$$|\partial_\xi^\alpha \partial_\eta^\beta (\sigma(x + h, \xi, \eta) - \sigma(x, \xi, \eta))| \le C_{\alpha,\beta} \, \frac{\omega(|h|)}{(1 + |\xi| + |\eta|)^{(|\alpha|+|\beta|)}},$$

for all $\alpha, \beta \in \mathbb{N}_0^n$, then T_σ can be extended as a bounded operator from

$$L^p(\mathbb{R}^n) \times L^q(\mathbb{R}^n) \text{ into } L^r(\mathbb{R}^n)$$

for $1 < p, q < \infty$ and $\frac{1}{p} + \frac{1}{q} = \frac{1}{r} \in (0, 1)$.

Various other well-known linear estimates for pseudo-differential operators have natural bilinear analogues (see, for instance, [1–4, 13, 24, 26, 32]). However, another example of a linear result that misses bilinearization is the case of the linear Marcinkiewicz multiplier theorem, whose natural bilinear version fails to hold true, as shown by Grafakos and Kalton in [23].

We will be concerned with an extension of Theorem 9 (see Theorem 10). For $\omega, \Omega : [0,\infty) \to [0,\infty)$, $m \in \mathbb{R}$, and $\rho \in (0,1]$ we write $\sigma \in BS^m_{\rho,\omega,\Omega}$ if

$$|\partial_\xi^\alpha \partial_\eta^\beta \sigma(x,\xi,\eta)| \leq C_{\alpha,\beta}(1+|\xi|+|\eta|)^{m-\rho(|\alpha|+|\beta|)} \text{and} \tag{7}$$

$$|\partial_\xi^\alpha \partial_\eta^\beta (\sigma(x+h,\xi,\eta)-\sigma(x,\xi,\eta))| \leq C_{\alpha,\beta} \frac{\omega(|h|)\Omega(|\xi|+|\eta|)}{(1+|\xi|+|\eta|)^{-m+\rho(|\alpha|+|\beta|)}}, \tag{8}$$

for all $x, \xi, \eta \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{N}_0^n$. Also, for $a > 0$, we write $\omega \in \text{Dini}(a)$ if $\omega : [0,\infty) \to [0,\infty)$, ω is non-decreasing, concave, and

$$|\omega|_{\text{Dini}(a)} := \int_0^1 \omega^a(t)\frac{\mathrm{d}t}{t} < \infty.$$

For $1 < p < \infty$, the discrete variant of Muckenhoupt's A_p class is denoted by $A_p(\mathbb{Z}^n)$ and consists of the positive sequences $\{w_z\}_{z\in\mathbb{Z}^n}$ such that

$$|w|_{A_p(\mathbb{Z}^n)} := \sup_{Q\in\mathbf{Q}} \left(\frac{1}{\#(Q\cap\mathbb{Z}^n)}\sum_{z\in Q\cap\mathbb{Z}^n} w_z\right)\left(\frac{1}{\#(Q\cap\mathbb{Z}^n)}\sum_{z\in Q\cap\mathbb{Z}^n} w_z^{1-p'}\right)^{p-1} < \infty.$$

For $z \in \mathbb{Z}^n$ set $Q_z := \{x \in \mathbb{R}^n : |x_i - z_i| \leq 1/2, i = 1,\dots,n\}$. Consider $1 \leq p, q \leq \infty$ and a positive sequence $\{w_z\}_{z\in\mathbb{Z}^n}$. We denote by l^q_w the space of all sequences $\{a_z\}_{z\in\mathbb{Z}^n}$ such that $\|a\|_{l^q_w} := (\sum_{z\in\mathbb{Z}^n} |a_z|^q w_z)^{1/q} < \infty$. In particular we write l^q instead of l^q_w when $w \equiv 1$. The *weighted amalgam space* (L^p, l^q_w) consists of the locally integrable functions f on $\mathbb{R}^n$ such that $\{\|f\|_{L^p(Q_z)}\}_{z\in\mathbb{Z}^n} \in l^q_w$, with norm

$$\|f\|_{(L^p,l^q_w)} := \left(\sum_{z\in\mathbb{Z}^n} \|f\|^q_{L^p(Q_z)} w_z\right)^{1/q}.$$

The usual interpretation applies when $q = \infty$.

We are now in position to state one of the main results from [33].

Theorem 10 ([33]). *Let $a \in (0,1)$, $\omega \in \text{Dini}(a/2)$, and $\Omega : [0,\infty) \to [0,\infty)$ non-decreasing such that*

$$\sup_{0<t<1} \omega^{1-a}(t)\Omega(1/t) < \infty.$$

Consider $1 \leq p, q \leq \infty$ *and* $\frac{1}{2} \leq r < \infty$ *such that* $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$. *Then, if* $\sigma \in BS^0_{1,\omega,\Omega}$, *with* $|\alpha| + |\beta| \leq 4n + 4$, *the bilinear pseudo-differential operator* T_σ *has the following boundedness properties:*

(i) If $1 < p, q$, *then*

$$\|T_\sigma(f,g)\|_{L^r(\mathbb{R}^n)} \leq C\|f\|_{L^p(\mathbb{R}^n)}\|g\|_{L^q(\mathbb{R}^n)},$$

where $L^p(\mathbb{R}^n)$ *or* $L^q(\mathbb{R}^n)$ *should be replaced by* $L^\infty_c(\mathbb{R}^n)$ *(bounded functions with compact support) if* $p = \infty$ *or* $q = \infty$, *respectively.*

(ii) If $p = 1$ *or* $q = 1$, *then*

$$\|T_\sigma(f,g)\|_{L^{r,\infty}(\mathbb{R}^n)} \leq C\|f\|_{L^p(\mathbb{R}^n)}\|g\|_{L^q(\mathbb{R}^n)},$$

where $L^p(\mathbb{R}^n)$ *or* $L^q(\mathbb{R}^n)$ *should be replaced by* $L^\infty_c(\mathbb{R}^n)$ *if* $p = \infty$ *or* $q = \infty$, *respectively.*

(iii)

$$\|T_\sigma(f,g)\|_{BMO(\mathbb{R}^n)} \leq C\|f\|_{L^\infty(\mathbb{R}^n)}\|g\|_{L^\infty(\mathbb{R}^n)}.$$

(iv) If $1 < p, q < \infty$, *and* $w \in A_{\min(p,q)}$, *then*

$$\|T_\sigma(f,g)\|_{L^r_w(\mathbb{R}^n)} \leq C\|f\|_{L^p_w(\mathbb{R}^n)}\|g\|_{L^q_w(\mathbb{R}^n)},$$

where A_r, $1 \leq r \leq \infty$, *denotes the Muckenhoupt weight class.*

(v) If $w \in A_1$, *the following endpoint estimates hold*

$$\|T_\sigma(f,g)\|_{L^{1/2,\infty}_w(\mathbb{R}^n)} \leq C\|f\|_{L^1_w(\mathbb{R}^n)}\|g\|_{L^1_w(\mathbb{R}^n)}$$

and

$$\|T_\sigma(f,g)\|_{L^{1/2}_w(\mathbb{R}^n)} \leq C\|f\|_{H^1_w(\mathbb{R}^n)}\|g\|_{H^1_w(\mathbb{R}^n)}.$$

(vi) Finally, if $1 < p, q < \infty$, $1 < s_1, s_2 < \infty$, $1/s_3 = 1/s_1 + 1/s_2$, *and* $w \in A_{\min(s_1,s_2)}(\mathbb{Z}^n)$, *then* T_σ *verifies the following inequality on weighted amalgam spaces*

$$\|T_\sigma(f,g)\|_{(L^r,l^{s_3}_w)} \leq C\|f\|_{(L^p,l^{s_1}_w)}\|g\|_{(L^q,l^{s_2}_w)}.$$

Our approach to the proof of Theorem 10, based on a bilinear interpretation of some of Yabuta's ideas in [42, 43], also applies to the study of some molecular paraproducts with mild regularity.

2 Molecular Paraproducts with Mild Regularity

In this section we address some mapping properties of paraproducts built from Dini-continuous molecules. For $\nu \in \mathbb{Z}$ and $k \in \mathbb{Z}^n$, let $P_{\nu k}$ be the dyadic cube

$$P_{\nu k} := \{(x_1, \dots, x_n) \in \mathbb{R}^n : k_i \leq 2^\nu x_i < k_i + 1, i = 1, \dots, n\}. \tag{9}$$

The *lower left corner* of $P = P_{\nu k}$ is $x_P = x_{\nu k} := 2^{-\nu} k$. Notice that the Lebesgue measure of P is $|P| = 2^{-\nu n}$. We set

$$\mathscr{D} := \{P_{\nu k} : \nu \in \mathbb{Z}, k \in \mathbb{Z}^n\}$$

as the collection of all dyadic cubes.

Let $\omega : [0, \infty) \to [0, \infty)$ be a nondecreasing function. Following [33], an *ω-molecule* associated to a dyadic cube $P = P_{\nu k}$ is a function $\phi_P = \phi_{\nu k} : \mathbb{R}^n \to \mathbb{C}$ such that, for some $A > 0$ and $N > n$, it satisfies the *decay* (or *concentration*) condition

$$|\phi_P(x)| \leq \frac{A 2^{\nu n/2}}{(1 + 2^\nu |x - x_P|)^N}, \quad x \in \mathbb{R}^n, \tag{10}$$

and the *mild regularity condition*

$$|\phi_P(x) - \phi_P(y)| \leq A 2^{\nu n/2} \omega(2^\nu |x - y|) \left[\frac{1}{(1 + 2^\nu |x - x_P|)^N} + \frac{1}{(1 + 2^\nu |y - x_P|)^N} \right] \tag{11}$$

for all $x, y \in \mathbb{R}^n$.

For instance, if ϕ is a Dini-continuous function with enough decay, then

$$\phi_{\nu k}(x) := 2^{\nu n/2} \phi(2^\nu x - k) \tag{12}$$

is an ω-molecule associated to P.

Given three families of ω-molecules $\{\phi_Q^j\}_{Q \in \mathscr{D}}$, $j = 1, 2, 3$, the *molecular paraproduct* $\Pi(f, g)$ associated to these families is defined by

$$\Pi(f, g) := \sum_{Q \in \mathscr{D}} |Q|^{-1/2} \langle f, \phi_Q^1 \rangle \langle g, \phi_Q^2 \rangle \phi_Q^3, \quad f, g \in \mathscr{S}(\mathbb{R}^n). \tag{13}$$

The term paraproduct was coined by Bony in [8] and ever since it has been used to denote superpositions of various time-frequency components of two functions. For a brief account on the evolution of the notion of paraproducts, see [6]. Paraproducts have found plenty of inspired applications: from Bony's paradifferential calculus (see [8]) and David–Journé's remarkable $T(1)$-theorem (see [18]), to their alliance with wavelet analysis in the study of PDEs (see, for instance, [10, 11, 23, 39, 40]) and their role as toy models or building blocks of classical operators in Fourier analysis (see, for instance, [21, 22, 30, 31, 35–37, 41]), just to mention a few. The

paraproducts we will work with are built from mildly regular molecules which come to cover the gap between the smooth molecules and paraproducts in [5, 19, 20], and the (discontinuous) Haar molecules and dyadic paraproducts studied, for instance, in [30, 41].

In [5], sufficient conditions on *smooth* molecules were established so that smooth paraproducts of the form (13) can be realized as bilinear Calderón–Zygmund operators. In this chapter we revisit the analysis of ω-molecules introduced in [33] and mention how the paraproducts they build can be realized as bilinear Calderón–Zygmund operators of type $\omega(t)$ (as defined in Sect. 3), provided that they have enough decay, suitable cancelation, and $\omega \in \text{Dini}(1/2)$.

Concerning Dini-continuous molecules, one of the main results in [33] reads:

Theorem 11 ([33]). *Consider $\omega \in \text{Dini}(1/2)$ and let $\{\phi_Q^j\}_{Q\in\mathscr{D}}$, $j = 1, 2, 3$ be three families of ω-molecules with decay $N > 10n$ such that at least two of them, say $j = 1, 2$, enjoy the cancelation property*

$$\int_{\mathbb{R}^n} \phi_Q^j(x)\,dx = 0, \quad Q \in \mathscr{D}, j = 1, 2.$$

Then, the paraproduct $\Pi(f, g)$ defined in (13) verifies the inequalities (i)–(vi) as in Theorem 10.

3 Bilinear Calderón–Zygmund Operators of Type $\omega(t)$

In [42, 43], Yabuta developed the notion of linear Calderón–Zygmund operator of type $\omega(t)$ (which includes the classical Calderón–Zygmund operators). In [33], a bilinear program inspired by Yabuta's work was carried out as follows. Let $\omega : [0, \infty) \to [0, \infty)$ be a nondecreasing function. We say that $K(x, y, z)$ defined on $\mathbb{R}^{3n} \setminus \{(x, y, z) \in \mathbb{R}^{3n} : x = y = z\}$ is a *bilinear Calderón–Zygmund kernel of type* $\omega(t)$ if for some constants $0 < \tau < 1$ (the specific value of $\tau \in (0, 1)$ is immaterial in the development of the theory), $C_K > 0$, and every $(x, y, z) \in \mathbb{R}^{3n} \setminus \{(x, y, z) \in \mathbb{R}^{3n} : x = y = z\}$ it holds

$$|K(x, y, z)| \leq \frac{C_K}{(|x - y| + |x - z|)^{2n}} \tag{14}$$

and

$$\begin{aligned} &|K(x + h, y, z) - K(x, y, z)| + |K(x, y + h, z) - K(x, y, z)| \\ &\qquad + |K(x, y, z + h) - K(x, y, z)| \\ &\qquad \leq \frac{C_K}{(|x-y|+|x-z|)^{2n}}\, \omega\left(\frac{|h|}{|x-y|+|x-z|}\right), \end{aligned} \tag{15}$$

whenever $|h| \leq \tau \max(|x-y|, |x-z|)$. A bilinear operator $T : \mathscr{S}(\mathbb{R}^n) \times \mathscr{S}(\mathbb{R}^n) \to \mathscr{S}(\mathbb{R}^n)'$ is said to be associated to a bilinear Calderón–Zygmund kernel of type $\omega(t)$, $K(x, y, z)$, if

$$T(f, g)(x) = \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} K(x, y, z) f(y) g(z) \, \mathrm{d}y \mathrm{d}z$$

whenever $x \notin \operatorname{supp}(f) \cap \operatorname{supp}(g)$ and $f, g \in C_0^\infty(\mathbb{R}^n)$. If, in addition, T maps

$$L^p(\mathbb{R}^n) \times L^q(\mathbb{R}^n) \to L^{r,\infty}(\mathbb{R}^n),$$

for some $1 < p, q < \infty$ and $r > 1$ with $1/p + 1/q = 1/r$, or

$$L^p(\mathbb{R}^n) \times L^q(\mathbb{R}^n) \to L^1(\mathbb{R}^n),$$

for some $1 < p, q < \infty$ with $1/p + 1/q = 1$, T is called a *bilinear Calderón–Zygmund operator of type* $\omega(t)$.

The multilinear Calderón–Zygmund theory, which corresponds to the case $\omega(t) = t^\epsilon$ for some $\epsilon \in (0, 1]$, was introduced by Coifman and Meyer in [13–15]. This theory was then further investigated by Grafakos and Torres [25, 26] and Kenig and Stein [28].

Next, we list the main theorems in [33].

Theorem 12 ([33]). *Consider $\omega \in \mathrm{Dini}(1/2)$ and let T be a bilinear operator associated to a bilinear Calderón–Zygmund kernel of type $\omega(t)$, $K(x, y, z)$. Assume that for some $1 \leq p, q \leq \infty$ and $0 < r < \infty$ satisfying*

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{r},$$

T maps $L^p(\mathbb{R}^n) \times L^q(\mathbb{R}^n)$ into $L^{r,\infty}(\mathbb{R}^n)$. Then, T can be extended to a bounded operator from $L^1(\mathbb{R}^n) \times L^1(\mathbb{R}^n)$ into $L^{\frac{1}{2},\infty}(\mathbb{R}^n)$.

By means of duality arguments, Theorem 12 implies.

Theorem 13 ([33]). *Consider $\omega \in \mathrm{Dini}(1/2)$ and T be a bilinear Calderón–Zygmund operator of type $\omega(t)$ in $\mathbb{R}^n$ with kernel K. Let $1 \leq p, q \leq \infty$, $\frac{1}{2} \leq r < \infty$ such that $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$. Then we have*

(i) If $p, q > 1$, then T can be extended to a bounded operator from $L^p(\mathbb{R}^n) \times L^q(\mathbb{R}^n)$ into $L^r(\mathbb{R}^n)$, where $L^p(\mathbb{R}^n)$ or $L^q(\mathbb{R}^n)$ should be replaced by $L_c^\infty(\mathbb{R}^n)$ if $p = \infty$ or $q = \infty$, respectively.

(ii) If $p = 1$ or $q = 1$, then T can be extended to a bounded operator from $L^p(\mathbb{R}^n) \times L^q(\mathbb{R}^n)$ into $L^{r,\infty}(\mathbb{R}^n)$, where $L^p(\mathbb{R}^n)$ or $L^q(\mathbb{R}^n)$ should be replaced by $L_c^\infty(\mathbb{R}^n)$ if $p = \infty$ or $q = \infty$, respectively.

(iii) T can be extended to a bounded operator from $L_c^\infty(\mathbb{R}^n) \times L_c^\infty(\mathbb{R}^n)$ into BMO.

Theorem 14 ([33]). *Let $1 < p, q < \infty$, $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$, and $w \in A_\infty$. Consider $\omega \in \mathrm{Dini}(1/2)$ and let T be a bilinear Calderón–Zygmund operator of type $\omega(t)$ in $\mathbb{R}^n$ with kernel K. Let C_K be the constant in (14) and (15), and let W denote the norm of T as a bounded operator from $L^1(\mathbb{R}^n) \times L^1(\mathbb{R}^n)$ into $L^{\frac{1}{2},\infty}(\mathbb{R}^n)$ (see Theorem 12). Then for f and g bounded and compactly supported,*

$$\|T(f,g)\|_{L^r_w(\mathbb{R}^n)} \le C_{p,n}(C_K + W)\|\mathscr{M} f\|_{L^p_w(\mathbb{R}^n)} \|\mathscr{M} g\|_{L^q_w(\mathbb{R}^n)}, \tag{16}$$

where $\mathscr{M}$ stands for the Hardy–Littlewood maximal operator. In particular, if $w \in A_{\min(p,q)}$, we have

$$\|T(f,g)\|_{L^r_w(\mathbb{R}^n)} \le C_{p,n}(C_K + W)\|f\|_{L^p_w(\mathbb{R}^n)} \|g\|_{L^q_w(\mathbb{R}^n)}, \tag{17}$$

and therefore, T extends as a bounded operator from $L^p_w(\mathbb{R}^n) \times L^q_w(\mathbb{R}^n)$ into $L^r_w(\mathbb{R}^n)$. Weighted endpoint estimates and weighted H^1 estimates also hold true; see Remark 5 and Theorem 6.9 in [33].

Theorem 15 ([33]). *Consider $\omega \in \mathrm{Dini}(1/2)$ and let T be a bilinear Calderón–Zygmund operator of type $\omega(t)$ with kernel K. If $1 < p, q < \infty$, $1 < s_1, s_2 < \infty$, $1/r = 1/p + 1/q$, $1/s_3 = 1/s_1 + 1/s_2$, and $w \in A_s(\mathbb{Z}^n)$, $s = \min\{s_1, s_2\}$, then*

$$\|T(f,g)\|_{(L^r, l^{s_3}_w)} \le C\|f\|_{(L^p, l^{s_1}_w)}\|g\|_{(L^q, l^{s_2}_w)}. \tag{18}$$

4 The Proofs of Theorems 10 and 11

In this section we come back to the bilinear pseudo-differential operator

$$T_\sigma(f,g)(x) = \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \sigma(x,\xi,\eta)\mathrm{e}^{\mathrm{i}x(\xi+\eta)}\hat{f}(\xi)\hat{g}(\eta)\,\mathrm{d}\xi\,\mathrm{d}\eta, \quad f,\ g \in \mathscr{S}(\mathbb{R}^n),$$

whose symbol $\sigma(x,\xi,\eta)$ satisfies the conditions

$$|\partial^\alpha_\xi \partial^\beta_\eta \sigma(x,\xi,\eta)| \le \frac{C_{\alpha,\beta}}{(1+|\xi|+|\eta|)^{|\alpha|+|\beta|}}, \tag{19}$$

$$|\partial^\alpha_\xi \partial^\beta_\eta (\sigma(x+h,\xi,\eta) - \sigma(x,\xi,\eta))| \le C_{\alpha,\beta}\ \omega(|h|)\ \frac{\Omega(|\xi|+|\eta|)}{(1+|\xi|+|\eta|)^{|\alpha|+|\beta|}}, \tag{20}$$

for all $x,\ \xi, \eta \in \mathbb{R}^n$ and multi-indices $\alpha,\ \beta \in \mathbb{N}^n_0$.

Theorems 16–18 establish sufficient conditions on ω and Ω so that the class $BS^0_{1,\omega,\Omega}$ produces pseudo-differential operators with bilinear Calderón–Zygmund

kernels of type $\omega^a(t)$ for some $a \in (0,1)$, the class $BS^0_{1,\omega,\Omega}$ produces pseudo-differential operators with bilinear Calderón–Zygmund operators of type $\omega^a(t)$ for some $a \in (0,1)$, and so that the paraproducts based on ω-molecules can be realized as bilinear Calderón–Zygmund kernels of type $\theta(t)$, for some appropriate $\theta(t)$.

Theorem 16 ([33]). *Let $\omega, \Omega : [0,\infty) \to [0,\infty)$ be nondecreasing functions with ω concave. Suppose that there exists $a \in (0,1)$ such that ω and Ω verify*

$$\sup_{0<t<1} \omega^{1-a}(t)\Omega(1/t) < \infty. \tag{21}$$

If $\sigma(x,\xi,\eta)$ verifies (19) and (20) with $|\alpha| + |\beta| \le 2n+2$, then T_σ has a bilinear Calderón–Zygmund kernel of type $\omega^a(t)$.

Theorem 17 ([33]). *Let $\Omega : [0,\infty) \to [0,\infty)$ be a nondecreasing function, $a \in (0,1)$, and $\omega \in \mathrm{Dini}(a/2)$ such that (21) holds. If $\sigma(x,\xi,\eta)$ verifies (19) and (20) with $|\alpha| + |\beta| \le 4n+4$, then T_σ is a bilinear Calderón–Zygmund operator of type $\omega^a(t)$.*

Theorem 18 ([33]). *Assume $\omega \in \mathrm{Dini}(1/2)$ and let $\{\phi_Q^j\}_{Q\in\mathscr{D}}$, $j = 1,2,3$ be three families of ω-molecules with decay $N > 10n$ and such that at least two of them have cancelation. Then, the paraproduct Π defined in (13) has a bilinear Calderón–Zygmund kernel of type $\theta(t)$ with*

$$\theta(t) := A^3 A_N \omega(C_N t), \quad t > 0,$$

for some positive constants A_N and C_N (hence, $\theta \in \mathrm{Dini}(1/2)$). Here A is as in (10) and (11). Moreover, Π has the mapping property

$$\Pi : L^2(\mathbb{R}^n) \times L^2(\mathbb{R}^n) \to L^1(\mathbb{R}^n).$$

In particular, Π is a bilinear Calderón–Zygmund operator of type $\theta(t)$.

The proofs of Theorems 10 and 11 now follow from the realization as bilinear Calderón–Zygmund operators of type $\omega(t)$ of the pseudo-differential operators and molecular paraproducts in Theorems 16, 17, and 18 and the boundedness properties in Theorems 12–15.

Acknowledgment Author supported by the NSF under grant DMS 0901587.

References

1. Bényi, Á.: Bilinear singular integrals and pseudodifferential operators. Ph.D. Thesis, University of Kansas (2002)
2. Bényi, Á.: Bilinear pseudodifferential operators on Lipschitz and Besov spaces. J. Math. Anal. Appl. **284**, 97–103 (2003)

3. Bényi, Á., Torres, R.H.: Symbolic calculus and the transposes of bilinear pseudodifferential operators. Comm. P.D.E. **28**, 1161–1181 (2003)
4. Bényi, Á., Torres, R.H.: Almost orthogonality and a class of bounded bilinear pseudodifferential operators. Math. Res. Lett. **11.1**, 1–12 (2004)
5. Bényi, Á., Maldonado, D., Nahmod, A., Torres, R.H.: Bilinear paraproducts revisited. Math. Nachr. **283**(9), 1257–1276 (2010)
6. Bényi, Á. Maldonado, D., Naibo, V.: What is a... paraproduct? Notices Amer. Math. Soc. **57**(07), 858–860 (2010)
7. Bényi, Á., Maldonado, D., Naibo, V., Torres, R.H.: On the Hörmander classes of bilinear pseudodifferential operators. Integr. Equat. Oper. Theor. **67**(3), 341–364 (2010)
8. Bony, J.M.: Calcul symbolique et propagation des singularités pour les équations aux dérivées partielles non-linéaires. Annales Scientifiques de l'École Normale Supérieure Sér. 4, **14**(2), 209–246 (1981)
9. Calderón, A., Vaillancourt, R.: A class of bounded pseudo-differential operators. Proc. Nat. Acad. Sci. USA **69**, 1185–1187 (1972)
10. Cannone, M.: Ondelettes, paraproduits et Navier-Stokes. (French) [Wavelets, paraproducts and Navier-Stokes]. Diderot Editeur, Paris (1995)
11. Cannone, M.: Harmonic analysis tools for solving the incompressible Navier-Stokes equations. Handbook of mathematical fluid dynamics, vol. III, pp. 161–244. North-Holland, Amsterdam (2004)
12. Ching, C.-H.: Pseudo-differential operators with non-regular symbols. J. Diff. Eqns. **11**, 436–447 (1972)
13. Coifman, R.R., Meyer, Y.: Au-delà des Opérateurs Pseudo-différentiels, 2nd edn. Astèrisque 57 (1978)
14. Coifman, R.R., Meyer, Y.: Wavelets: Calderón-Zygmund and Multilinear Operators. Cambridge University Press, Cambridge (1997)
15. Coifman, R.R., Meyer, Y.: Commutateurs d'intégrales singulières et opérateurs multilinéaires. Ann. l'institut Fourier, **28**(3), 177–202 (1978)
16. Coifman, R.R., Lions, P.-L., Meyer, Y., Semmes, S.: Compensated compactness and Hardy spaces. J. Math. Pures Appl. **72**, 247–286 (1993)
17. Coifman, R.R., Dobyinsky, S., Meyer, Y.: Opérateurs bilinéaires et renormalization. In: Fefferman, C., Fefferman, R., Wainger, S.: Essays on Fourier Analysis in Honor of Elias M. Stein. Princeton University Press, Princeton (1995)
18. David, G., Journé, J.-L.: A boundedness criterion for generalized Calderón–Zygmund operators. Ann. Math. **120**, 371–397 (1984)
19. Frazier, M., Jawerth, B.: A discrete transform and decompositions of distribution spaces. J. Func. Anal. **93**, 34–169 (1990)
20. Frazier, M., Jawerth, B., Weiss, G.: Littlewood–Paley theory and the study of function spaces. CBMS Regional Conference Series in Mathematics, vol. 79 (1991)
21. Gilbert, J., Nahmod, A.: Bilinear operators with non-smooth symbols. I. J. Fourier Anal. Appl. **5**, 435–467 (2001)
22. Gilbert, J., Nahmod, A.: L^p-boundedness of time-frequency paraproducts. II. J. Fourier Anal. Appl. **8**, 109–172 (2002)
23. Grafakos, L., Kalton, N.: The Marcinkiewicz multiplier condition for bilinear operators. Studia Math. **146**(2), 115–156 (2001)
24. Grafakos, L., Torres, R.H.: Discrete decompositions for bilinear operators and almost diagonal conditions. Trans. Amer. Math. Soc. **354**, 1153–1176 (2002)
25. Grafakos, L., Torres, R.H.: Maximal operator and weighted norm inequalities for multilinear singular integrals. Indiana Univ. Math. J. **51**(5), 1261–1276 (2002)
26. Grafakos, L., Torres, R.H.: Multilinear Calderón–Zygmund theory. Adv. Math. **165**, 124–164 (2002)
27. Hörmander, L.: Pseudo-differential operators and hypoelliptic equations. Proceedings of Symposium in Pure Mathematics, vol. X, pp. 138–183. American Mathematical Society, Providence (1967)

28. Kenig, C., Stein, E.: Multilinear estimates and fractional integration. Math. Res. Lett. **6**, 1–15 (1999). Erratum in Math. Res. Lett. **6**(3–4), 467 (1999)
29. Kohn, J., Nirenberg, L.: An algebra of pseudo-differential operators. Comm. Pure Appl. Math. **18**, 269–305 (1965)
30. Lacey, M.: Commutators with Riesz potentials in one and several parameters. Hokkaido Math. J. **36**, 175–191 (2007)
31. Lacey, M., Metcalfe, J.: Paraproducts in one and several parameters. Forum Math. **19**, 325–351 (2007)
32. Maldonado, D., Naibo, V.: On the boundedness of bilinear operators on products of Besov and Lebesgue spaces. J. Math. Anal. Appl. **352**, 591–603 (2009)
33. Maldonado, D., Naibo, V.: Weighted norm inequalities for paraproducts and bilinear pseudodifferential operators with mild regularity. J. Fourier Anal. Appl. **15**(2), 218–261 (2009)
34. Mikhlin, S.: On the multipliers of Fourier integrals. Doklady Akademii Nauk SSSR, n. Ser., **109**, 701–703, (1956) (in Russian)
35. Muscalu, C., Tao, T., Thiele, C.: Multilinear operators given by singular multipliers. J. Amer. Math. Soc. **15**, 469–496 (2002)
36. Muscalu, C., Pipher, J., Tao, T., Thiele, C.: Bi-parameter paraproducts. Acta Math. **193**, 269–296 (2004)
37. Petermichl, S.: Dyadic shifts and a logarithmic estimate for Hankel operators with matrix symbols. C. R. Acad. Sci. Paris Sèr. I Math. **330**, 455–460 (2000)
38. Stein, E.M.: Harmonic Analysis: Real Variable Methods, Orthogonality, and Oscillatory Integrals. Princeton University Press, Princeton (1993)
39. Taylor, M.: Pseudodifferential operators and nonlinear PDE. Progress in Mathematics, vol. 100. Birkhäuser, Boston, Inc., Boston, MA (1991)
40. Taylor, M.: Tools for PDE. Pseudodifferential operators, paradifferential operators, and layer potentials. Mathematical Surveys and Monographs, vol. 81. American Mathematical Society, Providence (2000)
41. Thiele, C.: Wave packet analysis. CBMS Regional Conference Series in Mathematics, vol. 105. American Mathematical Society (2006)
42. Yabuta, K.: Generalizations of Calderón–Zygmund operators. Studia Math. **82**(1), 17–31 (1985)
43. Yabuta, K.: Calderón–Zygmund operators and pseudodifferential operators. Comm. P.D.E. **10**(9), 1005–1022 (1985)
44. Youssif, A.: Bilinear operators and the Jacobian-determinant on Besov spaces. Indiana Univ. Math. J. **45**, 381–396 (1996)

Weighted Inequalities and Dyadic Harmonic Analysis

María Cristina Pereyra

Abstract We survey the recent solution of the so-called A_2 conjecture, that states: all Calderón–Zygmund singular integral operators are bounded on $L^2(w)$ with a bound that depends linearly on the A_2 characteristic of the weight w. We also survey corresponding results for commutators. We highlight the interplay of dyadic harmonic analysis in the solution of the A_2 conjecture, especially Hytönen's representation theorem for Calderón–Zygmund singular integral operators in terms of Haar shift operators. We describe Chung's dyadic proof of the corresponding quadratic bound on $L^2(w)$ for the commutator of the Hilbert transform with a BMO function, and we deduce *sharpness* of the bounds for the dyadic paraproduct on $L^p(w)$ that were obtained extrapolating Beznosova's linear bound on $L^2(w)$. We show that if an operator T is bounded on the weighted Lebesgue space $L^r(w)$ and its operator norm is bounded by a power α of the A_r characteristic of the weight, then its commutator $[T, b]$ with a function b in BMO will be bounded on $L^r(w)$ with an operator norm bounded by the increased power $\alpha + \max\{1, \frac{1}{r-1}\}$ of the A_r characteristic of the weight. The results are sharp in terms of the growth of the operator norm with respect to the A_r characteristic of the weight for all $1 < r < \infty$.

Keywords Dyadic harmonic analysis • Weighted inequalities • A_2 conjecture • Haar shift operators • Commutator • Dyadic paraproduct • A_p weights • Random dyadic grids • Haar basis • Hilbert transform • Fourier multiplier

1 Introduction

The main problem of study in this chapter is the *weighted L^p inequalities*

M.C. Pereyra (✉)
Department of Mathematics and Statistics, University of New Mexico, MSC03 21501, Albuquerque, NM 87131-0001, USA
e-mail: crisp@math.unm.edu

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_15,

$$\|Tf\|_{L^p(v)} \le C(u,v)\,\|f\|_{L^p(u)}, \tag{1}$$

where $f \in L^p(u)$ iff $\|f\|_{L^p(u)} := (\int |f(x)|^p u(x)\,dx)^{1/p} < \infty$, and u, v are locally integrable positive a.e. functions defined on $\mathbb{R}^n$. When $u \equiv 1$ we denote $\|f\|_p := \|f\|_{L^p(u)}$.

Two-weight problem: *Find necessary and sufficient conditions on the weights so that above inequality holds for a given operator or class of operators T, and find the optimal rate of dependence of the constant $C(w)$ on the weight.*

In this survey we will concentrate on *one-weight inequalities*, $u = v$, for Calderón–Zygmund singular integral operators, more specifically for the Hilbert transform $T = H$ and for the commutator of the Hilbert transform with a function b in the space BMO of bounded mean oscillation, namely $T = [b, H] := bH - Hb$.

The Hilbert transform is bounded on $L^p(w)$ if and only if the weight w belongs to the Muckenhoupt A_p class [31]. This is also true for Calderón–Zygmund singular integral operators [11]. A weight w is in the *Muckenhoupt A_p class* if

$$[w]_{A_p} := \sup_I \left(\frac{1}{|I|}\int_I w\right)\left(\frac{1}{|I|}\int_I w^{-1/(p-1)}\right)^{p-1} < \infty. \tag{2}$$

In the last decade there has been a flurry of activity trying to identify the exact dependence of the operator bound on the A_p characteristic, $[w]_{A_p}$, of the weight. This dependence was first proved to be linear in A_2 for a few dyadic operators [30, 79, 80], then for the Beurling–Ahlfors [70], Hilbert [67], and Riesz transforms [68], and for the dyadic paraproduct [4]. Finally Tuomas Hytönen solved in the positive the A_2 conjecture [33]: If T is a Calderón–Zygmund singular integral operator, $w \in A_2$, then the dependence on the A_2 characteristic of the weight is linear, that is,

$$\|Tf\|_{L^2(w)} \le C\,[w]_{A_2}\,\|f\|_{L^2(w)}. \tag{3}$$

Sharp extrapolation [20] then yields the correct L^p bounds for the class of Calderón–Zygmund singular integral operators:

$$\|Tf\|_{L^p(w)} \le C_p [w]_{A_p}^{\max\{1,\frac{1}{p-1}\}}\,\|f\|_{L^p(w)}.$$

Remark. The long-standing two-weight problem for the Hilbert transform "à la Muckenhoupt" is an outstanding open problem: Characterize the pairs of weights (u, v), in terms of conditions like the A_p condition in the one-weight problem, for which (1) holds. Recently there has been progress due to Lacey, Sawyer, Shen, and Uriarte-Tuero [45]. Note that Cotlar and Sadosky solved, years ago, the two-weight problem "à la Helson-Szëgo," that is, using complex analysis techniques [13, 14].

In this chapter we want to highlight the interplay with dyadic harmonic analysis [60] in the solution of the A_2 conjecture. Initially the A_2 conjecture was shown to hold, one at a time, for dyadic operators and for operators such as the Hilbert trans-

form that have lots of symmetries. Stephanie Petermichl showed, in groundbreaking work in 2000, that the Hilbert transform can be written as an appropriate average of dyadic shift operators [32,66], and later she showed, in a tour de force using Bellman function techniques, that for the dyadic shift operators the A_2 conjecture is true and therefore also for the Hilbert transform [67]. This work represented a quantum jump in our understanding of singular integral operators. Until then a simpler dyadic model, the *martingale transform*, was considered the toy model for singular integrals. One would first try to prove results for this model and then hope to prove them for a genuine singular integral operator, but the transition was by no means automatic [60]. Petermichl's representation theorem made this transition trivial for the Hilbert transform. For a while it seemed that the miracle of this representation theorem was a consequence of the symmetries of the operator. Similar constructions were found for other symmetric operators: the Riesz transform (n-dimensional analogue of the Hilbert transform) [68], the Beurling–Ahlfors transform [70], and for sufficiently smooth convolution Calderón–Zygmund singular integral operators [76]. The fact that for the Beurling–Ahlfors transform the A_2 conjecture holds for $p > 2$ (linear estimate in A_p characteristic in the range of $p > 2$) had important implications in the theory of quasiconformal mappings [2].

All these operators have a representation as averages of dyadic Haar shift operators of bounded complexity. In 2008, Oleksandra Beznosova showed that the linear bound on $L^2(w)$ also holds for the dyadic paraproduct, an operator not in the above class [4]. Hytönen was able to prove a representation theorem valid for *all* Calderón–Zygmund singular integral operators (not only convolution) in terms of dyadic Haar shift operators of arbitrary complexity, paraproducts, and adjoints of the paraproducts. Different groups of researchers had already shown that the A_2 conjecture was true for all these Haar shift operators [17, 18, 44], using techniques other than Bellman function which had dominated the scene until then. However, the dependence of the operator bound on the complexity was exponential and prevented one from deducing the A_2 conjecture for general Calderón–Zygmund singular operators. Only for those operators that were averages of dyadic shift operators of bounded complexity one could deduce the A_2 conjecture. Hytönen was able to overcome this obstacle as well, proving a polynomial dependence on the complexity and the linear dependence on the A_2 characteristic of the weight for Haar shift operators, therefore proving the A_2 conjecture [33]. Precursors to Petermichl's and Hytönen's results can be found in Figiel's work [24]. Nowadays some of the simpler arguments yielding polynomial and even linear dependence on the complexity use minimally Bellman functions [54, 74], or do not use them at all [37, 42].

The commutator $[b, H]$ is more singular than the operator H, and this is reflected on the nature of its bounds on weighted L^p spaces. Daewon Chung showed in [8] that

$$\|[b, H]f\|_{L^2(w)} \leq C\,[w]_{A_2}^2 \|f\|_{L^2(w)}. \tag{4}$$

That is, the dependence on the A_2 characteristic of the operator bound is now *quadratic* as opposed to the *linear* bound enjoyed by the Hilbert transform. Chung's proof can be labeled as a dyadic proof. It suffices to consider the commutator

with Petermichl's Haar shift operator [69]. Then known linear bounds for the shift operator [67] and for the dyadic paraproduct [4] can be used, and Bellman function arguments can be invoked as did all of Chung's predecessors until then. We observe that the sharp bounds for the commutator of the Hilbert transform imply that Beznosova's bounds [4] are the sharp bounds for the dyadic paraproduct in $L^p(w)$, which was not known until now. The author in collaboration with Chung and Carlos Pérez established a transference theorem that states that if a linear operator T obeys a linear bound on $L^2(w)$ then its commutator with a BMO function obeys a quadratic bound [10]. In light of Hytönen's theorem this means that *all* commutators of Calderón–Zygmund singular operators with BMO functions obey a quadratic bound as in inequality (4). The argument follows the classical Coifman, Rochberg, and Weiss argument [12] exploiting the Cauchy integral formula and some very precise quantitative results in the theory of A_2 weights and BMO functions. Generalizations of these results to commutators with fractional integrals and to the two-weight setting appear in [16], and weak-type estimates and strong estimates involving instead the A_1 characteristic of the weight appear in [59]. In this note we present the simple modifications necessary to state a transference theorem that provides bounds on $L^r(w)$, $r \neq 2$, for the commutator given corresponding bounds on $L^r(w)$ for the initial operator.

The author strongly believes that Petermichl and Hytönen's representation theorem in terms of dyadic operators could have important consequences in applications, in the same way that the $T(1)$ theorem [19] had repercussions in computational harmonic analysis via the Beylkin, Coifman, and Rokhlin algorithm to decompose singular integral operators [3].

This chapter is organized as follows. In Sect. 2 we define the Hilbert transform and the dyadic Haar shift operators, recall some of their basic properties, state Petermichl's representation theorem, and show how it provides a straightforward proof of the boundedness of the Hilbert transform on $L^p(\mathbb{R})$ (Riesz's theorem). In Sect. 3 we discuss weighted inequalities for the Hilbert transform and recount the prehistory of linear estimates for dyadic operators on $L^2(w)$. We state the sharp extrapolation theorem and deduce $L^p(w)$ bounds from linear bounds and observe that these bounds are sometimes sharp, but not always, as Buckley's estimates for the maximal function show. We then define the Haar shift operators of complexity (m,n), discuss their boundedness properties, and state Hytönen's theorem (the A_2 conjecture), as well as his representation theorem. In Sect. 4 we define the commutator, state its boundedness properties, and sketch Chung's dyadic proof of the quadratic estimate on $L^2(w)$. We note that this quadratic estimate is sharp, and we show that Chung's dyadic method of proof implies that Beznosova's bound for the dyadic paraproduct is sharp as well. Finally we state a variation of the transference theorem for commutators on $L^r(w)$ with $r \neq 2$ and present its proof in the Appendix.

2 Hilbert Transform Versus Dyadic Shift Operators

We define the Hilbert transform both on Fourier and space domains, and we describe its boundedness and symmetry properties. We introduce the dyadic intervals, the random dyadic grids, and corresponding Haar bases, and we emphasize some of the properties these bases share with wavelets such us being an unconditional basis on L^p spaces. We define Petermichl's Haar shift operators and describe their symmetry properties; we state Petermichl's representation theorem and show how it provides a straightforward proof of the boundedness of the Hilbert transform on $L^p(\mathbb{R})$.

2.1 Hilbert Transform

In this section we recall the definition of the Hilbert transform on Fourier domain as a Fourier multiplier and on space domain as a convolution with a singular kernel. We also recall how symmetry properties completely characterize the Hilbert transform. These are well-known facts that can be found in any Fourier analysis book such as [21, 26, 73]. You will also find here the definition of BMO, the space of functions of bounded mean oscillation.

2.1.1 Fourier Multiplier

The *Fourier transform* of a Schwartz function is defined by

$$\widehat{f}(\xi) := \int_{\mathbb{R}} f(x)\, \mathrm{e}^{-2\pi \mathrm{i}\xi x}\, \mathrm{d}x.$$

With some work one can define the Fourier transform on $L^2(\mathbb{R})$ and show that it is an isometry, that is, $\|\widehat{f}\,\|_2 = \|f\|_2$ (Plancherel's identity).

On Fourier side the *Hilbert transform* can be defined as a *Fourier multiplier*:

$$\widehat{Hf}(\xi) = -i\, \mathrm{sgn}(\xi)\, \widehat{f}(\xi), \tag{5}$$

where $\mathrm{sgn}(\xi) = 1$ if $\xi > 0$, $\mathrm{sgn}(\xi) = -1$ if $\xi < 0$, and is zero at $\xi = 0$.

The absolute value of the symbol $m_H(\xi) := -i\, \mathrm{sgn}(\xi)$ is 1 a.e., and Plancherel's identity used twice implies that $H : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ and that it is an isometry:

$$\|Hf\|_2 = \|\widehat{Hf}\,\|_2 = \|\widehat{f}\,\|_2 = \|f\|_2.$$

2.1.2 Singular Integral Operator

Since the Hilbert transform is given on Fourier side by

$$\widehat{Hf}(\xi) = m_H(\xi)\,\widehat{f}(\xi),$$

multiplication on Fourier side comes from convolution on space with the distributional kernel k_H which is the inverse Fourier transform of the multiplier m_H. A calculation yields

$$k_H(x) := (m_H)^\vee(x) = \text{p.v.}\frac{1}{\pi x}.$$

For a distributional kernel, the integration must be done in the principal value sense:

$$Hf(x) = k_H * f(x) = \text{p.v.}\frac{1}{\pi}\int \frac{f(x-y)}{y}\,\mathrm{d}y := \lim_{\epsilon\to 0}\frac{1}{\pi}\int_{|x-y|>\epsilon}\frac{f(y)}{x-y}\,\mathrm{d}y. \quad (6)$$

Had the kernel k_H been integrable, boundedness on $L^p(\mathbb{R})$ would be a consequence of the Hausdorff–Young's inequality for $p \geq 1$: if $g \in L^1(\mathbb{R})$, $f \in L^p(\mathbb{R})$, then $\|g * f\|_p \leq \|g\|_1 \|f\|_p$. But k_H is not in $L^1(\mathbb{R})$; despite this fact, H is bounded on $L^p(\mathbb{R})$ for all $1 < p < \infty$, as Marcel Riesz proved in 1927:

$$\|Hf\|_p \leq C_p \|f\|_p.$$

However, H is not bounded on $L^1(\mathbb{R})$ nor on $L^\infty(\mathbb{R})$, but there are appropriate substitutes: H is of weak type (1,1) and is bounded on BMO [21, 26, 73]. Recall that a function $b : \mathbb{R} \to \mathbb{R}$ belongs to BMO, the *space of bounded mean oscillation*, if and only if

$$\|b\|_{BMO} := \sup_I \frac{1}{|I|}\int_I |b(x) - m_I b|\,\mathrm{d}x < \infty, \quad (7)$$

where $m_I b$ denotes the integral average of b on the interval I, $m_I b = \frac{1}{|I|}\int_I b(x)\,\mathrm{d}x$. This space was introduced by John and Nirenberg in the 1960s [39]. The space of bounded functions $L^\infty(\mathbb{R})$ is a proper subset of BMO; the canonical example of a function that is not bounded but it is in BMO is $\log|x|$ [26].

2.1.3 Symmetries

The Hilbert transform commutes with translations and dilations and anticommutes with reflections, and it is essentially the only bounded linear operator in $L^2(\mathbb{R})$ that has those properties. In what follows $h \in \mathbb{R}$ and $\delta > 0$.

- Convolution $\Leftrightarrow$ H commutes with *translations* $\tau_h f(x) := f(x-h)$

$$\tau_h(Hf) = H(\tau_h f).$$

- Homogeneity of kernel $\Leftrightarrow$ H commutes with *dilations* $D_\delta f(x) = f(\delta x)$

$$D_\delta(Hf) = H(D_\delta f).$$

- Kernel odd $\Leftrightarrow$ H anticommutes with *reflections* $\tilde{f}(x) := f(-x)$

$$(H\tilde{f}) = -H(\tilde{f}).$$

Theorem 1 ([26, 73]). *Let T be a linear and bounded operator in $L^2(\mathbb{R})$ that commutes with translations and dilations and anticommutes with reflections, then T* must be *a constant multiple of the Hilbert transform: $T = cH$.*

Using this principle, Petermichl [66] showed that we can write H as a suitable "average of dyadic operators"; see also [32].

2.2 Dyadic Shift Operators

We first introduce the dyadic intervals and associated Haar basis, as well as random dyadic grids. We recall some important properties of the Haar basis shared with wavelet bases such us being an unconditional system in L^p spaces and weighted $L^p(w)$ whenever $w \in A_p$. We then describe Petermichl's averaging theorem and give some intuition why this should work. We deduce Riesz's theorem from this representation, that is, the boundedness on $L^p(\mathbb{R})$ of the Hilbert transform.

2.2.1 Dyadic Intervals

The *standard dyadic grid* $\mathscr{D}$ is the collection of intervals of the form $[k2^{-j}, (k+1)2^{-j})$, for all integers $k, j \in \mathbb{Z}$. They are organized by generations: $\mathscr{D} = \cup_{j\in\mathbb{Z}}\mathscr{D}_j$, and our labeling is such that $I \in \mathscr{D}_j$ iff $|I| = 2^{-j}$. They satisfy:

- *Trichotomy or nestedness:* $I, J \in \mathscr{D}$ then $I \cap J = \emptyset$, $I \subseteq J$, *or* $J \subset I$.
- *One parent, two children:* If $I \in \mathscr{D}_j$, then there is a unique interval $\tilde{I} \in \mathscr{D}_{j-1}$ such that $I \subset \tilde{I}$ and $|\tilde{I}| = 2|I|$. There are exactly two disjoint intervals, the right and left children $I_r, I_l \in \mathscr{D}_{j+1}$, such that $I = I_r \cup I_l$ and $|I| = 2|I_r| = 2|I_l|$.

2.2.2 Random Dyadic Grids

A dyadic grid in $\mathbb{R}$ is a collection of intervals, organized in generations, each of them being a partition of $\mathbb{R}$, that have the trichotomy and two children per interval property. For example, the shifted and rescaled regular dyadic grid will be a dyadic grid. However, these are *not* all possible dyadic grids.

The following parametrization will capture *all* dyadic grids. Consider the *scaling or dilation parameter* r with $1 \leq r < 2$ and the *random parameter* β with $\beta = \{\beta_i\}_{i\in\mathbb{Z}}$, $\beta_i = 0, 1$; let $x_j = \sum_{i<-j} \beta_i 2^i$ and then define

$$\mathscr{D}_j^\beta := x_j + \mathscr{D}_j, \qquad \text{and} \qquad \mathscr{D}_j^{r,\beta} := r\mathscr{D}_j^\beta.$$

The family of intervals $\mathscr{D}^{r,\beta}$ so defined is a dyadic grid. Here r is a dilation parameter, and β a random parameter that encode all possible dyadic grids. Notice that for the standard dyadic grid zero is never an interior point of a dyadic interval, and it is always on the right side of any dyadic interval it belongs to. If we translate $\mathscr{D}$ by a fixed number it will simply shift zero, and it will still have this singular property. The translated grids correspond to parameters β such that β_j is constant for all sufficiently large j. But these are not all the possible grids. Once we have an interval in a dyadic grid its descendants are completely determined, simply subdivide; however, there are two possible choices for the parent, four possible choices for the grandparent, and 2^n choices for the nth-parent. The parameter β captures all of these possibilities. Those β's that do not become eventually constant eliminate the presence of a singular point such as zero in the standard grid.

The random dyadic grids were introduced by Nazarov, Treil, and Volberg in their study of Calderón–Zygmund singular integrals on nonhomogeneous spaces [56] and are utilized by Hytönen in his representation theorem [32, 33]. The advantage of this parametrization is that there is a very natural probability space, say (Ω, P) associated to the parameters, and averaging here means calculating the expectation in this probability space, that is, $\mathbb{E}f = \int_\Omega f \, \mathrm{d}P$.

2.2.3 Haar Basis

Given an interval I, its associated *Haar function* is defined to be

$$h_I(x) := |I|^{-1/2}(\chi_{I_r}(x) - \chi_{I_l}(x)),$$

where $\chi_I(x) = 1$ if $x \in I$, zero otherwise. Note that $\|h_I\|_2 = 1$, and it has zero integral $\int h_I = 0$. One can check, from these integral properties and the nestedness properties of the dyadic intervals, that $\{h_I\}_{I\in\mathscr{D}}$ is an orthonormal system in $L^2(\mathbb{R})$. Furthermore, the system is complete, that is, it is an orthonormal basis in $L^2(\mathbb{R})$.

Alfred Haar introduced in 1910 the Haar basis in $L^2([0, 1])$ and showed that for continuous functions their Haar expansions converge uniformly [28], unlike their expansions in the trigonometric (Fourier) basis [21, 26, 73].

A basis is *unconditional in* $L^p(\mathbb{R})$ if and only if changes in the signs of the coefficients of a function keep it in the same space with comparable norms [82]. The trigonometric system $\{e^{2\pi inx}\}_{n\in\mathbb{Z}}$ does not form an unconditional basis in $L^p([0, 1))$ for $p \neq 2$ [26, 82]. On the other hand, the Haar basis $\{h_I\}_{I\in\mathscr{D}}$ is an unconditional basis in $L^p(\mathbb{R})$. More precisely we can define an operator, the *martingale transform*,

given by

$$T_\sigma f(x) = \sum_{I \in \mathscr{D}} \sigma_I \langle f, h_I \rangle h_I, \qquad \text{where} \quad \sigma_I = \pm 1. \tag{8}$$

Unconditionality of the Haar basis in $L^p(\mathbb{R})$ reduces then to show that the martingale transform is bounded in $L^p(\mathbb{R})$ with norm independent of the choice of signs:

$$\sup_\sigma \|T_\sigma f\|_p \le C_p \|f\|_p.$$

This was proved by Burkholder who also found the optimal constant C_p [7].

The Haar system $\{h_I\}_{I \in \mathscr{D}}$ *is an unconditional basis in* $L^p(w)$ *if and only if* $w \in A_p$. This fact is deduced from the boundedness of the martingale transform on $L^p(w)$ [75]. For sharp linear bounds in $L^2(w)$ for the martingale transform see [79].

The Haar basis is the first example of a wavelet basis, that is, a basis $\{\psi_{j,k}\}_{j,k \in \mathbb{Z}}$, that is found by translating and dilating appropriately a fixed function ψ, the wavelet, more precisely, $\psi_{j,k}(x) := 2^{-j/2}\psi(2^j x + k)$. The Haar functions are translates and dyadic dilates of the function $h(x) := \chi_{[0,1/2)}(x) - \chi_{[1/2,1)}(x)$. These unconditionality properties are shared by a large class of wavelets [29, 75, 82].

2.2.4 Petermichl's Dyadic Shift Operator

Petermichl's dyadic shift operator S associated to the standard dyadic grid $\mathscr{D}$ is defined for function $f \in L^2(\mathbb{R})$ by

$$Sf(x) := \sum_{I \in \mathscr{D}} \langle f, h_I \rangle H_I(x), \qquad \text{where} \quad H_I := 2^{-1/2}(h_{I_r} - h_{I_l}).$$

Petermichl's shift operator is an isometry in $L^2(\mathbb{R})$, that is, it preserves L^2-norms, $\|Sf\|_2 = \|f\|_2$. Notice that if $I \in \mathscr{D}$, $Sh_I(x) = H_I(x)$. A periodic version of the Hilbert transform that we denote by H_p, has the property that it maps cosines into sines, $H_p \cos(x) = \sin(x)$. Draw the profiles of h_I and H_I and you can view them as a localized sine and cosine. This indicates that this shift operator may be a good dyadic model for the Hilbert transform. More evidence comes from the way it interacts with translations, dilations, and reflections.

Denote by $S_{r,\beta}$ Petermichl's shift operator associated to the dyadic grid $\mathscr{D}_{r,\beta}$. Each shift operator $S_{r,\beta}$ does not commute with translations and dilations, nor does it anticommute with reflections; however, one can verify that the following symmetries for the family of shift operators $\{S_{r,\beta}\}_{(r,\beta)\in\Omega}$ hold:

- Translation: $\tau_h(S_{r,\beta} f) = S_{\tau_h(r,\beta)}(\tau_h f)$, where $\tau_h(r,\beta) \in \Omega$.
- Dilation: $D_\delta(S_{r,\beta} f) = S_{D_\delta(r,\beta)}(D_\delta f)$, where $D_\delta(r,\beta) \in \Omega$.
- Reflection: $\widetilde{S_{r,\beta} f} = S_{r,\tilde{\beta}}(\tilde{f})$, where $\tilde{\beta}_i = 1 - \beta_i$.

Where the maps $\tau_h, D_\delta : \Omega \to \Omega$ are bijections. Each shift dyadic operator does not have the symmetries that characterize the Hilbert transform, but the average over all dyadic grids will, therefore,

Theorem 2 (Petermichl's [32,66]).

$$\mathbb{E}S_{r,\beta} = \int_\Omega S_{r,\beta} \mathrm{d}P(r,\beta) = cH.$$

Petermichl's result then follows once one verifies that $c \neq 0$ (which she did!). Similar trick works for the *Beurling–Ahlfors* [70] and the *Riesz transforms* [68]. Vagharshakyan showed that sufficiently smooth one-dimensional Calderón–Zygmund convolution operators are averages of Haar shift operators of bounded complexity [76].

2.2.5 L^p Boundedness of the Hilbert Transform: A Dyadic Proof

Estimates for the Hilbert transform H follow from uniform estimates for Petermichl's shift operators.

Lemma 1 (Riesz [17]). *The Hilbert transform is bounded on L^p for $1 < p < \infty$.*

$$\|Hf\|_p \leq C_p \|f\|_p.$$

Proof. Suffices to check that

$$\sup_{(r,\beta)\in\Omega} \|S_{r,\beta} f\|_p \leq C_p \|f\|_p.$$

Case $p = 2$ follows from orthonormality of the Haar basis. First rewrite Petermichl's shift operator in the following manner, where $\tilde{I}$ is the parent of I in the dyadic grid $\mathscr{D}^{r,\beta}$:

$$S_{r,\beta} f = \sum_{I\in\mathscr{D}_{r,\beta}} \frac{1}{\sqrt{2}} \operatorname{sgn}(I,\tilde{I}) \langle f, h_{\tilde{I}}\rangle h_I,$$

where $\operatorname{sgn}(I,\tilde{I}) = 1$ if I is the right child of $\tilde{I}$ and -1 if I is the left child. We can now use Plancherel to compute the L^2 norm, and noticing that each parent has two children,

$$\|S_{r,\beta} f\|_2^2 = \sum_{I\in\mathscr{D}_{r,\beta}} \frac{|\langle f, h_{\tilde{I}}\rangle|^2}{2} = \|f\|_2^2.$$

Minkowski integral inequality then shows that

$$\left\|\mathbb{E}S_{r,\beta} f\right\|_2 \leq \mathbb{E}\|S_{r,\beta} f\|_2 \leq \|f\|_2.$$

Case $p \neq 2$ follows from the unconditionality of the Haar basis on $L^p(\mathbb{R})$. □

3 Weighted Inequalities and the A_2 Conjecture

In this section we discuss weighted inequalities for the Hilbert transform and recount the prehistory of linear estimates for dyadic operators on $L^2(w)$. We state the sharp extrapolation theorem and deduce $L^p(w)$ bounds from linear bounds and observe that these bounds are sometimes sharp, but not always, as Buckley's estimates for the maximal function show. We then define the Haar shift operators of complexity (m,n), discuss their boundedness properties, and finally state Hytönen's theorem (A_2 conjecture).

3.1 Boundedness on Weighthed L^p

The Hilbert transform is bounded on weighted $L^p(w)$; the celebrated 1973 Hunt–Muckenhoupt–Wheeden theorem says:

Theorem 3 (Hunt–Muckenhoupt–Wheeden [31]).

$$w \in A_p \Leftrightarrow \|Hf\|_{L^p(w)} \leq C_p(w)\|f\|_{L^p(w)}.$$

Dependence of the constant on the A_p characteristic was found 30 years later.

Theorem 4 (Petermichl [67]).

$$\|Hf\|_{L^p(w)} \leq C\,[w]_{A_p}^{\max\{1,\frac{1}{p-1}\}}\|f\|_{L^p(w)}.$$

Proof (Sketch of the proof). For $p = 2$ suffices to find uniform (on the grids) linear estimates for Petermichl's shift operator (this was the hard part which she did using Bellman functions and a bilinear Carleson embedding theorem due to Nazarov, Treil, and Volberg [55]). For $p \neq 2$ a sharp extrapolation theorem [20] that we will discuss in Sect. 3.1.2 automatically gives the result from the *linear estimate* in $L^2(w)$. □

3.1.1 Chronology of First Linear Estimates on $L^2(w)$

In 1993, Steve Buckley showed that the maximal function obeys a linear bound in $L^2(w)$ [6]. Starting in 2000, one at a time over a span of 10 years, a handful of dyadic operators or operators with enough symmetries that could be written as averages of dyadic operators were shown to obey a linear bound in $L^2(w)$; see (3):

- *Martingale transform* (Janine Wittwer [79] in 2000)
- *Dyadic square function* (Sanja Hukovic, Treil, Volberg [30], Wittwer [80] in 2000)

- *Beurling transform* (Petermichl, Volberg [70] in 2002)
- *Hilbert transform* (Stephanie Petermichl [67] in 2003, published 2007)
- *Riesz transforms* (Stephanie Petermichl [68] in 2008)
- *Dyadic paraproduct in* $\mathbb{R}$ (Oleksandra Beznosova [4] in 2008)

These estimates were based on Bellman functions and bilinear Carleson estimates by Nazarov, Treil, and Volberg [55]. See [61] for Bellman function extensions of the results for dyadic square functions to homogeneous spaces. See [9] for a neat Bellman function transference lemma that allows to use Bellman functions in $\mathbb{R}$ to deduce results in $\mathbb{R}^n$ with no sweat, similar considerations are used in [74]. There are now simpler Bellman function proofs that recover the estimates for the dyadic shift operators [54,74] and for the dyadic paraproduct [52]. The Bellman function method was introduced in harmonic analysis by Nazarov, Treil, and Volberg, and with their students and collaborators, they have been able to use this method to obtain a number of astonishing results not only in this area; see [77,78] and references.

3.1.2 Estimates in $L^p(w)$ via Sharp Extrapolation

The $L^p(w)$ inequalities can be deduced from the linear bounds on $L^2(w)$, thanks to a sharp version of Rubio de Francia's extrapolation theorem [25].

Theorem 5 (Sharp Extrapolation Theorem [20]). *If for all $w \in A_r$ there is $\alpha>0$, and $C > 0$ such that*

$$\|Tf\|_{L^r(w)} \leq C\,[w]_{A_r}^{\alpha}\,\|f\|_{L^r(w)},$$

then for all $w \in A_p$ and $1 < p < \infty$,

$$\|Tf\|_{L^p(w)} \leq C_{p,r}\,[w]_{A_p}^{\alpha \max\{1,\frac{r-1}{p-1}\}}\,\|f\|_{L^p(w)}.$$

Duoandikoetxea found recently a shorter proof of this theorem [22]. Sharp extrapolation from $r = 2$ is sharp for the martingale, Hilbert, Beurling–Ahlfors, and Riesz transforms for all $1 < p < \infty$ [20]. Therefore the theorem cannot be improved in terms of the power on the A_p characteristic of the weight. However, it is not necessarily sharp for each individual operator. The theorem is sharp for the dyadic square function and $1 < p \leq 2$, see [20], but it is not sharp for $p > 2$, see [46]. The optimal power for the square function is $\max\{\frac{1}{2}, \frac{1}{p-1}\}$ (see [18]), which corresponds to sharp extrapolation starting at $r = 3$ with square root power instead of starting at $r = 2$ with linear power; see also [50]. We conclude that *sharp extrapolation is not always sharp.* Buckley's estimates for the maximal function are a more dramatic example of the above statement.

Remember the Hardy–Littlewood maximal function is defined as

$$Mf(x) = \sup_{I \ni x} \frac{1}{|I|} \int_I |f(y)|\,\mathrm{d}y.$$

The maximal function is known to be bounded on $L^p(\mathbb{R})$ for $1 < p$; it is not bounded on $L^1(\mathbb{R})$, but it is of weak type $(1,1)$ [21, 26, 73]. Muckenhoupt showed in 1972 [53] that the maximal function is bounded on $L^p(w)$ if and only if $w \in A_p$. The optimal dependence on the A_p characteristic of the weight was discovered by Buckley 20 years later.

Theorem 6 (Buckley [6]). *Let $w \in A_p$ and $1 < p$, then*

$$\|Mf\|_{L^p(w)} \leq C_p [w]_{A_p}^{\frac{1}{p-1}} \|f\|_{L^p(w)}.$$

This estimate is key in the proof of the sharp extrapolation theorem. Observe that if we start with Buckley's estimate on $L^r(w)$, then sharp extrapolation will give the right power for all $1 < p \leq r$; however, for $p > r$, it will simply give $\frac{1}{r-1}$ which is bigger than the correct power $\frac{1}{p-1}$.

3.1.3 Estimates for Larger Classes of Operators

Petermichl's shift operator and the martingale transform are the simplest among a larger class of Haar shift operators that we now define.

A *Haar shift operator of complexity* (m,n), $S_{m,n}$, is defined as follows:

$$S_{m,n} f(x) := \sum_{L \in \mathscr{D}} \sum_{I \in \mathscr{D}_m(L), J \in \mathscr{D}_n(L)} c_{I,J}^L \langle f, h_I \rangle h_J(x), \tag{9}$$

where the coefficients $|c_{I,J}^L| \leq \frac{\sqrt{|I||J|}}{|L|}$ and $\mathscr{D}_m(L)$ denote the dyadic subintervals of L with length $2^{-m}|L|$.

The normalization of the coefficients ensures that $\|S_{m,n} f\|_2 \leq \|f\|_2$. The reader can now check that the martingale transform is a Haar shift operator of complexity $(0,0)$ and Petermichl's shift operator is a Haar shift operator of complexity $(0,1)$. However, the *dyadic paraproduct* π_b, which is defined for a function $b \in BMO$ as

$$\pi_b f(x) := \sum_{I \in \mathscr{D}} m_I f \, \langle b, h_I \rangle h_I(x), \quad \text{where} \quad m_I f = \frac{1}{|I|} \int_I f(x)\, \mathrm{d}x,$$

is not a Haar shift operator. The Haar shift operators were introduced in [44] and used in [17, 18]. Later, a larger class, the *generalized dyadic shift operators*, that included the paraproducts was defined [33, 37], where the Haar functions in (9) were replaced by $|I|^{-1/2}\chi_I(x)$ and boundedness on $L^2(\mathbb{R})$ is now part of the definition since it will not follow from the normalization of the coefficients. In this setting the dyadic paraproduct, the martingale transform, and Petermichl's Haar shift operator are generalized dyadic shift operators of complexity $(0,1)$, $(1,1)$, and $(1,2)$, respectively. The adjoint of the dyadic paraproduct, defined by

$$\pi_b^* f(x) = \sum_{I \in \mathscr{D}} \langle f, h_I \rangle \, \langle b, h_I \rangle \frac{\chi_I(x)}{|I|},$$

is a generalized dyadic shift operator of complexity $(1,0)$, and the composition $\pi_b^* \pi_b$ is of complexity $(0,0)$. On the other hand the composition $\pi_b \pi_b^*$ is not a generalized dyadic shift operator; localization has been lost.

The following authors either extend to other settings or recover most of the previous known results (the linear bounds on $L^2(w)$) and can extend them to the larger class of Haar shift operators, and in particular averaging appropriately, they can get Hilbert, Riesz, and Beurling–Ahlfors transforms:

- Lacey, Moen, Pérez, and Torres [43] obtain sharp bound on weighted L^p spaces for fractional integral operators.
- Lacey, Petermichl, and Reguera [44] use a *corona decomposition* and a *two-weight theorem for "well-localized operators"* of Nazarov, Treil, and Volberg, to recover linear bounds for Haar shifts operators on $L^2(w)$; they do not use Bellman functions. Dependence on the complexity is exponential. This result does not include dyadic paraproducts.
- Cruz-Uribe, Martell, and Pérez [17, 18] recover all results for Haar shift operators. No Bellman functions, no two-weight results. Instead they use a local median oscillation introduced by Lerner [47, 48]. The method is very flexible, they can get new results such as the sharp bounds for the square function for $p > 2$, they can recover also the result for the dyadic paraproduct, they can get results for vector-valued maximal operators and two-weight results as well. Dependence on complexity is exponential.

After these results were posted a lot of activity followed and results covering larger classes of operators appeared:

- Lerner [48, 50] showed that all standard convolution-type operators in arbitrary dimension gave the expected result for $p \in (1, 3/2] \cap [3, \infty)$. He also showed sharp estimates on $L^p(w)$ for all $p > 1$ and for all sort of square functions. This is based on controlling them with Wilson's intrinsic square function [81].
- Hytönen, Lacey, Reguera, Sawyer, Uriarte-Tuero, and Vagharshakyan posted a preprint in 2010 which was then replaced by a 2011 preprint with more authors [37]. They obtain the desired result for a general class of Calderón–Zygmund non-convolution operators, still requiring smoothness of the kernels.
- Pérez, Treil, and Volberg [65] showed that all Calderón–Zygmund operators obey an almost linear estimate on $L^2(w)$: $[w]_{A_2} \log(1 + [w]_{A_2})$. They identified the obstacle that would remove the log term.

3.1.4 The A_2 Conjecture (Now Theorem)

The A_2 conjecture said that all Calderón–Zygmund singular integral operators should obey a linear bound on $L^2(w)$. This was finally proved by Tuomas Hytönen in 2010.

Theorem 7 (Hytönen [33]). *Let $1 < p < \infty$ and let T be any Calderón–Zygmund singular integral operator in $\mathbb{R}^n$; then there is a constant $c_{T,n,p} > 0$ such that*

$$\|Tf\|_{L^p(w)} \leq c_{T,n,p}\,[w]_{A_p}^{\max\{1,\frac{1}{p-1}\}}\|f\|_{L^p(w)}.$$

It is enough to consider the case $p = 2$ thanks to sharp extrapolation. Hytönen proves the representation theorem, gets linear estimates on $L^2(w)$ with respect to the A_2 characteristic for Haar shift operators, and gets polynomial dependence in the complexity. Together these imply the theorem for $p = 2$. We consider the representation theorem to be of independent interest, and we state it here.

Theorem 8 (Hytönen [33]). *Let T be a Calderón–Zygmund singular integral operator, then*

$$Tf = \mathbb{E}\left(\sum_{(m,n)\in\mathbb{N}^2} a_{m,n} S_{m,n}^{r,\beta} f\right),$$

where the coefficients in the series are of the form $a_{m,n} = \mathrm{e}^{-(m+n)\alpha/2}$, α is the smoothness parameter of T, and $S_{m,n}^{r,\beta}$ are Haar shift operators of complexity (m,n) when $(m,n) \neq (0,0)$, and when $(m,n) = (0,0)$ they are a linear combination of a Haar shift of complexity $(0,0)$, a dyadic paraproduct, and the adjoint of the dyadic paraproduct, all based on the dyadic grid $\mathscr{D}_{r,\beta}$, and $\mathbb{E}$ is the expectation in the probability space (Ω, P) associated to the random dyadic grids $\mathscr{D}_{r,\beta}$.

Leading to the solution of the A_2 conjecture were the results of Pérez, Treil, and Volberg [65]. Since the appearance of Hytönen's theorem several simplifications of the argument have appeared [34, 38, 42, 54, 74], as well as an extension to metric spaces with geometric doubling condition [58]. There is also a very nice survey of the A_2 conjecture [41].

Can we expect more singular operators to have worst estimates? Yes, for example, the commutators of $b \in BMO$ with T a Calderón–Zygmund singular integral operator.

4 Sharp Weighted Inequalities for the Commutator

In this section we define the commutator, state its boundedness properties, and sketch Chung's dyadic proof of the quadratic estimate on $L^2(w)$. We note that this quadratic estimate is sharp, and we show that Chung's dyadic method of proof implies that Beznosova's bounds for the dyadic paraproduct are sharp as well.

Finally we state a variation of the transference theorem for commutators on $L^r(w)$ with $r \neq 2$ and present its proof in Appendix.

4.1 The Commutator

The commutator $[b, H]$ of $b \in BMO$ and H the Hilbert transform is defined:

$$[b, H]f = b(Hf) - H(bf).$$

It is well known that the commutator $[b, H]$ is bounded on $L^p(\mathbb{R})$.

Theorem 9 (Coifman et al. [12]). *Let $b \in BMO$ and $1 < p < \infty$, then*

$$\|[H, b]f\|_p \leq C_p \|b\|_{BMO} \|f\|_p.$$

However, the commutator is not of weak type $(1, 1)$ as Carlos Pérez showed [62]. The commutator $[b, H]$ is more singular than H. Another way to quantify this roughness is to observe that the maximal function M controls H; however, to control the commutator we need M^2 [63].

Observe that separately bH and Hb are *not* bounded on $L^p(\mathbb{R})$ when $b \in BMO$, simply because multiplication by a BMO function does not preserve $L^p(\mathbb{R})$ (one needs the multiplier to be bounded and $L^\infty(\mathbb{R}) \subsetneq BMO$). *The commutator introduces some key cancellation*. This is very much connected to the celebrated H^1 - BMO duality by Fefferman and Stein [23] (H^1 denotes the Hardy space on the line).

Coifman, Rochberg, and Weiss have a beautiful argument in [12] to prove boundedness on $L^p(\mathbb{R})$ of the commutator based on the boundedness of the Hilbert transform on $L^p(v)$ for $v \in A_2$; it is this argument that was exploited to obtain the following weighted inequalities for the commutator in quite a general framework; here we state the estimate for the Hilbert transform.

Theorem 10 (Alvarez et al. [1]). *If $w \in A_p$ and $b \in BMO$, then*

$$\|[H, b]f\|_{L^p(w)} \leq C_p(w) \|b\|_{BMO} \|f\|_{L^p(w)}.$$

4.2 Chung's Dyadic Argument

Daewon Chung proved the following sharp bound on $L^p(w)$ for the commutator of the Hilbert transform and a BMO function:

Theorem 11 (Chung [8]).

$$\|[H,b]f\|_{L^p(w)} \le C_p \|b\|_{BMO} [w]_{A_p}^{2\max\{1,\frac{1}{p-1}\}} \|f\|_{L^p(w)}.$$

The result is sharp in $L^2(w)$, meaning that in that case the quadratic power cannot be improved. Similar examples show extrapolated bounds are sharp in $L^p(w)$; see [8].

Chung's proof is based on a decomposition of the product bf using the dyadic paraproduct $\pi_b f$, its adjoint $\pi_b^* f$, and a related operator $\pi_f b$; this line of argument was suggested in [69]. He works with Petermichl's dyadic shift operator S instead of H, and Bellman functions. This argument works for dyadic shift operators (hence for Riesz and Beurling transforms, and it is sharp for them as well). We will sketch Chung's proof after some preliminaries on paraproducts.

4.2.1 Dyadic Paraproduct

Recall that dyadic paraproduct associated to the function $b \in BMO$ is defined by

$$\pi_b f(x) = \sum_{I \in \mathscr{D}} m_I f \, \langle b, h_I \rangle h_I(x), \qquad \text{where} \quad m_I f = \frac{1}{|I|} \int_I f(x)\, \mathrm{d}x.$$

The dyadic paraproduct is bounded on $L^p(\mathbb{R})$ for $1 < p < \infty$ and is of weak type $(1,1)$ [60]. Paraproducts appeared in the work of Bony [5] on paradifferential equations; they also appeared in the proof of the $T(1)$ theorem [19].

Theorem 12 (Beznosova [4]). *Let $b \in BMO$, $w \in A_2$, then for all $f \in L^2(w)$*

$$\|\pi_b f\|_{L^2(w)} + \|\pi_b^* f\|_{L^2(w)} \le C \|b\|_{BMO} [w]_{A_2} \|f\|_{L^2(w)}.$$

Ordinary multiplication $M_b f = bf$ is not bounded on $L^p(\mathbb{R})$ unless $b \in L^\infty(\mathbb{R})$. The space BMO includes unbounded functions. Hence the boundedness properties of the paraproduct are better than those of the ordinary product. It is well known that the following decomposition holds:

$$bf = \pi_b f + \pi_b^* f + \pi_f b. \tag{10}$$

The first two terms are not only bounded on $L^p(\mathbb{R})$ but are also bounded on $L^p(w)$ (follows by extrapolation from boundedness on $L^2(w)$) when $b \in BMO$ and $w \in A_p$; the enemy in this decomposition is the third term $\pi_f b$. It is because of this relation with the ordinary product that the name "paraproduct" was coined.

Proof (Sketch of Chung's proof of Theorem 11). Apply the decomposition (10) to the commutator with Petermichl's shift operator S:

$$[S,b]f = [S,\pi_b]f + [S,\pi_b^*]f + \big[S(\pi_f b) - \pi_{Sf}(b)\big]. \tag{11}$$

The first two terms give quadratic bounds from the linear bounds for S, π_b, and π_b^*. Boundedness of the commutator on $L^p(w)$ will be recovered from the uniform boundedness of the third commutator. Surprisingly (at the time this was discovered) the third term is better; it obeys a linear bound, and so do halves of the other two commutators:

$$\|S(\pi_f b) - \pi_{Sf}(b)\|_{L^2(w)} + \|S\pi_b f\|_{L^2(w)} + \|\pi_b^* S\|_{L^2(w)} \leq C\|b\|_{BMO}[w]_{A_2}\|f\|_{L^2(w)}.$$

Providing uniform *quadratic bounds* for the commutator $[S, b]$, hence

$$\|[H, b]\|_{L^2(w)} \leq C\|b\|_{BMO}[w]_{A_2}^2\|f\|_{L^2(w)}.$$

□

Chung proved his linear estimates using Bellman functions. A posteriori one realizes that the operators $[S(\pi_{\{.\}}b) - \pi_{S\{.\}}(b)]$, $S\pi_b$, and $\pi_b^* S$ are generalized Haar shift operators; hence, the linear bound is a particular case of the results in [33, 34, 37, 38]. For the commutator the bad terms are the nonlocal operators $\pi_b S$ and $S\pi_b^*$.

4.2.2 Commutators Versus Paraproducts

Beznosova proved the linear bound for the dyadic paraproduct, and then sharp extrapolation shows that the following bounds hold in $L^p(w)$ for $w \in A_p$:

$$\|\pi_b f\|_{L^p(w)} \leq C_p\|b\|_{BMO}[w]_{A_p}^{\max\{1, \frac{1}{p-1}\}}\|f\|_{L^p(w)}.$$

It was not known whether these were sharp for some or all $1 < p < \infty$.

Theorem 13. *The above estimate is optimal in the power* $\max\{1, \frac{1}{p-1}\}$.

Proof. Suppose there is an $\alpha < 1$ and a $p > 1$ such that for all $b \in BMO$ weights $w \in A_p$ and for all $f \in L^p(w)$ the following estimate holds:

$$\|\pi_b f\|_{L^p(w)} \leq C_p\|b\|_{BMO}[w]_{A_p}^{\alpha\max\{1, \frac{1}{p-1}\}}\|f\|_{L^p(w)}.$$

One can verify that the same estimate holds for π_b^*. Then we will obtain the following bound for the commutator of the Hilbert transform and b:

$$\|[b, H]f\|_{L^p(w)} \leq C_p\|b\|_{BMO}[w]_{A_p}^{(1+\alpha)\max\{1, \frac{1}{p-1}\}}\|f\|_{L^p(w)}.$$

And this is a contradiction because the power $2\max\{1, \frac{1}{p-1}\}$ is optimal for $[b, H]$.

□

4.3 Transference Theorem in $L^r(w)$ for Commutators

The following transference theorem holds:

Theorem 14 (Chung et al. [10]). *If a linear operator T obeys linear bounds in $L^2(w)$ for all $w \in A_2$*

$$\|Tf\|_{L^2(w)} \leq C\,[w]_{A_2}\|f\|_{L^2(w)},$$

then its commutator with $b \in BMO$ obeys quadratic bounds for all $w \in A_2$,

$$\|[T,b]f\|_{L^2(w)} \leq C\,[w]_{A_2}^2\|b\|_{BMO}\|f\|_{L^2(w)}. \tag{12}$$

Proof follows the beautiful Coifman–Rochberg–Weiss classical argument using the Cauchy integral formula and immediately generalizes to higher-order commutators $T_b^k := [b, T_b^{k-1}]$. Under the same assumptions of Theorem 14,

$$\|T_b^k f\|_{L^2(w)} \leq C\,k![w]_{A_2}^{k+1}\|b\|_{BMO}^k\|f\|_{L^2(w)}. \tag{13}$$

Extrapolation gives bounds on $L^p(w)$; they are sharp for all $1 < p < \infty$, all $k \geq 1$, and all dimensions, as examples involving the Riesz transforms show [10].

As a corollary of these and Hytönen's theorem we conclude that for each Calderón–Zygmund singular integral operators T there is a constant $C>0$ such that for all BMO functions b and for all A_2 weights w, (12) holds. Sharp extrapolation then shows that for all Calderón–Zygmund singular operators T,

$$\|[T,b]f\|_{L^p(w)} \leq C_p[w]_{A_p}^{2\max\{1,\frac{1}{p-1}\}}\|b\|_{BMO}\|f\|_{L^p(w)}. \tag{14}$$

A refinement of the argument in [10] shows that

Theorem 15. *If a linear operator T obeys a power bound in $L^r(w)$ for all $w \in A_r$,*

$$\|Tf\|_{L^r(w)} \leq C\,[w]_{A_r}^{\alpha}\|f\|_{L^r(w)},$$

then its commutator with $b \in BMO$ obeys the following bounds for all $w \in A_r$:

$$\|[T,b]f\|_{L^r(w)} \leq C_{n,r}[w]_{A_r}^{\alpha+\max\{1,\frac{1}{r-1}\}}\|b\|_{BMO}\|f\|_{L^r(w)}.$$

Notice that in the case of T a Calderón–Zygmund singular integral operator, we recover the $L^p(w)$ norm obtained from sharp extrapolation in [10], because the initial estimate on $L^p(w)$ corresponds to $\alpha = \max\{1, \frac{1}{p-1}\}$; hence in this case Theorem 15 gives (14). Because this bound is known to be sharp for the Hilbert and Riesz transforms, we deduce that the power obtained in Theorem 15 cannot be improved.

We present the proof of this result in the Appendix.

Generalizations and variations of these results have already appeared. Cruz-Uribe and Moen [16] prove corresponding estimates for commutators with fractional integrals (they also use the classical Coifman–Rochberg–Weiss argument). They use the machinery developed by Cruz-Uribe, Martell, and Pérez [18] and Lerner's local mean oscillation [48] to obtain two-weight estimates for the commutators with Calderón–Zygmund singular integral operators and fractional integrals. Carmen Ortiz-Caraballo [59] shows the following quadratic estimate for $b \in BMO$, and any Calderón–Zygmund operator T, on $L^p(w)$ where the weight is in $A_1 \subset \cap_{p>1} A_p$, the following estimate was obtained before Hytönen proved the A_2 conjecture, so it was the first nontrivial bound valid for all commutators of Calderón–Zygmund singular integral operators:

$$\|[T,b]\|_{L^p(w)} \leq C_n \|b\|_{BMO}\ p\,(p')^2 [w]^2_{A_1}.$$

There are now mixed A_p-A_∞ estimates that hold for all Calderón–Zygmund singular integral operators [34–36, 49]; inequality (15) is an example of such an estimate when $p = 2$. These estimates can be transferred to the commutators [36].

Theorem 16 (Hytönen and Pérez [36]). *If a linear operator T obeys the following bounds in $L^2(w)$ for all $w \in A_2$:*

$$\|Tf\|_{L^2(w)} \leq C\,[w]^{\frac{1}{2}}_{A_2} ([w]_{A_\infty} + [w^{-1}]_{A_\infty})^{\frac{1}{2}} \|f\|_{L^2(w)}, \tag{15}$$

then its commutator of order $k \geq 1$ with $b \in BMO$ obeys the following bounds for all $w \in A_2$:

$$\|T^k_b f\|_{L^2(w)} \leq C\,[w]^{\frac{1}{2}}_{A_2} ([w]_{A_\infty} + [w^{-1}]_{A_\infty})^{k+\frac{1}{2}} \|b\|_{BMO} \|f\|_{L^2(w)}.$$

The two-weight problem is still an outstanding open problem for most operators. Necessary and sufficient conditions are known for the maximal function via Sawyer-type conditions [51, 72], for the martingale transform and other dyadic operators [55] (these are of Sawyer type as well with respect to the dyadic operators), and for the dyadic square function (Beznosova, O., personal communication); compare to [81]. As for sufficient conditions many different sets are known, including several sets for the Hilbert transform [15, 40, 45, 57]. In all these cases the conditions are somehow inherent to the operator studied: "Sawyer-type conditions." An exception being sufficient conditions in terms of "bump conditions" in Orlicz spaces [16, 18]. Lacking are theorems of the nature; operator A is bounded from $L^p(u)$ into $L^p(v)$ if and only if operator B is bounded from $L^p(u)$ into $L^p(v)$.

Appendix

Proof (Sketch the proof of Theorem 15).

We "conjugate" the operator as follows: if z is any complex number we define

$$T_z(f) = \mathrm{e}^{zb} T(\mathrm{e}^{-zb} f).$$

Then, a computation gives (for instance for "nice" functions)

$$[b, T](f) = \frac{\mathrm{d}}{\mathrm{d}z} T_z(f)|_{z=0} = \frac{1}{2\pi i} \int_{|z|=\epsilon} \frac{T_z(f)}{z^2}\, \mathrm{d}z\,, \qquad \epsilon > 0$$

by the Cauchy integral theorem; see [1, 12].

Now, by Minkowski's inequality,

$$\|[b, T](f)\|_{L^r(w)} \leq \frac{1}{2\pi\, \epsilon^2} \int_{|z|=\epsilon} \|T_z(f)\|_{L^r(w)} |\mathrm{d}z|\,, \qquad \epsilon > 0.$$

The key point is to find the appropriate radius ϵ. First we look at the inner norm,

$$\|T_z(f)\|_{L^r(w)} = \|T(\mathrm{e}^{-zb} f)\|_{L^r(we^{rRez\, b})}\,,$$

and try to find appropriate bounds on z. To do this we use the main hypothesis, namely that T is bounded on $L^r(v)$ if $v \in A_r$ with

$$\|T\|_{L^r(v)} \leq C\, [v]_{A_r}^{\alpha}.$$

Let $v = we^{rRez\, b}$. We must check that if $w \in A_r$ then $v \in A_r$ for $|z|$ sufficiently small:

$$[v]_{A_r} = \sup_Q \left(\frac{1}{|Q|} \int_Q we^{rRez\, b(x)}\, \mathrm{d}x\right) \left(\frac{1}{|Q|} \int_Q w^{-\frac{1}{r-1}}(x) \mathrm{e}^{-\frac{r}{r-1} Rez\, b(x)}\, \mathrm{d}x\right)^{r-1}.$$

Now, since $w \in A_r$, then $w \in RH_q$ for some $q > 1$ [11]. Recall that $w \in RH_q$ if and only if there is a constant $C > 0$ such that for all cubes Q,

$$\left(\frac{1}{|Q|} \int_Q w^q \mathrm{d}x\right)^{\frac{1}{q}} \leq \frac{C}{|Q|} \int_Q w\,.$$

The following precise reverse Hölder condition for A_r weights holds [64]:

Lemma 2. *If $w \in A_r$ and $q = 1 + \dfrac{1}{2^{2r+n+1}[w]_{A_r}}$ (< 2), then $w \in RH_q$ and*

$$\left(\frac{1}{|Q|} \int_Q w^q \mathrm{d}x\right)^{\frac{1}{q}} \leq \frac{2}{|Q|} \int_Q w\,. \tag{16}$$

It is well known that if $w \in A_r$ then $\sigma := w^{-\frac{1}{r-1}} \in A_{r'}$ with $r' = \frac{r}{r-1}$ the dual exponent of r, and $[w]_{A_r}^{r'} = [\sigma]_{A_{r'}}^{r}$. Applying Lemma 2 to σ and r' we conclude then that if $s = 1 + \dfrac{1}{2^{2r'+n+1}[\sigma]_{A_{r'}}} < 2$ then $\sigma \in RH_s$ and

$$\left(\frac{1}{|Q|}\int_Q \sigma^s \mathrm{d}x\right)^{\frac{1}{s}} \leq \frac{2}{|Q|}\int_Q \sigma\,. \tag{17}$$

Let $t = \min\{q, s\}$, where q and s are as above, then $t \leq q$ and $t \leq s$. Holder's inequality with $p = q/t > 1$ and $p = s/t > 1$, respectively, implies that

$$\left(\frac{1}{|Q|}\int_Q w^t \mathrm{d}x\right)^{\frac{1}{t}} \leq \frac{2}{|Q|}\int_Q w, \qquad \left(\frac{1}{|Q|}\int_Q \sigma^t \mathrm{d}x\right)^{\frac{1}{t}} \leq \frac{2}{|Q|}\int_Q \sigma.$$

Using these and Holder's inequality twice with $p = t$, we have for an arbitrary Q

$$\left(\frac{1}{|Q|}\int_Q w(x)\mathrm{e}^{rRez\,b(x)}\,\mathrm{d}x\right)\left(\frac{1}{|Q|}\int_Q \sigma(x)\mathrm{e}^{-r'Rez\,b(x)}\,\mathrm{d}x\right)^{r-1} \leq 4\,[w]_{A_r}\,[\mathrm{e}^{t'r\,Rez\,b}]_{A_r}^{\frac{1}{t'}}.$$

Now, since $b \in BMO$, it is well known that $\mathrm{e}^{\eta b} \in A_r$ for η small enough [21,27]. We need a quantitative version of this result.

Lemma 3. *Given $b \in BMO$ then there are $0 < \alpha_n < 1$ and $\beta_n > 1$ such that if $\eta \leq \min\{1, r-1\}\frac{\alpha_n}{\|b\|_{BMO}}$, then $[\mathrm{e}^{\eta b}]_{A_r} \leq \beta_n^r$.*

This follows from a similar computation to the one done for $r = 2$ in [10]. In our case, we need to ensure that $|t'r\,Rez| \leq \min\{1, r-1\}\frac{\alpha_n}{\|b\|_{BMO}}$ to deduce that $[\mathrm{e}^{rt'Rez\,b}]_{A_r} \leq \beta_n^r$. That is, we are constrained to consider complex numbers z such that $|Rez| \leq \min\{\frac{1}{r}, \frac{r-1}{r}\}\frac{\alpha_n}{t'\|b\|_{BMO}}$.

Recall that $t = \min\{1 + \frac{1}{2^{2r+n+1}[w]_{A_r}}, 1 + \frac{1}{2^{2r'+n+1}[w]_{A_r}^{\frac{1}{r-1}}}\}$; a calculation now shows that

$$t' = \begin{cases} 1 + 2^{2r+n+1}[w]_{A_r} & p \geq 2 \\ 1 + 2^{2r'+n+1}[w]_{A_r}^{\frac{1}{r-1}} & p < 2 \end{cases}.$$

Furthermore $t' \sim [w]_{A_r}^{\max\{1, \frac{1}{r-1}\}}$ with comparability constant depending exponentially in the dimension n and $\max\{r, r'\}$.

For $|z| \leq \epsilon$, with $\epsilon^{-1} \sim \|b\|_{BMO}[w]_{A_r}^{\max\{1, \frac{1}{r-1}\}}$, and since $1 < t < 2$, thus $t' > 2$, we have that

$$[v]_{A_r} = [we^{rRez\,b}]_{A_r} \leq 4\,[w]_{A_r}\,\beta_n^{\frac{r}{t'}} \leq 4\,[w]_{A_r}\,\beta_n^{r/2}.$$

Observe that $\|e^{-zb} f\|_{L^r(v)} = \|e^{-zb} f\|_{L^r(we^{rRezb})} = \|f\|_{L^r(w)}$, and if $|z| \leq \frac{\alpha_n}{rt'\|b\|_{BMO}}$,

$$\|T_z(f)\|_{L^r(w)} = \|T(e^{-zb} f)\|_{L^r(v)} \leq [v]_{A_r}^{\alpha} \|f\|_{L^r(w)} \leq 4[w]_{A_r}^{\alpha} \beta_n^{r/2} \|f\|_{L^r(w)}.$$

Choose the radius $\epsilon = \dfrac{\alpha_n}{rt'\|b\|_{BMO}}$, and we can continue estimating the norm of the commutator

$$\|[b,T](f)\|_{L^r(w)} \leq \frac{1}{2\pi\,\epsilon^2} \int_{|z|=\epsilon} 4[w]_{A_r}^{\alpha} \beta_n^{r/2} \|f\|_{L^r(w)} |dz| = \frac{1}{\epsilon} 4[w]_{A_r}^{\alpha} \beta_n^{r/2} \|f\|_{L^r(w)}.$$

Finally, observe that ϵ^{-1} is essentially $[w]_{A_r}^{\max\{1,\frac{1}{r-1}\}} \|b\|_{BMO}$, so we conclude that

$$\|[b,T](f)\|_{L^r(w)} \leq C_{n,r}\, [w]_{A_r}^{\alpha+\max\{1,\frac{1}{r-1}\}} \|b\|_{BMO}.$$

This finishes the proof of the theorem. □

Acknowledgments The author would like to thank the organizers of the February Fourier Talks, at The Norbert Wiener Center for Harmonic Analysis and Applications, University of Maryland, for inviting her to deliver a talk in the fifth edition of the FFTs on February 18–19, 2010, that was the seed of this chapter. The author dedicates this chapter to the memory of her friend and mentor Cora Sadosky [1940–2010].

References

1. Alvarez, J., Bagby, R., Kurtz, D., Perez, C.: Weighted estimates for commutators of linear operators. Studia Mat. **104**(2), 195–209 (1994)
2. Astala, K., Iwaniec, T., Saksman, E.: Beltrami operators in the plane. Duke Math. J. **107**(1), 27–56 (2001)
3. Beylkin, G., Coifman, R., Rokhlyn, V., Fast Wavelet Transforms and Numerical Algorithms I. Comm. Pure Appl. Math. **44**(2), 141–183 (1991)
4. Beznosova, O., Linear bound for dyadic paraproduct on weighted Lebesgue space $L^2(w)$. J. Func. Anal. **255**(4), 994–1007 (2008)
5. Bony, J.: Calcul symbolique et propagation des singularités pour les equations aux dérivées non-linéaires. Ann. Sci. Ecole Norm. Sup. **14**, 209–246 (1981)
6. Buckley, S.: Estimates for operator norms and reverse Jensen's inequalities. Trans. Amer. Math. Soc. **340**(1), 253–272 (1993)
7. Burkholder, D.: Boundary value problems and sharp inequalities for martingale transforms. Ann. Probab. **12**(3), 647–702 (1984)
8. Chung, D.: Sharp estimates for the commutators of the Hilbert, Riesz and Beurling transforms on weighted Lebesgue spaces. To appear in Indiana U. Math. J. **60**(5), (2011)
9. Chung, D.: Weighted inequalities for multivariable dyadic paraproducts. Pub. Mat. **55**(2), 475–499 (2011)
10. Chung, D., Pereyra, M.C., Pérez, C.: Sharp bounds for general commutators on weighted Lebesgue spaces. Trans. Amer. Math. Soc. **364**(3), 1163–1177 (2012)
11. Coifman, R., Fefferman, C.: Weighted norm inequalities for maximal fuctions and singular integrals. Studia Math. **51**, 241–250 (1974)

12. Coifman, R., Rochberg, R., Weiss, G.: Factorization theorems for Hardy spaces in several variables. Ann. Math. **103**, 611–635 (1976)
13. Cotlar, M., Sadosky, C.: On the Helson–Szegö theorem and a related class of modified Toeplitz kernels. Harmonic analysis in Euclidean spaces (Proceedings of Symposium on Pure Mathematics, Williams College, Williamstown, MA, 1978), Part 1, pp. 383–407. Proceedings of Symposium on Pure Mathematics, vol. XXXV, Part, American Mathematical Society, Providence, RI (1979)
14. Cotlar, M., Sadosky, C.: On some L^p versions of the Helson-Szegö theorem. In: Conference on harmonic analysis in honor of Antoni Zygmund, Vol. I, II (Chicago, IL, 1981), pp. 306–317. Wadsworth Mathematical Series, Wadsworth, Belmont, CA (1983)
15. Cruz-Uribe, D., Pérez, C.: On the two-weight problem for singular integral operators. Ann. Scuola Norm. Super. Pisa Cl. Sci. (5) **I**, 821–849 (2002)
16. Cruz-Uribe, D., SFO, Moen, K., Sharp norm inequalities for commutators of classical operators. Pub. Mat. **56**(1), 147–190 (2012)
17. Cruz-Uribe, D., SFO, Martell, J.M., Pérez, C.: Sharp weighted estimates for approximating dyadic operators. Electron. Res. Announc. Math. Sci. **17**, 12–19 (2010)
18. Cruz-Uribe, D., SFO, Martell, J., Pérez, C.: Sharp weighted estimates for classical operators. Adv. Math. **229**(1), 408–441 (2012)
19. David, G., Journé, J.-L.: A boundedness criterion for generalized Calderón–Zygmund operators. Ann. Math. **120**, 371–397 (1984)
20. Dragičević, O., Grafakos, L., Pereyra, M.C., Petermichl, S.: Extrapolation and sharp norm estimates for classical operators on weighted Lebegue spaces. Publ. Math **49**(1), 73–91 (2005)
21. Duoandikoetxea, J., Fourier Analysis. Graduate Studies in Mathematics, vol. 29. American Mathematical Society, Providence (2001)
22. Duoandikoetxea, J.: Extrapolation of weights revisited: new proofs and sharp bounds. J. Funct. Anal. **260**(6), 1886–1901 (2011)
23. Fefferman, C., Stein, E.: H^p spaces of several variables. Acta Math. **129**, 137–193 (1972)
24. Figiel, T.: Singular intergral operators: a martingale approach. In: Geometry of Banach Spaces (Strbl, 1989), London Mathematical Society, vol. 158. Lecture Notes Series, pp. 95–110. Cambridge University Press, Cambridge (1990)
25. Garcia-Cuerva, J., Rubio de Francia, J.L.: Weighted norm inequalities and related topics. North-Holland Mathematics Studies, vol. 116. North-Holland, Amsterdam, Holland (1981)
26. Grafakos, L.: Classical Fourier Analysis. 2nd edn. Graduate Texts in Mathematics, vol. 249. Springer, New York (2008)
27. Grafakos, L.: Modern Fourier Analysis. Graduate Texts in Mathematics, vol. 250, 2nd edn. Springer, New York (2009)
28. Haar, A.: Zur Theorie der orthogonalen Funktionen systeme. Math. Ann. **69**, 331–371 (1910)
29. Hernandez, E., Weiss, G.: A First Course on Wavelets. 3rd edn. Academic, New York (2008)
30. Hukovic, S., Treil, S., Volberg, A.: The Bellman function and sharp weighted inequalities for square functions. In: Complex Analysis, Operators and Related Topics. Operator Theory: Advances and Applications, vol. 113, pp. 97–113. Birkhäuser, Basel (2000)
31. Hunt, R., Muckenhoupt, B., Wheeden, R.: Weighted norm inequalities for the conjugate function and the Hilbert transform. Trans. Amer. Math. Soc. **176**, 227–252 (1973)
32. Hytönen, T.P.: On Petermichl's dyadic shift and the Hilbert transform. C. R. Acad. de Sci. Paris, Ser. I **346**(21–22), 1133–1136 (2008)
33. Hytönen, T.P.: The sharp weighted bound for general Calderón–Zygmund operators. Ann. Math. (2) **175**(3), 1473–1506 (2012)
34. Hytönen, T.P.: Representation of singular integrals by dyadic operators, and the A_2 theorem. Preprint (2011) available at arXiv:1108.5119
35. Hytönen, T.P., Lacey, M.T.: The $A_p - A_\infty$ inequality for general Calderón–Zygmund operators. Preprint (2011) available at arXiv:1106.4797
36. Hytönen, T.P., Pérez, C.: Sharp weighted bounds involving A_∞. To appear Journal of Analysis and P.D.E., available at arXiv:1103.5562

37. Hytönen, T.P., Lacey, M.T., Martikainen, H., Orponen, T., Reguera, M.C., Sawyer, E.T., Uriarte-Tuero, I.: Weak and strong type estimates for maximal truncations of Calderón–Zygmund operators on A_p weighted spaces. Preprint (2011) available at arXiv:1103.5229
38. Hytönen, T.P., Pérez, C., Treil, S., Volberg, A.: Sharp weighted estimates for dyadic shifts and the A_2 conjecture. To appear in Journal für die Reine und Angewandte Mathematik, available at arXiv:1010.0755
39. John, F., Niremberg, L.: On functions of bounded mean oscillation. Comm. Pure Appl. Math. **14**, 415–426 (1961)
40. Katz, N.H., Pereyra, M.C.: On the two weights problem for the Hilbert transform. Rev. Mat. Iberoamericana **13**(1), 211–243 (1997)
41. Lacey, M.T.: The linear bound in A_2 for Calderón–Zygmund operators: a survey. Marcinkiewicz centenary volume, Banach Center Publ., 95, Polish Acad. Sci. Inst. Math., Warsaw, 97–114 (2011)
42. Lacey, M.T.: On the A_2 inequality for Calderón–Zygmund operators. Preprint (2011) available at arXiv:1106.4802
43. Lacey, M.T., Moen, K., Pérez, C., Torres, R.: The sharp bound the fractional operators on weighted L^p spaces and related Sobolev inequalities. J. Funct. Anal. **259**(5), 1073–1097 (2010)
44. Lacey, M.T., Petermichl, S., Reguera, M.C.: Sharp A_2 inequality for Haar shift operators. Math. Ann. **348**(1), 127–141 (2010)
45. Lacey, M.T., Sawyer, E., Shen, C.-Y., Uriarte-Tuero, I.: The two weight inequality for the Hilbert transform, coronas and energy conditions. Preprint (2011) available at arXiv:1108.2319
46. Lerner, A.K.: On some weighted norm inequalities for Littlewood–Paley operators. Illinois J. Math. **52**(2), 653–666 (2008)
47. Lerner, A.K.: A pointwise estimate for the local sharp maximal function with applications to singular integrals. Bull. London Math. Soc. **42**(5), 843–856 (2010)
48. Lerner, A.K.: A "local mean oscillation" and some of its applications. Function spaces, Approximation, Inequalities and Lineability, Lectures of the Spring School in Analysis, pp. 71–106. Matfyzpress, Prague (2011)
49. Lerner, A.K.: Mixed Ap-Ar inequalities for classical singular integrals and Littlewood–Paley operators. To appear in J. Geom. Anal., available at arXiv:1105.5735
50. Lerner, A.K.: Sharp weighted norm inequalities for Littlewood–Paley operators and singular integrals. Adv. Math. **226**, 3912–3926 (2011)
51. Moen, K.: Sharp one-weight and two-weight bounds for maximal operators. Studia Math. **194**, 163–180 (2009)
52. Moraes, J.C., Pereyra, M.C.: Weighted estimates for dyadic paraproducts and t-Haar multipliers with complexity (m, n). Submitted to Pub. Mat., available at arXiv:1108.3109
53. Muckenhoupt, B.: Weighted norm inequalities for the Hardy maximal function. Trans. Amer. Math. Soc. **165**, 207–226 (1972)
54. Nazarov, F., Volberg, A.: A simple sharp weighted estimate of the dyadic shifts on metric spaces with geometric doubling. Preprint (2011) available at arXiv:1104.4893
55. Nazarov, F., Treil, S., Volberg, A.: The Bellman functions and two-weight inequalities for Haar multipliers. J. Amer. Math. Soc. **12**(4), 909–928 (1999)
56. Nazarov, F., Treil, S., Volberg, A.: The Tb- theorem on non-homogeneous spaces. Acta Math. **190**(2), 151–239 (2003)
57. Nazarov, F., Treil, S., Volberg, A.: Two weight estimate for the Hilbert transform and corona decomposition for non-doubling measures. Preprint (2005) available at arXiv:1003.1596
58. Nazarov, F., Reznikov, A., Volberg, A.: The proof of A_2 conjecture in a geometrically doubling metric space. Preprint (2011), available at arXiv:1106.1342
59. Ortiz, C.: Quadratic A_1 bounds for commutators of singular integrals with BMO functions. Indiana U. Math. J., available at arXiv:1104.1069
60. Pereyra, M.C.: Lecture Notes on Dyadic Harmonic Analysis. Contemporary Mathematics, vol. 289. American Mathematical Society, Providence, RI, Chapter 1, pp. 1–61 (2001)
61. Pereyra, M.C.: Haar multipliers meet Bellman functions. Rev. Mat. Iberoamericana **25**(3), 799–840 (2009)

62. Pérez, C.: Endpoint Estimates for Commutators of Singular Integral Operators. J. Funct. Anal. **128**(1), 163–185 (1995)
63. Pérez, C.: Sharp estimates for commutators of singular integrals via iterations of the Hardy–Littlewood maximal function. J. Fourier Anal. Appl. **3**, 743–756 (1997)
64. Pérez, C.: A course on Singular Integrals and weights. To appear in Birkhäuser as part of the series "Advanced courses in Mathematics at the C.R.M., Barcelona".
65. Pérez, C., Treil, S., Volberg, A.: On A_2 conjecture and corona decomposition of weights. Preprint (2010) available at arXiv:1006.2630
66. Petermichl, S., Dyadic shift and a logarithmic estimate for Hankel operators with matrix symbol. C. R. Acad. Sci. Paris Sér. I Math. **330**(6), 455–460 (2000)
67. Petermichl, S.: The sharp bound for the Hilbert transform on weighted Lebesgue spaces in terms of the classical A_p characteristic. Amer. J. Math. **129**, 1355–1375 (2007)
68. Petermichl, S.: The sharp weighted bound for the Riesz transforms. Proc. Amer. Math. Soc. **136**, 1237–1249 (2008)
69. Petermichl, S., Pott, S.: An estimate for weighted Hilbert transform via square functions. Trans. Amer. Math. Soc. **354**(4), 1699–1703 (2002)
70. Petermichl, S., Volberg, A.: Heating of the Ahlfors-Beurling operator: Weakly quasiregular maps on the plane are quasiregular. Duke Math J. **112**, 281–305 (2002)
71. Riesz, M.: *Sur le fonctions conjuguées*. Math. Zeit. **27**, 218–244 (1927)
72. Sawyer, E.T.: A characterization of a two weight norm inequality for maximal operators. Studia Math. **75**(1), 1–11 (1982)
73. Stein, E.: Harmonic Analysis: real-variable methods, orthogonality, and oscillatory integrals. With the assistance of Timothy S. Murphy. Princeton Mathematical Series, vol. 43. Monographs in Harmonic Analysis, vol. III. Princeton University Press, Princeton (1993)
74. Treil, S.: Sharp A_2 estimates of Haar shifts via Bellman function. Preprint (2011) available at arXiv:1105.2252
75. Treil, S., Volberg, A.: Wavelets and the angle between past and future. J. Funct. Anal. **143**(2), 269–308 (1997)
76. Vagharsyakhan, A.: Recovering singular integrals from the Haar shifts. Proc. Amer. Math. Soc. **138**(12), 4303–4309 (2010)
77. Vasuynin, V., Volberg, A.: Notes on Bellman functions in harmonic analysis. Preprint (2010), available at http://sashavolberg.wordpress.com/
78. Volberg, A.: Bellman function technique in harmonic analysis. Lectures of INRIA Summer School in Antibes. France (2011), available at arXiv:1106.3899
79. Wittwer, J.: A sharp estimate on the norm of the martingale transform. Math. Res. Lett. **7**(1), 1–12 (2000)
80. Wittwer, J.: A sharp estimate on the norm of the continuous square function. Proc. Amer. Math. Soc. **130**(8), 2335–2342 (2002) (electronic)
81. Wilson, M.: Weighted Littlewood–Paley theory and exponential-square integrability. Lecture Notes in Mathematics, vol. 1924. Springer, Berlin (2008)
82. Wojtaszczyk, P.: A mathematical introduction to wavelets. London Mathematical Society Student Texts, vol. 37. Cambridge University Press, Cambridge (2003)

Part VIII
Biomathematics

Medical and biological sciences are nowadays spearheading important research and development efforts around the world. In this, they both benefit from and are benefactors of progress in applied mathematics, with many fundamental contributions arising in the field of harmonic analysis. Prominent examples of interactions between the mathematical and biomedical sciences include the role of the Kaczmarz algorithm in computed tomography and of Radon transforms in magnetic resonance imaging—both rewarded with Nobel Prizes in physiology and medicine. There is no doubt that this trend is going to continue. This is supported by the growing importance of fields such as systems biology, which reflects the fact that biological and medical models become increasingly more complex and involved, or bioinformatics, which addresses the issue of rapid growth in available medical data. With this in mind, we present some contributions describing state-of-the-art applications of harmonic analysis to current problems in the medical and biological sciences.

Alex Chen, Andrea L. Bertozzi, Paul D. Ashby, Pascal Getreuer, and Yifei Lou introduce us to the field of atomic force microscopy (AFM)—a powerful tool to study biological, chemical, and physical processes at the atomic level. The authors detail the role of mathematical advancements in this novel imaging modality. This includes the discussion of the role of sampling methodologies in AFM image reconstruction, as well as a review of a number of interpolation and inpainting approaches that are useful in AFM applications.

Gregory S. Chirikjian analyzes the role of representation theory and numerical harmonic analysis in the mechanics of double-helical DNA molecules. Through modeling of DNA as an elastic filament capable of bending and twisting, the author introduces the representation theory on unimodular Lie groups, with special emphasis placed on the three-dimensional group of rigid-body motions. The associated unitary irreducible representations are then used to provide explicit solutions of the diffusion equations describing the DNA structure. The result is a simplified model for a distribution of DNA poses.

Martin Ehler, Julia A. Dobrosotskaya, Emily J. King, and Robert F. Bonner present state-of-the-art mathematical applications in ophthalmology, the branch of medicine that deals with the human eye. It is not at all unexpected that the image analysis of our visual system poses a number of captivating problems of fundamental importance to our health. As an example the authors consider the early detection of age-related retinal diseases. Among such diseases is AMD (age-related macular degeneration), the most common cause of blindness among the elderly populations in the developed world. A number of image analysis tools is employed to understand its mechanics, including computational models for rhodopsin bleaching kinetics, variational inpainting techniques, and multispectral analysis.

The problems in magnetic resonance (MR) are analyzed by Evren Özarslan, Cheng G. Koay, and Peter J. Basser. Nuclear MR is a technique that allows us to obtain information about the imaged domain via the analysis of diffusion processes in that domain. The chapter focuses on MR performed in the q-space (i.e., after the mapping by the Fourier transform). This technique allows the researchers to analyze microscopic tissue structures, which otherwise are inaccessible to conventional MR

imaging. Hermite functions are one of their mathematical tools. Applications to reconstruction of certain two- and three-dimensional signals from one-dimensional measurements are also provided.

A long history of the use of Fourier analysis in the study of structured materials is revisited by Richard O. Prum and Rodolfo H. Torres. Their goal is to analyze the nature of nonpigmentary coloration in the tissues of living organisms. Their groundbreaking work mathematically establishes the fact that coherent light scattering can also be achieved as a result of reflection from quasi-ordered collagen fibers. This result manifests itself in our perception of certain nano-structures as colors.

Paul Hernandez-Herrera, David Jiménez, Ioannis A. Kakadiaris, Andreas Koutsogiannis, Demetrio Labate, Fernanda Laezza, and Manos Papadakis give us a harmonic analysis view on neuroscience imaging. The chapter begins with an extensive, historical, accessible overview of modern neuron imaging techniques. This is followed by a detailed study of the approximation errors due to the action of a group of orthogonal transformations on Euclidean space. These results depend on efficient directional representations, with examples including such novel representation systems as shearlets and curvelets. All this fascinating work culminates in an algorithm for computation of realistic models for naturally occurring neuronal dendrites.

Enhancement and Recovery in Atomic Force Microscopy Images

Alex Chen, Andrea L. Bertozzi, Paul D. Ashby, Pascal Getreuer, and Yifei Lou

Abstract Atomic force microscopy (AFM) images have become increasingly useful in the study of biological, chemical, and physical processes at the atomic level. The acquisition of AFM images takes more time than the acquisition of most optical images, so that the avoidance of unnecessary scanning becomes important. Details that are unclear from a scan may be enhanced using various image processing techniques. This chapter reviews various interpolation and inpainting methods and considers them in the specific application of AFM images. Lower-resolution AFM data is simulated by subsampling the number of scan lines in an image, and reconstruction methods are used to recreate an image on the original domain. The methods considered are classified in the categories of linear interpolation, nonlinear interpolation, and inpainting. These techniques are evaluated based on qualitative and quantitative measures, showing the extent to which scan times can be reduced while preserving the essence of the original features. A further application is in the removal of streaks, which can occur due to scanning errors and post-processing corrections. Identified streaks are removed, and the resulting unknown region is filled using inpainting.

A. Chen · A.L. Bertozzi (✉)
UCLA, Department of Mathematics, Los Angeles, CA, USA
e-mail: achen81@stanfordalumni.org bertozzi@math.ucla.edu

P.D. Ashby
Lawrence Berkeley National Laboratory, Molecular Foundry, Berkeley, CA, USA,
e-mail: pdashby@lbl.gov

P. Getreuer
CMLA, ENS Cachan, France
e-mail: getreuer@gmail.com

Y. Lou
Georgia Institute of Technology, School of Electrical and Computer Engineering, Atlanta, GA, USA
e-mail: louyifei@gmail.com

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_16,

Keywords Atomic force microscopy • Streak removal • Image inpainting • Variational methods • Edge preservation • Image interpolation • Subsampling • Image reconstruction • Dynamic imaging

1 Introduction

The atomic force microscope (AFM) is an extremely high-magnification microscope [5]. It achieves its high resolution by moving an atomically sharp probe over surfaces and recording the highly localized interaction force. Isolating specific interaction forces such as electrostatic, magnetic, specific chemical interactions, van der Waals attraction, and Pauli repulsion enables the AFM to measure many surface properties in addition to topography [2, 12, 17, 20, 22, 29]. The AFM is also able to measure surfaces in any environment from liquids to corrosive gases and vacuum. The high resolution, versatility, and broad information content make AFM a frequent choice for nanoscience imaging.

The current standard method of AFM data collection is the raster scan. The probe starts by traveling along the "fast scan direction," or in the $+x$ direction. As it reaches the end of the scan region, it takes a small step in the "slow scan direction," or $+y$ direction, and scans in the $-x$ direction until it retraces the x displacement. Another small step in the $+y$ direction is taken, and the scanner moves in the $+x$ direction to initiate another scan line. Continuing in this manner, an image is formed. The backward (retrace) scans are often displayed independently from the forward (trace) scans due to errors in position from scanner nonlinearities and hysteresis. A feedback mechanism maintains the probe–sample interaction at constant force to ensure that the probe is not damaged by contact with the sample.

Because the AFM is a local probe it must collect data serially to construct an image over time which can be a significant disadvantage. The sample and probe are massive objects that are difficult to accelerate requiring relatively slow scan velocities otherwise the feedback mechanism that holds the interaction force constant may not be able to compensate quickly enough, causing erroneous readings or damage to the probe. This problem is even more pronounced when the sample has sharp gradients. Another problem is thermal drift, the tendency for the probe and sample to move relative to each other due to temperature variations in the probe, sample, and substrate [19]. Since an image is formed point by point, the topography data may be skewed or distorted. As a result, it is challenging for AFM to record dynamic processes.

Figure 1 shows images of a chemical reaction that occurs faster than the AFM scan time. Shorter scan times are required to better capture sample dynamics. Before oxidation, the surface is atomically flat with a few step edges where the sample changes height by one atomic layer. After oxidation for the same region, the material at the step edges has been reacted leading to roughening, erosion, and migration. Imaging the surface during oxidation does not sufficiently resolve the oxidation edges, since data is collected at varying times.

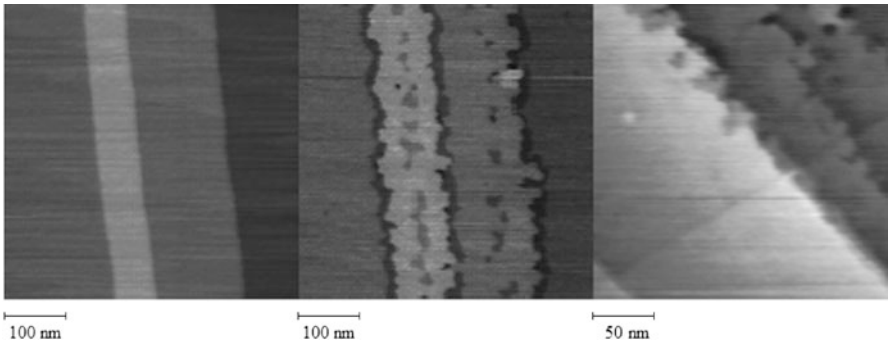


Fig. 1 AFM scans showing morphological changes to a potassium bromide surface when oxidized by ozone. *Left*: Before oxidation. *Middle*: After oxidation. *Right*: Image of another surface collected during oxidation in an attempt to observe the chemical reaction. However, the reaction happens during a single image. The time and location of specific oxidation events is unavailable

An active area of research to improve the temporal resolution of AFM includes building lighter and stronger scanners that can operate at higher frequencies while maintaining a safe probe–sample interaction force. The best-performing instruments can record images at video rate [1, 18]. However, this approach often has significant sample size and environmental limitations compromising the versatility of AFM. Methods that increase the instrument's temporal resolution while maintaining versatility are needed.

Recording fewer scan lines per image can make AFM image collection faster. Alternatively, tracking only the boundaries of important features can drastically reduce image times. The important question regarding these methods is whether such scans still resolve the areas of interest sufficiently.

There are two classes of image processing techniques, interpolation and inpainting, used to fill in missing data. Though there is significant overlap in the methods and approaches, we generally take interpolation to denote methods based on local averaging ideas and let inpainting refer to methods that detect important image features in a known region and seek to continue these into the unknown region.

One of the major applications of interpolation is to increase the resolution of an image by interpolating intermediate values between known data points. Interpolation is thus well suited to the problem of converting coarse raster scans to higher-resolution images.

This chapter considers various image reconstruction methods and the degree to which AFM images can be enhanced. The focus is on using interpolation and inpainting techniques on AFM images that have been obtained with fewer scan lines (subsampling along the slow scan direction). This is in contrast to the typical applications of interpolation, in which both axes are usually subsampled by the same factor, and inpainting, in which one typically has large connected known regions and unknown regions.

The rest of the chapter is organized as follows. Section 2 introduces the reconstruction problem and relevant terminology. Sections 3 and 4 review some commonly used interpolation algorithms and assess their strengths and weaknesses. Inpainting techniques, which look at the reconstruction from another perspective and are more readily generalizable, are addressed in Sect. 5. Several reconstructions on AFM images are presented in each of these sections.

2 Description of the Reconstruction Problem

The scenes underlying images are often taken to lie in continuous space. When images of these scenes are captured, they are sampled at a certain rate and thus mapped to discrete space. This sampling rate is directly related to the image resolution. If the sampling rate is increased (upsampling), the image resolution is increased. Similarly, downsampling decreases image resolution. Interpolation can thus be reinterpreted as the inverse problem of recreating a higher-resolution version of a given scene.

Let the given discrete image be denoted as $\bar{I}_{m,n}$ for $(m,n)\ \varepsilon\ \mathbf{Z}^2 \bigcap \Omega$. The problem is to reconstruct an image $I(x,y)$ with $(x,y)\ \varepsilon\ \Omega$. Such an image $I(x,y)$ should be close to $\bar{I}$ at the points $(m,n)\ \varepsilon\ \mathbf{Z}^2 \bigcap \Omega$ and continue the structural features present in $\bar{I}$.

2.1 Adapting Reconstruction Methods to AFM

The fastest AFM scans that still distinguish image features are the most preferable. Image reconstruction methods can then be used to obtain an approximation of a higher-resolution version. There are, however, further considerations in adapting these methods to AFM problems.

In general, the maximum sampling rate in space for data points is proportionally related to the speed of the tip, while the tip speed is limited by considerations such as feature height and scanning pattern. For a fixed tip speed, there is no benefit to sampling at a rate lower than the maximum rate since the tip still must traverse the same area [8]. This consideration fixes the number of points in each scan line for a specific velocity. However, decreasing the number of scan lines can bring practical benefits, as long as the important features are still being detected.

In terms of the reconstruction problem, this means that interpolating by different factors in the two dimensions of the image is of great interest. The usual treatment of the interpolation problem, however, is that both dimensions are scaled by the same factor. Most interpolation algorithms are valid when the scaling in the two dimensions is not equal, but in the AFM reconstruction problem, care must be taken to ensure that reconstructions consider features, such as interrupted edges, correctly.

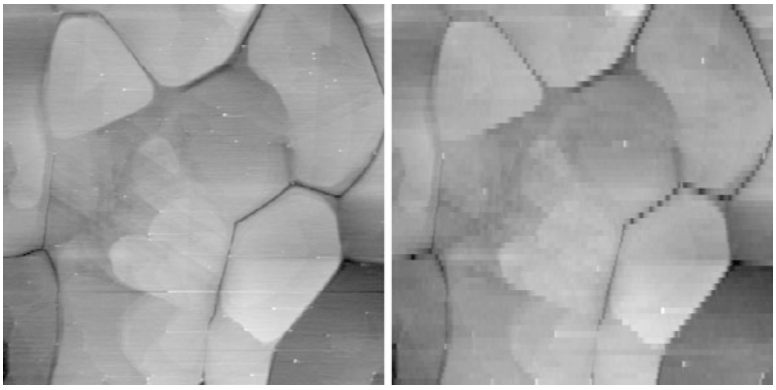


Fig. 2 Subsampling the "annealed gold" image in the slow scan (vertical) direction. *Left*: Original image. *Right*: Subsampling by a factor of 4 along the vertical axis. Nearest neighbor interpolation to preserve the aspect ratio

A further distinction between the AFM application and many other applications is in image acquisition. For each pixel, optical systems use, for example, a charge-coupled device to measure incoming light over a small photoactive region. Since lenses and filters also introduce blur, each sample represents a weighted average over a small area in continuous space. In acquiring an AFM image, however, the tip obtains a sample by visiting a point instead of averaging over a region. While this sample still represents a convolution between the tip and the surface, it is concentrated over a smaller area. In other words, optical images aggregate over the entire domain while AFM images capture concentrated point samples.

This subtle difference in image acquisition can result in the loss of image features from subsampling. Typically, a lower-resolution image can be obtained from a higher-resolution version by some averaging of the original data [21]. Such averaging results in the greatest retention of information. This method is also analogous to the taking of a lower-resolution camera image because the amount of light from a neighborhood of pixels is averaged. There is, however, significant information loss when the lower-resolution image is formed from simply discarding certain lines of data, as in AFM subsampling. Figure 2 shows the contrast between an image of annealed gold and the same image obtained by subsampling lines. In the latter image, the intensities along some edges vary due to the loss of information around the edge pixels, resulting in pixelation.

This method of sampling also makes AFM imaging more susceptible to *aliasing*. Aliasing is an effect where an oscillating pattern appears to change frequency after sampling. It occurs when sampling a pattern that is finer than what is representable with the image resolution and manifests in the sampled data as artificial oscillations and Moiré patterns. The blurring in optical systems cancels out ("anti-aliases") most high-frequency oscillations so that aliasing is limited. AFM imaging does not have as much blurring, so aliasing is a problem.

2.2 Desirable Traits of Reconstruction Methods

Image reconstruction methods are typically evaluated based on several criteria. Generally, preserving edge sharpness allows objects to be distinguished clearly from each other and from the background. At the same time, it is important to filter noise in order to remove random features that may obscure the image and make it difficult to evaluate. Unfortunately, the two goals are often antithetical since both edges and noise are typically defined by high gradients. In contrast to the randomness of noise, however, edges often can be identified as continuous contours traversing high-gradient regions. A good reconstruction algorithm keeps edges sharp while smoothing noise.

Another consideration is the "connectivity principle," [7] which states that edge curves should be connected through the unknown region whenever possible. This principle is based on human perception and experience, as well as the particular prevalence of long, thin objects in nature and in man-made applications. Particular examples include road inpainting [3] and the identification of bar codes [10, 11]. Reconstruction algorithms have historically had particular problems connecting such slim objects through an unknown region. Inpainting methods relying on the evolution of a fourth-order PDE [25] are more likely to satisfy the connectivity principle due to their penalization of high-curvature edges.

A related problem is the shape of such an edge connection. Often, edges are connected by the shortest path [6], which can result in unrealistic kinks. The same penalization on high-curvature edges discussed for the connectivity principle also fixes the overreliance on shortest path connections. These considerations are well studied for the typical inpainting problem, in which there are large unknown regions interrupting mostly known data.

The connectivity principle and staircasing pose particular challenges in the AFM inpainting problem. Since the known data is relatively more disconnected than in usual inpainting applications, it is unlikely to expect edge connections with the same degree of effectiveness. Indeed, experiments in Sect. 7 show that this is the case.

3 Linear Interpolation

Linear interpolation methods average values of nearby pixels to calculate values at intermediate points. Mathematically, the reconstructed image is the result of the input image convolved with a given kernel. Linear interpolation methods are linear in the sense that the relative weights on neighboring points in the average do not depend on their respective intensities, that is, the convolution kernel is not dependent on the image intensity values. Thus, these methods are fast and easy to calculate. At the same time, since they do not take any edge information into account, there is always a trade-off between artificial ripples and staircasing along diagonal edges.

3.1 Nearest Neighbor

Nearest neighbor interpolation is one of the simplest methods of interpolation. In this method, the value of the nearest known point is copied directly without regard to any other point. That is, if $(x, y)\ \varepsilon\ \Omega$, then $I(x, y) = I(m, n)$ where $(m, n) = \text{argmin}_{(i,j)\ \varepsilon \mathbf{Z}^2 \cap \Omega} \text{dist}((i, j), (x, y))$.

Since nearest neighbor interpolation copies data points $\bar{I}_{m,n}$ without alteration, any noise will also be copied. Similarly, edges are thickened, giving the image a blocky, pixelated appearance. Nearest neighbor interpolation is also useful in comparing images on varying domain sizes without altering the underlying quality of the image.

3.2 Bilinear and Bicubic Interpolation

Taking averages is one of the primary methods used to eliminate noise, one of the primary problems with nearest neighbor interpolation. Polynomial interpolation methods fill in intermediate points by taking a weighted average, with the weighting depending on their distances to nearby points. Bilinear and bicubic interpolation are examples of these methods.

Bilinear interpolation is a combination of two linear interpolation steps along the x-coordinate, then along the y-coordinate. First, linear interpolation is performed along the x-coordinate for each fixed y-coordinate, followed by linear interpolation along the y-coordinate.

If $x_1 = \lfloor x \rfloor$, $y_1 = \lfloor y \rfloor$, $c = x - x_1$, $d = y - y_1$, where $\lfloor \cdot \rfloor$ denotes the floor function, then bilinear interpolation is given by

$$I(x, y) = (1-c)(1-d)I_{x_1,y_1} + (1-c)d\,I_{x_1,y_1+1} + c(1-d)I_{x_1+1,y_1} + cd\,I_{x_1+1,y_1+1}.$$

An undesirable property of nearest neighbor interpolation is that the result is artificially discontinuous between pixels. Bilinear and other higher-order polynomial methods construct the interpolant from continuous piecewise polynomials, so the result is always continuous. This has an effect of smoothing the image, which improves interpolation of smooth regions and directional features.

Bicubic interpolation is similarly a two-dimensional version of cubic interpolation. The values in each square $(x, y)\ \varepsilon\ [m, m + 1] \times [n, n + 1]$ are approximated by a polynomial that has at most cubic terms in both x and y.

The resulting polynomial $I(x, y) = \sum_{i=0}^{3} \sum_{i=0}^{3} a_{i,j} x^i y^j$ can be calculated by using the values $\bar{I}$, $\frac{\mathrm{d}\bar{I}}{\mathrm{d}x}$, $\frac{\mathrm{d}\bar{I}}{\mathrm{d}y}$, and $\frac{\mathrm{d}^2\bar{I}_{x,y}}{\mathrm{d}x\mathrm{d}y}$ at the corners, with the derivatives being calculated numerically. Bicubic interpolation results in an even smoother reconstruction than bilinear interpolation. Since bicubic interpolation is also computationally efficient, it is often used for resizing applications [26].

3.3 Lanczos

Interpolation with a sinc kernel, also known as Whittaker–Shannon interpolation, has the remarkable property that the interpolation is exact when the underlying signal is bandlimited. Applied to image interpolation, this results in an extremely smooth image. Sinc interpolation also avoids the staircasing that can occur at diagonal edges. Unfortunately, since edges are the result of sharp changes, application of the sinc filter across them results in significant ripple artifacts as the edges are fitted to lower frequencies. The Lanczos filter is a windowed version of the sinc filter. The windowing of the sinc function allows for the higher frequency changes characteristic of edges. It thus provides a compromise between staircasing and ripple effects.

4 Nonlinear Interpolation

Nonlinear interpolation algorithms attempt to solve the problems of staircasing and rippling by taking an adaptive approach. Any image of practical interest has some structure, which can be used in the reconstruction. Thus, instead of an unbiased average of intensity values, averaging is based on the detection of edges. These methods generally attempt to average along edges to preserve edge sharpness. As with linear interpolation, noise is smoothed since these points are not identified as edge points.

4.1 Contour Stencils

Interpolation by contour stencils was introduced [13, 14] as a nonlinear method to detect the orientation of edges. Edges are detected in the input image by comparing each image patch to each element in a set of "contour stencils," which predict the location and direction of edges. The stencil which provides the best match to the image patch is selected. The reconstruction step then follows by interpolating according to the selected stencil.

The predicted orientation of edges follows from measuring the total variation along a curve,

$$\|u\|_{TV(C)} = \int_0^T \left| \frac{\partial}{\partial t} u(\gamma(t)) \right| \, \mathrm{d}t, \gamma : [0, T] \to C,$$

so that a small value for $\|u\|_{TV(C)}$ suggests that an edge lies along C.

In order to make the problem computationally efficient, a subset of possible image contours is used, the set Σ of "contour stencils." An example set of contour

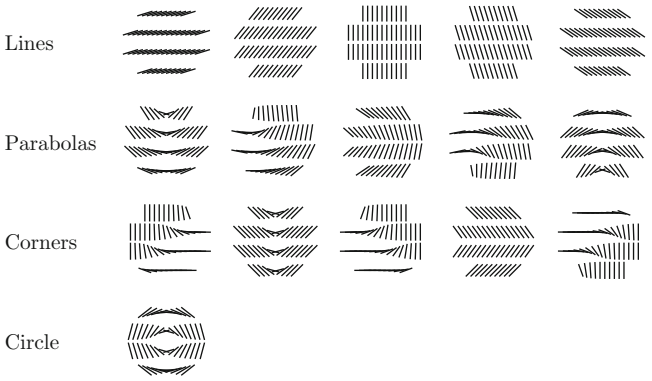


Fig. 3 Several contour stencils for a rectangular grid with pixel aspect ratio 4:1. The *lines* depict the orientation measured over each cell of the neighborhood. The stencil set comprises lines at 32 orientations, 16 parabolas, 8 corners, and a circle

stencils is shown in Fig. 3. This set of contour stencils can distinguish between eight different orientations for edges.

Once the contour orientations at every pixel have been estimated, the interpolant is constructed as a linear combination of oriented Gaussians. In this way, the interpolated image is encouraged to have the same contour orientations as those detected in the input image.

4.2 Prioritized Line Interpolation

This section introduces prioritized line interpolation (PLI), which is designed to connect edges that have been broken by the subsampling process. The idea is to assume that edges are locally linear in space and are locally nearly constant. Then, starting with the highest gradient points, which are more likely to be edge points, the algorithm searches a neighborhood for possible edge connections. Once a suitable location is found, unknown points along the connection are filled in by linear interpolation.

The PLI algorithm is as follows:

1. Each point in the interior of the image is placed in a priority queue $\{A_i\}_{i=1}^{N}$, where N is the number of interior pixels, in decreasing order of a function based on the discrete gradient; that is, the function

$$f(u_{i,j}) = |u_{i+1,j} - u_{i,j}| + |u_{i-1,j} - u_{i,j}| + |u_{i,j+1} - u_{i,j}| + |u_{i,j-1} - u_{i,j}|.$$

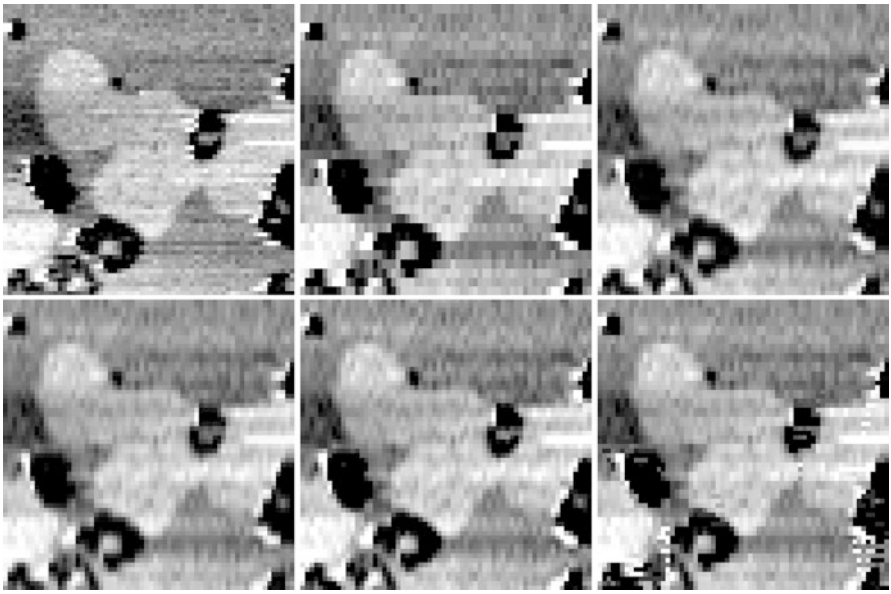


Fig. 4 Interpolation reconstructions from a section of an image of lipid bilayer domains on mica, subsampled by a factor of 2 on the vertical axis. *Left to right*, *top to bottom*: Original image; nearest neighbor interpolation; Lanczos-3 interpolation; bicubic interpolation; contour stencil interpolation; PLI

2. For a neighborhood N_i of the point A_i (a square neighborhood centered at A_i with radius r, for simplicity), the algorithm considers any known points $\{B_{i_j}\}$ with gray value within a certain threshold t_1.
3. For each B_{i_j}, the unknown points lying between B_{i_j} and A_i are determined by linear interpolation, as long as the sum of the absolute differences from the known points do not exceed a threshold t_2.
4. Repeat steps 2 and 3 until a significant portion of the image is filled.
5. Any remaining unknown points are filled with some interpolation or inpainting algorithm.

PLI is able to reconstruct edges and other large-scale features well. There are significant small-scale artifacts, however, making the algorithm less suitable for reconstructing small and thin features (Figs. 4 and 5).

In the examples that follow, $r = 10$, $t_1 = 5$, and $t_2 = 5$.

5 Variational Inpainting

The principle underlying inpainting is similar to that of nonlinear interpolation algorithms such as contour stencils. These algorithms generally detect features in

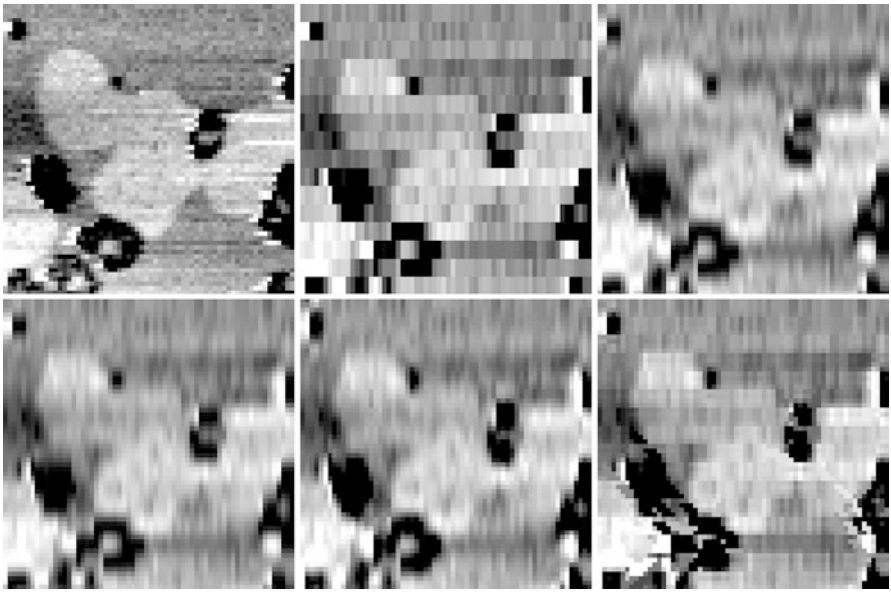


Fig. 5 Interpolation reconstructions from a section of an image of lipid bilayer domains, subsampled by a factor of 4 on the vertical axis. *Left to right*, *top to bottom*: Original image; nearest neighbor interpolation; Lanczos-3 interpolation; bicubic interpolation; contour stencil interpolation; PLI

the known region and continue them into the unknown region while preserving properties such as edge continuity and curvature.

A major advantage of inpainting over interpolation methods is that they are more readily generalizable to reconstructing information in general unknown regions. With the extra information available from subsampling on only one axis instead of both, inpainting algorithms can reconstruct many features in an image more accurately. This is especially relevant to AFM applications, in which the sampling rate can be increased in the fast scan direction much more readily in the slow scan direction. Thus, the fast scan direction typically has sufficient resolution, while the slow scan direction requires enhancement. Inpainting methods, however, are usually more computationally expensive than interpolation methods.

Variational image inpainting methods define energy functionals that seek to recreate plausible images given the known data. These energies generally have the following structure:

$$E(u) = R(u) + \lambda(x, y)F(d(f, u)),$$

where $R(u)$ is a "regularization" term that penalizes unlikely image features such as high gradients (relatively less common than smooth changes), and $\lambda(x, y)F(d(f, u))$ is a data "fidelity" term that penalizes deviation from known data, as measured

by the distance function $d(\cdot,\cdot)$. The weight $\lambda(x, y)$ is generally chosen to have a constant weight λ in the known region Ω_1 and 0 in the unknown region:

$$\lambda(\boldsymbol{x}) = \begin{cases} \lambda, \text{ if}(x, y)\ \varepsilon\ \Omega_1, \\ 0, \text{ if}(x, y)\ \varepsilon\ \Omega \backslash \Omega_1. \end{cases} \tag{1}$$

Minimization of the energy gives an image with the desired properties. One of the simplest models for variational inpainting with this structure is the H^1 (diffusion) model:

$$E(u) = \frac{1}{2}\int_{\Omega} |\nabla u|^2 \, \mathrm{d}x \, \mathrm{d}y + \frac{\lambda}{2}\int_{\Omega_1} (f-u)^2 \, \mathrm{d}x \, \mathrm{d}y.$$

The regularization term of the H^1 energy indeed penalizes the high gradients characteristic of noise. Unfortunately, edges are also excessively smoothed due to the squared penalty on gradients. The gradient descent equation also shows this fact:

$$u_t = \Delta u + \lambda(f-u),$$

which indicates that the propagation of information is by isotropic diffusion. Generally, features are reconstructed reasonably well, but since the diffusion is completely unbiased, significant blurring results.

A significant improvement is the total variation (TV) model of Rudin, Osher, and Fatemi [24], originally for image denoising. The TV inpainting energy is

$$E(u) = \int_{\Omega} |\nabla u| \, \mathrm{d}x \, \mathrm{d}y + \lambda \int_{\Omega} (f-u)^2 \, \mathrm{d}x \, \mathrm{d}y.$$

As in the H^1 model, large gradients are penalized, so that the model seeks smooth continuations of the data while removing noise. However, the lack of a square on the regularization term prevents excessive penalization. There are various methods to minimize this energy. Gradient descent has typically been used as a simple and straightforward method to find a minimizer. More recently, the Split Bregman method [15] and graph cuts [9] have made minimization more efficient.

5.1 Fourth-Order Inpainting Methods

The TV model significantly improves edge definition. There are several other factors, however, that are desirable in an inpainting model. The connectivity principle is introduced in Sect. 2.2. In the TV model, the connectivity principle is, in particular, often violated when connecting broken edges. If an unknown region separates two long, thin objects flowing toward each other, it is logical to assume that they should be connected through the unknown region. Yet the added amount of total variation needed to connect the objects may be high, so the TV model would keep the two objects separate.

One solution is to add a penalty on edge contour curvature, since a long, thin object which ends abruptly certainly has high curvature at its terminus. The curvature term adds to the complexity of the energy and the corresponding gradient descent. In fact, since curvature depends on second-order derivatives, the corresponding gradient descent equation is a fourth-order PDE.

A related problem is the shape of such an edge connection. Since the regularization term of the TV inpainting model depends only on the total variation, straightedge connections are preferred over curved edge connections, as the former would contain fewer pixels at high-gradient locations.

In fourth-order inpainting methods, boundary conditions on the evolved function and its gradient need to be specified. The boundary condition on the function itself tends to promote continuity of edges near the boundary of the inpainting region. The second boundary condition promotes continuity in the gradient and thus promotes the propagation of information along level lines in a smooth manner.

Low-curvature image simplifiers (LCIS) [25, 27] is a fourth-order inpainting method that provides many fine-scale features that are lost in methods such as H^1 and TV inpainting. The inpainted version follows the evolution

$$u_t = -\nabla \cdot (g(\Delta u)\nabla \Delta u) + \lambda(f - u),$$

where $g(s) = \frac{1}{1+s^2}$ is a "conductance threshold." For high values of Δu, $g(\Delta u)$ is small, so that there is little evolution across high gradients. However, low values of Δu give large values of $g(\Delta u)$, promoting the propagation of information.

LCIS is based on the Perona–Malik equation [23], a second-order PDE often used for image denoising tasks because it propagates information via anisotropic diffusion. Thus, edge sharpness is preserved while noise is smoothed. Unfortunately, the Perona–Malik equation is ill-posed in continuous space, making the model somewhat theoretically unsatisfying. On the other hand, LCIS preserves the anisotropic diffusion properties of the Perona–Malik model while being globally well posed [4, 16] and making more realistic curvature connections.

The results of several inpainting reconstructions are shown in Figs. 6 and 7 for an image of lipid bilayer domains that has been subsampled by factors of 2 and 4 on the vertical axis.

6 Reconstructing Damaged Scan Lines

With many AFM images, there are some artifacts related to the process of raster scanning. After each scan line is complete, flattening is done in order to adjust for effects such as tilt and thermal drift, mostly linear in their effects. In this way, a first-order polynomial is subtracted from each scan line. Flattening generally works well in compensating for tilt and thermal drift, but some errors still occur, particularly relating to streaks.

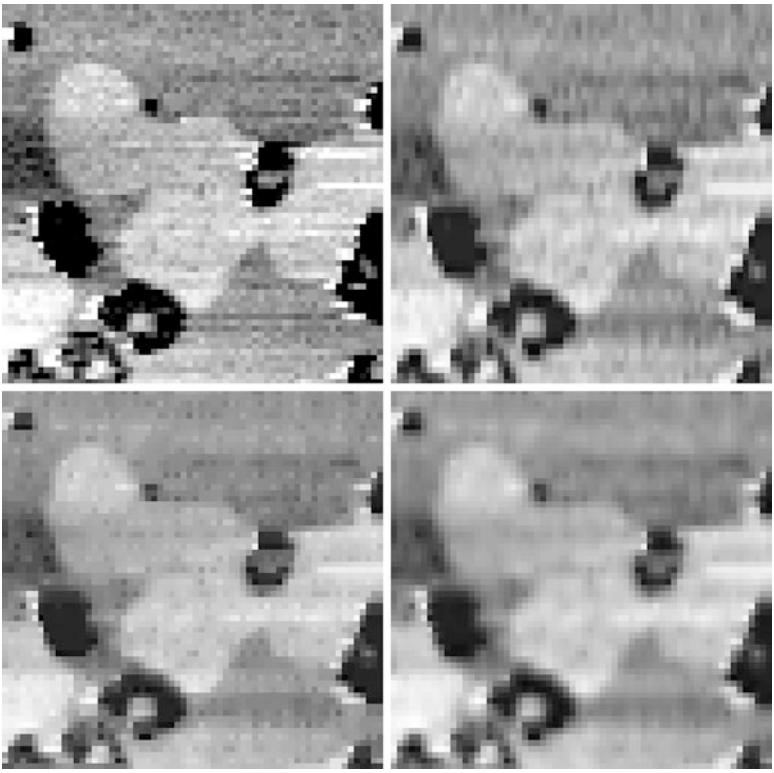


Fig. 6 Inpainting reconstructions from a section of a lipid bilayer image, subsampled by a factor of 2 on the vertical axis. *Left to right*, *top to bottom*: Original image; H1 inpainting; TV inpainting; LCIS inpainting

These streaks can occur for various reasons. In the course of scanning, the probe may be damaged, be temporarily changed from the addition of material from the sample, or be changed when thermal excitations cause jumps between stable states in the governing equations of the probe–sample interactions.

Another source of the streaks is when anomalous features are detected within a given scan line. In general, since the anomalous features are part of the image, it is useful to keep these in a processed image. One problem with this is that due to the flattening process, streaks can occur in scan lines directly following these contaminants. Since flattening is done by subtracting a polynomial function from each scan line, this can result in shifting the data around the feature. One further challenge is designing an automatic detection method that can distinguish between streaks due to features in the sample and streaks due to mistakes in the scanning process.

One of the current standard techniques to deal with these streak artifacts is removal of the entire line, followed by an average of the neighboring scan lines. This

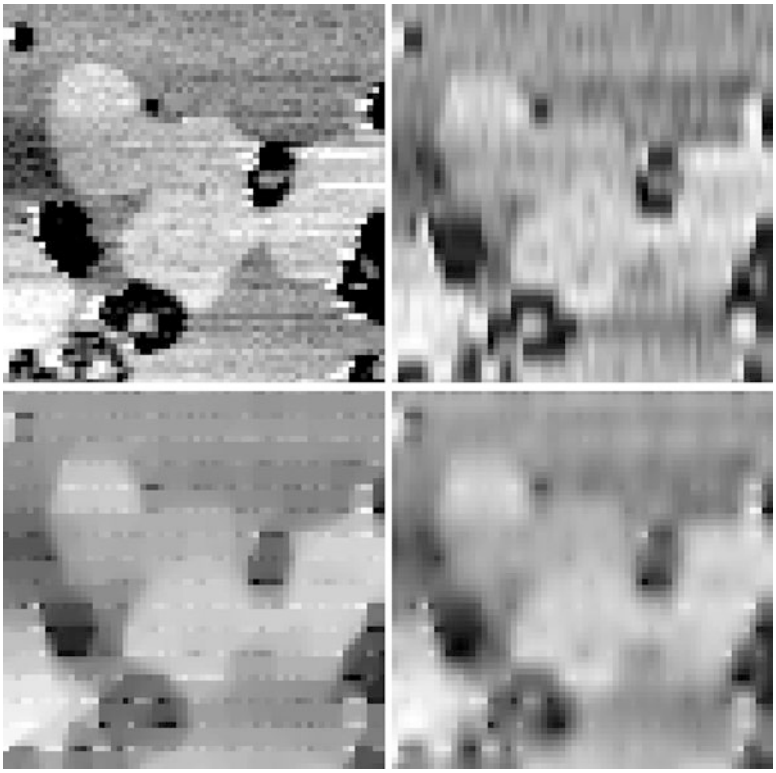


Fig. 7 Inpainting reconstructions from a section of a lipid bilayer image, subsampled by a factor of 4 on the vertical axis. *Left to right*, *top to bottom*: original image; H1 inpainting; TV inpainting; LCIS inpainting

can cause some distortions, particularly near edges. Figure 8 shows the removing of streaks from the image by manual identification of the inaccurate parts, followed by inpainting of the identified regions by LCIS inpainting and by averaging. The results look comparable, since the unknown regions are small. However, inpainting algorithms work slightly better near edges and when there are multiple streaks nearby.

7 Reconstructing Important Features

In this section, various image features are examined more closely and the various algorithms are compared and contrasted. Figure 9 shows zoomed versions of the most common type of edge, one that forms the boundary between two regions of contrasting intensity.

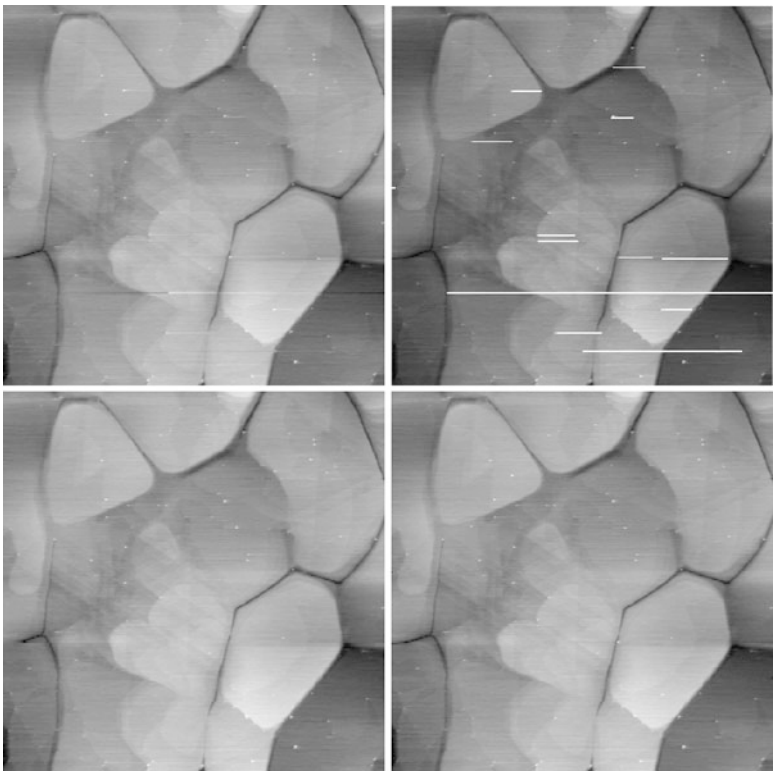


Fig. 8 Inpainting damaged scan lines from the "annealed gold" image. *Top left*: Original image. *Top right*: Manual identification of the damaged areas is displayed in white. *Bottom left*: Averaging of damaged lines. *Bottom right*: Recovery by LCIS inpainting

Since the nonlinear interpolation methods explicitly detect the orientation of edges, they are generally able to reconstruct sharper edges than the linear interpolation methods. Additionally, the staircasing effect is reduced as well. The edge is reconstructed more sharply in the PLI algorithm than in contour stencils at the expense of more artifacts.

Analogously, the TV and LCIS inpainting methods result in better edge reconstructions than the H1 method because information is designed to propagate along edges and not isotropically. These result in edges that are comparable to the nonlinear interpolation techniques in sharpness.

A second type of edge is that of a trench separating two regions of similar intensity. These types of edges are typically much more difficult for inpainting and interpolation methods due to the trench values making up a smaller portion of the neighborhood around an edge. Thus, both interpolation and inpainting methods tend to blur these edges by averaging with the surrounding values. Additionally,

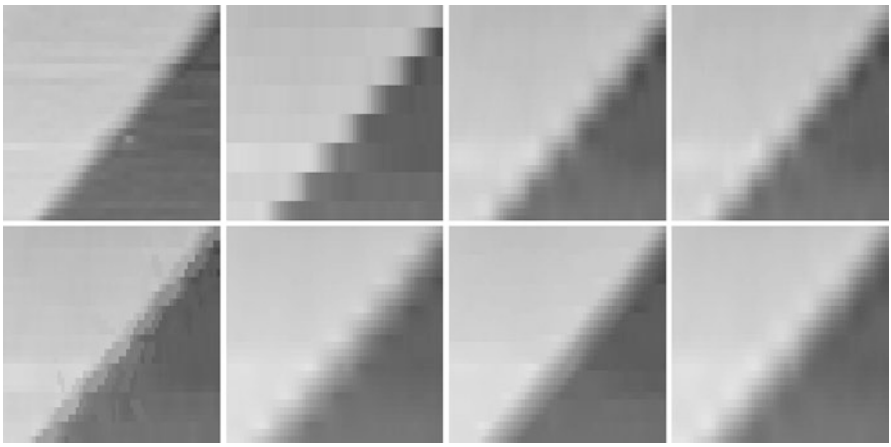


Fig. 9 A high-contrast edge from an image of annealed gold, subsampling by 4 in the vertical direction. *Left to right*, *top to bottom*: Original image; nearest neighbor interpolation; bicubic interpolation; contour stencil interpolation; PLI; H1 inpainting; TV inpainting; LCIS inpainting

Fig. 10 Edges from a trench between two regions of similar intensity from an image of annealed gold, subsampling by 4 in the vertical direction. *Left to right*, *top to bottom*: Original image; nearest neighbor interpolation; bicubic interpolation; contour stencil interpolation; PLI; H1 inpainting; TV inpainting; LCIS inpainting

variational inpainting places an extra penalty on having a double-sided gradient, so that edges often remain disconnected. The PLI algorithm connects this type of edge but at the expense of significant artifacts.

Example reconstructions on this type of edge are in Fig. 10.

Staircasing of diagonal edges occurs with both the interpolation and inpainting methods. Figure 11 shows a comparison between the recovery of several different interpolation and inpainting algorithms. The TV inpainting algorithm performs very well in straightening all edges, due to its tendency to connect edges in straight lines. Most of the algorithms give minimal staircasing for the longer edge of the InP nanowire but do much worse along the shorter side of the nanowire. This is due to the fact that the longer edge lies mostly along the vertical direction. In the space of the recovered image, the points along the edge are separated by a shorter distance and thus more easily connected.

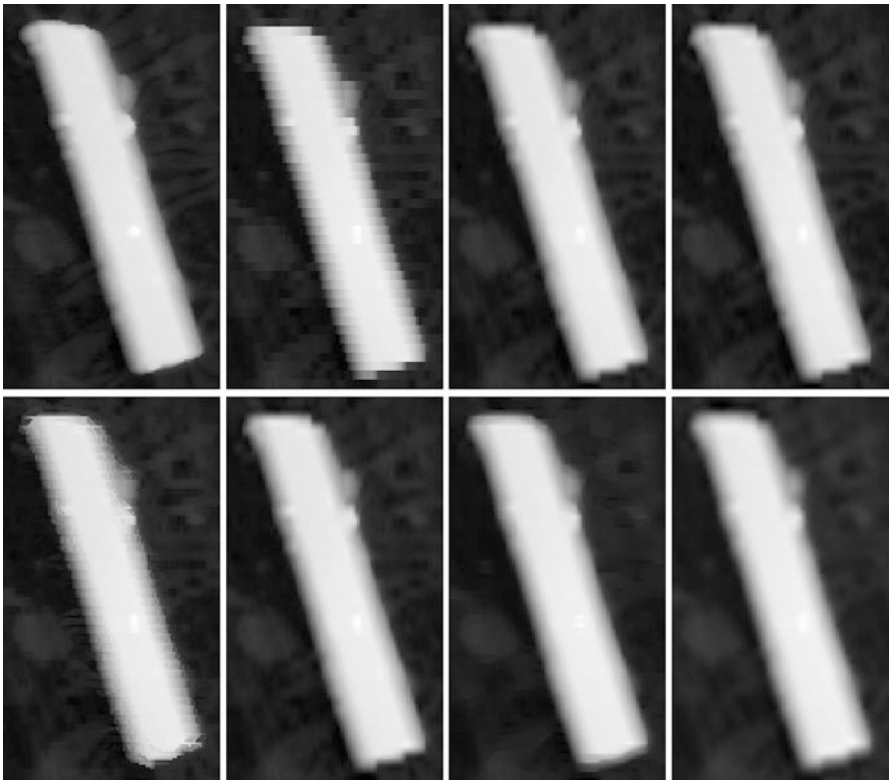


Fig. 11 A comparison of edge quality in the recovery from various inpainting and interpolation algorithms on the InP nanowire image, subsampled by a factor of 4 in one direction. *Left to right*, *top to bottom*: Original image; nearest neighbor interpolation; bicubic interpolation; contour stencil interpolation; PLI; H1 inpainting; TV inpainting; LCIS inpainting

7.1 Difference Images

Difference images can be helpful in determining where the largest errors in the reconstruction take place. They are computed by taking the absolute value of the difference between the reconstructed image and the ground truth image. Not surprisingly, they often occur near edges and noise. The lightest parts often indicate systematic errors in a certain method. Figures 12 and 13 show the difference images formed from the various methods on the InP nanowire and lipid bilayer images.

The various algorithms can also be measured objectively through the calculation of the peak signal-to-noise ratio (PSNR), which uses the root-mean-squared error (RMSE):

$$\text{RMSE} = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} |I_{m,n} - \bar{I}_{m,n}|^2.$$

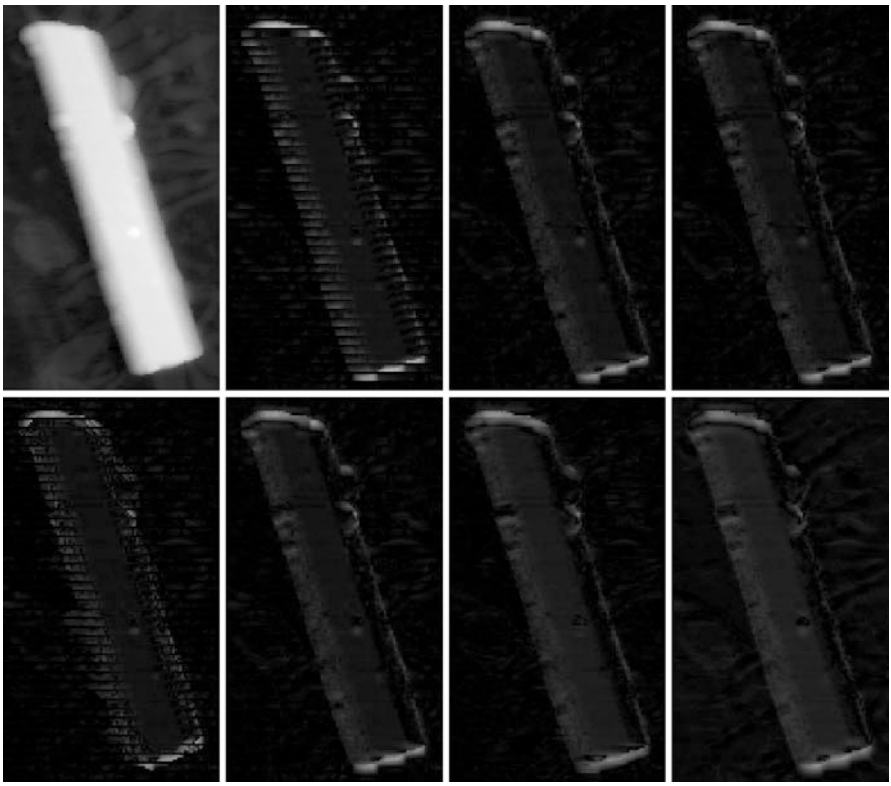


Fig. 12 Difference images of various inpainting and interpolation algorithms on the InP nanowire, subsampled by a factor of 4 in the vertical direction. The original image is shown in the *top left corner* for comparison. *Left to right*, *top to bottom*: Original image; nearest neighbor interpolation; bicubic interpolation; contour stencil interpolation; PLI; H1 inpainting; TV inpainting; LCIS inpainting

Then the PSNR is defined as

$$\mathrm{PSNR} = 20 \cdot \log_{10}\left(\frac{1}{\mathrm{RMSE}}\right).$$

Another method that compares the quality of an image reconstruction is the Mean Structural SIMilarity (MSSIM) index, which measures the similarity between two images by comparing luminance, contrast, and structure [28].

These measures give some indication of the effectiveness of a method, but they can sometimes be misleading as well. If an algorithm performs well near important features such as edges but badly in the background, its performance indicators will be worse than for an algorithm that does well in the background and worse near edges. Yet the former might be preferable in that the features of interest are reconstructed well.

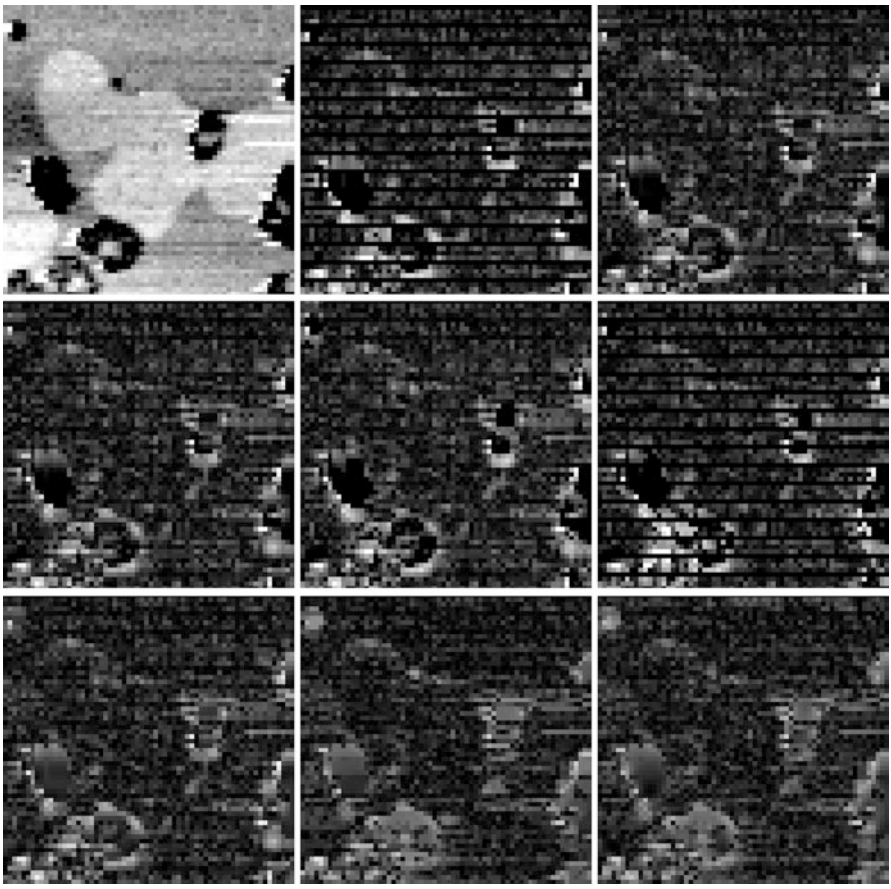


Fig. 13 Difference images of various inpainting and interpolation algorithms on the lipid bilayer image, subsampled by a factor of 4 in the vertical direction. The original image is shown in the *top left corner* for comparison. *Left to right*, *top to bottom*: Original image; nearest neighbor interpolation; Lanczos-3 interpolation; bicubic interpolation; contour stencil interpolation; PLI; H1 inpainting; TV inpainting; LCIS inpainting

A table of the PSNR and SSIM on the various reconstructions of the lipid bilayer image is shown in Table 1, and the same table is shown for the InP nanowire image in Table 2.

Acknowledgments The authors would like to thank Todd Wittman, Jef Huang, and Kevin Thompson for useful conversations on the AFM and inpainting. This research is supported by NSF grant CBET-0940417. Work at the Molecular Foundry was supported by the Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under contract no. DE-AC02-05CH11231. Image of lipid bilayer domain sample provided by Elaine DeMasi.

Table 1 A comparison of peak signal-to-noise ratio (PSNR) and Mean Structural SIMilarity (MSSIM) for various recovery algorithms on the lipid bilayer image

Objective measures of an algorithm's effectiveness		
Method	PSNR	MSSIM
Nearest neighbor	14.8636	0.4494
Bicubic	15.6739	0.4489
Contour stencils	15.1333	0.4396
PLI	14.9638	**0.4728**
H1	16.0301	0.4398
TV	16.5837	0.4356
LCIS	**16.6496**	0.4292

The best-performing algorithm in each column is in bold

Table 2 A comparison of peak signal-to-noise ratio (PSNR) and Mean Structural SIMilarity (MSSIM) for various recovery algorithms on the InP nanowire image

Objective measures of an algorithm's effectiveness		
Method	PSNR	MSSIM
Nearest neighbor	26.3990	0.9104
Bicubic	27.9555	0.9367
Contour stencils	28.0161	0.9364
PLI	**28.5414**	0.9262
H1	27.8737	0.9383
TV	27.9498	**0.9422**
LCIS	27.1007	0.9292

The best-performing algorithm in each column is in bold

References

1. Ando, T., Uchihashi, T., Kodera, N., Yamamoto, D., Miyagi, A., Taniguchi, M., Yamashita, H.: High-speed AFM and nano-visualization of biomolecular processes. In: Pflügers Archive European Journal of Physiology, vol. 456, pp. 211–225. Springer, Berlin (2008)
2. Ashby, P., Lieber, C.: Ultra-sensitive imaging and interfacial analysis of patterned hydrophilic SAM surfaces using energy dissipation chemical force microscopy. J. Am. Chem. Soc. **127**, 6814–6818 (2005)
3. Bertozzi, A., Esedoglu, S., Gillette, A.: Inpainting of binary images using the Cahn-Hilliard equation. IEEE Trans. Image Process. **16**(1), 285–291 (2007)
4. Bertozzi, A.L., Greer, J.B.: Low curvature image simplifiers: global regularity of smooth solutions and Laplacian limiting schemes. Comm. Pure Appl. Math. **57**(6), 764–790 (2004)
5. Binnig, G., Quate, C., Gerber, C.: Atomic force microscope. Phys. Rev. Lett. **56**, 930–933 (1986)
6. Chan, T., Kang, S., Shen, J.: Euler's elastica and curvature-based inpainting. SIAM J. Appl. Math. **63**(2), 564–592 (2002)
7. Chan, T., Shen, J.: Communications on Pure and Applied Mathematics. **58**(5), 579–619 (2005)
8. Chasiotis, I.: Atomic force microscopy in solid mechanics. In: Sharpe, W.N. Jr. (ed.) Springer Handbook of Experimental Solid Mechanics, pp. 409–443. Springer, New York (2008)
9. Darbon, J., Lefebvre, S., Chan, T., Esedoglu, S.: TV optimization and graph-cuts. Proc. Appl. Math. Mech. **7**(1), 1042,303–1042,304 (2007)
10. Dobrosotskaya, J., Bertozzi, A.: A wavelet-laplace variational technique for image deconvolution and inpainting. IEEE Trans. Image Process. **17**(5), 657–663 (2008)
11. Esedoglu, S.: Blind deconvolution of bar code signals. Inverse Probl. **20**, 121–135 (2004)

12. Florin, E., Moy, V., Gaub, H.: Adhesion forces between individual ligand-receptor pairs. Science **264**, 415–417 (1994)
13. Getreuer, P.: Contour stencils for edge-adaptive image interpolation. In: Proc. SPIE, vol. 7246 (2009)
14. Getreuer, P.: Image zooming with contour stencils. In: Proc. SPIE, vol. 7257 (2009)
15. Goldstein, T., Osher, S.: The split Bregman algorithm for L1 regularized problems. SIAM J. Imaging Sci. **2**(2), 323–343 (2009)
16. Greer, J.B., Bertozzi, A.L.: Traveling wave solutions of fourth order PDEs for image processing. SIAM J. Math. Anal. **36**(1), 38–68 (2004)
17. Hansma, P., Cleveland, J., Radmacher, M., Walters, D., Hillner, P., Bezanilla, M., Fritz, M., Vie, D., Hansma, H., Prater, C., Massie, J., Fukunaga, L., Gurley, J., Elings, V.: Tapping mode atomic force microscopy in liquids. Surface Sci. Lett. **64**(13), 1738–1740 (1994)
18. Kodera, N., Yamamoto, D., Ishikawa, R., Ando, T.: Video imaging of walking myosin v by high-speed atomic force microscopy. Nature **468**, 72–76 (2010)
19. Lapshin, R.V.: Feature-oriented scanning methodology for probe microscopy and nanotechnology. Nanotechnol. **15**, 1135–1151 (2004)
20. Martin, Y., Wickramasinghe, H.: Magnetic imaging by "force microscopy" with 1000 å resolution. Appl. Phys. Lett. **50**, 1455–1457 (1987)
21. Morel, J., Yu, G.: Is SIFT scale invariant? Inverse Probl. Imag. **5**(1), 115–136 (2011)
22. Noy, A., Vezenov, D., Kayyem, J., Meade, T., Lieber, C.: Stretching and breaking duplex DNA by chemical force microscopy. Chem. Biol. **4**, 519–527 (1997)
23. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Machine Intell. **12**(7), 629–639 (1990)
24. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D **60**, 259–268 (1992)
25. Schoenlieb, C.B., Bertozzi, A.: Unconditionally stable schemes for higher order inpainting. Comm. Math. Sci. **9**(2), 413–457 (2011)
26. Thévenaz, P., Blu, T., Unser, M.: Image interpolation and resampling. In: Handbook of Medical Imaging. Academic, Orlando (2000)
27. Tumblin, J., Turk, G.: LCIS: A boundary hierarchy for detail-preserving contrast reduction. In: Siggraph, Computer Graphics Proceedings, pp. 83–90 (1999)
28. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
29. Zhong, Q., Innis, D., Kjolle, K.: Fractured polymer/silica fiber surface studied by tapping mode atomic force microscopy. Surface Sci. Lett. **290**, L688–L692 (1993)

Numerical Harmonic Analysis and Diffusions on the 3D-Motion Group

Gregory S. Chirikjian

Abstract Representation theory and harmonic analysis on the group of proper motions in three-dimensional Euclidean space has been applied in a variety of areas ranging from robotics to DNA statistical mechanics. This theory can be used to implement noncommutative convolutions analytically and numerically, as well as to solve diffusion equations over this group. This chapter presents a brief review of this theory together with an emphasis on DNA applications involving diffusions and convolutions. Since representations of this noncompact group are infinite dimensional, quantification of numerical truncation errors is also addressed.

Keywords Noncommutative harmonic analysis • Group representation theory • Infinite-dimensional matrices • Exponential map

1 Introduction

This chapter is a review of applications of the representation theory and harmonic analysis on the group of rigid-body motions in three dimensional space and associated numerical issues. For a more detailed treatment and many other applications, see [4].

The group of proper/special motions in three-dimensional Euclidean space, $SE(3)$, consists of elements that can be thought of as 4×4 homogeneous transformation matrices of the form

$$g = \begin{pmatrix} R & \mathbf{t} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{pmatrix} = \begin{pmatrix} \mathbb{I} & \mathbf{t} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{pmatrix} \begin{pmatrix} R & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{pmatrix}, \tag{1}$$

G.S. Chirikjian (✉)
Department of Mechanical Engineering, Johns Hopkins University, 3400. N. Charles St., Baltimore, MD 21218, USA
e-mail: gregc@jhu.edu

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_17,

where $(\mathbf{t}, R) \in \mathbb{R}^3 \times SO(3)$, $\mathbf{0} \in \mathbb{R}^3$ is the zero vector, $\mathbf{0}^{\mathrm{T}} = [0,0,0]$ is its transpose, and $\mathbb{I}$ is the identity matrix (which in the above context is 3×3). Matrix multiplication, $g_1 \circ g_2$, is the group product, and this is noncommutative ($g_1 \circ g_2 \neq g_2 \circ g_1$). Since $SE(3)$ is the emphasis of this review, the notation $G = SE(3)$ is used for this six-dimensional Lie group. In the applications that will follow, functions of the form $f : G \to \mathbb{R}_{\geq 0}$ will arise. Since G is a connected unimodular matrix Lie group, a bi-invariant integration measure, $\mathrm{d}g$, exists such that

$$\int_G f(h \circ g)\mathrm{d}g = \int_G f(g \circ h)\mathrm{d}g = \int_G f(g^{-1})\mathrm{d}g = \int_G f(g)\mathrm{d}g \tag{2}$$

for any fixed $h \in G$ whenever $\int_G f(g)\mathrm{d}g$ is finite. The explicit form for the integration measure for G is the product $\mathrm{d}g = \mathrm{d}R\,\mathrm{d}\mathbf{t}$ where $\mathrm{d}R$ is the Haar measure for $SO(3)$ and $\mathrm{d}\mathbf{t}$ is the Lebesgue measure for $\mathbb{R}^3$. Explicitly, $\mathrm{d}\mathbf{t} = \mathrm{d}t_1\mathrm{d}t_2\mathrm{d}t_3$, and when ZXZ Euler angles α,β,γ are used to parameterize $SO(3)$, then to within an arbitrary constant, $\mathrm{d}R = \sin\beta\mathrm{d}\alpha\mathrm{d}\beta\mathrm{d}\gamma$. This will be relevant in the applications that follow where the emphasis will be probability density functions where

$$\int_G f(g)\mathrm{d}g = \int_{\mathbb{R}^3}\int_{SO(3)} f(\mathbf{t}, R)\mathrm{d}R\,\mathrm{d}\mathbf{t} = 1.$$

For small translational (rotational) displacements from the identity along (about) the kth coordinate axis in three-dimensional space, the 4×4 homogeneous transforms representing infinitesimal motions look like

$$g_k(\varepsilon) = \mathbb{I} + \varepsilon E_k,$$

where now $\mathbb{I}$ is the 4×4 identity and

$$E_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad E_2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad E_3 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix};$$

$$E_4 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad E_5 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad E_6 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

E_1, E_2, E_3 describe infinitesimal rotations, whereas E_4, E_5, E_6 are infinitesimal translations. These matrices $\{E_i\}$ serve as a basis for a Lie algebra under the matrix commutator $[A, B] = AB - BA$.

1.1 Probability Theory and Harmonic Analysis on Unimodular Lie Groups

Given two probability density functions $f_1(g)$ and $f_2(g)$, their convolution is the probability density function (pdf)

$$(f_1 * f_2)(g) = \int_G f_1(h) f_2(h^{-1} \circ g) \mathrm{d}h. \tag{3}$$

In general $(f_1 * f_2)(g) \neq (f_2 * f_1)(g)$ due to the noncommutative nature of G, but in special cases it is possible to define pdfs that commute under convolution.

A powerful generalization of classical Fourier analysis can be used to compute such convolutions. This theory is built on families of unitary matrix-valued functions of group-valued argument that are parameterized by values λ drawn from a set $\hat{G}$ (called the unitary dual of G) and satisfy the homomorphism property:

$$U(g_1 \circ g_2, \lambda) = U(g_1, \lambda) U(g_2, \lambda). \tag{4}$$

Using $*$ as a superscript to denote the Hermitian conjugate, it follows that

$$\mathbb{I} = U(e, \lambda) = U(g^{-1} \circ g, \lambda) = U(g^{-1}, \lambda) U(g, \lambda),$$

and so

$$U(g^{-1}, \lambda) = (U(g, \lambda))^{-1} = U^*(g, \lambda).$$

Here λ (which is analogous to frequency) indexes the complete set of all IURs.

In this generalized Fourier analysis (called noncommutative harmonic analysis) each $U(g, \lambda)$ is constructed to be *irreducible* in the sense that it is not possible to simultaneously block-diagonalize $U(g, \lambda)$ by the same similarity transformation for all values of g in the group. Such a matrix function $U(g, \lambda)$ is called an *irreducible unitary representation* (IUR). Completeness of a set of representations means that every (reducible) representation can be decomposed into a direct sum of the representations in the set.

Though IURs, $U(g, \lambda)$, are known for many Lie groups as described in [9, 10, 15, 18], the sole emphasis of this chapter is the case when $G = SE(3)$.

Once a complete set of IURs is known for a unimodular Lie group, the Fourier transform of a function on that group can be defined as

$$\hat{f}(\lambda) = \int_G f(g) U(g^{-1}, \lambda) \mathrm{d}g. \tag{5}$$

For unimodular groups such as the motion groups, an inversion formula can be used to recover the original function from all of the Fourier transforms as

$$f(g) = \int_{\hat{G}} \mathrm{tr}[\hat{f}(\lambda)U(g,\lambda)]d(\lambda). \tag{6}$$

In general, the integration measure $d(\lambda)$ on the dual (frequency) space $\hat{G}$ of a unimodular Lie group must be constructed on a case-by-case basis. In the case of $SE(3)$, λ is of the form $\lambda = (p,s) \in \mathbb{R}_{\geq 0} \times \mathbb{Z}$, $d(\lambda) = p^2 dp$, and $\int_{\hat{G}} = \sum_{s\in\mathbb{Z}} \int_{\mathbb{R}_{\geq 0}}$. The exact form of $U(g,\lambda)$ for $SE(3)$ will be explained in explicit detail in Sect. 3.

1.2 Operational Properties

In analogy with classical Fourier analysis, a number of useful operational properties exist for the noncommutative Fourier transform for $G = SE(3)$.

A convolution theorem follows from (4) as

$$\widehat{(f_1 * f_2)}(\lambda) = \hat{f}_2(\lambda)\hat{f}_1(\lambda) \tag{7}$$

and the Parseval/Plancherel formula

$$\int_G |f(g)|^2 \mathrm{d}g = \int_{\hat{G}} ||\hat{f}(\lambda)||^2 d(\lambda) \tag{8}$$

follows from (5) and (6) and the unitary nature of $U(g,\lambda)$. Here $||\cdot||$ is the Hilbert-Schmidt (Frobenius) norm, and $d(\lambda)$ is the integration measure on $\hat{G}$.

If $\mathscr{G}$ is the Lie algebra associated with G, then the exponential map $\exp : \mathscr{G} \to G$ can be used to parameterize a neighborhood in G around the identity. It turns out that in the case of $SE(3)$ this neighborhood is the whole group minus a set of measure zero. Moreover, if $\{E_i\}$ is a basis for $\mathscr{G}$, it is possible to define differential operators akin to directional derivatives as

$$(\tilde{E}_i f)(g) = \frac{\mathrm{d}}{\mathrm{d}t} f(g \circ \exp(tE_i))\Big|_{t=0}. \tag{9}$$

Another useful definition is

$$u(E_i,\lambda) = \frac{\mathrm{d}}{\mathrm{d}t} \left(U(\exp(tE_i),\lambda)\right)|_{t=0}. \tag{10}$$

As a consequence of these definitions, it can be shown that operational properties result as follows. By the definition of the group Fourier transform and operators $\tilde{E}_i$ reviewed earlier,

$$\widehat{(\tilde{E}_i f)}(\lambda) = \int_G \frac{\mathrm{d}}{\mathrm{d}t} f(g \circ \exp(tE_i))\Big|_{t=0} U(g^{-1},\lambda)\,\mathrm{d}g. \tag{11}$$

By performing the change of variables $h = g \circ \exp(tE_i)$ and using the homomorphism property of the representations $U(\cdot, \lambda)$,

$$\widehat{(\tilde{E}_i f)}(\lambda) = \int_G f(h) \left. \frac{\mathrm{d}}{\mathrm{d}t} \left(U(\exp(tE_i) \circ h^{-1}, \lambda) \right) \right|_{t=0} \mathrm{d}h \tag{12}$$

$$= \left(\left. \frac{\mathrm{d}}{\mathrm{d}t} U(\exp(tE_i), \lambda) \right|_{t=0} \right) \int_G f(h)\, U(h^{-1}, \lambda)\, \mathrm{d}h. \tag{13}$$

Then using the definition in (10),

$$\widehat{(\tilde{E}_i f)}(\lambda) = u(E_i, \lambda)\, \hat{f}(\lambda). \tag{14}$$

This is called an *operational property* because the differential operator is converted into an algebraic operation in Fourier space.

For a general treatment of connected unimodular matrix Lie groups and connections with information theory, see [2]. Applications of the motion group in the two-dimensional case are reviewed in [4,6,17,20]. For additional applications in the three-dimensional case, see [3, 16].

1.3 Structure of the Remainder of This Chapter

The remainder of this chapter is structured as follows. Section 2 reviews applications in DNA statistical mechanics in which diffusion equations on G arise. For many stiffness models, the diffusion equations describing the flexibility of double-helical DNA are degenerate. Section 3 reviews the Fourier transform for functions on G in explicit detail, and Sect. 4 then shows how the associated operational properties can be used to solve these diffusion equations. Section 5 reviews a theorem that relates exponentials of Lie algebra representation matrices that produce representations of the associated Lie group. Section 6 builds on this theory by reviewing how the effects of truncation of infinite-dimensional operators lead to numerical approximations of infinite-dimensional representations and diffusion operators.

2 DNA Mechanics

In this section, continuum filament models of DNA are reviewed together with associated diffusion equations that result from subjecting elastic filaments to ambient Brownian motion forcing. These diffusion equations result in probability densities that can be solved for using Fourier analysis on the Euclidean motion group.

2.1 Continuum Filament Model

Let s denote the arc length along an inextensible shearless elastic filament capable of bending and twisting. Such filaments are an appropriate model for double-helical DNA molecules. Attach reference frames along the filament with their local z-axis pointing along the tangent to the filament. When $s = 0$, the attached reference frame is viewed as the identity. Then the relative reference frame attached at another value of s will be $(\mathbf{a}(s), R(s)) \in G$.

One model for the elastic energy in such an elastic filament model of DNA with potential energy per unit length V is

$$E = \int_0^L V(s)\,\mathrm{d}s \text{ where } V = \frac{1}{2}\omega(s)^{\mathrm{T}} B(s)\omega(s) - \mathbf{b}^{\mathrm{T}}(s)\omega(s) + c,$$

where ω is the dual vector of $\Omega = R^{\mathrm{T}}\frac{\mathrm{d}R}{\mathrm{d}s}$, which is the unique vector such that $\omega \times \mathbf{x} = \Omega\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^3$. Here ω can be thought of as angular velocity with respect to the independent arc length variable s (rather than time). This vector is defined relative to the local reference frame attached to the filament at arc length s. Here B is a stiffness matrix and $\mathbf{b}$ is a vector, both of which can depend on s.

With the identity reference frame $e = (\mathbf{0}, \mathbb{I}) \in G$ attached at $s = 0$, the position to any point on the arc-length-parameterized helical backbone will be

$$\mathbf{a}(s) = \int_0^s R(\sigma)\,\mathbf{e}_3\,\mathrm{d}\sigma, \tag{15}$$

where $\mathbf{e}_3 = [0, 0, 1]^{\mathrm{T}}$. When there are no external forces applied to the filament, the lowest energy conformations are given by

$$\frac{\partial V}{\partial \omega} = \mathbf{0} \quad \Longrightarrow \quad \omega(s) = B^{-1}(s)\mathbf{b}(s).$$

In the special case when $\omega(s) = \omega_0$ is constant,

$$\frac{\mathrm{d}R}{\mathrm{d}s} = R\,\Omega_0 \quad \text{with} \quad R(0) = \mathbb{I} \quad \Longrightarrow \quad R(s) = \exp(s\Omega_0). \tag{16}$$

Substituting this into (15) then defines a helix. It can be shown that the corresponding backbone curve is written in closed form as [21]

$$\mathbf{a}(s) = \begin{pmatrix} \frac{n_2}{\|\omega_0\|}(1-\cos\|\omega_0\|s) + n_1 n_3\left(s - \frac{\sin\|\omega_0\|s}{\|\omega_0\|}\right) \\ \frac{n_1}{\|\omega_0\|}(\cos\|\omega_0\|s - 1) + n_2 n_3\left(s - \frac{\sin\|\omega_0\|s}{\|\omega_0\|}\right) \\ s - (n_1^2 + n_2^2)\left(s - \frac{\sin\|\omega_0\|s}{\|\omega_0\|}\right) \end{pmatrix}, \tag{17}$$

where $\mathbf{n} = \omega_0/\|\omega_0\|$.

If it happens that $B = B(s)$ and $\mathbf{b} = \mathbf{b}(s)$ vary with s in such a way that $\omega(s)$ is not constant, then the equilibrium conformation is still described by $g_0(s) = (\mathbf{a}(s), R(s)) \in G$, but this pair will generally not have a closed-form solution like the one given above in (16) and (17) .

2.2 Modeling the Effects of Brownian Motion

Consider the equilibrium statistics of a stochastically forced elastic filament. Let the evolution of the probability density of relative pose of reference frames attached to a stochastically forced elastic filament at values of curve parameter 0 and s be denoted as $f(g;0,s)$. Since it is a probability density, by definition

$$\int_G f(g;0,s)\,\mathrm{d}g = 1. \tag{18}$$

Clearly $f(g,s) \doteq f(g;0,s)$ must be related in some way to the equilibrium shape of the filament, its stiffness, and the strength of the Brownian motion forcing from the ambient solvent. And the strength of this noise should be related in some way to the temperature. In fact, since $f(g;0,s)$ is the function describing the distribution of poses for a filament at equilibrium, it can be represented exactly as a path integral [12], or equivalently, as a diffusion equation [5]:

$$\frac{\partial f}{\partial s} = \frac{1}{2}\sum_{k,l=1}^{6} D_{lk}(s)\,\tilde{E}_l\tilde{E}_k^r f - \sum_{l=1}^{6}\xi_l(s)\,\tilde{E}_l f \tag{19}$$

subject to the initial conditions

$$f(g;0,0) = \delta(g).$$

In the case of the inextensible model described above,

$$D(s) = 2k_B T\cdot B^{-1}(s) \oplus \mathbb{O} \in \mathbb{R}^{6\times 6} \qquad \text{and} \qquad \xi(s) = [\omega^{\mathrm{T}}(s), \mathbf{e}_3^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^6,$$

where k_B is the Boltzmann constant and T is temperature in degrees Kelvin, $\oplus$ is the direct sum of matrices, and $\mathbb{O}$ is the 3×3 zero matrix.

Under the extreme condition that $T \to 0$, no diffusion would take place, and $f(g;,0,s) \to \delta(g_0^{-1}(s) \circ g)$. For the biologically relevant case ($T \approx 300$), (19) can be solved using the harmonic analysis approach in [5, 21, 22]. If we make the shorthand notation $f_{s_1,s_2}(g) = f(g; s_1, s_2)$, then it will always be the case for $s_1 < s < s_2$ that

$$f_{s_1,s_2}(g) = (f_{s_1,s} * f_{s,s_2})(g) = \int_G f_{s_1,s}(h)\, f_{s,s_2}(h^{-1} \circ g)\, \mathrm{d}h. \tag{20}$$

This is the convolution of two pose distributions. Here h is a dummy variable of integration, and dh is the bi-invariant integration measure for $SE(3)$. While (20) will always hold for semiflexible phantom chains, in the homogeneous case when B and ω are independent of s, there is the additional convenient properties that

$$f(g; s_1, s_2) = f(g; 0, s_2 - s_1) \text{ and } f(g; s_2, s_1) = f(g^{-1}, s_1, s_2). \tag{21}$$

The first of these says that for a uniform chain the pose distribution only depends on the difference of arc length along the chain. The second provides a relationship between the pose distribution for a uniform chain resulting from taking the frame at s_1 to be fixed at the identity and recording the poses visited by s_2 and the distribution of frames that results when s_2 is fixed at the identity.

The description above is for a phantom model. That is, neither of these nor (20) will hold when excluded-volume interactions due to distal points in arc length coming into spatial proximity are taken into account. Such effects can be built in as explained in [1].

2.3 Solving Diffusion Equations on the Euclidean Group

The true benefit of the group-theoretic approach is realized when observing that in coordinate form, (19) is expressed as pages of complicated-looking (but essentially elementary) mathematical expressions. In contrast, it is possible to write out the solution very simply using results from group theory. One numerical approach that works well for dilute solutions of DNA of lengths in the range of 1/2–2 persistence lengths (60–300 base pairs at 300 K) is based on the group Fourier transform for $SE(3)$. The reason why this approach is most appropriate for this regime is that DNA of this length is flexible enough for Fourier methods (which work better for more spread out distributions than for highly focused ones) to be applicable, and it is short enough that the effects of self-contact can be neglected.

Applying (14) to (19) gives

$$\frac{\partial \hat{f}(\lambda, s)}{\partial s} = \mathscr{B}(\lambda, s)\, \hat{f}(\lambda, s) \tag{22}$$

where

$$\mathscr{B}(\lambda, s) = \frac{1}{2} \sum_{i,j=1}^{6} D_{ij}(s)\, u(E_i, \lambda)\, u(E_j, \lambda) \; - \; \sum_{k=1}^{6} \xi_k(s)\, u(E_k, \lambda).$$

In the case of a referential configuration that is helical and stiffness parameters that are uniform (and therefore independent of s), then $\mathscr{B}(\lambda, s) = \mathscr{B}_0(\lambda)$ is constant with respect to arc length, s, and the solution can be written in Fourier space as

$$\hat{f}(\lambda, s) = \exp(s\, \mathscr{B}_0(\lambda)), \tag{23}$$

and the inversion formula can be used to recover $f(g, s)$. The details of this procedure have been discussed in a number of the author's papers, together with the use of the convolution theorem for group Fourier transforms to "stitch together" the statistics of several segments of DNA connected by joints and/or kinks [21, 22]. In the case when $\mathscr{B}(\lambda; s)$ is not independent of s, the differential equation in (22), which is an ODE for each fixed value of λ, can be solved either as a product of exponentials or by numerical integration.

3 Fourier Analysis on the 3D-Motion Group

This section reviews the construction of IURs for $G = SE(3)$ together with the corresponding harmonic analysis and operational calculus. This is done at the explicit level in order to be a useful tool in solving the sorts of problems described in the previous section. The presentation in this section follows that in [6], which in turn followed [14, 15].

3.1 Induced Representations for SE(3)

Let the pair $(\mathbf{a}, A)$ denote a translation/position $\mathbf{a} \in \mathbb{R}^3$ and a rotation/orientation $A \in SO(3)$. Two such pairs, when viewed as elements of $SE(3)$, satisfy the group operation $(\mathbf{a}_1, A_1) \circ (\mathbf{a}_2, A_2) = (A_1 \mathbf{a}_2 + \mathbf{a}_1, A_1 A_2)$. Operators for the IURs of $SE(3)$ that act on functions on the sphere can be written in the form

$$(U(\mathbf{a}, A; p, s)\varphi)(\mathbf{u}) = \mathrm{e}^{-\mathrm{i}p\mathbf{u}\cdot\mathbf{a}}\, \Delta_s(R_{\mathbf{u}}^{-1}\, A\, R_{A^{-1}\mathbf{u}})\, \varphi(A^{-1}\mathbf{u}), \tag{24}$$

where $\mathbf{p} = p\mathbf{u}$ and $\mathbf{u}$ is a unit vector. Here $\varphi(\cdot)$ is defined on the unit sphere and

$$\Delta_s : \; \phi \; \rightarrow \mathrm{e}^{\mathrm{i}s\phi} \; , \; 0 \leq \phi \leq 2\pi$$

for $s = 0, \pm 1, \pm 2, \ldots$ (which is not to be confused with the earlier usage of s as arc length; here the pair $(p, s) = \lambda$ and they have nothing to do with arc length).

The irreducible representations of the motion group can be built on spaces $\varphi(\mathbf{p}) \in \mathscr{L}^2(S_p)$, with the inner product defined as

$$(\varphi_1, \varphi_2) = \int_{\Theta=0}^{\pi} \int_{\Phi=0}^{2\pi} \overline{\varphi_1(\mathbf{p})}\, \varphi_2(\mathbf{p}) \sin\Theta \, \mathrm{d}\Theta \, \mathrm{d}\Phi \ , \tag{25}$$

where $\mathbf{p} = (p \sin\Theta \cos\Phi,\ p \sin\Theta \sin\Phi,\ p \cos\Theta)$ and $p > 0$, $0 \leq \Theta \leq \pi$, $0 \leq \Phi \leq 2\pi$.

3.2 Matrix Elements of IURs

To obtain the matrix elements of the unitary representations, we use the group property

$$U(\mathbf{a}, A; p, s) = U(\mathbf{a}, \mathbb{I}; p, s) \cdot U(\mathbf{0}, A; p, s). \tag{26}$$

This can be written as [14, 15]

$$U_{l',m';l,m}(\mathbf{a}, A; p, s) = \sum_{j=-l}^{l} [l', m' \mid p, s \mid l, j](\mathbf{a})\, U_{j\,m}(A, l) \tag{27}$$

by using (26), where $U_{j\,m}(A, l)$ are the matrix elements of IURs for $SO(3)$ given in [2, 6]. The translational part of the matrix elements $U_{l',m';l,m}(\mathbf{a}, A; p, s)$ can be written in closed form as [14, 15][1]

$$\begin{aligned}
&[l', m' \mid p, s \mid l, m](\mathbf{a}) \\
&= (4\pi)^{1/2} \sum_{k=|l'-l|}^{l'+l} i^k \sqrt{\frac{(2l'+1)(2k+1)}{(2l+1)}}\, j_k(p\,a)\, C(k, 0; l', s \mid l, s) \\
&\quad \cdot C(k, m-m'; l', m' \mid l, m)\, Y_k^{m-m'}(\mathbf{u}(\phi, \theta)) \quad ,
\end{aligned} \tag{28}$$

where θ, ϕ are polar and azimuthal angles of the translation vector $\mathbf{a} = a \cdot \mathbf{u}(\phi, \theta)$, $Y_k^m(\mathbf{u})$ are the spherical harmonics defined according to the Condon and Shortley convention [8], and $C(k, m-m'; l', m' \mid l, m)$ are Clebsch–Gordan coefficients (see [11]).

The matrix elements of the transform are given in terms of matrix elements (27) as

[1] Here $j_k(\cdot)$ is the classical kth-order spherical Bessel function.

$$\hat{f}_{l',m';l,m}(p,s) \;=\; \int_{SE(3)} f(\mathbf{a},A)\,\overline{U_{l,m;l',m'}(\mathbf{a},A;p,s)}\,\mathrm{d}A\,\mathrm{d}\mathbf{a}, \tag{29}$$

where we have used the unitarity property.

The inverse Fourier transform is defined by

$$f(g) \;=\; \mathscr{F}^{-1}(\hat{f}) \;=\; \frac{1}{2\pi^2}\sum_{s=-\infty}^{\infty}\int_0^{\infty} \mathrm{trace}(\hat{f}(p,s)U(g;p,s))\,p^2\,\mathrm{d}p\,. \tag{30}$$

Explicitly

$$f(\mathbf{a},A) \;=\; \frac{1}{2}\pi^2 \sum_{s=-\infty}^{\infty}\sum_{l'=|s|}^{\infty}\sum_{l=|s|}^{\infty}\sum_{m'=-l'}^{l'}\sum_{m=-l}^{l}\int_0^{\infty} p^2\,\mathrm{d}p\;\hat{f}_{l,m;l',m'}(p,s)U_{l',m';l,m}(\mathbf{a},A;p,s)\;. \tag{31}$$

Representations (24), which can be viewed as infinite-dimensional matrices denoted as $U(g;p,s)$ with elements (27), satisfy the homomorphism properties

$$U(g_1\circ g_2;p,s) \;=\; U(g_1;p,s)\cdot U(g_2;p,s)\,,$$

where $\circ$ is the group operation.

4 Operational Matrices

The Lie algebra representation matrices

$$u(E_k;p,s) \;=\; \frac{\mathrm{d}}{\mathrm{d}t}\,(U(\exp(tE_k);p,s)\Big|_{t=0} \tag{32}$$

play an important role in operational properties. They can be derived explicitly as infinite-dimensional matrices by evaluating the expressions for $U(g;p,s)$ given in the previous section. Detailed calculations are given in [4, 19]. The result is that the matrix elements of each representation $u(E_k;p,s)$ for the Lie algebra $\mathscr{G} = se(3)$ can be explicitly written as

$$u_{l',m';l,m}(E_1;p,s) = -\frac{i}{2}c^l_{-m}\delta_{l,l'}\delta_{m'+1,m} - \frac{i}{2}c^l_m\delta_{l,l'}\delta_{m'-1,m} \tag{33}$$

$$u_{l',m';l,m}(E_2;p,s) = +\frac{1}{2}c^l_{-m}\delta_{l,l'}\delta_{m'+1,m} - \frac{1}{2}c^l_m\delta_{l,l'}\delta_{m'-1,m} \tag{34}$$

$$u_{l',m';l,m}(E_2;p,s) = -im\delta_{l,l'}\delta_{m',m} \tag{35}$$

$$
\begin{aligned}
&u_{l',m';l,m}(E_4;\,p,s)\\
&\quad=-\frac{ip}{2}\gamma^s_{l',-m'}\delta_{m',m+1}\delta_{l'-1,l}+\frac{ip}{2}\lambda^s_{l,m}\delta_{m',m+1}\delta_{l',l}+\frac{ip}{2}\gamma^s_{l,m}\delta_{m',m+1}\delta_{l'+1,l}\\
&\qquad+\frac{ip}{2}\gamma^s_{l',m'}\delta_{m',m-1}\delta_{l'-1,l}+\frac{ip}{2}\lambda^s_{l,-m}\delta_{m',m-1}\delta_{l',l}-\frac{ip}{2}\gamma^s_{l,-m}\delta_{m',m-1}\delta_{l'+1,l}
\end{aligned}
\tag{36}
$$

$$
\begin{aligned}
&u_{l',m';l,m}(E_5;\,p,s)\\
&\quad=-\frac{p}{2}\gamma^s_{l',-m'}\delta_{m',m+1}\delta_{l'-1,l}+\frac{p}{2}\lambda^s_{l,m}\delta_{m',m+1}\delta_{l',l}+\frac{p}{2}\gamma^s_{l,m}\delta_{m',m+1}\delta_{l'+1,l}\\
&\qquad-\frac{p}{2}\gamma^s_{l',m'}\delta_{m',m-1}\delta_{l'-1,l}-\frac{p}{2}\lambda^s_{l,-m}\delta_{m',m-1}\delta_{l',l}+\frac{p}{2}\gamma^s_{l,-m}\delta_{m',m-1}\delta_{l'+1,l}
\end{aligned}
\tag{37}
$$

and

$$
u_{l',m';l,m}(E_6;\,p,s)=ip\kappa^s_{l',m'}\delta_{m',m}\delta_{l'-1,l}+ip\frac{sm}{l(l+1)}\delta_{m',m}\delta_{l',l}+ip\kappa^s_{l,m}\delta_{m',m}\delta_{l'+1,l},
\tag{38}
$$

where

$$
\gamma^s_{l,m}=\left(\frac{(l^2-s^2)(l-m)(l-m-1)}{l^2(2l-1)(2l+1)}\right)^{1/2}
$$

$$
\lambda^s_{l,m}=\frac{s\sqrt{(l-m)(l+m+1)}}{l(l+1)}
$$

and

$$
\kappa^s_{l,m}=\left(\frac{(l^2-m^2)(l^2-s^2)}{l^2(2l-1)(2l+1)}\right)^{1/2}.
$$

Whereas the matrix elements $u_{l',m';l,m}(E_k;\,p,s)$ are obtained by evaluating the matrix elements $U_{l',m';l,m}(g;\,p,s)$ at special values of $g=\exp(tE_k)$, differentiating with respect to t, and evaluating the result at $t=0$, in the following section, it will be shown how to go in the opposite direction. Namely, the matrix $U(g;\,p,s)$ can be obtained by exponentiating a linear combination of matrices $u(E_k;\,p,s)$.

5 The Exponential Map and Representations

Given a matrix representation $U(g,\lambda)$ of G, a representation of the Lie algebra $\mathscr{G}$, $u(X,\lambda)$, results from (10). This is called a representation of $\mathscr{G}$ because it inherits the Lie bracket from the Lie algebra:

$$
u([X,Y],\lambda)=[u(X,\lambda),u(Y,\lambda)],
$$

where $X, Y \in \mathscr{G}$ and $[\cdot,\cdot]$ is the matrix commutator. Below, an explicit relationship between $u(X,\lambda)$ and $U(\exp X,\lambda)$ is reviewed which will be important when discussing truncation and numerical approximations of infinite-dimensional representations.

Theorem 1 ([9, 18]). *Given a connected unimodular n-dimensional Lie group G, associated Lie algebra $\mathscr{G}$, and $u(E_i,\lambda)$ as defined in (10), then*

$$U(\exp(tE_i);\lambda) = \exp[t\, u(E_i,\lambda)], \tag{39}$$

where $E_i \in \mathscr{G}$. Furthermore, if the matrix exponential parameterization

$$g(x_1,\ldots,x_n) = \exp\left(\sum_{i=1}^{n} x_i E_i\right) \tag{40}$$

is surjective, then when the $u(E_i,\lambda)$s are not simultaneously block-diagonalizable by some matrix $S(\lambda)$,

$$U(g;\lambda) = \exp\left(\sum_{i=1}^{n} x_i\, u(E_i,\lambda)\right) \tag{41}$$

is an irreducible representation for all $g \in G$.

Proof. For the exponential parameterization (40),

$$g(tx_1,\ldots,tx_n)\circ g(\tau x_1,\ldots,\tau x_n) = g((t+\tau)x_1,\ldots,(t+\tau)x_n)$$

for all $t,\tau \in \mathbb{R}$, i.e., the set of all $g(tx_1,\ldots,tx_n)$ forms a one-dimensional (Abelian) subgroup of G for fixed values of x_i. From the definition of a representation it follows that

$$\begin{aligned} U(g((t+\tau)x_1,\ldots,(t+\tau)x_n),\lambda) &= U(g(tx_1,\ldots,tx_n),\lambda)\, U(g(\tau x_1,\ldots,\tau x_n),\lambda) \\ &= U(g(\tau x_1,\ldots,\tau x_n),\lambda)\, U(g(tx_1,\ldots,tx_n),\lambda). \end{aligned}$$

Then differentiating the above expression with respect to τ and setting $\tau = 0$, and using the definition

$$\tilde{U}(x_1,\ldots,x_n;\lambda) = U(g(x_1,\ldots,x_n),\lambda)$$

gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{U}(tx_1,\ldots,tx_n;\lambda) = \frac{\mathrm{d}}{\mathrm{d}\tau}\tilde{U}(\tau x_1,\ldots,\tau x_n;\lambda)\Big|_{\tau=0}\tilde{U}(tx_1,\ldots,tx_n;\lambda).$$

But since infinitesimal operations commute, it follows from (10) that

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\tilde{U}(\tau x_1, \ldots, \tau x_n; \lambda)\Big|_{\tau=0} = \sum_{i=1}^{n} x_i\, u(E_i, \lambda).$$

Therefore the matrix differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{U}(t x_1, \ldots, t x_n; \lambda) = \left(\sum_{i=1}^{n} x_i\, u(E_i, \lambda)\right) \tilde{U}(t x_1, \ldots, t x_n; \lambda)$$

results, subject to the initial conditions

$$\tilde{U}(0, \ldots, 0; \lambda) = \mathbb{I}.$$

The solution is therefore

$$\tilde{U}(t x_1, \ldots, t x_n; \lambda) = \exp\left(t \sum_{i=1}^{n} x_i\, u(E_i, \lambda)\right).$$

Evaluating at $t = 1$, (41) results, and setting all $x_j = 0$ except x_i produces (39). The irreducibility of these representations follows from the assumed properties of $u(X, \lambda) = \sum_{i=1}^{n} x_i\, u(E_i, \lambda)$ and the fact that

$$\exp(S(\lambda)\, u(X, \lambda)\, [S(\lambda)]^{-1}) = S(\lambda)\, \exp(u(X, \lambda))\, [S(\lambda)]^{-1}.$$

The above works well for compact Lie groups, in which the representation matrices are finite. But for noncompact Lie groups such as $SE(3)$, the matrices are infinite dimensional. The next section addresses the errors that result from truncating infinite-dimensional matrices before exponentiating. This is relevant both to the previous section, and in the numerical evaluation of exponentiated diffusion matrices such as those arise in modeling DNA.

6 Bounding the Effects of Truncation in Infinite-Dimensional Systems

The matrices $U(g, \lambda)$ and $u(E_i, \lambda)$ are both infinite dimensional for the motion group of three-dimensional Euclidean space. In order to use the theory described in previous sections in a practical numerical framework, some sort of truncation is required. In this section, the effects of truncating the expressions for $U(g, \lambda)$ as well as the effects of truncating $u(E_i, \lambda)$, and the diffusion operator $\mathscr{B}_0$ in (23), before exponentiating are analyzed. The analysis reviewed here follows that in [7, 13].

6.1 How Close Is a Truncated Unitary Matrix to Being Unitary?

In the theory of infinite-dimensional matrices, it is common to consider the effects of truncation by partitioning the matrix into finite and infinite parts. In the context of IURs for $SE(3)$, the indices of the matrices extend to infinity in both directions, and a finite section is taken as the middle block.

Consider a $(2T_1 + 1)$-dimensional unitary matrix

$$U = \begin{pmatrix} U_{11} & U_{12} & U_{13} \\ U_{21} & U_{22} & U_{23} \\ U_{31} & U_{32} & U_{33} \end{pmatrix},$$

This can be considered to be either an infinite-dimensional IUR of G with U_{22} being a finite block or the above U could be the exponential of a truncated infinite-dimensional Lie algebra representation. Since the matrices $u(E_i, \lambda)$ are skew-Hermitian, so too are finite blocks centered on the middle element. In the former case T_1 would be infinite, whereas in the latter it would be finite.

In either case, U is unitary. Let $[U]_{T_2} = U_{22}$ be a $(2T_2 + 1)$-dimensional square block of this matrix where of course $T_2 < T_1$. Here we address how close $[U]_{T_2}$ is to being unitary. This is relevant because truncating unitary matrix representations of groups or exponentiating truncated Lie algebra representations are two ways to numerically approximate the infinite-dimensional quantities of interest.

Since U is unitary $UU^* = \mathbb{I}_{2T_1+1}$. The square of the Frobenius norm then gives $\|UU^*\|^2 = \|\mathbb{I}_{2T_1+1}\|^2 = 2T_1 + 1$. If we consider the block multiplication of the middle row of blocks in U with the corresponding collies in U^*, this gives

$$U_{21}U_{12}^* + U_{22}U_{22}^* + U_{23}U_{32}^* = \mathbb{I}_{2T_2+1}.$$

Taking the norm gives

$$\begin{aligned} &\|U_{21}U_{12}^* + U_{22}U_{22}^* + U_{23}U_{32}^*\| = \|\mathbb{I}_{2T_2+1}\| \\ &= \sqrt{2T_2+1} \leq \|U_{21}U_{12}^* + U_{23}U_{32}^*\| + \|U_{22}U_{22}^*\|. \end{aligned}$$

In other words,

$$\| [U]_{T_2}[U]_{T_2}^* \| \geq \sqrt{2T_2+1} - \|U_{21}U_{12}^* + U_{23}U_{32}^*\|. \tag{42}$$

On the other hand, since all of the rows and columns of U are unit under the Frobenius norm, it follows that

$$\begin{aligned} 2T_1 + 1 = \|U\|^2 &= 2(T_1 - T_2) + \|U_{12}\|^2 + \|U_{22}\|^2 + \|U_{32}\|^2 \\ &= 2(T_1 - T_2) + \|U_{21}\|^2 + \|U_{22}\|^2 + \|U_{23}\|^2. \end{aligned}$$

Since the norm of any matrix is always greater than or equal to zero, it follows that

$$\|U_{12}\|^2 + \|U_{32}\|^2 = \|U_{21}\|^2 + \|U_{23}\|^2 \leq 2T_2 + 1.$$

and

$$\|U_{22}\|^2 \leq 2T_2 + 1$$

or

$$\| [U]_{T_2} \| \leq \sqrt{2T_2 + 1}. \tag{43}$$

Using (42) and (43), it follows that

$$\frac{\| [U]_{T_2}[U]_{T_2}^* \|}{\| [U]_{T_2} \|} \geq \frac{\| [U]_{T_2}[U]_{T_2}^* \|}{\sqrt{2T_2 + 1}} \geq 1 - \frac{\|U_{21}U_{12}^* + U_{23}U_{32}^*\|}{\sqrt{2T_2 + 1}} \geq 0.$$

where

$$\begin{aligned}\|U_{21}U_{12}^* + U_{23}U_{32}^*\| &\leq \|U_{21}U_{12}^*\| + \|U_{23}U_{32}^*\| \leq \|U_{21}\|\|U_{12}^*\| + \|U_{23}\|\|U_{32}^*\| \\ &\leq (\|U_{21}\|^2 + \|U_{23}\|^2)^{\frac{1}{2}}(\|U_{12}\|^2 + \|U_{32}\|^2)^{\frac{1}{2}} \\ &= \|U_{21}\|^2 + \|U_{23}\|^2 = \|U_{12}\|^2 + \|U_{32}\|^2.\end{aligned}$$

This indicates that there is a lower bound on the achievable accuracy when using a finite section $[U]_{T_2}$ in place of an infinite-dimensional unitary representation.

6.2 *Using Stability Theory to Bound the Difference of Systems*

Here systems theory is used to bound the difference between the matrix exponential of an infinite-dimensional matrix and its truncated version. This is applicable both to finding finite-dimensional approximations of (23) and (41).

In error analysis, it is useful to consider the general problem of obtaining expressions that bound the difference of two systems. For example, given

$$\dot{X}_1 = AX_1 \text{ and } \dot{X}_2 = (A + B(t))X_2$$

with initial conditions $X_i(0)$, what can be said about $\|X_1(t) - X_2(t)\|$? Define $\Delta(t) = X_2(t) - X_1(t)$ and $\Delta_0 = \Delta(0) = X_2(0) - X_1(0)$. Suppose that the two original systems were solved. Then, we can define two new systems that both result from subtracting the first of the original systems from the second:

$$\dot{\Delta} = A\Delta + B(t)X_2$$

or

$$\dot{\Delta} = (A + B(t))\Delta + B(t)X_1.$$

Then, using a basic result from linear systems theory, the first of the above can be written as

$$\Delta(t) = \exp(At)\Delta_0 + \int_0^t \exp(A(t-\tau))B(\tau)X_2(\tau)\mathrm{d}\tau. \tag{44}$$

In the case when $B(t) = B_0$ is constant, we can also write

$$\Delta(t) = \exp((A + B_0)t)\Delta_0 + \int_0^t \exp((A + B_0)(t-\tau))B_0X_1(\tau)\mathrm{d}\tau.$$

Then norms can be applied.

Note that we will be comparing the "inner parts" of truncated and nontruncated systems. $B = B_0$ will correspond to the border that surrounds a truncated infinite-dimensional matrix. And we will not be concerned so much with $\|\Delta\|$ as with $\|\,[\Delta]_T\,\|$, which will provide a tighter estimate of the difference that we seek.

6.2.1 Bounds on Norm of Truncated Differences

Consider the following infinite-dimensional system

$$\begin{pmatrix} \dot{X}_{11} & \dot{X}_{12} & \dot{X}_{13} \\ \dot{X}_{21} & \dot{X}_{22} & \dot{X}_{23} \\ \dot{X}_{31} & \dot{X}_{32} & \dot{X}_{33} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & \mathbb{O} \\ A_{21} & [A]_T & A_{23} \\ \mathbb{O} & A_{32} & A_{33} \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{pmatrix}$$

and the corresponding finite section

$$[\dot{X}]_T = [A]_T[X]_T \longleftrightarrow \begin{pmatrix} \mathbb{O} & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & [\dot{X}]_T & \mathbb{O} \\ \mathbb{O} & \mathbb{O} & \mathbb{O} \end{pmatrix} = \begin{pmatrix} A_{11} & \mathbb{O} & \mathbb{O} \\ A_{21} & [A]_T & A_{23} \\ \mathbb{O} & \mathbb{O} & A_{33} \end{pmatrix} \begin{pmatrix} \mathbb{O} & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & [X]_T & \mathbb{O} \\ \mathbb{O} & \mathbb{O} & \mathbb{O} \end{pmatrix}.$$

The initial conditions are $X(0) = \mathbb{I}$ and $[X]_T(0) = \mathbb{I}_T$ (where $\mathbb{I}_T$ is shorthand for $[\mathbb{I}]_T$). The solutions to these systems are $X = \exp At$ and $[X]_T = \exp([A]_T t)$.

We are interested in knowing something about

$$\|X_{22} - [X]_T\| = \|\,[\exp At]_T - \exp([A]_T t)\,\|.$$

Let X_T denote $\mathbb{O} \oplus [X]_T \oplus \mathbb{O}$. Then by defining $\Delta = X - X_T$, it follows that $\|X_{22} - [X]_T\| = \|\mathbb{I}_T\,\Delta\,\mathbb{I}_T\|$. Using the results from the previous section, we can write

$$\|X_{22} - [X]_T\| = \|\mathbb{I}_T\, \Delta\, \mathbb{I}_T\| = \|\mathbb{I}_T\, \exp(At)\, \Delta_0\, \mathbb{I}_T + \int_0^t \mathbb{I}_T\, \exp(A(t-\tau))\, B\, X_T(\tau)\, \mathbb{I}_T\, \mathrm{d}\tau\|, \tag{45}$$

where

$$B = \begin{pmatrix} \mathbb{O} & A_{12} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & A_{32} & \mathbb{O} \end{pmatrix}.$$

Moreover, we can write

$$\begin{aligned} \|X_{22} - [X]_T\| &= \left\| \int_0^t \mathbb{I}_T \exp(A(t-\tau)) \begin{pmatrix} A_{12}[X]_T \\ \mathbb{O} \\ A_{32}[X]_T \end{pmatrix} \mathrm{d}\tau \right\| \\ &\leq \int_0^t \left\| \mathbb{I}_T\, \exp(A(t-\tau)) \begin{pmatrix} A_{12}[X]_T \\ \mathbb{O} \\ A_{32}[X]_T \end{pmatrix} \right\| \mathrm{d}\tau \\ &\leq \int_0^t \|\mathbb{I}_T\, \exp(A(t-\tau))\| \|A_{12} \oplus A_{32}\| \|[X]_T\| \mathrm{d}\tau, \end{aligned}$$

where for the Frobenius norm,

$$\|A_{12} \oplus A_{32}\| = (\|A_{12}\|^2 + \|A_{32}\|^2)^{\frac{1}{2}} \leq \|A_{12}\| + \|A_{32}\|.$$

In the case when A and $[A]_T$ are skew-Hermitian and if the Frobenius norm is used, $\|[X]_T\| = T$ and likewise $\|\mathbb{I}_T\, \exp(A(t-\tau))\| = T$ (since $\mathbb{I}_T$ picks off T rows of this infinite-dimensional unitary matrix, each row having a unit length). Therefore, in this case,

$$\| [\exp At]_T - \exp([A]_T\, t) \| \leq T^2 \|A_{12} \oplus A_{32}\| t. \tag{46}$$

This is a bound that we can use to quantify the error in the truncation and exponentiation method for approximating infinite-dimensional IURs for $SE(3)$.

On the other hand, if we want to find bounds on the effects of truncation and exponentiation of $SE(3)$ diffusion matrices, if we can obtain the eigenvalues of A and $[A]_T$, and if all of these eigenvalues have negative real part, then the eigenvalue with smallest magnitude negative real part will limit. Namely, if

$$\| \exp At\| \leq C \mathrm{e}^{-ct} \text{ and } \| \exp[A]_T t\| \leq C_T \mathrm{e}^{-c_T t}$$

then

$$\begin{aligned} \| [\exp At]_T - \exp([A]_T\, t) \| &\leq T \cdot C \cdot C_T \|A_{12} \oplus A_{32}\| \mathrm{e}^{-ct} \int_0^t \mathrm{e}^{(c-c_T)\tau} \mathrm{d}t \\ &= \frac{T \cdot C \cdot C_T \|A_{12} \oplus A_{32}\|}{c - c_T} [\mathrm{e}^{-c_T t} - \mathrm{e}^{-ct}]. \end{aligned} \tag{47}$$

This provides an upper bound on the error due to truncation and exponentiation of infinite-dimensional matrices.

6.3 Numerical Harmonic Analysis and Diffusions on Groups

The solution to the linear diffusion equation with drift $\partial f/\partial s = \Delta^* f$ subject to the initial conditions $f(g,0) = \delta(g)$ can be written using the operational properties of the group Fourier transform and inversion formula as

$$f(g,s) = \int_{\hat{G}} \mathrm{tr}\left[\mathrm{e}^{s\,\hat{\Delta}^*(\lambda)}\mathrm{e}^{u(X,\lambda)}\right]\mathrm{d}\lambda, \tag{48}$$

where $g = \exp X$ with $X = \sum_{i=1}^{6} x_i E_i$, $u(X,\lambda) = \sum_{i=1}^{6} x_i u(E_i,\lambda)$, and

$$\hat{\Delta}^*(\lambda) = \frac{1}{2}\sum_{i=1}^{6}\sum_{j=1}^{6} D_{ij}\,u(E_i,\lambda)u(E_j,\lambda) - \sum_{k=1}^{6}\xi_k u(E_k,\lambda).$$

When considering the numerical evaluation of (48) and properties of the solution $f(g,s)$, a number of issues can be explored. For example, when can the exponentials $\mathrm{e}^{s\,\hat{\Delta}^*(\lambda)}$ and $\mathrm{e}^{u(X,\lambda)}$ be computed efficiently? When can their product be computed efficiently? Is there a way to compute tr[·] without explicitly computing the matrix product? Would bringing the integral inside of tr[·] lead to internal cancelations that aid in speeding up computations? Even though in general $[\mathrm{e}^{s\,\hat{\Delta}^*(\lambda)}, \mathrm{e}^{u(X,\lambda)}] \neq 0$, are there situations in which $\mathrm{tr}[\mathrm{e}^{s\,\hat{\Delta}^*(\lambda)}\mathrm{e}^{u(X,\lambda)}] \approx \mathrm{tr}[\exp[s\,\hat{\Delta}^*(\lambda) + u(X,\lambda)]]$ is a close approximation?

In addition, in some applications, it is not necessary to know $f(g,s)$ at many values of g but rather only at $g = e$, in which case all that needs to be computed is

$$f(e,s) = \int_{\hat{G}} \mathrm{tr}\left[\mathrm{e}^{s\,\hat{\Delta}^*(\lambda)}\right]\mathrm{d}\lambda. \tag{49}$$

And since $\|f(g,s)\|$ (the integral of the square of $f(g,s)$ over G) is a measure of how concentrated $f(g,s)$ is, it is sometimes useful to compute

$$\|f(g,s)\|^2 = \int_{\hat{G}} \mathrm{tr}\left[\mathrm{e}^{s\,\hat{\Delta}^*(\lambda)}\mathrm{e}^{s\,\hat{\Delta}(\lambda)}\right]\mathrm{d}\lambda, \tag{50}$$

which follows from the Parseval equality.

When $[\hat{\Delta}^*(\lambda), \hat{\Delta}(\lambda)] = \mathbb{O}$, e.g., when Δ has no drift terms, this can be written as a single exponential.

In [7] issues related to the efficient computation of $\mathrm{e}^{s\,\hat{\Delta}^*(\lambda)}$, $\mathrm{e}^{u(X,\lambda)}$, and $f(e,s)$ and $\|f(g,s)\|^2$ are addressed together with exact "closed-form" lower and upper

bounds on integrals such as (48)–(50). These bounds can serve either as alternatives to approximate numerical evaluation or as checks on their accuracy.

7 Conclusions

This chapter has reviewed how harmonic analysis on the group of rigid-body motions in three-dimensional Euclidean space can be used to solve diffusion equations related to DNA statistical mechanics. Since IURs of the motion group are operators that can be expressed as infinite-dimensional matrices, truncation is required when using them in numerical codes. This chapter therefore reviews different issues related to truncation effects, thereby bringing together theory, applications, and numerical analysis issues.

References

1. Chirikjian, G.S.: Group theory and biomolecular conformation, I.: mathematical and computational models. J. Phys.: Condens. Matter **22** (2010) 323103 (21pp)
2. Chirikjian, G.S.: Information-theoretic inequalities on unimodular Lie groups. J. Geomet. Mech. **2**(2), 119–158 (2010)
3. Chirikjian, G.S.: Modeling loop entropy. Methods Enzymol. Part C **487**, 101–130 (2011)
4. Chirikjian, G.S.: Stochastic Models, Information Theory, and Lie Groups, vol. 2. Birkhäuser, Boston (2012)
5. Chirikjian, G.S., Wang, Y.F.: Conformational statistics of stiff macromolecules as solutions to PDEs on the rotation and motion groups. Phys. Rev. E **62**(1), 880–892 (2000)
6. Chirikjian, G.S., Kyatkin, A.B.: Engineering Applications of Noncommutative Harmonic Analysis. CRC Press, Boca Raton (2001)
7. Chirikjian, G.S., Liu, Y.: Truncation of infinite-dimensional group representations. In preparation.
8. Condon, E.U., Shortley, Q.W.: The Theory of Atomic Spectra. Cambridge University Press, Cambridge (1935)
9. Gelfand, I.M., Minlos, R.A., Shapiro, Z.Ya.: Representations of the rotation and Lorentz groups and their applications. Macmillan, New York (1963)
10. Gurarie, D.: Symmetry and Laplacians. Introduction to Harmonic Analysis, Group Representations and Applications. Elsevier Science, The Netherlands (1992) (Dover edition 2008)
11. Jones, M.N.: Spherical Harmonics and Tensors for Classical Field Theory. Research Studies Press, England (1985)
12. Kleinert, H.: Path Integrals in Quantum Mechanics, Statistics, and Polymer Physics, 2nd edn. World Scientific, Singapore (1995)
13. Liu, Y.: Probability density estimation on rotation and motion groups. Ph.D. Dissertation, JHU, (2007)
14. Miller, W. Jr.: Some applications of the representation theory of the Euclidean group in three-space. Commun. Pure App. Math. **17**, 527–540 (1964)
15. Miller, W. Jr.: Lie Theory and Special Functions. Academic, New York (1968)
16. Park, W., Wang, Y., Chirikjian, G.S.: The path-of-probability algorithm for steering and feedback control of flexible needles. Internat. J. Robot. Res. **29**(7), 813–830 (2010)

17. Park, W., Midgett, C.R., Madden, D.R., Chirikjian, G.S.: A stochastic kinematic model of class averaging in single-particle electron microscopy. Internat. J. Robot. Res. (Special issue on Stochasticity in Robotics and Biological Systems) **30**(6), 730–754 (2011)
18. Vilenkin, N.Ja., Klimyk, A.U.: Representation of Lie Groups and Special Functions, vols. 1–3. Kluwer, Dordrecht, Holland (1991)
19. Wang, Y.: Applications of diffusion processes in robotics, optical communications and polymer science. Ph.D. Dissertation, JHU (2001)
20. Wang, Y., Chirikjian, G.S.: Engineering applications of the motion-group Fourier transform. In: Rockmore, D.N., Healy, D.M., Jr. (eds.) Modern Signal Processing, pp. 63–78. MSRI Publications 46, Cambridge University Press (2004)
21. Zhou, Y., Chirikjian, G.S.: Conformational statistics of bent semiflexible polymers. J. Chem. Phys. **119**(9), 4962–4970 (2003)
22. Zhou, Y., Chirikjian, G.S.: Conformational statistics of semi-flexible macromolecular chains with internal joints. Macromolecules **39**(5), 1950–1960 (2006)

Quantification of Retinal Chromophores Through Autofluorescence Imaging to Identify Precursors of Age-Related Macular Degeneration

M. Ehler, J. Dobrosotskaya, E.J. King, and R.F. Bonner

Abstract Age-related macular degeneration is a common disease that impairs central vision. To better understand early disease progression, we quantified two families of retinal chromophores: macular pigments in retinal axons and rod photoreceptor rhodopsin, whose changes have been associated with age-related maculopathy progression. First, we introduced noninvasive multispectral fluorescence imaging of the human retina and quantified macular pigments from those multispectral image sets. Second, we modeled the brightening of the lipofuscin autofluorescence in confocal scanning laser ophthalmoscopy imaging sequences to map local rod rhodopsin density.

1 Introduction

As the elderly demographic in many industrialized countries is growing, age-related diseases are becoming more common. One such disease is age-related macular degeneration (AMD), which is the most common cause of blindness among the elderly in the developed world, cf. [4, 5, 20]. In a majority of Americans over

M. Ehler (✉)
Helmholtz Zentrum Munich, Institute of Biomathematics and Biometry, Neuherberg, Germany
e-mail: martin.ehler@helmholtz-muenchen.de

J. Dobrosotskaya
Mathematics Department, University of Maryland, College Park, MD, USA
e-mail: juliadob@math.umd.edu

E.J. King
Department of Mathematics, Technical University Berlin, Berlin, Germany
e-mail: eking@math.umd.edu

R.F. Bonner
Section on Medical Biophysics, NICHD, National Institutes of Health, Bethesda, MD, USA
e-mail: bonnerr@mail.nih.gov

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_18,

the age of 60, the earliest clinical signs of retinal pigment epithelium (RPE) dysfunction are observed in color fundus photographs as *drusen*—bright, highly reflective extracellular deposits arising from the RPE. Macular drusen increase in number and size with advancing age. Larger, irregularly shaped, perifoveal sub-RPE drusen ("soft"); regions of supra-RPE reticular drusen; and elevated levels of RPE bisretinoids are considered to confer the greatest risk for progression to advanced AMD. Currently, pathologists in reading centers classify drusen based on size and shape [2] in reflection color fundus images, and many retinal diseases are diagnosed and evaluated by subjective examination of retinal images and lower resolution visual field testing [26]. To this end, we are developing automated analysis tools to obtain molecular maps of these biomarkers from multispectral autofluorescence images.

Effective prevention of disease progression requires the identification of precursor lesions and the ability to quantify biomarker imbalances. Retinal imaging modalities have intrinsically high resolutions ($\approx 5\mu$m) that are inherent to the evolutionary design of the eye and its ocular media. Improved quantitative automated image analysis tools should increase understanding of mechanisms of early retinal disease and provide a means to assess disease prevention strategies. Noninvasive autofluorescence imaging of RPE fluorophores and absorbing molecules in the overlying retina offers the possibility to sensitively monitor early changes in retinal function and early pathophysiology.

In the present chapter[1] we present approaches based on autofluorescence imaging to quantitatively measure three families of retinal chromophores, macular pigments, rod photoreceptor rhodopsin, and bisretinoids (lipofuscin fluorophores). Macular pigments (lutein and zeaxanthin) concentrate within the photoreceptor nerve fibers, and low macular pigment levels have been identified as risk factors for AMD. To quantify macular pigment distributions, we developed noninvasive multi-spectral fluorescence imaging of the human retina by adding selected interference filter sets to standard fundus cameras, cf. [7–11,17–19]. By exciting the fluorescent lipofuscin granules within the RPE, the Beer–Lambert model for the double-path penetration enables us to effectively measure the spatial macular pigment distribution from a set of multi-spectral autofluorescence images.

Localized rod photoreceptor and rhodopsin losses have been observed in postmortem histology of age-related maculopathy. Degraded dark adaptation during aging is due to reduction in the rate of RPE regeneration of *cis* retinal to rod rhodopsin and thus may locally reflect early RPE dysfunction. We extended a model for rod rhodopsin bleaching and regeneration proposed in [3, 21, 23, 28] to create high-resolution maps of the spatial rod rhodopsin distribution from confocal scanning laser ophthalmoscope (cSLO) autofluorescence movies. Although there are virtually no rods in the center of the foveola, rod photoreceptors gradually appear in the foveola outskirts, where macular pigments are present. In contrast

[1] Unless stated otherwise, images were derived from NEI cameras or modified from resources at the National Eye Institute (NEI) or the Canadian National Institute for the Blind (CNIB).

to [3, 21, 23, 28], we explicitly incorporate the signal attenuation caused by macular pigments, which enables rod rhodopsin measurements in the entire macula.

The outline is as follows: We give a brief overview of AMD in Sect. 2 and introduce retinal autofluorescence imaging together with its physical model in Sect. 3. Macula pigment is quantified in Sect. 4. The model and physical measurements for rod rhodopsin quantification are presented in Sect. 5, and the computational aspects are discussed in Sect. 5.4. Conclusions are given in Sect. 6.

2 Age-Related Macular Degeneration

Aging of the human retina is universally associated with microscopic changes within the RPE, which is located at the back of the retina. In particular, the accumulation of fluorescent bisretinoids within RPE lipofuscin granules appears to be a chronic stressor [24] that, on the other hand, provides a direct means of imaging local changes in the RPE and the overlying retina via visible light autofluorescence.

AMD is a disease associated with aging that affects the macula, the part of the retina that enables central vision. Its incidence increases geometrically with increasing age above age 60 and may reflect an advanced pathophysiology in a continuum of progressive RPE dysfunction associated with aging. "Dry" AMD refers to loss of photoreceptors in the macula caused by local regional atrophy in the RPE that reduces its support of the overlying retina, which then atrophies. Other patients with severe visual loss develop "wet" AMD, which refers to choroidal neovascularization and subretinal hemorrhage. Here, new choroidal microvessels invade Bruch's membrane and, when they leak or hemorrhage, cause irreversible damage to the overlying retinal layers that can lead to rapid vision loss.

Early signs of dry AMD are blurry vision and loss of sensitivity. A standard symptom of wet AMD is that straight lines appear wavy. The Amsler grid in Fig. 1 is used to detect the first visual signs of maculopathy, and, often, a person with wet AMD may view the Amsler grid as in Fig. 2b. Although enough peripheral vision can remain, such macular degeneration hampers reading or recognizing faces, because the central vision is impaired, cf. Fig. 2d.

3 Retinal Autofluorescence Imaging

The retina is a multilayer neural tissue, uniquely suited for noninvasive optical imaging with high resolution due to its evolutionary design with its back-side photodetection. As light penetrates the retina, it is largely unscattered and only locally absorbed by retinal chromophores (first, hemoglobin within large retinal vessels, then macular pigments largely in the photoreceptor axons, then the unbleached opsins within the photoreceptor outer segments) before reaching the lipofuscin granules in the RPE. The emitted fluorescence from unoxidized lipofuscin bisretinoids

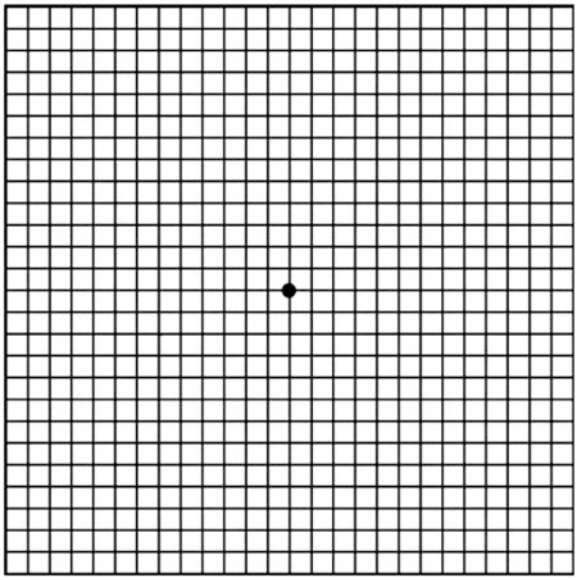


Fig. 1 While focusing on the dot in the center of the grid, with one eye covered, ask yourself the following questions: Am I able to see the corners and sides of the square? Do I see any crooked lines? Are there any holes or missing areas?

can be excited by a range of visible wavelengths (blue to yellow) to yield long-wavelength emission in the red, which is detected noninvasively by the fluorescence imaging camera. Generally, the excitation and emission wavelengths are partially absorbed by the various retinal chromophores (hemoglobin, macular pigments, rhodopsin) in the double passage through the retina requiring a double-path model. If $\mathrm{AF}(\Lambda, \lambda)$ denotes the measured autofluorescence, where Λ is the excitation and λ the emission wavelength, then the Beer–Lambert law for the double-path yields

$$\mathrm{AF}(\Lambda, \lambda) = I(\Lambda)\Phi(\Lambda, \lambda)\mathrm{e}^{-(D(\Lambda)+D(\lambda))}, \tag{1}$$

where D is the optical density of the underlying tissue, $\Phi(\Lambda, \lambda)$ is the fluorescence efficiency of lipofuscin, and $I(\Lambda)$ is the radiant power of the excitation light.

We use two different imaging techniques: First, to measure macular pigment, we developed autofluorescence imaging of the human retina at varying emission and excitation wavelengths by modifying standard fundus cameras. Secondly, we quantify rod rhodopsin from images recorded at a commercial cSLO camera (Heidelberg Retinal Angiograph 4.0) that delivers an average of $\approx 3\mu\mathrm{W/mm}^2$ at 488 nm, by rapidly scanning a small laser beam (10μ) over the retina. The intensity of excitation in the cSLO is $>$ 100-fold less than in the fundus camera so that each cSLO image is acquired over a 100 ms scan with an incremental rhodopsin bleaching ($\approx$ 1 %). After $\approx$ 40 s of cSLO imaging, the rhodopsin is completely

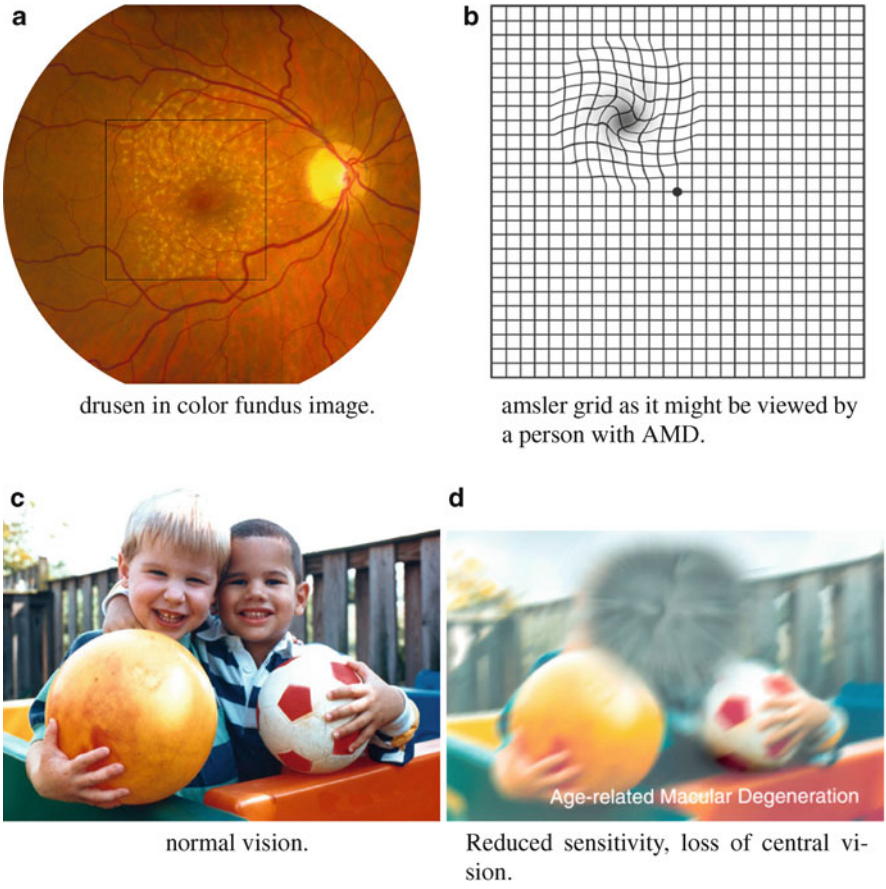


Fig. 2 Noticeable visual changes in AMD (**a**) Drusen in color fundus image (**b**) Amsler grid as it might be viewed by a person with AMD (**c**) Normal vision (**d**) Reduced sensitivity, loss of central vision

bleached, and its initial attenuation of the excitation light at 488 nm is removed. The magnitude of the brightening of the autofluorescence in registered movies allows mapping of rhodopsin density with high resolution.

4 Quantifying Macular Pigment

Macular pigment is composed of lutein and the related carotenoid zeaxanthin, and low macular pigment levels in the retina have been identified as risk factors for AMD. Neither lutein nor zeaxanthin is formed within the body and so can only be obtained from the diet. Both pigments are found in green, leafy vegetables (see Table 1).

Table 1 Lutein/zeaxanthin content per serving (1 cup) [29]

Food	Weight (g)	Content per serving (μg)
Spinach (cooked)	190	29,811
Kale (cooked)	130	25,606
Collards (cooked)	170	18,527
Peas (cooked)	160	3,840
Spinach (raw)	30	3,659
Pumpkin (cooked)	245	2,484
Corn (cooked)	156	2,429
Brussels sprouts	155	2,389
Broccoli (cooked)	156	2,015
Asparagus (cooked)	180	1,112
Carrots (cooked)	156	1,072
Beans (cooked)	125	886

Measurements of macular pigment based on two-wavelength autofluorescence images have been introduced by Delori et al. in [13]. To more robustly analyze the spatial macular pigment distribution, we introduce a multiple-wavelength model that enables more effective self-consistency tests.

Let $\mathrm{AF}_f(\Lambda,\lambda)$ and $\mathrm{AF}_p(\Lambda,\lambda)$ be the autofluorescence measured at the fovea and the perifovea, respectively. While AF_f depends on the specific location within the fovea, the term AF_p is often replaced by a circular average at 6 degrees [13]. We denote the optical density of the foveal and perifoveal tissue by D_f and D_p, respectively. Let Φ_f and Φ_p be the fluorescence efficiencies of lipofuscin in the foveal and perifoveal regions. According to (1), the foveal and perifoveal autofluorescence are given by

$$\mathrm{AF}_f(\Lambda,\lambda) = I(\Lambda)\Phi_f(\Lambda,\lambda)\mathrm{e}^{-(D_f(\Lambda)+D_f(\lambda))},$$
$$\mathrm{AF}_p(\Lambda,\lambda) = I(\Lambda)\Phi_p(\Lambda,\lambda)\mathrm{e}^{-(D_p(\Lambda)+D_p(\lambda))}.$$

Since there is no macular pigment in the perifoveal region, the optical density of macular pigment D_{MP} at 460 nm is the difference

$$D_{\mathrm{MP}}(460) = D_f(460) - D_p(460).$$

We use the relative extinction coefficient k_{MP} of macular pigment, scaled to $k_{\mathrm{MP}}(460) = 1$, such that

$$D_{\mathrm{MP}}(\lambda) = k_{\mathrm{MP}}(\lambda)\, D_{\mathrm{MP}}(460),$$

and we obtain

$$\log\left(\frac{\mathrm{AF}_p(\Lambda,\lambda)}{\mathrm{AF}_f(\Lambda,\lambda)}\right) = \log\left(\frac{\Phi_p(\Lambda,\lambda)}{\Phi_f(\Lambda,\lambda)}\right) + D_{\mathrm{MP}}(460)(k_{\mathrm{MP}}(\Lambda) + k_{\mathrm{MP}}(\lambda)).$$

We choose n pairs of excitation and emission wavelengths $\{(\Lambda_j, \lambda_j)\}_{j=1}^n$ and weights $\{\omega_j\}_{j=1}^n$ such that

$$\sum_{j=1}^{n} \omega_j = 0.$$

We apply the above equations for each wavelength pair (Λ_j, λ_j), multiply by ω_j, and add them up to obtain

$$\sum_{j=1}^{n} \omega_j \log\left(\frac{\mathrm{AF}_p(\Lambda_j, \lambda_j)}{\mathrm{AF}_f(\Lambda_j, \lambda_j)}\right) = \sum_{j=1}^{n} \omega_j \log\left(\frac{\Phi_p(\Lambda_j, \lambda_j)}{\Phi_f(\Lambda_j, \lambda_j)}\right) + D_{\mathrm{MP}}(460) \sum_{j=1}^{n} \omega_j (k_{\mathrm{MP}}(\Lambda_j) + k_{\mathrm{MP}}(\lambda_j)).$$

We can assume that the fluorophore at the fovea has the same composition as that at the perifovea (constant shape of its spectrum over $\{(\Lambda_j, \lambda_j)\}_{j=1}^n$), and that foveal-perifoveal differences in absorption by other pigments (retinal blood, visual pigments, and RPE melanin) are negligible, so that the ratio

$$\frac{\Phi_p(\Lambda_j, \lambda_j)}{\Phi_f(\Lambda_j, \lambda_j)}$$

does not depend on j. Therefore, we obtain

$$\sum_{j=1}^{n} \omega_j \log\left(\frac{\mathrm{AF}_p(\Lambda_j, \lambda_j)}{\mathrm{AF}_f(\Lambda_j, \lambda_j)}\right) = D_{\mathrm{MP}}(460) \sum_{j=1}^{n} \omega_j (k_{\mathrm{MP}}(\Lambda_j) + k_{\mathrm{MP}}(\lambda_j)),$$

which enables us to determine $D_{\mathrm{MP}}(460)$ by

$$D_{\mathrm{MP}}(460) = \frac{1}{\sum_{j=1}^{n} \omega_j (k_{\mathrm{MP}}(\Lambda_j) + k_{\mathrm{MP}}(\lambda_j))} \log\left(\prod_{j=1}^{n} \frac{\mathrm{AF}_p^{\omega_j}(\Lambda_j, \lambda_j)}{\mathrm{AF}_f^{\omega_j}(\Lambda_j, \lambda_j)}\right). \quad (2)$$

If we only choose two excitation wavelengths $\Lambda_1 = 480$ and $\Lambda_2 = 520$ with the weights 1 and -1, respectively, and keep the emission wavelength λ fixed, then our formula leads to

$$D_{\mathrm{MP}}(460) = \frac{1}{k_{\mathrm{MP}}(480) - k_{\mathrm{MP}}(520)} \log\left(\frac{\mathrm{AF}_p(480, \lambda)\mathrm{AF}_f(520, \lambda)}{\mathrm{AF}_p(520, \lambda)\mathrm{AF}_f(480, \lambda)}\right) \quad (3)$$

as originally proposed in [13].

In contrast to (3), the formula (2) enables several tests for self-consistency by removing or adding wavelength pairs and by changing the weights. Figure 3a shows

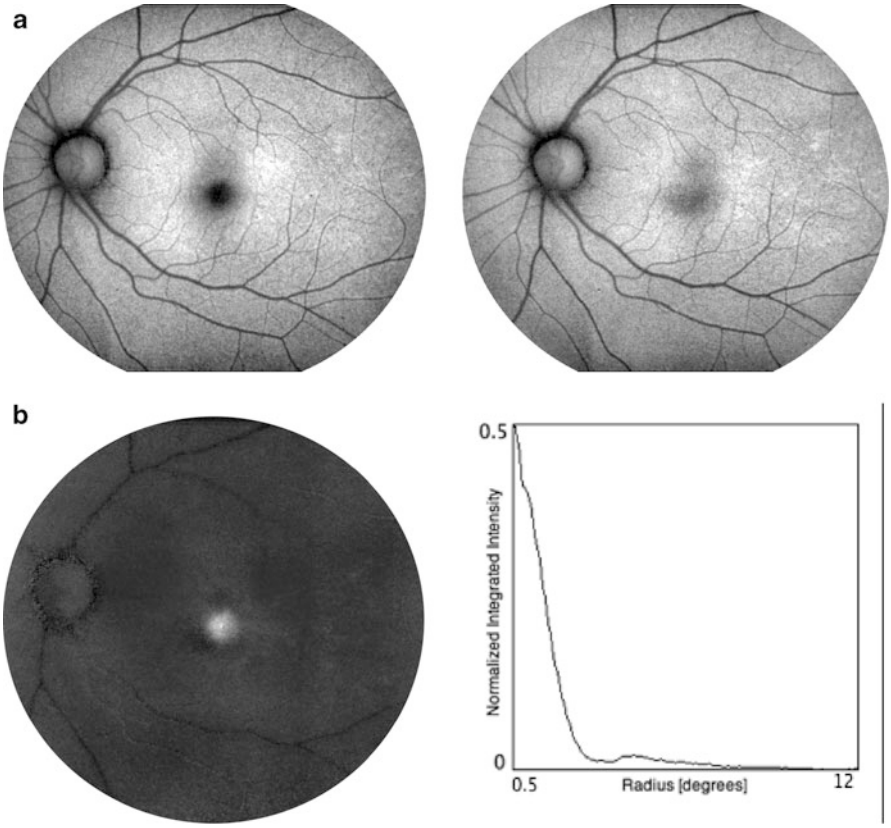


Fig. 3 Macular pigment measurements (**a**) Two bands (*blue and yellow excitation*) of a multispectral image set (**b**) (*Left*) Spatial map of MP density, (*right*) radial profile of MP density

two wavelengths of a multi-spectral autofluorescence image set. The spatial macular pigment map is given in Fig. 3b, where we optimized the map over several choices of weights to maximize self-consistency of the measurement. As expected, the macular pigment density is concentrated in the fovea and decreases with distance from the center.

5 Measuring Rod Rhodopsin Through Retinal Bleaching

We aim to map rhodopsin density within the human retina with sufficient accuracy to characterize local rod rhodopsin loss relative to surrounding areas. Such high-resolution maps are a measure of local rod function that could be uniquely

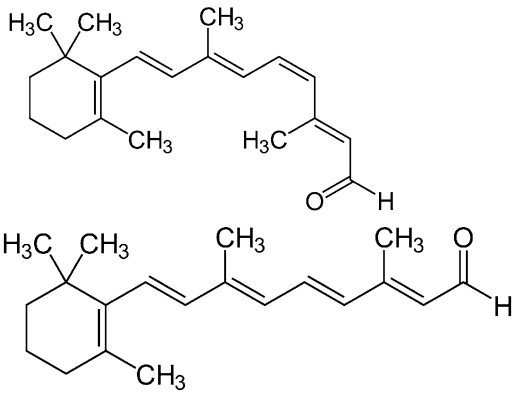


Fig. 4 (*Top*) 11-*cis* retinal, (*bottom*) all-*trans* retinal. In the outer segment, opsin is bound to 11-*cis* retinal that undergoes a conformational change into all-*trans* retinal when activated by blue light

sensitive to early changes in function of the RPE and rods within ophthalmoscopically (e.g., fundus camera or OCT) identified lesions. A model for rod rhodopsin bleaching and regeneration has been proposed in [3,21,23,28]. We extend this model to describe one-minute-long cSLO movies that show autofluorescence brightening to a steady-state level as the rhodopsin bleaches. Retinal bleaching has long been observed in the literature [3, 6, 14–16, 21–24, 27, 28] but has not yet been used to quantify or map local rod rhodopsin density changes in retinal disease.

5.1 Bleaching Model

A dark-adapted retina corresponds to a state when opsin only appears in its bound form with 11-*cis* retinal. Rod rhodopsin is opsin bound to 11-*cis* retinal that is activated by blue light and undergoes a conformational change into all-*trans* retinal, cf. Fig. 4. While 11-*cis* retinal absorbs blue light, the resulting photo-isomer all-*trans* retinal plus opsin is transparent at 488–507 nm. This photo-isomerization from 11-*cis* to all-*trans* is referred to as rhodopsin bleaching [21, 23, 28].

If we start with a dark-adapted retina, then the intensity of the local autofluorescence AF within a small circle of pixels should increase as the rhodopsin bleaches (becomes transparent to the $488nm$ exciting laser light). The unbleached rhodopsin in the photoreceptor layer initially attenuates the lipofuscin excitation by $\approx 50\,\%$ outside the central fovea, and the excitation light reaching the RPE then progressively increases as the overlying rhodopsin is locally bleached. Thus, we need to incorporate the time course in (1), so that we have

$$\mathrm{AF}(\Lambda,\lambda,t) = I(\Lambda)\Phi(\Lambda,\lambda)\mathrm{e}^{-(D(\Lambda,t)+D(\lambda,t))},$$

where we assume a constant radiation power $I(\Lambda)$. The optical density changes over time, and the major chromophores are macular pigment in the fovea and rhodopsin in the perifoveal region. There is, however, a region within the fovea, where both are present at a significant density. Although the macular pigment contribution was ignored in [21, 23, 28], here, we explicitly model the signal attenuation caused by macular pigment absorption to obtain rod rhodopsin maps covering the entire macula. Therefore, we obtain

$$D(\Lambda,t) + D(\lambda,t) = D_{\mathrm{MP}}(\Lambda) + D_{\mathrm{MP}}(\lambda) + D_{\mathrm{rh}}(\Lambda,t) + D_{\mathrm{rh}}(\lambda,t),$$

where D_{rh} denotes the optical density of present rhodopsin. If $R(t)$ denotes the fraction of rhodopsin remaining at time t, then we obtain

$$D(\Lambda,t) + D(\lambda,t) = D_{\mathrm{MP}}(\Lambda) + D_{\mathrm{MP}}(\lambda) + (D_{\mathrm{rh}}(\Lambda) + D_{\mathrm{rh}}(\lambda))R(t),$$

where $R(0) = 1$ corresponds to the dark-adapted retina and

$$D_{\mathrm{rh}}(\Lambda) + D_{\mathrm{rh}}(\lambda)$$

is the double-path optical density of rhodopsin present at time $t = 0$. If $\rho_{\mathrm{rh}}(0)$ denotes the physical density of present rhodopsin at time $t = 0$ and its molar extinction coefficient is k_{rh}, then we have

$$D_{\mathrm{rh}}(\Lambda) + D_{\mathrm{rh}}(\lambda) = \rho_{\mathrm{rh}}(0)(k_{\mathrm{rh}}(\Lambda) + k_{\mathrm{rh}}(\lambda)),$$

which yields

$$\mathrm{AF}(\Lambda,\lambda,t) = I(\Lambda)\Phi(\Lambda,\lambda)\mathrm{e}^{-(D_{\mathrm{MP}}(\Lambda)+D_{\mathrm{MP}}(\lambda)+\rho_{\mathrm{rh}}(0)(k_{\mathrm{rh}}(\Lambda)+k_{\mathrm{rh}}(\lambda))R(t))}.$$

In order to determine $\rho_{\mathrm{rh}}(0)$, we still need to specify $R(t)$.

5.2 Regeneration Model

To specify the fraction of rhodopsin remaining at time t, we need to study rhodopsin regeneration, a process that competes with bleaching and in which 11-*cis* retinal binds to free opsin so that rhodopsin is rebuilt [21, 23]. In the RPE cells, *cis* retinol (vitamin A_1) is built from trans-ester through isomerase. The enzyme 11-*cis* retinol dehydrogenase (RDH) yields *cis* retinal from *cis* retinol. The interphotoreceptor binding protein (IRBP) escorts *cis* retinal to the outer segment where it binds to opsin so that rhodopsin is rebuilt, see Fig. 5. To specify $R(t)$, we need to model the regeneration process, for which we started with Michaelis–Menten kinetics of the production of *cis* retinal from *cis* retinol. The change in the fraction of unbound opsin

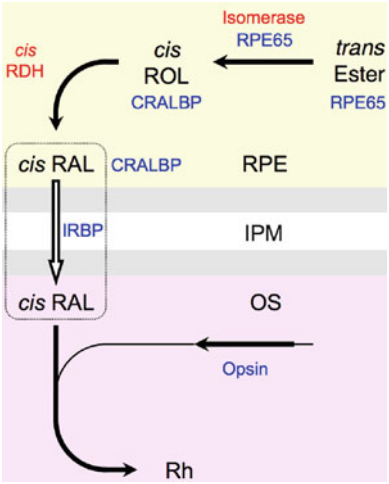


Fig. 5 Rhodopsin regeneration scheme, modified from images used in [21]. Retinol = vitamin A_1 (ROL), retinaldehyde (RAL), 11-*cis* retinol dehydrogenase (RDH), cellular retinaldehyde-binding protein (CRALBP), interphotoreceptor binding protein (IRBP), and rhodopsin (Rh)

$$\mathrm{Ops}(t) = 1 - R(t)$$

is proportional to the 11-*cis* retinal concentration that binds to opsin:

$$S + E \leftrightarrows [ES] \leftrightarrows E + P,$$

where S is 11-*cis* retinol, E is 11-cis-RDH (RDH5), and P is 11-*cis* retinal. The fractional change of unbound opsin satisfies

$$\frac{\partial}{\partial t}\mathrm{Ops}(t) = -kP(t)\mathrm{Ops}(t).$$

Thus, when rhodopsin is bleached by a steady light of illuminance I, its regeneration resembles a first-order reaction

$$\frac{1 - R(t)}{\tau},$$

where τ is the time constant. Therefore, the fractional change of the rhodopsin concentration satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}R(t) = -\frac{IR(t)}{L} + \frac{1 - R(t)}{\tau}, \tag{4}$$

where L is a "bleaching constant" that corresponds to the reciprocal of "photosensitivity" and has been measured by retinal densitometry [1,25]. Equation (4) has the analytical solution

$$R(t) = \frac{\frac{L}{\tau}}{I + \frac{L}{\tau}} + \frac{I}{I + \frac{L}{\tau}} \exp\left(-\left(1 + \frac{\tau I}{L}\right)\frac{t}{\tau}\right). \tag{5}$$

In cSLO measurements, the retinal illuminance I is much bigger than $\frac{L}{\tau}$, and we evaluate R only at t much smaller than τ. Therefore, (5) reduces to

$$R(t) \approx \exp\left(-\frac{I}{L}t\right).$$

5.3 Protocol

We follow a simple protocol in which the subject wears vermillion sunglasses (rod-protecting) while waiting and the photographer performs focus adjustment using the infrared reflection imaging (nonbleaching) in the cSLO. A $488nm$ excited autofluorescence movie ($\approx$ 8 frames/s, 1 min long) is started that is recorded from the start with blinks every 10 s, which refresh the tear film layer on the cornea. The average photon flux bleaches the rod rhodopsin after $>$ 25 s. We record the cSLO movie until steady-state rhodopsin bleaching.

Remark 1. Note that the bleaching kinetics and differences between foveal and perifoveal rhodopsin distribution do not affect macular pigment measurements. Each flash of the fundus camera in macular pigment measurements fully bleaches rhodopsin so that our subsequent multi-spectral images are not affected by the bleaching kinetics and hence do not depend on the rod rhodopsin distribution.

5.4 Computational Aspects

Combining the findings in Sects. 5.1 and 5.2 leads to

$$\mathrm{AF}(\Lambda, \lambda, t) = \alpha \mathrm{e}^{-\gamma \mathrm{e}^{-\beta t}},$$

where $\alpha := I(\Lambda)\Phi(\Lambda, \lambda)$, $\beta := \frac{I}{L}$, and

$$\gamma := D_{\mathrm{MP}}(\Lambda) + D_{\mathrm{MP}}(\lambda) + \rho_{\mathrm{rh}}(0)(k_{\mathrm{rh}}(\Lambda) + k_{\mathrm{rh}}(\lambda)).$$

To derive $\rho_{\mathrm{rh}}(0)$, we first determine the optical density of macular pigment $D_{\mathrm{MP}}(460)$ so that we can compute

$$D_{\mathrm{MP}}(\Lambda) + D_{\mathrm{MP}}(\lambda) = (k_{\mathrm{MP}}(\Lambda) + k_{\mathrm{MP}}(\lambda))D_{\mathrm{MP}}(460).$$

In order to determine $\rho_{\mathrm{rh}}(0)$, we shall compute the model parameters α, β, and γ as a minimum of the energy functional

$$E_f(\alpha, \beta, \gamma) = \int_{T_0}^{T} \left(\alpha e^{-\gamma e^{-\beta t}} - f(t)\right)^2 dt, \tag{6}$$

where $f(t)$ is the image data derived from the cSLO measurements. A gradient descent method determines the optimal values of α, β, and γ as the steady-state solutions of the following system of ODEs:

$$\frac{d}{dt}\alpha = -\int_{T_0}^{T} \left(\alpha \exp(-\gamma \exp(-\beta t)) - f(t)\right) \exp(-\gamma \exp(-\beta t)) dt. \tag{7}$$

$$\frac{d}{dt}\beta = -\alpha\gamma \int_{T_0}^{T} \left(\alpha \exp(-\gamma \exp(-\beta t)) - f(t)\right) \exp(-\gamma \exp(-\beta t)) \exp(-\beta t) t\, dt. \tag{8}$$

$$\frac{d}{dt}\gamma = \alpha \int_{T_0}^{T} \left(\alpha \exp(-\gamma \exp(-\beta t)) - f(t)\right) \exp(-\gamma \exp(-\beta t)) \exp(-\beta t) dt. \tag{9}$$

Due to the high non-linearity of the model, the differential equations are solved using an explicit finite difference scheme with the time stepping chosen individually for each parameter. When aiming for the recovery of the actual parameters of each bleaching curve, we consider the parameters that we recover as optimal approximations within the declared model to the actual parameters. The recovered parameters can coincide with the actual ones (α, β, γ), differ by a numerical-error margin, or differ due to the fact that in certain areas the proposed model does not describe all physiological phenomena completely. The details of the minimization procedure and its mathematical foundations are described in [12]. A spatial map of γ is shown in Fig. 6, and spatial variations within a bleaching movie are visualized in Fig. 7.

6 Conclusion

The analysis of cSLO lipofuscin autofluorescence image sequences in dark-adapted subjects provides a new means for high-resolution mapping of the state of human rod photoreceptors. We extended a validated physiological model of rhodopsin bleaching kinetics by incorporating the macular pigment density, which we can reliably quantify from multispectral autofluorescence images. As opposed to [3, 21, 23, 28], we can therefore determine the correct rod rhodopsin density even within the central fovea, where both macular pigments and rhodopsin are present at a significant density. The physiological variables $\alpha(x, y)$, $\beta(x, y)$, and $\gamma(x, y)$ appear to be consistent as we optimize the model parameters using the steady state of a system of ODEs. Our model and parameter optimization provide plausible high-resolution maps of rod rhodopsin distributions in normal individuals using a clinically simple $1min$ noninvasive imaging sequence.

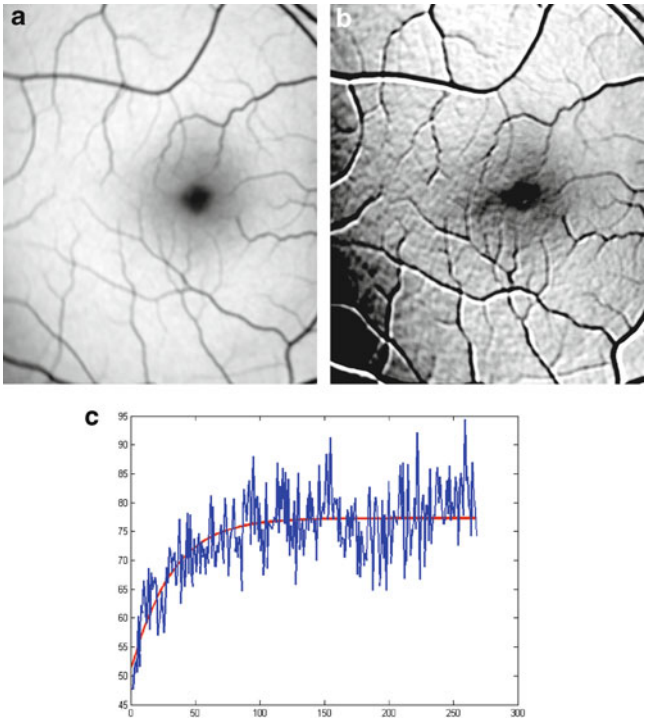


Fig. 6 Values of the parameters α and γ determined via the gradient descent minimization with a fixed value of $\beta = 0.0368$ (**a**) Values of a between 43 and 108 (**b**) Values of c between 0.1 and 0.5 (**c**) A typical temporal sequence of the intensity values (*blue*) at one pixel and the corresponding fit (*red*)

Our new noninvasive imaging and analysis approaches appear well suited for measuring localized changes in macular pigments and rod photoreceptors and to correlate them at high spatial resolution with localized pathological changes of the RPE seen in steady-state local autofluorescence images. Further refinement and validation should lead to better methods for evaluating very early microscopic lesions, their natural progression, and their responses to benign disease prevention strategies that are most likely to be effective with early "preclinical" changes. For instance, the retinal irradiance during cSLO lipofuscin autofluorescence imaging provides $> 97\,\%$ complete rhodopsin bleaching and therefore is insensitive to the visual cycle. If one uses a $576nm$ laser wavelength, for which the rhodopsin absorption is about eightfold less, then a longer movie at $576nm$ or intermittent imaging at $488nm$ might result in lower steady-state bleaching fractions and slower kinetics, enabling us to extract the visual cycle time. Additionally, the Heidelberg Spectralis HRA+OCT allows simultaneous optical coherence tomography and cSLO auto fluorescence measurements, cf. Fig. 8, in which 3-dimensional structure (OCT) can be correlated with local changes in rhodopsin and fluorescence bisretinoids

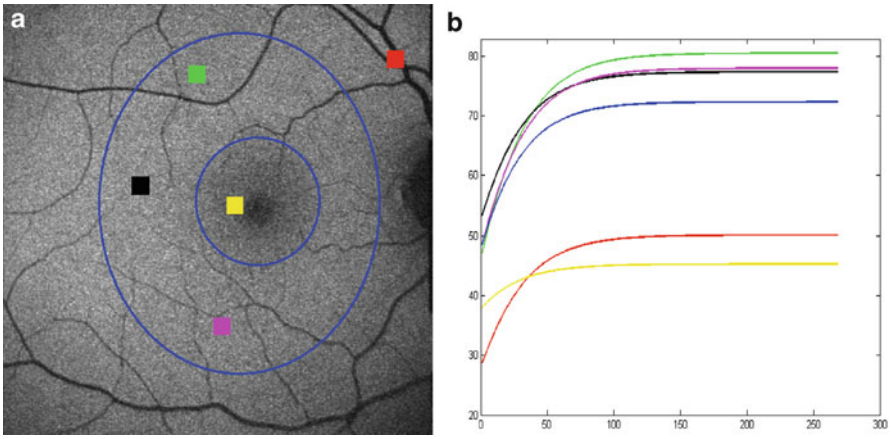


Fig. 7 Fitted bleaching curves at various regions (**a**) Regions marked on the first image of the bleaching stack. Note that the model is off on the blood vessel (**b**) Fitted bleaching curves, the additional blue curve corresponds to the fitting of an average over the annulus-like area between the two blue ovals shown in (a)

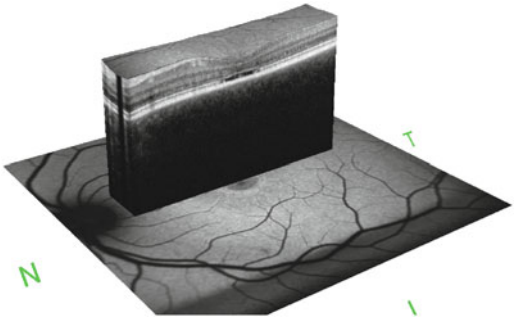


Fig. 8 The Heidelberg Spectralis HRA+OCT enables us to complement local autofluorescence measurements with optical coherence tomography images, recorded in parallel and well aligned

by our new autofluorescence image analysis methods. Such integrated quantitative analysis could allow improved characterization of the natural history of early lesion progression and responses to therapies.

Acknowledgment The research was funded by intramural research funds from the National Institute of Child Health and Human Development, National Institutes of Health. M. E. is supported by the NIH/DFG Research Career Transition Awards Program (EH 405/1-1/575910). J.D. was supported by NSF (CBET0854233). E.J.K. is supported by the Alexander von Humboldt Foundation.

References

1. Alpern, M., Pugh, E.N. Jr.: The density and photosensitivity of human rhodopsin in the living retina. J. Physiol. **237**, 341–370 (1974)
2. Bird, A.C., Bressler, N.M., Bressler, S.B., Chisholm, I.H., Coscas, G., Davis, M.D., de Jong, P.T., Klaver, C.C., Klein, B.E., Klein, R.: An international classification and grading system for age-related maculopathy and age-related macular degeneration. The international ARM epidemiological study group. Surv. Ophthalmol. **39**(5), 367–374 (1995)
3. Cameron, A.M., Miao, L., Ruseckaite, R., Pianta, M.J., Lamb, T.D.: Dark adaptation recovery of human rod bipolar cell response kinetics estimated from scotopic b-wave measurements. J. Physiol. **586**(Pt 22), 5419–5436 (2008)
4. Chew, E.Y., Lindblad, A.S., Clemons, T., and Age-Related Eye Disease Study Research Group. Summary results and recommendations from the age-related eye disease study. Arch. Ophthalmol. **127**(12), 1678–1679 (2009)
5. Coleman, H.R., Chan, C., Ferris, F.L., Chew, E. Y.: Age-related macular degeneration. Lancet **372**(9652), 1835–1845 (2008)
6. Cunningham, D., Caruso, R.C., Ferris, F.L.: Case report: Retinal bleaching during fundus autofluorescence using a confocal scanning laser ophthalmoscope. J. Ophthalmic Photogr. **29**, 93–94 (2007)
7. Dobrosotskaya, J., Ehler, M., et al.: Sparse representation and variational methods in retinal image processing. In: International Federation for Medical & Biological Engineering. Springer Proceedings Series. 26th Southern Biomedical Engineering Conference. Springer, Berlin (2010)
8. Dobrosotskaya, J., Ehler, M., et al.: Modeling of the rhodopsin bleaching with variational analysis of retinal images. SPIE Medical Imaging, Image Processing **7962**(1), 79624N (2011)
9. Ehler, M., Dobrosotskaya, J., et al.: Modeling photo-bleaching kinetics to create high resolution maps of rod rhodopsin in the human retina. Preprint (2012)
10. Ehler, M., Kainerstorfer, J., Cunningham, D., et al.: Extended correction model for optical imaging. In: IEEE International Conference on Computational Advances in Bio and Medical Sciences, pp. 93–98 (2011)
11. Ehler, M., Dobrosotskaya, J., et al.: Modeling photo-bleaching kinetics to map local variations in rod rhodopsin density. SPIE Medical Imaging, Computer-Aided Diagnosis, **7963**(1), 79633R (2011)
12. Ehler, M., Dobrosotskaya, J., King, E., Czaja, W., Bonner, R.F.: Modeling photo-bleaching kinetics to create high resolution maps of rod rhodopsin in the human retina. Preprint (2012)
13. Delori, F.C., et al.: Macular pigment density measured by autofluorescence spectrometry: comparison with reflectometry and heterochromatic flicker photometry. J. Opt. Soc. Am. A. Opt. Image Sci. Vis. **18**(6), 1212–1230 (2001)
14. Gloster, J., Greaves, D.P.: A study of bleaching of the fundus oculi in pigmentary degenerations of the retina. Br. J. Ophthalmol. **48**, 260–273 (1964)
15. Jackson, G.R., Owsley, C., McGwin, G. Jr.: Aging and dark adaptation. Vision Res. **39**(23), 3975–3982 (1999)
16. McGwin, G. Jr., Jackson, G.R., Owsley, C.: Using nonlinear regression to estimate parameters of dark adaptation. Behav. Res. Methods Instrum. Comput. **31**(4), 712–717 (1999)
17. Kainerstorfer, J., Amyot, F., Ehler, M., et al.: Direct curvature correction for non-contact imaging modalities - applied to multi-spectral imaging. J. Biomed. Opt. **15**(4), 046013 (2010)
18. Kainerstorfer, J., Ehler, M., et. al.: Principal component model of multi spectral data for near real-time skin chromophore mapping. J. Biomed. Opt. **15**(4), 046007 (2010)
19. Kainerstorfer, J., Riley, J.D., Ehler, M., et al.: Quantitative principal component model for skin chromophore mapping using multi spectral images and spatial priors. Biomed. Opt. Exp., **2**(5), 1040–1058 (2011)
20. Krishnadev, N., Meleth, A.D., Chew, E.Y.: Nutritional supplements for age-related macular degeneration. Curr. Opin. Ophthalmol. **21**(3), 184–189 (2010)

21. Lamb, T.D., Pugh, E.N. Jr.: Phototransduction, dark adaptation, and rhodopsin regeneration the proctor lecture. Invest. Ophthalmol. Vis. Sci. **47**(12), 5137–52 (2006)
22. Lyubarsky, A.L., Daniele, L.L., Pugh, E.N. Jr.: From candelas to photoisomerizations in the mouse eye by rhodopsin bleaching in situ and the light-rearing dependence of the major components of the mouse erg. Vision Res. **44**(28), 3235–3251 (2004)
23. Mahroo, O.A.R., Lamb, T.D.: Recovery of the human photopic electroretinogram after bleaching exposures: estimation of pigment regeneration kinetics. J. Physiol. **554**(Pt 2), 417–437 (2004)
24. Meyers, S.M., Ostrovsky, M.A., Bonner, R.F.: A model of spectral filtering to reduce photochemical damage in age-related macular degeneration. Trans. Am. Ophthalmol. Soc. **102**, 83–93; discussion 93–95 (2004)
25. Rushton, W.A.H., Powell, D.S.: The rhodopsin content and the visual threshold of human rods. Vision Res. **12**, 1073–1081 (1972)
26. Sandberg, M.A., Weigel-DiFranco, C., Rosner, B., Berso, E.L.: The relationship between visual field size and electroretinogram amplitude in retinitis pigmentosa. Invest. Ophthalmol. Vis. Sci. **37**(8), 1693–1698 (1996)
27. Theelen, T., Berendschot, T.T.J.M., Boon, C.J.F., Hoyng, C.B., Klevering, B.J.: Analysis of visual pigment by fundus autofluorescence. Exp. Eye Res. **86**(2), 296–304 (2008)
28. Thomas, M.M., Lamb, T.D.: Light adaptation and dark adaptation of human rod photoreceptors measured from the a-wave of the electroretinogram. J. Physiol. **518**(Pt 2), 479–496 (1999)
29. U.S. Department of Agriculture: USDA National Nutrient Database for Standard Reference, Release 24. Agricultural Research Service (2011)

Simple Harmonic Oscillator Based Reconstruction and Estimation for One-Dimensional q-Space Magnetic Resonance (1D-SHORE)

Evren Özarslan, Cheng Guan Koay, and Peter J. Basser

Abstract The movements of endogenous molecules during the magnetic resonance acquisition influence the resulting signal. By exploiting the sensitivity of diffusion on the signal, q-space MR has the ability to transform a set of diffusion-attenuated signal values into a probability density function or propagator that characterizes the diffusion process. Accurate estimation of the signal values and reconstruction of the propagator demand sophisticated tools that are well suited to these estimation and reconstruction problems. In this work, a series representation of one-dimensional q-space signals is presented in terms of a complete set of orthogonal Hermite functions. The basis possesses many interesting properties relevant to q-space MR, such as the ability to represent both the signal and its Fourier transform. Unlike the previously employed cumulant expansion, bi-exponential fit, and similar methods, this approach is linear and capable of reproducing complicated signal profiles, e.g., those exhibiting diffraction peaks. The estimation of the coefficients is fast and accurate while the representation lends itself to a direct reconstruction of ensemble average propagators as well as calculation of useful descriptors of it, such as the return-to-origin probability and its moments. In axially symmetric and isotropic geometries, respectively, two- and three-dimensional propagators can be

E. Özarslan
Section on Tissue Biophysics and Biomimetics (STBB), PITS, NICHD, NIH

Center for Neuroscience and Regenerative Medicine, USUHS, Bethesda, MD, USA
e-mail: evren@helix.nih.gov

C.G. Koay
Department of Medical Physics, University of Wisconsin, Madison, WI, USA
e-mail: cgkoay@wisc.edu

P.J. Basser (✉)
Section on Tissue Biophysics and Biomimetics (STBB), PITS, NICHD, NIH, Bethesda, MD, USA
e-mail: pjbasser@helix.nih.gov

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_19,

reconstructed from one-dimensional q-space data. Useful relationships between the one- and higher-dimensional propagators in such environments are derived.

Keywords Magnetic resonance • q-space • Diffusion • Propagator • MR • Hermite polynomial • Return-to-origin • Return-to-axis • Return-to-plane • Axial symmetry • Isotropy

1 Introduction

Diffusion is a transport process characterized by the spontaneous and incessant movements of particles. The characteristics of the diffusion process are determined by the structure of the host matrix. As such, one can obtain information about the domain in which diffusion is taking place by observing diffusion. One widespread method for measuring diffusion involves the nuclear magnetic resonance (NMR or MR) technique whose sensitivity to diffusion of spin-bearing molecules was realized in its earliest days [9]. Later on, it was demonstrated that by incorporating a pair of pulsed magnetic field gradients into conventional MR acquisitions, one can observe diffusion in a convenient and controllable way [37]. This "pulsed-field-gradient" (PFG) MR technique enabled the examination of numerous substances in diverse areas. A spin that is moving between the application of the two diffusion sensitization pulses of the PFG experiment suffers a net phase shift. A population of randomly moving spins yields an incoherent phase profile, which leads to an attenuation of the MR signal [17].

In diffusion MR, the net displacement vector $\mathbf{R}$ is a Fourier conjugate to an experimentally controlled variable $\mathbf{q} = \gamma\delta\mathbf{G}/(2\pi)$, where γ is the gyromagnetic ratio, δ is the duration of the diffusion gradient pulses, and $\mathbf{G}$ is the diffusion gradient vector, i.e., [13, 38],

$$P_{3D}(\mathbf{R}) = \int \mathrm{d}\mathbf{q}\, \mathrm{e}^{\mathrm{i}2\pi\mathbf{q}\cdot\mathbf{R}}\, E(\mathbf{q})\ . \tag{1}$$

Here, $E(\mathbf{q})$ is the MR signal attenuation and, when δ is small, $P_{3D}(\mathbf{R})$ is an ensemble average propagator indicating the probability for molecules to undergo a displacement $\mathbf{R}$ in the interval between the two pulses. Therefore, $P_{3D}(\mathbf{R})$ can be estimated from data obtained via sampling the three-dimensional "q-space" and then employing a Fourier transform scheme.

Frequently, because of experimental limitations or because the desired characteristics of the specimen can be extracted from one-dimensional data, the entire three-dimensional q-space is not sampled. Instead, a one-dimensional version of the q-space acquisition is performed by keeping the direction of the diffusion gradients fixed, and varying only their strength. If the x-axis is defined to be the direction of the gradients, then a one-dimensional average propagator can be obtained from the relationship

$$P(x) = \int \mathrm{d}q\, \mathrm{e}^{\mathrm{i}2\pi q x}\, E(q). \tag{2}$$

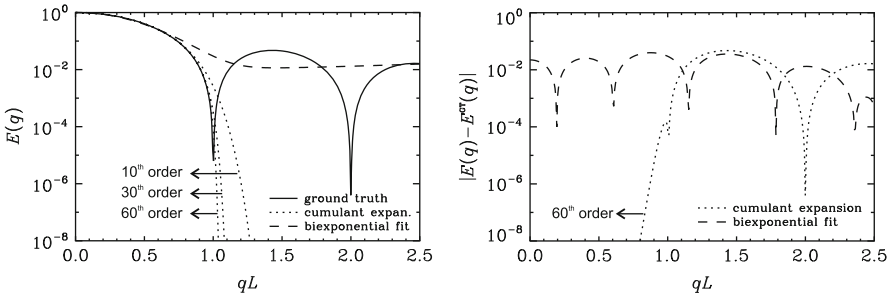


Fig. 1 MR signal attenuation expected from spins diffusing inside a rectangular pore of length L is depicted via the continuous line on the *left panel*. Also shown on this panel are the curves obtained from a bi-exponential fit (*dashed line*) and cumulant expansion with 10th-, 30th-, and 60th-order approximations (*dotted lines*). The *right panel* shows the errors acquired in the bi-exponential fit and the 60th order cumulant expansion. In these simulations, the bi-exponential fit was obtained from 400 points uniformly covering a qL interval of [0,2.5]. The cumulant expansions are obtained by analytically expanding the logarithm of the signal attenuation in a power series

By sensitizing the signal to the random motion of the molecules, the q-space MR technique enables the study of microscopic compartments whose dimensions cannot be resolved by conventional MR imaging and microscopy. Moreover, the one-dimensional average propagator was shown to provide information about diffusion, flow, restrictions to motion, and even spatially dependent relaxation sinks [6]. In one application of diffusion acquisitions involving specimens with an ordered microstructure, the non-monotonic dependence of the q-space signal on q [4, 23] was exploited to determine cell sizes. In another application, the q-space signal has been used to estimate scaling exponents that were related to the fractal dimension of disordered media [25]. Since the diffusion propagator is a probability density function, among its descriptors are the moments of this density function. Another important quantity is the probability for no net displacement, or more commonly referred to as the return-to-origin probability [10]. These quantities are all indicators of tissue microstructure, which could be altered by changes due to development, aging, and disease.

Estimation of the derived quantities and reconstruction of the propagators can be significantly improved if the signal decay can be expressed parametrically. For this purpose, bi-exponential fitting [5, 26] and cumulant expansion [11, 16, 18, 39] techniques have been applied to q-space data. However, both of these approaches are limited in their ability to reproduce general $E(q)$ profiles. For example, bi-exponential functions are monotonic by design, and as such, they can not possibly model non-monotonic diffraction-like features. The cumulant expansion method is bound to fail as well, because the signal minima are typically at or beyond the radius of convergence [7, 14] for such expansions. Moreover, Pawula's theorem guarantees that the propagators reconstructed from a cumulant expansion terminated beyond the Gaussian term will have unacceptable properties [32]. Figure 1 illustrates how both of these methods fail in reproducing the exact MR signal attenuation

from spins diffusing inside a rectangular pore. Other parameterizations of the q-space signal include the assignment of a continuous spectrum of diffusivities [33, 36, 42] and fitting stretched exponential [2, 15] or Rigaut-type asymptotic fractal expressions [12, 15] to diffusion-attenuated MR data. These parametric representations also suffer from the above-mentioned problems.

In this work, we propose expressing the one-dimensional q-space MR signal in terms of the eigenfunctions of the quantum-mechanical simple harmonic oscillator Hamiltonian, sometimes called the Hermite functions, which form a complete orthogonal basis for the space of square-integrable functions [21]. Because Fourier transforms of these functions are Hermite functions themselves, our approach directly yields a propagator expressed in the same set of basis functions. Estimation of probability distributions in a series of Hermite functions is well studied in the statistics literature [35], and such expansions were shown to possess powerful properties, such as rapid convergence in both real and Fourier spaces [41], which suit problems of q-space signal analysis and average propagator estimation.

After introducing the basis and a numerical estimation method for its coefficients in the next section, in Sect. 3, we evaluate the accuracy of the signal, propagator, moment, and return-to-origin probability estimates. Several important and useful relationships regarding the employed basis and geometries with axial symmetry or isotropic environments are derived in the appendices.

2 Theory

We propose expressing the diffusion-weighted MR signal as

$$S(q) = \sum_{n=0}^{N-1} a'_n \, \phi_n(u, q), \tag{3}$$

with

$$\phi_n(u, q) = \frac{i^{-n}}{\sqrt{2^n \, n!}} \, \mathrm{e}^{-2\pi^2 q^2 u^2} \, H_n(2\pi u q). \tag{4}$$

Here $H_n(x)$ is the nth-order Hermite polynomial and u is a characteristic length to be determined. The MR signal attenuation, defined to be $E(q) = S(q)/S(0)$, can be expressed in the same basis as

$$E(q) = \sum_{n=0}^{N-1} a_n \, \phi_n(u, q), \tag{5}$$

where

$$a_n = \frac{a'_n}{S_0}, \tag{6}$$

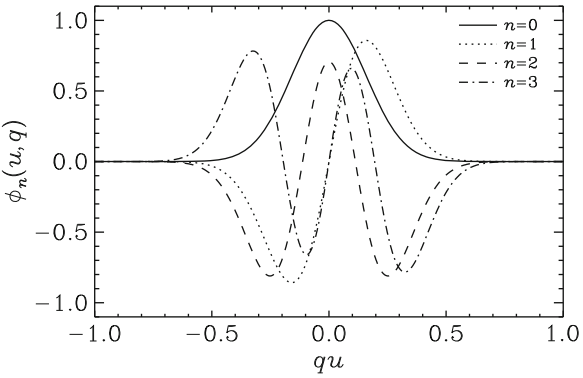


Fig. 2 First four functions used in the expansion of the MR signal profiles, $S(q)$. Note that the real and imaginary parts of, respectively, the even- and odd-ordered functions are plotted as the other parts are 0

with $S_0 = S(0)$ is the signal with no diffusion weighting, which can be estimated from the coefficients a'_n:

$$S_0 = \sum_{n=0}^{N-1} a'_n \, \phi_n(u,0) = \sum_{n=0,2,4,\ldots}^{N-1} \frac{(n-1)!!}{\sqrt{n!}} \, a'_n. \tag{7}$$

Note that the ϕ_n functions are related to the eigenfunctions of the quantum-mechanical simple harmonic oscillator Hamiltonian. It is well known that these functions form a complete orthogonal basis for the space of square-integrable functions [21]. Figure 2 depicts the first few of these functions. One important property of these functions is that their Fourier transforms are also Hermite functions; this characteristic enables direct estimation of the propagator through the expression

$$P(x) = \sum_{n=0}^{N-1} a_n \, \psi_n(u,x), \tag{8}$$

where

$$\begin{aligned} \psi_n(u,x) &= \frac{i^n}{\sqrt{2\pi}\,u} \, \phi_n\!\left(\frac{1}{2\pi u}, x\right) \\ &= \frac{1}{\sqrt{2^{n+1}\,\pi\,n!}\,u} \, \mathrm{e}^{-x^2/(2u^2)} \, H_n(x/u). \end{aligned} \tag{9}$$

Note that the functions $\psi_n(u,x)$ are real-valued, which assures that the probabilities will be real-valued when the a_n are real. This is a consequence of the phase convention we have employed in Eq. 4, which ensures that the real and imaginary parts of the signal are even and odd, respectively. Moreover, Eq. 6 guarantees that the

total probability, i.e., the integral of the function $P(x)$, will be unity. See Appendix 1 for some additional properties of this basis.

2.1 Implementation

A set of a'_n coefficients can be estimated by solving a set of linear equations. To see this, we shall denote by $\mathbf{S}$ the M-dimensional vector of signal values. The mth component of this vector is $S_m = S(q_m)$. Similarly, an $M \times N$ dimensional matrix, $\mathbf{Q}$, can be defined with components $Q_{mn} = \phi_n(u, q_m)$. The estimation problem is reduced simply to a matrix equation $\mathbf{S} = \mathbf{Q}\mathbf{a}'$, where the N-dimensional vector of a'_n coefficients is denoted as $\mathbf{a}'$. In our implementation, this equation was solved by computing the pseudoinverse of $\mathbf{Q}$ via singular value decomposition [34]. S_0 is computed subsequently using Eq. 7, which is inserted into Eq. 6 to determine the coefficients a_n.

It is important to note that a prior estimate of u is necessary for the above estimation scheme. An adequate choice of u is necessary to obtain a reasonable approximation of the signal with few terms in the series. To this end, in our implementation, we first estimated a maximum value for u from the first few points of $S(q)$; in this range of q-values, the signal was assumed to undergo Gaussian attenuation. Note that when the signal is Gaussian, all coefficients except a_0 vanish, and the signal attenuation is given by $E(q) = \exp(-2\pi^2 q^2 u^2)$. Starting with this estimate of u, we gradually reduced it, and at each value of u, we estimated the a_n coefficients using the above scheme. Next, the signal attenuation values at the data points corresponding to the particular u and a_n values were computed. These values shall be denoted by $E^{\text{est}}(u, q)$. The mean error defined by

$$\epsilon(u) = \frac{1}{M} \sum_{i=1}^{M} (E^{\text{est}}(u, q_i) - E^{\text{data}}(q_i))^2, \tag{10}$$

was evaluated at each step, where $E^{\text{data}}(q_i)$ are the original data points. The search for the optimal u was discontinued when a local minimum was achieved, or $\epsilon(u)$ fell below 1×10^{-15}. The last set of u and a_n values were used in subsequent analysis.

The average probability estimates can be computed for arbitrary values of x using Eqs. 8–9. A return-to-yz-plane probability can be estimated either from the $x = 0$ point of $P(x)$, or directly from Eqs. 26 with 22. The moments, $\langle x^m \rangle$, can be computed from the coefficients a_n using Eq. 25.

Many examples of media of interest to the MR community exhibit certain levels of symmetry such as full isotropy or axial symmetry. For such specimens, one-dimensional $E(q)$ data is sufficient to reconstruct two- and three-dimensional average propagators in cases of axial symmetry and isotropy, respectively. It was shown that such higher-dimensional propagators may be more meaningful than the one-dimensional propagator [29]. For the case of axial symmetry, several general

relationships exist between the one- and two-dimensional propagators and their moments along with the expressions of the same quantities in terms of the a_n coefficients. These are derived in Appendix 3. Finally, Appendix 4 includes the derivations of similar relationships between one- and three-dimensional propagators for isotropic geometries.

3 Results

We test our estimation and reconstruction scheme on six different signal attenuations with analytically (i.e., exactly) available average probability profiles and moments. Table 1 includes the percentage deviations of the zero-displacement probabilities as well as even-order moments, estimated using the proposed series representation, from the exact ground-truth values for five of the geometries considered.

From its first days, PFG-MR techniques have been used to measure the bulk diffusion coefficients of fluids. In this case, the signal as well as the average propagator is Gaussian. Similarly, when the sample has two distinct, non-exchanging, but freely diffusing compartments, the signal and the average propagators can be written in terms of sums of two Gaussians. It has been shown in numerous studies that bi-exponential fits are quite satisfactory in modeling typical signal decays observed from real tissue [19, 20]. Therefore, it is very important for a new reconstruction scheme to fit mono- and bi-exponential decays. Figure 3 illustrates the performance of the approach in mono- and bi-exponential signal attenuations where the diffusion coefficient was taken to be $1.0 \times 10^{-3}\,\mathrm{mm^2/s}$ in the mono-exponential case. In the simulations of bi-exponential attenuation, the diffusion coefficients were taken to be $1.5 \times 10^{-3}\,\mathrm{mm^2/s}$ and $2.5 \times 10^{-4}\,\mathrm{mm^2/s}$ with volume fractions of 0.6 and 0.4, respectively. 33 sampling points were used, and a total of 12 even-ordered a_n coefficients were kept in the series representation under the assumption that the propagator is symmetric. The first row depicts the signal attenuation values (left) as well as the deviation of the estimated signals from the ground truth (right). In the case of mono-exponential attenuation, the scheme is exact up to numerical precision while the performance is very accurate for bi-exponential decay as well.

Table 1 Percent (%) deviations of the estimated quantities from their exact values

	Mono-exponential	Bi-exponential	Rectangular pore	Cylindrical pore	Spherical pore
S_0	3.0×10^{-14}	7.0×10^{-7}	4.2×10^{-12}	4.3×10^{-12}	1.9×10^{-13}
$P(0)$	5.7×10^{-13}	4.0×10^{-2}	3.3	1.7×10^{-1}	1.3×10^{-2}
$P_{2D}(0)$	4.0×10^{-12}	2.2×10^{-1}	–	4.3	–
$P_{3D}(0)$	1.9×10^{-11}	6.9×10^{-1}	–	–	5.7
$\langle x^0 \rangle$	0	4.4×10^{-14}	1.6×10^{-6}	7.4×10^{-7}	5.1×10^{-10}
$\langle x^2 \rangle$	4.1×10^{-13}	4.3×10^{-5}	5.1×10^{-5}	2.4×10^{-5}	1.0×10^{-7}
$\langle x^4 \rangle$	5.0×10^{-12}	5.6×10^{-4}	6.7×10^{-4}	2.9×10^{-4}	2.5×10^{-6}
$\langle x^6 \rangle$	3.4×10^{-11}	3.9×10^{-3}	6.7×10^{-3}	3.0×10^{-3}	3.8×10^{-5}

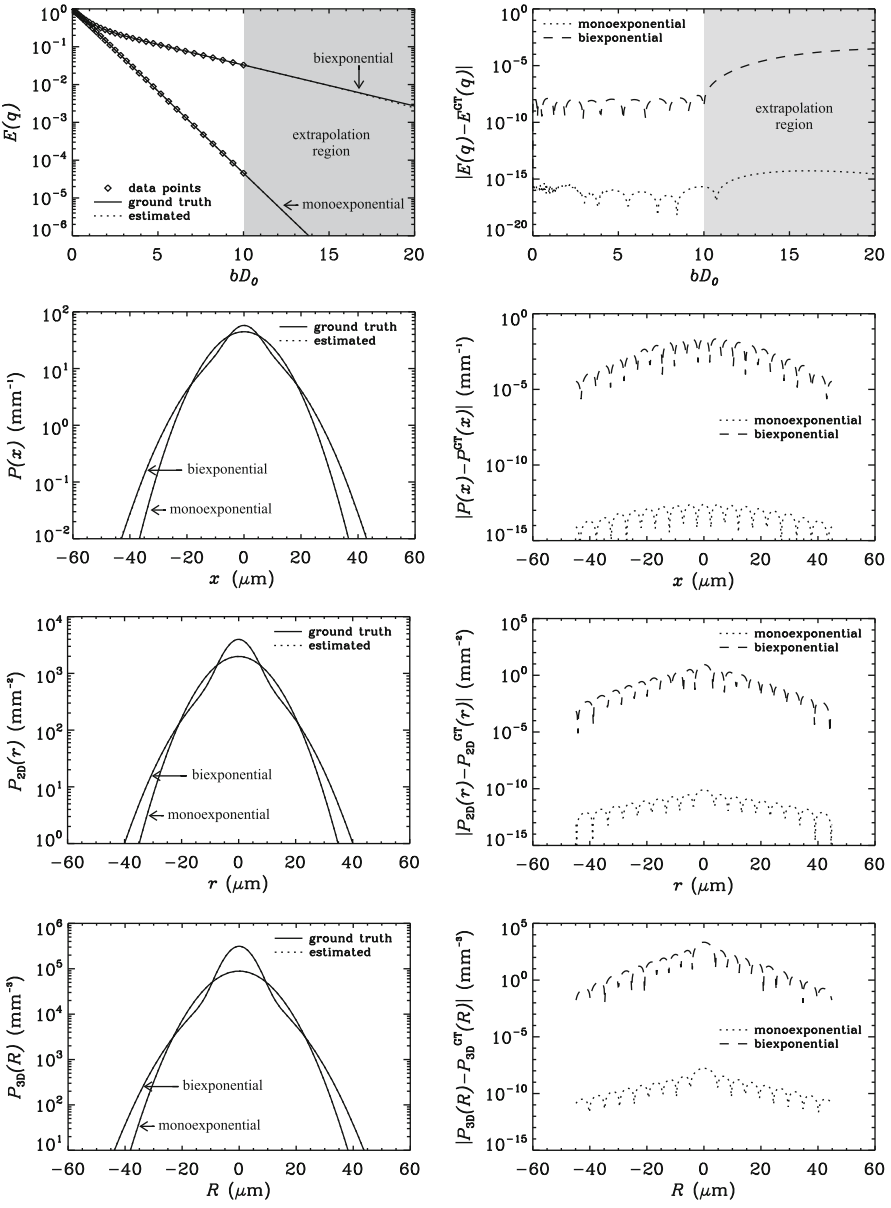


Fig. 3 Signal decay profiles and one-, two-, and three-dimensional propagators (*left column*, *from top to bottom*) including both the ground-truth and estimated curves from mono-exponential as well as bi-exponential diffusion. The *right column* shows the associated errors in the estimates. Note that the two- and three-dimensional propagators are symmetrized around the 0 radius to make comparisons with the one-dimensional propagator convenient

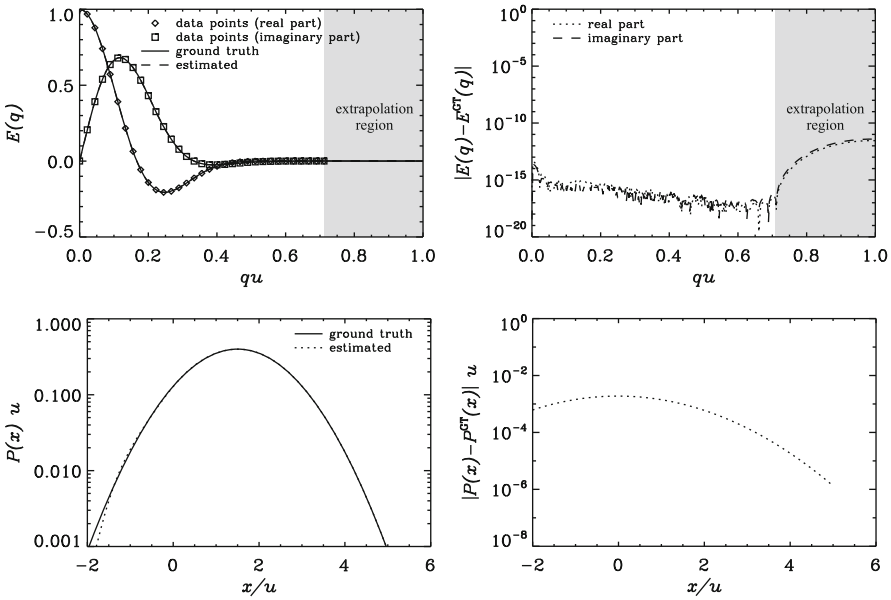


Fig. 4 Diffusion signal decay, and the corresponding average propagator (*left column*, *from top to bottom*) from the simulations of Gaussian diffusion with flow. Note that the signal is complex-valued due to flow, which yields a horizontal shift in the reconstructed average propagator. The *right column* shows the associated errors in the estimates

The figure demonstrates that the scheme yields not only a good approximation within the sampling window, but also a satisfactory extrapolation of the decay curve outside the sampling window. The second, third, and fourth rows of the figure illustrate the results obtained from the one-, two-, and three-dimensional Fourier transforms of the signal decay curves. The two-dimensional Fourier transform was performed under the assumption that the signal originated from an axially symmetric environment, while the three-dimensional Fourier transform assumed isotropy. In these cases, one-dimensional q-space data are sufficient to reconstruct these higher-dimensional propagators as detailed in Appendices 3 and 4. In all cases, the reconstructed propagators are indistinguishable from the ground-truth propagators.

To show the performance of the scheme for non-symmetric displacement probabilities, we simulated a flowing fluid with the assumption that the molecules undergo a net coherent displacement of $1.5u$. In the presence of flow, the expected signal attenuation is complex-valued, and the odd-ordered coefficients of the series in Eq. 3 have to be retained. Our simulations started with 33 complex data points and N was set to 23. The results shown in Fig. 4 indicate that the errors in the signal decay as well as in the reconstructed average propagator are negligible and the peak of the displacement probability shifted by the correct distance. Note that because of the lack of symmetry, the zero-displacement probabilities

are not meaningful for this set of simulations. Additionally, unlike in the other geometries considered, the propagator in the presence of flow has non-zero odd-ordered moments. Consequently, we did not include the percentage deviations from this simulation in Table 1. The percentage deviation of the estimated S_0 from the correct one was only 9.7×10^{-14}. The percent deviations in the moments $\langle x^0 \rangle$ through $\langle x^7 \rangle$ were 4.4×10^{-14}, 9.7×10^{-11}, 3.5×10^{-9}, 2.6×10^{-8}, 3.6×10^{-7}, 1.4×10^{-6}, 1.1×10^{-5}, and 3.0×10^{-5}.

Next we tackle three different scenarios of restricted diffusion. All of these simulations start with generating 33 data points, and a total of 28 terms in the series of Eq. 3 are kept. First, we simulate the signal attenuation from a one-dimensional geometry in which the molecules are trapped between two parallel plates separated from each other by a distance L. When the diffusion time is long, the diffraction-like features are apparent, which leads to a challenging signal decay profile to estimate. However, as shown in the first row of Fig. 5, the proposed basis performs well not only in the sampling window but also in the extrapolation region. Note the tremendous improvement over the results obtained from bi-exponential fitting as well as cumulant expansion as was illustrated in Fig. 1. This improvement was achieved in spite of the fact that the analytical form of the cumulant expansion was used and the bi-exponential fitting was performed on 400 data points. In contrast, our SHORE simulation was numerical and used less than 10% of the data points within the same window.

The corresponding propagator is given by a triangular function, which is not differentiable at three points. Although our basis is composed of smooth functions, the approximation is quite successful for this piecewise smooth function (see the second row of Fig. 5). Finally, in the last row of the same figure, we consider the diffraction pattern predicted for a double-PFG experiment [23]. In such experiments, rather than bouncing back from the horizontal axis, the signal decay is expected to cross the horizontal axis and become negative at exactly half the q-value of the corresponding single-PFG experiment [23]. The satisfactory performance of the approach for this more oscillatory signal attenuation profile suggests that the approximation can be used to model signal decays obtained from multiple PFG sequences.

The second simulated restricted diffusion scenario is diffusion taking place inside a cylinder of radius r_0. The simulations of this axially symmetric geometry were performed with identical parameters, and the two-dimensional axially symmetric Fourier transform was computed both exactly and also from the a_n coefficients as described in Appendix 3. The results are presented in Fig. 6. Finally in Fig. 7 we depict the results obtained from simulations of diffusion inside a sphere of radius R_0. For this geometry both the one- and three-dimensional average propagators are included. Note that the two- and three-dimensional average propagators obtained from the cylindrical and spherical pores resemble the triangular propagator obtained from the rectangular pore. This observation suggests that the true displacement profile is approximately linear in these geometries—a finding that cannot be gleaned by studying the form of one-dimensional propagators.

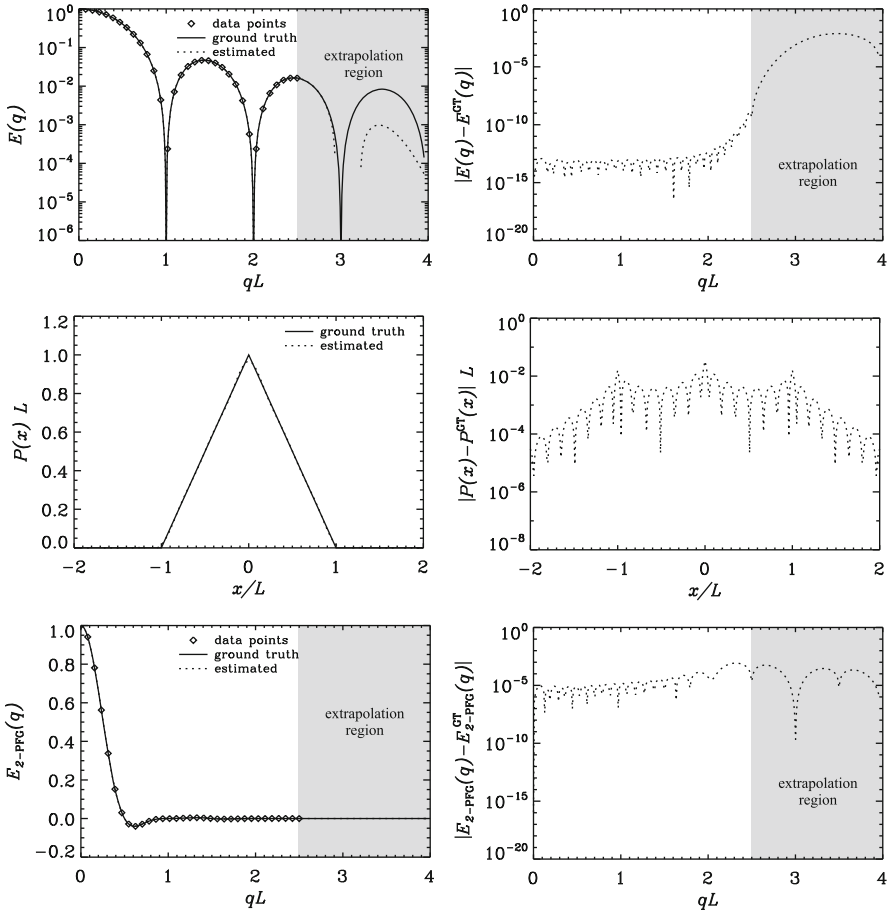


Fig. 5 Signal decay expected from a rectangular pore obtained via a single-PFG experiment and the error in its SHORE estimate (*top row*). These images are the SHORE counterpart of the cumulant expansion and bi-exponential fitting results presented in Fig. 1. The associated average propagator and the error in the reconstruction are shown in the *middle row*. The *bottom row* shows the signal expected from a double-PFG experiment. Note that unlike in the case of the single-PFG experiment, this plot is not logarithmic to accommodate negative portions of the curve

The results presented in Table 1 demonstrate the accuracy of the quantities that are computed directly from the a_n coefficients using the relations derived in the appendices. Note that zero-displacement probabilities are typically more difficult to estimate because of their sensitivity to the signal values over the entire q-axis. Therefore, extrapolations become more significant in these estimations. Similarly, the higher-order moments are related to the higher-order derivatives of the signal decay at the origin. Therefore, the accuracy in the estimates of the moments is an indication of the accuracy in the derivatives of $E(q)$ at $q = 0$. Finally, we would

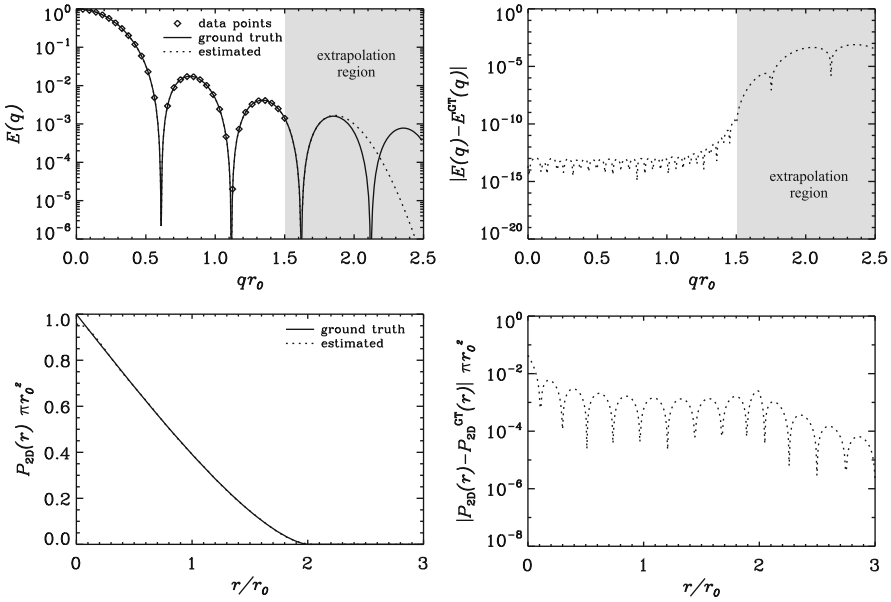


Fig. 6 Signal decay curve and two-dimensional axially symmetric propagator (*left column*, *from top to bottom*) for the cylindrical pore with radius r_0. The associated errors are depicted in the *right column*

like to note that because the one-dimensional moments are directly proportional to the two- and three-dimensional radial moments, as implied by Eqs. 41 and 56, the percent deviations in the radial moments are identical to those in one-dimensional moments. Consequently, these deviations are not included in Table 1.

4 Discussion

We would like to point out that the two- and three-dimensional average propagators were not qualitatively different from their one-dimensional counterparts in the simulations of mono- and bi-exponential attenuations. However, in the case of restricted diffusion, e.g., inside a spherical pore, as seen in Fig. 7, the propagator obtained from the three-dimensional Fourier transform resembles the one-dimensional triangular propagator for diffusion inside a rectangular pore although the one-dimensional propagator of the spherical pore is smoother and Gaussian-like. This is an indication that the violation of the Gaussian phase approximation is more severe in one-dimensional geometries, because in one-dimensional propagators of higher-dimensional pores, displacements are projected onto one of the axes leading to a smoothing effect in such environments. Moreover, since the propagator is the autocorrelation function of the shape function, it is

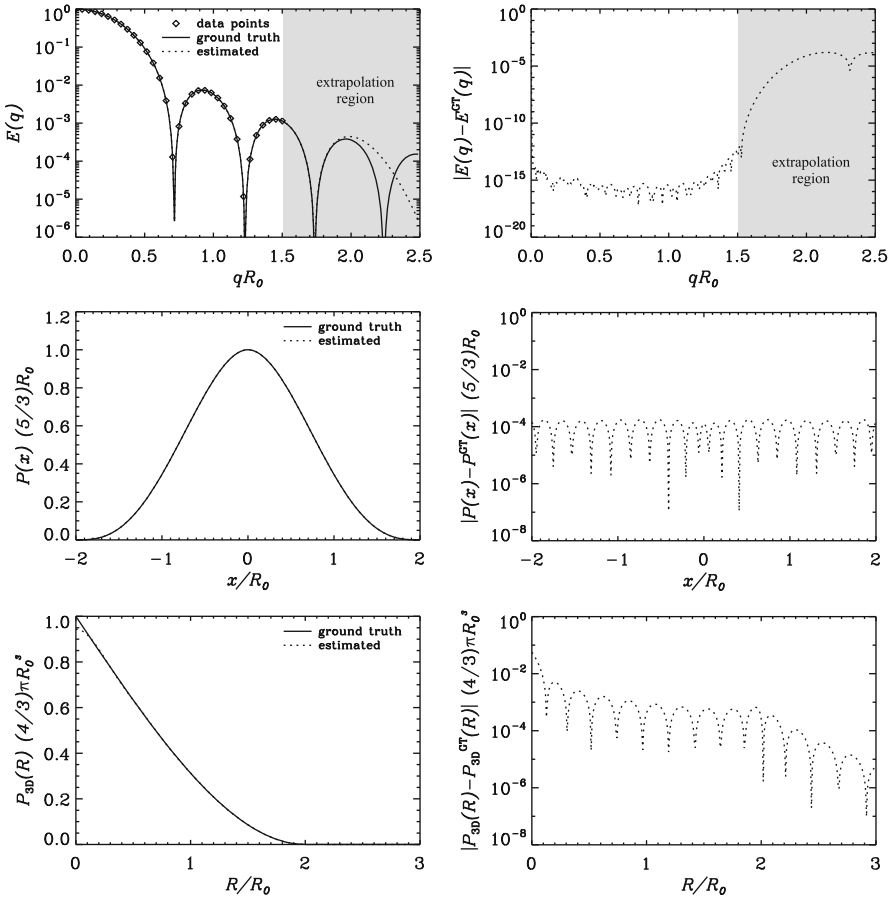


Fig. 7 Signal decay profile and one-dimensional and three-dimensional isotropic propagator (*left column*, *from top to bottom*) for the spherical pore with radius R_0. The associated errors in the estimates are included in the *right column*

straightforward to prove that, in closed pores, the zero-displacement probability is just the reciprocal of the pore "volume." As can be seen in Eqs. 29, 46, and 61, this was exactly the case for the zero-displacement values of the one-, two-, and three-dimensional propagators for rectangular, cylindrical, and spherical pores, respectively. The $x = 0$ values of the one-dimensional propagators for cylindrical and spherical pores (see Eqs. 47 and 63) suggest that there may not be such a shape-independent relation between the $P(x = 0)$ value of a higher-dimensional geometry and the shape of the pore.

The technique we have presented here is linear because the estimated coefficients, a'_n, are expressible as a linear combination of the signal values. In fact, to estimate the coefficients of the series representation, we posed the problem as a matrix equation. The scheme demands an a priori estimate of the characteristic

length u. This can be done in many different ways; the approach we have described in the Theory section starts by fitting the small-q section of the $E(q)$ data to a Gaussian decay curve, i.e., $E(q) \approx \mathrm{e}^{-2\pi^2 q^2 u^2}$, and gradually reducing the u-value until the mean error, i.e., the average squared deviation of the signal attenuation estimates from the original values, reaches a minimum or drops below a very small value comparable to the machine precision. Note that the estimation of a maximum value for u from the first few points of the signal profile is nonlinear, and because it is performed on fewer data points, the estimates are prone to error. However, the completeness of the employed basis, regardless of u, guarantees the convergence of the series although a deviation in the estimated u-value from its ground-truth value may affect the rate of this convergence. In our simulations, we observed that even 10% error in the estimation of u was tolerable and did not change the quality of the results significantly.

As discussed above, the constant u is a characteristic length proportional to the square root of the diffusion time. Because the basis is symmetric under the interchange of u and q, the same formalism can be applied to data obtained by varying the diffusion time while keeping the diffusion gradient strength fixed.

The analyses we have presented have focused entirely on one-dimensional data and we have provided the details of the 1D-SHORE framework introduced for the first time in Ref. [27]. Although we presented results from reconstructions of two- and three-dimensional propagators, these results were based on axially symmetric or isotropic geometries, respectively. In these geometries, having the q-space data along one direction is tantamount to having it on the entire plane or within the entire three-dimensional space, making it possible to compute the higher-dimensional propagators using one-dimensional transforms (see Eqs. 33 and 49). However, because of the separation of variables of the higher-dimensional analog of the simple harmonic oscillator Hamiltonian in Eq. 11, our scheme has a trivial extension to two- and three-dimensional q-space signals even in the absence of axial symmetry or isotropy as we showed in [30]. A similar approach was introduced in Ref. [1] that uses Gauss–Laguerre functions. We envision that representing the MR signal attenuation analytically in a series of orthogonal functions will have many applications. Most recently, the 1D-SHORE framework was shown to be useful in accurately estimating the moments of the underlying compartment size distributions, which could be employed to obtain new forms of MR image contrast [31].

5 Conclusion

We have introduced a new basis to represent one-dimensional q-space signal and reconstruct the average propagators from it. The basis is well known in quantum mechanics, but some characteristics of the basis make it particularly relevant to and useful for q-space MR. Among these is its capability to accommodate

complex-valued signals while ensuring a real and normalized average propagator. Additionally, the Fourier transform of each component is readily available making it possible to reconstruct the average propagators immediately. Similarly, useful descriptors of the propagator such as return-to-origin probabilities and its moments can be computed from the basis representation. On several simulations, the accuracy of the estimations was assessed, and we demonstrated that it successfully represents signal profiles even when the signal and the propagators have unusual forms such as in the presence of diffraction-like features. Unlike the previously employed cumulant expansion, multi-exponential fitting, and similar approaches, the basis functions are complete and orthogonal, and the estimation/reconstruction scheme is linear with a wider range of applicability.

Acknowledgements This research was supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) at the National Institutes of Health (NIH) and the Department of Defense in the Center for Neuroscience and Regenerative Medicine (CNRM).

Appendix 1: Remarks on the Basis Functions

Note that the functions $\psi_n(u, x)$, defined in Eq. 9, are the solutions to the eigenvalue equation

$$\left(-u^2 \frac{\partial^2}{\partial x^2} + \frac{x^2}{u^2}\right) \psi_n(u, x) = \lambda_n \, \psi_n(u, x), \tag{11}$$

with eigenvalues $\lambda_n = (2n + 1)$. Taking the Fourier transform of both sides, it is easy to show that the functions $\phi_n(u, q)$ are the solutions to the eigenvalue equation

$$\left(-\frac{1}{(2\pi u)^2} \frac{\partial^2}{\partial q^2} + (2\pi u)^2 q^2\right) \phi_n(u, q) = \lambda_n \, \phi_n(u, q). \tag{12}$$

Since Eqs. 11 and 12 are identical upon the transformations $x \longrightarrow q$ and $u \longrightarrow (2\pi u)^{-1}$, ψ_n and ϕ_n have the same form up to a multiplicative factor (see Eq. 9). In fact, the operator on the left-hand side of these eigenvalue equations is the Hamiltonian operator with a quadratic potential, which describes the simple harmonic oscillator problem in quantum mechanics. However, our definitions of the eigenfunctions are slightly different from their forms as commonly used in quantum mechanics. Specifically, our basis is not normalized, but the scaling is such that when diffusion is Gaussian, $a_n = \delta_{n0}$, where δ_{ij} is the Kronecker delta.

Despite these minor differences from the basis used in quantum mechanics, our basis functions still satisfy the relationships

$$A \, \psi_n(u, z) = \begin{cases} 0 & , n = 0 \\ \sqrt{n} \, \psi_{n-1}(u, z), & n \geq 1 \end{cases} \tag{13}$$

and

$$\tilde{A}\,\phi_n(u,q) = \begin{cases} 0 & ,n=0 \\ \sqrt{n}\,\phi_{n-1}(u,q), & n \geq 1 \end{cases}, \tag{14}$$

where A and $\tilde{A}$ are the "lowering operators" defined by

$$A = \frac{1}{\sqrt{2}}\left(\frac{z}{u} + u\frac{\mathrm{d}}{\mathrm{d}z}\right) \tag{15}$$

and

$$\tilde{A} = \frac{i}{\sqrt{2}}\left(2\pi u q + \frac{1}{2\pi u}\frac{\mathrm{d}}{\mathrm{d}q}\right). \tag{16}$$

Writing the Polynomials in the Definitions of $E(q)$ and $P(x)$ in Power Series

It is possible to show that the Hermite polynomials can be written as [8]

$$H_n(x) = \sum_{m=0,2,4,\ldots}^{n} (-1)^{m/2}\frac{2^{n-m}\,n!}{(n-m)!\,(m/2)!}\,x^{n-m}. \tag{17}$$

Inserting this expression into Eq. 5, the following power series expansion for the signal decay is obtained:

$$E(q) = \mathrm{e}^{-2\pi^2 q^2 u^2}\sum_{n=0}^{N-1} a_n \sum_{m=0,2,4,\ldots}^{n} \frac{i^{-n-m}\,2^{-m+n/2}\,\sqrt{n!}}{(n-m)!\,(m/2)!}\,(2\pi q u)^{n-m}. \tag{18}$$

The double summation in the above expression can be recast by using the transformations $k = n - m$ ($k = 0, 1, 2, \ldots, N-1$) and $l = m$ ($l = 0, 2, \ldots, N-k-1$) as shown in Fig. 8:

$$E(q) = \mathrm{e}^{-2\pi^2 q^2 u^2}\sum_{k=0}^{N-1} b_{Nk}(u)\,q^k, \tag{19}$$

where

$$b_{Nk}(u) = \frac{i^{-k}\,(2\pi u)^k}{k!}\sum_{l=0,2,\ldots}^{N-k-1}\frac{\sqrt{2^{k-l}\,(k+l)!}}{(l/2)!}\,a_{k+l}. \tag{20}$$

Using the same approach, the propagator can be written as

$$P(x) = \mathrm{e}^{-x^2/(2u^2)}\sum_{k=0}^{N-1} c_{Nk}(u)\,x^k, \tag{21}$$

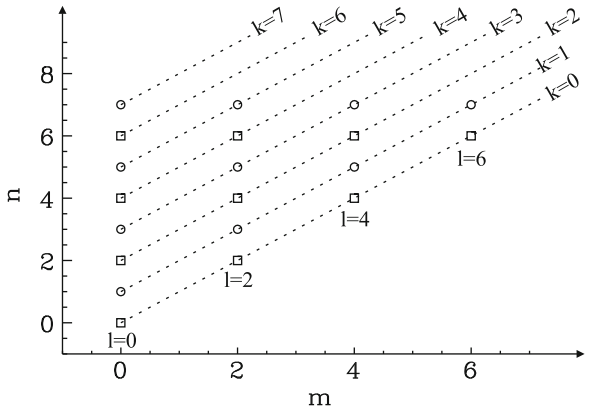


Fig. 8 Transformation of the indices n and m into k and l. Note that $N = 8$ in this figure

where

$$c_{Nk}(u) = \frac{1}{\sqrt{2\pi}\, u^{k+1}\, k!} \sum_{l=0,2,\ldots}^{N-k-1} (-1)^{l/2}\, \frac{\sqrt{2^{k-l}\,(k+l)!}}{(l/2)!}\, a_{k+l}. \tag{22}$$

Moments of $P(x)$ and $P(0)$

The mth-order moment of $P(x)$ is defined to be

$$\langle x^m \rangle = \int_{-\infty}^{\infty} \mathrm{d}x\, x^m\, P(x)\ . \tag{23}$$

The usual strategy to compute the moments of a propagator from its $E(q)$ profile involves the power series expansion of the plane wave in the Fourier relationship between the signal and the probability, i.e.,

$$\begin{aligned} E(q) &= \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}^{-\mathrm{i}2\pi q x}\, P(x) \\ &= 1 + \frac{(-i\,2\pi q)}{1!}\langle x \rangle + \frac{(-i\,2\pi q)^2}{2!}\langle x^2 \rangle + \frac{(-i\,2\pi q)^3}{3!}\langle x^3 \rangle + \ldots \end{aligned} \tag{24}$$

Therefore, a power series representation of $E(q)$ data upon a term-by-term comparison with Eq. 24 would yield the moments of $P(x)$.

However, the moments can be computed directly using the Hermite function representation of the $E(q)$ profile as well. This can be done by inserting Eqs. 21–22 into 23 yielding

$$\langle x^m \rangle = u^m \sum_{k=0,2,\ldots}^{N-1} \frac{(k+m-1)!!}{k!} \sum_{l=0,2,\ldots}^{N-k-1} (-1)^{l/2} \frac{\sqrt{2^{k-l}\,(k+l)!}}{(l/2)!}\, a_{k+l}, \tag{25}$$

when m is even. Odd-ordered moments can be computed using essentially the same expression where the index k takes odd values, i.e., $k = 1, 3, 5, \ldots$.

Note that $P(0)$ can be evaluated conveniently by setting $x = 0$ in Eq. 21, i.e.,

$$P(0) = c_{N0}(u). \tag{26}$$

Note that $P(x = 0)$ is not a true return-to-origin probability, but, since $P(x)$ is obtained through a one-dimensional Fourier transform, it is the probability for the molecules to return to the yz-plane—a consequence of the Fourier slice theorem.

Appendix 2: The Rectangular Pore

When the spins are trapped between two infinite plates, one located at $x = 0$ and the other at $x = L$, the expected signal intensity at long diffusion times is given by [40]

$$E^{\text{rect}}(q) = \frac{\sin^2(\pi q L)}{(\pi q L)^2}, \tag{27}$$

where it is implied that the diffusion gradients are applied perpendicular to the infinite plates. The corresponding average propagator is

$$P^{\text{rect}}(x) = \begin{cases} \dfrac{L - |x|}{L^2}, |x| \leq L \\ 0, |x| > L \end{cases}. \tag{28}$$

Obviously, the return-to-yz-plane probability is simply

$$P^{\text{rect}}(0) = \frac{1}{L}. \tag{29}$$

Finally, the even-order moments of the propagator are given as

$$\langle x^m \rangle^{\text{rect}} = \frac{2L^m}{(m+1)\,(m+2)}, \tag{30}$$

while the odd-ordered moments vanish.

Appendix 3: Axially Symmetric Geometries and the Cylindrical Pore

General Results

Many geometries of interest have an anisotropic structure with an oblate or prolate shape, where the environment possesses a symmetry axis. In our treatment we shall take the z-axis to be along this symmetry axis. In such an axially symmetric or transversely isotropic process, the same signal attenuation profile is obtained when the diffusion gradient is applied in any direction (which defines the x-axis in our treatment) perpendicular to the symmetry axis. In this case, a two-dimensional isotropic Fourier transform can be evaluated from one-dimensional q-space data, i.e.,

$$P_{2\mathrm{D}}(\mathbf{r}) = \int_{-\infty}^{\infty} \mathrm{d}q_x \int_{-\infty}^{\infty} \mathrm{d}q_y \, \mathrm{e}^{\mathrm{i}2\pi\mathbf{q}\cdot\mathbf{r}} \, E(q), \tag{31}$$

where the two-dimensional vectors $\mathbf{q}$ and $\mathbf{r}$ reside on the xy-plane. The radial and polar coordinates of these vectors shall be denoted to be (q, θ_q) and (r, θ_r), respectively. Inserting the Rayleigh expansion for two-dimensional plane waves,

$$\mathrm{e}^{\mathrm{i}2\pi\mathbf{q}\cdot\mathbf{r}} = \sum_{m=-\infty}^{\infty} i^m \, J_m(2\pi qr) \, \mathrm{e}^{\mathrm{i}m(\theta_r - \theta_q)}, \tag{32}$$

into Eq. 31, the two-dimensional isotropic propagator for axially symmetric environments is obtained to be

$$P_{2\mathrm{D}}(r) = 2\pi \int_0^{\infty} \mathrm{d}q \, q \, J_0(2\pi qr) \, E(q). \tag{33}$$

The same analysis can be repeated for the inverse Fourier transform, yielding

$$E(q) = 2\pi \int_0^{\infty} \mathrm{d}r \, r \, J_0(2\pi qr) \, P_{2\mathrm{D}}(r). \tag{34}$$

The one-dimensional average propagator, obtained from a one-dimensional Fourier transform, is related to the two-dimensional propagator via the relation

$$\begin{aligned} P(x) &= \int_{-\infty}^{\infty} \mathrm{d}y \, P_{2\mathrm{D}}(x, y) \\ &= 2 \int_{|x|}^{\infty} P_{2\mathrm{D}}(r) \, \frac{r}{\sqrt{r^2 - x^2}} \, \mathrm{d}r \ , \end{aligned} \tag{35}$$

which is a consequence of the Fourier slice theorem. Clearly, the above expression is just the Abel transform [3] of $P_{2\mathrm{D}}(r)$. Therefore, the inverse Abel transform

of the one-dimensional projection reveals the two-dimensional axially symmetric propagator to be

$$P_{2\mathrm{D}}(r) = -\frac{1}{\pi}\int_r^\infty \frac{P'(x)}{\sqrt{x^2-r^2}}\,\mathrm{d}x. \tag{36}$$

The return-to-yz-plane probability can be estimated from the two-dimensional axially symmetric propagator:

$$P(x=0) = 2\int_0^\infty \mathrm{d}r\; P_{2\mathrm{D}}(r). \tag{37}$$

On the other hand, a return-to-z-axis probability can be calculated by setting $r = 0$ in Eq. 33, i.e.,

$$P_{2\mathrm{D}}(0) = 2\pi\int_0^\infty \mathrm{d}q\, q\, E(q). \tag{38}$$

The radial moments of the two-dimensional axially symmetric propagator are defined as

$$\langle r^m\rangle_{2\mathrm{D}} = 2\pi\int_0^\infty \mathrm{d}r\, r^{m+1}\, P_{2\mathrm{D}}(r). \tag{39}$$

Similar to what is done in Eq. 24, the Bessel function in Eq. 34 can be written as a power series, yielding

$$E(q) = 1 - \frac{(2\pi q)^2}{2^2}\langle r^2\rangle_{2\mathrm{D}} + \frac{(2\pi q)^4}{(2\cdot 4)^2}\langle r^4\rangle_{2\mathrm{D}} - \frac{(2\pi q)^6}{(2\cdot 4\cdot 6)^2}\langle r^6\rangle_{2\mathrm{D}} + \ldots. \tag{40}$$

A term-by-term comparison of the series in Eqs. 24 and 40 suggests that the radial moments are given in terms of the one-dimensional moments by the relationship

$$\langle r^m\rangle_{2\mathrm{D}} = \frac{m!!}{(m-1)!!}\langle x^m\rangle. \tag{41}$$

Note that this relationship holds only when m is even; axial symmetry implies that odd-ordered moments of the one-dimensional propagator, $\langle x^m\rangle$, will vanish.

Estimates in Terms of a_n Coefficients

Inserting Eq. 19 into Eq. 33 yields [8]

$$P_{2\mathrm{D}}(r) = \sum_{k=0}^{N-1} b_{Nk}(u)\,\frac{\Gamma(k/2+1)}{2^{k/2+1}\,\pi^{k+1}\,u^{k+2}}\;{}_1F_1\left(\frac{k}{2}+1,\,1,\,-\frac{r^2}{2u^2}\right), \tag{42}$$

where ${}_1F_1(\alpha,\gamma;z)$ is the confluent hypergeometric function of the first kind. Standard computational libraries do not include an implementation of these functions.

However, a simple and accurate implementation can be performed by exploiting the recurrence relation [8]

$$\alpha\, {}_1F_1(\alpha+1,\gamma;z) = (z+2\alpha-\gamma)\, {}_1F_1(\alpha,\gamma;z) + (\gamma-\alpha)\, {}_1F_1(\alpha-1,\gamma;z). \quad (43)$$

Since ${}_1F_1(\alpha,\gamma;0) = 1$, a return-to-$z$-axis probability can be computed simply by summing up the factors before the confluent hypergeometric function in Eq. 42. Note that the radial moments, $\langle r^m\rangle_{2D}$, can be computed from a_n by using Eq. 25 along with Eq. 41.

The Cylindrical Pore

In this section we shall consider restricted diffusion within a cylinder of radius r_0, which is an example of an axially symmetric process. The MR signal attenuation is given by [17]

$$E^{\mathrm{cyl}}(q) = \left(\frac{J_1(2\pi q r_0)}{\pi q r_0}\right)^2. \quad (44)$$

By inserting Eq. 44 into 33, the two-dimensional axially symmetric propagator can be evaluated to be

$$P_{2D}^{\mathrm{cyl}}(r) = \begin{cases} \dfrac{4\cos^{-1}\left(\frac{r}{2r_0}\right) - \frac{r}{r_0}\sqrt{4-\left(\frac{r}{r_0}\right)^2}}{2\pi^2 r_0^2}, & r \le 2r_0 \\ 0, & r > 2r_0 \end{cases}. \quad (45)$$

It immediately follows that the return-to-z-axis probability is given by

$$P_{2D}^{\mathrm{cyl}}(0) = \frac{1}{\pi r_0^2}. \quad (46)$$

Moreover, the return-to-yz-plane probability was calculated, by inserting Eq. 45 into Eq. 37, to be

$$P^{\mathrm{cyl}}(x=0) = \frac{16}{3\pi^2 r_0}. \quad (47)$$

Finally, the radial moments are given by

$$\langle r^m\rangle_{2D}^{\mathrm{cyl}} = \frac{2^{m+4}\,(m+1)!!}{(m+2)\,(m+4)!!}\, r_0^m. \quad (48)$$

Note that $\langle x^m\rangle$ can be calculated by inserting this expression into Eq. 41.

Appendix 4: Isotropic Geometries and the Spherical Pore

General Results

Many specimens of interest in pulsed-field-gradient MR are isotropic. Even in the presence of local anisotropy [22, 24, 28], the randomness in the shape and orientation of the pores would lead to isotropy due to the averaging of signals from individual compartments. In such environments, having the q-space data with diffusion gradients applied along a single direction is tantamount to having the data all across the three-dimensional q-space. Therefore, it is possible to characterize the entire average propagators and related parameters via one-dimensional sampling. In fact, the resulting three-dimensional isotropic propagator can be computed through the relationship

$$P_{3\mathrm{D}}(R) = \frac{2}{R}\int_0^\infty \mathrm{d}q\, q\, \sin(2\pi qR)\, E(q), \tag{49}$$

which is obtained by inserting the Rayleigh expansion of three-dimensional plane waves [26]

$$\mathrm{e}^{\mathrm{i}2\pi\mathbf{q}\cdot\mathbf{R}} = 4\pi\sum_{l=0}^{\infty} i^l\, j_l(2\pi qR)\sum_{m=-l}^{l} Y_{lm}(\mathbf{R}/R)\, Y_{lm}(\mathbf{q}/q)^* \tag{50}$$

into the 3D Fourier transform relationship between $E(\mathbf{q})$ and $P(\mathbf{R})$ in Eq. 1, where $q = |\mathbf{q}|$, $R = |\mathbf{R}|$, and $\mathbf{q}$ and $\mathbf{R}$ are three-dimensional vectors. The Fourier slice theorem enables establishment of the relation between the three-dimensional isotropic propagator and one-dimensional propagator:

$$\begin{aligned} P(x) &= \int_{-\infty}^{\infty}\mathrm{d}y\int_{-\infty}^{\infty}\mathrm{d}z\, P_{3\mathrm{D}}(\sqrt{x^2+y^2+z^2}) \\ &= 2\pi\int_0^\infty \mathrm{d}\rho\,\rho\, P_{3\mathrm{D}}(\sqrt{\rho^2+x^2}) \\ &= 2\pi\int_{|x|}^\infty \mathrm{d}R\, R\, P_{3\mathrm{D}}(R). \end{aligned} \tag{51}$$

Here the first step involves the change of variables $\rho^2 = y^2 + z^2$. Similarly, the transformation $R^2 = x^2+\rho^2$ was employed in the second step. Taking the derivative of both sides with respect to x and subsequently employing the fundamental theorem of calculus, one obtains

$$P_{3\mathrm{D}}(R) = \left(-\frac{1}{2\pi x}\frac{\mathrm{d}P(x)}{\mathrm{d}x}\right)\Bigg|_{x=R}. \tag{52}$$

Eq. 51 implies that the return-to-yz-plane probability can be calculated using the relationship

$$P(x = 0) = 2\pi \int_0^\infty \mathrm{d}R\, R\, P_{3\mathrm{D}}(R). \tag{53}$$

Note that Eq. 49 leaves the return-to-origin probability undetermined, which should be calculated using the relationship

$$P_{3\mathrm{D}}(0) = 4\pi \int_0^\infty \mathrm{d}q\, q^2\, E(q). \tag{54}$$

The radial moments of the three-dimensional isotropic propagator are defined as

$$\langle R^m \rangle_{3\mathrm{D}} = 4\pi \int_0^\infty \mathrm{d}R\, R^{m+2}\, P_{3\mathrm{D}}(R). \tag{55}$$

Inserting Eq. 52 into the above expression and performing integration by parts, it is straightforward to show that the radial moments of the three-dimensional isotropic propagator and the moments of the one-dimensional propagator are related through the relationship

$$\langle R^m \rangle_{3\mathrm{D}} = (m + 1)\, \langle x^m \rangle. \tag{56}$$

Note that this relationship holds only when m is even. odd-ordered moments of the one-dimensional propagator, $\langle x^m \rangle$, vanish due to isotropy.

Estimates in Terms of a_n Coefficients

By inserting the expansion of the one-dimensional propagator in Eq. 8 into Eq. 52 and differentiating by using the relationships in Eqs. 13 and 15, one can expand the three-dimensional isotropic propagator as

$$P_{3\mathrm{D}}(R) = \sum_{n=0}^{N-1} a_n\, \xi_n(u, R), \tag{57}$$

where

$$\xi_n(u, R) = \begin{cases} \dfrac{1}{2\pi u^2}\, \psi_0(u, R) & , n = 0 \\ \dfrac{1}{2\pi u^2}\, \psi_n(u, R) - \sqrt{\dfrac{n}{2}}\, \dfrac{1}{\pi u R}\, \psi_{n-1}(u, R) & , n \geq 1 \end{cases}. \tag{58}$$

Note that the return-to-origin probability can be estimated from the coefficients a_n by setting $R = 0$ in the above expression. Finally, the radial moments, $\langle R^m \rangle_{3D}$, can be computed from a_n by using Eq. 25 along with Eq. 56.

The Spherical Pore

Diffusion inside a spherical pore of radius R_0 yields the following MR signal attenuation at long diffusion times [40]:

$$E^{\text{sph}}(q) = \left[\frac{3}{(2\pi q R_0)^2} \left(\frac{\sin(2\pi q R_0)}{2\pi q R_0} - \cos(2\pi q R_0) \right) \right]^2. \tag{59}$$

By inserting Eq. 59 into 49, one can evaluate the three-dimensional isotropic propagator to be

$$P_{3D}^{\text{sph}}(R) = \begin{cases} \dfrac{3\,(2R_0 - R)^2\,(4R_0 + R)}{64\pi R_0^6}, & R \le 2R_0 \\ 0, & R > 2R_0 \end{cases}. \tag{60}$$

It immediately follows that the return-to-origin probability is given by

$$P_{3D}^{\text{sph}}(0) = \frac{3}{4\pi R_0^3}. \tag{61}$$

The one-dimensional propagator can be obtained via a one-dimensional Fourier transform of $E^{\text{sph}}(q)$ or by inserting Eq. 60 into Eq. 51. In either case, it is given by

$$P^{\text{sph}}(x) = \begin{cases} \dfrac{3\,(2R_0 - |x|)^3\,(4R_0^2 + 6R_0|x| + x^2)}{160R_0^6}, & |x| \le 2R_0 \\ 0, & |x| > 2R_0 \end{cases}, \tag{62}$$

which implies that the return-to-yz-plane probability is given by

$$P^{\text{sph}}(x = 0) = \frac{3}{5R_0}. \tag{63}$$

Since isotropic geometries are also axially symmetric, the expressions derived in Appendix 3 apply also to this appendix. For brevity, we shall include only the result for the return to long-axis probability predicted for spherical pores:

$$P_{2D}^{\text{sph}}(0) = \frac{9}{8\pi R_0^2}. \tag{64}$$

Finally, the radial moments are given by

$$\langle R^m \rangle_{3D}^{sph} = \frac{9 \cdot 2^{m+3}}{m^3 + 13m^2 + 54m + 72} R_0^m. \quad (65)$$

Note that $\langle x^m \rangle$ can be calculated using this expression along with Eq. 56.

References

1. Assemlal, H.E., Tschumperlé, D., Brun, L.: Efficient and robust computation of PDF features from diffusion MR signal. Med. Image Anal. **13**(5), 715–729 (2009). DOI 10.1016/j.media.2009.06.004. URL http://dx.doi.org/10.1016/j.media.2009.06.004
2. Bennett, K.M., Schmainda, K.M., Bennett (Tong), R., Rowe, D.B., Lu, H., Hyde, J.S.: Characterization of continuously distributed cortical water diffusion rates with a stretched-exponential model. Magn. Reson. Med. **50**(4), 727–734 (2003)
3. Bracewell, R.N.: The Fourier Transform and Its Applications. McGraw-Hill, New York (1978)
4. Callaghan, P.T., Coy, A., MacGowan, D., Packer, K.J., Zelaya, F.O.: Diffraction-like effects in NMR diffusion studies of fluids in porous solids. Nature **351**, 467–469 (1991)
5. Cohen, Y., Assaf, Y.: High b-value q-space analyzed diffusion-weighted MRS and MRI in neuronal tissues—a technical review. NMR Biomed. **15**, 516–542 (2002)
6. Cory, D.G., Garroway, A.N.: Measurement of translational displacement probabilities by NMR: An indicator of compartmentation. Magn. Reson. Med. **14**(3), 435–444 (1990)
7. Frøhlich, A.F., Østergaard, L., Kiselev, V.G.: Effect of impermeable boundaries on diffusion-attenuated MR signal. J. Magn. Reson. **179**(2), 223–233 (2006). DOI 10.1016/j.jmr.2005.12.005. URL http://dx.doi.org/10.1016/j.jmr.2005.12.005
8. Gradshteyn, I.S., Ryzhik, I.M.: Table of Integrals, Series and Products. Academic, New York (2000)
9. Hahn, E.L.: Spin echoes. Phys. Rev. **80**, 580–594 (1950)
10. Hürlimann, M.D., Schwartz, L.M., Sen, P.N.: Probability of return to origin at short times: A probe of microstructure in porous media. Phys. Rev. B **51**(21), 14,936–14,940 (1995)
11. Jensen, J.H., Helpern, J.A., Ramani, A., Lu, H., Kaczynski, K.: Diffusional kurtosis imaging: the quantification of non-Gaussian water diffusion by means of magnetic resonance imaging. Magn. Reson. Med. **53**, 1432–1440 (2005)
12. Jian, B., Vemuri, B.C., Özarslan, E., Carney, P.R., Mareci, T.H.: A novel tensor distribution model for the diffusion-weighted MR signal. NeuroImage **37**(1), 164–176 (2007). DOI 10.1016/j.neuroimage.2007.03.074. URL http://dx.doi.org/10.1016/j.neuroimage.2007.03.074
13. Kärger, J., Heink, W.: The propagator representation of molecular transport in microporous crystallites. J. Magn. Reson. **51**(1), 1–7 (1983)
14. Kiselev, V.G., Il'yasov, K.A.: Is the "biexponential diffusion" biexponential? Magn. Reson. Med. **57**(3), 464–469 (2007). DOI 10.1002/mrm.21164. URL http://dx.doi.org/10.1002/mrm.21164
15. Köpf, M., Metzler, R., Haferkamp, O., Nonnenmacher, T.F.: NMR studies of anomalous diffusion in biological tissues: Experimental observation of Lévy stable processes. In: Losa, G.A., Merlini, D., Nonnenmacher, T.F., Weibel, E.R. (eds.) Fractals in Biology and Medicine, vol. 2, pp. 354–364. Birkhäuser, Basel (1998)
16. Liu, C.L., Bammer, R., Moseley, M.E.: Generalized diffusion tensor imaging (GDTI): A method for characterizing and imaging diffusion anisotropy caused by non-Gaussian diffusion. Isr. J. Chem. **43**(1–2), 145–154 (2003)
17. McCall, D.W., Douglass, D.C., Anderson, E.W.: Self-diffusion studies by means of nuclear magnetic resonance spin-echo techniques. Ber. Bunsenges. Phys. Chem. **67**, 336–340 (1963)

18. Mitra, P.P., Sen, P.N.: Effects of microgeometry and surface relaxation on NMR pulsed-field-gradient experiments: simple pore geometries. Phys. Rev. B **45**, 143–156 (1992)
19. Mulkern, R.V., Gudbjartsson, H., Westin, C.F., Zengingönül, H.P., Gartner, W., Guttmann, C.R., Robertson, R.L., Kyriakos, W., Schwartz, R., Holtzman, D., Jolesz, F.A., Maier, S.E.: Multi-component apparent diffusion coefficients in human brain. NMR Biomed. **12**(1), 51–62 (1999)
20. Niendorf, T., Dijkhuizen, R.M., Norris, D.G., van Lookeren Campagne, M., Nicolay, K.: Biexponential diffusion attenuation in various states of brain tissue: implications for diffusion-weighted imaging. Magn. Reson. Med. **36**(6), 847–857 (1996)
21. Ohanian, H.C.: Principles of Quantum Mechanics. Prentice-Hall, Englewood Cliffs (1990)
22. Özarslan, E.: Compartment shape anisotropy (CSA) revealed by double pulsed field gradient MR. J. Magn. Reson. **199**(1), 56–67 (2009). DOI 10.1016/j.jmr.2009.04.002. URL http://dx.doi.org/10.1016/j.jmr.2009.04.002
23. Özarslan, E., Basser, P.J.: MR diffusion—"diffraction" phenomenon in multi-pulse-field-gradient experiments. J. Magn. Reson. **188**(2), 285–294 (2007). DOI 10.1016/j.jmr.2007.08.002. URL http://dx.doi.org/10.1016/j.jmr.2007.08.002
24. Özarslan, E., Basser, P.J.: Microscopic anisotropy revealed by NMR double pulsed field gradient experiments with arbitrary timing parameters. J. Chem. Phys. **128**(15), 154,511 (2008). DOI 10.1063/1.2905765. URL http://dx.doi.org/10.1063/1.2905765
25. Özarslan, E., Basser, P.J., Shepherd, T.M., Thelwall, P.E., Vemuri, B.C., Blackband, S.J.: Observation of anomalous diffusion in excised tissue by characterizing the diffusion-time dependence of the MR signal. J. Magn. Reson. **183**(2), 315–323 (2006). DOI 10.1016/j.jmr.2006.08.009. URL http://dx.doi.org/10.1016/j.jmr.2006.08.009
26. Özarslan, E., Shepherd, T.M., Vemuri, B.C., Blackband, S.J., Mareci, T.H.: Resolution of complex tissue microarchitecture using the diffusion orientation transform (DOT). NeuroImage **31**(3), 1086–1103 (2006). DOI 10.1016/j.neuroimage.2006.01.024. URL http://dx.doi.org/10.1016/j.neuroimage.2006.01.024
27. Özarslan, E., Koay, C.G., Basser, P.J.: Simple harmonic oscillator based estimation and reconstruction for one-dimensional q-space MR. In: Proceedings of the International Society for Magnetic Resonance in Medicine, vol. 16, p. 35 (2008)
28. Özarslan, E., Nevo, U., Basser, P.J.: Anisotropy induced by macroscopic boundaries: Surface-normal mapping using diffusion-weighted imaging. Biophys. J. **94**(7), 2809–2818 (2008). DOI 10.1529/biophysj.107.124081. URL http://dx.doi.org/10.1529/biophysj.107.124081
29. Özarslan, E., Koay, C.G., Basser, P.J.: Remarks on q-space MR propagator in partially restricted, axially-symmetric, and isotropic environments. Magn. Reson. Imaging **27**(6), 834–844 (2009). DOI 10.1016/j.mri.2009.01.005. URL http://dx.doi.org/10.1016/j.mri.2009.01.005
30. Özarslan, E., Koay, C.G., Shepherd, T.M., Blackband, S.J., Basser, P.J.: Simple harmonic oscillator based reconstruction and estimation for three-dimensional q-space MRI. In: Proceedings of the International Society for Magnetic Resonance in Medicine, vol. 17, p. 1396 (2009)
31. Özarslan, E., Shemesh, N., Koay, C.G., Cohen, Y., Basser, P.J.: Nuclear magnetic resonance characterization of general compartment size distributions. New J. Phys. **13**, 015,010 (2011)
32. Pawula, R.F.: Approximation of the linear Boltzmann equation by the Fokker-Planck equation. Phys. Rev. **162**, 186–188 (1967)
33. Pfeuffer, J., Provencher, S.W., Gruetter, R.: Water diffusion in rat brain in vivo as detected at very large b values is multicompartmental. MAGMA **8**(2), 98–108 (1999)
34. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C: The Art of Scientific Computing. Cambridge Press, Cambridge (1992)
35. Schwartz, S.C.: Estimation of probability density by an orthogonal series. Ann. Math. Statist. **38**, 1261–1265 (1967)
36. Silva, M.D., Helmer, K.G., Lee, J.H., Han, S.S., Springer, C.S., Sotak, C.H.: Deconvolution of compartmental water diffusion coefficients in yeast-cell suspensions using combined T_1 and diffusion measurements. J. Magn. Reson. **156**(1), 52–63 (2002). DOI 10.1006/jmre.2002.2527. URL http://dx.doi.org/10.1006/jmre.2002.2527

37. Stejskal, E.O., Tanner, J.E.: Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. J. Chem. Phys. **42**(1), 288–292 (1965)
38. Stejskal, E.O.: Use of spin echoes in a pulsed magnetic-field gradient to study anisotropic, restricted diffusion and flow. J. Chem. Phys. **43**(10), 3597–3603 (1965)
39. Stepišnik, J.: Analysis of NMR self-diffusion measurements by a density matrix calculation. Phys. B & C **104**, 350–364 (1981)
40. Tanner, J.E., Stejskal, E.O.: Restricted self-diffusion of protons in colloidal systems by the pulsed-gradient, spin-echo method. J. Chem. Phys. **49**(4), 1768–1777 (1968)
41. Walter, G.G.: Properties of Hermite series estimation of probability density. Ann. Statist. **5**, 1258–1264 (1977)
42. Yablonskiy, D.A., Bretthorst, G.L., Ackerman, J.J.: Statistical model for diffusion attenuated MR signal. Magn. Reson. Med. **50**(4), 664–669 (2003)

Fourier Blues: Structural Coloration of Biological Tissues

Richard O. Prum and Rodolfo H. Torres

Abstract The non-pigmentary colors of the tissues of living organisms are produced by the physical interaction of light with nanostructures in the tissues. Contrary to what has been previously assumed for many decades, it has been established now that many of the beautiful blue and green colors observed in the tissues of mammals, birds, and butterflies are the result of coherent scattering or constructive interference. Using Fourier analysis one can show that many structurally colored tissues are *quasi-ordered* on the appropriate nanoscale to produce the observed colors by constructive interference. Understanding the mechanisms of coloration in animals is very important because of the role that bright colors play in communication, courtship display, and mate selection in many species of the animal kingdom. In this note we give an exposition of some of the extensive work done recently on nanomaterials with noncrystalline, local scale order. The focus of this article is, in particular, on a truly fascinating manifestation of Fourier analysis and synthesis in nature, which provides a way to explain coloration phenomena that are of interest in behavioral and evolutionary biology.

Keywords Fourier Transform • FFT • Quasi-order • Nano-scale • Nano-structure(s)(d) • Crystallography • Coloration • Scattering • Iridescent • Interference • Electron micrograph(s) • X-ray(s) • Bragg's law • Rayleigh scattering • Benedek

R.O. Prum
Department of Ecology and Evolutionary Biology, Yale University,
New Haven, CT 06520-8105, USA
e-mail: richard.prum@yale.edu

R.H. Torres (✉)
Department of Mathematics, University of Kansas, Lawrence, KS 66045-7594, USA
e-mail: torres@math.ku.edu

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_20,

1 Introduction

The study of the forms of coloration in different materials is a rich, intricate, and multidisciplinary activity. The classic book by Nassau [15] presents a detailed account of at least fifteen different forms of coloration found in our physical world. From a scientific point of view, the explanation of the origin of the colors observed belongs mainly to the métiers of physics and chemistry, but the implications of the presence of coloration in different materials extend to many other disciplines. In particular, coloration as mean of communication plays a crucial role in many areas of biology and the study of species capable of analyzing the complicated color signals. Among such species are certainly humans, and color and coloration play a central role in many situations extending from the scientific, through the practical, to the aesthetic aspects of our lives. Colors allow us to discover and understand physico/chemical phenomena taking place both at microscopic scales invisible to our eyes and at intergalactic distance in our universe; they code, guide, warn, and help us in many aspects of our everyday lives, and they are also capable to stimulate our minds, provoke emotions, and move our souls through the plastic arts.

For biologists it has become clear that the analysis of the mechanisms of coloration, their functions, and evolution can only be studied in an integrative way if one is to fully understand the amazing color displays in many species. In particular, birds are animals capable of communicating though coloration, and the analysis of bird coloration in recent times has refocused some of its efforts to this comprehensive approach. We refer to the two volumes [11] for an extensive account of some of the state of the art in the subject.

Mathematics could not be absent in the explanation of the phenomena of coloration. It is not only present as the universal language of science, but also, through the powerful lenses of Fourier analysis, it provides new explanations and understanding of certain forms of colorations. In this expository note, we will describe some of the developments in which we have been involved in our interdisciplinary collaborations in [19–27]. We will illustrate how Fourier analysis naturally appears in the theoretical formulation of mechanisms of structural coloration through coherent scattering. We will concentrate here on aspects of the coloration of the skin of birds, but the same tools and techniques apply to the study of feathers and other tissues of living organisms.

1.1 Bird Coloration

Both chemical pigments and the physical aspects of the wavelike behavior of light are responsible for the coloration of birds. Pigments have the property of absorbing and emitting selective wavelengths of the ambient light. The resulting colors are determined by the molecular structure of the pigments. Such pigments may be synthesized by the birds themselves or acquired by the birds through their

diet. By removing the pigmentary substance from the tissues the colors disappear, verifying that the pigments are the cause of the coloration. A typical example of pigmentary coloration is provided by flamingos, whose recognizable pink color tends to fade out in captivity through a modification of their diet from the one they have in the wild. Likewise, the black or brown colors of the feathers of a crow or a robin are produced by melanin pigments synthesized by the animal, just as in human black or red hair.

Unlike pigmentary colors (usually yellows, oranges, reds, browns, and blacks) structurally produced colors in avian tissues (often blues and greens) are the result of the physical interaction of light with optical heterogeneities of the tissues. Incoherent Rayleigh scattering has been erroneously assumed to be responsible for the observed non-pigmentary colors of many birds. Rayleigh (or Tyndall) scattering occurs when small, light-scattering objects are randomly distributed without a spatial pattern in the path of the light. Small objects will preferentially scatter smaller wavelengths, giving rise to a bluish or violet color. This mechanism is the explanation for the color of the blue sky. According to this conception of biological structural color, small melanin granules present in the feathers or skin of bird tissues will reflect back short waves, such as violet and blue, but will let pass through longer waves such as red and yellow. The physical and biological literature in the subject can be found in the classical works [10,13,15,35]. A key feature of Rayleigh scattering is that it lacks iridescence or color change with angle of observation, so it was originally applied to all the biological examples of structural color that lack iridescence. Nevertheless, the Rayleigh scattering hypothesis was never supported by spectrophotometric data or microscopic observation of the tissues.

The Rayleigh hypothesis was questioned by Raman [28] in the thirties, but his speculations that color in a certain bird from southern India was produced by constructive interference were dismissed because of the lack of crystalline structure of the bird tissue; see [17]. Dyck [6, 7] in the 1970s was the first to document that the reflectance spectrum of many bird feathers presents a clear peak within the visible spectrum matching the color observed. This is in contradiction with the continuous increase of energy distribution in the direction of the ultraviolet (UV) part of the spectrum that Rayleigh scattering would produce. It was not until the turn of the century that a new explanation for non-iridescent coloration in many animal tissues emerged. The more recent research has established that most greens, blues, and violets observed in birds are in fact structural colors produced by coherent scattering.

1.2 Fourier Analysis Comes in to the Picture

The new explanations about coloration involve Fourier analysis and follow a model by Benedek [1]. The first use of these techniques was our study of the blue feather barbs of a South American bird called the Plum-throated Coating, *Cotinga maynana* (Cotingidae) [22]. The intense blue color of the cotinga is produced by closely

packed spherical air bubbles in the protein of the feathers. This was followed by numerous other works in the study of many other types of structurally colored tissues. The tissues that have been analyzed by now present a big diversity of nanostructures at scales comparable to the wavelengths of visible light. A certain order or periodicity in these structures permits a predictable phase relationship between the light waves scattered and the coherent scattering of certain reinforced specific wavelengths.

Traditionally, the classification of the color-producing structural tissues has been based on the particular physical model used to explain idealized perfectly periodic structures similar to the ones in nature. Nanomaterials may be periodic (or crystalline) in one, two, or three dimensions. These highly periodic materials produce iridescent colors which change in hue with the angle of observation, as typically seen in hummingbirds or peacocks. However, quasi-ordered materials lack periodicity at longer spatial scales, but are still substantially ordered at local spatial scales. There were no traditional physical methods for analysis of constructive interference by materials with only local order, which led to the application of Fourier analysis to the problem.

In our approach we use Fourier analysis to study the geometric nanostructure of 2D transmission electron microscope images of these color-producing tissues. This gives as a frequency content analysis of the images that we use to produce a prediction or modeling of the coherent scattering behavior of the tissue. We also compare these predictions with the reflectance spectrum of the colorful tissues measured with a spectrophotometer. The reflectance spectrum gives the relative intensity of energy at different bandwidths within the visible spectrum.

The use of Fourier analysis in the study of structured materials has a long history. For example, the structure of crystals and quasicrystals can be studied by looking at the diffraction patterns obtained when a crystalline material is illuminated with X-rays. Mathematically, this essentially accounts for the analysis of the Fourier transform of the characteristic function of the crystal or the density of mass function. We refer the reader to the book [29] for a very nice introduction to the subject.

Some of the patterns obtained in crystallography can be explained by *Bragg's law*, named after the only father-son team of Nobel laureates. They were the first to describe the phenomena of X-ray diffraction by crystals [2]. In very ordered materials two parallel incident electromagnetic waves will bounce from scatterers in the material and arrive at a distant observer with a lag in phase produced by the different distances traveled (path addition). This difference in phase produces the reinforcement of certain waves with appropriate wave numbers and the cancelation of others. For this to happen, the wavelengths and the physical distances defining the ordered structures in the material have to be of comparable size. Simple trigonometry shows that for light incident at an angle θ onto parallel atomic planes separated by a distance d, the first peak of diffraction takes place for wavelengths λ given by Bragg's law:

$$\lambda = 2d \sin\theta. \tag{1}$$

It is important to note that (1) relates a physical dimension of the illuminated material with the wavelength of the light.

For very short X-ray waves with wavelengths of the order of $10^{-10}m$ Bragg's effect takes place at the atomic level. Diffraction photographs of crystals produced very ordered patterns proving the existence of a very particular arrangement of the atoms in the material. In crystallography one has to deal with an inverse problem. The structure of a crystal, or at least its symmetries, is to be determined by looking at the crystal spectrum, i.e., the patterns in their Fourier transforms.

In biological tissues, structural color production takes place at a much larger spatial scale than the inter atomic distance in crystals. Nevertheless, the situation is similar to Bragg's law. As expressed by Benedek in [1], it is a general principle that

> "*. . .light is scattered only by those fluctuations in the index of refraction whose wavelengths are larger than one-half of the wavelength of the light in the medium.*"

The structure in the material originating those fluctuations can clearly be observed in electronic microscope images of the tissues, and their Fourier spectrum can be computed and related to the spectral measurements made with a spectrophotometer. Essentially, the predominant spatial periodicity of the tissues, as quantified by the Fourier transform, gives a prediction of the wavelengths of light scattered the most. The direct problem of computing the Fourier transform is simpler than the inverse problem of crystallography and can be carried out numerically using the fast Fourier transform (FFT). (But this truth has interesting biological implications, i.e., there are multiple biological nanostructures that can make the same color!)

1.3 More About Color

In describing colors and forms of coloration it is convenient to recall the difference between the production of color by addition or subtraction, which sometimes produces some confusions. Coloration by addition is the result of the combination of light of different wavelengths. For example, the superposition of red and blue lights over a white screen produces the so-called color magenta. If one adds light of its complementary color, green, one obtains white light. On the other hand, the coloration produced in the presence of pigments is due to color subtraction. The color attributed to a pigment, the one observed, is the one complementary to the one absorbed. For example, if we mix a green pigment (one that absorbs blue and red) with a magenta one (one that absorbs green) the result is black.

It is important also to recall that the visible spectrum of humans ranges approximately between 400 nm and 700 nm. Our optical systems posses three color

receptors most sensible to different sets of wavelengths around the red, green, and blue colors. It is the combined excitation of these receptors together with the amount of luminosity and ambient light conditions that determines the final interpretation of colors that our brain makes of certain electromagnetic waves. The colors observed in birds due to coherent scattering result from the constructive superposition of the wavelengths scattered the most. In the study of the resulting hues observed and measured by spectrophotometry, the rules of coloration by addition take place. However, unlike our ears, which let our brain distinguish between each individual note played as part of a cord, our visual system only interprets the final result of the superposition of light of different wavelengths. That is, the same perceived color can be created by addition in different ways.

It is interesting to note that birds have a broader visible spectrum with a fourth receptor and are able to see into the UV (320–400 nm) part of the electromagnetic spectrum [11]. It is perhaps impossible for us to image how do the colors seen by birds actually look like to them because of this ability to see UV ones, but we can still study the full spectral content of the signals. This detailed spectrum, undetected by our eyes but measurable by a spectrophotometer and predicted by our Fourier analysis, is what helps us explain the physical mechanisms taking place in the production of the color.

In the rest of this expository article we chose to describe some of the physical and mathematical models employed in the description of structural colors in the skin of some birds. The same models apply to feathers and other living tissues. We refer to the already cited literature for more technical details. This note also overlap in part with a more elementary exposition translated into Spanish presented in [32].

2 A Physical Model for Coherent Scattering

To explain how Fourier methods can be used to predict the color produced, we based our analysis on some of the work in [1, 33, 34] and the references therein. A mathematical and physical explanation of the transparency of the human cornea (a biological tissues similar in structure to the wattles of some birds), as well as the reasons of its turbidity due to swollen pathological abnormalities, was given by Benedek in [1]. The cornea is made of long and thin parallel collagen fibers immersed in a ground substance of mucopolysaccharide. A cross section of a bundle of such fibers looks very much like the cross section of the tissues of some birds, though at a smaller scale. According to Benedek, Maurice [14] was one of the first researchers to realize that, to explain the transparency of the cornea, it was important to understand the relationship among the phases of waves scattered by each of the fibers in the tissue. Maurice first speculated that the fibers should be equal in diameter and have their longitudinal axis centered on the points in a perfect lattice. Maurice thought the absence of the perfect crystalline periodicity in electron micrographs of the cornea was an experimental artifact, but soon it was realized that the corneal collagen fibers were not arranged in a perfect crystal lattice which

required a new theoretical explanation. A series of experimental, numerical, and theoretical works culminated then with Benedek's explanation that a perfect lattice arrangement is not necessary. Again, fluctuations in the index of refraction whose wavelength are equal to or larger than one-half the wavelength in the medium of an incident light are responsible for most of the coherently scattered light. In the case of the cornea these fluctuations from the fibers to the ground substance in which they are immersed are of very small physical dimensions and produce most of the scattered energy at very small wavelengths [33]. Wavelengths in the visible part of the spectrum are then almost completely transmitted, giving the transparency of the cornea.

The fibers in the tissue can be modeled as very long and thin cylinders or rather needles. Benedek described the propagation of a scattered electromagnetic field in the plane perpendicular to these fibers. Because of the particular geometric arrangement, further physical considerations imply that most of the scattered field by each fiber propagates only in this plane, and a two-dimensional analysis is a reasonable approximation to the physical situation. A brief and simplified description to illustrate the arguments in [1] is as follows. To model the situation, imagine then a distribution of point masses M_j at positions x_j in the plane and an incident light wave

$$E(x,t) = E_0 \mathrm{e}^{\mathrm{i}(k_0 x - \omega t)}, \tag{2}$$

where the two-dimensional wave vector k_0 has length

$$|k_0| = 2\pi\eta/\lambda, \tag{3}$$

η is the mean index of refraction, λ is the wavelength of the incident beam, and ω is the angular time frequency of the incident light. The incident electric field induces oscillating dipoles in the medium which in turn irradiate new electric fields in every direction, and also part of the field is transmitted. The scattered field at a particular position in the plane is determined by the superposition of all the individual scattered fields. The rays emanating from different fibers travel different distances to a given fixed point. Moreover, the oscillations induced by the incident field at the different positions R_j take place at different times producing also a retardation in time. Appropriately using this time delay and path addition, the field scattered by M_j at a position R in the direction given by the vector k_R, with $|k_R| = |k_0|$ and forming an angle θ with the incident wave, is computed in [1] to be

$$E_j = E_0 \mathrm{e}^{\mathrm{i}(k_0 R - \omega t)} \mathrm{e}^{-\mathrm{i}k x_j}. \tag{4}$$

Here $k = k_0 - k_R$ is called the scattering vector and

$$|k| = 2|k_0| \sin(\theta/2) = \frac{4\pi\eta}{\lambda} \sin(\theta/2). \tag{5}$$

Or, in terms of wavelengths, we have that

$$\lambda = 2\eta\lambda_k \sin(\theta/2), \tag{6}$$

where

$$\lambda_k = \frac{2\pi\eta}{|k|}. \tag{7}$$

The total scattered field is then given by

$$E_T = E_0 \mathrm{e}^{\mathrm{i}(k_0 R - \omega t)} \sum_j \mathrm{e}^{-\mathrm{i}k x_j}. \tag{8}$$

Note that, formally using delta distributions and the Fourier transform, the last factor in (8) is the interference function which can be seen as a Fourier transform

$$I(k) = \sum_j \widehat{\delta}_{x_j}(k) = \widehat{(\Sigma_j \delta_{x_j})}(k) = \widehat{f}(k), \tag{9}$$

and where

$$f = \sum_j \delta_{x_j} \tag{10}$$

can be viewed as a density distribution of mass.

The intensity of the scattered light is proportional to the square of the scattered electric field. Thus, as argued by Benedek, the intensity will be large for those spacial frequencies k so that

$$|I(k)|^2 = |\widehat{f}(k)|^2 \tag{11}$$

is large. For example, when we measure backward scattering (that is the one back to a distant observer) which correspond $\theta = \pi$, the scattering will be very intense if f has a large Fourier component with wavelength

$$\lambda_k = \lambda/(2\eta), \tag{12}$$

i.e., *half the wavelength of the wavelength in the medium of the incident light.* This is a restatement of Bragg's law in this context, which permits again to relate the wavelength of the constructively reinforced scattered light with a physical dimension in the material.

Like with crystals or quasicrystals, if the density function f is very *ordered*, then the Fourier transform $\hat{f}$ will show clear *peaks* at certain frequencies. Loosely speaking (see [29]), a quasicrystal can be defined to be a countable set Λ such that there exits another (dual) set $\widehat{\Lambda}$, with the property that

$$\widehat{(\Sigma_{x_j \in \Lambda} \delta_{x_j})} = \sum\nolimits_{y_j \in \widehat{\Lambda}} \delta_{y_j} + \text{“small continuous spectrum.”} \tag{13}$$

When computed numerically the size of the Fourier transform of a quasicrystal reveals very high values at the (approximate) positions in the set $\widehat{\Lambda}$ with the continuously distributed spectrum as a background noise.

The physical description we gave above is only an approximation of the real situation. In the case of many tissue we no longer have very thin needles or even fibers. The density f in Benedek's theory is then replaced by the fluctuations in density from an average value. We refer again the reader to [1]. The density function f can be seen in the electronic microscope observations of the tissues. The predominant components in the Fourier transform of the density function are what we still claim determine to some extend and via (12) the hue and the distribution of energy observed in the spectrophotometer.

In tissues that lack a perfect crystal structure, the observed peaks will not necessarily be on a lattice, but they will still occur around a certain characteristic frequency within the visible spectrum. To determine theoretically the exact position of such peaks would require a precise knowledge of the dimensions and arrangements of the fibers. Because of the diversity of tissues and variations in specimens making assumptions about the exact diameter and position of fibers is too rigid to model many real-life situations, and we perform instead a numerical calculation of the Fourier transform. It is not possible to characterize all functions which will produce a noticeable peak in their spectra within a certain bandwidth. We are only interested in a particular scale that affects the distribution of energy of the scattered light in the visible part of the spectrum. What we want to corroborate is that the numerous tissues examined do posses the necessary order to produce such peaks.

3 Fourier Analysis of Nano-Structured Tissues and Color Prediction

We illustrate this application of Fourier analysis with some results already in the literature. Our first study on bird's skin was from the brilliantly colored patches around and above the eyes of a small group of perching birds from Madagascar—the asities (Eurylaimidae, Aves)—shown in Fig. 1 which we reproduce from [23]. As described in [23] the asities are a group of suboscine perching fruit and nectar-feeding birds endemic to the tropical forests of Madagascar. Adult males of the asities have brilliantly colored, sexually dimorphic facial skin during the breeding season. The colorful patches of facial skin play an important role in inter-sexual communication and mate choice of these birds.

The caruncles in the dermis of these tissues (Fig. 1b) are composed of numerous bundles of macrofibrils arrays of long parallel collagen fibers of similar diameters and separated by a mucopolysaccharide matrix. At large scales, the macrofibrils have little apparent order and run through the tissue in different directions. However, cross sections at 10k–50k magnification of any macrofibril reveal the circular shape of the cross sections of the parallel collagen fibers and their uniform distribution.

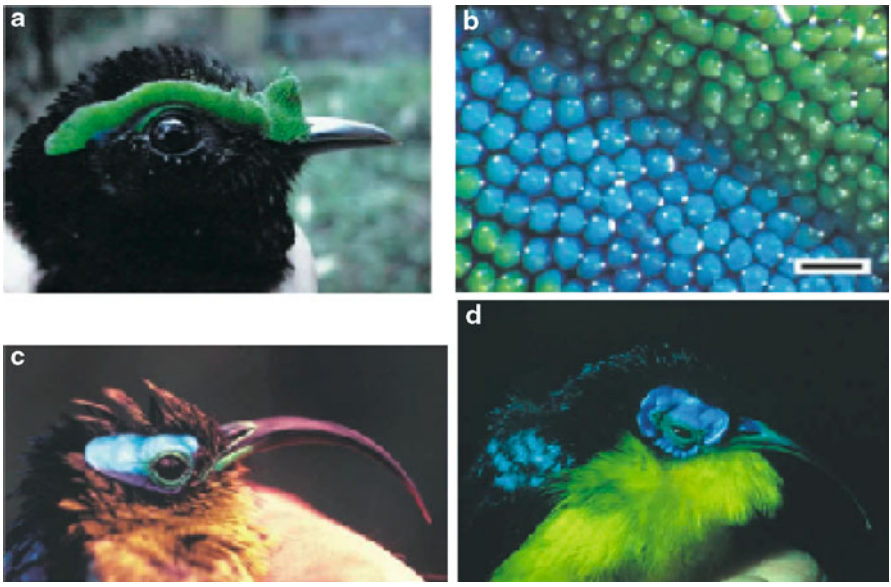


Fig. 1 *Blue and green* facial caruncles of asities. (**a**) *Phileipitta castanea*. (**b**) Close-up of the supraorbital caruncles of *Phileipitta castanea*. Scale bar approximately 500 μm. (**c**) *Neodrepanis coruscans*. (**d**) *Neodrepanis hypoxantha*

Figure 2 (also reproduced from [23]) shows a cross section of the collagen fibers of a typical tissue and the corresponding (modulus square of the) FFT of the image. The fibers show small variations in their diameters and center- to- center distances. The fibers in this image are not arranged in a crystal-like array, and the Fourier transforms shows certain concentric ring structures and have a radial-like symmetry (although it obviously cannot be perfectly radially symmetric). The intensity of the rings decreases as we move away from the origin on the Fourier transform domain. Intuitively the images can be thought as being made up of certain predominant periodicities of a particular length in every direction. The location of peaks in the side of the Fourier transform indicates that the fluctuations in the density function are rather homogeneous and similar in all directions in the tissues at least at a particular small scale. This is clearly observed in the images of the tissues. We call this arrangements in the tissue a *quasi-order*. The distance between nearest neighboring fibers does not change much from place to place, though there is very little correlation among fibers that are further away.

A big diversity of tissues and their corresponding FFTs can be seen in Figs. 3 and 4 below. They are from a larger study of many other birds that we carried out in [20]. Interestingly, some tissues analyzed do present an almost perfectly periodic structure which is clearly present too in the FFT of the images of the tissues. Note, for example, the image of tissue from *Philepitta castanea* (bird photo in Fig. 1a),

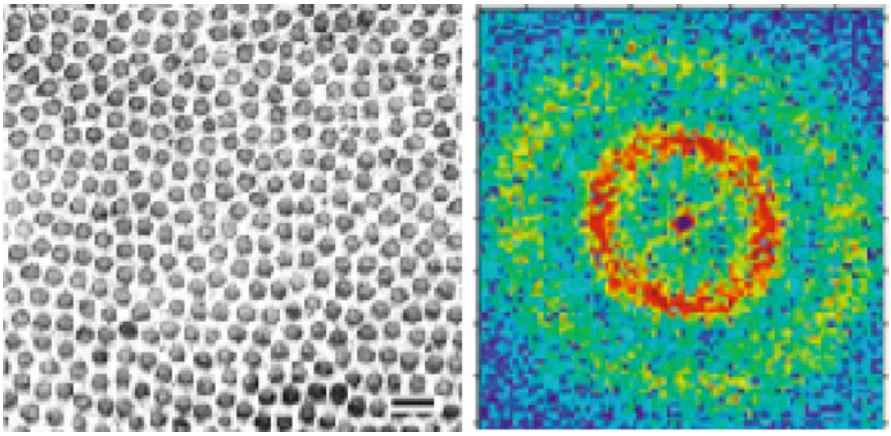


Fig. 2 Typical transmission electron micrograph of a cross section of an array of collagen fibers from the caruncle tissues. Scale bar approximately 200 nm. The colors map from *blue* to *red* indicate the magnitude of the squared Fourier components. *Blur* indicates small values and *red* high ones

which is given in Fig. 3i, and its FFT given in Fig. 4f. (For photos of the other birds mentioned in the figures please see [20].)

What is observed in the FFT images can be explained as follows: The radially decay of intensity is determined, in part, by the fact that the fibers are not needles whose cross sections are determined by delta masses at certain points but rather have approximately circular cross sections of a certain radius R. Though the physical problem is different, the mathematical problem of computing the Fourier transform of such collection of circles (or rather the characteristic function of the cross section) is equivalent to determine the Franunhofer diffraction patterns produced by a number of circular apertures of similar size.

Let B_0 be a circle of radius R centered at the origin in two-dimensional Euclideian space and let χ_{B_0} be its characteristic function. Let B_j be the circle with same radius but with center translated to the point x_j and let χ_{B_j} be the corresponding characteristic function. The Fourier transform of a the images of the tissue is then the Fourier transform of the characteristic function of a collection of circles $\{B_j\}_{j=0}^N$. Using the properties of the Fourier transform this is easily computed to be

$$\widehat{\Sigma_{j=0}^N \chi_{B_j}}(y) = \hat{\chi}_{B_0}(y)\, \widehat{\Sigma_{j=0}^N \delta_{x_j}}. \tag{14}$$

Therefore, the Fourier transform is determined by the product of two factors. One is determined by the shape of the aperture, while the other is determined only by the position of them. For a circle, the first factor is a radially decaying or damped wave. The sum of deltas in (14) is part of what is sometimes called a Dirac comb. For appropriate distribution of the deltas, the modulus square of the Fourier transform of them presents very high peaks (almost new deltas) at a particular position.

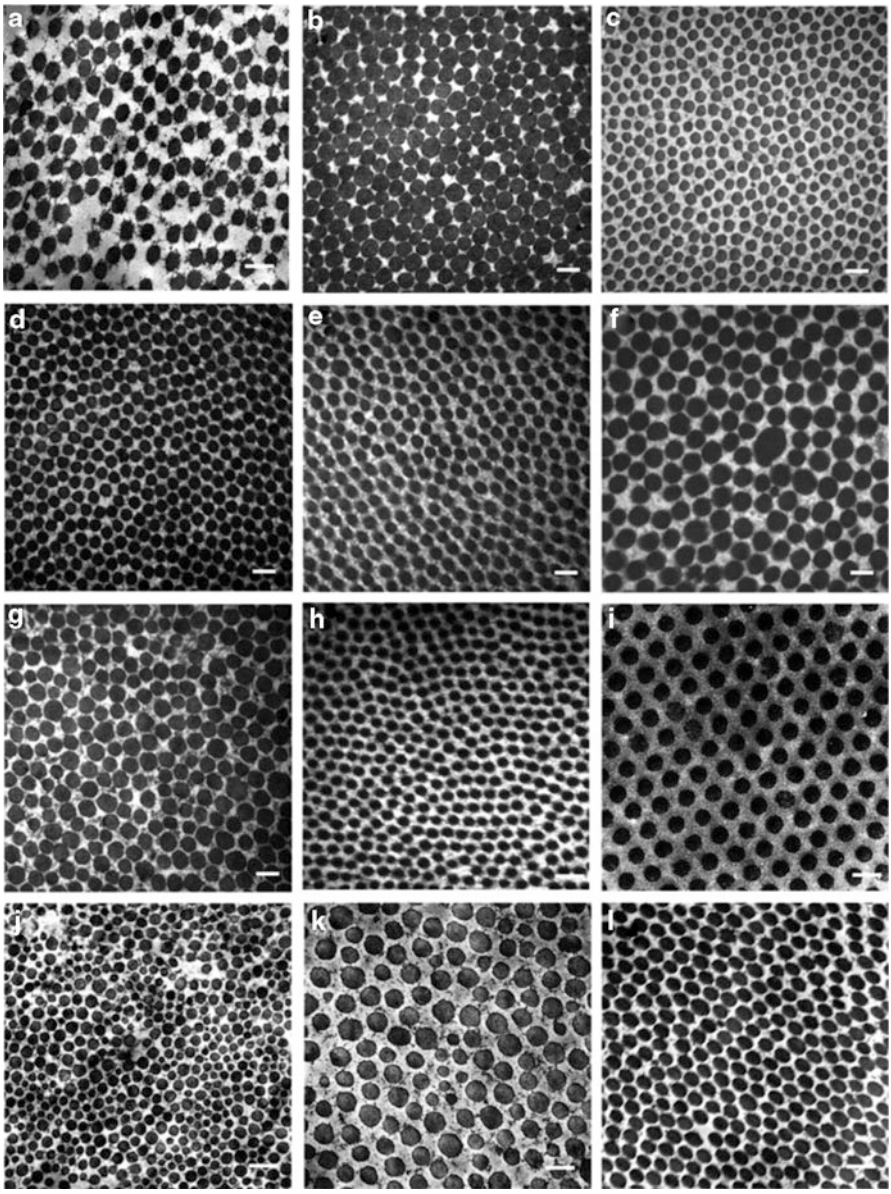


Fig. 3 Transmission electron micrographs of nano-structured arrays of dermal collagen from several species of birds of different colors. (**a**) *Oxyura jamaicensis*, *light blue*; (**b**) *Numida meleagris*, *dark blue*; (**c**) *Tragopan satyra*, *dark blue*; (**d**) *Tragopan caboti*, *dark blue*; (**e**) *Tragopan caboti*, *light blue*; (**f**) *Tragopan caboti*, *orange*; (**g**) *Syrigma sibilatrix*, *blue*; (**h**) *Ramphastos toco*, *dark blue*; (**i**) *Philepitta castanea*, *light blue*; (**j**) *Gymnopithys leucapsis*, *light blue*; (**k**) *Procnias nudicollis*, *green*; and (**l**) *Terpsiphone mutata*, *dark blue*. All scale bars represent 200 nm

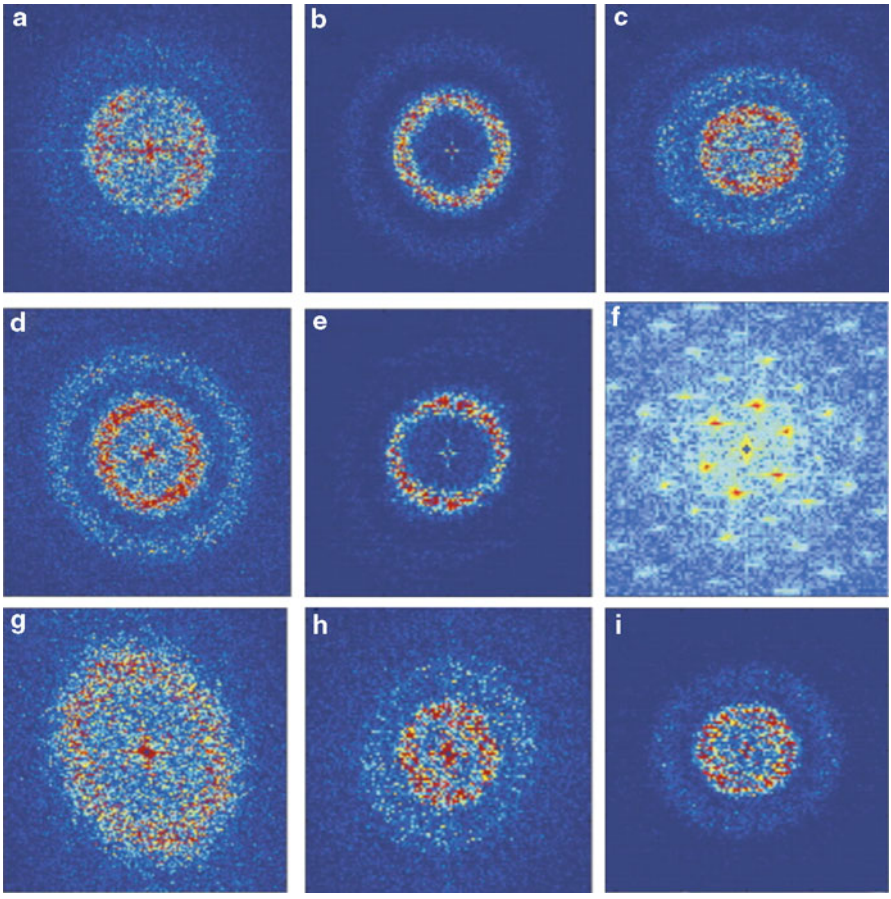


Fig. 4 Two-dimensional FFT spectra of transmission electron micrographs of nano-structured collagen arrays from the tissues of different birds. (**a**) *Dromaius novaehollandiae*, *blue*; (**b**) *Tragopan satyra*, *dark blue*; (**c**) *Pilherodius pileatus*, *light blue*; (**d**) *Coua reynaudii*, *dark blue*; (**e**) *Ramphastos toco*, *dark blue*; (**f**) *Philepitta castanea*, *light blue*; (**g**) *Gymnopithys leucapsis*, *light blue*; (**h**) *Procnias nudicollis*, *green*; and (**i**) *Dyaphorophyia concreta*, *yellow green*. The colors from *blue* to *red* indicate the magnitude of the squared Fourier components

For example, if we consider an infinite dimensional lattice of points in two dimensions generated by linear combinations with integer coefficients of two linearly independent vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, then the Fourier transform of the sum of the deltas at the points of the lattice is a sum of deltas at a dual lattice. This fact is just a restatement of Poisson summation formula. See, for example, [29] or [3]. The dual lattice is generated by the vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ satisfying $\boldsymbol{u}_k \cdot \boldsymbol{v}_j = \delta_{kj}$, where now δ_{kj} is the Kronecker delta, $\delta_{jk} = 0$ if $j \neq k$ and $\delta_{jj} = 1$. If $A = (\boldsymbol{v}_1, \boldsymbol{v}_2)$

is the matrix of the linear transformation that maps the standard square lattice onto the lattice generated by $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ then $(A^{-1})^* = (\boldsymbol{u}_1, \boldsymbol{u}_2)$, where * denotes the transposed of a matrix. See again [29] and [3] for details.

If we consider instead a finite portion of an infinite lattice and compute its Fourier transform, one observes distinct peaks at the points of the dual lattice rather than deltas, but an *echo*, as called in [29], or a *finite size effect* is observed as a variation in intensity. The location of the peaks is not affected much by this echo, but the height and width of the peaks are. In particular, the height is determined by the number of points in the finite region of the lattice analyzed and, hence, in our case, the physical length of the image of the tissue. In a perfect lattice as the number of points increases to infinity the Fourier transform converges locally around the peaks to delta distributions. However, when analyzing biological tissues, the finite effect should not be completely disregarded because the tissues do have specific finite dimension.

In the quasi-ordered tissue a precise mathematical description is harder to state. Except for the local order extended to the next neighboring fibers, the tissues have no order that can be analytically quantified in an obvious way. To quantify such order or (lack of it) we compute the Fourier transform numerically. Intuitively, the lack of order at larger distances makes the predominant frequencies to be mostly associated to the nearest-neighbored order, and the quasi-homogeneity of the tissues (the tissues look the same in any orientation) makes the peaks in the side of the Fourier transforms to be uniformly distributed in a ring at particular frequencies. The peak at the origin of the Fourier transform corresponds to the transmitted energy of the incident field that is not scattered. The first peak outside the origin occurs at a frequency determined in part by the average distance from the center of a fiber to the center of the nearest one and the size of the fibers and the overall arrangement, and represents the main physical periodicity in the tissue.

Using Benedek's theory we can try to use the Fourier transforms of the images of the tissues to give some prediction of the dimensions of the spatial variation in refractive index and hence the predominant wavelengths to be constructively scattered. The refractive indices of the collagen and the mucopolysaccharide are known (approximately 1.55 and 1.35, respectively). With these indexes of refraction and the density function as observed in the electronic microscope image, the average refractive index used can be estimated numerically from the micrographs (by looking at regions of black and white). Using the peaks observed in the FFT of the image of the tissue, one can predict the wavelength of the predominant color observed using the formula (12). For backward scattering, wavelengths of about twice the spatial periodicity measured by the FFT will be scattered the most.

It is not only the peaks in the Fourier transform what matters in the model but also the general distribution of energy. In order to exploit further information encoded in the Fourier transform of the images, we want to make some kind of comparison of the distribution of energy of our predictions with the actual spectrophotometer measurements. As mentioned in the introduction colors can be made up in different ways, but the whole spectral distribution helps falsify the Rayleigh hypothesis.

For the comparison, we first need to have a one-dimensional distribution of energy as the one given by the spectrophotometer. One can think of several ways to do this. One is, for example, to select an arbitrary radial direction. A similar approach to this was carried out by Vaezy et al. [33, 34]. However, the radial symmetry of the Fourier transform of the quasi-ordered tissues suggests that we consider instead a different analysis. For the comparison we want to further accentuate the radial symmetry, and hence we replace $|\widehat{f}|^2$ by its average on small concentric rings. Though artificially imposed, this radial (or azimuthal) average certainly reflects the ringlike structure observed in the images of the Fourier transform of the quasi-ordered tissue. To obtain a one-dimensional distribution of energy we use the radial average distribution to compute the total energy in each frequency band. We normalize the total energy or Fourier power (the L^2 norm of the Fourier transform) to be one over the visible part of the spectrum. Finally we plotted the amount of energy on a certain bandwidth as a percentage of the total energy and compare it with the spectrophotometer measurements (reflectance spectrum).

Figure 5 reproduced from [20] shows the comparisons between the actual reflectance spectrum measured with a spectrophotometer and our predictions using the radial average FFT. These were done for several tissues whose images appear in Fig. 3 and whose FFT are given in Fig. 4. See also [20, 23] for further technical details.

We observe that the resulting general profile of distribution of relative energy is similar to the one obtained by spectrophotometry. The quantitative discrepancy in the actual numerical computation is not surprising given the many elements involved in the collection and preparation of the specimens that could slightly modify their structure and the numerous approximations and analytical simplifications we have made, both physically and mathematically. The qualitative similitudes in the observed and predicted spectrum (both in terms of locations of peaks and general shape) are, on the other hand, quite noticeable and are a reasonable experimental corroboration of the validity of the physical model used. It is evident that only certain wavelengths are coherently scattered, and their range of values (and hence the colors observed) are determined by the physical periodicities in the tissues.

4 Lack of Iridescence and Other Works that Followed

Our experiments also clearly put in evidence that the color is not produced by Rayleigh scattering (which would produce for spectrum a ramp toward the UV). Our analysis based on the existing theory by Benedek was the first to provide an alternative explanation for the phenomena. Further, the general radial symmetry of the Fourier transform of the quasi-ordered tissues explains why these tissues are not highly iridescent: the spatial frequency of variation in refractive index remains similar in all directions within the quasi-ordered tissues and thus produces a uniform hue for backward scattering of light independent of the angle of incidence. (See

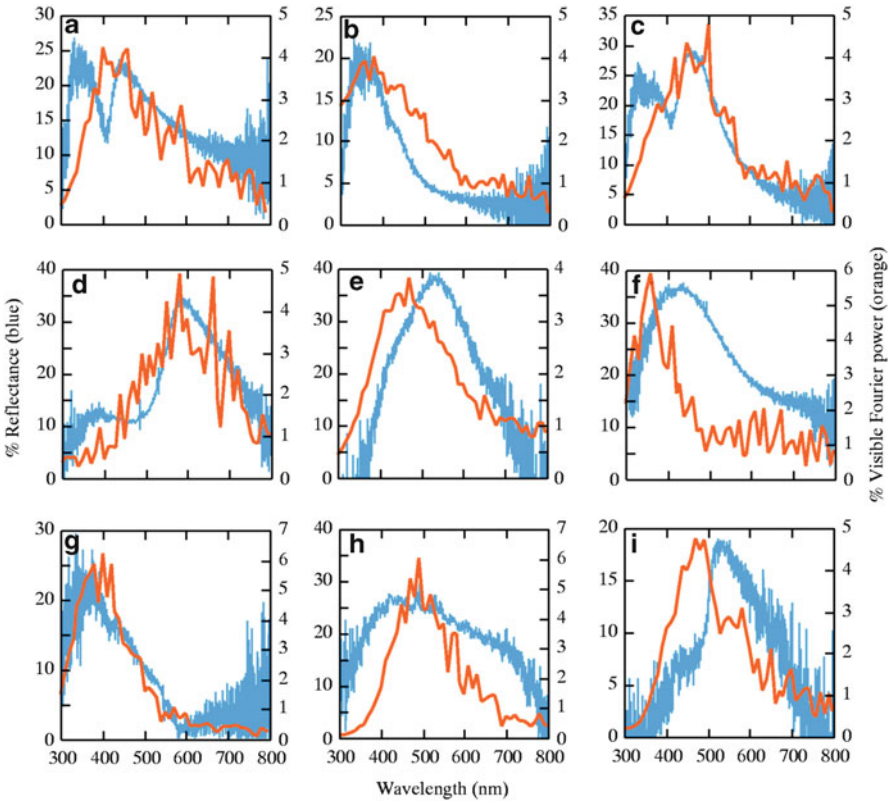


Fig. 5 Comparisons of reflectance spectra (*blue*) measured with a spectrophotometer and the Fourier-predicted spectra (*orange*) for samples of tissues from different birds. (**a**) *Lophophorus impejanus, dark blue*; (**b**) *Tragopan temminckii*, *dark blue*; (**c**) *Tragopan temminckii*, *light blue*; (**d**) *Tragopan caboti, orange*; (**e**) *Syrigma sibilatrix*, *light blue*; (**f**) *Coua caerulea*, *dark blue*; (**g**) *Ramphastos toco*, *dark blue*; (**h**) *Selenidera culik*, *green*; and (**i**) *Dyaphorophyia concreta*, *yellow-green*. The reflectance spectra are reported as a percentage reflectance (*blue*, *left axis*), and the predicted spectra are reported as a percentage of Fourier power (*orange*, *right axis*)

Noh et al. [17] for a rigorous demonstration of this fact.) Interestingly, it was this feature of coherent scattering from quasi-ordered materials that originally led to the confusion with Rayleigh scattering. Researchers in the field traditionally conceived of only two alternative sorts of order: complete crystalline periodicity or complete random distribution of particles. Within this framework, iridescent structural colors were associated with the interference from crystalline materials, and non-iridescent colors were associated with Rayleigh scattering from random distributions. The possibility of order only at the local scale and its optical consequences were not considered.

In very ordered materials to perform the radial average is, perhaps, not fully justified, but the relative intensity of the peaks is so large that we still get a good

match with the measured reflectance spectrum, see again [23]. In fact, the color of the extremely ordered tissues is more brilliant and *pure tone* than those in some of the quasi-ordered ones. A natural question to ask at this point is why then the very ordered tissues are non-iridescent. The answer lays again in the more complicated structure of the tissue at other larger scales and which is hard to incorporate in our first analysis. As mentioned before, the tissue is made of several layers of fibers in different directions and, as explained in [1], the total scattered field is some average field made from the contributions of all the layers. The lack of organization of the different layers has a similar effect to what is already observed in the quasi-ordered tissue. Though within an array of parallel fibers running in a particular direction we have an almost perfect hexagonal lattice, the same lattice in another set of parallel fibers may appear rotated by an arbitrary angle. The total intensity then will have a substantially uniform peak in its frequency content in a dense distribution of angles around the origin. An average effect is still what we observe or measure with the spectrophotometer. In other words, if all the cross section of parallel arrays of fibers would have the same orientation for the observed hexagonal lattice, the tissues would be iridescent. But, as we just explained, this is not the case. For comparison, we mention again hummingbirds, whose structurally color tissues have an almost perfect parallel laminar morphology resembling thin parallel films, and hence they do produce iridescent coloration according to Bragg's law.

Blue, green, and violets produced by coherent scattering have been documented by now by our methods. One could speculate that warmer colors are harder to produce by constructive interference since they would require a spatial order at a larger scale, which is perhaps too difficult to achieve in a biological tissue that needs to keeps such order as it grows. Otherwise, it may be that rarity of blue or green pigments prevents animals from making pigmentary blues or greens, but that the availability of long wavelength pigments favors those outcomes.

We have also analyzed feather barbs which look black to human eyes but possess vivid UV peaks (approximately 350 nm) that are not visible to humans but are easily perceived by birds. See [25]. In feathers the periodicities in the tissues take place in three dimensions and are provided by a distribution of air bubbles inside the tissues.

The methods have also been applied with similar results to the study of coloration in primates [21], butterflies [27], and dragonflies [26]. In addition, the Fourier method has been applied by Shawkey et al. [30] to a three-dimensional bird feather data set that was acquired by electron tomography. The empirical result provided some advance over 2D Fourier analysis of electron micrographs when dealing with 3D arrangements, but it still had inaccuracies due to systemic distortions in the 3D tomographic reconstruction.

The works mentioned above were some of the first approaches to the understanding of so many diverse structurally colored biological tissues. Since then many other works have appeared in the literature. The results using the relative simple model of Benedek have been by now corroborated with other more comprehensive techniques and experimentation. In particular, Prum and a multidisciplinary team at Yale University have recently employed in their studies small-angle X-ray scattering (SAXS) carried out at the Argonne National Labs. These studies hope to improve

upon the empirical limitations of Fourier transforms of electron micrographs by direct measurement of the Fourier transform of electron density variations in these nanostructures. See the works [4,5,12] for technical details and further explanations.

In our original analysis unidirectional light was assumed, and we were concerned only with backward scattering. A more delicate analysis and experimentation using omni-directional lighting in the quasi-ordered structures of birds feathers was recently carried out in [17] using SAXS. It was shown in [17] that in fact, under directional light, the scattering peak occurs in the backward direction. Moreover the authors in [17] also showed that under omni-directional lighting the colors observed remain unchanged with the angle of observation. See the cited reference for further information.

Likewise, our original analyses only concerned single scattering, i.e., interactions of photons that were each scattered only a single time by the scattering objects. But it became apparent that some inaccuracies in experimental comparisons of Fourier predicted and measured reflectance spectra were the result of multiple scattering: i.e., interactions among photons scattered two or more times by the nanostructures. This led to new physical theory and tests on double scattering by quasi-ordered nanostructures [16,18]. These works show that multiple scattering by quasi-ordered nanostructures produces new optical phenomena (e.g., double-peaked reflectance spectra) that were not anticipated in traditional optics. Although they require a new experimental method, the new X-ray scattering studies demonstrate the fundamental relevance and accuracy of the Fourier transform to the analysis of this optical phenomenon in nature.

The study of nature made structured tissues also relates to the study of photonic materials. A lot of activity in this area has taken place as groups of researchers try to fabricate photonic crystals and understand their properties. See [8] for references. In addition the tissues we studied have resemblances and similar physical properties to *hyperuniform* systems as studied, for example, by Torquato and Sillinger in [31]. These systems are theoretical arrangements of distribution of points that produce complete band gaps at low frequencies. The understanding of the fluctuation of density in materials and their scattering and transmitting properties will certainly continue to be an intense area of research in the immediate future. It is interesting to see how such materials are already present in biological tissues and are used in nature for a variety of purposes.

5 Summary

We wanted to illustrate here how Fourier analysis and numerical experiment with numerous tissues sustain the claim that quasi-ordered systems can produce non-iridescent structural colors by coherent scattering. Such color production occurs when only some wavelengths of visible light are selectively reinforced. The Fourier transform becomes an ideal analytical tool because it is a mathematical analog of the actual physical process of light interacting with the optically heterogeneous tissues.

The application presented renewed our appreciation of the ability of the Fourier transform to codify order or the lack of it, which makes Fourier analysis a very valuable tool for studies in material sciences.

Lastly, we find the use of Fourier analysis in biological questions addressing physical phenomena that affect communication and behavior in animals rather thought-provoking. We marvel at this beautiful manifestation of Fourier analysis in nature and the role it may play in sexual selection in many bird species. In fact, the animals' sexual preference for a specific color is not really based on the physical reason for the coloration, which is the collagen fiber order at invisible nano-scales. Instead, preference is based on the observable features of the reflectance spectrum resulting from such order. We can say that, essentially, preference is based on the Fourier transform of the invisible structures!

As it is well-known, Fourier introduced his groundbreaking analysis of the heat equation (by now called Fourier analysis) in his famous *Analytic theory of heat* [9]. We have mentioned in other occasions (e.g., [32]) a favorite quote from his work, which we want to repeat here one more time:

> "*...if the order which is established in this phenomena could be grasped by our senses, it would produce in as an impression comparable to the sensation of musical sound.*"

With this quote in mind, we would like to conclude by pointing out that the order in the nanostructures of the biological tissues studied can indeed be perceived by our senses as vivid colors, and these colors can certainly be as aesthetically pleasing to the observer as the sensation of musical sound referred to in Fourier's words.

Acknowledgements This note is based in part on the lecture *FFT blues: Fourier analysis and the structural colors of biological tissues*, which was presented by the second-named author at the 2007 FFT talks. He would like to thank the organizers for the opportunity to present the talk and for providing with the FFT series of conferences such a stimulating interdisciplinary environment for the interaction of mathematics with other fields of research. The authors would also like to thanks the organizers for their invitation to write this article.

The research reported here has been supported in part by the National Science Foundation under the grants DBI-0078376, DMS-0070514, DMS-0112375, and DMS-0400423. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Benedek, G.B.: Theory of transparency of the eye. Appl. Opt. **10**, 459–473 (1971)
2. Bragg, W.H., Bragg, W.L.: X-rays and Crystal Structure. G. Bell, London (1915)
3. Córdoba, A.: La formule sommatoire de Poisson. C. R. Acad. Sci. Paris Ser. I Math. **306**(8), 373–376 (1988)

4. Dalba, L., Saranathan, V., Clarke, J.A., Vinther, J.A., Prum, R.O., Shawkey, M.D.: Colour-producing β-keratin nanofibres in Blue Penguin (Eudyptula minor) feathers. Biol. Lett. **7**(4), 543–546 (2011)
5. Dufresne, E.R., Noh, H., Saranathan, V., Mochrie, S., Cao, H., Prum, R.O.: Self-assembly of biophotonic nanostructures by phase separation, Soft Matter **5**, 1792–1795 (2009). doi:10.1039/b902775k
6. Dyck, J.: Structure and colour-production of the blue barbs of Agapornis roseicollis and Cotinga maynana. Z. Zellforsch. **115**, 17–29 (1971)
7. Dyck, J.: Structural colours. In: Proceedings of 16th International Orthnithological Congress, pp. 426–437. Australian Academy of Science, Canberra (1976)
8. Forster, J.D. Noh, H., Liew, S.F., Saranathan, V., Schreck, C.F., Yang, L., Park, J.G., Prum, R.O., Mochrie, S.G.J., O'Hern, C.S., Cao, H., Dufresne, E.R.: Biomimetic isotropic nanostructures for structural coloration. Adv. Mater. **22**, 2939–2944 (2010)
9. Fourier, J.: Analytic Theory of Heat. 1822, Translation by A. Freeman in Great Books of the Western World, Encyclopedia Britannica (1990)
10. Fox, D.L.: Animal Biochromes and Structural Colors. University of California Press, Berkeley, California (1976)
11. Hill, G.E., McGraw, K.J. (eds.): Bird Coloration. Harvard University Press, Cambridge (2006)
12. Liew, S.F., Forster, J.D., Noh, H., Schreck, C.F., Saranathan, V., Lu, X., Yang, L., Prum, R.O., OHern, C.S.O., Dufresne, E.R., Cao, H.: Short-range order and near-field effects of optical scattering and structural coloration. Opt. Exp. **19**(9), 8208–8217 (2011). doi:10.1364/OE.19.008208
13. Macleod, H.: Thin-film Optical Filters. Adam Hilger, Bristol (1986)
14. Maurice, D.M.: The structure and transparency of the cornea. J. Physiol. London **136**, 268–286 (1957)
15. Nassau, K.: The Physics and Chemistry of Color. Wiley, New York (1983)
16. Noh, H., Liew, S.F., Saranathan, V., Prum, R.O., Dufresne, E.R., Mochrie, S.G.J., Cao, H.: Double scattering of light from biophotonic nanostructures with short-range order. Opt. Exp. **18**, 11942–11948 (2010). doi:10.1364/OE.18.011942
17. Noh, H., Liew, S.F., Saranathan, V., Prum, R.O., Mochrie, S.G.J., Dufresne, E.R., Cao, H.: How non-iridescent colors are generated by quasi-ordered structures of bird feathers. Adv. Mater. doi:10.1002/adma.200903693
18. Noh, H., Liew, S.F., Saranathan, V., Prum, R.O., Mochrie, S.G.J., Dufresne, E.R., Cao, H.: Contribution of double scattering to structural coloration in quasi-ordered nanostructures of bird feathers. Phys. Rev. E **81**, 051923 (2010) [8 pages]. doi:10.1103/PhysRevE.81.051923
19. Prum, R.O., Torres, R.H.: A Fourier tool for the analysis of coherent light scattering by bio-optical nanostructures. Integrative and Comparative Biology **43**, 591–610 (2003)
20. Prum, R.O., Torres, R.H.: Structural colouration of avian skin: convergent evolution of coherently scattering dermal collagen arrays. J. Exp. Biol. **206**, 2409–2429 (2003)
21. Prum, R.O., Torres, R.H.: Structural colouration of mammalian skin: convergent evolution of coherently scattering dermal collagen arrays. J. Exp. Biol. **207**, 2157–2172 (2004)
22. Prum, R.O., Torres, R.H., Willamson, S., Dyck, J.: Coherent light scattering by blue feather barbs. Nature **396**, 28–29 (1998)
23. Prum, R.O., Torres, R.H., Kovach, C., Willamson, S., Goodman, S.: Coherent light scattering by nanostructures collagen arrays in the caruncles of the Malagasy Asities (Eurylaimidae: Aves). J. Exp. Biol. **202**, 3507–3522 (1999)
24. Prum, R.O., Torres, R.H., Willamson, S., Dyck, J.: Two-dimensional Fourier analysis of the spongy medullary keratin of structurally coloured feather barbs. Proc. Royal Soc. London, Series B **266**, 13–22 (1999)
25. Prum, R.O., Andersson, S., Torres, R.H.: Coherent light scattering of ultraviolet light by avian feather barbs. Auk **120**, 163–170 (2003)
26. Prum, R.O., Cole, J.A., Torres, R.H.: Blue integumentary structural colours in dragonflies (Odonata) are not produced by incoherent Tyndall scattering. J. Exp. Biol. **207**, 3999–4009 (2004)

27. Prum, R.O., Quinn, T., Torres, R.H.: Anatomically Diverse Butterfly Scales Produce Structural Colors by Coherent Scattering. J. Exp. Biol. **209**, 748–765 (2006)
28. Raman, C.V.: The origin of the colours in the plumage of birds. Proc. Ind. Acad. Sci. (A) **1**, 1–7 (1934)
29. Senechal, M.: Quasicrystals and geometry. Cambridge University Press, Cambridge (1996)
30. Shawkey, M.D., Saranathan, V., Pálsdóttir, H., Crum, J., Ellisman, M., Auer, M., Prum, R.O.: Electron tomography, three-dimensional Fourier analysis and colour prediction of a three-dimensional amorphous biophotonic nanostructure. J. R. Soc. Interface **6**, S213-S220 (2009). doi:10.1098/rsif.2008.0374.focus
31. Torquato, S., Stillinger, F.H.: Local density fluctuations, hyperuniformity, and order metrics. Phys. Rev. E **68**, 041113 (2003)
32. Torres, R.H., Prum, R.O.: Análisis espectral de nanoestructuras en tejidos biológicos. Matematicalia **1**(2) (2005). http://www.matematicalia.net/
33. Vaezy, S., Clark, J.I.: Quantitative analysis of the microstructure of the human cornea and sclera using 2-D Fourier methods. J. Microsc. **175**, 93–99 (1993)
34. Vaezy, S., Smith, L.T., Milaninia, A., Clark, J.I.: Two-dimensional Fourier analysis of electron micrographs of human skin for quantification of the collagen fiber organization in the dermis. J. Electron Microsc. **44**, 358–364 (1995)
35. Van de Hulst, H.C.: Light Scattering by Small Particles. Dover, New York (1981)

A Harmonic Analysis View on Neuroscience Imaging

Paul Hernandez–Herrera, David Jiménez, Ioannis A. Kakadiaris, Andreas Koutsogiannis, Demetrio Labate, Fernanda Laezza, and Manos Papadakis

Abstract After highlighting some of the current trends in neuroscience imaging, this work studies the approximation errors due to varying directional aliasing, arising when 2D or 3D images are subjected to the action of orthogonal transformations. Such errors are common in 3D images of neurons acquired by confocal microscopes. We also present an algorithm for the construction of synthetic data (computational phantoms) for the validation of algorithms for the morphological reconstruction of neurons. Our approach delivers synthetic data that have a very high degree of fidelity with respect to their ground-truth specifications.

Keywords Synthetic tubular data • Synthetic dendrites • Directional aliasing • Approximation error • Dendritic arbor segmentation • Confocal microscopy

1 Overture

What is the substance of knowledge and memory? These fundamental questions have been at the center of philosophical debate for over three millennia, but only

P. Hernandez–Herrera (✉) • I.A. Kakadiaris
Computational Biomedicine Lab, Department of Computer Science,
University of Houston, Houston, TX 77204, USA
e-mail: jalip1985@gmail.com; ioannisk@uh.edu

D. Jiménez • D. Labate • M. Papadakis
Department of Mathematics, University of Houston, Houston, TX 77204-3008, USA
e-mail: djimenez@cbl.uh.edu; diabate@math.uh.edu; mpapadak@math.uh.edu

A. Koutsogiannis
Department of Mathematics, University of Athens, Greece, GR-15784 Zografou, Greece
e-mail: akoutsos@math.uoa.gr

F. Laezza
Department of Pharmacology and Toxicology, University of Texas Medical Branch,
Galveston, TX 77555-1031, USA
e-mail: felaezza@utmb.edu

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5_21,

during the last fifty years our understanding of these essential human cognitive functions is finally becoming concrete. The quest for answers takes us back to the philosopher Plato (424/423 BC–348/347 BC) who, in the dialogue "Theaetetus," written circa 360 BC when Athens' glory was in decline amidst the Peloponnesian war, attempts to define knowledge from a philosophical viewpoint. In the dialogue, Euclid (not the famous geometer from Alexandria) recounts a discussion between Socrates and Theaetetus aiming to discover the nature of knowledge. Around the middle of their conversation Socrates refers to knowledge as being a series of "engrams", impressions on the "wax of the soul":

Socrates: And the origin of truth and error is as follows: When the wax in the soul of any one is deep and abundant, and smooth and perfectly tempered, then the impressions which pass through the senses and sink into the heart of the soul, as Homer says in a parable, meaning to indicate the likeness of the soul to wax (κηρός); these, I say, being pure and clear, and having a sufficient depth of wax, are also lasting, and minds, such as these, easily learn and easily retain, and are not liable to confusion, but have true thoughts, for they have plenty of room, and having clear impressions of things, as we term them, quickly distribute them into their proper places on the block. And such men are called wise. Do you agree?

Theaetetus : Entirely.

Socrates: But when the heart of any one is shaggy a quality which the all-wise poet commends, or muddy and of impure wax, or very soft, or very hard, then there is a corresponding defect in the mind the soft are good at learning, but apt to forget; and the hard are the reverse; the shaggy and rugged and gritty, or those who have an admixture of earth or dung in their composition, have the impressions indistinct, as also the hard, for there is no depth in them; and the soft too are indistinct, for their impressions are easily confused and effaced. Yet greater is the indistinctness when they are all jostled together in a little soul, which has no room. These are the natures which have false opinion; for when they see or hear or think of anything, they are slow in assigning the right objects to the right impressions in their stupidity they confuse them, and are apt to see and hear and think amiss and such men are said to be deceived in their knowledge of objects, and ignorant.

Theaetetus: No man, Socrates, can say anything truer than that.[1]

With the "wax of the soul" theory, Greek philosophers anticipated the impressively modern concept of the human brain and its plastic neuronal network connections as the site of memory engrams formation and knowledge retention [21, 22]. Despite the impressive advances of modern science, however, our journey towards the comprehension of the physical nature of the "wax of the soul" and of the memory engrams is still at the "end of the beginning". We are optimistic that through interdisciplinary, collective scientific efforts, this mystery will be finally unlocked.

[1] Translated by Benjamin Jowett [27].

1.1 Outline

This article is organized as follows. In Sect. 2, we provide a brief historical overview of neuroscience and describe the challenges and opportunities opened up by the recent advances in microscopy. In particular, we discuss the significance of developing computational tools for the morphological reconstruction of neurons. Next, in Sect. 3, we give an overview of the algorithms currently available for the segmentation and morphological reconstruction of neurons, including a brief account of online reconstruction and functional imaging of neurons (ORION), a suite of algorithms and software developed by some of the authors of this paper which provides semiautomatic segmentation and morphological reconstruction of dendritic arbors in neurons. In Sect. 4 we examine the aliasing errors arising when images are subjected to the action of orthogonal transformations. Such errors are common in 3D images of neurons acquired by confocal microscopes. The action of those orthogonal transformations modifies the frequency content of images during the conversion of an image from analog to digital as the high-frequency content may be enhanced or attenuated solely due to the action of an orthogonal transformation. We also provide error estimates and sufficient conditions for the sampling kernels guaranteeing that these reconstruction error estimates are not affected by the action of a group of orthogonal transformations. Finally, in Sect. 5, we use the results of Sect. 4 to develop an algorithm for the construction of highly accurate phantoms of tubular 3D structures which are useful to model realistic phantoms of dendritic arbors of arbitrary topological complexity and can be used for the benchmarking of segmentation and tracing algorithms.

2 The Saga

2.1 Historical Background

The concept of the neuron as the primary structural unit of the central nervous system was introduced as early as the Nineteenth century by the ground-breaking studies of Camillo Golgi (1843–1926) and Ramón y Cajal (1852–1934). Utilizing an ingenious tissue staining technique developed by Golgi, Ramón y Cajal provided the earliest evidence of the neuron as the primary discrete unit of the central nervous system, and defined its microanatomy using light microscopy. By examining the structure of thousands of neurons in every region of the brain, Ramón y Cajal discovered the universal structure of neurons consisting of a cell body (also called the *soma*), dendrites, and an axon. With impressive accuracy, he also postulated that dendrites, which are multiple branching structures that arise from the cell body, and the axon, a single elongated cellular protrusion stemming from the cell body, retain different functions and mediate specialized connections between neurons. In following studies in 1933, Ramón y Cajal conjectured that neuronal *spines*, which

are the protuberances appearing along the dendrites (similar to rose thorns hence called *espinas*), are the points where these specialized connections through which the axon of one neuron contacts a neighboring neuron are established. Remarkably, he postulated that spines are a manifestation of the economy of nature: they increase the surface of a dendrite enabling stronger connections between neural cells via the dendrite–axon route [72, pp. 3,101]. These connections are now referred to as *synapses*[2] and constitute the fundamental structures that permit a neuron to transmit electrochemical signals to another neighboring cell (usually to another neuron). With these fundamental studies, Ramón y Cajal laid the foundations of modern neuroscience. While neuroanatomy studies were flourishing, in the 1930s, Curtis, Cole, Hodgkin, and Huxley, four neurophysiologists, were investigating the electrical properties of the axon of the Atlantic giant squid, a large and easily accessible tissue preparation, and provided the first recordings of the action potential, a form of regenerative electrical waveform that propagates down the axon [14, 25]. With the use of the voltage clamp technique, Hodgkin and Huxley discovered that the action potential arises from sequential changes in the cell membrane permeability to Na^+ and K^+ ions and developed the first mathematical model of the action potential propagation using nonlinear differential equations. For the first time in history, these experiments revealed the basis of electric function in neurons. Later studies in the 1950s by Fatt, Katz, and del Castillo [31] established that, by propagating down the axon, the action potential mediates synaptic transmission. Once it reaches the presynaptic *bouton* (the large ending of the axon), the action potential is decoded into a chemical signal through the release of discrete quanta of neurotransmitter molecules which eventually reach the postsynaptic side of the synapse (spine) and bind to specific membrane ion channels, called *receptors*. Upon binding to the neurotransmitter, receptors change conformation allowing specific ions to permeate into the postsynaptic cell and generate an electric charge called the excitatory postsynaptic potential (EPSP). If this electric charge exceeds a certain threshold, the EPSP elicits an action potential in the postsynaptic cell (the receiving cell). It is through this sequence of electrochemical chain reactions that the information is transmitted and stored in the brain through a connectome of neuronal networks. Although these basic concepts of neurophysiology are very well established, the explosive development of ultrasophisticated, unprecedented resolution imaging technologies in modern times have revealed new fascinating aspects of synaptic transmission and specifically has highlighted the critical role of synaptic spines as the main integrators of neuronal information.

[2]From the Greek prefix "$\sigma\upsilon\nu$-" and the root of the verb '$\acute{\alpha}\pi\tau o\mu\alpha\iota$', to touch; by adding the prefix the verb $\sigma\upsilon\nu\acute{\alpha}\pi\tau\omega$ means to clasp together but in ancient and in modern Greek, it means "to form an accord or to establish a formal relationship".

2.2 Modern Neuroscience

The emergence of fluorescence-based technologies and the development of sophisticated fluorescence microscopes, such as confocal and multiphoton, facilitate the acquisition of high-resolution fluorescent images of dendrites and spines both *in vitro* and *in vivo* and allow the monitoring of their dynamic structural changes in real time. It is now well documented that spines can grow or disappear in response to rapid and local changes in synaptic transmission (*spine plasticity*) or to more global and prolonged effects induced by network activity (*neuronal homeostasis*). These dynamic morphological changes of spines, associated with plasticity and homeostasis, are considered to be the structure-function link in the heart of learning and memory formation and are associated with different behavioral states or chronic neuropathologies [21, 22, 58, 59, 72–74]. It is through structure-function changes of synaptic spines throughout neuron networks that we retain what we have learned, we respond to external stimuli, and eventually we adapt to the surrounding environment. With no doubts, the ability to accurately capture the morphological information of dendrites and spines and track their dynamic changes will rapidly translate into a better understanding of brain function. Towards futuristic applications this improved knowledge of cellular and subcellular neuronal morphologies could be included in electrophysiological computer simulations, so that quantitative and qualitative effects of dendritic and spine structure under stimulation can be extensively characterized. With the current computational capabilities, these models can be implemented into supercomputers to allow the generation of *virtual neurons* which retain all anatomical and functional characteristics of their real counterparts. When modeled neurons are organized into complex structures under appropriate rules governing their anatomical and functional connectivity, in principle, entire portions of the nervous system would be simulated into realistic neural networks, leading to what G. Ascoli phrased as: "A detailed computer model of a *virtual brain* that was truly equivalent to the biological structure" and "could in principle allow scientists to carry out experiments that could not be performed on real nervous systems because of physical constraints" [2].

While the technological advances in fluorescence-based microscopy have opened up exciting avenues of investigation in neuroscience and have set high-standard goals to modern neurophysiology, this area of research has also raised a number of computational, algorithmic, and mathematical challenges involving the acquisition and modeling of high-resolution data acquired through confocal microscopy, the preprocessing of the data (which are typically affected by blurring and Poisson noise), and the morphological reconstruction of dendritic structures and spines. Capturing and accurately modeling the morphological transitions of spines and dendrites in response to various functional states of a neuron will bring us a step closer to identify the physical nature of the "wax of the soul" anticipated by Plato and to unravel how memory traces or engrams are formed, retained, translated into human cognitive functions [21, 22].

3 Imaging Neurons

3.1 The State of the Art

As indicated by the observations in the section above, the ability to produce accurate morphological reconstruction of dendritic arbors and spines in neurons is of fundamental importance to the goal of generating a virtual neuron. During the last ten years, a flurry of activity was aimed at the development of automatic or semiautomatic computational tools for delivering morphological reconstructions of neurons, and there are currently several academic and a few commercial and freeware imaging suites available. All of these algorithms depict branching and terminal points, diameters of dendritic branches and the soma and output the results in 3D visualizations. Their performance varies and depends on the level of training per data set, the noise that affects the data, and on the level of manual intervention required. These reconstructions rely on a tracing of the dendritic arbor which has been a hot research topic, the least, in the last ten years, e.g., [1, 19, 20, 24, 29, 38–41, 43, 44, 48, 49, 51, 55, 58, 59, 70, 71, 73]. More recently, significant work on the tracing and morphological reconstruction of dendritic arbors emerged as a result of the DIADEM competition [7, 40]. Although, all of them are designed to capture the 3D structure of the dendritic arbor with sufficient accuracy, they usually miss the spatially localized detail of the surface of the dendritic branches, and in particular they ignore spines [56] as the common goal of all of the dendrite-tracing methods is to detect the centerline of dendritic branches. This naturally and reasonably becomes a new system of coordinates for navigating the dendrite. Although several of the dendrite-tracing algorithms estimate the dendritic diameter locally, their estimations cannot capture the localized details of the dendritic surface with the exception of [29,30,35,36,52–55] which generates a probabilistic segmentation of the volume of the dendritic arbor; thus the likelihood of the association of a voxel to the dendritic surface is obtained.

The existing spine detection capabilities of current 3D algorithms build upon the type of centerline tracing we previously described. Using the detected centerline for navigation within the dendrite, they typically apply the Rayburst detection algorithm [48, 61, 70]. Several other methods rely on detecting spines on 2D maximum intensity projections, but those methods frequently miss significant spines and the many of the weaker ones as they are obscured by the projection of the higher intensity parts of the dendritic volume onto them [3, 13]. In particular, Fan et al. use maximum intensity projections for in vivo spine detection and analysis [18]. There is very limited work on in vivo spine detection primarily because of the necessity to use tracking algorithms when 2D image analysis methods are employed. There are also pseudo-3D approaches in the sense that spines are detected on each scanning plane and then the results of the detection are fused to create a 3D image stack [3, 11, 12, 18, 45, 75, 76]. Classification of spines according to their types, estimation of volume, and of head diameter is mainly being done with the Rayburst algorithm

[17]—often applied in 2D only [34, 51]—which counts voxels whose intensity exceeds an operator-chosen threshold in certain directions. However, there are two main problems in all of these approaches:

1. The use of intensity thresholds applied on the original image in order to detect the surface of spines. Since Poisson noise corrupts images, intensity thresholds are increasingly unreliable as the concentration of the fluorophores decreases. This is often the case in undeveloped or thin spines, but more importantly, it affects spine necks resulting into detached spines which are harder to distinguish from leaking fluorophores or plain noise spikes.
2. Constraints of technical nature as well as the need to decrease image acquisition duration lead to the use of anisotropic voxels, typically of aspect ratio 1:1:3–1:1:4. Although images are corrected to account for blurring introduced by the microscope, dendritic volumes are reconstructed with voxels of these aspect ratios. This implies that objects such as spines which are oriented in an arbitrary way in 3D and have a diameter of 10–13 voxels with their neck being less than 2–3 voxels thick at the highest resolution will not be properly classified according to their shape as their shape is distorted by this anisotropic sampling grid. A partial heuristic remedy utilizing the Rayburst algorithm is proposed in [17] to mitigate this problem, but it can only have limited success since the data are severely undersampled at off the xy-plane orientations.

3.2 Online Reconstruction and Functional Imaging of Neurons

ORION [29, 30, 35, 36, 52–55, 67] is a suite of algorithms and integrated software that can be used for the morphological reconstruction of dendritic arbors from 3D images obtained by multiphoton or confocal microscopes. ORION can identify dendritic centerlines their branching and terminal points and estimates the diameter of branches at every centerline point; however, it does not identify or classify spines, and 3D visualizations of the morphological reconstructions of dendritic arbors do not include spines [10]. ORION segmentation of the dendritic arbor is based on extracting the eigenvalues of the Hessian of an ensemble of low-pass Gaussian-filtered outputs of the original 3D volume and by learning how these eigenvalues depend on a tubular model estimated from the data. The segmented volume results from a probability 3D map conditioned on the learned model. Dendritic centerlines and branching points represent the unique solution of a certain optimization problem. ORION has been successfully tested on synthetic data, on real data where the system outperformed experts, and on several DIADEM competition image sets. Notice however that, since ORION is designed to work primarily on dendritic arbors that are acyclic connected graphs, it cannot be applied to some of the DIADEM data sets. Figure 1 illustrates an application of ORION.

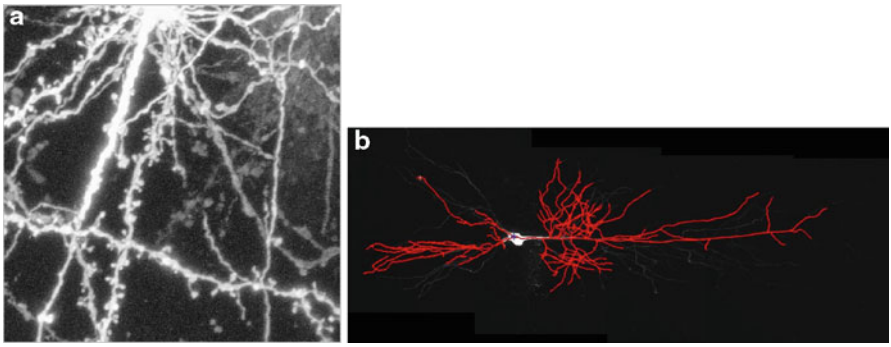


Fig. 1 (**a**) MIP of the *logarithm* of a raw image of CA1 hippocampal pyramidal neuron labeled with DiI fluorescent dye demonstrating the nature of the noise in voxels next to the dendritic branches. (**b**) Morphological reconstruction of a pyramidal neuron with ORION. The segmented dendrite's voxels are color-coded *red*. Note the absence of spines

4 Approximations Under the Action of a Group of Orthogonal Transformations

Since all images have compact support in the space domain $\mathbb{R}^d$ ($d = 2$ or $d = 3$), their conversion from analog to digital form occurring during acquisition requires truncating the image in the frequency domain. Typically, this process is modeled by convolving the given image, say $f \in L^2(\mathbb{R}^d)$, with a kernel function ϕ_{a}, referred to as the *analysis kernel*. Hence, if $0 < \varepsilon < 1$ is a preselected constant representing the level of the desired relative error, there is a compact subset of the frequency domain, say Ω, such that $\int_{\Omega^c} |\widehat{f}(\xi)|^2 \mathrm{d}\xi < \varepsilon \|\widehat{f}\|_2^2$. The set Ω is called the *essential bandwidth* of f. With no loss of generality we assume $\Omega \subseteq \mathbb{T}^d = [-1/2, 1/2]^d$. In particular, we define

$$\mathscr{B}_{\Omega}^{\varepsilon} := \left\{ f \in L^2(\mathbb{R}^d) : \hat{f} \in W^{1,2}(\mathbb{R}^d) \text{ and } \int_{\Omega^c} |\widehat{f}(\xi)| \mathrm{d}\xi < \varepsilon \|\widehat{f}\|_1 \right\}, \quad (1)$$

where $W^{1,2}(\mathbb{R}^d)$ is the Sobolev space containing all functions h whose distributional partial derivatives up to second order are contained in $L^1(\mathbb{R}^d)$. We can view $\mathscr{B}_{\Omega}^{\varepsilon}$ as a family of functions that are *almost bandlimited*, as for a function bandlimited in Ω one would have $\int_{\Omega^c} |\widehat{f}(\xi)| \mathrm{d}\xi = 0$. This observation motivates us to generalize classical sampling theory approaches in the spirit of [4]. Specifically, we adopt an oversampling approach implementing the digitization of the input image and its reconstruction from its samples. To this end we use two kernels, the analysis kernel ϕ_{a} and the synthesis kernel ϕ_{s}. Two compact sets are associated with this pair of kernels, B_{a} and B_{s}. We assume $\Omega \subset B_{\mathrm{a}}^{\circ}$, $B_{\mathrm{a}} \subset B_{\mathrm{s}}^{\circ}$, and $B_{\mathrm{s}} \subset (\mathbb{T}^d)^{\circ}$, where the superscript $\circ$ indicates the interior of a set, and:

1. $\widehat{\phi}_{\mathrm{a}}, \widehat{\phi}_{\mathrm{s}} \subset C^{\infty}(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$.
2. All partials derivatives of $\widehat{\phi}_{\mathrm{a}}$ and $\widehat{\phi}_{\mathrm{s}}$ up to second order are bounded.
3. $|\widehat{\phi}_{\mathrm{a}}(\xi)| \leq 1 + \varepsilon$, $|\widehat{\phi}_{\mathrm{s}}(\xi)| \leq 1 + \varepsilon$ for all $\xi \in \mathbb{R}^d$.
4. $|\widehat{\phi}_{\mathrm{a}}(\xi) - 1| \leq \varepsilon$ if $\xi \in B_{\mathrm{a}}$, $|\widehat{\phi}_{\mathrm{a}}(\xi)| \leq \varepsilon$ if $\xi \notin B_{\mathrm{s}}$.
5. $|\widehat{\phi}_{\mathrm{s}}(\xi) - 1| \leq \varepsilon$ if $\xi \in B_{\mathrm{s}}$, $|\widehat{\phi}_{\mathrm{s}}(\xi)| \leq \varepsilon$ if $\xi \notin \mathbb{T}^d$.
6. $\int_{B_{\mathrm{a}}^c} |\widehat{\phi}_{\mathrm{a}}(\xi)| \mathrm{d}\xi < \varepsilon$ and $\int_{(\mathbb{T}^d)^c} |\widehat{\phi}_{\mathrm{s}}(\xi)| \mathrm{d}\xi < \varepsilon$.
7. There exists $C > 0$ such that $\sum_{k \in \mathbb{Z}^d} |\hat{\phi}_{\mathrm{a}}(\xi + k)|^2 \leq C$ and $\sum_{k \in \mathbb{Z}^d} |\hat{\phi}_{\mathrm{s}}(\xi + k)|^2 \leq C$ for a.e. $\xi \in \mathbb{R}^d$.

With these conditions in mind we define

$$\tilde{f} := \sum_{n \in \mathbb{Z}^d} \langle f, T_n \phi_{\mathrm{a}} \rangle T_n \phi_{\mathrm{s}}, \tag{2}$$

where the right-hand side of (2) converges with respect to the L^2-norm due to Property (7) . Notice that Property (1) guarantees that both kernels have good spatial localization. Properties (3) and (4) indicate that B_{a} and B_{s} are the pass-bands of ϕ_{a} and ϕ_{s}, respectively, while their stop bands are both contained in $\mathbb{T}^d$. The digitization process gives the sequence $\{\langle f, T_n \phi_{\mathrm{a}} \rangle : n \in \mathbb{Z}^d\}$, while the inversion of this process, the reconstruction of original analog image from its samples $\langle f, T_n \phi_{\mathrm{a}} \rangle$ is referred to as the digital to analog conversion. Typically, in imaging, we only use the first part the analog to digital conversion, while the reverse process has only theoretical value. In general, $f \neq \tilde{f}$. It is one of our goals to estimate $\|f - \tilde{f}\|$, called the *reconstruction error*, with respect to different meaningful norms. In the applications presented in this chapter, the L^{∞} norm is the proper norm because it guarantees the uniform fidelity of the reconstruction of f throughout the spatial domain. In practice though, it is impossible to keep an infinite number of the samples $\{\langle f, T_n \phi_{\mathrm{a}} \rangle\}_{n \in \mathbb{Z}^d}$. Therefore, it becomes necessary to make a choice of a finite set $\Lambda \subset \mathbb{Z}^d$ such that the only values kept belong to $\{\langle f, T_n \phi_{\mathrm{a}} \rangle\}_{\Lambda}$. Hence,

$$f_{\Lambda} = \sum_{n \in \Lambda} \langle f, T_n \phi_{\mathrm{a}} \rangle T_n \phi_{\mathrm{s}} \tag{3}$$

gives an approximation of the original input signal or image f. Specializing to images, their finite extend and the limitations of the acquisition devices prescribe a certain size of voxels/pixels. This mathematically amounts to prescribing a certain essential bandwidth which has the form of a parallelepiped in $\mathbb{R}^d$ $(d = 2, 3)$ and $\Lambda = \prod_{s=1}^{d} [-N_s, N_s]$, where all N_s are integers. So it is important to study the overall *approximation error* $\|f - f_{\Lambda}\|$, with respect to various norms. In this chapter we are interested in the approximation error with respect to the L^{∞} norm, and in particular, we propose how to control this error when f varies, due to the action of a group of orthogonal transformations defined on $\mathbb{R}^d$. *Specifically, given a group G of orthogonal transformations acting on $\mathbb{R}^d$, e.g., $G = SO(d)$, we want to be able to find suitable kernels ϕ_{a} and ϕ_{s} so that if for a choice of $\Lambda \subseteq \mathbb{Z}^d$ (e.g., $\Lambda = \prod_{s=1}^{d} [-N_s, N_s]$) the error $\|f - f_{\Lambda}\|_{\infty} < \epsilon$, that is, $\|f - f_{\Lambda}\|_{\infty}$ is small enough, then*

$$\sup_{M\in G}\|\rho(M)f-(\rho(M)f)_{\Lambda}\|_{\infty}\leq\epsilon\ ,$$

where $\rho(M)f(x)=f(Mx)$, $x\in\mathbb{R}^d$.

Since,

$$\|\rho(M)f-(\rho(M)f)_{\Lambda}\|_{\infty}\leq\|\rho(M)f-\widetilde{\rho(M)f}\|_{\infty}+\|\widetilde{\rho(M)f}-(\rho(M)f)_{\Lambda}\|_{\infty} \tag{4}$$

for all $M\in G$ and $f\in L^2(\mathbb{R}^d)\cap L^{\infty}(\mathbb{R}^d)$, it becomes apparent that we need to control the growth of each one of the terms in the right-hand side of the previous inequality as M varies. The first of the two terms is known as *reconstruction error* while the other is called *truncation error*. Throughout the rest of the section we assume $f\in L^2(\mathbb{R}^d)$ and $\hat{f}\in L^1(\mathbb{R}^d)$.

There is an abundance of work on the study of decay estimates of these two types of errors, e.g., [4, 5, 16, 26, 28, 41, 44, 60] and of the approximation error as well in one and multidimensions both in the context of linear (when $\Lambda=\prod_{s=1}^{d}[-N_s,N_s]$), nonlinear and n-term approximation, e.g., [33, 62–66]. An excellent tutorial on nonlinear approximations [15] provides several more references that we did not include in this chapter. The novel concept we introduce in this section is that the proper selection of approximants should take into account the variations of an image due to the action of groups of orthogonal transformations on it, e.g., rotations. This kind of variation affects the rate of convergence of linear approximations as we demonstrate with an example at the end of the section. As Table 1 indicates that nonlinear and n-term approximations may be affected as well as the high-pass content of the image increases due to its rotation and the non-isotropy of the analysis kernel. In particular, if we keep Λ fixed, the error $\|\rho(M)f-(\rho(M)f)_{\Lambda}\|$ in any relevant norm may vary with M. Nevertheless, the error estimate provided by Theorem 1 provides a search range for Λ that does not depend on the individual transformations M, but it rather depends on the group to which M belong to.

Before continuing with the analysis of both errors, we give an example demonstrating the practical significance of this problem in images acquired by confocal microscopy, when ϕ_{a} is anisotropic.

The most common practice among neuroscientists is to acquire their data by using an anisotropic sampling grid of the form $\mathbb{Z}^2\times(N\mathbb{Z})$, where $N=3,4$ [17], which amounts to using anisotropic analysis and synthesis kernels. The use of this grid saves time and overcomes limitations due to the quantum nature of light but reduces the resolution to the point that spine volumes cannot be accurately estimated [17]. Scanning time increases nonlinearly [17] as the resolution in the z-direction increases and at xy is kept high.

Heuristic methods, popular among neuroscientists, have been proposed to resolve this issue [17], but those methods ignore the real mathematical problem, the undersampling in the z-direction. Figure 2 shows exactly how the volumes of spines are consistently ignored in the binary segmentation of a hippocampal CA1 neuron.

Let us now return to our analysis. Take $0<\varepsilon<1$, $\Omega\subseteq\mathbb{T}^d$ and $f\in\mathscr{B}_{\Omega}^{\varepsilon}$. We also assume that ϕ_{a} and ϕ_{s} are analysis and synthesis kernels satisfying Properties (1) through (7), $\Omega\subset B_{\mathrm{a}}^{\circ}$, $B_{\mathrm{a}}\subset B_{\mathrm{s}}^{\circ}$, and $B_{\mathrm{s}}\subset(\mathbb{T}^d)^{\circ}$. By taking the Fourier

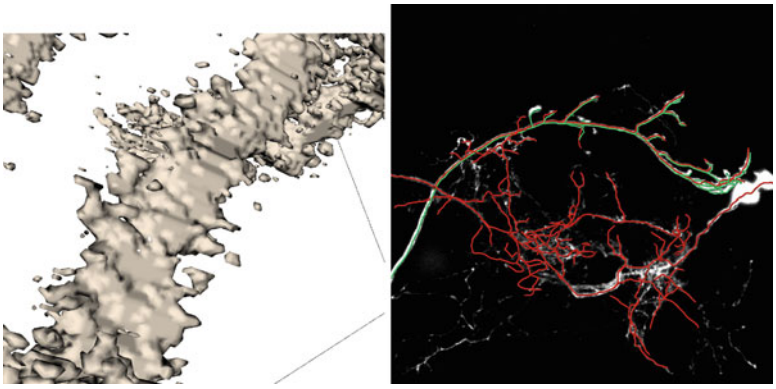


Fig. 2 *Left*: Part of the binary segmentation of the hippocampus CA1 neuron shown in Fig. 1. Notice the smoothness on the dendrite's side while its "other side" is rougher. The smoothness on the smooth side is due to the undersampling in the z-direction. This data set was sampled on the grid $\mathbb{Z}^2 \times (4\mathbb{Z})$. *Right*: MIP results of the tracing of an olfactory cell (OP2) from the Diadem competition data sets. The centerline annotated by an expert is marked with *green*. Centerline tracing results with ORION are marked with *red*. The raw image is in the background. There are two cells in this image stack although only a single cell should have been included in the image stack. ORION traces both of them, but by default it considers them as a single cell

transform on both sides of Eq. (2) we obtain

$$\widehat{\tilde{f}}(\xi) = \left(\sum_{n \in \mathbb{Z}^d} \langle f, T_n \phi_{\mathrm{a}} \rangle \mathrm{e}^{-2\mathrm{i}\pi n \cdot \xi} \right) \widehat{\phi}_{\mathrm{s}}(\xi) := A(\xi) \widehat{\phi}_{\mathrm{s}}(\xi) \,, \tag{5}$$

where A is a $\mathbb{Z}^d$-periodic function verifying

$$A(\xi) = \left(\sum_{n \in \mathbb{Z}^d} \langle f, T_n \phi_{\mathrm{a}} \rangle \mathrm{e}^{-2\mathrm{i}\pi n \cdot \xi} \right) = \sum_{\ell \in \mathbb{Z}^d} \widehat{f}(\xi + \ell) \overline{\widehat{\phi}_{\mathrm{a}}(\xi + \ell)}. \tag{6}$$

The next observation is critical for estimating the error bounds.

4.1 Bounds for the Coefficients of A

Several of the estimates that will be given below critically depend on the coefficients $\langle f, T_n \phi_{\mathrm{a}} \rangle$. By Parseval's theorem, we have

$$\langle f, T_n \phi_{\mathrm{a}} \rangle = \langle \widehat{f}, \mathrm{e}^{-2\pi \mathrm{i} n \cdot (\cdot)} \widehat{\phi}_{\mathrm{a}} \rangle = \int_{\mathbb{R}^d} \widehat{f}(\xi) \overline{\widehat{\phi}_{\mathrm{a}}(\xi)} \mathrm{e}^{2\pi \mathrm{i} n \cdot \xi} \mathrm{d}\xi = \left(\widehat{f} \overline{\widehat{\phi}_{\mathrm{a}}} \right)^{\vee} (n).$$

In general, for a function g such that $\widehat{g} \in C^{2m}(\mathbb{R}^d)$, where $m = 1, 2, \ldots$, and all of its derivatives are integrable, we have

$$(\Delta^m \widehat{g})^\vee(x) = (2\pi \mathrm{i})^{2m} \left(x_1^2 + \cdots + x_d^2\right)^m g(x) = (2\pi \mathrm{i})^{2m} \|x\|_2^{2m} g(x).$$

Since, Properties (1) and (2) and the definition of $\mathscr{B}_\Omega^\varepsilon$ imply that the distributional Laplacian $\Delta\left(\widehat{f}\overline{\widehat{\phi}_\mathrm{a}}\right)$ is integrable, we assert

$$|\langle f, T_n \phi_\mathrm{a}\rangle| = \left|\left(\widehat{f}\overline{\widehat{\phi}_\mathrm{a}}\right)^\vee(n)\right| \leq \frac{\left\|\Delta\left(\widehat{f}\overline{\widehat{\phi}_\mathrm{a}}\right)\right\|_1}{(2\pi)^2 \|n\|_2^2}.$$

Therefore, $\sum_{n\in\mathbb{Z}^d} |\langle f, T_n\phi_\mathrm{a}\rangle| < \infty$ and A belongs to $A(\mathbb{T}^d)$.

So, if in addition to the previous assumptions for f and for the analysis kernel ϕ_a, we have $\hat{f} \in W^{1,2m}$ and all partial derivatives of ϕ_a up to order $2m$, where $m = 1, 2, \ldots$, are bounded, then

$$(2\pi)^{2m}\|x\|_2^{2m}\left|\left(\widehat{f}\overline{\widehat{\phi}_\mathrm{a}}\right)^\vee(x)\right| \leq \left\|\left(\Delta^m\left(\widehat{f}\overline{\widehat{\phi}_\mathrm{a}}\right)\right)^\vee\right\|_\infty \leq \left\|\Delta^m\left(\widehat{f}\overline{\widehat{\phi}_\mathrm{a}}\right)\right\|_1$$

therefore, if $x \in \mathbb{R}^d$

$$|\langle f, T_x\phi_\mathrm{a}\rangle| = \left|\left(\widehat{f}\overline{\widehat{\phi}_\mathrm{a}}\right)^\vee(x)\right| \leq \frac{\|\Delta^m(\widehat{f}\overline{\widehat{\phi}_\mathrm{a}})\|_1}{(2\pi)^{2m}\|x\|_2^{2m}} \tag{7}$$

4.2 Estimation of the Approximation Error $\|f - f_A\|_\infty$

First, we proceed with the estimation of $\|\widehat{f} - \widehat{\tilde{f}}\|_1$. Note that

$$\begin{aligned}\|\widehat{f} - \widehat{\tilde{f}}\|_1 &= \int_{\mathbb{R}^d} |\widehat{f}(\xi) - A(\xi)\widehat{\phi}_\mathrm{s}(\xi)|\mathrm{d}\xi \\ &= \int_{\mathbb{T}^d} |\widehat{f}(\xi) - A(\xi)\widehat{\phi}_\mathrm{s}(\xi)|\mathrm{d}\xi + \int_{(\mathbb{T}^d)^c} |\widehat{f}(\xi) - A(\xi)\widehat{\phi}_\mathrm{s}(\xi)|\mathrm{d}\xi.\end{aligned}$$

Using Property (6) and the fact $f \in \mathscr{B}_\Omega^\varepsilon$, the second term in the previous sum can be bounded by

$$\int_{(\mathbb{T}^d)^c} |\widehat{f}(\xi) - A(\xi)\widehat{\phi}_\mathrm{s}(\xi)|\mathrm{d}\xi \leq \int_{(\mathbb{T}^d)^c} |\widehat{f}(\xi)|\mathrm{d}\xi + \sup_{\mathbb{R}^d}|A(\xi)| \int_{(\mathbb{T}^d)^c} |\widehat{\phi}_\mathrm{s}(\xi)|\mathrm{d}\xi$$

$$\leq \varepsilon \|\widehat{f}\,\|_1 + \left(\sum_{n\in\mathbb{Z}^d} |\langle f, T_n\phi_{\mathrm{a}}\rangle| \right) \varepsilon.$$

Next we estimate the contribution of the term $\int_{\mathbb{T}^d} |\widehat{f}(\xi) - \widehat{\tilde{f}}(\xi)| \mathrm{d}\xi$ to the reconstruction error of f:

$$\begin{aligned}
\int_{\mathbb{T}^d} |\widehat{f}(\xi) - A(\xi)\widehat{\phi}_{\mathrm{s}}(\xi)| \mathrm{d}\xi &= \int_{\mathbb{T}^d} \left| \widehat{f}(\xi) - \left(\sum_{\ell\in\mathbb{Z}^d} \widehat{f}(\xi+\ell)\overline{\widehat{\phi}_{\mathrm{a}}(\xi+\ell)} \right) \widehat{\phi}_{\mathrm{s}}(\xi) \right| \mathrm{d}\xi \\
&\leq \int_{\mathbb{T}^d} \left| (1 - \overline{\widehat{\phi}_{\mathrm{a}}(\xi)}\widehat{\phi}_{\mathrm{s}}(\xi))\widehat{f}(\xi) \right| \mathrm{d}\xi \\
&\quad + \int_{\mathbb{T}^d} \left| \sum_{\ell\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \widehat{f}(\xi+\ell)\overline{\widehat{\phi}_{\mathrm{a}}(\xi+\ell)}\widehat{\phi}_{\mathrm{s}}(\xi) \right| \mathrm{d}\xi.
\end{aligned}$$

Now,

$$\begin{aligned}
&\int_{\mathbb{T}^d} \left| \sum_{\ell\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \widehat{f}(\xi+\ell)\overline{\widehat{\phi}_{\mathrm{a}}(\xi+\ell)}\widehat{\phi}_{\mathrm{s}}(\xi) \right| \mathrm{d}\xi \leq (1+\varepsilon)\varepsilon \sum_{\ell\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \int_{\mathbb{T}^d} \left| \widehat{f}(\xi+\ell) \right| \mathrm{d}\xi \\
&\quad \leq (1+\varepsilon)\varepsilon^2 \|\widehat{f}\,\|_1 \leq 2\varepsilon \|\widehat{f}\,\|_1 .
\end{aligned}$$

Using Property (3) we infer $|1 - \overline{\widehat{\phi}_{\mathrm{a}}(\xi)}\widehat{\phi}_{\mathrm{s}}(\xi)| \leq 1 + |\overline{\widehat{\phi}_{\mathrm{a}}(\xi)}\widehat{\phi}_{\mathrm{s}}(\xi)| \leq 1 + (1+\varepsilon)^2 \leq 5$. On the other hand, if $\xi \in \Omega$, Properties (4) and (5) imply $|1 - \overline{\widehat{\phi}_{\mathrm{a}}(\xi)}\widehat{\phi}_{\mathrm{s}}(\xi)| < 2\varepsilon + \varepsilon^2 < 3\varepsilon$. Therefore

$$\begin{aligned}
\int_{\mathbb{T}^d} \left| (1 - \overline{\widehat{\phi}_{\mathrm{a}}(\xi)}\widehat{\phi}_{\mathrm{s}}(\xi))\widehat{f}(\xi) \right| \mathrm{d}\xi &= \int_{\Omega} \left| (1 - \overline{\widehat{\phi}_{\mathrm{a}}(\xi)}\widehat{\phi}_{\mathrm{s}}(\xi))\widehat{f}(\xi) \right| \mathrm{d}\xi \\
&\quad + \int_{\mathbb{T}^d\setminus\Omega} \left| (1 - \overline{\widehat{\phi}_{\mathrm{a}}(\xi)}\widehat{\phi}_{\mathrm{s}}(\xi))\widehat{f}(\xi) \right| \mathrm{d}\xi \\
&\leq 3\varepsilon \int_{\Omega} \left| \widehat{f}(\xi) \right| \mathrm{d}\xi + 5 \int_{\mathbb{T}^d\setminus\Omega} \left| \widehat{f}(\xi) \right| \mathrm{d}\xi \leq 8\varepsilon \|\widehat{f}\,\|_1 .
\end{aligned}$$

Collecting terms we conclude

$$\|f - \tilde{f}\|_\infty \leq \left[\left(\sum_{n\in\mathbb{Z}^d} |\langle f, T_n\phi_{\mathrm{a}}\rangle| \right) + 11 \|\widehat{f}\,\|_1 \right] \varepsilon. \tag{8}$$

Now, let Λ be a finite subset of $\mathbb{Z}^d$. Then,

$$\|\tilde{f} - f_\Lambda\|_\infty \leq \left(\sum_{n \notin \Lambda} |\langle f, T_n \phi_\mathrm{a} \rangle| \right) \|\hat{\phi}_\mathrm{s}\|_1 \,. \tag{9}$$

Next, take G to be a group of orthogonal transformations acting on $\mathbb{R}^d$ and $M \in G$. If $\rho(M) f \in \mathscr{B}_\Omega^\varepsilon$, then Eqs. (4), (8), and (9) imply that for every Λ be a finite subset of $\mathbb{Z}^d$ we have

$$\begin{aligned}
&\|\rho(M) f - (\rho(M) f)_\Lambda\|_\infty \\
&\quad \leq \left[\left(\sum_{n \in \mathbb{Z}^d} |\langle \rho(M) f, T_n \phi_\mathrm{a} \rangle| \right) + 11 \|\widehat{f}\|_1 \right] \varepsilon + \left(\sum_{n \notin \Lambda} |\langle \rho(M) f, T_n \phi_\mathrm{a} \rangle| \right) \|\hat{\phi}_\mathrm{s}\|_1 \\
&\quad = \left[\left(\sum_{n \in \mathbb{Z}^d} |\langle f, T_{Mn} \rho(M^T) \phi_\mathrm{a} \rangle| \right) + 11 \|\widehat{f}\|_1 \right] \varepsilon + \left(\sum_{n \notin \Lambda} \left| \langle f, T_{Mn} \rho(M^T) \phi_\mathrm{a} \rangle \right| \right) \\
&\|\hat{\phi}_\mathrm{s}\|_1 .
\end{aligned}$$

Assuming $\hat{f} \in W^{1,2m}$ and that $\hat{\phi}_\mathrm{a}$ has bounded partial derivatives up to order $2m$ with $m \geq 1$, Eq. (7) gives

$$|\langle f, T_{Mn} \rho(M^T) \phi_\mathrm{a} \rangle| \leq \frac{\left\| \Delta^m \left(\widehat{f} \, \overline{\rho(M^T) \widehat{\phi}_\mathrm{a}} \right) \right\|_1}{(2\pi)^m \|Mn\|_2^m}, \quad n \in \mathbb{Z}^d. \tag{10}$$

Since, $\|Mn\| = \|n\|$ for all grid points n, we conclude that the estimate of the error $\|\rho(M) f - (\rho(M) f)_\Lambda\|_\infty$ provided above depends on the norm $\left\| \Delta^m \left(\widehat{f} \, \overline{\rho(M^T) \widehat{\phi}_\mathrm{a}} \right) \right\|_1$, which depends on M.

We can now summarize the previous discussion in the following theorem.

Theorem 1. *Assume that G is a group of orthogonal transformations acting on $\mathbb{R}^d$, $\Omega \subseteq (\mathbb{T}^d)^\circ$, $0 < \varepsilon < 1$. Suppose, $M(\Omega) \subseteq \Omega$ for all $M \in G$. We also assume that ϕ_a and ϕ_s are analysis and synthesis kernels satisfying Properties (1)–(7), $\Omega \subset B_\mathrm{a}^\circ$, $B_\mathrm{a} \subset B_\mathrm{s}^\circ$, and $B_\mathrm{s} \subset (\mathbb{T}^d)^\circ$ and $\rho(M)\phi_\mathrm{a} = \phi_\mathrm{a}$ for all $M \in G$. In addition, we assume that $\hat{\phi}_\mathrm{a}$ has bounded partial derivatives up to order $2m$ with $m \in \mathbb{Z}^+$. Then, for every $f \in \mathscr{B}_\Omega^\varepsilon$ such that $\hat{f} \in W^{1,2m}$ the following estimate holds:*

$$\|\rho(M) f - (\rho(M) f)_{\Lambda_N}\|_\infty \leq C_1 \varepsilon + C_2 \sum_{\|n\|_2 \geq N} \|n\|_2^{-2m}, \qquad M \in G,$$

where $\Lambda_N = \prod_{s=1}^d [-N_s, N_s]$ with $N_s = N$ for all $s = 1, 2, \ldots, d$. The constants C_1 and C_2 depend only on f.

Proof. First, observe that $M(\Omega) \subseteq \Omega$ for all $M \in G$ implies that, if $f \in \mathscr{B}^{\varepsilon}_{\Omega}$, then, $\rho(M)f \in \mathscr{B}^{\varepsilon}_{\Omega}$. Since, $\rho(M)\phi_{\mathrm{a}} = \phi_{\mathrm{a}}$ for all $M \in G$ using (10), we conclude

$$|\langle f, T_{Mn}\rho(M^T)\phi_{\mathrm{a}}\rangle| \leq \frac{\left\| \Delta^m \left(\widehat{f}\, \widehat{\phi_{\mathrm{a}}} \right) \right\|_1}{(2\pi)^m \|n\|_2^m}, \quad n \in \mathbb{Z}^d.$$

So, the factors $C_1 := \left(\sum_{n\in\mathbb{Z}^d} |\langle f, T_{Mn}\rho(M^T)\phi_{\mathrm{a}}\rangle|\right) + 11\|\widehat{f}\,\|_1$ and $C_2 := (2\pi)^{-m}\|\hat{\phi}_{\mathrm{s}}\|_1 \left\| \Delta^m \left(\widehat{f}\, \widehat{\phi_{\mathrm{a}}} \right) \right\|_1$ depend only on f. The conclusion follows from (10) and

$$\begin{aligned} \|\rho(M)f - (\rho(M)f)_{\Lambda}\|_{\infty} \leq & \left[\left(\sum_{n\in\mathbb{Z}^d} |\langle f, T_{Mn}\rho(M^T)\phi_{\mathrm{a}}\rangle| \right) + 11\|\widehat{f}\,\|_1 \right] \varepsilon \\ & + \left(\sum_{n\notin\Lambda_N} \left|\langle f, T_{Mn}\rho(M^T)\phi_{\mathrm{a}}\rangle\right| \right) \|\hat{\phi}_{\mathrm{s}}\|_1. \leq C_1\varepsilon + C_2 \\ & \sum_{\|n\|_2\geq N} \|n\|_2^{-2m} \end{aligned}$$

since M preserves norms.

Remark 1.

1. Theorem 1 requires analysis kernels to be invariant under the action of the groups of orthogonal transformations that may affect an image of interest. A slight modification of the statement of Theorem 1 can make it applicable to data representations defined by families of analysis and synthesis kernels instead of a single pair of kernels. Popular examples of these families are shearlegs [23, 32] and curvelets [8]. Condition $\rho(M)\phi_{\mathrm{a}} = \phi_{\mathrm{a}}$ for all $M \in G$ is now replaced by the requirement that the family of analysis filters must remain invariant under the action of G. In other words, if one ϕ_{a} works well for f, then $\rho(M^T)\phi_{\mathrm{a}}$ must be another analysis filter in the family of filters used by the data representation to allow to maintain control over the size of Λ when approximating $\rho(M)f$ by $(\rho(M)f)_{\Lambda}$ and thus maintain the sparsity of the representation.
2. The hypothesis $M(\Omega) \subseteq \Omega$ for all $M \in G$ in the statement of the previous theorem is not redundant. Indeed, assume that ϕ_{a} and ϕ_{s} satisfy all of the assumptions of the previous theorem. In addition, we assume $B_{\mathrm{s}} = \mathbb{T}^2$ and that $\hat{\phi}_{\mathrm{a}}$ vanishes outside $\mathbb{T}^2$. Take $0 < d < \frac{\sqrt{3-2\sqrt{2}}}{2\sqrt{2}}$ and assume that $B_{\mathrm{a}} = [-\frac{1}{2} + d, \frac{1}{2} - d]^2$ and $\Omega = B_{\mathrm{a}}^{\circ}$. Now pick f so that $\hat{f}$ is smooth and vanishes outside $[\frac{\sqrt{2}}{4}, \frac{1}{2} - d] \times [\frac{\sqrt{2}}{4}, \frac{1}{2} - d]$. If M is the rotation by $\pi/4$, then,

Table 1 Experiment results

N	$\frac{\|\rho(M)f-(\rho(M)f)_{\Lambda_N}\|_\infty}{\|f-f_{\Lambda_N}\|_\infty}$	$\|f-f_{\Lambda_N}\|_\infty$	$\|\rho(M)f-(\rho(M)f)_{\Lambda_N}\|_\infty$
35	$1.3023\cdot 10^{15}$	$6.3961\cdot 10^{-31}$	$8.3299\cdot 10^{-16}$
45	$2.0835\cdot 10^{163}$	$3.8408\cdot 10^{-180}$	$8.0026\cdot 10^{-17}$
50	$2.6247\cdot 10^{267}$	$1.2856\cdot 10^{-298}$	$3.3746\cdot 10^{-31}$
55	∞	0	$1.3558\cdot 10^{-108}$
60	∞	0	$2.8357\cdot 10^{-228}$
65	undefined	0	0

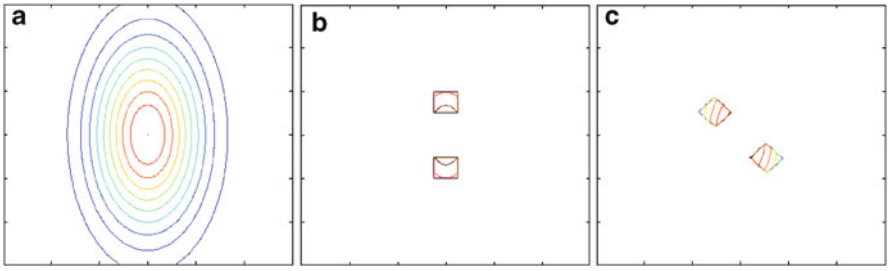


Fig. 3 (**a**) $\hat{\phi}_{\mathrm{a}}$, (**b**) $\hat{f}\cdot\hat{\phi}_{\mathrm{a}}$, (**c**) $[\rho(M)\hat{f}]\cdot\hat{\phi}_{\mathrm{a}}$

$(\rho(M)\hat{f})\overline{\hat{\phi}_{\mathrm{a}}}=0$ therefore $\langle\rho(M)f,T_n\phi_{\mathrm{a}}\rangle=0$, for all $n\in\mathbb{Z}^2$. In this case $\|\rho(M)f-(\rho(M)f)_\Lambda\|_\infty=\|f\|_\infty$ for every $\Lambda\subset\mathbb{Z}^2$.

We close this section with a simulation intending to demonstrate how rotations affect the rate of decay of $\|\rho(M)f-(\rho(M)f)_{\Lambda_N}\|_\infty$ as $N\to\infty$.

Let M be a rotation by $\pi/4$ in $\mathbb{R}^2$. Consider f such that $\widehat{f}(\xi_1,\xi_2)=\chi_{I_1}(\xi_1)\chi_{I_2}(\xi_2)$, where $I_1=[-\sigma_1,-\sigma_2]\cup[\sigma_2,\sigma_1]$ and $I_2=[-\sigma_3,\sigma_3]$. Let

$$\widehat{\phi}_{\mathrm{a}}(\xi_1,\xi_2)=\mathrm{e}^{-\frac{\xi_1^2}{2\sigma_4^2}-\frac{\xi_2^2}{2\sigma_5^2}}\quad\text{and}\quad\widehat{\phi}_{\mathrm{s}}(\xi_1,\xi_2)=\mathrm{e}^{-\frac{(\xi_1^2+\xi_2^2)}{2\sigma_6^2}},$$

where $0\le\sigma_2\le\sigma_1\le\sigma_4\le\sigma_6\le 1$ and $0\le\sigma_3\le\sigma_5\le\sigma_6$. In this example we set $\sigma_6=0.6,\sigma_5=0.55,\sigma_2=0.15,\sigma_4=0.5,\sigma_1=0.25,\sigma_3=0.12$. We compute the errors $\|\rho(M)f-(\rho(M)f)_{\Lambda_N}\|_\infty$ and $\|f-f_{\Lambda_N}\|_\infty$, as N grows, where $\Lambda_N=\{(m,n)\in\mathbb{Z}^2:\ \max(|m|,|n|)\le N\}$.

In Table 1 we can observe the results of this experiment. Notice that $(\rho(M)f)_{\Lambda_N}$ converges to zero more slowly than f_{Λ_N}.

Figure 3 shows the contour figures for $\widehat{\phi}_{\mathrm{a}}$, $\hat{f}\cdot\widehat{\phi}_{\mathrm{a}}$, and $[\rho(M)\hat{f}]\cdot\widehat{\phi}_{\mathrm{a}}$.

5 Construction of Synthetic Tubular 3D Data Sets

As mentioned above, a fundamental step in the development of a computational platform for neuronal reconstructions is the segmentation of the dendritic arbor and the extraction of its centerline. Validating the accuracy of the performance of these two tasks heavily relies on the manual segmentation of dendritic arbors which is time consuming, tedious, and often quite subjective. Therefore, the benchmarking of dendritic arbor segmentation and centerline extraction algorithms often becomes controversial, as most of the times the "gold standard" entirely relies on the experience of the user and cannot be verified against histology. Indeed, it is very common in neuroscience imaging that segmentation results manually obtained by experts working on the same data sets significantly differ, and automatic segmentation algorithms may outperform experts. Hence, there is a real need for highly accurate computational phantoms representing tubular structures in 3D that can be used to benchmark the baseline performance of segmentation and morphological reconstruction algorithms. To this end, we introduce a new method for the construction of highly accurate computational phantoms that represent the geometry of realistic tubular 3D structures. Our method yields very complex 3D data sets emulating with high accuracy at the resolution level normally used in confocal microscopy, and the prescribed morphological properties are centerline, branching points and branch diameter. Thus such data sets enable the reliable validation of segmentation and centerline extraction algorithms. It is clear that noisy data sets can easily be derived from our algorithm using standard methods like those in [68].

One feature of our approach is the ability to simulate varying fluorescence intensity values even within the same cross-section of the volume. The basic scenario for the spatial distribution of the fluorescence intensity values assumes that at any cross-section the maximum intensity occurs only at centerline voxel. The intensity values for each voxel in a cross section perpendicular to the centerline (*transversal cross section*) decreases almost radially, in the sense that voxels in the same transversal cross section and equidistant from the centerline voxel have the same intensity values. Moreover, for any two transversal cross-sections whose centerline voxels have the same intensity values, the spatial distributions of the intensity values in these cross sections are identical. We refer to this model of spatial distribution of intensity values as the *ideal tubular intensity distribution model*. This radial symmetry of the intensity function can only be implemented approximately in a digital phantom, since voxels are not dimensionless in the 3D space. Moreover, the centerline must be smooth everywhere except possibly at branching points.

5.1 Related Work

Only few methods to generate synthetic data for tubular objects can be found in the literature. In particular, Canero et al. [9] proposed a method to generate images

of synthetic vessels. After generating a random centerline, the intensity for each pixel in the vessel is modeled as a function of the distance from the pixel to the centerline and the radius of the vessel. Vasilkoski et al. [68] proposed a method to generate a 3D image stack of a neuron assuming that the infused fluorophore is distributed uniformly throughout the neuron and that the background intensity is zero. Then, they convolve the volume with a Gaussian point spread function to simulate the photon count $\mu(x, y, z)$ in the image stack. The actual photon count $n(x, y, z)$ for each voxel was randomly generated using the Poisson distribution with mean equal to $\mu(x, y, z)$. Bouix et al. [6] created synthetic tubular structures by using a predefined centerline. They slide a sphere centered on the centerline starting at a seed point. Voxel intensities are all the same inside the tubular volume while noise may be added at the final stage. Unfortunately, all these methods tend to suffer from a significant degree of geometric artifacts.

5.2 Methods

The first step in our approach, to create synthetic data volumes such as dendritic arbor phantoms is to construct the prescribed volume at a very high spatial resolution level, much higher than the resolution level used in confocal microscopy. Next, to reduce the resolution of those volumes, we downsample the data by a factor of two per dimension. To bring the data set to the desired resolution level we typically repeat the downsampling step as needed (typically three or four times). The *problem that often arises with this reduction of resolution is aliasing causing image degradation with errors directionally varying in 3D*. This problem compromises the radial symmetry of the ideal tubular intensity distribution model. To reduce the effect of such aliasing we first filter the input volume with a low-pass antialiasing filter. One might naturally wonder *what are the properties that the antialiasing filters should have in order to minimize the adverse effects of this reduction of resolution on the symmetry properties imposed by the ideal tubular intensity distribution model*. Although we do not directly address this question, we will provide below a family of antialiasing filters and justify why they are suitable for this application by invoking Theorem 1. We are now ready to proceed with the description of our approach.

Using the prescribed geometric properties (i.e., centerline points, branching and terminal points, and thickness) of the sought tubular structure (e.g., a dendrite) we first create a very high resolution approximation of the desired volume in the physical domain. In the language of multiscale analysis, this high-resolution image provides an approximation of the physical structure at a very high scale and, so, it may not be distinguished from the prototype structure living in the physical "continuous" domain $\mathbb{R}^3$. We denote this initial volume by I_0. To create the centerline of I_0 we use cubic spline interpolation in 3D. Using this centerline and the diameter information we create a "mask" M_0 which is an indicator function taking two values only, 0 if a voxel does not belong to I_0 and 1 otherwise. To create M_0 we superimpose spheres of radii matching the diameter of I_0 at the location

where the sphere is centered. The centers of these spheres belong to the centerline of I_0. The centers of these spheres are not uniformly distributed on the centerline of I_0. Their distribution varies depending on the spatial accuracy needed for M_0 (Fig. 4a). The use of spheres helps to successfully and accurately create the curved parts of M_0 with low computational cost. Each voxel in the interior of each of one of these spheres is assigned the value 1 (Fig. 4b). Thus, M_0 is defined to be a characteristic function of the set of voxels whose value is at least equal to 1. To algorithmically define the transversal cross sections in I_0, for each voxel $\mathbf{v}$ with mask value $M_0(\mathbf{v}) = 1$, we assign a "tag", $\mathbf{c}(\mathbf{v})$, where $\mathbf{c}(\mathbf{v})$ is the most proximal centerline voxel to $\mathbf{v}$. We thus partition I_0 into cross sections via the equivalence relation $\mathbf{v_1} \sim \mathbf{v_2}$ if and only if $\mathbf{c}(\mathbf{v_1}) = \mathbf{c}(\mathbf{v_2})$. Figure 4e depicts how this equivalence relation defines cross sections in a digital high-resolution volume I_0. So far, the intensity values of I_0 are identical with the values of M_0.

To complete the construction of I_0 we assign the desired intensity values for each voxel at which M_0 is equal to 1. For all other voxels the intensity is set constant to a fixed "background" value. For dendritic arbor phantoms the luminosity intensity on the centerline may decay as the distance of a point in the dendrite from the soma increases in order to simulate the decaying concentration of the infused fluorophore in distal branches from the point of infusion. However, several other models may be chosen to simulate the fluorescence intensity values induced by various fluorophore administration protocols.

In our approach, we assume that both rates of decay of the intensity values, radially, in any transversal cross section and along the centerline, are constant. However, the theoretical model of the luminosity intensity at cross sections assumes that this function is an isotropic Gaussian. This assumption is standard across all proposed models for confocal microscopy data. Implementing though an isotropic Gaussian in small transversal cross sections gives results no different from those obeying the linear decay model both radially in transversal cross section and along the centerline. Figure 4e depicts an example of the luminosity intensity in a transversal cross section with radius $R = 50$ pixels, background intensity $I_{\mathrm{BG}} = 10$, and maximum intensity at the center of the cross section $I_{\max} = 150$.

Figures 5a–d depicts the intensity obtained for a synthetic tubular structure: shown are images for the plane $x - y$ at $z = 160, 168, 176, 180$. The maximum intensity is obtained at $z = 160$ because the centerline is on this plane. The intensity value at a voxel is high if the voxel is close to the skeleton and the intensity decays as we approach the boundary.

The original synthetic volume I_0 created so far has 8 or even 16 times higher resolution than that of a typical data set acquired using confocal microscopy. In order to generate a 3D data volume useful for our purposes we need to drastically reduce the resolution by a factor of 8 at least. Simple downsampling is the first obvious, yet bad choice. Downsampling following the application of a special antialiasing filter is the right approach. In the following, we argue about the properties of this filter that mitigate undesirable directional aliasing.

A plain cylinder in $\mathbb{R}^3$ can be modeled using the tensor product of two Gaussians,

$$f_{\sigma_1,\sigma_2}(x, y, z) = \mathrm{e}^{-\frac{x^2}{2\sigma_1^2}} \mathrm{e}^{-\frac{y^2+z^2}{2\sigma_2^2}} \quad x, y, z, \in \mathbb{R}.$$

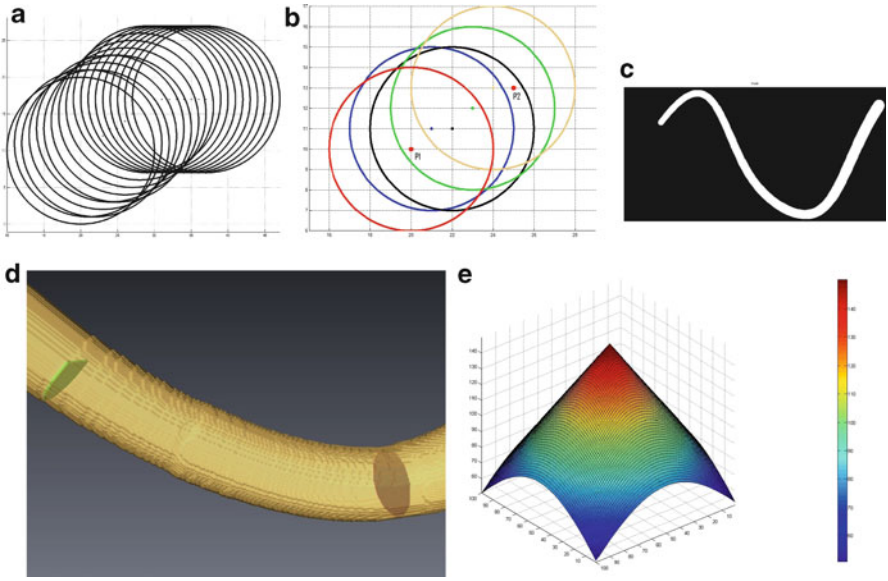


Fig. 4 (**a**) Side view of cross sections of spheres whose centers belong to the centerline. (**b**) A snapshot of the resulting mask M_0 from the same observation point. The small marching step of centers of the spheres yields a smooth digitized mask. (**c**) Binary mask with radius increasing along the centerline. (**d**) Circular cross section for two points on the centerline. (**e**) Graph of fluorescent intensity on a circular cross section

The centerline of this cylinder is the x-axis. We take $\sigma_1, \sigma_2 > 0$. The first Gaussian factor controls the length of the cylinder while the second controls the decay of the intensity values of this structure. The cylinder can be oriented to any different centerline by applying a 3D rotation R on the argument of f and must be digitized in a way that, ideally, does not generate artifacts due to its spatial orientation. The original tubular structure $\mathscr{I}$ can then be considered as finite sum of the form

$$\mathscr{I} = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} T_{x_i} R_k a_{i,k,j_1,j_2} f_{\sigma_{j_1},\sigma_{j_2}} \ , \quad R_k \in \mathrm{SO}(3), \quad a_{i,k,J_1,j_2} > 0.$$

It is this volume defined on $\mathbb{R}^3$ and from this volume we essentially create I_0 by applying Theorem 1. Since, the rotations R_k may be random we must assume that the set Ω the theorem requires must be invariant under all 3D rotations. Pick a desirable $0 < \varepsilon < 1$. Since

$$\hat{f}(\xi_1, \xi_2, \xi_3) = (2\pi)^{\frac{3}{2}} \sigma_{j_1} \sigma_{j_2}^2 \mathrm{e}^{-2\pi^2 \sigma_{j_1}^2 \xi_1^2} \mathrm{e}^{-2\pi^2 \sigma_{j_2}^2 (\xi_2^2 + \xi_3^2)},$$

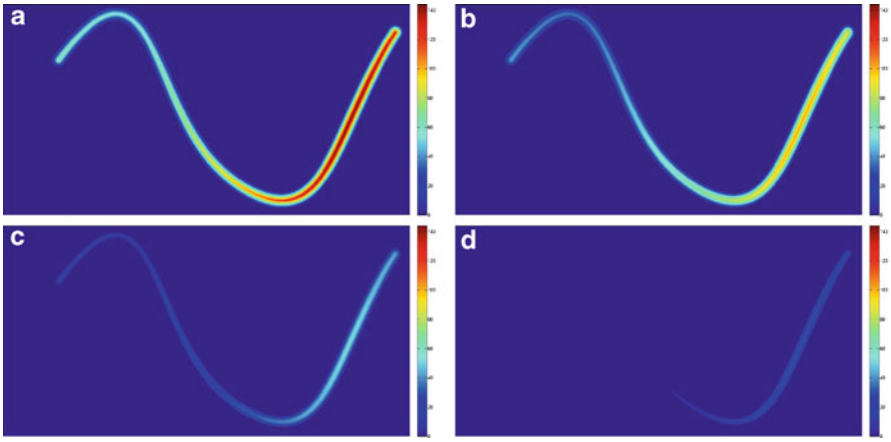


Fig. 5 (**a**)–(**d**) Intensity values on the high-resolution synthetic volume at different x-y planes (z=160,168,176,180)

it is not hard to observe that Ω must contain all sets of the form $[-\frac{a_{j_1,j_2}}{\sigma_{j_1}}, \frac{a_{j_1,j_2}}{\sigma_{j_1}}] \times [-\frac{a_{j_1,j_2}}{\sigma_{j_2}}, \frac{a_{j_1,j_2}}{\sigma_{j_2}}] \times [-\frac{a_{j_1,j_2}}{\sigma_{j_2}}, \frac{a_{j_1,j_2}}{\sigma_{j_2}}]$, where $a_{j_1,j_2} > 0$ depends on ε and all rotations of these parallelepipeds. This implies that Ω must be a sphere centered at the origin whose radius is greater than all $\frac{a_{j_1,j_2}}{\sigma_{j_1}}$ and all $\frac{a_{j_1,j_2}}{\sigma_{j_2}}$. Then, according to Theorem 1, the analysis kernel we use (theoretically only) to derive I_0 from $\mathscr{I}$ must be radial. To this end we use a refinable function ϕ_{a} which defines an Isotropic Multiresolution Analysis [50] which is an MRA with the additional property that each resolution space V_j is invariant under rotations as well. The use of the MRA will soon become clear. Take

$$\widehat{\phi^a}(\xi) := \begin{cases} 1, & |\xi| < 1/4 \\ \frac{1+\cos\left(6\pi|\xi|-\frac{3\pi}{2}\right)}{2}, & 1/4 < |\xi| < 5/12 \\ 0, & |\xi| > 5/12 \end{cases}$$

and consider $\phi_{\mathrm{a}}^j := 2^{3j/2}\phi_{\mathrm{a}}(2^j \cdot)$ as the analysis kernel of Theorem 1 where j is the appropriate scale required by the theorem. Note that in this case $\Omega = B_{\mathrm{a}} = B(0, 2^j/4)$. This condition determines the scale j. The synthesis kernel is of similar form, but we do not need it here, because the volume I_0 consists of the values $\{\langle \mathscr{I}, T_{2^{-j}n}\phi_{\mathrm{a}}^j \rangle : n \in \mathbb{Z}^3\}$. Thus, we will make no further reference to it. There is one added benefit which we obtain for free. Since ϕ_{a}^j has compact support in the frequency domain, I_0 is covariant to translations. Simply, one does not need worry about the effect of translations in this digitization process.

The Isotropic Multiresolution Analysis allows to reduce the resolution as needed. This is where we make use of the fact that this construct is an MRA. To do so, we use as the antialiasing filter the mask H_0^a of the refinable function ϕ_{a}. This is given, in the frequency domain, by

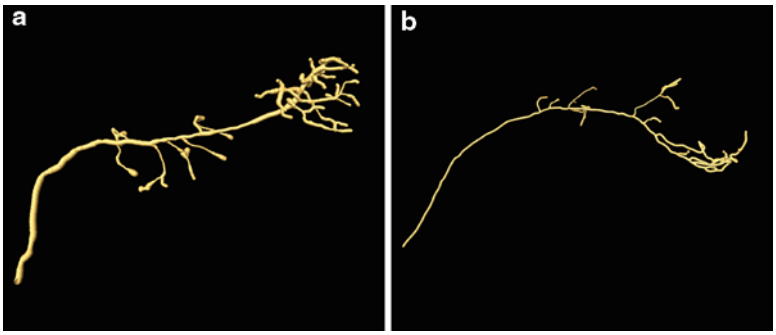


Fig. 6 Phantoms of olfactory dendrites (OP1 (**a**) and OP2 (**b**)) generated from information from the Diadem competition site

$$H_0^a(\xi) = \begin{cases} 1, & |\xi| < 1/8 \\ \frac{1+\cos\left(12\pi|\xi|-\frac{3\pi}{2}\right)}{2}, & 1/8 < |\xi| < 5/24 \\ 0, & |\xi| > 5/ < 24. \end{cases}$$

To summarize the previous discussion we list the steps of the proposed algorithm for generating synthetic 3D data sets of tubular structures.

Algorithm 1

Require: Manual reconstruction of a neuron.
Ensure: A computational phantom of a neuron.
Step 1: Create a high resolution volume

1.1 Refine manual reconstruction: compute for each branch new centerline points using cubic interpolation.
1.2 Create neuron's shape: center a sphere at each centerline and assign value equal to 1 to each voxel inside the sphere.
1.3 Compute the intensity for each voxel: The intensity is a function of the distance from each voxel to the centerline and radius of tubular structure and it satisfies the ideal tubular intensity distribution model.

Step 2: Downsample volume

2.1 Decrease the resolution: Apply an isotropic low-pass filter, e.g. H_0^a and downsample.

5.3 Experiments

We performed two sets of experiments to illustrate our algorithm. For the first set of experiments we construct simple volumes such as straight cylinder whose centerline lies on a circle. For the second set of experiments, we constructed three synthetic

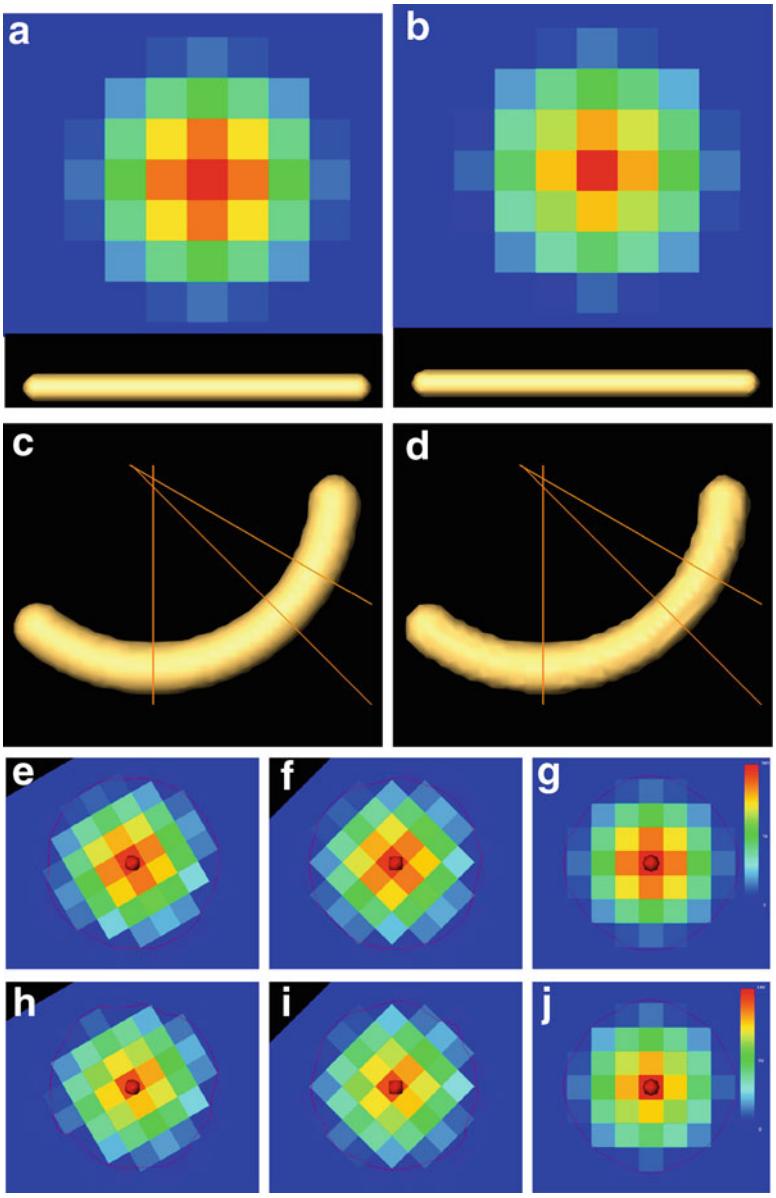


Fig. 7 (**a**) and (**b**) Depiction of a cross section and isosurface of cylinder using our method and symlets. (**c**) and (**d**) Isosurface of the second volume for first set of experiments using our method and symlets; slices at three angles (30°, 45°, and 90°) with respect to the arc of the circle depicting the centerline. (**e**)–(**g**) Depiction of a cross section at the slices shown on (**c**) by our method; (**h**)–(**j**) cross section at the slices shown on (**d**) by symlets

dendrite volumes using specifications from the DIADEM competition (Fig. 6). The reader can observe how the radial symmetry of the cross-sectional intensity function is achieved regardless of the incidence angle of the cross section, due to the use of isotropic filters with small transition band, such as the proposed IMRA filter H_0^a.

On these sets we evaluated the capabilities of the proposed method. We focused on the following three desirable properties. (1) *The symmetry of the luminosity intensity function in every cross section*: this function must satisfy $I(n) = I(m)$ if $\|n - c\| = \|n - m\|$, where c is the center of the cross section. (2) *The smoothness of the centerline*: in a realistic volume, the centerline must be a polygonal line. (3) *The variation of the angle of the normal vector at any point on the boundary isosurface and the centerline*. Typically, this angle must be equal to 90° except at bifurcation points. We used these criteria to qualitatively evaluate the performance of the proposed method using both the isotropic low-pass-IMRA filters H and filters obtained from a tensor product of 1D symlets. It can be observed from Fig. 7 that the isotropic filter performs better than the symlet filter counterpart.

Acknowledgements This work was supported in part by NSF grants DMS 0915242, DMS 1005799, and DMS 1008900 and by NHARP grant 003652-0136-2009.

References

1. Al-Kofahi, K., Lasek, S., Szarowski, D., Pace, C., Nagy, G.: Rapid automated three-dimensional tracing of neurons from confocal image stacks. IEEE Trans. Information Technology in Biomedicine **6**(2), 171–187 (2002)
2. Ascoli, G.A.: Progress and perspectives in computational neuroanatomy. Anat. Record. **257**(6), 195–207 (1999)
3. Bai, W., Zhou, X., Ji, L., Cheng, J., Wong, S.T.C.: Automatic dendritic spine analysis in two-photon laser scanning microscopy images. Cytometry Part A **71A**(10), 818–826 (2007). DOI 10.1002/cyto.a.20431. URL http://dx.doi.org/10.1002/cyto.a.20431
4. Bodmann, B., Melas, A., Papadakis, M., Stavropoulos, T.: Analog to digital revisited: Controlling the accuracy of reconstruction. Sampl. Theory Signal Image Process. **5**(3), 321–340 (2006)
5. Boor, C.D., DeVore, R., Ron, A.: Approximation from shift-invariant subspaces of $l_2(\mathbb{R}^d)$. Trans. Amer. Math. Soc. **341**(2), 787–806 (1994)
6. Bouix, S., Siddiqi, K., Tannenbaum, A.: Flux driven automatic centerline extraction. Med. Image Anal. **9**(3), 209–221 (2005)
7. Brown, K., Barrionuevo, G., Canty, A., Paola, V., Hirsch, J., Jefferis, G., Lu, J., Snippe, M., Sugihara, I., Ascoli, G.: The DIADEM data sets: representative light microscopy images of neuronal morphology to advance automation of digital reconstructions. Neuroinform. **9**(2–3), 143–157 (2011)
8. Candès, E.J., Donoho, D.L.: New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. Comm. Pure Appl. Math. **57**(2), 219–266 (2004)
9. Caṇero, C., Radeva, P.: Vesselness enhancement diffusion. Pattern Recogn. Lett. **24**(16), 3141–3151 (2003)
10. CBL: ORION: Online Reconstruction and functional Imaging Of Neurons (2008). URL http://www.cbl.uh.edu/ORION

11. Cheng, J., Zhou, X., Miller, E., Witt, R., Zhu, J., Sabatini, B., Wong, S.: A novel computational approach for automatic dendrite spines detection in two-photon laser scan microscopy. J. Neurosci. Methods **165**(1), 122–134 (2007)
12. Cheng, J., Zhou, X., Miller, E., Alvarez, V., Sabatini, B., Wong, S.: Oriented markov random field based dendritic spine segmentation for fluorescence microscopy images. Neuroinformatics **8**, 157–170 (2010). URL http://dx.doi.org/10.1007/s12021-010-9073-y. 10.1007/s12021-010-9073-y
13. Choy, S., Chen, K., Zhang, Y., Baron, M., Teylan, M., Kim, Y., Tong, C.S., Song, Z., Wong, S.: Multi scale and slice-based approach for automatic spine detection. In: Engineering in Medicine and Biology Society (EMBC), pp. 4765–4768 (2010)
14. Curtis, H.J., Cole, K.S.: Transverse electric impedance of the squid giant axon. J. General Physiol. **21**, 757–765 (1938)
15. DeVore, R.: Non-linear approximation. Acta Numer. **7**, 51–150 (1998)
16. DeVore, R.A., Jawerth, B., Popov, V.: Compression of wavelet decomposition. Am. J. Math. **114**(4), 737–785 (1992)
17. Dumitriu, D., Rodriguez, A., Morrison, J.H.: High-throughput, detailed, cell-specific neuroanatomy of dendritic spines using microinjection and confocal microscopy. Nat. Protocols **6**(9), 1391–1411 (2011). DOI 10.1038/nprot.2011.389. URL http://dx.doi.org/10.1038/nprot.2011.389
18. Fan, J., Zhou, X., Dy, J., Zhang, Y., Wong, S.: An automated pipeline for dendrite spine detection and tracking of 3d optical microscopy neuron images of in vivo mouse models. Neuroinformatics **7**, 113–130 (2009). URL http://dx.doi.org/10.1007/s12021-009-9047-0. 10.1007/s12021-009-9047-0
19. Glaser, J., Glaser, E.: Neuron imaging with neurolucida-a pc-based system for image combining microscopy. Comput. Med. Imaging Graph. **14**(5), 307–317 (1990)
20. Gonzalez, G., Fleuret, F., Fua, P.: Automated delineation of dendritic networks in noisy image stacks. In: Proceedings of European Conference on Computer Vision, pp. 214–227. Marseille, France (2008)
21. Govindarajan, A., Kelleher, R.J., Tonegawa, S.: A clustered plasticity model of long-term memory engrams. Nat. Rev. Neurosci. **7**(7), 575–583 (2006). DOI 10.1038/nrn1937. URL http://dx.doi.org/10.1038/nrn1937
22. Govindarajan, A., Israely, I., Huang, S.Y., Tonegawa, S.: The dendritic branch is the preferred integrative unit for protein synthesis-dependent ltp. Neuron **69**(1), 132–146 (2011). DOI 10.1016/j.neuron.2010.12.008. URL http://www.sciencedirect.com/science/article/pii/S0896627310009931
23. Guo, K., Labate, D.: Optimally sparse multidimensional representation using shearlets. SIAM J. Math. Anal. **39** (2007)
24. Hines, M., Carnevale, N.: NEURON: a tool for neuroscientists. Neuroscientist **7**, 123–135 (2001)
25. Hodgkin, A.L., Huxley, A.F., Katz, B.: Measurement of current-voltage relations in the membrane of the giant axon of loligo. J. Physiol. **116**, 424448 (1952)
26. Jingfan, L., Gensun, F.: On truncation error bound for multidimensional sampling expansion Laplace transform. Anal. Theory Appl. **1**, 52–57 (2004)
27. Jowet, B.: Plato: Theaetetus. http://ebooks.adelaide.edu.au/p/plato/p71th/index.html
28. Jetter, K., Plonka, G.: A survey on L^2-approximation order from shift-invariant spaces. In: Dyn, N., Leviatan, D., Levin, D., Pinkus, A. (eds.) Multivariate Approximation and Applications, pp. 73–111. Cambridge University Press, Cambridge (2001)
29. Kakadiaris, I., Santamaría-Pang, A., Colbert, C., Saggau, P.: Morphological reconstruction of living neurons. In: Rittscher, J., Machiraju, R., Wong, S. (eds.) Microscopic Image Analysis for Life Science Applications. Artech House Publishers, Norwood (2007)
30. Kakadiaris, I., Santamaría-Pang, A., Colbert, C., Saggau, P.: Automatic 3-D morphological reconstruction of neuron cells from multiphoton images. In: Rittscher, J., Machiraju, R., Wong, S. (eds.) Microscopic Image Analysis for Life Science Applications, pp. 389–399. Artech House, Norwood (2008). DOI 978-1-59693-236-4
31. Katz, B.: Nerve, Muscle and Synapse. McGraw-Hill, New York (1966)

32. Labate, D., Lim, W., Kutyniok, G., Weiss, G.: Sparse multidimensional representation using shearlets. In: Unser, M. (ed.) Proceedings of Wavelets XI, SPIE Proceedings, vol. 5914, pp. 247–255 (2005)
33. Leviatan, D., Temlyakov, V.N.: Simultaneous approximation by greedy algorithms. Advances in Computational Mathematics **25**(1), 73–90 (2006)
34. Li, Q., Zhou, X., Deng, Z., Baron, M., Teylan, M., Kim, Y., Wong, S.: A novel surface-based geometric approach for 3d dendritic spine detection from multi-photon excitation microscopy images. In: Proceedings of Biomedical Imaging: From Nano to Macro, 2009. ISBI '09. IEEE International Symposium on, pp. 1255–1258 (2009). DOI 10.1109/ISBI.2009.5193290
35. Losavio, B., Reddy, G., Colbert, C., Kakadiaris, I., Saggau, P.: Combining optical imaging and computational modeling to analyze structure and function of living neurons. In: Proceedings of 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 1, pp. 668–670. New York, NY (2006). DOI 10.1109/IEMBS.2006.259552
36. Losavio, B., Liang, Y., Santamaria-Pang, A., Kakadiaris, I., Colbert, C., Saggau, P.: Live neuron morphology automatically reconstructed from multiphoton and confocal imaging data. J. Neurophysiol. **100**, 2422–2429 (2008). DOI 10.1152/jn.90627.2008
37. Lu, J.: Neuronal tracing for connectomic studies. Neuroinformatics **9**(2–3), 159–166 (2011)
38. Luisi, J., Narayanaswamy, A., Galbreath, Z., Roysam, B.: The farsight trace editor: an open source tool for 3-D inspection and efficient pattern analysis aided editing of automated neuronal reconstructions. Neuroinformatics **9**(2–3), 305–315 (2011). DOI 10.1007/s12021-011-9115-0
39. Meijering, E.: Neuron tracing in perspective. Cytometry Part A **77A**(7), 693–704 (2010). DOI 10.1002/cyto.a.20895. URL http://dx.doi.org/10.1002/cyto.a.20895
40. Narayanaswamy, A., Wang, Y., Roysam, B.: 3-D image pre-processing algorithms for improved automated tracing of neuronal arbors. Neuroinformatics **9**(2–3), 219–231 (2011)
41. Olenko, A., Pogány, T.: A precise upper bound for the error of interpolation of stochastic processes. Theory Probab. Math. Stat. **71**, 151–163 (2005)
42. Pelt, J.v., Schierwagen, A.: Morphological analysis and modeling of neuronal dendrites. Math. Biosci. **188**(1–2), 147–155 (2004)
43. Peng, H., Ruan, Z., Atasoy, D., Sternson, S.: Automatic reconstruction of 3D neuron structures using a graph-augmented deformable model. Bioinformatics **26**(12) (2010). URL http://www.biomedsearch.com/nih/Automatic-reconstruction-3D-neuron-structures/20529931.html
44. Plonka, G.: Approximation order provided by refinable function vectors. Constr. Approx. **13**(2), 221–244 (1997)
45. Q., L., Z., D.: A surface-based 3d dendritic spine detection approach from confocal microscopy images. IEEE Trans. Image Process. (2011). To appear
46. Reddy, G.D., Saggau, P.: Development of a random-access multi-photon microscope for fast three-dimensional functional recording of neuronal activity (2007).
47. Rodriguez, A., Ehlenberger, D., Kelliher, K., Einstein, M., Henderson, S., Morrison, J., Hof, P., Wearne, S.: Automated reconstruction of three-dimensional neuronal morphology from laser scanning microscopy images. Methods **30**(1), 94–105 (2003)
48. Rodriguez, A., Ehlenberger, D., Hof, P., Wearne, S.: Rayburst sampling, an algorithm for automated three-dimensional shape analysis from laser-scanning microscopy images. Nat. Protoc. **1**, 2156–2161 (2006)
49. Rodriguez, A., Ehlenberger, D., Hof, P., Wearne, S.: Three-dimensional neuron tracing by voxel scooping. J. Neurosci. Methods **184**(1), 169–175 (2009). DOI 10.1016/j.jneumeth.2009.07.021
50. Romero, J., Alexander, S., Baid, S., Jain, S., Papadakis, M.: The geometry and the analytic properties of isotropic multiresolution analysis. Adv. Computat. Math. **31**, 283–328 (2009). DOI 10.1007/s10444-008-9111-6
51. Rusakov, D., Stewart, M.: Quantification of dendritic spine populations using image analysis and a tilting dissector. J. Neurosci. Methods **60**, 11–21 (1995)
52. Santamaria-Pang, A., Bildea, T., Colbert, C., Saggau, P., Kakadiaris, I.: Towards segmentation of irregular tubular structures in 3D confocal microscope images. In: Proceedings of MICCAI Workshop in Microscopic Image Analysis and Applications in Biology, pp. 78–85. Denmark, Copenhagen (2006). DOI 10.1.1.97.3671

53. Santamaria-Pang, A., Colbert, C., Losavio, B., Saggau, P., Kakadiaris, I.: Automatic morphological reconstruction of neurons from optical images. In: Proceedings of International Workshop in Microscopic Image Analysis and Applications in Biology. Piscataway, NJ (2007)
54. Santamaria-Pang, A., Colbert, C., Saggau, P., Kakadiaris, I.: Automatic centerline extraction of irregular tubular structures using probability volumes from multiphoton imaging. In: Proceedings of Medical Image Computing and Computer-Assisted Intervention, pp. 486–494. Brisbane, Australia (2007). DOI 10.1007/978-3-540-75759-759
55. Santamaria-Pang, A., Herrera, P.H., Papadakis, M., Prott, A., Shah, S., Kakadiaris, I.: Automatic morphological reconstruction of neurons from multiphoton and confocal microscopy images (2011). Submitted
56. Scorcioni, R., Polavaram, S., Ascoli, G.A.: L-measure: a web-accessible tool for the analysis, comparison and search of digital reconstructions of neuronal morphologies. Nat. Protoc. **3**(5), 866–876 (2008). DOI doi:10.1038/nprot.2008.51
57. Senft, S.: A brief history of neuronal reconstruction. Neuroinformatics **9**(2–3), 119–128 (2011)
58. Shen, H., Sesack, S., Toda, S., Kalivas, P.: Automated quantification of dendritic spine density and spine head diameter in medium spiny neurons of the nucleus accumbens. Brain Struct. Funct. **213**, 149–157 (2008). URL http://dx.doi.org/10.1007/s00429-008-0184-2. 10.1007/s00429-008-0184-2
59. Shen, H.W., Toda, S., Moussawi, K., Bouknight, A., Zahm, D.S., Kalivas, P.W.: Altered dendritic spine plasticity in cocaine-withdrawn rats. J. Neurosci. **29**(9), 2876–2884 (2009). DOI 10.1523/JNEUROSCI.5638-08.2009. URL http://www.jneurosci.org/content/29/9/2876.abstract
60. Strohmer, T., Tanner, J.: Implementations of Shannon's sampling theorem, a time-frequency approach. Sampl. Theory Signal Image Process. **4**(1), 1–17 (2005)
61. Swanger, S., Yao, X., Gross, C., Bassell, G.: Automated 4D analysis of dendritic spine morphology: applications to stimulus-induced spine remodeling and pharmacological rescue in a disease model. Mol. Brain **4**, 1–14 (2011). URL http://dx.doi.org/10.1186/1756-6606-4-38. 10.1186/1756-6606-4-38
62. Temlyakov, V.N.: Greedy algorithms with regard to multivariate systems with special structure. Constr. Approx. **16**(3), 399–425 (2000)
63. Temlyakov, V.N.: Weak greedy algorithms. Adv. Comput. Math. **12**(2–3), 213–227 (2000)
64. Temlyakov, V.N.: Greedy algorithms in banach spaces. Adv. Comput. Math. **14**(3), 277–292 (2001)
65. Tropp, J.: Algorithms for simultaneous sparse approximation. part ii: convex relaxation. Signal Processing, special issue "Sparse approximations in signal and image processing" **86**, 589–602 (2006)
66. Tropp, J., Gilbert, A., Strauss, M.: Algorithms for simultaneous sparse approximation. part I: greedy pursuit. Signal Processing, special issue "Sparse approximations in signal and image processing" **86**, 572–588 (2006)
67. Uehara, C., Colbert, C.M., Saggau, P., Kakadiaris, I.: Towards automatic reconstruction of dendrite morphology from live neurons. In: Proceedings of 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1798–1801. San Francisco, CA (2004). DOI 10.1109/IEMBS.2004.1403537
68. Vasilkoski, Z., Stepanyants, A.: Detection of the optimal neuron traces in confocal microscopy images. J. Neurosci. Methods **178**(1), 197–204 (2009)
69. Wang, Y., Narayanaswamy, A., Tsai, C.L., Roysam, B.: A broadly applicable 3-D neuron tracing method based on open-curve snake. Neuroinformatics **9**(2–3), 193–217 (2011). DOI 10.1007/s12021-011-9110-5
70. Wearne, S., Rodriguez, A., Ehlenberger, D., Rocher, A., Henderson, S., Hof, P.: New techniques for imaging, digitization and analysis of three-dimensional neural morphology on multiple scales. Neuroscience **136**(3), 661–680 (2005). DOI 10.1016/j.neuroscience.2005.05.053. Quantitative Neuroanatomy: from molecules to system. A special issue in honor of the late Professor Theodor W. Blackstad

71. Weaver, C.M., Hof, P.R., Wearne, S.L., Lindquist, W.B.: Automated algorithms for multiscale morphometry of neuronal dendrites. Neural Comput. **16**(7), 1353–1383 (2004). DOI 10.1162/089976604323057425. URL http://dx.doi.org/10.1162/089976604323057425
72. Yuste, R.: Dendritic Spines. MIT, Cambridge (2009)
73. Yuste, R., Denk, W.: Dendritic spines as basic functional units of neuronal integration. Nature **375**(6533), 682–684 (1995)
74. Yuste, R., Bonhoeffer, T.: Genesis of dendritic spines: insights from ultrastructural and imaging studies. Nat. Rev. Neurosci. **5**(1), 24–34 (2004). DOI 10.1038/nrn1300.
75. Zhang, Y., Zhou, X., Witt, R., Sabatini, B., Adjeroh, D., Wong, S.: Dendritic spine detection using curvilinear structure detector and LDA classifier. Neuroimage **36**(2), 346–360 (2007) URL http://dx.doi.org/10.1038/nrn1300
76. Zhang, Y., Chen, K., Baron, M., Teylan, M., Kim, Y., Song, Z., Greengard, P., Wong, S.: A neurocomputational method for fully automated 3d dendritic spine detection and segmentation of medium-sized spiny neurons. NeuroImage **50**(4), 1472–1484 (2010). DOI DOI:10.1016/j.neuroimage.2010.01.048. URL http://www.sciencedirect.com/science/article/B6WNP-4Y7P6W0-3/2/7224d8d28ad16bd42e865849ed68810a

Index

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 2*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8379-5,