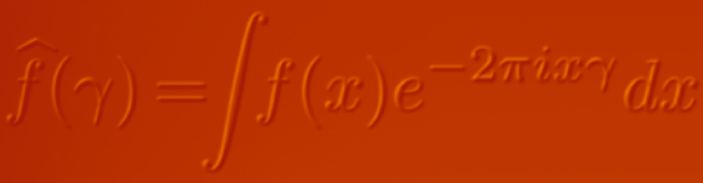
Applied and Numerical Harmonic Analysis



Radu Balan, John J. Benedetto,
Wojciech Czaja, Matthew Dellatorre,
Kasso A. Okoudjou, Editors

Excursions in Harmonic Analysis, Volume 5

The February Fourier Talks at the Norbert Wiener Center

Birkhäuser

Applied and Numerical Harmonic Analysis

More information about this series at http://www.springer.com/series/4968

Radu Balan • John J. Benedetto • Wojciech Czaja
Matthew Dellatorre • Kasso A. Okoudjou
Editors

Excursions in Harmonic Analysis, Volume 5

The February Fourier Talks at the Norbert Wiener Center

Birkhäuser

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

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

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

Matthew Dellatorre
Norbert Wiener Center
University of Maryland
College Park, MD, USA

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

ISSN 2296-5009 ISSN 2296-5017 (electronic)
Applied and Numerical Harmonic Analysis
ISBN 978-3-319-54710-7 ISBN 978-3-319-54711-4 (eBook)
DOI 10.1007/978-3-319-54711-4

Library of Congress Control Number: 2012951313

Printed on acid-free paper

This book is published under the trade name Birkhäuser, www.birkhauser-science.com
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Dedicated to

Ying Wang and Dennis Healy,

inspiring members of our harmonic analysis family
and lost to us when they were so young.

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-the-art *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 with 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:

Antenna theory	*Prediction theory*
Biomedical signal processing	*Radar applications*
Digital signal processing	*Sampling theory*
Fast algorithms	*Spectral estimation*
Gabor theory and applications	*Speech processing*
Image processing	*Time-frequency and time-scale analysis*
Numerical partial differential equations	*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, for example, 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 this Vol. 5 have at least one author who spoke at the February Fourier Talks during the period 2014–2016. Volumes 1–4 cover the February Fourier Talks during the period 2002–2013.

The February Fourier Talks (FFT)

The *FFT*s were initiated in 2002 and 2003 as small meetings on harmonic analysis and applications, held at the University of Maryland, College Park. There were no *FFT*s in 2004 and 2005. The Norbert Wiener Center (NWC) for Harmonic Analysis and Applications was founded in 2004 in the Department of Mathematics at the university, and, since 2006, the *FFT* has been organized by the NWC. The *FFT* has developed into a major annual conference that brings together applied and pure harmonic analysts along with scientists and engineers from universities, industry, and government for an intense and enriching 2-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 2-day single-track meeting consisting of 30-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	Keynote	Colloquium
• Robert Calderbank	• Peter Carr	• Richard Baraniuk
• Ronald Coifman	• Barry Cipra	• Rama Chellappa
• Ingrid Daubechies	• James Coddington	• Margaret Cheney
• Ronald DeVore	• Nathan Crone	• Charles Fefferman
• Richard Kadison	• Ali Hirsa	• Robert Fefferman
• Peter Lax	• Mario Livio	• Gerald Folland
• Elias Stein	• William Noel	• Christopher Heil
• Gilbert Strang	• Steven Schiff	• Peter Jones
	• Mark Stopfer	• Thomas Strohmer
	• Frederick Williams	• Victor Wickerhauser

In 2013, the February Fourier Talks was followed by a workshop on phaseless reconstruction, also hosted by the Norbert Wiener Center and intellectually in the spirit of the *FFT*.

The Norbert Wiener Center

The Norbert Wiener Center for Harmonic Analysis and Applications provides a national focus for the broad area of applied harmonic analysis. Its theoretical underpinnings form the technological basis for many applications. Further, the applications themselves impel the study of fundamental harmonic analysis issues in topics such as signal and image processing, machine learning, data mining, waveform design, and dimension reduction.

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 website: www.norbertwiener.umd.edu

The Structure of the Volumes

To some extent, the four parts for each of these 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 part 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

Volume 3

IX Special Topics in Harmonic Analysis
X Applications and Algorithms in the Physical Sciences
XI Gabor Theory
XII RADAR and Communications: Design, Theory, and Applications

Volume 4

XIII Theoretical Harmonic Analysis
XIV Sparsity
XV Signal Processing and Sampling
XVI Spectral Analysis and Correlation

Volume 5

XVII Theoretical Harmonic Analysis
XVIII Image and Signal Processing
XIX Quantization
XX Algorithms and Representations

Acknowledgments

The editors of Vol. 5 gratefully acknowledge additional editorial assistance by Dr. Alfredo Nava-Tudela, as well as the support of Benjamin Levitt, editor for Birkhäuser Science in New York.

The Norbert Wiener Center also gratefully acknowledges the indispensable support of the following groups: Birkhäuser and Springer Publishers; the IEEE Baltimore Section; MiMoCloud, Inc.; Radyn, Inc.; the SIAM Washington-Baltimore Section; and Reality Analytics, Inc. 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 XVII
Theoretical Harmonic Analysis

Real analysis, harmonic analysis, and representation theory have all been present and well represented at the FFT conferences over the years. This volume contains contributions from three utmost distinguished researchers (and their collaborators) in these areas.

The Heisenberg group is intimately connected with the quantum mechanics. More recently it has been recognized as a central tool in time-frequency signal processing. Folland's chapter presents a framework for analysis and of the discrete Heisenberg group based on a direct integral decomposition of its irreducible representations on $L^2(\mathbb{R})$. The author shows in the rational case the Zak transform represents the unitary operators of time and frequency shifts as a collection of finite dimensional unitary matrices of same size acting independently on spaces indexed by points of a 2-dimensional square. To illustrate this result, consider the simpler case of integer time and frequency shifts. In this case the unitary shifts commute and the Zak transform diagonalizes simultaneously these operators. The irrational case is more complicated and the author shows that the inequivalent representations cannot be indexed by Lebesgue measurable sets.

Grafakos' chapter on fractional differentiation is a beautiful exposition of a hard core real analysis problem. We all take for granted the Leibniz's product rule of differentiation and the Hölder inequality. When put together, they control the L^1 norm of a higher order derivative of a product of two functions by dual L^p norms of lower order derivatives. The question the author studies is what happens if the regular derivative is replaced by a fractional derivative defined using the Fourier transform. In his chapter he elegantly summarizes the state of affair for this problem and presents the sharpest result for a Kato-Ponce type inequality.

The chapter by Farsi, Gillaspy, Kang, and Packer presents an overview of C^*-algebras theory based wavelet constructions. The Cuntz C^*-algebras representations had been shown to be closely related to construction of Multi Resolution Analysis (MRA) orthonormal wavelets. The current chapter starts with a survey of these known results. Next the authors consider several ways to generalize these results by using C^*-algebras associated to higher-rank graphs. Their construction generalizes previous approaches using graph Laplacian wavelets and MRA. One targeted application is the spatial traffic analysis on k-graphs in network engineering.

Time-Frequency Analysis and Representations of the Discrete Heisenberg Group

Gerald B. Folland

Abstract The operators $[\varrho_\omega(j,k,l)f](t) = e^{2\pi i\omega l}e^{2\pi i\omega kt}f(t+j)$ on $L^2(\mathbb{R})$ constitute a representation of the discrete Heisenberg group. We investigate how this representation decomposes as a direct integral of irreducible representations. The answer is quite different depending on whether ω is rational or irrational, and in the latter case it provides illustrations of some interesting pathological phenomena.

Keywords Discrete Heisenberg group • Unitary representations • Direct integral decompositions

1 Introduction

Among the most basic operators in signal analysis are the translations in time and frequency space, also known as translations and modulations: these are the unitary operators T_x and M_y ($x, y \in \mathbb{R}$) on $L^2(\mathbb{R})$ defined by

$$T_x f(t) = f(t+x), \qquad M_y f(t) = e^{2\pi iyt} f(t). \tag{1}$$

Since $T_x M_y = e^{2\pi ixy} M_y T_x$, the collection of operators $\{e^{2\pi iz} M_y T_x : x, y, z \in \mathbb{R}\}$ forms a group. If one considers the group structure in the abstract, one is led to the (real) *Heisenberg group* H, which is $\mathbb{R}^3$ equipped with the group law

$$\begin{aligned}(x,y,z)(x',y',z') &= (x+x',\, y+y',\, z+z'+xy'),\\ (x,y,z)^{-1} &= (-x,-y,-z+xy).\end{aligned} \tag{2}$$

More precisely, the group generated by the translations and modulations consists of the image of H under the unitary representation $R : H \to \mathcal{U}(L^2(\mathbb{R}))$ defined by $R(x,y,z) = e^{2\pi iz} M_y T_x$, that is,

G.B. Folland (✉)
Department of Mathematics, University of Washington, Seattle, WA 98195, USA
e-mail: folland@uw.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4_1

$$R(x,y,z)f(t) = e^{2\pi iz}e^{2\pi iyt}f(t+x). \tag{3}$$

The representation R is irreducible, i.e., there are no nontrivial closed subspaces of $L^2(\mathbb{R})$ that are invariant under it. Indeed, suppose $f, g \in L^2(\mathbb{R})$ and $g \perp R(x,y,z)f$ for all x, y, z. Then

$$0 = \langle R(x,y,0)f, g\rangle = \int e^{2\pi iyt}f(t+x)\overline{g(t)}\,dt$$

for all x, y. By Fourier uniqueness, $f(t+x)\overline{g(t)} = 0$ for a.e. (x,t). Taking the abolute square of both sides and integrating first in x and then in t, we see that $\|f\|_2\|g\|_2 = 0$, that is, either $f = 0$ or $g = 0$.

For future reference we note that there is a more symmetric way to describe the group law of H. Namely, let $\widetilde{H}$ be $\mathbb{R}^3$ equipped with the group law

$$\begin{aligned}(x,y,z)(x',y',z') &= \big(x+x',\ y+y',\ z+z'+\tfrac{1}{2}(xy'-yx')\big),\\ (x,y,z)^{-1} &= (-x,-y,-z).\end{aligned} \tag{4}$$

It is easy to check that the map $(x,y,z) \mapsto (x,y,z+\frac{1}{2}xy)$ is an isomorphism from $\widetilde{H}$ to H.

Ever since Gabor's fundamental paper [4], it has been of interest to study the discrete group of operators generated by the translations and modulations by integer multiples of some fundamental quantities τ and ω, that is, $T_{j\tau}$ and $M_{k\omega}$ with $j, k \in \mathbb{Z}$. (Note that since $T_{j\tau}M_{k\omega} = e^{2\pi i\tau\omega jk}M_{k\omega}T_{j\tau}$, the scalars needed here to fill out the group are $e^{2\pi iz}$ with z an integer multiple of $\tau\omega$.) The abstract group structure in this situation is that of the *discrete Heisenberg group* H, which is $\mathbb{Z}^3$ equipped with the group law (2) — but we shall write elements of H as (j,k,l) rather than (x,y,z).

By rescaling the real line, we may and shall assume that $\tau = 1$. Thus, for a given $\omega > 0$, we are considering the unitary representation ϱ_ω of H on $L^2(\mathbb{R})$ defined by $\varrho_\omega(j,k,l) = e^{2\pi i\omega l}M_{k\omega}T_j$, that is,

$$\varrho_\omega(j,k,l)f(t) = e^{2\pi i\omega l}e^{2\pi i\omega kt}f(t+j). \tag{5}$$

The representations ϱ_ω, in contrast to R, are highly reducible, and it is natural to ask how they decompose into irreducible representations. These decompositions involve not direct sums but direct integrals, a concept that we shall review briefly in section 2. When ω is rational, the solution to this problem turns out to be a nice exercise in Fourier analysis that involves one of the signal analysts' favorite devices, the Zak transform; we shall present it in section 3. When ω is irrational, however, one has to confront the fact that H is a "non-type-I" group, which means that its representation theory displays various pathologies (see Folland [3]). One of them is that the set of unitary equivalence classes of irreducible representations is geometrically bizarre and cannot, in general, be used as a parameter space for direct integral decompositions. Another one is that it may be possible to express a representation as a direct integral of irreducibles in many completely

different ways. The analysis of our representations ϱ_ω in section 4 provides easily accessible illustrations of these phenomena. In particular, we recover some results of Kawakami [7] concerning the non-uniqueness, in a way that is simpler and more transparent than his original constructions.

Some terminology: we shall be concerned only with *unitary, strongly continuous* representations of locally compact groups G on *separable* Hilbert spaces. Two representations π and π' of G on Hilbert spaces $\mathcal{H}$ and $\mathcal{H}'$ are *equivalent* if there is a unitary operator $U : \mathcal{H} \to \mathcal{H}'$ that *intertwines* them, i.e., $U\pi(g) = \pi'(g)U$ for all $g \in G$; in this case, we write $\pi \sim \pi'$. Let Z be the center of G. If π is a representation of G such that Z acts by scalar multiples of the identity. i.e., $\pi(z) = \chi(z)I$ where $\chi : Z \to U(1) = \{\zeta \in \mathbb{C} : |\zeta| = 1\}$ (this always happens if π is irreducible, by Schur's lemma), χ is called the *central character* of π.

The center of the discrete Heisenberg group H is $\{(0, 0, l) : l \in \mathbb{Z}\}$, and the central character of the representation ϱ_ω defined by (5) is $\chi_\omega(l) = e^{2\pi i\omega l}$. The decomposition of ϱ_ω into irreducibles will involve only irreducible representations with the same central character. This is true on general grounds, but we will verify it by explicit calculations.

2 Direct Integrals

The general theory of direct integrals of Hilbert spaces and operators on them involves some measure-theoretic technicalities that need not concern us; for our purposes the following will suffice. Suppose $\{\mathcal{H}_\alpha : \alpha \in \mathbb{R}^n\}$ is a family of separable Hilbert spaces with inner products $\langle\cdot,\cdot\rangle_\alpha$ parametrized by an n-tuple α of real variables. We assume that the $\mathcal{H}_\alpha$ are all continuously embedded in a topological vector space $\mathcal{V}$ and that there is a family of vectors $\{e_j^\alpha : \alpha \in \mathbb{R}^n,\ j \in J\}$ in $\mathcal{V}$ (where J is a suitable index set) that depend continuously on α, such that for each α, $\{e_j^\alpha : j \in J\}$ is an orthonormal basis for $\mathcal{H}_\alpha$. For any Borel set $A \subset \mathbb{R}^n$, we then define the direct integral

$$\mathcal{H} = \int_A^\oplus \mathcal{H}_\alpha \, d\alpha$$

to be the set of all $f : A \to \mathcal{V}$ such that

(i) $f(\alpha) \in \mathcal{H}_\alpha$ for all $\alpha \in A$;

(ii) for each j, $\alpha \mapsto \langle f(\alpha), e_j^\alpha\rangle_\alpha$ is Borel measurable;

(iii) $\|f\|^2 \equiv \int_A \|f(\alpha)\|_\alpha^2 \, d\alpha < \infty$.

(The integrand in (iii) is measurable since $\|f(\alpha)\|_\alpha^2 = \sum_j |\langle f, e_j^\alpha\rangle_\alpha|^2$.) We identify two functions f and g in $\mathcal{H}$ if $f(\alpha) = g(\alpha)$ for Lebesgue-almost every α; $\mathcal{H}$ is then easily seen to be a Hilbert space.

If the $\mathcal{H}_\alpha$ all coincide with a fixed Hilbert space $\mathcal{H}_0$, we may take $\mathcal{V} = \mathcal{H}_0$ and $\{e_j^\alpha\} = \{e_j\}$ to be a fixed orthonormal basis for $\mathcal{H}_0$, and $\int_A^\oplus \mathcal{H}_\alpha \, d\alpha$ is simply $L^2(A, \mathcal{H}_0)$, the space of square-integrable $\mathcal{H}_0$-valued functions on A.

Now suppose that for each α, T_α is a unitary operator on $\mathcal{H}_\alpha$, depending measurably on α in the sense that $\langle T_\alpha e_j^\alpha, e_k^\alpha \rangle_\alpha$ is Borel measurable for all j, k. It is easy to check that if f is in $\int_A^\oplus \mathcal{H}_\alpha \, d\alpha$ then so is $\alpha \mapsto T_\alpha[f(\alpha)]$, so we can define the direct integral

$$T = \int_A^\oplus T_\alpha \, d\alpha$$

to be the unitary operator on $\int_A^\oplus \mathcal{H}_\alpha \, d\alpha$ given by

$$[Tf](\alpha) = T_\alpha[f(\alpha)].$$

Finally, if π_α is a unitary representation of a locally compact group G on $\mathcal{H}_\alpha$ for each $\alpha \in A$, depending measurably on α in the sense described above, we obtain the direct integral representation $\int_A^\oplus \pi_\alpha \, d\alpha$ of G on $\int_A^\oplus \mathcal{H}_\alpha \, d\alpha$ by applying this construction to each family of operators $\pi_\alpha(g)$, $g \in G$.

3 The Rational Case

We begin our analysis of the representations ϱ_ω of H defined by (5).

The simplest situation is where ω is a positive integer p. In this case the central character of ϱ_p is trivial, so that $\varrho_p(j, k, l)$ depends only on j and k, and ϱ_p is effectively a representation of the group $\mathbb{Z}^2$. The irreducible representations of this group, or of H with trivial central character, are the one-dimensional ones, that is, the characters

$$\chi_{u,v}(j, k, l) = \chi_{u,v}(j, k) = e^{2\pi i(ju+kv)}. \tag{6}$$

(Here we may regard u and v as elements of $\mathbb{R}$ or of $\mathbb{R}/\mathbb{Z}$ as convenience dictates; the same understanding will apply in similar situations below.)

The operation that relates ϱ_p to the characters $\chi_{u,v}$ is the *Zak transform*, the map $\mathcal{Z}$ from (reasonable) functions on $\mathbb{R}$ to functions on $\mathbb{R}^2$ defined by

$$\mathcal{Z}f(u, v) = \sum_{n \in \mathbb{Z}} e^{2\pi i n u} f(v - n). \tag{7}$$

Note that for $m \in \mathbb{Z}$,

$$\mathcal{Z}f(u + m, v) = \mathcal{Z}f(u, v), \qquad \mathcal{Z}f(u, v + m) = e^{2\pi i m u} \mathcal{Z}f(u, v),$$

so $\mathcal{Z}f$ is determined by its values on $[0,1)\times[0,1)$. Moreover, by the Parseval identity,

$$\int_0^1\int_0^1 |\mathcal{Z}f(u,v)|^2\,du\,dv = \sum_n \int_0^1 |f(v-n)|^2\,dv = \int_{\mathbb{R}} |f(t)|^2\,dt,$$

so $\mathcal{Z}$ is an isometry from $L^2(\mathbb{R})$ to $L^2([0,1])^2$ that is easily seen to be surjective, hence unitary. Finally, since $\varrho_p(j,k,l)f(t) = e^{2\pi ikpt}f(t+j)$, a simple calculation shows that

$$\mathcal{Z}\varrho_p(j,k,l)f(u,v) = e^{2\pi iju}e^{2\pi ikpv}\mathcal{Z}f(u,v) = \chi_{u,pv}(j,k,l)\mathcal{Z}f(u,v).$$

But this says that $\mathcal{Z}$ intertwines ϱ_p with the direct integral

$$\int_{[0,1)\times[0,1)}^{\oplus} \chi_{u,pv}\,du\,dv$$

(acting on $L^2([0,1)^2) = \int_{[0,1)\times[0,1)}^{\oplus} \mathbb{C}\,du\,dv$). By the rescaling $v \mapsto v/p$, this is equivalent to

$$\int_{[0,1)\times[0,p)}^{\oplus} \chi_{u,v}\,du\,dv.$$

Finally, by the periodicity of $\chi_{u,v}$ in u and v, this integral over $[0,1)\times[0,p)$ is the direct sum of p copies of the integral over $[0,1)^2$, or, more naturally, of the integral over $(\mathbb{R}/\mathbb{Z})^2$. In short, we have proved:

Theorem 1 *If p is a positive integer, ϱ_p is equivalent to the direct sum of p copies of the direct integral $\int_{(\mathbb{R}/\mathbb{Z})^2}^{\oplus} \chi_{u,v}\,du\,dv$, where $\chi_{u,v}$ is the one-dimensional representation defined by (6).*

The situation where ω is rational but not integral is similar but not quite so simple. For the rest of this section we assume that $\omega = p/q$ where p and q are relatively prime positive integers with $q > 1$. In this case we have $\varrho_\omega(0,0,l) = I$ when q divides l, so ϱ_ω is really a representation of the quotient group of H in which the central variable l is taken to be an integer modulo q. It is easy to obtain a complete list of irreducible representations of this group (up to equivalence) with central character $e^{2\pi i\omega l}$ by an application of the "Mackey machine." The details are worked out in Folland [3, §6.8]; here, we shall just quote the results.

For $u \in \mathbb{R}$, let

$$\mathcal{H}_u = \big\{f : \mathbb{Z} \to \mathbb{C} : f(m+nq) = e^{-2\pi iunq}f(m) \text{ for all } m,n \in \mathbb{Z}\big\}. \tag{8}$$

Any $f \in \mathcal{H}_u$ is completely determined by its values at $1,\dots,q$ (or any set of q consecutive integers), so $\mathcal{H}_u$ is q-dimensional. The norm on it is defined in the

obvious way: $\|f\|^2 = \sum_1^q |f(m)|^2$ ($= \sum_{M+1}^{M+q} |f(m)|^2$ for any M). Observe that this family of Hilbert spaces satisfies the conditions in section 2: we can take $\mathcal{V}$ to be $l^\infty(\mathbb{Z})$ and define e_j^u for $j = 1, \dots, q$ by $e_j^u(m) = e^{-2\pi i u n q}$ if $m = j + nq$ and $e_j^u(m) = 0$ if $m \not\equiv j \bmod q$.

Now let v be another real number. We define the representation $\pi_{u,v}$ of H on $\mathcal{H}_u$ by

$$\pi_{u,v}(j,k,l)f(m) = e^{2\pi i(p/q)l} e^{2\pi i k[v-(p/q)m]} f(m-j). \tag{9}$$

(This formula for $\pi_{u,v}$ depends only on v; the u-dependence comes from the space on which it acts.) A proof of the following result can be found in Folland [3, §6.8]:

Proposition 1 *Suppose p and q are relatively prime positive integers with $q > 1$. The representations $\pi_{u,v}$ of H defined by (8) and (9) are irreducible, and every irreducible representation of H with central character $e^{2\pi i(p/q)l}$ is equivalent to one of them.*

It is obvious that $\mathcal{H}_u = \mathcal{H}_{u'}$ if $u' \equiv u \bmod q^{-1}\mathbb{Z}$, and in this case $\pi_{u,v} = \pi_{u',v'}$ if $v' \equiv v \bmod \mathbb{Z}$. However, up to equivalence even more is true: $\pi_{u,v} \sim \pi_{u',v'}$ if $u' \equiv u$ and $v' \equiv v \bmod q^{-1}\mathbb{Z}$. Indeed, in this case we can write $v' = v + (r/q)$ for some $r \in \mathbb{Z}$, and hence $v' \equiv v + n(p/q) \bmod \mathbb{Z}$ for some $n \in \mathbb{Z}$. (Indeed, since p and q are relatively prime, there are integers a and b with $ap + bq = 1$, so that $v' = v + ar(p/q) + br$.) Then $\pi_{u',v'} = \pi_{u,v+n(p/q)}$, and it is easily verified that the map $Uf(m) = f(m+n)$ intertwines $\pi_{u,v+n(p/q)}$ with $\pi_{u,v}$. Hence:

Corollary 1 *The equivalence class of $\pi_{u,v}$ depends only on the image of (u,v) in $(\mathbb{R}/q^{-1}\mathbb{Z})^2$. In this way, the set of equivalence classes of irreducible representations of H with central character $e^{2\pi i(p/q)l}$ is in one-to-one correspondence with $(\mathbb{R}/q^{-1}\mathbb{Z})^2$.*

With this result in hand, the direct integral decomposition of $\varrho_{p/q}$ into irreducibles takes almost exactly the same form as Theorem 1: $\varrho_{p/q}$ is the direct sum of p copies of the integral of the $\pi_{u,v}$ over a complete set of equivalence classes. To derive this result, it will be convenient to start from the other end by building the direct integral in question. We begin by considering the direct integral

$$\pi_v = \int_{[0,1/q)}^{\oplus} \pi_{u,v}\, du = \int_{\mathbb{R}/q^{-1}\mathbb{Z}}^{\oplus} \pi_{u,v}\, du,$$

which acts on the Hilbert space

$$\begin{aligned}\mathcal{H} &= \int_{\mathbb{R}/q^{-1}\mathbb{Z}}^{\oplus} \mathcal{H}_u\, du \\ &= \left\{ f : (\mathbb{R}/q^{-1}\mathbb{Z}) \times \mathbb{Z} \to \mathbb{C} : f(u, m+kq) = e^{-2\pi i k q u} f(u,m),\ \|f\|_{\mathcal{H}} < \infty \right\}\end{aligned}$$

where

$$\|f\|_{\mathcal{H}}^2 = \sum_{m=1}^{q} \int_0^{1/q} |f(u,m)|^2 \, du.$$

An element of $\mathcal{H}$ is, in essence, a q-tuple of $1/q$-periodic functions of u. We wish to trade such a q-tuple in for a single 1-periodic function of u. To this end, observe that if $f \in \mathcal{H}$, we have $e^{2\pi i(m+kq)u} f(u, m+kq) = e^{2\pi imu} f(u,m)$, so the latter function (call it $f_m(u)$) depends only on the residue of m modulo q. It is 1-periodic in u, and its Fourier coefficients (defined by $f_m(u) = \sum_n c_{mn} e^{2\pi inu}$) are nonzero only for $n \equiv m \bmod q$ since $f(u,m)$ is $1/q$-periodic. We define $T : \mathcal{H} \to L^2(\mathbb{R}/\mathbb{Z})$ by

$$Tf(u) = q^{-1/2} \sum_{1}^{q} f_m(u) = q^{-1/2} \sum_{1}^{q} e^{2\pi imu} f(u,m). \tag{10}$$

By the preceding remarks, the sum could be taken over any set of q consecutive integers, and the terms in this sum are pairwise orthogonal. Hence

$$\int_0^1 |Tf(u)|^2 \, du = q^{-1} \sum_{1}^{q} \int_0^1 |f(u,m)|^2 \, du = \sum_{1}^{q} \int_0^{1/q} |f(u,m)|^2 \, du = \|f\|_{\mathcal{H}}^2,$$

so T is an isometry. In fact it is unitary: if $g \in L^2(\mathbb{R}/\mathbb{Z})$ has the Fourier series $\sum_{\nu \in \mathbb{Z}} c_\nu e^{2\pi i\nu u}$, we group the ν's according to their residues mod q and get

$$g(u) = \sum_{m=1}^{q} e^{2\pi imu} \sum_{n \in \mathbb{Z}} c_{m+nq} e^{2\pi inqu} = Tf(u),$$

where $f(u,m) = q^{1/2} \sum_j c_{m+nq} e^{2\pi inqu}$ for all $m \in \mathbb{Z}$ (not just $m = 1, \ldots, q$). This f does indeed belong to $\mathcal{H}$, for

$$\begin{aligned} f(u, m+kq) = q^{1/2} \sum_n c_{m+(n+k)q} e^{2\pi inqu} &= q^{1/2} \sum_n c_{m+nq} e^{2\pi i(n-k)qu} \\ &= e^{-2\pi ikqu} f(u,m). \end{aligned}$$

We now transfer the representation π_v to $L^2(\mathbb{R}/\mathbb{Z})$ by means of T, that is, we set

$$\widetilde{\pi}_v(j,k,l) = T\pi_v(j,k,l)T^{-1}.$$

Since

$$[\pi_v(j,k,l)f](u,m) = e^{2\pi i(p/q)l} e^{2\pi ik[v-(p/q)m]} f(u, m-j),$$

for $g(u) = \sum c_v e^{2\pi i v u}$ we have

$$[\widetilde{\pi}_v(j,k,l)g](u) = e^{2\pi i(p/q)l} \sum_{m=1}^{q} \sum_{n\in\mathbb{Z}} c_{m-j+nq} e^{2\pi i[mu+k(v-\omega m)+nqu]}.$$

By the observation following (10), we can replace m by $m+j$ while still summing from 1 to q, so using the fact that $e^{2\pi i nq(p/q)k} = 1$, we see that

$$\begin{aligned}
&[\widetilde{\pi}_v(j,k,l)g](u) \\
&= e^{2\pi i[(p/q)(l-jk)+ju+kv]} \sum_{m=1}^{q} e^{2\pi i m(u-(p/q)k)} \sum_{n\in\mathbb{Z}} c_{m+nq} e^{2\pi i nq(u-(p/q)k)} \\
&= e^{2\pi i[(p/q)(l-jk)+ju+kv]} g(u-(p/q)k).
\end{aligned}$$

Since the representations $\pi_{u,v}$ are 1-periodic in v, and their equivalence classes are $1/q$-periodic in v, the same is true of π_v and hence of $\widetilde{\pi}_v$. As the reader may verify, the intertwinings for $\widetilde{\pi}_v$ are given by the operators M_n defined by

$$M_n f(u) = e^{2\pi i n u} f(u). \tag{11}$$

More precisely, if $v' \equiv v \bmod q^{-1}\mathbb{Z}$, we have (as before) $v' \equiv v + n(p/q) \bmod \mathbb{Z}$ for some $n \in \mathbb{Z}$, and

$$\widetilde{\pi}_{v+n(p/q)}(j,k,l)M_n = M_n \widetilde{\pi}_v(j,k,l). \tag{12}$$

Next, we form the direct integral

$$\Pi = \int_{[0,p/q)}^{\oplus} \widetilde{\pi}_v \, dv,$$

which acts on $L^2((\mathbb{R}/\mathbb{Z}) \times [0,p/q))$ (we take the variable of integration to be the *second* variable in this product) by

$$[\Pi(j,k,l)h](\cdot,v) = \widetilde{\pi}_v(j,k,l)[h(\cdot,v)], \tag{13}$$

that is,

$$[\Pi(j,k,l)h](u,v) = e^{2\pi i[(p/q)(l-jk)+ju+kv]} h(u-k(p/q),v). \tag{14}$$

To put this into final form and remove the slightly artificial use of the interval $[0,p/q)$, we extend functions defined on $(\mathbb{R}/\mathbb{Z}) \times [0,p/q)$ to $(\mathbb{R}/\mathbb{Z}) \times \mathbb{R}$ by the prescription

$$h(u, v + n(p/q)) = e^{2\pi inu} h(u, v) = [M_n h](u, v) \qquad (n \in \mathbb{Z}), \tag{15}$$

where M_n is defined by (11) and it is understood that M_n acts in the first variable. Note that the extension of h is accomplished by initially taking $v \in [0, p/q)$ in (15), but once this is done, (15) holds for all $v \in \mathbb{R}$. In view of (12), this extension is compatible with the formula (13); that is, for $v \in [0, p/q)$ and $n \in \mathbb{Z}$,

$$\begin{aligned} \widetilde{\pi}_{v+n(p/q)}(j, k, l)[h(\cdot, v + n(p/q))] &= \widetilde{\pi}_{v+n(p/q)}(j, k, l) M_n[h(\cdot, v)] \\ &= M_n[\widetilde{\pi}_v(j, k, l)[h(\cdot, v)]. \end{aligned}$$

Thus we may, and shall, consider the representation Π as acting on functions on $\mathbb{R}^2$ that are 1-periodic in the first variable, quasi-periodic in the second variable according to (15), and square-integrable on $[0, 1) \times [0, p/q)$, with the action now given by (14) for all $u, v \in \mathbb{R}$. But now we can use (15) to give this a final reformulation:

$$\begin{aligned} [\Pi(j, k, l)h]\,(u, v) &= e^{2\pi i[(p/q)l + j(u - k(p/q)) + kv]} h(u - k(p/q), v) \\ &= e^{2\pi i(p/q)l} e^{2\pi ikv} [M_j h](u - k(p/q), v) \\ &= e^{2\pi i(p/q)l} e^{2\pi ikv} h(u - k(p/q), v + j(p/q)). \end{aligned} \tag{16}$$

It is now a simple matter to relate Π to our original representation $\varrho_{p/q}$ by a rescaling of the Zak transform. Namely, with $\mathcal{Z}$ given by (7), we define

$$\mathcal{Z}_{p/q} f(u, v) = (q/p)^{1/2} \mathcal{Z} f(u, (q/p)v) = (q/p)^{1/2} \sum_{n \in \mathbb{Z}} e^{2\pi inu} f((q/p)v - n).$$

Thus $\mathcal{Z}_{p/q}$ maps functions on $\mathbb{R}$ functions on $\mathbb{R}^2$ that are 1-periodic in the first variable and satisfy (15), and it is unitary from $L^2(\mathbb{R})$ to $L^2([0, 1) \times [0, p/q))$. Moreover,

$$\begin{aligned} &[\mathcal{Z}_{p/q} \varrho_{p/q}(j, k, l) f](u, v) \\ &= (q/p)^{1/2} \sum_{n \in \mathbb{Z}} e^{2\pi i(p/q)l} e^{2\pi i[k(v - n(p/q)) + nu]} f((q/p)v - n + j) \\ &= e^{2\pi i(p/q)l} e^{2\pi ikv} \mathcal{Z}_{p/q} f(u - k(p/q), v + j(p/q)) \\ &= [\Pi(j, k, l) \mathcal{Z}_{p/q} f](u, v). \end{aligned}$$

In short, $\mathcal{Z}_{p/q}$ intertwines $\varrho_{p/q}$ with Π, so these representations are equivalent. Moreover, by the $1/q$-periodicity of the equivalence class of $\widetilde{\pi}_v$ in v, $\Pi = \int_{[0,p/q)}^{\oplus} \widetilde{\pi}_v \, dv$ is equivalent to the direct sum of p copies of $\int_{[0,1/q)}^{\oplus} \widetilde{\pi}_v \, dv$, and the latter representation in turn is equivalent to $\int_{[0,1/q)^2}^{\oplus} \pi_{u,v} \, du \, dv$. We have proved:

Theorem 2 *If p and q are relatively prime positive integers with $q > 1$, the representation $\varrho_{p/q}$ is equivalent to the direct sum of p copies of the direct integral $\int_{[0,1/q)^2}^{\oplus} \pi_{u,v}\, du\, dv$, where $\pi_{u,v}$ is the representation defined by (8) and (9). In this integral, the integrand $\pi_{u,v}$ ranges over a complete set of inequivalent irreducible representations of H with central character $e^{2\pi i(p/q)l}$.*

4 The Irrational Case

In this section ω will denote a fixed positive *irrational* number, and our goal is again to analyze the decomposition of the representation ϱ_ω of H defined by (5) into irreducibles. The problem of classifying all the equivalence classes of irreducible representations of H with central character $e^{2\pi i\omega l}$ in a concrete way is completely intractable (see Folland [3, §6.8] for a fuller explanation of this issue), but we do not need all of them. To get started, it will suffice to consider the following one-parameter family of representations, which is analogous to the family $\{\pi_{u,v}\}$ in the rational case. To wit, for $v \in \mathbb{R}$ we define the representation σ_v of H on $l^2(\mathbb{Z})$ by

$$[\sigma_v(j,k,l)f](m) = e^{2\pi i\omega l} e^{2\pi ik(v-\omega m)} f(m-j). \tag{17}$$

Proposition 2 *The representations σ_v are all irreducible. Moreover, $\sigma_{v'} \sim \sigma_v$ if and only if $v' \equiv v \bmod \mathbb{Z} + \omega\mathbb{Z}$.*

Proof Suppose $\mathcal{V}$ is a nonzero σ_v-invariant subspace of $l^2(\mathbb{Z})$ and $0 \neq f \in \mathcal{V}$. If $g \perp \mathcal{V}$, then for all $j, k \in \mathbb{Z}$,

$$0 = \langle \sigma_v(j,k,0)f, g\rangle = e^{2\pi ivk} \sum_m e^{-2\pi i\omega km} f(m-j)\overline{g(m)}.$$

Since $f(\cdot -)\overline{g(\cdot)} \in l^1(\mathbb{Z})$, the function $\phi_j(\theta) = \sum_m e^{-2\pi im\theta} f(m-j)\overline{g(m)}$ is continuous on $\mathbb{R}/\mathbb{Z}$, and it vanishes at $\theta = \omega k$ for all k. Since ω is irrational, $\{\omega k : k \in \mathbb{Z}\}$ is dense in $\mathbb{R}/\mathbb{Z}$; hence, ϕ_j vanishes identically, so its Fourier coefficients $f(m-j)\overline{g(m)}$ are all zero. This being true for all j, it follows that $g = 0$, so $\mathcal{V} = l^2(\mathbb{Z})$. Thus σ_v is irreducible.

The standard basis vectors $e_n(m) = \delta_{mn}$ for $l^2(\mathbb{Z})$ are eigenvectors for the operators $\sigma_v(0,k,0)$ with eigenvalues $e^{2\pi ik(v-\omega n)}$. Thus if $v' \not\equiv v \bmod \mathbb{Z} + \omega\mathbb{Z}$, the operators $\sigma_{v'}(0,k,0)$ and $\sigma_v(0,k,0)$ have different eigenvalues, and so $\sigma_{v'} \not\sim \sigma_v$. On the other hand, if $v' = v + n + \omega p$ $(n, p \in \mathbb{Z})$, then obviously $\sigma_{v'} = \sigma_{v+\omega p}$, and it is easily checked that the operator $U_p f(m) = f(m+p)$ intertwines $\sigma_{v+\omega p}$ and σ_v, so $\sigma_{v'} \sim \sigma_v$. □

Proposition 3 (Baggett [1]) *If $\omega > 0$ is irrational and σ_v is given by (17),*

$$\varrho_\omega \sim \int_{[0,\omega)}^{\oplus} \sigma_v\, dv. \tag{18}$$

Proof The direct integral $\sigma = \int_{[0,\omega)}^{\oplus} \sigma_v \, dv$ acts on $L^2([0,\omega) \times \mathbb{Z})$ (with respect to Lebesgue measure times counting measure) by

$$\sigma(j,k,l)g(v,m) = e^{2\pi i\omega l} e^{2\pi ik(v-m\omega)} g(v, m-j).$$

We identify $\mathbb{R}$ with $[0,\omega) \times \mathbb{Z}$ by cutting up $\mathbb{R}$ into disjoint intervals of length 1 and dilating them by a factor of ω, and thereby define the unitary map $U : L^2(\mathbb{R}) \to L^2([0,\omega) \times \mathbb{Z})$ by

$$Uf(v,m) = \omega^{-1/2} f(\omega^{-1} v - m).$$

A simple calculation then shows that $U\varrho_\omega(j,k,l) = \sigma(j,k,l)U$, so $\varrho_\omega \sim \sigma$. □

Thus we have a simple-looking direct integral decomposition of ϱ_ω into irreducibles. Unlike the integrals for the rational case in Theorems 1 and 2, however, the representations σ_v in (18) are not all inequivalent, and *there is no Lebesgue measurable way to separate out the equivalence classes to obtain an integral over inequivalent representations with multiplicities*. We can, of course, use the fact that $\sigma_v = \sigma_{v+n}$ for $n \in \mathbb{Z}$ to reduce to integrals over $[0,\alpha)$ with $\alpha \leq 1$ or over $\mathbb{R}/\mathbb{Z}$. But by Proposition 2, the equivalence classes for the σ_v's are given by the cosets of $(\mathbb{Z} + \omega\mathbb{Z})/\mathbb{Z}$ in $\mathbb{R}/\mathbb{Z}$, and *none of the cross-sections for these cosets are Lebesgue measurable*. Indeed, these are essentially the classic examples of non-measurable sets. (Most textbooks use $\mathbb{Q}$ rather than $\mathbb{Z}+\omega\mathbb{Z}$ to build examples, but any subgroup of $\mathbb{R}$ whose image in $\mathbb{R}/\mathbb{Z}$ is countably infinite will serve the purpose.)

We now turn to the question of non-uniqueness.

The representation ϱ_ω of H is the restriction to H of an irreducible representation R_ω of the real Heisenberg group H, a close relative of the representation R defined by (3):

$$R_\omega(x,y,z)f(t) = R(x,\omega y,\omega z)f(t) = e^{2\pi i\omega z} e^{2\pi i\omega yt} f(t+x).$$

If Φ is an automorphism of H that leaves the center pointwise fixed, then $R_\omega \circ \Phi$ is another irreducible representation of H with the same central character $e^{2\pi i\omega z}$. But irreducible representations of H with nontrivial central character are completely determined up to equivalence by that character: this is the Stone-von Neumann theorem (see Folland [2, §1.5] or [3, §6.7]). Thus $R_\omega \circ \Phi \sim R_\omega$.

Now, if Φ preserves the discrete subgroup H, we also have $\varrho_\omega \circ \Phi \sim \varrho_\omega$, and hence

$$\int_{[0,\omega)}^{\oplus} \sigma_v \, dv \sim \varrho_\omega \sim \varrho_\omega \circ \Phi \sim \int_{[0,\omega)}^{\oplus} \sigma_v \circ \Phi \, dv.$$

The point is that *each of the representations $\sigma_v \circ \Phi$ may be inequivalent to all of the representations σ_v*, in which case we have two entirely different direct integral decompositions of ϱ_ω.

We are going to analyze this phenomenon for the group of automorphisms of H arising from an action of the group $SL(2,\mathbb{R})$ of 2×2 real matrices of determinant 1 on H. Our standard notation for elements of $SL(2,\mathbb{R})$ will be

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad (ad - bc = 1),$$

and for future reference we record a couple of simple facts about the subgroup $SL(2,\mathbb{Z})$ of matrices in $SL(2,\mathbb{R})$ with integer entries.

Lemma 1 *Suppose* $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL(2,\mathbb{Z})$. *Then:*

a. a and b are relatively prime.
b. ac and bd cannot both be odd.

Proof (a) Any number that divides a and b also divides $ad-bc$. (b) If ac and bd are odd, then a, b, c, d are all odd and hence $ad - bc$ is even. □

In terms of the symmetric form $\widetilde{H}$ of H with the group law (4), the action of $SL(2,\mathbb{R})$ is simply its natural action on the first pair of variables: for $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL(2,\mathbb{R})$,

$$\widetilde{\Phi}_A(x,y,z) = (ax + by,\ cx + dy,\ z).$$

The condition $\det A = 1$ is precisely what is needed to guarantee that the quantity $xy' - yx'$ (the signed area of the parallelogram spanned by (x,y) and (x',y')) in the group law (4) is preserved and hence that $\widetilde{\Phi}_A$ is an automorphism of $\widetilde{H}$. For our purposes we need to transport this action of $SL(2,\mathbb{R})$ to the group H with group law (2) by conjugating with the isomorphism $(x,y,z) \mapsto (x,y,z+\frac{1}{2}xy)$ from $\widetilde{H}$ to H; the reader may verify that the (somewhat uglier) result is

$$\Phi_A(x,y,z) = \big(ax + by,\ cx + dy,\ z + \tfrac{1}{2}(acx^2 + 2bcxy + bdy^2)\big). \tag{19}$$

Incidentally, the operators that intertwine R_ω and $R_\omega \circ \Phi_A$ are essentially given by the *metaplectic representation* μ of $SL(2,\mathbb{R}) = Sp(1,\mathbb{R})$ on $L^2(\mathbb{R})$; see Folland [2, §4.2] for a detailed description of μ. More precisely, $\mu(A)$ intertwines R and $R\circ\Phi_A$; the corresponding intertwiner for R_ω and $R_\omega \circ \Phi_A$ is $\mu_\omega(A) = D_\omega \mu(A) D_\omega^{-1}$, where $D_\omega f(t) = \omega^{1/4} f(\omega^{-1/2} t)$.

When does Φ_A preserve the discrete subgroup H? To preserve the integer lattice in the first two variables, it is necessary and sufficient that $A \in SL(2,\mathbb{Z})$, but the annoying factor of $\frac{1}{2}$ in (19) creates a problem in the third variable — fortunately, a minor one. By Lemma 1(b), at least one of ac and bd must be even, so there are three possibilities. If both are even, $\Phi_A|\mathsf{H}$ is an automorphism of H. If ac (resp. bd) is odd, it is an isomorphism from H to H' (resp. H''), where

$$\mathsf{H}' = \big\{(j,k,l) : j,k \in \mathbb{Z},\ l \in \tfrac{1}{2}\mathbb{Z},\ 2l \equiv j \bmod 2\big\},$$
$$\mathsf{H}'' = \big\{(j,k,l) : j,k \in \mathbb{Z},\ l \in \tfrac{1}{2}\mathbb{Z},\ 2l \equiv k \bmod 2\big\}.$$

But the formula (17) for $\sigma_v(j,k,l)$ makes perfectly good sense when l is a half-integer and thus defines a representation of H' or H''. Hence, in all cases $\sigma_v \circ \Phi_A$ is an irreducible representation of H.

Theorem 3 *Suppose* $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ *and* $A' = \left(\begin{smallmatrix} a' & b' \\ c' & d' \end{smallmatrix}\right)$ *are in* $SL(2,\mathbb{Z})$.

a. If $(a',b') = \pm(a,b)$, *then* $\sigma_v \circ \Phi_{A'} \sim \sigma_{\pm v} \circ \Phi_A$.
b. If $(a',b') \neq \pm(a,b)$, *then* $\sigma_{v'} \circ \Phi_{A'} \not\sim \sigma_v \circ \Phi_A$ *for all* $v, v' \in \mathbb{R}$.

Proof (a): It is easily checked that if $(a',b') = \pm(a,b)$, then $B = A'A^{-1}$ has the form $\pm \left(\begin{smallmatrix} 1 & 0 \\ r & 1 \end{smallmatrix}\right)$ for some $r \in \mathbb{Z}$, and $\sigma_v \circ \Phi_{A'} \sim \sigma_{\pm v} \circ \Phi_A$ if and only if $\sigma_v \circ \Phi_B \sim \sigma_{\pm v}$. The reader may verify that the intertwining operator that implements the latter relation is

$$Uf(m) = e^{\pi i r \omega m^2 \mp 2\pi i r v m} f(\pm m).$$

(This is related to the metaplectic operator $\mu_\omega(\pm B)$, which is given by $\mu_\omega(\pm B)f(t) = e^{-\pi i r \omega t^2} f(\pm t)$.)

(b): The operator $\sigma_v \circ \Phi_A(j,k,l)$ has the form

$$[\sigma_v(\Phi_A(j,k,l))f](m) = C(a,b,c,d,j,k,l,m) f(m - aj - bk)$$

where $C(a,b,c,d,j,k,l,m)$ is a complex number of absolute value 1. Thus, if $aj + bk = 0$, the standard basis $e_n(m) = \delta_{mn}$ $(n \in \mathbb{Z})$ for $l^2(\mathbb{Z})$ consists of eigenvectors for $\sigma_v \circ \Phi_A(j,k,l)$. On the other hand, if $aj + bk \neq 0$, $\sigma_v \circ \Phi_A(j,k,l)$ is a weighted shift operator with weights of absolute value 1, so it has *no* eigenvectors. (If f were an eigenvector, it would satisfy $|f(m)| = |f(m+n)|$ for every n that is a multiple of $aj+bk$, which is impossible for $f \in l^2(\mathbb{Z})$.) But by Lemma 1(a), if $(a',b') \neq \pm(a,b)$ we have $b'/a' \neq b/a$, so the equations $ax+by = 0$ and $a'x+b'y = 0$ define different lines in the xy-plane. It follows that if $aj+bk = 0$ and $(j,k) \neq (0,0)$, $\sigma_v \circ \Phi_A(j,k,l)$ has an eigenbasis and $\sigma_{v'} \circ \Phi_{A'}(j,k,l)$ does not, so these operators are not unitarily equivalent. □

It remains to make two more remarks to complete the picture. First, concerning the case $(a',b') = (-a,-b)$: we have $\sigma_{-v} \sim \sigma_{\omega - v}$, and $v \mapsto \omega - v$ is a bijection on $(0,\omega)$, so the replacement of v by $-v$ does not affect the set of equivalence classes in the direct integral $\int_{[0,\omega)}^{\oplus} \sigma_v \circ \Phi_A \, dv$. Second, if a and b are relatively prime integers, there always exist integers c and d such that $ad - bc = 1$ and hence $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL(2,\mathbb{Z})$.

In summary, *we have found an infinite family of direct integral decompositions of* ϱ_ω *involving sets of irreducible representations coming from completely disjoint sets of equivalence classes, parametrized by pairs of relatively prime integers* (a,b) *modulo the equivalence* $(-a,-b) \sim (a,b)$.

In [7], S. Kawakami constructed two families of irreducible representations of H that he denoted by $U^{(\omega,q,\lambda)}$ and $V^{(\omega,d,r)}$ $(\omega \in \mathbb{R} \setminus \mathbb{Q};\ q,d \in \mathbb{Z};\ \lambda, r \in \mathbb{R})$ and showed that ϱ_ω is equivalent to the direct integrals $\int_{[0,\omega)}^{\oplus} U^{(\omega,q,\lambda)}\, d\lambda$ and $\int_{[0,1)}^{\oplus} V^{(\omega,d,r)}\, dr$ for all $q,d \in \mathbb{Z}$. (Actually, Kawakami's formula for $V^{(\omega,d,r)}$ on [7, p. 559] needs a small

correction; as it stands, it defines an antirepresentation rather than a representation.) Kawakami derived these results from more general constructions of representations of certain non-regular semi-direct product groups in [5] and [6], which use some abstract machinery appropriate to this situation.

Straightforward calculations show that $U^{(\omega,q,\lambda)}$ and $V^{(\omega,d,r)}$ are equivalent to our representations $\sigma_\lambda \circ \Phi_{A_q}$ and $\sigma_{\omega r} \circ \Phi_{B_d}$, respectively, where

$$A_q = \begin{pmatrix} -q & -1 \\ 1 & 0 \end{pmatrix}, \qquad B_d = \begin{pmatrix} 1 & -d \\ 0 & 1 \end{pmatrix}.$$

Thus Kawakami's results are special cases of ours. Our efficient use of the $SL(2,\mathbb{Z})$ action yields a simple new derivation of them as well as an even larger family of different direct integral decompositions of ϱ_ω.

References

1. L.W. Baggett, Processing a radar signal and representations of the discrete Heisenberg group. Colloq. Math. **60/61**, 195–203 (1990)
2. G.B. Folland, *Harmonic Analysis in Phase Space* (Princeton University Press, Princeton, NJ, 1989)
3. G.B. Folland, *A Course in Abstract Harmonic Analysis*, 2nd edn. (CRC Press, Boca Raton, FL, 2015)
4. D. Gabor, Theory of communication. J. Inst. Electr. Eng. **93**(III), 429–457 (1946)
5. S. Kawakami, Irreducible representations of some non-regular semi-direct product groups. Math. Jpn. **26**, 667–693 (1981)
6. S. Kawakami, On decompositions of some factor representations. Math. Jpn. **27**, 521–534 (1982)
7. S. Kawakami, Representations of the discrete Heisenberg group. Math. Jpn. **27**, 551–564 (1982)

Fractional Differentiation: Leibniz Meets Hölder

Loukas Grafakos

Abstract We discuss how to estimate the fractional derivative of the product of two functions, not in the pointwise sense, but on Lebesgue spaces whose indices satisfy Hölder's inequality

Keywords Kato-Ponce inequality • bilinear operators • Riesz and Bessel potentials

1991 Mathematics Subject Classification. Primary 42B20. Secondary 35Axx.

1 Introduction

We recall Leibniz's product rule of differentiation

$$(fg)^{(m)} = \sum_{k=0}^{m} \binom{m}{k} f^{(m-k)} g^{(k)} \tag{1}$$

which is valid for $\mathcal{C}^m$ functions f, g on the real line. Here $g^{(k)}$ denotes the kth derivative of the function g on the line. This rule can be extended to functions of n variables. For a given multiindex $\alpha = (\alpha_1, \dots, \alpha_n) \in (\mathbf{Z}^+ \cup \{0\})^n$ we set

$$\partial^\alpha f = \partial_1^{\alpha_1} \cdots \partial_n^{\alpha_n} f = \frac{\partial^{\alpha_1}}{\partial {x_1}^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial {x_n}^{\alpha_n}} f \,.$$

The nth dimensional extension of the Leibniz rule (2) is

Grafakos acknowledges the support of the Simons Foundation.

L. Grafakos (✉)
Department of Mathematics, University of Missouri, Columbia, MO 65211, USA
e-mail: grafakosl@missouri.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4_2

$$\partial^\alpha(fg) = \sum_{\beta\le\alpha} \binom{\alpha}{\beta} (\partial^{\alpha-\beta} f)\,(\partial^\beta g) \tag{2}$$

where $\beta \le \alpha$ means $\beta_j \le \alpha_j$ for all $j = 1, \dots, n$, and

$$\binom{\alpha}{\beta} = \binom{\alpha_1}{\beta_1} \cdots \binom{\alpha_n}{\beta_n} = \prod_{j=1}^n \frac{\alpha_j!}{\beta_j!(\alpha_j-\beta_j)!}\,. \tag{3}$$

Identity (3) can be used to control the Lebesgue norm of $\partial^\alpha(fg)$ in terms of Lebesgue norms of partial derivatives of f and g via Hölder's inequality:

$$\|FG\|_{L^r(\mathbf{R}^n)} \le \|F\|_{L^p(\mathbf{R}^n)} \|G\|_{L^q(\mathbf{R}^n)}$$

where $0 < p, q, r \le \infty$ and $1/r = 1/p + 1/q$.

Unlike convolution, which captures the smoothness of its smoother input, multiplication inherits the smoothness of the rougher function. In this note we study the smoothness of the product of two functions of equal smoothness. The results we prove are quantitative and we measure smoothness in terms of Sobolev spaces. We focus on a version of (3) in which the multiindex α is replaced by a non-integer positive number, for instance a fractional power. Fractional powers are defined in terms of the Fourier transform.

We denote by $\mathcal{S}(\mathbf{R}^n)$ the space of all rapidly decreasing functions on $\mathbf{R}^n$, called Schwartz functions. The Fourier transform of an integrable function f on $\mathbf{R}^n$ (in particular of a Schwartz function) is defined by

$$\widehat{f}(\xi) = \int_{\mathbf{R}^n} f(x) e^{-2\pi i x\cdot\xi}\,dx$$

and its inverse Fourier transform is defined by

$$f^\vee(\xi) = \widehat{f}(-\xi)$$

for all $\xi \in \mathbf{R}^n$. The Laplacian of a $\mathcal{C}^2$ function f on $\mathbf{R}^n$ is defined by

$$\Delta f = \sum_{j=1}^n \partial_j^2 f$$

and this can be expressed in terms of the Fourier transform as follows:

$$\Delta f = \big(-4\pi^2|\xi|^2 \widehat{f}(\xi)\big)^\vee\,. \tag{4}$$

Identity (4) can be used to define fractional derivatives of f as follows: given $s > 0$ define

$$\Delta^{s/2} f = \left((2\pi|\xi|)^s \widehat{f}(\xi)\right)^\vee$$

and

$$J^s(f) = (1-\Delta)^{s/2}(f) = \left((1+4\pi^2|\xi|^2)^{s/2}\widehat{f}(\xi)\right)^\vee.$$

The operators $\Delta^{s/2}$ and $J^s = (1-\Delta)^{s/2}$ are called the Riesz potential and Bessel potential on $\mathbf{R}^n$, respectively. Heuristically speaking, $\Delta^{s/2} f$ is the "sth derivative" of f, while $J^s(f)$ "captures" all derivatives of all f of orders up to and including s.

Concerning J^s, in [14], Kato and Ponce obtained the commutator estimate

$$\|J^s(fg) - f(J^s g)\|_{L^p(\mathbf{R}^n)} \le C\Big[\|\nabla f\|_{L^\infty(\mathbf{R}^n)} \left\|J^{s-1} g\right\|_{L^p(\mathbf{R}^n)} + \|J^s f\|_{L^p(\mathbf{R}^n)} \|g\|_{L^\infty(\mathbf{R}^n)}\Big]$$

for $1 < p < \infty$ and $s > 0$, where ∇ is the n-dimensional gradient, f, g are Schwartz functions, and C is a constant depending on n, p, and s. This estimate was motivated by a question stated in Remark 4:1 in Kato's work [13].

Using the Riesz potential $D^s = (-\Delta)^{s/2}$, Kenig, Ponce, and Vega [16] obtained the related estimate

$$\|D^s[fg] - fD^s g - gD^s f\|_{L^r} \le C(s, s_1, s_2, r, p, q)\|D^{s_1} f\|_{L^p} \|D^{s_2} g\|_{L^q},$$

where $s = s_1 + s_2$ for $s, s_1, s_2 \in (0,1)$, and $1 < p, q, r < \infty$ such that $\frac{1}{r} = \frac{1}{p} + \frac{1}{r}$.

Instead of the original statement given by Kato and Ponce, the following variant is known in the literature as the Kato-Ponce inequality (also *fractional Leibniz rule*)

$$\|J^s(fg)\|_{L^r(\mathbf{R}^n)} \le C\Big[\|f\|_{L^{p_1}(\mathbf{R}^n)} \|J^s g\|_{L^{q_1}(\mathbf{R}^n)} + \|J^s f\|_{L^{p_2}(\mathbf{R}^n)} \|g\|_{L^{q_2}(\mathbf{R}^n)}\Big] \tag{5}$$

where $s > 0$ and $\frac{1}{r} = \frac{1}{p_1} + \frac{1}{q_1} = \frac{1}{p_2} + \frac{1}{q_2}$ for $1 < r < \infty$, $1 < p_1, q_2 \le \infty$, $1 < p_2, q_1 < \infty$ and $C = C(s, n, r, p_1, p_2, q_1, q_2)$. It is important to note that in the preceding formulation the L^∞ norm does not fall on the terms with the Bessel potential. There is an analogous Kato-Ponce version of (5) in which the Bessel potential is replaced by the Riesz potential

$$\|D^s(fg)\|_{L^r(\mathbf{R}^n)} \le C\Big[\|f\|_{L^{p_1}(\mathbf{R}^n)} \|D^s g\|_{L^{q_1}(\mathbf{R}^n)} + \|D^s f\|_{L^{p_2}(\mathbf{R}^n)} \|g\|_{L^{q_2}(\mathbf{R}^n)}\Big] \tag{6}$$

and the indices are as before. In this note we study (5) and (6) focusing on (6).

There are further generalizations of the aforementioned Kato-Ponce inequalities. For instance, Muscalu, Pipher, Tao, and Thiele [18] extended this inequality to allow for partial fractional derivatives in $\mathbf{R}^2$. Bernicot, Maldonado, Moen, and Naibo [1] proved the Kato-Ponce inequality in weighted Lebesgue spaces under certain restrictions on the weights. The last authors also extended the Kato-Ponce inequality to indices $r < 1$ under the assumption $s > n$. Additional work on the Kato-Ponce inequality was done by Christ and Weinstein [3], Gulisashvili and Kon [12], and

Cordero and Zucco [6], present a way to obtain the homogeneous inequality from the inhomogeneous via a limiting process. There is also a discussion about the Kato-Ponce inequality in [8].

When $r \geq 1$ inequality (5) is valid for all $s \geq 0$ but when $r < 1$ the author and Oh [9] showed that there is a restriction $s > n/r - n$. Moreover, there is an example that inequality (5) fails for $s \leq n/r - n$. This restriction in the case $r < 1$ was independently obtained by Muscalu and Schlag [17].

2 The counterexample

In this section we provide an example to determine the range of $r < 1$ for which $D^s(fg)$ can lie in $L^r(\mathbf{R}^n)$ and in particular (6) can hold.

We set

$$f(x) = g(x) = e^{-\pi|x|^2/2}.$$

Then

$$fg(x) = e^{-\pi|x|^2}$$

and for $s > 0$ we have

$$(-\Delta)^{s/2}(fg)(x) = (2\pi)^s \int_{\mathbf{R}^n} \big(e^{-\pi|x|^2}\big)^\wedge(\xi)|\xi|^s e^{2\pi i x\xi} dx = (2\pi)^s \int_{\mathbf{R}^n} e^{-\pi|\xi|^2}|\xi|^s e^{2\pi i x\xi} dx.$$

We now consider two cases: (a) s is an even integer. In this case we have that $(-\Delta)^{s/2}(fg)$ is a Schwartz function and thus it has fast decay at infinity. (b) s is not an even integer. In the second case $(-\Delta)^{s/2}(fg)$ is given by multiplication on the Fourier transform side by $c\,|\xi|^s$ and thus it is given by convolving $e^{-\pi|x|^2}$ with the distribution

$$W_s = c\,\frac{\pi^{-\frac{s}{2}}}{\Gamma(-\frac{s}{2})}\,\frac{\Gamma(\frac{s+n}{2})}{\pi^{\frac{s+n}{2}}}|x|^{-n-s}.$$

Notice that $e^{-\pi|\cdot|^2} * W_s$ is the convolution of a Schwartz function with a tempered distribution and thus it is a smooth function with at most polynomial growth at infinity [7]. Thus $e^{-\pi|\cdot|^2} * W_s$ is a smooth and bounded function on any compact set.

Next we study the decay of this function as $|x| \to \infty$. To do so, we introduce a nonnegative function ϕ with support contained in $|x| \leq 2$ and equal to 1 on the ball $|x| \leq 1$. Then we have

$$e^{-\pi|\cdot|^2} * W_s = e^{-\pi|\cdot|^2} * \phi W_s + e^{-\pi|\cdot|^2} * (1-\phi)W_s.$$

First we notice that for $|x| \geq 10$ we have

$$\begin{aligned}
\left|(1-\phi)W_s * e^{-\pi|\cdot|^2}(x)\right| &= \left|c_{s,n}\int_{\mathbf{R}^n}(1-\phi(y))|y|^{-n-s}e^{-\pi|x-y|^2}dy\right| \\
&\geq |c_{s,n}|\int_{|y|\geq 1}|y|^{-n-s}e^{-\pi|x-y|^2}dy \\
&\geq |c_{s,n}|\int_{\substack{|y|\geq 1\\ |x-y|\leq 1}}|y|^{-n-s}e^{-\pi|x-y|^2}dy \\
&\geq |c_{s,n}|\, e^{-\pi}\int_{\substack{|y|\geq 1\\ |x-y|\leq 1}}|y|^{-n-s}dy \\
&\geq c'\,|x|^{-n-s}\,.
\end{aligned}$$

As for the other term, for $|x| \geq 10$ and for $N = [s] + 1$ we have

$$\begin{aligned}
&\int_{|y|\leq 2} e^{-\pi|x-y|^2}\phi(y)\frac{|y|^{-s-n}}{\Gamma(-\frac{s}{2})}\,dy \\
&= \int_{|y|\leq 2}\left[e^{-\pi|x-y|^2} - \sum_{|\gamma|\leq N}\frac{\partial^\gamma(e^{-\pi|\cdot|^2})}{\gamma!}(x)y^\gamma\right]\phi(y)\frac{|y|^{-s-n}}{\Gamma(-\frac{s}{2})}\,dy \\
&\qquad + \sum_{|\gamma|\leq N}\frac{\partial^\gamma(e^{-\pi|\cdot|^2})}{\gamma!}(x)\int_{|y|\leq 2}\phi(y)y^\gamma\frac{|y|^{-s-n}}{\Gamma(-\frac{s}{2})}\,dy \\
&= \sum_{|\gamma|=N+1}\int_{|y|\leq 2}\frac{\partial^\gamma(e^{-\pi|\cdot|^2})}{\gamma!}(x-\theta_y y)\;\phi(y)\frac{y^\gamma\,|y|^{-s-n}}{\Gamma(-\frac{s}{2})}\,dy \\
&\qquad + \sum_{|\gamma|\leq N}\frac{\partial^\gamma(e^{-\pi|\cdot|^2})}{\gamma!}(x)\int_{|y|\leq 2}\phi(y)y^\gamma\frac{|y|^{-s-n}}{\Gamma(-\frac{s}{2})}\,dy\,,
\end{aligned}$$

where $\theta_y \in [0,1]$. Notice that

$$\partial^\gamma(e^{-\pi|\cdot|^2})(x) = P_\gamma(x)e^{-\pi|x|^2},$$

where P_γ is a polynomial of n variables depending on γ. Also $|x-\theta_y y| \geq \frac{1}{2}|x|$, hence $\partial^\gamma(e^{-\pi|\cdot|^2})(x-\theta_y y)$ has exponential decay at infinity, while the integral

$$\int_{|y|\leq 2}\phi(y)y^\gamma\frac{|y|^{-s-n}}{\Gamma(-\frac{s}{2})}\,dy$$

is absolutely convergent when $|\gamma| = N + 1 = [s] + 2 > s + 1$ which is equivalent to $N - s - n > -n$. Moreover, for $|\gamma| \le [s] + 1$, the quantities

$$\int_{|y|\le 2} \phi(y) y^\gamma \frac{|y|^{-s-n}}{\Gamma(-\frac{s}{2})}\, dy = \int_{|y|\le 1} y^\gamma \frac{|y|^{-s-n}}{\Gamma(-\frac{s}{2})}\, dy + \int_{1\le|y|\le 2} \phi(y) y^\gamma \frac{|y|^{-s-n}}{\Gamma(-\frac{s}{2})}\, dy$$

are finite constants in view of the following observation: If $|\gamma|$ is odd, then the displayed expression below is zero, while if $|\gamma| = 2m$ is an even integer, we have

$$\int_{|y|\le 1} y^\gamma \frac{|y|^{-s-n}}{\Gamma(-\frac{s}{2})}\, dy = \Big(\int_{\mathbf{S}^{n-1}} \varphi^\gamma \, d\varphi \Big) \frac{\int_0^1 r^{|\gamma|-s-1} dr}{\Gamma(-\frac{s}{2})} = \Big(\int_{\mathbf{S}^{n-1}} \varphi^\gamma \, d\varphi \Big) \frac{1/2}{(m-\frac{s}{2})\Gamma(-\frac{s}{2})}$$

and this is a well-defined constant, since the entire function $\Gamma(-w)^{-1}$ has simple zeros at the positive integers and thus $(m-w)^{-1}\Gamma(-w)^{-1}$ is also entire in w for any $m < |w|$. An important observation here is that $w = s/2$ is not an integer, since s is not an even integer, hence m can never be equal to $w = s/2$.

The outcome of this discussion is that $e^{-\pi|\cdot|^2} * \phi W_s$ is a smooth function that decays exponentially as $|x| \to \infty$. Combining this result with the corresponding obtained for $e^{-\pi|\cdot|^2} * (1-\phi) W_s$, we deduce that

$$\big| (e^{-\pi|\cdot|^2} * W_s)(x) \big| \ge c'' |x|^{-n-s}$$

as $|x| \to \infty$, provided s is not an even integer.

Finally, this assertion is not valid if s is an even integer, since in this case we have that

$$(-\Delta)^{s/2}(e^{-\pi|\cdot|^2}) = \underbrace{\Big(\sum_{j=1}^n \partial_j^2\Big) \cdots \Big(\sum_{j=1}^n \partial_j^2\Big)}_{s/2 \text{ times}} (e^{-\pi|\cdot|^2})$$

has obviously exponential decay at infinity, as obtained by a direct differentiation.

The preceding calculation imposes a restriction on the p for which $(-\Delta)^{s/2}(fg)$ lies in $L^p(\mathbf{R}^n)$. In fact the simple example $f(x) = g(x) = e^{-\pi|x|^2/2}$ introduced at the beginning of this section provides a situation in which f, g, $(-\Delta)^{s/2}(f)$, and $(-\Delta)^{s/2}(g)$ lie in all the L^{p_j} spaces for $p_j > 1$ when $s > 0$, but $(-\Delta)^{s/2}(fg)$ lies in $L^r(\mathbf{R}^n)$ only if

$$(-s-n)r < -n,$$

that is, when $r > \frac{n}{n+s}$. Thus, when $1 \le p_1, p_2, q_1, q_2 \le \infty$ and

$$\frac{1}{r} = \frac{1}{p_1} + \frac{1}{q_1} = \frac{1}{p_2} + \frac{1}{q_2}$$

the inequality (6) fails when

$$r \leq \frac{n}{n+s}.$$

Obviously, since $r > 1/2$, so this restriction is relevant only when $0 < s \leq n$. Recall that Bernicot, Maldonado, Moen, and Naibo [1] showed that (6) holds when $s > n$, so this work fills in the gap $0 < s < n$.

3 The sharp Kato-Ponce inequalities and preliminaries

The following theorem is contained in the joint article of the author with Seungly Oh [9]. In this section we discuss some preliminary facts needed to prove the first inequality below.

Theorem 1 *Let* $\frac{1}{2} < r < \infty$, $1 < p_1, p_2, q_1, q_2 \leq \infty$ *satisfy* $\frac{1}{r} = \frac{1}{p_1} + \frac{1}{q_1} = \frac{1}{p_2} + \frac{1}{q_2}$. *Given* $s > \max\left(0, \frac{n}{r} - n\right)$ *or* $s \in 2\mathbf{N}$, *there exists* $C = C(n, s, r, p_1, q_1, p_2, q_2) < \infty$ *such that for all* $f, g \in \mathcal{S}(\mathbf{R}^n)$ *we have*

$$\|D^s(fg)\|_{L^r(\mathbf{R}^n)} \leq C\left[\|D^s f\|_{L^{p_1}(\mathbf{R}^n)}\|g\|_{L^{q_1}(\mathbf{R}^n)} + \|f\|_{L^{p_2}(\mathbf{R}^n)}\|D^s g\|_{L^{q_2}(\mathbf{R}^n)}\right], \tag{7}$$

$$\|J^s(fg)\|_{L^r(\mathbf{R}^n)} \leq C\left[\|f\|_{L^{p_1}(\mathbf{R}^n)}\|J^s g\|_{L^{q_1}(\mathbf{R}^n)} + \|J^s f\|_{L^{p_2}(\mathbf{R}^n)}\|g\|_{L^{q_2}(\mathbf{R}^n)}\right]. \tag{8}$$

Moreover if $r < 1$ *and any one of the indices* p_1, p_2, q_1, q_2 *is equal to* 1, *then* (7) *and* (8) *hold when the* $L^r(\mathbf{R}^n)$ *norms on the left-hand side of the inequalities are replaced by the* $L^{r,\infty}(\mathbf{R}^n)$ *quasi-norm.*

We remark that the statement above does not include the endpoints $L^1 \times L^\infty \to L^{1,\infty}$, $L^\infty \times L^1 \to L^{1,\infty}$, and $L^\infty \times L^\infty \to L^\infty$. However, the endpoint case $p_1 = p_2 = q_1 = q_2 = r = \infty$ was completed by Bourgain and Li [2]. An earlier version of this endpoint inequality was obtained by Grafakos, Maldonado, and Naibo [11].

As a consequence of (8) we obtain Hölder's inequality for Sobolev spaces. We have

Corollary 1 *Let* $s > 0$, $\frac{n}{s+n} < r < \infty$, $1 < p, q < \infty$ *satisfy* $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$. *Then there exists* $C = C(n, s, p, q) < \infty$ *such that for all* $f, g \in \mathcal{S}(\mathbf{R}^n)$ *we have*

$$\|J^s(fg)\|_{L^r(\mathbf{R}^n)} \leq C\|J^s f\|_{L^p(\mathbf{R}^n)}\|J^s g\|_{L^q(\mathbf{R}^n)}. \tag{9}$$

We note that (9) is an easy consequence of (8), since

$$\|f\|_{L^p} \leq C\|J^s f\|_{L^p}$$

for $1 < p < \infty$.

In the rest of this section we discuss some background material needed to prove Theorem 1. First we recall the classical multiplier result of Coifman and Meyer [4] (see also [5]) for the case $r > 1$ and its extension to the case $r \leq 1$ by Grafakos and Torres [10] and independently by Kenig and Stein [15].

Theorem A *Let $m \in L^\infty(\mathbf{R}^{2n})$ be smooth away from the origin. Suppose that there exists a constant $A > 0$ satisfying*

$$|\partial_\xi^\alpha \partial_\eta^\beta m(\xi,\eta)| \leq A\,(|\xi|+|\eta|)^{-|\alpha|-|\beta|} \tag{10}$$

for all $\xi, \eta \in \mathbf{R}^n$ with $|\xi|+|\eta| \neq 0$ and $\alpha, \beta \in \mathbf{Z}^n$ multi-indices with $|\alpha|, |\beta| \leq 2n+1$. Then the bilinear operator

$$T_m(f,g)(x) = \int_{\mathbf{R}^{2n}} m(\xi,\eta)\widehat{f}(\xi)\widehat{g}(\eta) e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta\,,$$

defined for all $f, g \in \mathcal{S}(\mathbf{R}^n)$, satisfies

$$\|T_m(f,g)\|_{L^r(\mathbf{R}^n)} \leq C(p,q,r,A)\,\|f\|_{L^p(\mathbf{R}^n)}\|g\|_{L^q(\mathbf{R}^n)}$$

where $\frac{1}{2} < r < \infty$, $1 < p, q \leq \infty$ and $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$. Furthermore, when either p or q is equal to 1, then the $L^r(\mathbf{R}^n)$ norm on left-hand side can be replaced by the $L^{r,\infty}(\mathbf{R}^n)$ norm.

In the sequel we will use the notation $\Psi_t(x) = t^{-n}\Psi(x/t)$ when $t > 0$ and $x \in \mathbf{R}^n$. The following result will be of use to us.

Theorem 2 *Let $m \in \mathbf{Z}^n \setminus \{0\}$ and $\Psi(x) = \psi(x+m)$ for some Schwartz function ψ whose Fourier transform is supported in the annulus $1/2 \leq |\xi| \leq 2$. Let $\Delta_j(f) = \Psi_{2^{-j}} * f$. Then there is a constant C_n such that for $1 < p < \infty$ we have*

$$\Big\| \Big(\sum_{j\in\mathbf{Z}} |\Delta_j(f)|^2 \Big)^{\frac{1}{2}} \Big\|_{L^p(\mathbf{R}^n)} \leq C_n \ln(1+|m|) \max\big(p, (p-1)^{-1}\big) \|f\|_{L^p(\mathbf{R}^n)}. \tag{11}$$

There also exists $C_n < \infty$ such that for all $f \in L^1(\mathbf{R}^n)$,

$$\Big\| \Big(\sum_{j\in\mathbf{Z}} |\Delta_j(f)|^2 \Big)^{\frac{1}{2}} \Big\|_{L^{1,\infty}(\mathbf{R}^n)} \leq C_n \ln(1+|m|) \|f\|_{L^1(\mathbf{R}^n)}. \tag{12}$$

Proof We recall the following form of the Littlewood-Paley theorem: Suppose that Ψ is an integrable function on $\mathbf{R}^n$ that satisfies

$$\sum_{j\in\mathbf{Z}} |\widehat{\Psi}(2^{-j}\xi)|^2 \leq B^2 \tag{13}$$

and

$$\sup_{y\in\mathbf{R}^n\setminus\{0\}}\sum_{j\in\mathbf{Z}}\int_{|x|\geq 2|y|}\left|\Psi_{2^{-j}}(x-y)-\Psi_{2^{-j}}(x)\right|dx\leq B \tag{14}$$

Then there exists a constant $C_n<\infty$ such that for all $1<p<\infty$ and all f in $L^p(\mathbf{R}^n)$,

$$\left\|\Big(\sum_{j\in\mathbf{Z}}|\Delta_j(f)|^2\Big)^{\frac{1}{2}}\right\|_{L^p(\mathbf{R}^n)}\leq C_nB\max\left(p,(p-1)^{-1}\right)\|f\|_{L^p(\mathbf{R}^n)} \tag{15}$$

where $\Delta_j(f)=\Psi_{2^{-j}}*f$. There also exists a $C_n'<\infty$ such that for all f in $L^1(\mathbf{R}^n)$,

$$\left\|\Big(\sum_{j\in\mathbf{Z}}|\Delta_j(f)|^2\Big)^{\frac{1}{2}}\right\|_{L^{1,\infty}(\mathbf{R}^n)}\leq C_n'B\|f\|_{L^1(\mathbf{R}^n)}. \tag{16}$$

We make a few remarks about the proof. Clearly the required estimate holds when $p=2$ in view of (13). To obtain estimate (16) and thus the case $p\neq 2$, we define an operator $\vec{T}$ acting on functions on $\mathbf{R}^n$ as follows:

$$\vec{T}(f)(x)=\{\Delta_j(f)(x)\}_j\,.$$

The inequalities (15) and (16) we wish to prove say simply that $\vec{T}$ is a bounded operator from $L^p(\mathbf{R}^n,\mathbf{C})$ to $L^p(\mathbf{R}^n,\ell^2)$ and from $L^1(\mathbf{R}^n,\mathbf{C})$ to $L^{1,\infty}(\mathbf{R}^n,\ell^2)$. We indicated that this statement is true when $p=2$, and therefore the first hypothesis of Theorem 4.6.1 in [7] is satisfied. We now observe that the operator $\vec{T}$ can be written in the form

$$\vec{T}(f)(x)=\left\{\int_{\mathbf{R}^n}\Psi_{2^{-j}}(x-y)f(y)\,dy\right\}_j=\int_{\mathbf{R}^n}\vec{K}(x-y)(f(y))\,dy,$$

where for each $x\in\mathbf{R}^n$, $\vec{K}(x)$ is a bounded linear operator from $\mathbf{C}$ to ℓ^2 given by

$$\vec{K}(x)(a)=\{\Psi_{2^{-j}}(x)a\}_j. \tag{17}$$

We clearly have that $\left\|\vec{K}(x)\right\|_{\mathbf{C}\to\ell^2}=\left(\sum_j|\Psi_{2^{-j}}(x)|^2\right)^{\frac{1}{2}}$, and to be able to apply Theorem 4.6.1 in [7] we need to know that

$$\int_{|x|\geq 2|y|}\left\|\vec{K}(x-y)-\vec{K}(x)\right\|_{\mathbf{C}\to\ell^2}dx\leq C_nB,\qquad y\neq 0. \tag{18}$$

We clearly have

$$\begin{aligned}\big\|\vec{K}(x-y)-\vec{K}(x)\big\|_{\mathbf{C}\to\ell^2} &= \bigg(\sum_{j\in\mathbf{Z}}\big|\Psi_{2^{-j}}(x-y)-\Psi_{2^{-j}}(x)\big|^2\bigg)^{\frac12}\\ &\le \sum_{j\in\mathbf{Z}}\big|\Psi_{2^{-j}}(x-y)-\Psi_{2^{-j}}(x)\big|\end{aligned}$$

and so condition (14) implies (18).

Note

$$\widehat{\Psi}(\xi)=\widehat{\psi}(\xi)e^{2\pi i m\cdot\xi}.$$

The fact that $\widehat{\psi}$ is supported in the annulus $1/2\le|\xi|\le 2$ implies condition (13) for $\widehat{\Psi}$. We now focus on condition (14) for Ψ.

We fix a nonzero y in $\mathbf{R}^n$ and $j\in\mathbf{Z}$. We look at

$$\int_{|x|\ge 2|y|}\big|\Psi_{2^{-j}}(x-y)-\Psi_{2^{-j}}(x)\big|dx=\int_{|x|\ge 2|y|}2^{jn}\big|\psi(2^jx-2^jy+m)-\psi(2^jx+m)\big|dx$$

Changing variables we can write the above as

$$I_j=\int_{|x|\ge 2|y|}\big|\Psi_{2^{-j}}(x-y)-\Psi_{2^{-j}}(x)\big|dx=\int_{|x-m|\ge 2^{j+1}|y|}\big|\psi(x-2^jy)-\psi(x)\big|dx$$

Case 1: $2^j\ge 2\,|m|\,|y|^{-1}$. In this case we estimate I_j by

$$\begin{aligned}&\int_{|x-m|\ge 2^{j+1}|y|}\frac{c}{(1+|x-2^jy|)^{n+2}}dx+\int_{|x-m|\ge 2^{j+1}|y|}\frac{c}{(1+|x|)^{n+2}}\,dx\\ &=\int_{|x+2^jy-m|\ge 2^{j+1}|y|}\frac{c}{(1+|x|)^{n+2}}dx+\int_{|x-m|\ge 2^{j+1}|y|}\frac{c}{(1+|x|)^{n+2}}\,dx\end{aligned}$$

Suppose that x lies in the domain of integration of the first integral. Then

$$|x|\ge|x+2^jy-m|-2^j|y|-|m|\ge 2^{j+1}|y|-2^j|y|-\frac12\,2^j|y|=\frac12\,2^j|y|.$$

If x lies in the domain of integration of the second integral, then

$$|x|\ge|x-m|-|m|\ge 2^{j+1}|y|-|m|\ge 2^{j+1}|y|-\frac12\,2^j|y|=\frac32\,2^j|y|.$$

In both cases we have

$$I_j \le 2 \int_{|x| \ge \frac{1}{2}\, 2^j |y|} \frac{c}{(1+|x|)^{n+2}}\, dx \le \frac{C}{2^j |y|} \int_{\mathbf{R}^n} \frac{1}{(1+|x|)^{n+1}}\, dx \le \frac{C_n}{2^j |y|},$$

and clearly

$$\sum_{j:\ 2^j |y| \ge 2|m|} I_j \le \sum_{j:\ 2^j |y| \ge 2} I_j \le C_n.$$

Case 2: $|y|^{-1} \le 2^j \le 2\,|m|\,|y|^{-1}$. The number of j's in this case are $O(\ln |m|)$. Thus, uniformly bounding I_j by a constant, we obtain

$$\sum_{j: 1 \le 2^j |y| \le 2|m|} I_j \le C_n (1 + \ln |m|).$$

Case 3. $2^j \le |y|^{-1}$. In this case we have

$$\left|\psi(x - 2^j y) - \psi(x)\right| = \left| \int_0^1 2^j \nabla \psi (x - 2^j t y) \cdot y\, dt \right| \le 2^j |y| \int_0^1 \frac{c}{(1 + |x - 2^j t y|)^{n+1}}\, dt\,.$$

Integrating over $x \in \mathbf{R}^n$ gives the bound $I_j \le C_n\, 2^j |y|$. Thus, we obtain

$$\sum_{j: 2^j |y| \le 1} I_j \le C_n.$$

Overall, we obtain the bound $C_n \ln(1 + |m|)$ for (14), which yields the desired statement using the Littlewood-Paley theorem.

4 The proof of the homogeneous inequality (7)

In this section we prove the homogeneous inequality (7) of Theorem 1.

Proof We fix a function $\widehat{\Phi} \in \mathcal{S}(\mathbf{R}^n)$ such that $\widehat{\Phi}(\xi) \equiv 1$ on $|\xi| \le 1$ and which is supported in $|\xi| \le 2$. We define another function

$$\widehat{\Psi}(\xi) = \widehat{\Phi}(\xi) - \widehat{\Phi}(2\xi)$$

and we note that $\widehat{\Psi}$ is supported on the annulus $\{\xi :\ 1/2 < |\xi| < 2\}$ and satisfies

$$\sum_{k \in \mathbf{Z}} \widehat{\Psi}(2^{-k} \xi) = 1$$

for all $\xi \ne 0$.

Given $f, g \in \mathcal{S}(\mathbf{R}^n)$, we decompose $D^s[fg]$ as follows:

$$\begin{aligned} D^s[fg](x) &= \int_{\mathbf{R}^n}\int_{\mathbf{R}^n} |\xi+\eta|^s \widehat{f}(\xi)\widehat{g}(\eta) e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta \\ &= \int_{\mathbf{R}^n}\int_{\mathbf{R}^n} |\xi+\eta|^s \left(\sum_{j\in\mathbf{Z}} \widehat{\Psi}(2^{-j}\xi)\widehat{f}(\xi)\right)\left(\sum_{k\in\mathbf{Z}} \widehat{\Psi}(2^{-k}\eta)\widehat{g}(\eta)\right) e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta \\ &= \Pi_1[f,g](x) + \Pi_2[f,g](x) + \Pi_3[f,g](x)\,, \end{aligned}$$

where

$$\Pi_1[f,g](x) = \sum_{j\in\mathbf{Z}} \sum_{k:k\le j-2} \int_{\mathbf{R}^n}\int_{\mathbf{R}^n} |\xi+\eta|^s \widehat{\Psi}(2^{-j}\xi)\widehat{f}(\xi)\widehat{\Psi}(2^{-k}\eta)\widehat{g}(\eta) e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta$$

$$\Pi_2[f,g](x) = \sum_{k\in\mathbf{Z}} \sum_{j:j\le k-2} \int_{\mathbf{R}^n}\int_{\mathbf{R}^n} |\xi+\eta|^s \widehat{\Psi}(2^{-j}\xi)\widehat{f}(\xi)\widehat{\Psi}(2^{-k}\eta)\widehat{g}(\eta) e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta$$

$$\Pi_3[f,g](x) = \sum_{k\in\mathbf{Z}} \sum_{j:|j-k|\le 1} \int_{\mathbf{R}^n}\int_{\mathbf{R}^n} |\xi+\eta|^s \widehat{\Psi}(2^{-j}\xi)\widehat{f}(\xi)\widehat{\Psi}(2^{-k}\eta)\widehat{g}(\eta) e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta\,.$$

For Π_1, we can write

$$\Pi_1[f,g](x) = \int_{\mathbf{R}^n}\int_{\mathbf{R}^n} \left\{\sum_{j\in\mathbf{Z}} \widehat{\Psi}(2^{-j}\xi)\widehat{\Phi}(2^{-j+2}\eta)\frac{|\xi+\eta|^s}{|\xi|^s}\right\} \widehat{D^s f}(\xi)\widehat{g}(\eta) e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta.$$

In Π_1 the variable ξ dominates η and so the ratio $\frac{|\xi+\eta|^s}{|\xi|^s}$ vanishes only at the origin in $\mathbf{R}^{2n}$. It is easy to verify that the expression in the square bracket above is a bilinear Coifman-Meyer multiplier, hence Theorem A implies that $\Pi_1[f,g]$ satisfies the inequality

$$\|\Pi_1[f,g]\|_{L^r} \le C\|D^s f\|_{L^{p_1}} \|g\|_{L^{q_1}}$$

and thus (7) holds for this term. The argument for Π_2 is identical under the apparent symmetry and one obtains

$$\|\Pi_2[f,g]\|_{L^r} \le C\|f\|_{L^{p_2}} \|D^s g\|_{L^{q_2}}\,.$$

For $\Pi_3[f,g]$, note that the summation in j is finite, we may only focus on one term, say $j=k$ and in this case it suffices to show estimate (7) for the term

$$\left\| \sum_{k\in\mathbf{Z}} \int_{\mathbf{R}^n}\int_{\mathbf{R}^n} |\xi+\eta|^s \widehat{\Psi}(2^{-k}\xi)\widehat{f}(\xi)\widehat{\Psi}(2^{-k}\eta)\widehat{g}(\eta) e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta \right\|_{L^r(\mathbf{R}^n)}. \tag{19}$$

When $s \in 2\mathbf{N}$, (19) can be written as

$$\left\| \int_{\mathbf{R}^n} \int_{\mathbf{R}^n} \left\{ \sum_{k\in\mathbf{Z}} \frac{|\xi+\eta|^s}{|\eta|^s} \widehat{\Psi}(2^{-k}\xi)\widehat{\Psi}(2^{-k}\eta) \right\} \widehat{f}(\xi)\widehat{D^s g}(\eta) e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta \right\|_{L^r(\mathbf{R}^n)}.$$

The expression in the bracket above belongs to Coifman-Meyer class, i.e., it satisfies (10), so the claimed inequality is a consequence of Theorem A in this case. When s is not an even integer we argue differently. In this case, the estimate for Π_3 requires a more careful analysis. We consider the following cases:

Case 1: $\frac{1}{2} < r < \infty$, $1 < p, q < \infty$ or $\frac{1}{2} \le r < 1$, $1 \le p, q < \infty$.

In this case, we may have the strong L^r norm on the left-hand side of (7) when $p, q > 1$ or the weak L^r norm instead when either p or q is equal to 1. In view of Theorem A and Theorem 2, the strategy for the proof in both of these subcases will be identical. For simplicity, we will only prove the estimate with a strong L^r norm on the left-hand side.

Notice that when $|\xi|, |\eta| \le 2 \cdot 2^k$, then $|\xi+\eta| \le 2^{k+2}$ and thus

$$\widehat{\Phi}(2^{-k-2}(\xi+\eta)) = 1$$

on the support of the integral giving Π_3. In view of this we may write

$$\begin{aligned}
&\Pi_3[f,g](x)\\
&= \int_{\mathbf{R}^n}\int_{\mathbf{R}^n} \sum_{k\in\mathbf{Z}} |\xi+\eta|^s \widehat{\Psi}(2^{-k}\xi)\widehat{f}(\xi)\widehat{\Psi}(2^{-k}\eta)\widehat{g}(\eta) e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta\\
&= \int_{\mathbf{R}^n}\int_{\mathbf{R}^n} \sum_{k\in\mathbf{Z}} |\xi+\eta|^s \widehat{\Phi}(2^{-k-2}(\xi+\eta))\widehat{\Psi}(2^{-k}\xi)\widehat{f}(\xi)\widehat{\Psi}(2^{-k}\eta)\widehat{g}(\eta) e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta\\
&= 2^{2s}\sum_{k\in\mathbf{Z}} \int_{\mathbf{R}^n}\int_{\mathbf{R}^n} \widehat{\Phi_s}(2^{-k-2}(\xi+\eta))\widehat{\Psi}(2^{-k}\xi)\widehat{f}(\xi)\widetilde{\Psi}(2^{-k}\eta)\widehat{D^s g}(\eta) e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta\\
&= 2^{2s}\sum_{k\in\mathbf{Z}} 2^{2nk} \int_{\mathbf{R}^n}\int_{\mathbf{R}^n} \widehat{\Phi_s}(2^{-2}(\xi+\eta))\widehat{\Psi}(\xi)\widehat{f}(2^k\xi)\widehat{\widetilde{\Psi}}(\eta)\widehat{D^s g}(2^k\eta) e^{2\pi i 2^k(\xi+\eta)\cdot x}\, d\xi\, d\eta,
\end{aligned}$$

where

$$\widehat{\widetilde{\Psi}}(\xi) = |\xi|^{-s}\widehat{\Psi}(\xi)$$

$$\widehat{\Phi_s}(\xi) = |\xi|^s \widehat{\Phi}(\xi)$$

Now the function $\xi \mapsto \widehat{\Phi_s}(2^{-2}\xi)$ is supported in $[-8,8]^n$ and can be expressed in terms of its Fourier series multiplied by the characteristic function of the set $[-8,8]^n$, denoted $\chi_{[-8,8]^n}$.

$$\widehat{\Phi_s}(2^{-2}(\xi+\eta)) = \sum_{m\in\mathbf{Z}^n} c_m^s e^{\frac{2\pi i}{16}(\xi+\eta)\cdot m}\chi_{[-8,8]^n}(\xi+\eta),$$

where

$$c_m^s = \frac{1}{16^n}\int_{[-8,8]^n} |y|^s \widehat{\Phi}(2^{-2}y)e^{-\frac{2\pi i}{16}y\cdot m}\,dy\,.$$

It is an easy calculation that

$$c_m^s = O((1+|m|)^{-s-n}) \tag{20}$$

as $|m|\to\infty$ and c_m^s is uniformly bounded for all $m\in\mathbf{Z}$. This calculation is similar to the one in Section 2.

Due to the support of $\widehat{\Psi}$ and $\widehat{\widetilde{\Psi}}$, we also have

$$\chi_{[-8,8]^n}(\xi+\eta)\widehat{\Psi}(\xi)\widehat{\widetilde{\Psi}}(\eta) = \widehat{\Psi}(\xi)\widehat{\widetilde{\Psi}}(\eta),$$

so that the characteristic function may be omitted from the integrand. Using this identity, we write $\Pi_3[f,g](x)$ as

$$= 2^{2s}\sum_{k\in\mathbf{Z}} 2^{2nk}\int_{\mathbf{R}^n}\int_{\mathbf{R}^n}\sum_{m\in\mathbf{Z}^n} c_m^s e^{\frac{2\pi i}{16}(\xi+\eta)\cdot m}\widehat{\Psi}(\xi)\widehat{f}(2^k\xi)\widehat{\widetilde{\Psi}}(\eta)\widehat{D^s g}(2^k\eta)e^{2\pi i 2^k(\xi+\eta)\cdot x}\,d\xi\,d\eta$$

$$= 2^{2s}\sum_{m\in\mathbf{Z}^n} c_m^s \sum_{k\in\mathbf{Z}} \Delta_k^m(f)(x)\widetilde{\Delta_k^m}(D^s g)(x),$$

where Δ_k^m is the Littlewood-Paley operator given by multiplication on the Fourier transform side by $e^{2\pi i 2^{-k}\xi\cdot\frac{m}{16}}\widehat{\Psi}(2^{-k}\xi)$, while $\widetilde{\Delta_k^m}$ is the Littlewood-Paley operator given by multiplication on the Fourier side by $e^{2\pi i 2^{-k}\xi\cdot\frac{m}{16}}\widehat{\widetilde{\Psi}}(2^{-k}\xi)$. Both Littlewood-Paley operators have the form:

$$\int_{\mathbf{R}^n} 2^{nk}\Theta(2^k(x-y)+\tfrac{1}{16}m)f(y)\,dy$$

for some Schwartz function Θ whose Fourier transform is supported in some annulus centered at zero.

Let $r_* = \min(r,1)$. Taking the L^r norm of the right-hand side above, we obtain

$$\|D^s[fg]\|_{L^r}^{r_*} \le \sum_{m\in\mathbf{Z}^n} |c_m^s|^{r_*}\left\|\sum_{k\in\mathbf{Z}}\Delta_k^m(f)(x)\widetilde{\Delta_k^m}D^s(g)(x)\right\|_{L^r(\mathbf{R}^n)}^{r_*}$$

$$\leq \sum_{m\in\mathbf{Z}^n} |c_m^s|^{r_*} \left\| \sqrt{\sum_{k\in\mathbf{Z}} |\Delta_k^m(f)|^2} \right\|_{L^p(\mathbf{R}^n)}^{r_*} \left\| \sqrt{\sum_{k\in\mathbf{Z}} |\widetilde{\Delta_k^m}(D^s g)|^2} \right\|_{L^q(\mathbf{R}^n)}^{r_*}$$

whenever $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$. By Theorem 2, the preceding expression is bounded by a constant multiple of

$$\sum_{m\in\mathbf{Z}^n} |c_m^s|^{r_*} [\ln(2+|m|)]^{2r_*} \|f\|_{L^p}^{r_*} \|D^s g\|_{L^q}^{r_*}$$

if $1 < p, q < \infty$ and the preceding series converges in view of (20) and of the assumption $r_*(n+s) > n$. This concludes Case 1.

Case 2: $1 < r < \infty$, $(p, q) \in \{(r, \infty), (\infty, r)\}$
In this case we use an argument inspired by Christ and Weinstein [3]. We have

$$\|\Pi_3[f,g]\|_{L^r(\mathbf{R}^n)} \leq C(r,n) \left\| \left(\sum_{j\in\mathbf{Z}} |\Delta_j(\Pi_3[f,g])|^2 \right)^{\frac{1}{2}} \right\|_{L^r(\mathbf{R}^n)}.$$

The summand in j above can be estimated as follows:

$$\begin{aligned}
&\Delta_j \Pi_3[f,g](x) \\
&= \int_{\mathbf{R}^n}\int_{\mathbf{R}^n} |\xi+\eta|^s \widehat{\Psi}(2^{-j}(\xi+\eta)) \sum_{k\geq j-2} \widehat{\Psi}(2^{-k}\xi)\widehat{f}(\xi)\widehat{\Psi}(2^{-k}\eta)\widehat{g}(\eta)\, e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta \\
&= \int_{\mathbf{R}^n}\int_{\mathbf{R}^n} 2^{js}\widehat{\widetilde{\Psi}_s}(2^{-j}(\xi+\eta))\widehat{\Psi}(2^{-k}\xi)\widehat{f}(\xi) \sum_{k\geq j-2} 2^{-ks}\widehat{\widetilde{\Psi}_{-s}}(2^{-k}\eta)\widehat{D^s g}(\eta)\, e^{2\pi i(\xi+\eta)\cdot x}\, d\xi\, d\eta \\
&= 2^{js} \sum_{k\geq j-2} 2^{-ks} \widetilde{\Delta_j^s}\Big[\Delta_k(f)\widetilde{\Delta_k^{-s}}(D^s g)\Big](x) \\
&\leq 2^{js} \left(\sum_{k\geq j-2} 2^{-2ks} \right)^{\frac{1}{2}} \left(\sum_{k\geq j-2} \left| \widetilde{\Delta_j^s}\left[\Delta_k(f)\widetilde{\Delta_k^{-s}}(D^s g)\right](x) \right|^2 \right)^{\frac{1}{2}} \\
&\leq C(s) \left(\sum_{k\geq j-2} \left| \widetilde{\Delta_j^s}\Big[\Delta_k(f)\widetilde{\Delta_k^{-s}}(D^s g)\Big](x) \right|^2 \right)^{\frac{1}{2}},
\end{aligned}$$

where

$$\widehat{\widetilde{\Psi}_s}(\xi) := |\xi|^s \widehat{\Psi}(\xi)$$

and

$$\widetilde{\Delta_k^s}(f)^\wedge(\xi) := \widehat{\widetilde{\Psi}_s}(2^{-k}\xi)\widehat{f}(\xi).$$

Thus we have

$$\|\Pi_3[f,g]\|_{L^r} \le C(r,n,s)\left\| \left(\sum_{j\in\mathbf{Z}} \sum_{k\in\mathbf{Z}} \left| \widetilde{\Delta_j^s}[\Delta_k(f)\, \widetilde{\Delta_k^{-s}}(D^s g)] \right|^2 \right)^{\frac{1}{2}} \right\|_{L^r}.$$

We apply [7, Proposition 4.6.4] to extend $\{\widetilde{\Delta_k^s}\}_{k\in\mathbf{Z}}$ from $L^r \to L^r\ell^2$ to $L^r\ell^2 \to L^r\ell^2\ell^2$ for $1 < r < \infty$. This gives

$$\begin{aligned}
\|\Pi_3[f,g]\|_{L^r} &\le C(r,n,s)\left\| \left(\sum_{k\in\mathbf{Z}} \left| \Delta_k(f)\, \widetilde{\Delta_k}(D^s g) \right|^2 \right)^{\frac{1}{2}} \right\|_{L^r} \\
&\le C(r,n,s)\left\| \sup_{k\in\mathbf{Z}} \widetilde{\Delta_k}(D^s g) \right\|_{L^\infty} \left\| \left(\sum_{k\in\mathbf{Z}} |\Delta_k(f)|^2 \right)^{\frac{1}{2}} \right\|_{L^r} \\
&\le C(r,n,s) \sup_{k\in\mathbf{Z}} \left\| \widetilde{\Delta_k}(D^s g) \right\|_{L^\infty} \|f\|_{L^r} \\
&\le C(r,n,s) \left\| \widetilde{\Psi_{-s}} \right\|_{L^1} \|D^s g\|_{L^\infty} \|f\|_{L^r}.
\end{aligned}$$

This proves the case when $(p,q) = (r,\infty)$ while the case $(p,q) = (\infty,r)$ follows by symmetry. □

5 Final remarks

We make a few comments about the inhomogeneous Kato-Ponce inequality (7). It can be obtained from the corresponding homogeneous inequality (8) via the following observation

$$\|J^s f\|_{L^p} = \|(I-\Delta)^{s/2} f\|_{L^p} \approx \|f\|_{L^p} + \|(-\Delta)^{s/2} f\|_{L^p} \tag{21}$$

which is valid for $1 < p < \infty$ and $s > 0$.

Then in the case where $r > 1$ we may use (21) and Hölder's inequality

$$\|fg\|_{L^r} \le \|f\|_{L^{p_j}} \|g\|_{L^{q_j}}$$

for $j = 1, 2$ to obtain (8). In the case $r \le 1$ another argument is needed, similar to that given in Section 4.

Inequalities (7) and (8) are also valid for complex values of s with nonnegative real part. This is an easy consequence of fact that the functions

$$|\xi|^{it}, \qquad (1+|\xi|^2)^{it/2}$$

for t real are L^p Fourier multipliers for $1 < p < \infty$. Finally, it is worth noting that the inequality (7) fails when $s < 0$; see [9].

References

1. F. Bernicot, D. Maldonado, K. Moen, V. Naibo, Bilinear Sobolev-Poincaré inequalities and Leibniz-type rules. J. Geom. Anal. **24**, 1144–1180 (2014)
2. J. Bourgain, D. Li, On an endpoint Kato-Ponce inequality. Differ. Integr. Equ. **27**(11/12), 1037–1072 (2014)
3. F. Christ, M. Weinstein, Dispersion of small-amplitude solutions of the generalized Korteweg-de Vries equation. J. Funct. Anal. **100**, 87–109 (1991)
4. R.R. Coifman, Y. Meyer, Commutateurs d' intégrales singulières et opérateurs multilinéaires. Ann. Inst. Fourier (Grenoble) **28**, 177–202 (1978)
5. R.R. Coifman, Y. Meyer, *Au délà des opérateurs pseudo-différentiels*. Astérisque No. 57 (Societé Mathematique de France, Paris, 1979)
6. E. Cordero, D. Zucco, Strichartz estimates for the vibrating plate equation. J. Evol. Equ. **11**(4), 827–845 (2011)
7. L. Grafakos, *Classical Fourier Analysis*, 2nd edn. Graduate Texts in Mathematics, no 249 (Springer, New York, 2008)
8. L. Grafakos, Multilinear operators in harmonic analysis and partial differential equations. Harmonic analysis and nonlinear partial differential equations, Research Institute of Mathematical Sciences (Kyoto), 2012
9. L. Grafakos, S. Oh, The Kato-Ponce inequality. Commun. Partial Differ. Equ. **39**, 1128–1157 (2014)
10. L. Grafakos, R.H. Torres, Multilinear Calderón-Zygmund theory. Adv. Math. **165**, 124–164 (2002)
11. L. Grafakos, D. Maldonado, V. Naibo, A remark on an endpoint Kato-Ponce inequality. Differ. Integr. Equ. **27**, 415–424 (2014)
12. A. Gulisashvili, M. Kon, Exact smoothing properties of Schröndinger semigroups. Am. J. Math. **118**, 1215–1248 (1996)
13. T. Kato, Remarks on the Euler and Navier-Stokes equations in $\mathbf{R}^2$, in *Nonlinear Functional Analysis and Its Applications, Part 2* (Berkeley, California, 1983). Proceedings of Symposia in Pure Mathematics, vol. 45, Part 2 (American Mathematical Society, Providence, RI, 1986), pp. 1–7
14. T. Kato, G. Ponce, Commutator estimates and the Euler and Navier-Stokes equations. Commun. Pure Appl. Math. **41**, 891–907 (1988)
15. C.E. Kenig, E.M. Stein, Multilinear estimates and fractional integration. Math. Res. Lett. **6**, 1–15 (1999)
16. C.E. Kenig, G. Ponce, L. Vega, Well-posedness and scattering results for the generalized Korteweg-de-Vries equation via the contraction principle. Commun. Pure Appl. Math. **46**(4), 527–620 (1993)
17. C. Muscalu, W. Schlag, *Classical and Multilinear Harmonic Analysis, Volume 2*. Cambridge Studies in Advanced Mathematics, 138 (Cambridge University Press, Cambridge, 2013)
18. C. Muscalu, J. Pipher, T. Tao, C. Thiele, Bi-parameter paraproducts. Acta Math. **193**, 269–296 (2004)

Wavelets and Graph C^*-Algebras

Carla Farsi, Elizabeth Gillaspy, Sooran Kang, and Judith Packer

Abstract Here we give an overview on the connection between wavelet theory and representation theory for graph C^*-algebras, including the higher-rank graph C^*-algebras of A. Kumjian and D. Pask. Many authors have studied different aspects of this connection over the last 20 years, and we begin this paper with a survey of the known results. We then discuss several new ways to generalize these results and obtain wavelets associated to representations of higher-rank graphs. In Farsi et al. (J Math Anal Appl 425:241–270, 2015), we introduced the "cubical wavelets" associated to a higher-rank graph. Here, we generalize this construction to build wavelets of arbitrary shapes. We also present a different but related construction of wavelets associated to a higher-rank graph, which we anticipate will have applications to traffic analysis on networks. Finally, we generalize the spectral graph wavelets of Hammond et al. (Appl Comput Harmon Anal 30:129–150, 2011) to higher-rank graphs, giving a third family of wavelets associated to higher-rank graphs.

Keywords Graph wavelets • Representations of C^*-algebras • Higher-rank graphs

2010 Mathematics Subject Classification. 46L05, 42C40

C. Farsi (✉) • J. Packer
Department of Mathematics, University of Colorado at Boulder, Boulder, CO 80309-0395, USA
e-mail: farsi@euclid.colorado.edu; packer@euclid.colorado.edu

E. Gillaspy
Mathematisches Institut der Universität Münster, Einsteinstrasse 62, Münster, 48149, Germany
e-mail: gillaspy@uni-muenster.de

S. Kang
Department of Mathematics, Sungkyunkwan University, Seobu-ro 2066, Jangan-gu, Suwon, 16419, Republic of Korea
e-mail: sooran@skku.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4_3

1 Introduction

Wavelets were developed by S. Mallat, Y. Meyer, I. Daubechies, A. Grossman, and J. Mallet in the late 1980s [40] and early 1990s as functions on $L^2(\mathbb{R}^n)$ that were well localized in either the "time" or "frequency" domain, and thus could be used to form an orthonormal basis for $L^2(\mathbb{R}^n)$ that behaved well under compression algorithms, for the purpose of signal or image storage. Mallat and Meyer developed a very important algorithm, the so-called multiresolution analysis algorithm, as a way to construct the so-called father wavelets and mother wavelets on $L^2(\mathbb{R})$ from associated "filter functions" [40, 58].

Beginning with the initial work of O. Bratteli and P. Jorgensen in the mid-1990s, which gave a relationship between multiresolution analyses for wavelets on $L^2(\mathbb{R})$ and certain types of representations of the Cuntz algebra $\mathcal{O}_N$, the representations of certain graph C^*-algebras and the constructions of wavelets on $L^2(\mathbb{R})$ were shown to be related. To be more precise, in 1996, Bratteli and Jorgensen first announced in [4] that there was a correspondence between dilation-translation wavelets of scale N on $L^2(\mathbb{R})$ constructed via the multiresolution analyses of Mallat and Meyer, and certain representations of the Cuntz algebra $\mathcal{O}_N$. Later, together with D. Dutkay, Jorgensen extended this analysis to describe wavelets on L^2-spaces corresponding to certain inflated fractal sets [15] constructed from iterated function systems. The material used to form these wavelets also gave rise to representations of $\mathcal{O}_N$. Recently, in [16], Dutkay and Jorgensen were able to relate representations of the Cuntz algebra of $\mathcal{O}_N$ that they termed "monic" to representations on L^2 spaces of other non-Euclidean spaces carrying more locally defined branching operations related to dilations. The form that monic representations take has similarities to earlier representations of $\mathcal{O}_N$ coming from classical wavelet theory.

Initially, the wavelet function or functions were made into an orthonormal basis by applying translation and dilation operators to a fixed family of functions, even in Dutkay and Jorgensen's inflated fractal space setting. However, already the term "wavelet" had come to have a broader meaning as being a function or finite collection of functions on a measure space (X, μ) that could be used to construct either an orthonormal basis or frame basis of $L^2(X, \mu)$ by means of operators connected to algebraic or geometric information relating to (X, μ).

In 1996, A. Jonsson described collections of functions on certain finite fractal spaces that he defined as wavelets, with the motivating example being Haar wavelets restricted to the Cantor set [30]. Indeed, Jonsson had been inspired by the fact that the Haar wavelets, which are discontinuous on $[0, 1]$, are in fact continuous when restricted to the fractal Cantor set, and therefore can be viewed as a very well-behaved orthonormal basis giving a great deal of information about the topological structure of the fractal involved. Also in 1996, just slightly before Jonsson's work, R. Strichartz analyzed wavelets on Sierpinski gasket fractals in [59], and noted that since fractals built up from affine iterated function systems such as the Sierpinski gasket fractal had locally defined translations, isometries, and dilations, they were good candidates for an orthonormal basis of wavelets. Strichartz's wavelets were defined by constructing an orthonormal basis from functions that at each stage of

the iteration building the fractal had certain properties (such as local constance) holding in a piecewise fashion [59].

With Jonsson's and Strichartz's constructions in mind, but starting from an operator-algebraic viewpoint, in 2011, M. Marcolli and A. Paolucci looked at representations of the Cuntz-Krieger C^*-algebras $\mathcal{O}_A$ on certain L^2-spaces, and showed that one could construct generalized "wavelet" families, by using the isometries and partial isometries naturally generating the C^*-algebras $\mathcal{O}_A$ to operate on the zero-order and first-order scaling functions and wavelet functions, thus providing the orthonormal basis for the Hilbert space in question. The Cuntz-Krieger C^*-algebras $\mathcal{O}_A$ are C^*-algebras generated by partial isometries, where the relations between the isometries are determined by the matrix A; interpreting the matrix A as the adjacency matrix of a graph allows us to view the Cuntz-Krieger C^*-algebras as graph algebras. Thus, in the wavelet constructions of Marcolli and Paolucci, the partial isometries coming from the graph algebra act in a sense similar to the localized dilations and isometries observed by Strichartz in [59]. More precisely, by showing that it was possible to represent certain Cuntz-Krieger C^*-algebras on L^2-spaces associated to non-inflated fractal spaces, Marcolli and Paolucci related the work of Bratteli and Jorgensen and Dutkay and Jorgensen to the work of Jonsson and Strichartz. Moreover, they showed that in this setting, certain families related to Jonsson's wavelets could be constructed by acting on the so-called scaling functions and wavelets by partial isometries geometrically related to the directed graph in question.

In this paper, in addition to giving a broad overview of this area, we will discuss several new ways to generalize these results and obtain wavelets associated to representations of directed graphs and higher-rank graphs. For a given directed graph E, the graph C^*-algebra $C^*(E)$ is the universal C^*-algebra generated by a collection of projections associated to the vertices and partial isometries associated to the edges that satisfy certain relations, called the Cuntz-Krieger relations. It has been shown that graph C^*-algebras not only generalize Cuntz-Krieger algebras, but they also include (up to Morita equivalence) a fairly wide class of C^*-algebras such as the AF-algebras, Kirchberg algebras with free K_1-group, and various noncommutative algebras of functions on quantum spaces. One of the benefits of studying graph C^*-algebras is that very abstract properties of C^*-algebras can be visualized via concrete characteristics of underlying graphs. See the details in the book "Graph Algebras" by Iain Raeburn [48] and the references therein.

Higher-rank graphs, also called k-graphs, were introduced in [35] by Kumjian and Pask as higher-dimensional analogues of directed graphs, and they provide a combinatorial model to study the higher-dimensional Cuntz-Krieger algebras of Robertson and Steger [51, 52]. Since then, k-graph C^*-algebras have been studied by many authors and have provided many examples of various classifiable C^*-algebras, and the study of fine structures and invariants of k-graph C^*-algebras can be found in [8, 18, 32, 44, 49, 50, 53, 54, 56, 57]. Also, k-graph C^*-algebras provide many examples of non-self-adjoint algebras and examples of crossed products. (See [7, 13, 14, 19, 21, 33, 47]). Recently, twisted k-graph C^*-algebras have been developed in [36–39, 55]; these provide many important examples of C^*-algebras including noncommutative tori. Moreover, specific examples of dynamical systems

on k-graph C^*-algebras have been studied in [42, 43], and the study of KMS states with gauge dynamics can be found in [23–27]. Furthermore, the works in [45, 46] show that k-graph C^*-algebras can be realized as noncommutative manifolds and have the potential to enrich the study of noncommutative geometry.

The first examples of directed graph algebras are the Cuntz C^*-algebras $\mathcal{O}_N$ defined for any integer $N \geq 2$, which are generated by N isometries satisfying some elementary relations. In the late 1990s it was realized by Bratteli and Jorgensen [4, 5] that the theory of multiresolution analyses for wavelets and the theory of certain representations of $\mathcal{O}_N$ could be connected through filter functions, or quadrature mirror filters, as they are sometimes called. We review this relationship in Section 2, since this was the first historical connection between wavelets and C^*-algebras. In this section, we also relate the filter functions associated to fractals coming from affine iterated function systems, as first defined by Dutkay and Jorgensen in [15], as well as certain kinds of representations of $\mathcal{O}_N$ defined by Bratteli and Jorgensen called *monic* representations, as all three of these representations of $\mathcal{O}_N$ (those coming from [5], from [15], and from [16]) correspond to what we call a *Cuntz-like* family of functions on $\mathbb{T}$. Certain forms of monic representations, when moved to L^2-spaces of Cantor sets associated to $\mathcal{O}_N$, can be viewed as examples of semibranching function systems, and thus are precursors of the types of representations of Cuntz-Krieger algebras studied by Marcolli and Paolucci in [41]. In Section 3 we give an overview of the work of Marcolli and Paolucci from [41], discussing semibranching function systems satisfying a Cuntz-Krieger condition and the representations of Cuntz-Krieger C^*-algebras on the L^2-spaces of fractals. We state the main theorem of Marcolli and Paolucci from [41] on the construction of wavelets on these spaces, which generalizes the constructions of Jonsson and Strichartz, but we omit the proof of their theorem. However, we give the proof that, given any Markov probability measure on the fractal space K_N associated to $\mathcal{O}_N$, there exist an associated representation of $\mathcal{O}_N$ and a family of related wavelets. In Section 4, we review the definition of directed graph algebras and also review C^*-algebras associated to finite higher-rank graphs (first defined by Kumjian and Pask in [35]) and then generalize the notion of semibranching function systems to higher-rank graph algebras via the definition of Λ-semibranching function systems, first introduced in [20]. In Section 5, we use the representations arising from Λ-semibranching function systems to construct wavelets of an arbitrary rectangular shape on the L^2-space of the infinite path space Λ^∞ of any finite strongly connected k-graph Λ. In so doing we generalize a main theorem from [20] and answer in the affirmative a question posed to one of us by Aidan Sims. In Section 6, motivated by the work of Marcolli and Paolucci for wavelets associated to Cuntz-Krieger C^*-algebras, we discuss the use of k-graph wavelets in the construction of (finite-dimensional) families of wavelets that can hopefully be used in traffic analysis on networks, and also discuss generalizations of wavelets on the vertex space of a k-graph that can be viewed as eigenvectors of the (discrete) Laplacian on this vertex space. We analyze the wavelets and the wavelet transform in this case, thereby generalizing some results of Hammond, Vanderghynst, and Gribonval from [29].

This work was partially supported by a grant from the Simons Foundation (#316981 to Judith Packer).

2 C^*-Algebras and Work by Bratteli and Jorgensen and Dutkay and Jorgensen on Representations of $\mathcal{O}_N$

We begin by giving a very brief overview of C^*-algebras and several important constructions in C^*-algebras that will prove important in what follows. Readers interested in further detail can examine B. Blackadar's book [3] (to give just one reference).

Definition 2.1 A C^*-algebra is a Banach algebra which we shall denote by $\mathcal{A}$ that has assigned to it an involution $*$ such that the norm of $\mathcal{A}$ satisfies the so-called C^*-identity:

$$\|a^*a\| \;=\; \|a\|^2,\ \forall a \in \mathcal{A}.$$

By a celebrated theorem of I. Gelfand and M. Naimark, every C^*-algebra can be represented faithfully as a Banach $*$-subalgebra of the algebra of all bounded operators on a Hilbert space.

C^*-algebras have a variety of important and very useful applications in mathematics and physics. C^*-algebras can be used to study the structure of topological spaces, as well as the algebraic and representation-theoretic structure of locally compact topological groups. Indeed, C^*-algebras provide one framework for a mathematical theory of quantum mechanics, with observables and states being described precisely in terms of self-adjoint operators and mathematical states on C^*-algebras. When there are also symmetry groups involved in the physical system, the theory of C^*-algebras allows these symmetries to be incorporated into the theoretical framework as well.

In this paper, we will mainly be concerned with C^*-algebras constructed from various relations arising from directed graphs and higher-rank graphs. These are combinatorial objects satisfying certain algebraic relations that are most easily represented by projections and partial isometries acting on a Hilbert space. The C^*-algebras that we will study will also contain within them certain naturally defined commutative C^*-algebras, as fixed points of a canonical gauge action. These commutative C^*-algebras can be realized as continuous functions on Cantor sets of various types, and under appropriate conditions there are measures on the Cantor sets, and associated representations of the C^*-algebras being studied on the L^2-spaces of the Cantor sets. These will be the representations that we shall study, but we will first briefly review the notion of C^*-algebras characterized by universal properties.

Definition 2.2 ([3]) Let $\mathcal{G}$ be a (countable) set of generators, closed under an involution $*$, and $\mathcal{R}(\mathcal{G})$ a set of algebraic relations on the elements of $\mathcal{G}$, which have as a restriction that it must be possible to realize $\mathcal{R}(\mathcal{G})$ among operators on a Hilbert space $\mathcal{H}$. It is also required that $\mathcal{R}(\mathcal{G})$ must place an upper bound on the norm of each generator when realized as an operator. A **representation**

$(\pi, \mathcal{H})$ of the pair $(\mathcal{G}, \mathcal{R}(\mathcal{G}))$ is a map $\pi : \mathcal{G} \to \mathcal{B}(\mathcal{H})$ such that the collection $\{\pi(g) : g \in \mathcal{G}\}$ satisfies all the relations of $\mathcal{R}(\mathcal{G})$. The smallest C^*-subalgebra of $\mathcal{B}(\mathcal{H})$ containing $\{\pi(g) : g \in \mathcal{G}\}$ is called a C^*-algebra that represents $(\mathcal{G}, \mathcal{R}(\mathcal{G}))$; we denote this C^*-algebra by $\mathcal{A}_\pi$. A representation $(\pi_\mathcal{U}, \mathcal{H}_\mathcal{U})$ of $(\mathcal{G}, \mathcal{R}(\mathcal{G}))$ is said to be **universal** and $\mathcal{A}_{\pi_\mathcal{U}}$ is called the **universal C^*-algebra** associated to $(\mathcal{G}, \mathcal{R}(\mathcal{G}))$ if for every representation $(\pi, \mathcal{H})$ of the pair $(\mathcal{G}, \mathcal{R}(\mathcal{G}))$ there is a $*$-homomorphism $\rho : \mathcal{A}_{\pi_\mathcal{U}} \to \mathcal{A}_\pi$ satisfying

$$\pi(g) = \rho \circ \pi_\mathcal{U}(g), \ \forall g \in \mathcal{G}.$$

The general theory found in Blackadar [3] can be used to show that this universal C^*-algebra exists, and is unique (the bounded-norm condition is used in the existence proof).

Example 2.3 Let $\mathcal{G} = \{u, u^*, v, v^*\}$ and fix $\lambda \in \mathbb{T}$ with $\lambda = e^{2\pi i\alpha}$, $\alpha \in [0, 1)$. Let $\mathcal{R}(\mathcal{G})$ consist of the following three identities, where I denotes the identity operator in $B(\mathcal{H})$:

(1) $uu^* = u^*u = I$.
(2) $vv^* = v^*v = I$.
(3) $uv = \lambda vu$.

We note that relations (1) and (2) together with the C^*-norm condition force $\|u\| = \|v\| = 1$. When $\lambda \neq 1$, relation (3) implies the universal C^*-algebra involved is noncommutative, and our universal C^*-algebra in this case is the well-known noncommutative torus $\mathcal{A}_\alpha$.

Example 2.4 Fix $N > 1$ and let $\mathcal{G} = \{s_0, s_0^*, \cdots, s_{N-1}, s_{N-1}^*\}$. Let $\mathcal{R}(\mathcal{G})$ consist of the relations

(1) $s_i^* s_i = 1, \ 0 \leq i \leq N-1$.
(2) $s_i^* s_j = 0, \ 0 \leq i \neq j \leq N-1$.
(3) $s_1 s_1^* + s_2 s_2^* + \cdots + s_{N-1} s_{N-1}^* = I$.

Again the first collection of relations (1) implies that $\|s_i\| = 1, \ 0 \leq i \leq N-1$, and also imply that the s_i will be isometries and the s_i^* will be partial isometries, $0 \leq i \leq N-1$.[1] The universal C^*-algebra constructed via these generators and relations was first discovered by J. Cuntz in the late 1970s. Therefore it is called the Cuntz algebra and is commonly denoted by $\mathcal{O}_N$.

We now wish to examine several different families of representations of $\mathcal{O}_N$ that take on a related form on the Hilbert space $L^2(\mathbb{T})$ where $\mathbb{T}$ is equipped with Haar measure. These types of representations were first studied by Bratteli and Jorgensen in [4] and [5], who found that certain of these representations could be formed from wavelet filter functions. They also appear as representations coming from inflated

[1] Recall that an isometry in $B(\mathcal{H})$ is an operator T such that $T^*T = I$; a partial isometry S satisfies $S = SS^*S$. A projection in $B(\mathcal{H})$ is an operator that is both self-adjoint and idempotent.

fractal wavelet filters and the recently defined monic representations of Dutkay and Jorgensen [15, 16, respectively].

We now discuss a common theme for all of the representations of $\mathcal{O}_N$ mentioned above, as was first done by Bratteli and Jorgensen in [5]. Fix $N > 1$, and suppose a collection of N essentially bounded measurable functions $\{h_0, h_1, \cdots, h_{N-1}\} \subseteq L^\infty(\mathbb{T})$ is given. We define bounded operators $\{T_i\}_{i=0}^{N-1}$ on $L^2(\mathbb{T})$ associated to the functions $\{h_0, h_1, \cdots, h_{N-1}\}$ by

$$(T_i\xi)(z) \;=\; h_i(z)\xi(z^N), \tag{1}$$

and we ask the question: when do the $\{T_i\}_{i=0}^{N-1}$ give a representation of $\mathcal{O}_N$ on $L^2(\mathbb{T})$?

We first compute that for $i \in \mathbb{Z}_N = \{0, 1, \ldots, N-1\}$, the adjoint of each T_i is given by

$$(T_i^*\xi)(z) \;=\; \frac{1}{N} \sum_{\omega \in \mathbb{T}:\omega^N = z} \overline{h_i(\omega)}\xi(\omega). \tag{2}$$

If we denote the N (measurable) branches of the N^{th} root function by $\tau_j : \mathbb{T} \to \mathbb{T}$, where

$$\tau_j(z = e^{2\pi it}) = e^{\frac{2\pi i(t+j)}{N}},\ t \in [0, 1) \ \text{ and } \ j \in \mathbb{Z}_N,$$

then we can rewrite our formula for T_i^* as:

$$(T_i^*\xi)(z) \;=\; \frac{1}{N} \sum_{j=0}^{N-1} \overline{h_i(\tau_j(z))}\xi(\tau_j(z)). \tag{3}$$

(Note we have chosen specific branches for the N^{th} root functions, but in our formula for the adjoint T_i^* we could have taken any measurable branches and obtained the same result.)

We now give necessary and sufficient conditions on the functions $\{h_0, h_1, \ldots, h_{N-1}\}$, as stated in [5], that the $\{T_i\}_{i=0}^{N-1}$ generate a representation of $\mathcal{O}_N$.

Proposition 2.5 Fix $N > 1$, let $\{h_i\}_{i=0}^{N-1} \subset L^\infty(\mathbb{T})$ and define $\{T_i\}_{i=0}^{N-1}$ as in Equation (1). Then the operators $\{T_i\}_{i=0}^{N-1}$ give a representation of the Cuntz algebra if and only if the map

$$z \;\mapsto\; \left(\frac{h_i\left(ze^{\frac{2\pi ij}{N}}\right)}{\sqrt{N}} \right)_{0 \le i,j \le N-1} \tag{4}$$

is a map from $\mathbb{T}$ into the unitary $N \times N$ matrices for almost all $z \in \mathbb{T}$.

Proof See Section 1 of Bratteli and Jorgensen's seminal paper [5] for more details on this. □

The above proposition motivates the next definition:

Definition 2.6 Let $\{h_j\}_{j=0}^{N-1}$ be a subset of $L^\infty(\mathbb{T})$. We say that this family is a **Cuntz-like family** if the matrix of Equation (4) is unitary for almost all $z \in \mathbb{T}$.

Bratteli and Jorgensen were the first to note, in [4], that certain wavelets on $L^2(\mathbb{R})$, the so-called multiresolution analysis wavelets, could be used to construct representations of the Cuntz algebra $\mathcal{O}_N$, by examining the filter function families, and showing that they were "Cuntz-like." Their representations used low- and high-pass filters associated to the wavelets to construct the related isometries as above. Filter functions on the circle $\mathbb{T}$ are used to define wavelets in the frequency domain (see [58] for an excellent exposition). We thus give our initial definitions of "dilation-translation wavelet families" in the frequency domain rather than the time domain. We note that we restrict ourselves to integer dilations on $L^2(\mathbb{R})$; more general dilation matrices giving rise to unitary dilations on $L^2(\mathbb{R}^d)$ are described in the Strichartz article [58].

Fix an integer $N > 1$. Define the operator D of dilation by N on $L^2(\mathbb{R})$ by:

$$D(f)(t) = \sqrt{N}, f(Nt) \text{ for } f \in L^2(\mathbb{R}).$$

and define the translation operator T on $L^2(\mathbb{R})$ by

$$T(f)(t) = f(t-1) \text{ for } f \in L^2(\mathbb{R}), \text{ and let } T_v = T^v, \ v \in \mathbb{Z}.$$

Let $\mathcal{F}$ denote the Fourier transform on $L^2(\mathbb{R})$ defined by

$$\mathcal{F}(f)(x) = \int_{\mathbb{R}} f(t) e^{2\pi i t x} dt.$$

Set

$$\widehat{D} = \mathcal{F} D \mathcal{F}^* \text{ and } \widehat{T} = \mathcal{F} T \mathcal{F}^*.$$

Then

$$\widehat{D}(f)(x) = \frac{1}{\sqrt{N}} f\left(\frac{x}{N}\right) \text{ and } \widehat{T}(f)(x) = e^{-2\pi i x} f(x) \text{ for } f \in L^2(\mathbb{R}).$$

Definition 2.7 A **wavelet family in the frequency domain** for dilation by $N > 1$ is a subset $\{\Phi\} \cup \{\Psi_1, \cdots, \Psi_m\} \subseteq L^2(\mathbb{R})$ such that

$$\{\widehat{T}_v(\Phi) : \ v \in \mathbb{Z}\} \cup \{\widehat{D}^j \widehat{T}_v(\Psi_i) : \ 1 \le i \le m, \ j \in \mathbb{N}, \ v \in \mathbb{Z}\} \tag{5}$$

is an orthonormal basis for $L^2(\mathbb{R})$. If $m = N-1$ and the set (5) is an orthonormal basis for $L^2(\mathbb{R})$, the family $\{\Phi\} \cup \{\Psi_1, \cdots, \Psi_{N-1}\}$ is called an **orthonormal wavelet family** for dilation by N.

In other words, wavelet families are finite subsets of the unit ball of a Hilbert space $L^2(\mathbb{R})$ that, when acted on by specific operators (in this case unitary operators corresponding to dilation and translation), give rise to a basis for the Hilbert space.

A fundamental algorithm for constructing wavelet families is the concept of multiresolution analysis (MRA) developed by Mallat and Meyer in [40], and key tools for constructing the MRAs are filter functions for dilation by N.

Definition 2.8 Let N be a positive integer greater than 1. A **low-pass filter** m_0 for dilation by N is a function $m_0 : \mathbb{T} \to \mathbb{C}$ which satisfies the following conditions:

(i) $m_0(1) = \sqrt{N}$ ("low-pass condition")
(ii) $\sum_{\ell=0}^{N-1} |m_0\left(ze^{\frac{2\pi i\ell}{N}}\right)|^2 = N$ a.e.;
(iii) m_0 is Hölder continuous at 1;
(iv) (Cohen's condition) m_0 is nonzero in a sufficiently large neighborhood of 1 (e.g., it is sufficient that m_0 be nonzero on the image of $[-\frac{1}{2N}, \frac{1}{2N}]$ under the exponential map from $\mathbb{R}$ to $\mathbb{T}$).

Sometimes in the above definition, condition (iv) Cohen's condition is dropped and thus **frame wavelets** are produced instead of orthonormal wavelets; these situations can be studied further in Bratteli and Jorgensen's book [6].

Given a low-pass filter m_0 for dilation by N, we can naturally view m_0 as a $\mathbb{Z}$-periodic function on $\mathbb{R}$ by identifying $\mathbb{T}$ with $[0, 1)$ and extending $\mathbb{Z}$-periodically. Then there is a canonical way to construct a "scaling function" associated to the filter m_0. We set

$$\Phi(x) = \prod_{i=1}^{\infty} \left[\frac{m_0\left(N^{-i}(x)\right)}{\sqrt{N}} \right].$$

Then the infinite product defining Φ converges a.e. and gives an element of $L^2(\mathbb{R})$. We call Φ a **scaling function in the frequency domain** for dilation by N. (The function $\mathcal{F}^{-1}(\Phi) = \phi$ is the scaling function in the sense of the original definition.)

Given a low-pass filter m_0 and the associated scaling function Φ for dilation by N, then if we have $N-1$ other functions defined on $\mathbb{T}$ which satisfy appropriate conditions described in the definition that follows, we can construct the additional members of a wavelet family for dilation by N.

Definition 2.9 Let N be a positive integer greater than 1, and let m_0 be a low-pass filter for dilation by N satisfying all the conditions of Definition 2.8. A set of essentially bounded measurable $\mathbb{Z}$-periodic functions $m_1, m_2, \cdots, m_{N-1}$ defined on $\mathbb{R}$ are called **high-pass filters** associated to m_0, if

$$\sum_{\ell=0}^{N-1} \overline{m_i\left(ze^{\frac{2\pi i\ell}{N}}\right)} m_j\left(ze^{\frac{2\pi i\ell}{N}}\right) \;=\; \delta_{i,j}N \;\text{ for }\; 0 \leq\; i,j \;\leq N-1.$$

Given a low-pass filter m_0, it is always possible to find *measurable* functions $m_1, m_2, \cdots, m_{N-1}$ that serve as high-pass filters to m_0. The functions $m_1,\ m_2, \cdots, m_{N-1}$ can then be seen as $\mathbb{Z}$-periodic functions on $\mathbb{R}$ as well. The connection between filter functions and wavelet families was provided by Mallat and Meyer for $N = 2$ in [40] and then extended to more general dilation matrices. We consider only integer dilations $N > 1$, and rely on the exposition of both Strichartz [58] and Bratteli and Jorgensen [6] in the material that follows below:

Theorem 2.10 (*[40], Section 1.5 of [58], [6]*) *Let N be a positive integer greater than* 1, *let $(m_0, m_1, \cdots, m_{N-1})$ be a classical system of low and associated high-pass filters for dilation by N, where m_0 satisfies all the conditions of Definition 2.8, and let Φ be the scaling function in the frequency domain constructed from m_0 as above. Then*

$$\{\Phi\} \bigcup \{\Psi_1 = \widehat{D}\,(m_1\Phi)\,,\ \Psi_2 = \widehat{D}\,(m_2\Phi)\,,\ \cdots,\ \Psi_{N-1} = \widehat{D}\,(m_{N-1}\Phi)\} \quad (6)$$

is an orthonormal wavelet family in the frequency domain for dilation by N. The wavelets $\{\Psi_1,\ \Psi_2,\ \cdots,\ \Psi_{N-1}\}$ are called the "wavelets" in the frequency domain for dilation by N. If Cohen's condition is satisfied, the family (6) is an orthonormal wavelet family. (Again, the functions $\{\psi_1 = \mathcal{F}^{-1}(\Psi_1), \psi_2 = \mathcal{F}^{-1}(\Psi_2), \cdots, \psi_{N-1} = \mathcal{F}^{-1}(\Psi_{N-1})\}$ form the "wavelets" in the original sense of the definition.)

Remark 2.11 It follows that filter systems are very important in the construction of wavelets arising from a multiresolution analysis. In their proof of the result above, Bratteli and Jorgensen used a representation of the Cuntz algebra $\mathcal{O}_N$ arising from the filter system.

It is then clear the filter conditions expressed as above can just be formulated as stating that the functions $\{m_0, m_1, \cdots, m_{N-1}\}$ can be used to construct the following function mapping $z \in \mathbb{R}/\mathbb{Z} \cong \mathbb{T}$ into the $N \times N$ unitary matrices over $\mathbb{C}$, given by the formula

$$z \;\mapsto\; \left(\frac{m_j\left(ze^{\frac{2\pi i\ell}{N}}\right)}{\sqrt{N}}\right)_{0\leq j,\ell\leq N-1}, \quad (7)$$

and therefore give a Cuntz-like family in the sense of Definition 2.6. As noted earlier, Bratteli and Jorgensen proved that the operators $\{S_i\}_{i=0}^{N-1}$ defined on $L^2(\mathbb{T})$ by

$$(S_i\xi)(z) \;=\; m_i(z)\xi(z^N), \quad (8)$$

for $\xi \in L^2(\mathbb{T})$, $z \in \mathbb{T}$ and $i = 0, 1, \cdots, N-1$, satisfy the relations

$$S_j^* S_i = \delta_{i,j} I, \tag{9}$$

$$\sum_{i=0}^{N-1} S_i S_i^* = I, \tag{10}$$

which we saw in Proposition 2.5; and thus we obtain exactly the Cuntz relations for the Cuntz algebra $\mathcal{O}_N$.

This gives the **Bratteli-Jorgensen mapping** from a wavelet family $\{\Phi\} \bigcup \{\Psi_1, \cdots, \Psi_{N-1}\}$ in $L^2(\mathbb{R})$ arising from a multiresolution analysis into a representations of $\mathcal{O}_N$.

We now recall the inflated fractal wavelets of Dutkay and Jorgensen [15], which also have a multiresolution analysis structure, and therefore also have related generalized filter functions that will satisfy Definition 2.6 and a weakened low-pass condition. Thus, these filter functions will also give rise to representations of $\mathcal{O}_N$ on $L^2(\mathbb{T})$. We review here only the case where the fractals embed inside $[0, 1]$, although the work in [15] generalizes to fractals sitting inside $[0, 1]^d$ constructed from affine iterated function systems. We note that a fine survey of the relationship between quadrature mirror filters of all types and representations of $\mathcal{O}_N$ can be found in the recent paper [17].

Fix an integer $N > 1$. Recall that $\mathbb{Z}_N = \{0, 1, \cdots, N-1\}$; let $B \subset \mathbb{Z}_N$ be a proper subset of $\mathbb{Z}_N$. Recall from [28] that there is a unique fractal set $\mathbf{F} \subset [0, 1]$ satisfying

$$\mathbf{F} = \bigsqcup_{i \in B} (\frac{1}{N}[\mathbf{F} + i]).$$

The Hausdorff dimension of $\mathbf{F}$ is known to be $\log_N(|B|)$ [28, Theorem 1 of Section 5.3].

Definition 2.12 ([15]) Let N, $B \subset \mathbb{N}$, and $\mathbf{F}$ be as described above. We define the **inflated fractal set** $\mathcal{R}$ associated to $\mathbf{F}$ by:

$$\mathcal{R} = \bigcup_{j \in \mathbb{Z}} \bigcup_{v \in \mathbb{Z}} N^{-j}(\mathbf{F} + v).$$

The Hausdorff measure μ of dimension $\log_N(|B|)$, restricted to $\mathcal{R} \subset \mathbb{R}$, gives a Borel measure on $\mathcal{R}$, but it is not a Radon measure on $\mathcal{R}$, because bounded measurable subsets of $\mathcal{R}$ need not have finite μ-measure. A dilation operator D and translation operators $\{T_v : v \in \mathbb{Z}\}$ on $L^2(\mathcal{R}, \mu)$ are defined as follows: for $f \in L^2(\mathcal{R}, \mu)$,

$$D(f)(x) = \sqrt{|B|} f(Nx),$$
$$T_v(f)(x) = f(x - v).$$

There is a natural multiresolution analysis (MRA) structure on $L^2(\mathcal{R}, \mu)$, which can be described as follows. We define a scaling function or "father wavelet" ϕ by $\phi = \chi_{\mathbf{F}}$. Translates of ϕ are orthonormal, and we define the core subspace V_0 of the MRA to be the closure of their span,

$$V_0 = \overline{\text{span}}\{T_v(\phi) : v \in \mathbb{Z}\}.$$

For $j \in \mathbb{Z}$, set $V_j = D^j(V_0)$. It was shown in Proposition 2.8 of [15] (using slightly different notation) that $\bigcup_{j\in\mathbb{Z}} V_j$ is dense in $L^2(\mathcal{R}, \mu)$ and $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$. The inclusion $V_j \subset V_{j+1}$ follows from the fact that

$$\phi = \frac{1}{\sqrt{|B|}} \sum_{i\in B} DT_i(\phi). \tag{11}$$

We note that the refinement equation (11) above gives a *weakened* low-pass filter for dilation by N, defined by $h_0(z) = \sum_{i\in B} \frac{1}{\sqrt{|B|}} z^i$ for $z \in \mathbb{T}$. It is weakened in that conditions (i) and (iv) of Definition 2.8 will not be satisfied in general, but it will satisfy

$$\sum_{\{w:\, w^N=z\}} |h_0(w)|^2 = N \quad \text{for } z \in \mathbb{T},$$

and $h_0(z)$ will be nonzero in a neighborhood of $z = 1$. Using linear algebra, it is then possible to find $N-1$ corresponding "high-pass" filters $\{h_1, h_2, \cdots, h_{N-1}\}$ defined as Laurent polynomials in z (see Theorem 3.4 of [12] for details) such that the condition of Definition 2.6 is satisfied for the family $\{h_0, h_1, \cdots, h_{N-1}\}$, and one thus obtains a representation of $\mathcal{O}_N$ to go along with the wavelet family. Moreover, the high-pass filters $\{h_1, \cdots, h_{N-1}\}$ are constructed in such a way to allow one to construct a subset $\{\psi_1, \psi_2, \cdots, \psi_{N-1}\}$ of $W_0 = V_1 \ominus V_0$ that serves as the generalized wavelet family for $L^2(\mathcal{R}, \nu)$ in the sense that

$$\{D^j T_v(\psi_i) : \ 1 \le i \le N-1 \ \text{ and } \ j, v \in \mathbb{Z}\}$$

form an orthonormal basis for $L^2(\mathcal{R}, \nu)$. See [15] and [12] for further details on this construction.

Finally we wish to briefly discuss the relationship of the above representations of $\mathcal{O}_N$ coming from Cuntz-like families of functions, to the *monic* representations of $\mathcal{O}_N$ defined by Dutkay and Jorgensen in [16].

Fix an integer $N > 1$. Let K_N denote the infinite product space $\prod_{j=1}^{\infty} \mathbb{Z}_N$, which has the topological structure of the Cantor set. Denote by σ the one-sided shift on K_N :

$$\sigma\left((i_j)_{j=1}^{\infty}\right) = (i_{j+1})_{j=1}^{\infty} \tag{12}$$

and let $\sigma_k,\ k \in \mathbb{Z}_N$ denote the inverse branches to σ :

$$\sigma_k\left((i_j)_{j=1}^{\infty}\right) \ = \ (ki_1i_2\cdots i_j\cdots) \tag{13}$$

Definition 2.13 ([16]) A **monic system** is a pair $(\mu, \{f_i\}_{i\in\mathbb{Z}_N})$, where μ is a finite Borel measure on K_N and $\{f_i\}_{i\in\mathbb{Z}_N}$ are functions on K_N such that for $j \in \mathbb{Z}_N$, $\mu \circ (\sigma_j)^{-1} << \mu$ and

$$\frac{d(\mu \circ (\sigma_j)^{-1})}{d\mu} \ = \ |f_j|^2, \tag{14}$$

and the functions $\{f_j\}$ have the property that

$$|f_j(x)| \ \neq \ 0,\ \mu - \text{a.e. } x \in \sigma_j(K_N).$$

A monic system is called nonnegative if $f_j \geq 0$ for all $j \in \mathbb{Z}_N$.

Given a monic system $(\mu, \{f_i\}_{i\in\mathbb{Z}_N})$, in [16] Dutkay and Jorgensen associated to it a representation of the Cuntz algebra $\mathcal{O}_N$ on $L^2(K_N, \mu)$ defined by:

$$S_j(\xi)(x) \ = \ f_j(x)\cdot \xi\circ\sigma(x) \ \text{ for } \xi \in L^2(K_N,\mu) \text{ and } j \in \mathbb{Z}_N,$$

and they proved that this representation is what they termed a *monic representation* (c.f. Theorem 2.7 of [16] for details).

Recall we have a map $\iota : K_N \to \mathbb{T}$ defined by

$$\iota((i_j)_{j=1}^{\infty}) \ = \ e^{2\pi i\sum_{j=1}^{\infty}\frac{i_j}{N^j}}.$$

We also have an inverse map $\theta : \mathbb{T} \to K_N$ where $\theta(e^{2\pi it}) = (i_j)_{j=1}^{\infty}$ for $t = \sum_{j=1}^{\infty}\frac{i_j}{N^j}$. Although the rational numbers admit more than one N-adic expansion, such anomalies form a set of measure 0 in $\mathbb{T}$. With respect to this correspondence, the map σ looks like τ, where $\tau(z) = z^N$, and the maps σ_j correspond to the maps $\tau_j(e^{2\pi it}) = e^{2\pi i\frac{t+j}{N}}$, i.e.

$$\iota\circ\sigma \ = \ \tau\circ\iota \ \text{ and } \ \tau_j\circ\iota \ = \ \iota\circ\sigma_j \ \text{ for } \ 0 \leq j \leq \ N-1.$$

So, if the measure μ on K_N is equal to Haar measure ν on K_N (thought of as an infinite product of the cyclic groups $\mathbb{Z}_N$), a monic system of functions $\{f_i\}_{i\in\mathbb{Z}_N}$ on (K_N, ν) gives a collection of functions $\{h_i = f_i \circ \theta\}_{i\in\mathbb{Z}_N}$ on $\mathbb{T}$.

The most relevant aspect of Dutkay and Jorgensen's work on monic representations to this paper is that, using the fact that (K_N, ν) can be measure-theoretically identified with $(\mathbb{T}, \nu_{\mathbb{T}})$, where $\nu_{\mathbb{T}}$ is Haar measure on the circle group $\mathbb{T}$, by using the maps θ and ι defined above, it is possible to identify a system of essentially bounded functions $\{h_j = f_j \circ \theta\}_{j=0}^{N-1}$ on $\mathbb{T}$, and one can check that these functions will satisfy

the condition of Definition 2.6. The key relevant point in the proof of this is that by Theorem 2.9 of [16], the support of each f_j is precisely $\sigma_j(K_N)$, and $|f_j((i_j)_{j=1}^{\infty})|^2 = N$ on its support, so that the support of each h_j is precisely $\tau_j(\mathbb{T})$, with $|h_j(z)| = \sqrt{N}$ for $z \in \tau_j(\mathbb{T})$ and 0 otherwise. It therefore follows that monic systems of functions on (K_N, ν) moved over to $\mathbb{T}$ via the map θ all give rise to Cuntz-like systems of functions on $\mathbb{T}$. However, these monic systems will only give rise to filter functions (and hence to classical wavelets) in isolated conditions (e.g., for $N = 2$ it is possible to obtain the Shannon wavelet via a monic system of two functions that is equivalent to the filter functions $m_0(z) = \sqrt{2}\chi_{E_0}$ and $m_1(z) = \sqrt{2}\chi_{E_1}$, where E_0 is the image of $[0, \frac{1}{4}) \cup [\frac{3}{4}, 1]$ under the exponential map from $[0, 1]$ to $\mathbb{T}$, and E_1 is the image in $\mathbb{T}$ of $[\frac{1}{4}, \frac{3}{4})$).

Further analysis of monic representations can be found in [16]. We mention them here because they are the closest analog in the Cuntz C^*-algebra case to the sorts of representations of the higher-rank graph algebras that are used to construct wavelets in [20].

3 Marcolli-Paolucci Wavelets

In the 2011 article [41], Marcolli and Paolucci constructed representations of (finite) Cuntz-Krieger C^*-algebras on L^2-spaces of certain fractals, and then in certain cases went on to define wavelets generalizing the wavelets of A. Jonsson [30]. We recall their basic constructions.

Definition 3.1 Fix an integer $N > 1$. Let $A = \left(A_{i,j}\right)_{i,j\in\mathbb{Z}_N}$ be an $N \times N$ matrix whose entries $A_{i,j}$ take on only values in $\{0, 1\}$. The **Cuntz-Krieger C^*-algebra** $\mathcal{O}_A$ is the universal C^*-algebra generated by partial isometries $\{T_i\}_{i\in\mathbb{Z}_N}$ satisfying

$$T_i^* T_i = \sum_{j=0}^{N-1} A_{i,j} T_j T_j^*, \tag{15}$$

$$T_i^* T_j = 0 \quad \text{for } i \neq j, \tag{16}$$

and

$$\sum_{i=0}^{N-1} T_i T_i^* = \mathrm{I}. \tag{17}$$

We note that these Cuntz-Krieger C^*-algebras $\mathcal{O}_A$ are examples of C^*-algebras associated to certain special finite directed graphs, namely, those directed graphs admitting at most one edge with source v and range w for any pair of vertices (v, w). Indeed (cf. [48], Remark 2.8) one can show that the directed graph in this case would have N vertices in a set E_A^0, labeled $E_A^0 = \{v_0, v_1, \cdots, v_{N-1}\}$, with edge set $E_A^1 = \{e_{(i,j)} \in \mathbb{Z}_N^2 : A_{i,j} = 1\}$; there is a (directed) edge $e_{(i,j)}$ beginning at v_j

and ending at v_i iff $A_{i,j} = 1$. The matrix A then becomes the vertex matrix of the associated directed graph. In the case where A is the matrix that has 1 in every entry, the C^*-algebra $\mathcal{O}_A$ is exactly the Cuntz algebra $\mathcal{O}_N$.

As had been done previously by K. Kawamura [34], Marcolli and Paolucci constructed representations of $\mathcal{O}_A$ by employing the method of "semibranching function systems." We note for completeness that the semibranching function systems of Kawamura [34] were for the most part defined on finite Euclidean spaces, e.g. the unit interval $[0, 1]$, whereas the semibranching function systems used by Marcolli and Paolucci [41] were mainly defined on Cantor sets.

Definition 3.2 (c.f. [34, 41] Definition 2.1, [1] Theorem 2.22) Let (X, μ) be a measure space and let $\{D_i\}_{i\in\mathbb{Z}_N}$ be a collection of μ-measurable sets and $\{\sigma_i : D_i \to X\}_{i\in\mathbb{Z}_N}$ a collection of μ-measurable maps. Let A be an $N \times N$ $\{0, 1\}$-matrix. The family of maps $\{\sigma_i\}_{i\in\mathbb{Z}_N}$ is called a *semibranching function system* on (X, μ) with coding map $\sigma : X \to X$ if the following conditions hold:

1. For $i \in \mathbb{Z}_N$, set $R_i = \sigma_i(D_i)$. Then we have

$$\mu(X \setminus \cup_{i\in\mathbb{Z}_N} R_i) = 0 \quad \text{and} \quad \mu(R_i \cap R_j) = 0 \ \text{ for } \ i \neq j.$$

2. For $i \in \mathbb{Z}_N$, we have $\mu \circ \sigma_i << \mu$ and

$$\frac{d(\mu \circ \sigma_i)}{d\mu} \ > \ 0, \ \mu - \text{a.e. on} \quad D_i. \tag{18}$$

3. For $i \in \mathbb{Z}_N$ and a.e. $x \in D_i$, we have

$$\sigma \circ \sigma_i(x) \ = \ x.$$

4. (Cuntz-Krieger (C-K) condition:) For $i, j \in \mathbb{Z}_N$, $\mu(D_i \Delta \cup_{i:A_{i,j}=1} R_j) \ = \ 0$.

Example 3.3 ([34]*)* Take $N > 1$, $(X, \mu) = (\mathbb{T}, \nu)$ where ν is Haar measure on $\mathbb{T}$, $D_i = \mathbb{T}$ for $i \in \mathbb{Z}_N$, and $\sigma_j(z = e^{2\pi it}) = e^{\frac{2\pi i(t+j)}{N}}$ for $t \in [0, 1)$; then $R_j = \{e^{2\pi it} : t \in [\frac{j}{N}, \frac{j+1}{N})\}$. With the coding map given by $\sigma(z) = z^N$, we obtain a semibranching function system satisfying the (C-K) condition for the $N \times N$ matrix consisting of all 1's.

Example 3.4 ([41] Proposition 2.6*)* Take $N > 1$, and fix an $N \times N$ $\{0, 1\}$-matrix A. Let $\Lambda_A \subset \prod_{j=1}^{\infty}[\mathbb{Z}_N]_j$ be defined by

$$\Lambda_A \ = \ \{(i_1 i_2 \cdots i_j \cdots) : \ A_{ij i_{j+1}} = 1 \ \text{ for } \ j \in \mathbb{N}\}.$$

Marcolli and Paolucci have shown that, using the N-adic expansion map, Λ_A can be embedded in $[0, 1]$ as a fractal set and thus has a corresponding Hausdorff probability measure μ_A defined on its Borel subsets. For each $i \in \mathbb{Z}_N$, let

$$D_i \ = \ \{(i_1 i_2 \cdots i_j \cdots) : \ A_{i,i_1} = 1\} \ \subset \Lambda_A,$$

and define σ_j for $j \in \mathbb{Z}_N$ by

$$\sigma_j : D_j \to \Lambda_A : \ \sigma_j((i_1 i_2 \cdots i_k \cdots)) \ = \ (j i_1 i_2 \cdots i_k \cdots).$$

Then

$$R_j := \sigma_j(D_j) = \{(j i_1 i_2 \cdots i_k \cdots) \ : (i_1 i_2 \cdots i_k \cdots) \in D_j\},$$

and denoting by σ the one-sided shift on Λ_A :

$$\sigma\left((i_1 i_2 \cdots i_k \cdots)\right) = (i_2 i_3 \cdots i_{k+1} \cdots)$$

we have that $\sigma \circ \sigma_j(x) = x$ for $x \in D_j$ and $j \in \mathbb{Z}_N$. Marcolli and Paolucci show in Section 2.1 of [41] that this data gives a semibranching function system satisfying the (C-K) condition on (Λ_A, μ_A). If A is the matrix consisting entirely of 1s, we obtain a monic system in the sense of [16]. Moreover, in this case, $D_i = K_N$ for all $i \in \mathbb{Z}_N$ and $R_i = Z(i) = \{(i_j)_{j=1}^{\infty} : \ i_1 = i\}$.

Kawamura and then Marcolli and Paolucci observed the following relationship between semibranching function systems satisfying the (C-K) condition and representations of $\mathcal{O}_A$:

Proposition 3.5 (*c.f. [41] Proposition 2.5*) *Fix a non-trivial $N \times N$ $\{0,1\}$-matrix A with $A_{i,i} = 1$ for all $i \in \mathbb{Z}_N$. Let (X, μ) be a measure space, and let $\{D_i\}_{i \in \mathbb{Z}_N}$, $\{\sigma_i : D_i \to X\}_{i \in \mathbb{Z}_N}$ and $\{R_i = \sigma_i(D_i)\}_{i \in \mathbb{Z}_N}$ be a semibranching function system satisfying the (C-K) condition on (X, μ) with coding map $\sigma : X \to X$. For each $i \in \mathbb{Z}_N$ define $S_i : L^2(X, \mu) \to L^2(X, \mu)$ by*

$$S_i(\xi)(x) \ = \ \chi_{R_i}(x)\Big(\frac{d(\mu \circ \sigma_i)}{d\mu}(\sigma(x))\Big)^{-\frac{1}{2}} \xi(\sigma(x)) \ \text{ for } \xi \in L^2(X, \mu) \text{ and } x \in X.$$

Then the family $\{S_i\}_{i \in \mathbb{Z}_N}$ of partial isometries satisfies the Cuntz-Krieger relations Equations (15), (16), and (17), and therefore generates a representation of the Cuntz-Krieger algebra $\mathcal{O}_A$.

We now discuss the construction of wavelets for Cuntz-Krieger C^*-algebras as developed by Marcolli and Paolucci. In the setting of Example 3.4, suppose in addition that the matrix A is irreducible, i.e. for every pair $(i, j) \in \mathbb{Z}_N \times \mathbb{Z}_N$ there exists $n \in \mathbb{N}$ with $A^n_{i,j} \neq 0$.

In this case, Marcolli and Paolucci proved that the Hausdorff measure μ_A on Λ_A is exactly the probability measure associated to the normalized Perron-Frobenius eigenvector of A. Namely, suppose $(p_0, p_1, \ldots, p_{N-1})$ is a vector in $(0, \infty)^N$ satisfying $\sum_{i=0}^{N-1} p_i = 1$, and such that

$$A(p_0, p_1, \ldots, p_{N-1})^T = \rho(A)(p_0, p_1, \cdots, p_{N-1})^T,$$

where $\rho(A)$ is the spectral radius of A. (The existence of such a vector $(p_0, \ldots, p_{N-1})$, called the **Perron-Frobenius eigenvector** of A, follows from the irreducibility of A.) Then we have:

Theorem 3.6 (*[41], Theorem 2.17*) *Let $N > 1$ be fixed, and suppose that A is an irreducible $\{0,1\}$-matrix. Let $\{\sigma_i : D_i \to R_i\}$ with $\sigma : \Lambda_A \to \Lambda_A$ be the semibranching function system satisfying the (C-K) condition associated to (Λ_A, μ_A) as in Example 3.4. Then the Hausdorff measure μ_A on Λ_A is exactly the probability measure associated to the Perron-Frobenius eigenvector $(p_0, \ldots, p_{N-1})$ of A. To be precise, for $i \in \mathbb{Z}_N$, $\mu_A(R_i) = p_i$ and*

$$\frac{d(\mu \circ \sigma_i)}{d\mu} = N^{-\delta_A}, \text{ a.e. on } D_i,$$

where δ_A is the Hausdorff dimension of Λ_A, and the spectral radius $\rho(A)$ of A is equal to N^{δ_A}.

Given an irreducible $\{0,1\}$-matrix A as in Theorem 3.6, Marcolli and Paolucci were able to construct families of $\mathcal{O}_A$-wavelets on $L^2(\Lambda_A, \mu_A)$ generalizing splines. We describe their construction here (see also Section 3 of [41]). For the purposes of this survey, we concentrate here on the wavelets whose scaling functions or "father wavelets" are constant on the subsets R_i of Λ_A.

We denote by $\mathcal{V}_0$ the (finite-dimensional) subspace of $L^2(\Lambda_A, \mu_A)$ given by

$$\mathcal{V}_0 = \overline{\text{span}}\{\chi_{R_i} : i \in \mathbb{Z}_N\}.$$

For each k, $0 \le k \le N-1$, let $\mathcal{D}_k = \{j : A_{kj} = 1\}$, and let $d_k = |\mathcal{D}_k|$. Enumerate the elements of $\mathcal{D}_k$ by setting $\mathcal{D}_k = \{n_0 < n_1 < \cdots < n_{d_k-1}\}$. For each $k \in \mathbb{Z}_N$, define the following inner product on $\mathbb{C}^{d_k}$:

$$\langle (x_j), (y_j) \rangle_{PF} = \sum_{j=0}^{d_k-1} \overline{x_j} y_j p_{n_j},$$

where (p_{n_j}) are the appropriate coefficients of the Perron-Frobenius eigenvector of A. We now define vectors $\{c^{j,k} : 0 \le j \le d_k - 1, \ 0 \le k \le N-1\}$, where $c^{j,k} = (c_1^{j,k}, \ldots, c_{d_k-1}^{j,k})$, such that for each $k \in \mathbb{Z}_N$, $\{c^{j,k} : 0 \le j \le d_k - 1\}$ is an orthonormal basis for $\mathbb{C}^{d_k-1}$ with respect to the inner product $\langle \cdot, \cdot \rangle_{PF}$, so that

$$c_\ell^{0,k} = c_{\ell'}^{0,k} \text{ for } 0 \le \ell, \ell' \le d_k - 1 \text{ and } k \in \mathbb{Z}_N,$$

and for each fixed $k \in \mathbb{Z}_N$,

$$\text{span}\{c^{j,k} : 1 \le j \le d_k - 1\} = \{(1, 1, \cdots, 1)\}^\perp$$

with respect to the inner product $\langle \cdot, \cdot \rangle_{PF}$ defined above.

We now note that we can write each set R_k as a disjoint union:

$$R_k = \bigcup_{j=0}^{d_k-1} R_{[kn_j]},$$

where

$$R_{[kn_j]} \;=\; \{(i_1 i_2 \cdots i_n \cdots) \in \Lambda_A : \;\; i_1 = k \text{ and } i_2 = n_j\}.$$

Thus in terms of characteristic functions,

$$\chi_{R_k} \;=\; \sum_{j=0}^{d_k-1} \chi_{R_{[kn_j]}} \;\text{ for }\; k \in \mathbb{Z}_N.$$

Now for each $k \in \mathbb{Z}_N$ we define functions $\{f^{j,k}\}_{j=0}^{d_k-1}$ on Λ_A by

$$f^{j,k}(x) \;=\; \frac{1}{\sqrt{p_k}} \sum_{\ell=0}^{d_k-1} c_\ell^{j,k} \chi_{R_{[kn_\ell]}}(x).$$

We note that each function $f^{j,k}$ is μ_A-measurable. Also, for each $k \in \mathbb{Z}_N$,

$$f^{0,k} = \frac{1}{\sqrt{p_k}} \sum_{\ell=0}^{d_k-1} c_\ell^{0,k} \chi_{R_{[kn_\ell]}} = \frac{c_1^{0,k}}{\sqrt{p_k}} \sum_{\ell=0}^{d_k-1} \chi_{R_{[kn_\ell]}} = \frac{c_1^{0,k}}{\sqrt{p_k}} \chi_{R_k}$$

is a scalar multiple of χ_{R_k}, since the vector $c^{0,k}$ is a constant vector. It follows that

$$\operatorname{span}\{f^{0,k}\}_{k=0}^{N-1} \;=\; \operatorname{span}\{\chi_{R_k}\}_{k=0}^{N-1} \;=\; \mathcal{V}_0.$$

We are now nearly ready to state our simplified version of the main theorem on wavelets in [41]. First, a definition: Fix an integer $n > 1$. We say that a word $w = w_1 w_2 \cdots w_n$ in $\prod_{k=1}^{n} \mathbb{Z}_N$ is *admissible* for our $\{0,1\}$-matrix A if, for all $1 \le i \le n-1$, we have $A_{w_i w_{i+1}} = 1$. If w is admissible, we write S_w for the partial isometry in $B(L^2(\Lambda_A, \mu_A))$ given by

$$S_w = S_{w_1} S_{w_2} \cdots S_{w_n}.$$

We also remark that in order to be consistent with standard notation from multiresolution analysis theory and also with our notation for the higher-rank graph C^*-algebra wavelets, we have changed the notation for the orthogonal subspaces from the original notation used in [41].

Theorem 3.7 (*[41], Theorem 3.2*) *Fix $N > 1$. Let A be an $N \times N$, irreducible, $\{0,1\}$-matrix, let (Λ_A, μ_A) be the associated fractal space with Hausdorff measure, and let $\{\sigma_j : D_j \to R_j\}_{j\in\mathbb{Z}_N}$ and σ be the associated semibranching function system satisfying the (C-K) condition defined on (Λ_A, μ_A). Let $\{S_k\}_{k\in\mathbb{Z}_N}$ be the set of operators on $L^2(\Lambda_A, \mu_A)$ given by the formula in Proposition 3.5. Let $\{f^{j,k} : k \in \mathbb{Z}_N,\ 0 \le j \le d_k - 1\}$ be the functions on Λ_A defined in the above paragraphs. For $k \in \mathbb{Z}_N$, let*

$$\phi_k = f^{0,k}.$$

Define

$$\mathcal{W}_0 = \overline{\text{span}}\{f^{j,k} : k \in \mathbb{Z}_N,\ 1 \le j \le d_{k-1}\};$$

$\mathcal{W}_n = \overline{\text{span}} S_w(f^{j,k}) : k \in \mathbb{Z}_N, 1 \le j \le d_k - 1$ and w is an admissible word of length n. Then the subspaces $\mathcal{V}_0$ and $\{\mathcal{W}_n\}_{n=0}^{\infty}$ are mutually pairwise orthogonal in $L^2(\Lambda_A, \mu_A)$ and

$$L^2(\Lambda_A, \mu_A) = \text{span}\left(\mathcal{V}_0 \oplus \Big[\bigoplus_{n=0}^{\infty} \mathcal{W}_n\Big]\right).$$

The ϕ_k are called the scaling functions (or "father wavelets") and the $f^{j,k}$ are called the wavelets (or "mother wavelets") for the system.

Since the proof of the above theorem can be read in [41], we do not include it here. However, as we did in the second paragraph of Section 4 of [20], we do wish to remark upon the fact that the emphasis on the Perron-Frobenius measure in [41] does not appear to be crucial for the construction of the orthogonal subspaces. To illustrate this further, we now construct wavelets for $\mathcal{O}_N$ corresponding to any Markov measure on K_N, and here we will include the proof so as to illustrate our techniques. Note also that by taking tensor products, the wavelets below will produce wavelets on k-graph algebras of tensor-product type, for example, in $\mathcal{O}_N \otimes \mathcal{O}_M$, as studied in Example 3.8 of [32].

Recall from Section 2 that K_N is the infinite product space $\prod_{j=1}^{\infty} \mathbb{Z}_N$ which can be realized as the Cantor set. Let σ and $\{\sigma_j\}_{j\in\mathbb{Z}_N}$ be the one-sided shift on K_N and its inverse branches given in (12) and (13). Following Example 3.11 of [16], fix $\{p_i \in (0,1) : i \in \mathbb{Z}_N\}$, with $\sum_{i\in\mathbb{Z}_N} p_i = 1$, and define the Markov measure

$$\mu(Z(i_1 i_2 \cdots i_n)) = \prod_{j=1}^{n} p_{i_j},$$

where $i_j \in \mathbb{Z}_N$ for $1 \le j \le n$, and

$$Z(i_1 i_2 \cdots i_n) \;=\; \{(x_1 x_2 \cdots x_j \cdots) : \; x_1 = i_1, \; x_2 = i_2, \cdots, \; x_n = i_n\}.$$

As described at the end of Example 3.4, for the $N \times N$ matrix consisting of all 1's, the standard semibranching function system on K_N satisfying the (C-K) condition is given by $\{\sigma_i : D_i \to R_i\}_{i \in Z_N}$, where $\sigma : K_N \to K_N$ satisfies $D_i = K_N$ for all $i \in \mathbb{Z}_N$ and $R_i = Z(i)$.

Theorem 3.8 *Fix $N > 1$, let $\{p_i\}_{i=0}^{N-1}$ be a collection of positive numbers with $\sum_{i=0}^{N-1} p_i = 1$, and let μ be the associated Markov Borel probability measure on (K_N, μ) defined as above. For $i \in \mathbb{Z}_N$, let $\{\sigma_i : K_N \to R_i = Z(i)\}$ and $\sigma : K_N \to K_N$ be the associated semibranching function system satisfying the (C-K) condition associated to the $N \times N$ matrix of all $1'$s, and define $S_i \in B(L^2(K_N, \mu))$ by*

$$S_i(f)(w) = \chi_{Z(i)}(w) p_i^{-1/2} f(\sigma(w)).$$

Then as in Theorem 3.7, there are scaling functions $\{\phi_k\}_{k=0}^{N-1} \subset L^2(K_N, \mu)$ and "wavelets" $\{\psi_{j,k} : \; k \in \mathbb{Z}_N, \; 1 \le j \le N-1\}$ such that setting

$$\mathcal{V}_0 = \overline{\text{span}}\{\phi_k : \; k \in \mathbb{Z}_N\},$$

$$\mathcal{W}_0 \;=\; \overline{\text{span}}\{\psi_{j,k} : k \in \mathbb{Z}_N, \; 1 \le j \le N-1\} \;\text{ and}$$

$$\mathcal{W}_n = \overline{\text{span}}\{S_w(\psi_{j,k}) : k \in \mathbb{Z}_N, \; 1 \le j \le N-1, \; w \;\text{ a word of length }\; n\} \;\text{ for }\; n \ge 1,$$

we obtain

$$L^2(K_N, \mu) = \;\text{span}\left(\mathcal{V}_0 \oplus [\bigoplus_{n=0}^{\infty} \mathcal{W}_n]\right).$$

Proof Following the method of Theorem 3.7, we define an inner product on $\mathbb{C}^N$ by setting

$$\langle (x_j)_j, (y_j)_j \rangle \;=\; \sum_{j \in \mathbb{Z}_N} \overline{x_j} \cdot y_j \cdot p_j. \tag{19}$$

For fixed $k \in \mathbb{Z}_N$, we let $c^{0,k}$ be the vector in $\mathbb{C}^N$ defined by

$$c^{0,k} = (1, 1, \cdots, 1),$$

and let $\{c^{j,k}\}_{1 \le j \le N-1}$, with $c^{j,k} \;=\; (c_\ell^{j,k})_{\ell \in \mathbb{Z}_N}$, be any orthonormal basis for $\{(1, 1, \cdots, 1)\}^{\perp}$ with respect to the inner product (19). For fixed $k \in \mathbb{Z}_N$, define functions $\{f^{j,k} : 0 \; \le j \le N-1\}$ on K_N by:

$$f^{j,k}(x) \;=\; \frac{1}{\sqrt{p_k}} \sum_{\ell=0}^{N-1} c_\ell^{j,k} \chi_{Z(k\ell)}(x).$$

Note that $f^{0,k}$ is a normalized version of $\chi_{Z(k)}(x)$. We claim that setting

$$\phi_k(x) = f^{0,k}(x) \text{ for } k \in \mathbb{Z}_N,$$

and

$$\psi_{j,k}(x) = f^{j,k}(x) \text{ for } 1 \leq k \leq N-1 \text{ and } 1 \leq j \leq N-1,$$

we will obtain a wavelet family for $L^2(K_N, \mu)$ where μ is the Markov measure determined by

$$\mu(Z(i_1 i_2 \cdots i_n)) = \prod_{j=1}^{n} p_{i_j}.$$

We first note that if $i_1 \neq i_2$, the integral

$$\int_{K_N} \phi_{i_1} \overline{\phi_{i_2}} d\mu$$

is a scalar multiple of the integral

$$\int_{K_N} \chi_{Z(i_1)}(x) \chi_{Z(i_2)}(x) d\mu$$

and this latter integral is equal to zero because the functions in question have disjoint support.

We also remark that for $1 \leq j \leq N-1$ and $k \in \mathbb{Z}_N$,

$$\sum_{\ell=0}^{N-1} c_\ell^{j,k} p_\ell = 0.$$

Multiplying by p_k we get:

$$\sum_{\ell=0}^{N-1} c_\ell^{j,k} p_\ell p_k = 0, \quad \text{so that} \quad \int_{K_N} \Big[\sum_{\ell=0}^{N-1} c_\ell^{j,k} \chi_{Z(k\ell)}(x) \Big] d\mu = 0.$$

We can write this as:

$$\int_{K_N} \Big[\sum_{\ell=0}^{N-1} c_\ell^{j,k} \chi_{Z(k\ell)}(x) \Big] \overline{\chi_{Z(k)}(x)} d\mu = 0,$$

i.e.,

$$\int_{K_N} \psi_{j,k}(x) \overline{\phi_k(x)} d\mu = 0 \text{ for } 1 \leq j \leq N-1.$$

We now check and calculate:

$$\begin{aligned} S_i(\psi_{j,k})(x) &= S_i\Big(\frac{1}{\sqrt{p_k}}\sum_{\ell=0}^{N-1} c_\ell^{j,k}\chi_{Z(k\ell)}\Big)(x) = \frac{1}{\sqrt{p_k}}\sum_{\ell=0}^{N-1} c_\ell^{j,k} S_i(\chi_{Z(k\ell)})(x) \\ &= \frac{1}{\sqrt{p_k}}\sum_{\ell=0}^{N-1} c_\ell^{j,k}\chi_{Z(i)}(x)\frac{1}{\sqrt{p_i}}\chi_{Z(k\ell)}(\sigma(x)) \\ &= \frac{1}{\sqrt{p_k}}\frac{1}{\sqrt{p_i}}\sum_{\ell=0}^{N-1} c_\ell^{j,k}\chi_{Z(ik\ell)}(x). \end{aligned}$$

Let $\mathcal{V}_0 = \operatorname{span}\{\phi_i : i \in \mathbb{Z}_N\}$ and let

$$\mathcal{W}_0 = \operatorname{span}\{\psi_{j,k} : k \in \mathbb{Z}_N,\ 1 \le j \le N-1\}.$$

We have shown $\mathcal{V}_0 \perp \mathcal{W}_0$. We now define

$$\mathcal{W}_1 = \operatorname{span}\{S_i(\psi_{j,k}) : i,\ k \in \mathbb{Z}_N,\ 1 \le j\ \le N-1\}.$$

A straightforward calculation shows that $\{S_i(\psi_{j,k}) : i,\ k \in \mathbb{Z}_N,\ 1 \le j\ \le N-1\}$ is an orthonormal set of functions.

We prove now that $(\mathcal{V}_0 \oplus \mathcal{W}_0) \perp \mathcal{W}_1$.

Let us first fix pairs (j,k) and (j',k') with $k,\ k' \in \mathbb{Z}_N$ and $1 \le j,\ j' \le N-1$. Fix $i \in \mathbb{Z}_N$. Then

$$\begin{aligned} &\int_{K_N} S_i(\psi_{j,k})(x)\overline{\psi_{j',k'}(x)}d\mu = \frac{1}{\sqrt{p_i}}\int_{K_N}\Big[\frac{1}{\sqrt{p_k}}\sum_{\ell=0}^{N-1} c_\ell^{j,k}\chi_{Z(ik\ell)}(x)\Big]\overline{\psi_{j',k'}(x)}\,d\mu \\ &= \frac{1}{\sqrt{p_i}}\int_{K_N}\Big[\frac{1}{\sqrt{p_k}}\sum_{\ell=0}^{N-1} c_\ell^{j,k}\chi_{Z(ik\ell)}(x)\Big]\overline{\frac{1}{\sqrt{p_{k'}}}\sum_{\ell'=0}^{N-1} c_{\ell'}^{j',k'}\chi_{Z(k'\ell')}(x)}d\mu \\ &= \frac{1}{\sqrt{p_i}}\frac{1}{\sqrt{p_k p_{k'}}}\int_{K_N}\sum_{\ell=0}^{N-1}\sum_{\ell'=0}^{N-1}\delta_{i,k'}\delta_{k,\ell'}c_\ell^{j,k}\overline{c_{\ell'}^{j',k'}}\chi_{Z(ik\ell)}(x)d\mu \\ &= \frac{1}{\sqrt{p_i}}\frac{1}{\sqrt{p_k p_{k'}}}\int_{K_N}\sum_{\ell=0}^{N-1}\delta_{i,k'}c_\ell^{j,k}\overline{c_k^{j',k'}}\chi_{Z(ik\ell)}(x)d\mu = \frac{1}{\sqrt{p_i}}p_i p_k\delta_{i,k'}\frac{1}{\sqrt{p_k p_{k'}}}\sum_{\ell=0}^{N-1} c_\ell^{j,k}\overline{c_k^{j',k'}}p_\ell \\ &= \frac{1}{\sqrt{p_i}}p_i p_k\delta_{i,k'}\overline{c_k^{j',k'}}\frac{1}{\sqrt{p_k p_{k'}}}\sum_{\ell=0}^{N-1} c_\ell^{j,k}p_\ell = 0, \end{aligned}$$

since $\sum_{\ell=0}^{N-1} c_\ell^{j,k}p_\ell = 0$ for $1 \le j \le N-1$.

It follows that

$$\mathcal{W}_0 \perp \mathcal{W}_1.$$

We now show that $\mathcal{V}_0 \perp \mathcal{W}_1$. Fix $i \in \mathbb{Z}_N$. Let $i',\ k \in \mathbb{Z}_N$ and fix $j \in \{1, 2, \cdots, N-1\}$. Then:

$$\begin{aligned}
\langle \phi_i, S_{i'}(\psi_{j,k})\rangle &= \frac{1}{\sqrt{p_i}} \int_{K_N} \chi_{Z(i)}(x) \overline{\frac{1}{\sqrt{p_{i'}}} \frac{1}{\sqrt{p_k}} \sum_{\ell=0}^{N-1} c_\ell^{j,k} \chi_{Z(i'k\ell)}(x) d\mu} \\
&= \frac{1}{\sqrt{p_i p_{i'}}} \frac{1}{\sqrt{p_k}} \sum_{\ell=0}^{N-1} \int_{K_N} \chi_{Z(i)}(x) \overline{c_\ell^{j,k}} \chi_{Z(i'k\ell)}(x) d\mu \\
&= \frac{1}{\sqrt{p_i p_{i'} p_k}} \delta_{i,i'} \sum_{\ell=0}^{N-1} \int_{K_N} \overline{c_\ell^{j,k}} \chi_{Z(i'k\ell)}(x) d\mu \\
&= \delta_{i,i'} \frac{1}{\sqrt{p_i p_{i'} p_k}} \sum_{\ell=0}^{N-1} \overline{c_\ell^{j,k}} p_{i'} p_k p_\ell \\
&= \delta_{i,i'} \frac{1}{\sqrt{p_i p_{i'} p_k}} p_{i'} p_k \Big[\sum_{\ell=0}^{N-1} \overline{c_\ell^{j,k}} p_\ell \Big].
\end{aligned}$$

In order for this value to have a chance of being nonzero we need $i = i'$. But even if that happens we get:

$$\langle \phi_i, S_i(\psi_{j,k})\rangle = p_k \Big[\frac{1}{\sqrt{p_k}} \sum_{\ell=0}^{N-1} \overline{c_\ell^{j,k}} p_\ell \Big] = \sqrt{p_k} \cdot \langle c^{j,k}, (1, 1, \cdots, 1)\rangle_{\mathbb{C}^N},$$

which is equal to 0 for $j \in \mathbb{Z}_N \backslash \{0\}$. Thus $\mathcal{V}_0$ is orthogonal to $\mathcal{W}_1$.

We now prove by induction that if for every $n \geq 0$ we define

$$\mathcal{W}_n \ = \ \text{span}\{S_w(\psi_{j,k}) : \ w \text{ is a word of length } n,\ k \in \mathbb{Z}_N \text{ and } 1 \leq j \leq N-1\},$$

then for all $n \geq 0$,

$$\mathcal{W}_{n+1} \perp \Big[\mathcal{V}_0 \oplus \bigoplus_{k=0}^{n} \mathcal{W}_k \Big].$$

We have proven this for $n = 0$ directly. We now assume it is true for $\ell = n$ and prove it is true for $\ell = n+1$, i.e. let us prove that $\mathcal{W}_{n+2}$ is orthogonal to $[\mathcal{V}_0 \oplus \bigoplus_{k=0}^{n+1} \mathcal{W}_k]$. We first note that if w is a word of length $n+2$ and w' is a word of length s where $1 \leq s \leq n+1$, and if $k, k' \in \mathbb{Z}_N$ and $1 \leq j, j' \leq N-1$, then there are unique $i, i' \in \mathbb{Z}_N$ such that

$$\langle S_w(\psi_{j,k}), S_{w'}(\psi_{j',k'})\rangle_{L^2(K_N,\mu)} = \langle S_i S_{w_1}(\psi_{j,k}), S_{i'} S_{w_1'}(\psi_{j',k'})\rangle_{L^2(K_N,\mu)}.$$

where w_1 is a word of length $n+1$ and w_1' is a word of length $s-1$ for $0 \le s-1 \le n$. This then is equal to

$$\langle S_{w_1}(\psi_{j,k}), S_i^* S_{i'} S_{w_1'}(\psi_{j',k'})\rangle_{L^2(K_N,\mu)};$$

since $S_i(\psi)(w) = \chi_{Z(i)}(x) p_i^{-1/2} \psi(\sigma(x))$, one can check that

$$S_i^*(\psi)(w) = p_i^{1/2} \psi(iw).$$

It follows that $S_i^* S_{i'} = \delta_{i,i'} I$. If $i = i'$ so that $S_i^* S_{i'} = I$, we obtain:

$$\langle S_w(\psi_{j,k}), S_{w'}(\psi_{j',k'})\rangle_{L^2(K_N,\mu)} = \langle S_{w_1}(\psi_{j,k}), S_{w_1'}(\psi_{j',k'})\rangle_{L^2(K_N,\mu)},$$

which is equal to 0 by the induction hypothesis.

Thus, in either case,

$$\mathcal{W}_{n+2} \perp \Big[\bigoplus_{k=1}^{n+1} \mathcal{W}_k\Big].$$

Now suppose $\psi_{j,k} \in \mathcal{W}_0$, w is a word of length $n+2$, $k' \in \mathbb{Z}_N$ and $1 \le j' \le N-1$. Then,

$$\langle \psi_{j,k}, S_w(\psi_{j',k'})\rangle_{L^2(K_N,\mu)} = \langle \frac{1}{\sqrt{p_k}} \sum_{\ell=0}^{N-1} c_\ell^{j,k} \chi_{Z(k\ell)}, S_w(\psi_{j',k'})\rangle_{L^2(K_N,\mu)}.$$

Write $w = i_1 i_2 \cdots i_{n+1} i_{n+2}$. Then

$$S_w(\chi_{Z(k'\ell')}) = S_{i_1} S_{i_2} \ldots S_{i_{n+2}}(\chi_{Z(k'l')}) = \frac{1}{\prod_{v=1}^{n+2} \sqrt{p_{i_v}}} \chi_{Z(i_1 i_2 \cdots i_{n+1} i_{n+2} k'\ell')},$$

so that

$$\begin{aligned} &\langle \frac{1}{\sqrt{p_k}} \sum_{\ell=0}^{N-1} c_\ell^{j,k} \chi_{Z(k\ell)}, S_w(\psi_{j',k'})\rangle_{L^2(K_N,\mu)} \\ &= \frac{1}{\prod_{v=1}^{n+2} \sqrt{p_{i_v}}} \frac{1}{\sqrt{p_k p_{k'}}} \sum_{\ell=0}^{N-1} \sum_{\ell'=0}^{N-1} \langle c_\ell^{j,k} \chi_{Z(k\ell)},\ c_{\ell'}^{j',k'} \chi_{Z((i_1 i_2 \cdots i_{n+1} i_{n+2} k'\ell')}\rangle_{L^2(K_N,\mu)} \end{aligned}$$

$$= \frac{1}{\prod_{v=1}^{n+2}\sqrt{p_{i_v}}}\frac{1}{\sqrt{p_k p_{k'}}}\sum_{\ell=0}^{N-1}\sum_{\ell'=0}^{N-1}\delta_{k,i_1}\delta_{\ell,i_2}\langle c_\ell^{j,k}\chi_{Z(k\ell)},\ c_{\ell'}^{j',k'}\chi_{Z((i_1i_2\cdots i_{n+1}i_{n+2}k'\ell')}\rangle_{L^2(K_N,\mu)}$$

$$= \frac{1}{\prod_{v=1}^{n+2}\sqrt{p_{i_v}}}\delta_{k,i_1}\frac{1}{\sqrt{p_k p_{k'}}}\sum_{\ell'=0}^{N-1}\langle c_{i_2}^{j,k}\chi_{Z(ki_2)},\ c_{\ell'}^{j',k'}\chi_{Z(i_1i_2\cdots i_{n+1}i_{n+2}k'\ell')}\rangle_{L^2(K_N,\mu)}.$$

This quantity will only be nonzero if $k = i_1$; in this case we get:

$$\begin{aligned}
\langle\psi_{j,k}, S_w(\psi_{j',k'})\rangle_{L^2(K_N,\mu)} &= \frac{1}{\prod_{v=1}^{n+2}\sqrt{p_{i_v}}}\frac{1}{\sqrt{p_{i_1}p_{k'}}} \\
&\quad \sum_{\ell'=0}^{N-1}\langle c_{i_2}^{j,i_1}\chi_{Z(i_1i_2)},\ c_{\ell'}^{j',k'}\chi_{Z((i_1i_2\cdots i_{n+1}i_{n+2}k'\ell')}\rangle_{L^2(K_N,\mu)} \\
&= \frac{1}{\prod_{v=1}^{n+2}\sqrt{p_{i_v}}}\frac{1}{\sqrt{p_{i_1}p_{k'}}} \\
&\quad \sum_{\ell'=0}^{N-1}\int_{K_N} c_{i_2}^{j,i_1}\overline{c_{\ell'}^{j',k'}}\chi_{Z(i_1i_2)}(x)\chi_{Z((i_1i_2\cdots i_{n+1}i_{n+2}k'\ell')}(x)d\mu \\
&= \frac{1}{\prod_{v=1}^{n+2}\sqrt{p_{i_v}}}\sqrt{p_{k'}}\frac{1}{\sqrt{p_{i_1}}}\Big(\prod_{v=1}^{n+2}p_{i_v}\Big)c_{i_2}^{j,i_1}\sum_{\ell'=0}^{N-1}\overline{c_{\ell'}^{j',k'}}p_{\ell'} \\
&= \frac{\sqrt{p_{k'}}}{\sqrt{p_{i_1}}}c_{i_2}^{j,i_1}\Big(\prod_{v=1}^{n+2}\sqrt{p_{i_v}}\Big)\sum_{\ell'=0}^{N-1}1\cdot\overline{c_{\ell'}^{j',k'}}p_{\ell'} \ = \ 0.
\end{aligned}$$

So in all cases, $\langle\psi_{j,k}, S_w(\psi_{j',k'})\rangle_{L^2(K_N,\mu)} = 0$, and we have $\mathcal{W}_{n+2} \perp \mathcal{W}_0$.

Finally, we want to show that $\mathcal{W}_{n+2} \perp \mathcal{V}_0$. Let $\phi_k \in \mathcal{V}_0$ be fixed and let $S_w(\psi_{j',k'}) \in \mathcal{W}_{n+2}$ for $w = i_1i_2\cdots i_{n+2}$ a word of length $n+2$ and $k' \in \mathbb{Z}_N$, $j' \in \{1, 2, \cdots, N-1\}$. Then

$$\begin{aligned}
\langle\phi_k, S_w(\psi_{j',k'})\rangle_{L^2(K_N,\mu)} &= \frac{1}{\sqrt{p_k}} \\
&\quad \int_{K_N}\chi_{Z(k)}\overline{\Big(\prod_{v=1}^{n+2}\frac{1}{\sqrt{p_{i_v}}}\Big)\frac{1}{\sqrt{p_{k'}}}\sum_{\ell=0}^{N-1}c_\ell^{j',k'}\chi_{Z(i_1i_2\cdots i_{n+1}i_{n+2}k'\ell)}}d\mu \\
&= \frac{1}{\sqrt{p_k p_{k'}}}\Big(\prod_{v=1}^{n+2}\frac{1}{\sqrt{p_{i_v}}}\Big)\delta_{k,i_1}\int_{K_N}\overline{\sum_{\ell=0}^{N-1}c_\ell^{j',k'}\chi_{Z(i_1i_2\cdots i_{n+1}i_{n+2}k'\ell)}}d\mu \\
&= \frac{1}{\sqrt{p_k p_{k'}}}\Big(\prod_{v=1}^{n+2}\frac{1}{\sqrt{p_{i_v}}}\Big)\delta_{k,i_1}\Big[\prod_{v=1}^{n+2}p_{i_v}\Big]p_{k'}\sum_{\ell=0}^{N-1}\overline{c_\ell^{j',k'}}p_\ell
\end{aligned}$$

$$= \delta_{k,i_1} \frac{1}{\sqrt{p_k}} \sqrt{p_{k'}} \Big[\prod_{v=1}^{n+2} \sqrt{p_{i_v}} \Big] \sum_{\ell=0}^{N-1} 1 \cdot \overline{c_\ell^{j',k'}} p_\ell \;=\; 0.$$

It follows that $\mathcal{W}_{n+2} \perp \mathcal{V}_0$, and we have proved the desired result by induction. □

Remark 3.9 Notice that the proof of Theorem 3.8 also extends to any other measure with shift operators having constant Radon-Nykodym derivative on cylinder sets.

4 C^*-Algebras Corresponding to Directed Graphs and Higher-Rank Graphs

4.1 Directed Graphs, Higher-Rank Graphs, and C^*-Algebras

A **directed graph** E consists of a countable collection of vertices E^0 and edges E^1 with range and source maps $r, s : E^1 \to E^0$. We view an edge e as being directed from its source $s(e)$ to its range $r(e)$. A **path** is a string of edges $e_1 e_2 \dots e_n$ where $s(e_i) = r(e_{i+1})$ for $i = 1, 2, \dots, n-1$. The **length** of a path is the number of edges in the string. As mentioned in the introduction, the graph C^*-algebra $C^*(E)$ is the universal C^*-algebra generated by a set of projections $\{p_v : v \in E^0\}$ and a set of partial isometries $\{s_e : e \in E^1\}$ that satisfy the Cuntz-Krieger relations. (These are relations (CK1)–(CK4) in (21) below).

Higher-rank graphs, also called k-graphs, are higher-dimensional analogues of directed graphs. By definition, a higher-rank graph is a small category Λ with a functor d from the set Λ of morphisms to $\mathbb{N}^k$ satisfying the **factorization property** : if $d(\lambda) = m + n$, then there exist unique $\alpha, \beta \in \Lambda$ such that $d(\alpha) = m$, $d(\beta) = n$, and $\lambda = \alpha\beta$.[2] Note that we write $\{e_1, \dots, e_k\}$ for the standard basis of $\mathbb{N}^k$. We often call a morphism $\lambda \in \Lambda$ a **path** (or an element) in Λ, and call $\Lambda^0 := d^{-1}(0)$ the set of **vertices** of Λ; then the factorization property gives us **range** and **source** maps $r, s : \Lambda \to \Lambda^0$. For $v, w \in \Lambda^0$ and $n \in \mathbb{N}^k$, we write

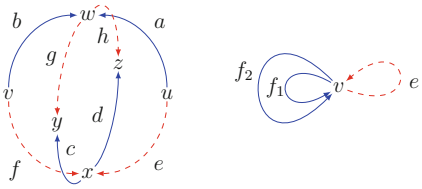
$$v\Lambda^n w := \{\lambda \in \Lambda : d(\lambda) = n, r(\lambda) = v, s(\lambda) = w\}.$$

Thus $\lambda \in v\Lambda^n w$ means that λ, is a path that starts at w, ends at v, and has shape n. Given two paths $\lambda, \nu \in \Lambda$, we can think of λ as a k-cube in a k-colored graph as in the next example.

Example 4.1 Consider the following two 2-colored graphs Γ_1 on the left and Γ_2 on the right. In both graphs, the dashed edges are red and the solid edges are blue. (The sphere-like 2-graph picture below is taken from [36] and we would like to thank

[2] We think of $\mathbb{N}^k$ as a category with one object, namely 0, and with composition of morphisms given by addition.

them for sharing their picture). We will explain how Γ_1, Γ_2 give rise to 2-graphs Λ_i; the degree functor $d : \Lambda_i \to \mathbb{N}^2$ will count the number of red and blue edges in a path $\lambda \in \Lambda_i$.

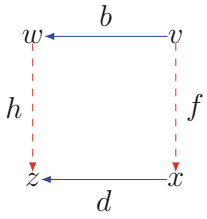


Depending on the choice of factorization rules, these 2-colored graphs can give rise to several different 2-graphs.

There is only one 2-graph Λ_1 with the 2-colored graph Γ_1; the factorization rules of Λ_1 are given by

$$hb = df, \quad ha = de, \quad gb = cf, \quad \text{and} \quad ga = ce.$$

Note that the path hb has degree $e_1 + e_2 = (1, 1) \in \mathbb{N}^2$. The factorization rule $hb = df$ means that the element $hb = df$ of Λ_1 can be understood as the following square; the 2-graph Λ_1 has four such squares (paths, or elements).



However, on Γ_2, there are two 2-graphs Λ_2 and Λ_3 associated to the 2-colored graph Γ_2. The factorization rules for Λ_2 are given by

$$f_1 e = e f_1 \quad \text{and} \quad f_2 e = e f_2.$$

The factorization rules for Λ_3 are given by

$$f_1 e = e f_2 \quad \text{and} \quad f_2 e = e f_1.$$

We leave it to the reader to check that both choices of factorization rules give rise to a well-defined functor $d : \Lambda_i \to \mathbb{N}^2$ satisfying the factorization property, where $d(\lambda) = (m, n)$ implies that the path λ contains m red edges and n blue edges.

We say that Λ is **finite** if Λ^n is finite for all $n \in \mathbb{N}^k$ and is **strongly connected** if $v\Lambda w \neq \emptyset$ for all $v, w \in \Lambda^0$. We say that k-graph has **no sources** if $v\Lambda^{e_i} \neq \emptyset$ for all $v \in \Lambda^0$ and for all $1 \leq i \leq k$. Note that we only consider finite k-graphs with no

sources in this section. Define an **infinite path** in Λ to be a morphism from Ω_k to Λ. To be more precise, consider the set

$$\Omega_k := \{(p,q) \in \mathbb{N}^k \times \mathbb{N}^k : p \leq q\}.$$

Then Ω_k is a k-graph with $\Omega_k^0 = \mathbb{N}^k$; the range and source maps $r, s : \Omega_k \to \mathbb{N}^k$ given by $r(p,q) := p$ and $s(p,q) := q$; and the degree functor d given by $d(p,q) = q - p$. Note that the composition is given by $(p,q)(q,m) = (p,m)$ and Ω_k has no sources. An infinite path in a k-graph Λ is a k-graph morphism $x : \Omega_k \to \Lambda$ and the infinite path space Λ^∞ is the collection of all infinite paths. The space Λ^∞ is equipped with a compact open topology generated by the cylinder sets $\{Z(\lambda) : \lambda \in \Lambda\}$, where

$$Z(\lambda) = \{x \in \Lambda^\infty : x(0, d(\lambda)) = \lambda\}.$$

For $p \in \mathbb{N}^k$, there is a shift map σ^p on Λ^∞ given by $\sigma^p(x)(m,n) = x(m+p, n+p)$ for $x \in \Lambda^\infty$. For more details on the above constructions, see Section 2 of [35].

For each $1 \leq i \leq k$, we write A_i for the **vertex matrices** for Λ, where the entries $A_i(v,w)$ are the number of paths from w to v with degree e_i. Because of the factorization property, the vertex matrices A_i commute, and if Λ is strongly connected, Lemma 4.1 of [26] establishes that there is a unique positive normalized Perron-Frobenius eigenvector for the matrices A_i. The Perron-Frobenius eigenvector x^Λ is the unique vector $x^\Lambda \in (0,\infty)^{|\Lambda^0|}$ with ℓ^1-norm 1 which is a common eigenvector of the matrices A_i. It is well known now (see [26] Theorem 8.1) that for a strongly connected finite k-graph Λ, there is a unique Borel probability measure M on Λ^∞, called the Perron-Frobenius measure, such that

$$M(Z(\lambda)) = \rho(\Lambda)^{-d(\lambda)} x^\Lambda_{s(\lambda)} \quad \text{for all} \quad \lambda \in \Lambda, \tag{20}$$

where $\rho(\Lambda) = (\rho(A_1), \ldots, \rho(A_k))$. See [26] for the construction of the measure M.

For a finite k-graph with no sources, the Cuntz-Krieger C^*-algebra $C^*(\Lambda)$, often called a k-graph C^*-algebra, is a universal C^*-algebra generated by a collection of partial isometries $\{t_\lambda : \lambda \in \Lambda\}$ satisfying the following Cuntz-Krieger relations:

$$\begin{aligned}
&\text{(CK1)}\{t_v : v \in \Lambda^0\} \text{ is a family of mutually orthogonal projections,}\\
&\text{(CK2)} t_\mu t_\lambda = t_{\mu\lambda} \text{ whenever } s(\mu) = r(\lambda),\\
&\text{(CK3)} t_\mu^* t_\mu = t_{s(\mu)} \text{ for all } \mu, \text{ and}\\
&\text{(CK4) for all } v \in \Lambda^0 \text{ and } n \in \mathbb{N}^k, \text{ we have} \quad t_v = \sum_{\lambda \in v\Lambda^n} t_\lambda t_\lambda^*.
\end{aligned} \tag{21}$$

Also we can show that

$$C^*(\Lambda) = \overline{\text{span}}\{t_\lambda t_\nu^* : \lambda, \nu \in \Lambda, s(\lambda) = s(\nu)\}.$$

4.2 Λ-Semibranching Function Systems and Representations of $C^*(\Lambda)$

We briefly review the definition of a Λ-semibranching function system given in [20], then discuss the recent results in [20].

Compare the following definition with the definition of a semibranching function system given in Definition 3.2.

Definition 4.2 Let Λ be a finite k-graph and let (X, μ) be a measure space. A **Λ-semibranching function system** on (X, μ) is a collection $\{D_\lambda\}_{\lambda\in\Lambda}$ of measurable subsets of X, together with a family of **prefixing maps** $\{\tau_\lambda : D_\lambda \to X\}_{\lambda\in\Lambda}$, and a family of coding maps $\{\tau^m : X \to X\}_{m\in\mathbb{N}^k}$, such that

(a) For each $m \in \mathbb{N}^k$, the family $\{\tau_\lambda : d(\lambda) = m\}$ is a semibranching function system, with coding map τ^m.
(b) If $v \in \Lambda^0$, then $\tau_v = id$, and $\mu(D_v) > 0$.
(c) Let $R_\lambda = \tau_\lambda D_\lambda$. For each $\lambda \in \Lambda, \nu \in s(\lambda)\Lambda$, we have $R_\nu \subseteq D_\lambda$ (up to a set of measure 0), and

$$\tau_\lambda \tau_\nu = \tau_{\lambda\nu} \text{a.e.}$$

(Note that this implies that up to a set of measure 0, $D_{\lambda\nu} = D_\nu$ whenever $s(\lambda) = r(\nu)$).
(d) The coding maps satisfy $\tau^m \circ \tau^n = \tau^{m+n}$ for any $m, n \in \mathbb{N}^k$. (Note that this implies that the coding maps pairwise commute.)

Remark 4.3 (1) The key condition of a Λ-semibranching function system is the condition (c). The immediate consequence is that $D_\lambda = D_{s(\lambda)}$ and $R_\lambda \subset R_{r(\lambda)}$ for all $\lambda \in \Lambda$. Also for $\lambda, \nu \in \Lambda$, if $s(\lambda) = r(\nu)$, then $x \in R_{\lambda\nu}$ if and only if $x \in R_\lambda$ and $\tau^{d(\lambda)}(x) \in R_\nu$.

(2) When E is a finite directed graph, the definition of an E-semibranching function system in Definition 4.2 is not equivalent to the semibranching function system of E in Definition 3.2. First of all, the set of domains $\{D_e : e \in E\}$ in Definition 3.2 neither have to be mutually disjoint nor the union to be the whole space X up to a set of measure zero. But since Definition 4.2(b) requires that $D_v = R_v$ for $v \in E^0$, the condition (a) of Definition 4.2 implies that $\mu(D_v \cap D_w) = \mu(R_v \cap R_w) = 0$ for $v \neq w$, and $\mu(X \setminus \bigcup_{v\in E^0} R_v) = \mu(X \setminus \bigcup_{v\in E^0} D_v) = 0$. As seen in Remark 4.3, $D_e = D_{s(e)}$ for any $e \in E$, and hence $\mu(D_e \cap D_f) = 0$ if $s(e) \neq s(f)$.

(3) It turned out that the conditions of Definition 4.2 are a lot stronger than what we expected. In particular, when we have a finite directed graph E, the conditions of Definition 4.2 imply what is called condition (C-K) in [6]:

$$D_v = \bigcup_{e\in E^m} R_\lambda \quad \text{for all } v \in E^0 \text{ and } m \in \mathbb{N}$$

up to a measure zero set. The condition (C-K) was assumed in Theorem 2.22 in [6] to obtain a representation of $C^*(\Lambda)$ on $L^2(X,\mu)$, where a semibranching function system is given on the measure space (X,μ).

Example 4.4 Let Λ be a strongly connected finite k-graph. As seen before, there is a Borel probability measure M on Λ^∞ given by the formula of (20). To construct a Λ-semibranching function system on (Λ^∞, M), we define, for $\lambda \in \Lambda$, prefixing maps $\sigma_\lambda : Z(s(\lambda)) \to Z(\lambda)$ by

$$\sigma_\lambda(x) = \lambda x,$$

where we denote by $y := \lambda x$ the unique infinite path $y : \Omega_k \to \Lambda$ such that $y((0, d(\lambda))) = \lambda$ and $\sigma^{d(\lambda)}(y) = x$.

For $m \in \mathbb{N}^k$ we define the coding maps $\sigma^m : \Lambda^\infty \to \Lambda^\infty$ by

$$\sigma^m(x) = x(m, \infty).$$

Then $\{\sigma_\lambda\}_{\lambda\in\Lambda}$ with $\{\sigma^m\}_{m\in\mathbb{N}^k}$ form a Λ-semibranching function system on (Λ^∞, M) as shown in Proposition 3.4 of [20].

When a k-graph Λ is finite and has no sources, one of the main theorems of [20], Theorem 3.5, says that the operators S_λ associated to a Λ-semibranching function system on a measure space (X,μ) given by

$$S_\lambda \xi(x) := \chi_{R_\lambda}(x)(\Phi_{\tau_\lambda}(\tau^{d(\lambda)}(x)))^{-1/2}\xi(\tau^{d(\lambda)}(x)) \tag{22}$$

generate a representation of $C^*(\Lambda)$ on $L^2(X,\mu)$, where

$$\Phi_{\tau_\lambda} = \frac{d(\mu \circ \tau_\lambda)}{d\mu}$$

is positive a.e. on D_λ.

In addition, if we have a strongly connected finite k-graph Λ, then the Λ-semibranching function system of Example 4.4 on the Borel probability measure space (Λ^∞, M) gives rise to a representation of $C^*(\Lambda)$ on $L^2(\Lambda^\infty, M)$ which is faithful if and only if Λ is aperiodic. (See Theorem 3.6 of [20]).

Moreover, if the vertex matrices A_i associated to a strongly connected finite k-graph Λ are all $\{0,1\}$-matrices, then we can construct Λ-semibranching function systems on a fractal subspace X of $[0,1]$. In particular, let $N = |\Lambda^0|$ and label the vertices of Λ by the integers, $0, 1, \ldots, N-1$. Let $\rho(A)$ denote the spectral radius of the product $A := A_1 \ldots A_k$. Then consider the embedding $\Psi : \Lambda^\infty \to [0,1]$ given by interpreting the sequence of vertices of a given infinite path as an N-adic decimal. Then $X = \Psi(\Lambda^\infty)$ is a Cantor-type fractal subspace of $[0,1]$ and the Hausdorff measure μ on X is given by the Borel probability measure M on Λ^∞ via Ψ. The prefixing maps $\{\tau_\lambda\}$ and coding maps $\{\tau^{d(\lambda)}\}$ on (X,μ) are induced from the prefixing maps $\{\sigma_\lambda\}$ and coding maps $\{\sigma^m\}$ on (Λ^∞, M) given in Example 4.4.

Moreover, if s denotes the Hausdorff dimension of X, we have

$$N^{ks} = \rho(A), \quad \text{and} \quad s = \frac{1}{k}\frac{\ln \rho(A)}{\ln N}.$$

See Section 3.2 of [20] for further details.

5 Wavelets on $L^2(\Lambda^\infty, M)$

Let Λ be a strongly connected finite k-graph. As seen in the previous section, there is a Borel probability measure M on the infinite path space Λ^∞ given by, for $\lambda \in \Lambda$,

$$M(Z(\lambda)) = \rho(\Lambda)^{-d(\lambda)} x^\Lambda_{s(\lambda)},$$

where $\rho(\Lambda) = (\rho(A_i))_{1\le i\le k}$ and x^Λ is the unimodular Perron-Frobenius eigenvector of Λ. We now proceed to generalize the wavelet decomposition of $L^2(\Lambda^\infty, M)$ that we constructed in Section 4 of [20]. In that paper, we built an orthonormal decomposition of $L^2(\Lambda^\infty, M)$, which we termed a **wavelet decomposition**, following Section 3 of [41]. Here, our wavelet decomposition is constructed by applying (some of) the operators S_λ of Example 4.4 and Equation (22) to a basic family of functions in $L^2(\Lambda^\infty, M)$. Instead of choosing the finite paths λ whose degrees are associated to k-cubes, we will construct them from isometries given by paths whose degrees are given by k-rectangles. One way to interpret our main result below (Theorem 5.2) is to say that for any rectangle $(j_1, j_2, \ldots, j_k) \in \mathbb{N}^k$ with no zero entries, the cofinal set $\{n \cdot (j_1, j_2, \ldots, j_k) : n \in \mathbb{N}\} \subseteq \mathbb{N}^k$ gives rise to an orthonormal decomposition of $L^2(\Lambda^\infty, M)$.

While we can use the same procedure to obtain a family of orthonormal functions in $L^2(X, \mu)$ whenever we have a Λ-semibranching function system on (X, μ), we cannot establish in general that this orthonormal decomposition densely spans $L^2(X, \mu)$ – we have no analogue of Lemma 5.1 for general Λ-semibranching function systems. Moreover, by Corollary 3.12 of [20], every Λ-semibranching function system on Λ^∞ with constant Radon-Nikodym derivative is endowed with the Perron-Frobenius measure M. Thus, in this section, we restrict ourselves to the case of (Λ^∞, M). We also note that our proofs in this section follow the same ideas found in the proof of Theorem 3.2 of [41].

For a path $\lambda \in \Lambda$, let Θ_λ denote the characteristic function of $Z(\lambda) \subseteq \Lambda^\infty$. Recall that M is the unique Borel probability measure on Λ^∞ satisfying our desired properties. For the rest of this section, we fix a k-tuple

$$(j_1, j_2, \cdots, j_k) \in \mathbb{N}^k$$

all of whose coordinates are *positive* integers.

Lemma 5.1 *Let Λ be a strongly connected k-graph and fix $J = (j_1, j_2, \ldots, j_k) \in (\mathbb{Z}^+)^k$. Then the span of the set*

$$S^J := \{\Theta_\lambda : d(\lambda) = (n \cdot j_1, n \cdot j_2, \ldots, n \cdot j_k) \text{ for some } n \in \mathbb{N}\}$$

is dense in $L^2(\Lambda^\infty, M)$.

Proof Let $\mu \in \Lambda$. We will show that we can write Θ_μ as a linear combination of functions from S^J.

Suppose $d(\mu) = (m_1, \ldots, m_k)$. Let $m = \min\{N > 0 : N \cdot j_i - m_i \geq 0 \text{ for } 1 \leq i \leq k\}$, and let $n = (m \cdot j_1, m \cdot j_2, \ldots, m \cdot j_k) - d(\mu)$. Let

$$C_\mu = \{\lambda \in \Lambda : r(\lambda) = s(\mu), d(\lambda) = n\}.$$

In words, C_μ consists of the paths that we could append to μ such that $\mu\lambda \in S^J$: if $\lambda \in C_\mu$ then the product $\mu\lambda$ is defined and

$$d(\mu\lambda) = d(\mu) + d(\lambda) = (m \cdot j_1, m \cdot j_2, \ldots, m \cdot j_k).$$

Similarly, since $d(\mu\lambda) = d(\mu\lambda') = (m \cdot j_1, \ldots, m \cdot j_k) = mJ$, if $x \in Z(\mu\lambda) \cap Z(\mu\lambda')$ then the fact that $x(0, mJ)$ is well defined implies that

$$x(0, mJ) = \mu\lambda = \mu\lambda' \Rightarrow \lambda = \lambda'.$$

It follows that if $\lambda \neq \lambda' \in C_\mu$, then $Z(\mu\lambda) \cap Z(\mu\lambda') = \emptyset$. Since every infinite path $x \in Z(\mu)$ has a well-defined "first segment" of shape $(m \cdot j_1, \ldots, m \cdot j_k)$ – namely $x(0, mJ)$ – every $x \in Z(\mu)$ must live in $Z(\mu\lambda)$ for precisely one $\lambda \in C_\mu$. Thus, we can write $Z(\mu)$ as a disjoint union,

$$Z(\mu) = \bigsqcup_{\lambda \in C_\mu} Z(\mu\lambda).$$

It follows that $\Theta_\mu = \sum_{\lambda \in C_\mu} \Theta_{\mu\lambda}$, so the span of functions in S^J includes the characteristic functions of cylinder sets. Since the cylinder sets $Z(\mu)$ form a basis for the topology on Λ^∞ with respect to which M is a Borel measure, it follows that the span of S^J is dense in $L^2(\Lambda^\infty, M)$ as claimed. □

Since the span of the functions in S^J is dense in $L^2(\Lambda^\infty, M)$, we will show how to decompose $\overline{\text{span}}\, S^J$ as an orthogonal direct sum,

$$\overline{\text{span}}\, S^J = \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^{\infty} \mathcal{W}_{j,\Lambda}^J,$$

where $\mathcal{V}_{0,\Lambda}$ will be equal to the subspace spanned by the functions $\{\Theta_v : v \in \Lambda^0\}$. We then will construct $\mathcal{W}^J_{j,\Lambda}$ for each $j > 1$ from the functions in $\mathcal{W}^J_{0,\Lambda}$ and (some of) the operators S_λ discussed in Section 3 of [20]. The construction of $\mathcal{W}^J_{0,\Lambda}$ generalizes that given in Section 4 of [20], which in turn was similar to that given in Section 3 of [41] for the case of a directed graph.

We recall from [20] that the functions $\{\Theta_v : v \in \Lambda^0\}$ form an orthogonal set in $L^2(\Lambda^\infty, M)$, whose span includes those functions that are constant on Λ^∞:

$$\int_{\Lambda^\infty} \Theta_v \overline{\Theta_w}\, dM = \delta_{v,w} M(Z(v)) = \delta_{v,w} x_v^\Lambda,$$

and

$$\sum_{v\in\Lambda^0} \Theta_v(x) \equiv 1.$$

Thus, the set $\{\frac{1}{\sqrt{x_v^\Lambda}}\Theta_v : v \in \Lambda^0\}$ is an orthonormal set in S^J. We define

$$\mathcal{V}_{0,\Lambda} := \overline{\text{span}}\{\frac{1}{\sqrt{x_v^\Lambda}}\Theta_v : v \in \Lambda^0\}.$$

To construct $\mathcal{W}^J_{0,\Lambda}$, let $v \in \Lambda^0$ be arbitrary. Let

$$D_v^J = \{\lambda \in \Lambda : d(\lambda) = J \text{ and } r(\lambda) = v\},$$

and write d_v^J for $|D_v^J|$ (note that by our hypothesis that Λ is a finite k-graph we have $d_v^J < \infty$).

Define an inner product on $\mathbb{C}^{d_v^J}$ by

$$\langle \vec{v}, \vec{w}\rangle = \sum_{\lambda\in D_v^J} \overline{v_\lambda} w_\lambda \rho(\Lambda)^{(-j_1,\ldots,-j_k)} x^\Lambda_{s(\lambda)}, \tag{23}$$

and let $\{c^{m,v}\}_{m=1}^{d_v^J-1}$ be an orthonormal basis for the orthogonal complement of $(1,\ldots,1) \in \mathbb{C}^{d_v^J}$ with respect to this inner product. Let $c^{0,v}$ be the unique vector of norm one with respect to this inner product with (equal) positive entries that is a multiple of $(1,\ldots,1) \in \mathbb{C}^{d_v^J}$. Thus, $\{c^{m,v}\}_{m=0}^{d_v-1}$ is an orthonormal basis for $\mathbb{C}^{d_v^J}$.

We explain the importance of $(1,\ldots,1) \in \mathbb{C}^{d_v^J}$ further. We index the λ's in D_v^J :

$$D_v^J = \{\lambda_1, \lambda_2, \cdots \lambda_{d_v^J}\}.$$

We need to stress here that

$$\sum_{j=1}^{d_v^J} \Theta_{\lambda_j} = \Theta_v.$$

In this way, we have identified Θ_v with $(1, 1, \cdots, 1) \in \mathbb{C}^{d_v^J}$. (When we do this, we identify Θ_{λ_1} with $(1, 0, 0, \cdots, 0)$, Θ_{λ_2} with $(0, 1, 0, \cdots, 0)$, and $\Theta_{\lambda_{d_v^J}}$ with $(0, 0, 0, \cdots, 1) \in \mathbb{C}^{d_v^J}$.)

Now, for each pair (m, v) with $0 \le m \le d_v^J - 1$ and v a vertex in Λ^0, define

$$f^{m,v} = \sum_{\lambda \in D_v^J} c_\lambda^{m,v} \Theta_\lambda.$$

Note that by our definition of the measure M on Λ^∞, since for $1 \le m \le d_v^J - 1$, the vectors $c^{m,v}$ are orthogonal to $(1, \ldots, 1)$ in the inner product (23), we have

$$\begin{aligned} \int_{\Lambda^\infty} f^{m,v}\, dM &= \sum_{\lambda \in D_v^J} c_\lambda^{m,v} M(Z(\lambda)) \\ &= \sum_{\lambda \in D_v^J} c_\lambda^{m,v} \rho(\Lambda)^{(-j_1,\ldots,j_k)} x^\Lambda_{s(\lambda)} \\ &= 0 \end{aligned}$$

for each (m, v) with $m \ge 1$. On the other hand, if $m = 0$, it is easy to see that

$$f^{0,v} = \sum_{\lambda \in D_v^J} c_\lambda^{0,v} \Theta_\lambda$$

is a constant multiple of Θ_v, since $c_\lambda^{0,v} = c_{\lambda'}^{0,v}$ for $\lambda,\ \lambda' \in D_v^J$, and $\sum_{\lambda \in D_v^J} \Theta_\lambda = \Theta_v$. Moreover, the arguments of Lemma 5.1 tell us that $\Theta_\lambda \Theta_{\lambda'} = \delta_{\lambda,\lambda'} \Theta_\lambda$ for any λ, λ' with $d(\lambda) = d(\lambda') = (j_1, \ldots, j_k)$. Consequently, if $\lambda \in D_v^J, \lambda' \in D_{v'}^J$ for $v \ne v'$, we have $\Theta_\lambda \Theta_{\lambda'} = 0$. It follows that

$$\begin{aligned} \int_{\Lambda^\infty} f^{m,v} \overline{f^{m',v'}}\, dM &= \delta_{v,v'} \sum_{\lambda \in D_v^J} c_\lambda^{m,v} \overline{c_\lambda^{m',v}} M(Z(\lambda)) \\ &= \delta_{v,v'} \delta_{m,m'} \end{aligned}$$

since the vectors $\{c^{m,v}\}$ form an orthonormal set with respect to the inner product (23). Thus, the functions $\{f^{m,v}\}$ are an orthonormal set in $L^2(\Lambda^\infty, M)$. We define

$$\mathcal{W}_{0,\Lambda}^J := \overline{\text{span}}\{f^{m,v} : v \in \Lambda^0, 1 \le m \le d_v^J - 1\}.$$

Note that $\mathcal{V}_{0,\Lambda}$ is orthogonal to $\mathcal{W}_{0,\Lambda}^J$. To see this, let $g \in \mathcal{V}_{0,\Lambda}$ be arbitrary, so $g = \sum_{v \in \Lambda^0} g_v \Theta_v$ with $g_v \in \mathbb{C}$ for all v. Then

$$\int_{\Lambda^\infty} \overline{f^{m,v'}(x)} g(x)\, dM = \sum_{v \in V_0} \delta_{v',v}\, g_v \sum_{\lambda \in D_{v'}^J} \overline{c_\lambda^{m,v'}} M(Z(\lambda))$$

$$= 0,$$

since $\sum_{\lambda \in D_v^J} c_\lambda^{m,v} M(Z(\lambda)) = 0$ for all fixed v, and $1 \le m \le d_v^J - 1$. Thus, g is orthogonal to every basis element $f^{m,v}$ of $\mathcal{W}_{0,\Lambda}^J$.

The basis $\{f^{m,v} : v \in \Lambda^0, 1 \le m \le d_v^J - 1\}$ for $\mathcal{W}_{0,\Lambda}^J$ generalizes the analogue for k-graphs of the **graph wavelets** of [41], as described in Section 4 of [20]. As the following Theorem shows, by shifting these functions using the operators

$$\{S_\lambda : d(\lambda) = nJ \text{ for some } n \in \mathbb{N}\},$$

we obtain an orthonormal basis for $L^2(\Lambda^\infty, M)$. Thus, each $J \in (\mathbb{Z}^+)^k$ gives a different family of k-graph wavelets associated to the representation of $C^*(\Lambda)$ described in Theorem 3.5 of [20].

Theorem 5.2 (*Compare to Theorem 4.2 of [20]*) *Let Λ be a strongly connected finite k-graph and fix $J \in (\mathbb{Z}^+)^k$. For each fixed $j \in \mathbb{N}^+$ and $v \in \Lambda^0$, let*

$$C_{j,v}^J := \{\lambda \in \Lambda : s(\lambda) = v, d(\lambda) = jJ\},$$

and let S_λ be the operator on $L^2(\Lambda^\infty, M)$ described in Theorem 3.5 of [20]; for $\xi \in L^2(\Lambda^\infty, M)$,

$$S_\lambda \xi(x) = \Theta_\lambda(x) \rho(\Lambda)^{d(\lambda)/2} \xi(\sigma^{d(\lambda)}(x)).$$

Then

$$\{S_\lambda f^{m,v} : v \in \Lambda^0,\ \lambda \in C_{j,v}^J,\ 1 \le m \le d_v^J - 1\}$$

is an orthonormal set. Moreover, if $\lambda \in C_{j,v}^J$, $\mu \in C_{i,v'}^J$ for $0 < i < j$, we have

$$\int_{\Lambda^\infty} S_\lambda f^{m,v} \overline{S_\mu f^{m',v'}}\, dM = 0 \quad \text{for } 1 \le m, m' \le d_v^J - 1.$$

It follows that defining

$$\mathcal{W}_{j,\Lambda}^J := \overline{\text{span}}\{S_\lambda f^{m,v} : v \in \Lambda^0, \lambda \in C_{j,v}^J, 1 \le m \le d_v^J - 1\},$$

for $j \ge 1$, we obtain an orthonormal decomposition

$$L^2(\Lambda^\infty, M) = \overline{\text{span}}\ S^J = \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^{\infty} \mathcal{W}_{j,\Lambda}^J.$$

Proof We first observe that if $s(\lambda) = v$, then

$$S_\lambda f^{m,v} = \sum_{\mu \in D_v^J} c_\mu^{m,v} \rho(\Lambda)^{d(\lambda)/2} \Theta_{\lambda\mu},$$

because the Radon-Nikodym derivatives Φ_{σ_λ} are constant on $Z(s(\lambda))$ for each $\lambda \in \Lambda$, thanks to Proposition 3.4 of [20]. In particular, if $d(\lambda) = 0$, then $S_\lambda f^{m,v} = f^{m,v}$. Thus, if $d(\lambda) = d(\lambda') = (j \cdot j_1, \ldots, j \cdot j_k)$, the factorization property and the fact that $d(\lambda\mu) = d(\lambda'\mu') = ((j+1) \cdot j_1, \ldots, (j+1) \cdot j_k)$ for every $\mu \in D^J_{s(\lambda)}, \mu' \in D^J_{s(\lambda')}$ implies that

$$\Theta_{\lambda\mu}\Theta_{\lambda'\mu'} = \delta_{\lambda,\lambda'}\delta_{\mu,\mu'} \quad \text{for all } \mu \in D^J_{s(\lambda)}, \mu' \in D^J_{s(\lambda')}.$$

In particular, $S_\lambda f^{m,v} \overline{S_{\lambda'} f^{m',v'}} = 0$ unless $\lambda = \lambda'$ (and hence $v = v'$). Moreover,

$$\begin{aligned}
\int_{\Lambda^\infty} S_\lambda f^{m,v} \overline{S_\lambda f^{m',v}}\, dM &= \sum_{\mu \in D_v^J} c_\mu^{m,v} \overline{c_\mu^{m',v}} \rho(\Lambda)^{d(\lambda)} M(Z(\lambda\mu)) \\
&= \sum_{\mu \in D_v^J} c_\mu^{m,v} \overline{c_\mu^{m',v}} \rho(\Lambda)^{-d(\mu)} x^\Lambda_{s(\mu)} \\
&= \delta_{m,m'},
\end{aligned}$$

by the definition of the vectors $c_\mu^{m,v}$, since $d(\mu) = (j_1, j_2, \ldots, j_k)$ for each $\mu \in D_v^J$.

Now, suppose $\lambda \in C^J_{1,v}$. Observe that $S_\lambda f^{m,v} \overline{f^{m',v'}}$ is nonzero only when $v' = r(\lambda)$, and also that

$$(S_\lambda f^{m,v})(x)\overline{f^{m',v'}}(x) = \sum_{\mu \in D_v^J} c_\mu^{m,v} \rho(\Lambda)^{d(\lambda)/2} \Theta_{\lambda\mu}(x) \sum_{\mu' \in D^J_{v'}} \overline{c_{\mu'}^{m',v'}} \Theta_{\mu'}(x).$$

Note that $\Theta_{\lambda\mu}(x)\Theta_{\mu'}(x) \neq 0$ if $x = \lambda\mu y = \mu' y'$ for some $y, y' \in \Lambda^\infty$, and $\lambda \in C^J_{1,v}$ implies $d(\lambda) = J = d(\mu')$. So the factorization property implies that $\mu' = \lambda$, and hence we obtain

$$\begin{aligned}
\int_{\Lambda^\infty} S_\lambda f^{m,v} \overline{f^{m',v'}}\, dM &= \sum_{\mu \in D_v^J} c_\mu^{m,v} \overline{c_\lambda^{m',v'}} \rho(\Lambda)^{d(\lambda)/2} M(Z(\lambda\mu)) \\
&= \overline{c_\lambda^{m',v'}} \rho(\Lambda)^{-d(\lambda)/2} \sum_{\mu \in D_v^J} c_\mu^{m,v} \rho(\Lambda)^{-d(\mu)} x^\Lambda_{s(\mu)} \\
&= 0.
\end{aligned}$$

Thus, $\mathcal{W}^J_{0,\Lambda}$ is orthogonal to $\mathcal{W}^J_{1,\Lambda}$.

In more generality, suppose that $\lambda \in C^J_{j,v}, \lambda' \in C^J_{i,v'}$, $j > i \geq 1$. We observe that $S_\lambda f^{m,v}\overline{S_{\lambda'} f^{m',v'}}$ is nonzero only when $\lambda = \lambda' \nu$ with $\nu \in C^J_{j-i,v}$, so we have

$$S_\lambda f^{m,v}\overline{S_{\lambda'} f^{m',v'}} = S_{\lambda'}(S_\nu f^{m,v})\overline{S_{\lambda'} f^{m',v'}}.$$

Consequently,

$$\begin{aligned}
\int_{\Lambda^\infty} S_\lambda f^{m,v}\overline{S_{\lambda'} f^{m',v'}}\, dM &= \int_{\Lambda^\infty} S_{\lambda'}(S_\nu f^{m,v})\overline{S_{\lambda'} f^{m',v'}}\, dM \\
&= \int_{\Lambda^\infty} (S_\nu f^{m,v})\overline{S^*_{\lambda'} S_{\lambda'} f^{m',v'}}\, dM \\
&= \int_{\Lambda^\infty} (S_\nu f^{m,v})\overline{f^{m',v'}}\, dM \\
&= \sum_{\mu \in D^J_v} c^{m,v}_\mu \overline{c^{m',v'}_\nu}\, \rho(\Lambda)^{d(\nu)/2} M(Z(\nu\mu)) \\
&= \overline{c^{m',v'}_\nu}\, \rho(\Lambda)^{-d(\nu)/2} \sum_{\mu \in D^J_v} c^{m,v}_\mu \rho(\Lambda)^{-d(\mu)} x^\Lambda_{s(\mu)} \\
&= 0.
\end{aligned}$$

Thus, the sets $\mathcal{W}^J_{j,\Lambda}$ are mutually orthogonal as claimed.

We now need to show that $L^2(\Lambda^\infty, M) = \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^\infty \mathcal{W}^J_{j,\Lambda}$. We will do this by showing that

$$S^J \subset \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^{\infty} \mathcal{W}^J_{j,\Lambda}.$$

We first note that if $\lambda \in \Lambda$ and $d(\lambda) = (j_1, j_2, \cdots, j_k)$, then $\Theta_\lambda \in \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^\infty \mathcal{W}^J_{j,\Lambda}$. Let $r(\lambda) = v$, so that $\lambda \in D^J_v$. Write $\lambda = \lambda_i$ for some specific $i \in \{1, 2, \cdots, d^J_v\}$. We identify Θ_{λ_i} with $(0, 0, \ldots, 1(\text{in } i_{th} \text{ spot}), 0, 0, \ldots, 0) = e_i \in \mathbb{C}^{d^J_v}$.

As we observed above, identifying Θ_{λ_i} with e_i induces an isomorphism between the (finite-dimensional) Hilbert spaces

$$\text{span}\{\Theta_{\lambda_1},\ \Theta_{\lambda_2},\ \cdots, \Theta_{\lambda_{d^J_v}}\} \subset L^2(\Lambda^\infty, M)$$

and $\mathbb{C}^{d^J_v}$ equipped with the inner product (23). By using this isomorphism, we can identify the function $f^{m,v} = \sum_{i=1}^{d^J_v} c^{m,v}_{\lambda_i} \Theta_{\lambda_i}$, with the vector $(c^{m,k}_{\lambda_i})_i \in \mathbb{C}^{d^J_v}$. This identification allows us to write

$$\Theta_{\lambda_i} = C\langle\Theta_{\lambda_i}, c^{0,v}\rangle\Theta_v + \sum_{m=1}^{d_v^J-1}\langle\Theta_{\lambda_i}, f^{m,v}\rangle f^{m,v}$$

for some $C \in \mathbb{C}$, using the orthonormality of the basis $\{c^{m,v}\}_{m=0}^{d_v^j-1}$. In other words, $\Theta_{\lambda_i} \in \mathcal{V}_{0,\Lambda} \oplus \mathcal{W}_{0,\Lambda}^J$. It follows that $\Theta_\lambda \in \mathcal{V}_{0,\Lambda} \oplus \mathcal{W}_{0,\Lambda}^J$ for all $\lambda \in \Lambda$ such that $d(\lambda) = (j_1, j_2, \cdots, j_k)$.

We now assume that for $1 \leq j \leq m$, if $\lambda \in \Lambda$ and $d(\lambda) = jJ$, then for any vertex $w \in \Lambda^0$,

$$S_\lambda(\Theta_w) \in \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^{m-1} \mathcal{W}_{j,\Lambda}^J; \text{ and} \tag{24}$$

$$\Theta_\lambda \in \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^{m-1} \mathcal{W}_{j,\Lambda}^J. \tag{25}$$

We have already established the base case $m = 1$.

Let us use induction to show that if $\lambda_0 \in \Lambda$ and $d(\lambda_0) = (m+1)J$, then

$$\Theta_{\lambda_0} \in \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^{m} \mathcal{W}_{j,\Lambda}^J, \text{ and } S_{\lambda_0}(\Theta_w) \in \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^{m} \mathcal{W}_{j,\Lambda}^J.$$

Fix a vertex $w \in \Lambda^0$. Let us calculate, using our standard formulas for our representation of $C^*(\Lambda)$ on $L^2(\Lambda^\infty, M)$,

$$S_{\lambda_0}(\Theta_w(x)) = \Theta_{\lambda_0}(x)(\rho(\Lambda)^{d(\lambda_0)/2})\Theta_w(\sigma^{d(\lambda_0)}(x)).$$

We first note: for this to have any chance of being nonzero, we need $x \in Z(\lambda_0)$ and $\sigma^{d(\lambda_0)}(x)$ must be in $Z(w)$. In other words, $s(\lambda_0) = w$. So we obtain: $S_{\lambda_0}(\Theta_w)$ is a constant multiple of Θ_{λ_0} if $w = s(\lambda_0)$, and $S_{\lambda_0}(\Theta_w) = 0$ if $w \neq s(\lambda_0)$.

So, assuming that $w = s(\lambda_0)$, we have that $S_{\lambda_0}(\Theta_w)$ is a constant multiple of $\chi_{Z(\lambda_0)} = \Theta_{\lambda_0}$. Using the factorization property, now write $\lambda_0 = \lambda_1\lambda_2$ with $s(\lambda_2) = s(\lambda_0) = w$ and

$$d(\lambda_1) = (j_1, j_2, \cdots, j_k)$$

and

$$d(\lambda_2) = (m \cdot j_1, m \cdot j_2, \cdots, m \cdot j_k).$$

Recall

$$S_{\lambda_0} = S_{\lambda_1\lambda_2} = S_{\lambda_1}S_{\lambda_2}.$$

By our induction hypothesis,

$$S_{\lambda_2}(\Theta_w) \in \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^{m-1} \mathcal{W}_{j,\Lambda}^J.$$

Therefore we can write

$$S_{\lambda_2}(\Theta_w) = g_0 + \sum_{j=0}^{m-1} h_j,$$

where $g_0 \in \mathcal{V}_{0,\Lambda}$ and $h_j \in \mathcal{W}_{j,\Lambda}^J$ for $0 \le j \le m-1$. So,

$$S_{\lambda_0}(\Theta_w) = S_{\lambda_1}\Big(g_0 + \sum_{j=0}^{m-1} h_j\Big) = S_{\lambda_1}(g_0) + \sum_{j=0}^{m-1} S_{\lambda_1}(h_j).$$

We have proved directly that $S_{\lambda_1}(g_0) \in \mathcal{V}_{0,\Lambda} \oplus \mathcal{W}_{0,\Lambda}$, and it follows from the definition of $\mathcal{W}_{j,\Lambda}^J$ that

$$S_{\lambda_1}(h_j) \in \mathcal{W}_{j+1,\Lambda}^J \quad \text{for} \quad 0 \le j \le m-1.$$

It follows that

$$S_{\lambda_0}(\Theta_w) \in \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^{m} \mathcal{W}_{j,\Lambda}^J.$$

Since $S_{\lambda_0}(\Theta_{s(\lambda_0)})$ is a constant multiple of Θ_{λ_0}, we have that

$$\Theta_{\lambda_0} \in \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^{m} \mathcal{W}_{j,\Lambda}^J,$$

as desired. It follows that the spanning set

$$S^J \subset \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^{\infty} \mathcal{W}_{j,\Lambda}^J,$$

and thus by Lemma 5.1,

$$L^2(\Lambda^\infty, M) = \mathcal{V}_{0,\Lambda} \oplus \bigoplus_{j=0}^{\infty} \mathcal{W}_{j,\Lambda}^J.$$

□

We now partially answer a question posed by A. Sims, who asked about the importance of the shape $(j, j, \ldots, j)$ of the "cubical wavelets" introduced in [20]. As we have now shown, we can construct wavelets of any non-trivial rectangular shape, not only cubes. Sims also asked if there was a relationship between the dimension of the spaces $\mathcal{W}_{j,\Lambda}^J$ and the fixed rectangular shape $J = (j_1, \ldots, j_k)$. The answer is "Not necessarily." We recall that for $v \in \Lambda^0$,

$$D_v^J = \{\lambda \in \Lambda : d(\lambda) = (j_1, j_2, \cdots, j_k) \text{ and } r(\lambda) = v\},$$

and $d_v^J = |D_v^J|$. The dimension of the wavelet space $\mathcal{W}_{j,\Lambda}^J$ is equal to

$$\sum_{v \in \Lambda^0} (d_v^J - 1).$$

Since each d_v^J depends on both $v \in \Lambda_0$ and $(j_1, j_2, \cdots, j_k) \in [\mathbb{N}^+]^k$, the dimensions obviously could change with different choices of degrees. On the other hand, if you take a degree that is ℓ times another degree $(j_1, j_2, \cdots, j_k)$, it would be interesting to check whether or not the wavelet space of level 0 corresponding to ℓJ, $\mathcal{W}_{0,\Lambda}^{\ell J}$, is equal to

$$\bigoplus_{j=0}^{\ell-1} \mathcal{W}_{j,\Lambda}^J.$$

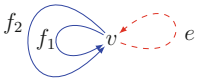
We also observe that, since $j_i \geq 1 \ \forall \ i$, the factorization property implies that every $\lambda \in D_v^J$ is associated to $\lambda_1 \in D_v^{(1,1,\ldots,1)}$, namely, $\lambda = \lambda_1 v$. In other words, $\lambda_1 = \lambda(0, (1, \ldots, 1))$ is the initial segment of λ of shape $(1, \ldots, 1)$. Thus, $d_v^J \geq d_v^{(1,\ldots,1)}$ for all v. In fact, by mapping the basis vector $\Theta_\mu \in \mathbb{C}^{d_v^{(1,\ldots,1)}}$ to the vector

$$\Psi_\mu = \sum_{v : \mu v \in D_v^J} \Theta_{\mu v} \in \mathbb{C}^{d_v^J}$$

we can transfer our orthonormal basis $\{c^{m,v}\}_m$ for $\mathbb{C}^{d_v^{(1,\ldots,1)}}$ to an orthonormal set in $\mathbb{C}^{d_v^J}$; then we can complete this orthonormal set to form the orthonormal basis for $\mathbb{C}^{d_v^J}$ that we use to construct the wavelet functions $f^{m,v}$.

In other words, whenever $J \geq (1, 1, \ldots, 1)$, not only can we form a wavelet basis for $L^2(\Lambda^\infty, M)$ by starting with paths of shape J, but we can use the data of the $(1, \ldots, 1)$-wavelets as the foundation for the J-shape wavelets.

Example 5.3 Here we consider the example introduced in Example 4.1 (and denoted by Λ_3 there) and compute some wavelets in this case. The corresponding 2-colored graph is given as the following;



and our factorization rules are:

$$f_2 e = e f_1 \text{ and } f_1 e = e f_2$$

By these factorization rules, we see that any particular infinite path in $x \in \Lambda^\infty$ can be chosen to be of the form

$$e f_{i_1} e f_{i_2} e f_{i_3} \cdots .$$

Setting "color 1" to be red and dashed, and "color 2" to be blue and solid, the two incidence matrices of this 2-graph are 1×1 and we have $(A_1) = (1)$, $(A_2) = (2)$. Therefore the Perron Frobenius-measure on cylinder sets is:

$$M(Z(e)) = 1,\ M(Z(ef_i)) = 1/2,\ M(Z(f_i e)) = 1/2,\ M(Z(f_i e f_j)) = 1/4,\ \text{etc},$$

where $i, j \in \{1, 2\}$.

Using Theorem 3.5 of [20], we construct isometries S_e, S_{f_1}, and S_{f_2} on $L^2(\Lambda^\infty, M)$ satisfying

$$\begin{aligned} S_e^* S_e &= S_{f_1}^* S_{f_1} = S_{f_2}^* S_{f_2} = I, \\ S_e S_e^* &= S_{f_1} S_{f_1}^* + S_{f_2} S_{f_2}^* = I. \end{aligned}$$

and finally

$$S_e S_{f_1} = S_{f_2} S_e \text{ and } S_e S_{f_2} = S_{f_1} S_e.$$

Fix $\xi \in L^2(\Lambda^\infty, M)$ and $x \equiv f_{i_1} e f_{i_2} e f_{i_3} \cdots$, where $i_j \in \{1, 2\}$. Note that our factorization rules imply that $x = f_{i_1+1} e f_{i_2+1} e f_{i_3+1} \ldots$, where the addition in the subscript of f is taken modulo 2.

We define

$$\begin{aligned} S_e(\xi)(x) &= \chi_{Z(e)}(x) 1^{1/2} 2^{0/2} \xi(\sigma^{(1,0)} x) = \xi(e f_{i_1+1} e f_{i_2+1} e f_{i_3+1} \cdots); \\ S_{f_1}(\xi)(x) &= \chi_{Z(f_1)}(x) 1^{0/2} 2^{1/2} \xi(\sigma^{(0,1)} x) = 2^{1/2} \chi_{Z(f_1)}(x) \xi(e f_{i_2+1} e f_{i_3+1} \cdots); \\ S_{f_2}(\xi)(x) &= \chi_{Z(f_2)}(x) 1^{0/2} 2^{1/2} \xi(\sigma^{(0,1)} x) = 2^{1/2} \chi_{Z(f_2)}(x) \xi(e f_{i_2+1} e f_{i_3+1} \cdots)). \end{aligned}$$

We further calculate:

$$\begin{aligned} S_e^*(\xi)(x) &= \chi_{Z(v)}(x) 1^{-1/2} 2^{0/2} \xi(ex) = \xi(e f_{i_1+1} e f_{i_2+1} e f_{i_3+1} \cdots); \\ S_{f_1}^*(\xi)(x) &= 2^{-1/2} \xi(f_1 x) = 2^{-1/2} \xi(e f_2 e f_{i_1+1} e f_{i_2+1} e f_{i_3+1} \cdots); \\ S_{f_2}^*(\xi)(x) &= 2^{-1/2} \xi(f_2 x) = 2^{-1/2} \xi(e f_1 e f_{i_1+1} e f_{i_2+1} e f_{i_3+1} \cdots). \end{aligned}$$

One can easily verify that the partial isometries satisfy the appropriate commutation relations.

We now construct wavelets for this example, using the method of Theorem 5.2. Recall M is the Perron-Frobenius measure on Λ^∞, and define ϕ to be the constant function 1 on Λ^∞. Take

$$(j_1, j_2) = (1, 1),$$

and let

$$\psi = \chi_{Z(ef_1)} - \chi_{Z(ef_2)}.$$

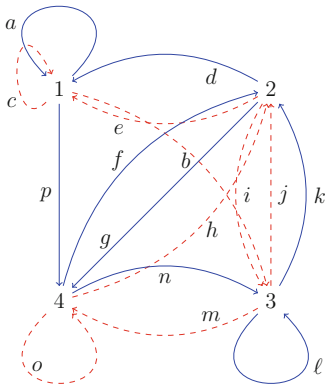
By using the main theorem of this section or direct calculation we verify that

$$\{\phi\} \cup \bigcup_{j=0}^{\infty} \{S_\lambda(\psi) : \lambda \in \Lambda, d(\lambda) = (j, j)\}$$

is an orthonormal basis for $L^2(\Lambda^\infty, M)$.

Example 5.4 In this example we describe how to construct the wavelets of this section for the Ledrappier 2-graph introduced in [43].

The skeleton of this 2-graph is



If we define "color 1" to be blue and solid, and "color 2" red and dashed, the adjacency matrices are

$$A_1 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \qquad A_2 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

Thus, there is a unique choice of factorization rules, since for each blue-red path (of length 2) between vertices v and w, there is exactly one red-blue path of length 2 between v and w.

For this 2-graph, one can check that $\rho(A_1) = \rho(A_2) = 2$ and that $x^\Lambda = \frac{1}{4}(1,1,1,1)$. Let $J = (1,2)$; then

$$D^J_{v_1} = \{acc, ace, aej, aeh, dhm, dho, djb, dji\}.$$

Similarly, $d^J_{v_i} = 8$ for all i, and the inner product (19) is given by

$$\langle \vec{x}, \vec{y} \rangle = \frac{1}{32} \sum_{j=1}^{8} x_j \overline{y_j}.$$

Thus, for each i, an orthonormal basis $\{c^{m,v_i}\}_{m=1}^7$ for the orthogonal complement of $\vec{1} \in \mathbb{C}^{d^J_{v_i}}$ is given by

$$c^{1,v_i} = (4,-4,0,0,0,0,0,0) \qquad c^{2,v_i} = (0,0,4,-4,0,0,0,0)$$
$$c^{3,v_i} = (0,0,0,0,4,-4,0,0) \qquad c^{4,v_i} = (0,0,0,0,0,0,4,-4)$$
$$c^{5,v_i} = \sqrt{2}(2,2,-2,-2,0,0,0,0) \qquad c^{6,v_i} = \sqrt{2}(0,0,0,0,2,2,-2,-2)$$
$$c^{7,v_i} = (2,2,2,2,-2,-2,-2,-2).$$

We will not list all of the 28 functions in $\mathcal{W}_{0,J}$ associated to the vectors $\{c^{m,v_i}\}$; however, we observe that

$$f^{1,v_1} = 4\Theta_{acc} - 4\Theta_{ace}; \qquad f^{4,v_1} = 4\Theta_{djb} - 4\Theta_{dji};$$
$$f^{5,v_1} = 2\sqrt{2}(\Theta_{acc} + \Theta_{ace} - \Theta_{aeh} - \Theta_{aej}).$$

6 Traffic Analysis Wavelets on $\ell^2(\Lambda^0)$ for a Finite Strongly Connected k-Graph Λ, and Wavelets from Spectral Graph Theory

Crovella and Kolaczyk argue in [11] that many crucial problems facing network engineers can profitably be approached using wavelets that reflect the structure of the underlying graph. They give axioms that such **graph wavelets** must satisfy and provide some examples; Marcolli and Paolucci use semibranching function systems to construct another example of graph wavelets in [41].

We begin this section by showing how to construct, from a Λ-semibranching function system, a family of wavelets on a higher-rank graph Λ which meets the specifications given in Section IV.A of [11]. In other words, our wavelets $g^{m,J}$ of Section 6.1 are orthonormal functions supported on the vertices Λ^0 of the k-graph Λ, which have finite support and zero integral. We thus hope that these wavelets will be of use for spatial traffic analysis on k-graphs, or, more generally, on networks with k different types of links.

In a complementary perspective to the graph wavelets discussed in [11], Hammond, Vandergheynst, and Gribonval use the graph Laplacian in [22] to construct wavelets on graphs. We show in Section 6.2 how to extend their construction to higher-rank graphs, and we compare the wavelets thus constructed with the wavelets from Section 5 and Section 6.1.

6.1 Wavelets for Spatial Traffic Analysis

Suppose that Λ is a finite strongly connected k-graph. Fix $v \in \Lambda^0$ once and for all; for every vertex $w \in \Lambda$, fix a "preferred path" $\lambda_w \in v\Lambda w$. We will use the Perron-Frobenius eigenvector x^Λ of Λ, and the vector $\rho(\Lambda) \in (0,\infty)^k$ of eigenvalues of the adjacency matrices A_i, to construct our traffic analysis wavelets.

For each $J \in \mathbb{N}^k$, let

$$D_J = \{\lambda \in v\Lambda : d(\lambda) = J \text{ and } \lambda = \lambda_{s(\lambda)}\}.$$

Observe that D_J might be empty. We will assume that we can (and have) chosen our preferred paths λ_w so that, for at least one $J \in \mathbb{N}^k$, $|D_J| \geq 2$.

If $|D_J| \geq 2$, define an inner product on $\mathbb{C}^{D_J}$ by

$$\langle \vec{v}, \vec{w} \rangle = \sum_{\lambda \in D_J} \overline{v_\lambda} w_\lambda \rho(\Lambda)^{-J} x^\Lambda_{s(\lambda)} \tag{26}$$

and let $\{(c^{m,J}_\lambda)_{\lambda \in D_J}\}_{m=1}^{|D_J|-1}$ be an orthonormal basis for the orthogonal complement of $(1,\dots,1) \in \mathbb{C}^{D_J}$ with respect to this inner product.

Define a measure $\tilde{\nu}$ on Λ^0 by a variation on counting measure: if $E \subseteq \Lambda^0$, set

$$\tilde{\nu}(E) = \sum_{w \in E} \rho(\Lambda)^{-d(\lambda_w)} x^\Lambda_w.$$

For each (m, J) with $J \in \mathbb{N}^k$, $|D_J| > 1$, and $m \leq |D_J| - 1$, define $g^{m,J} \in L^2(\Lambda^0, \tilde{\nu})$ by

$$g^{m,J}(w) = \begin{cases} 0, & d(\lambda_w) \neq J \\ c^{m,J}_{\lambda_w}, & d(\lambda_w) = J \end{cases}$$

Since the vectors $c^{m,J}$ are orthogonal to $(1,\dots,1)$ in the inner product (26), we have

$$\int_{\Lambda^0} g^{m,J}\, d\tilde{\nu} = \sum_{w \in \Lambda^0} g^{m,J}(w)\tilde{\nu}(w)$$

$$= \sum_{w:\lambda_w \in D_J} c^{m,J}_{\lambda_w} \rho(\Lambda)^{-J} x^{\Lambda}_w$$
$$= 0$$

for each (m, J). Moreover, if $g^{m,J}(w)\overline{g^{m',J'}(w)} \neq 0$, we must have $d(\lambda_w) = J = J'$; it follows that

$$\int_{\Lambda^0} g^{m,J}\overline{g^{m',J'}}\, d\tilde{\nu} = \delta_{J,J'} \sum_{w:\lambda_w \in D_J} c^{m,J}_{\lambda_w}\overline{c^{m',J}_{\lambda_w}} \rho(\Lambda)^{-J} x^{\Lambda}_w$$
$$= \delta_{J,J'}\delta_{m,m'}$$

since the vectors $\{c^{m,J}\}$ form an orthonormal set with respect to the inner product (26).

In other words, $\{g^{m,J}\}_{m,J}$ is an orthonormal set in $L^2(\Lambda^0, \tilde{\nu})$. However, we observe that the wavelets $g^{m,J}$ will not span $L^2(\Lambda^0, \tilde{\nu})$; at most, we will have $|\Lambda^0| - 1$ vectors $g^{m,J}$, which occurs when all the preferred paths λ_w are in the same D_J. In this case, $\{g^{m,J}\}_m \cup \{f\}$ is an orthonormal basis for $L^2(\Lambda^0, \tilde{\nu})$, where f is the constant function

$$f(w) = \frac{1}{\sqrt{\tilde{\nu}(\Lambda^0)}} = (\rho(\Lambda)^J)^{1/2}.$$

As an example, we consider the Ledrappier 2-graph of Example 5.4. Define $v := v_1$ and observe that every vertex v_i admits two paths $\lambda_i \in v\Lambda^{(1,2)}v_i$, so we can choose one of these for our "preferred paths" $\lambda_{v_i} := \lambda_i$. In this case, $g^{m,J} = 0$ unless $J = (1, 2)$; if we set

$$c^{1,(1,2)} = (4, -4, 0, 0), \qquad c^{2,(1,2)} = (0, 0, 4, -4), \qquad c^{3,(1,2)} = \sqrt{2}(2, 2, -2, -2),$$

then the vectors $\{c^{m,(1,2)}\}_m$ form an orthonormal basis for the orthogonal complement of $(1, 1, 1, 1)$ with respect to the inner product

$$\langle \vec{x}, \vec{y} \rangle = \sum_{i=1}^{4} x_{\lambda_i}\overline{y_{\lambda_i}} \rho(\Lambda)^{(-1,-2)} x^{\Lambda}_{v_i} = \frac{1}{32} \sum_{i=1}^{4} x_{\lambda_i}\overline{y_{\lambda_i}}.$$

Thus, our wavelets $g^{m,(1,2)}$ are given by

$$g^{1,(1,2)}(w) = \begin{cases} 4, & w = v_1 \\ -4, & w = v_2 \\ 0, & \text{else.} \end{cases} \qquad g^{2,(1,2)}(w) = \begin{cases} 4, & w = v_3 \\ -4, & w = v_4 \\ 0, & \text{else.} \end{cases}$$

$$g^{3,(1,2)}(w) = \begin{cases} 2\sqrt{2}, & w = v_1 \text{ or } v_2 \\ -2\sqrt{2}, & w = v_3 \text{ or } v_4. \end{cases}$$

Since $|\Lambda^0| = 4$ and all of the functions $g^{m,(1,2)}$ are orthogonal (in $L^2(\Lambda^0, \tilde{\nu})$) to each other and to the constant function $f(w) = (\rho(\Lambda)^{(1,2)})^{1/2} = 2\sqrt{2}$, the set $\{g^{m,(1,2)}\}_m \cup \{f\}$ is an orthonormal basis for $L^2(\Lambda^0, \tilde{\nu})$.

6.2 *Wavelets on $\ell^2(\Lambda^0)$ Coming from Spectral Graph Theory*

In this section we extend the definition of the graph Laplacian given by Hammond, Vandergheynst, and Gribonval in [22] to define a Laplacian for higher-rank graphs. For a graph (or k-graph) on N vertices, the (higher-rank) graph Laplacian is an $N \times N$ positive definite matrix. While the construction of the higher-rank graph Laplacian, given in Definition 6.1 below, differs slightly from that of the graph Laplacian of [22], the two matrices share many of the same structural properties. Consequently, the majority of the results from [22] apply to the higher-rank graph Laplacian as well, with nearly verbatim proofs. Thus, we include very few proofs in this section, instead referring the reader to [22].

There are many definitions of the graph Laplacian in the literature (cf. [2, 10, 31]); using the graph Laplacian to construct wavelets is also common.

Our definition of the k-graph Laplacian more closely parallels those of [2, 10] than that of [22], because the latter requires that the vertex matrix of the graph be symmetric. While this is always the case for an undirected graph, it is rarely the case for a k-graph, so we have chosen to define the k-graph Laplacian following the lines indicated in [2, 10]. We observe that in the case when the vertex matrices are indeed symmetric, the definitions in [22] and [2] of the graph Laplacian coincide.

Definition 6.1 (see [2, Definition 4.2], [10]) Let Λ be a finite k-graph with $N = |\Lambda^0|$ vertices. For each $1 \leq s \leq k$, let $N_1^s = |\Lambda^{e_s}|$ be the number of edges of color s. Define the incidence matrix $M_s = (m_{i,j}^s)_{i=1,\ldots,N; j=1,\ldots,N_1^s}$, where

$$m_{i,j}^s := \begin{cases} +1 \text{ if } r(e_j) \neq s(e_j) \text{ and } r(e_j) = v_i \\ -1 \text{ if } r(e_j) \neq s(e_j) \text{ and } s(e_j) = v_i \\ 0 \qquad\qquad \text{otherwise} \end{cases}$$

We then define the Laplacian Δ_Λ of Λ to be

$$\Delta_\Lambda := \sum_{s=1}^{k} M_s M_s^T.$$

Remark 6.2 When $k = 1$ and both definitions apply, Proposition 4.8 of [2] tells us that Definition 6.1 agrees with the definition of the graph Laplacian given in [22].

Furthermore, each summand $M_s M_s^T$ is a positive definite symmetric matrix; it follows (cf. [9]) that Δ_Λ has an orthonormal basis of eigenvectors and that the eigenvalues of Δ_Λ are all nonnegative.

Hammond, Vandergheynst, and Gribonval point out in [22] that the graph wavelets they describe can be viewed as arising from the graph Laplacian in the same way that continuous wavelets arise from the one-dimensional Laplacian operator d^2/dx^2. By slightly modifying the normalizations and definitions to make them consistent with our previous formulas for the Fourier transform, we obtain that the set of functions $\{e^{2\pi i\omega x} : \omega \in \mathbb{R}\}$ used to define the Fourier transform on $\mathbb{R}$ are also eigenfunctions of the Laplacian d^2/dx^2; thus, one could interpret the inverse Fourier transform

$$f(x) = \int \hat{f}(\omega) e^{2\pi i\omega x} d\omega$$

as providing the coefficients of f with respect to the eigenfunctions of the Laplacian. We define the **higher-rank graph Fourier transform** analogously.

To be precise, let $\{\vec{v}_i\}_{i=1}^N$ be a basis of eigenvectors for Δ_Λ.

Henceforth, we assume that we have ordered the eigenvalues $\lambda_1, \ldots, \lambda_N$ such that

$$\lambda_1 \leq \lambda_2 \leq \lambda_3 \cdots \leq \lambda_N.$$

The **higher-rank graph Fourier transform** of a function $f \in C(\Lambda^0)$ is the function $\hat{f} \in C(\Lambda^0)$ given by

$$\hat{f}(\ell) = \langle \vec{v}_\ell, f \rangle = \sum_{n=1}^{N_0} \vec{v}_\ell(n) f(n).$$

The motivation for the following definition comes from the calculations in Section 5.2 of [22]. Specific choices for wavelet kernels, and motivations for these choices, can be found in Section 8 of the same article.

Definition 6.3 Let Λ be a finite k-graph. A **wavelet kernel** is a function $g : \mathbb{R} \to \mathbb{R}$ such that

1. g is $(M+1)$-times continuously differentiable for some $M \in \mathbb{N}$, and $g^{(M)}(0) =: C \neq 0$;
2. On a neighborhood of 0, g is "well approximated" (as in Lemma 5.4 of [22]) by cx^M, where $c = C/M!$;
3. $\int_0^\infty \frac{g^2(x)}{x}\, dx =: C_g < \infty$.

Given a wavelet kernel g, the **k-graph wavelet operator** $T_g = g(\Delta_\Lambda)$ acts on $f \in C(\Lambda^0)$ by

$$T_g(f)(m) = \sum_{\ell=1}^{N} g(\lambda_\ell)\hat{f}(\ell)\vec{v}_\ell(m) = \sum_{\ell,n=1}^{N} g(\lambda_\ell)\vec{v}_\ell(n)\vec{v}_\ell(m) f(n).$$

For any $t \in \mathbb{R}$ we also have a **time scaling** T_g^t given by

$$T_g^t(f) = g(t\Delta_\Lambda)(f) = m \mapsto \sum_{\ell=1}^{N} g(t\lambda_\ell)\hat{f}(\ell)\vec{v}_\ell(m).$$

For each k-graph wavelet operator T_g and each $t \in \mathbb{R}$ we obtain a family $\{\psi_{g,t,n}\}_{1\le n\le N}$ of **higher-rank graph wavelets**: If $\delta_n \in C(\Lambda^0)$ is the indicator function at the nth vertex of Λ,

$$\psi_{g,t,n} := T_g^t\delta_n = m \mapsto \sum_{\ell=1}^{N} g(t\lambda_\ell)\vec{v}_\ell(n)\vec{v}_\ell(m).$$

Proposition 6.4 (*[22, Lemma 5.1]*) *Suppose $g : \mathbb{R} \to \mathbb{R}$ is a wavelet kernel and $g(0) = 0$. Then every function $f \in C(\Lambda^0)$ can be reconstructed from $\{\psi_{g,t,n}\}_{t,n}$:*

$$f = \frac{1}{C_g}\sum_{n=1}^{N}\int_0^\infty \frac{\langle\psi_{g,t,n}, f\rangle}{t}\psi_{g,t,n}\,dt.$$

Proof Recall that

$$\langle\psi_{g,t,n}, f\rangle = \sum_{\ell=1}^{N}\psi_{g,t,n}(\ell)f(\ell) = \sum_{\ell,m=1}^{N} g(t\lambda_m)\vec{v}_m(\ell)\vec{v}_m(n)f(\ell)$$

since the eigenvectors $\vec{v}_m$ are real-valued. Thus,

$$\begin{aligned}\sum_{n=1}^{N}\langle\psi_{g,t,n}, f\rangle\psi_{g,t,n}(k) &= \sum_{j,\ell,m,n=1}^{N} f(\ell)g(t\lambda_m)g(t\lambda_j)\vec{v}_m(\ell)\vec{v}_m(n)\vec{v}_j(n)\vec{v}_j(k)\\ &= \sum_{\ell,m=1}^{N} f(\ell)g(t\lambda_m)^2\vec{v}_m(\ell)\vec{v}_m(k)\end{aligned}$$

since the orthonormality of the eigenvectors $\{\vec{v}_m\}_m$ implies that

$$\langle\vec{v}_m, \vec{v}_j\rangle = \sum_n \vec{v}_m(n)\vec{v}_j(n) = \delta_{m,j}.$$

It follows that

$$\begin{aligned}\sum_{n=1}^{N}\int_0^\infty \frac{\langle \psi_{g,t,n}, f\rangle}{t}\psi_{g,t,n}(k)\,dt &= \sum_{\ell,m=1}^{N} f(\ell)\vec{v}_m(\ell)\vec{v}_m(k)\int_0^\infty \frac{g(t\lambda_m)^2}{t}\,dt \\ &= \sum_m \hat{f}(m)\vec{v}_m(k)\int_0^\infty \frac{g(t\lambda_m)^2}{t}\,dt \\ &= \sum_m \hat{f}(m)\vec{v}_m(k)\int_0^\infty \frac{g(u)^2}{u}\,du = \sum_m \hat{f}(m)\vec{v}_m(k)C_g.\end{aligned}$$

The symmetry of the Fourier transform implies that $f(k) = \sum_m \hat{f}(m)\vec{v}_m(k)$, which finishes the proof. □

Our hypothesis that the wavelet kernel g be well approximated by cx^M for some $M \in \mathbb{N}$ ensures that the wavelet $\psi_{g,t,n}$ is nearly zero on vertices more than M steps away from n. In other words, the wavelets $\psi_{g,t,n}$ are localized near the vertex n. The proof of this result is identical to that given in [22] for the case $k = 1$.

Proposition 6.5 (*[22, Theorem 5.5]*) *If $d(m,n) > M$, and if there exists $t' \in \mathbb{R}$ such that $|g^{(M+1)}(x)|$ is uniformly bounded for $x \in [0, t'\lambda_M]$, then there exist constants D, t'' such that for all $t < \min\{t', t''\}$,*

$$\frac{\psi_{g,t,n}(m)}{\|\psi_{g,t,n}\|} \leq Dt.$$

Example 6.6 We now construct spectral k-graph wavelets for the Ledrappier 2-graph of Example 5.4. Ordering the edges alphabetically, and assigning "color 1" to the blue, solid edges and "color 2" to the red, dashed edges, we obtain

$$M_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & -1 & 1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 \end{bmatrix} \qquad M_2 = \begin{bmatrix} -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & -1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Thus,

$$\begin{aligned}\Delta_\Lambda &= M_1M_1^T + M_2M_2^T \\ &= \begin{bmatrix} 2 & -1 & 0 & -1 \\ -1 & 4 & -1 & -2 \\ 0 & -1 & 2 & -1 \\ -1 & -2 & -1 & 4 \end{bmatrix} + \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 4 & -2 & -1 \\ -1 & -2 & 4 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 4 & -2 & -1 & -1 \\ -2 & 8 & -3 & -3 \\ -1 & -3 & 6 & -2 \\ -1 & -3 & -2 & 6 \end{bmatrix}.\end{aligned}$$

Computing the eigenvalues and eigenvectors (to two decimal places of accuracy), we obtain

$$\lambda_1 = 0, \lambda_2 = 5.17, \lambda_3 = 10.83, \lambda_4 = 8$$

and

$$\vec{v_1} = (1,1,1,1), \quad \vec{v_2} = (-0.85, 0.15, 0.35, 0.35),$$
$$\vec{v_3} = (-.15, 0.85, -0.35, -0.35), \quad \vec{v_4} = (0, 0, -0.71, 0.71).$$

Then the wavelets $\psi_{g,t,n}$ in $\ell^2(\Lambda^0)$ are given by

$$\psi_{g,t,n}(m) = \sum_{\ell=1}^{4} g(t\lambda_\ell)\vec{v}_\ell(n)\vec{v}_\ell(m).$$

As in [22], one possible wavelet kernel (with $N = 2$) is

$$g(x) := \begin{cases} x^2, & 0 \le x \le 1 \\ -5 + 11x - 6x^2 + x^3, & 1 < x < 2 \\ 4x^{-2}, & x \ge 2 \end{cases}$$

Observe that $g(x) > 0 \; \forall \; x > 0$.

To distinguish these wavelets $\psi_{g,t,n}$ from those of Section 6.1, we observe that (for fixed $t \in \mathbb{R}$) each of the four wavelets $\psi_{g,t,n}$ is supported on all four vertices of the Ledrappier 2-graph.

References

1. S. Bezuglyi, P.E.T. Jorgensen, Representations of Cuntz-Krieger algebras, dynamics on Bratteli diagrams, and path-space measures, in *Trends in Harmonic Analysis and Its Applications*, pp. 57–88, Contemp. Math., vol. 650 (Amer. Math. Soc., Providence, RI, 2015)
2. N. Biggs, *Algebraic Graph Theory*. Cambridge Tracts in Mathematics, vol. 67 (Cambridge University Press, London, 1974), vii+170 pp.
3. B. Blackadar, *Operator Algebras: Theory of C^*-Algebras and von Neumann Algebras*. Encyclopedia of Mathematical Sciences, vol. 122 (Springer, Berlin, 2006), xx+517 pp.
4. O. Bratteli, P.E.T. Jorgensen, A connection between multiresolution wavelet theory of scale N and representations of the Cuntz algebra $\mathcal{O}_N$. *Operator Algebras and Quantum Field Theory (Rome, 1996)* (Int. Press, Cambridge, MA, 1997), pp. 151–163
5. O. Bratteli, P.E.T. Jorgensen, Isometries, shifts, Cuntz algebras and multiresolution wavelet analysis of scale N. Integr. Equ. Oper. Theory **28**, 382–443 (1997)
6. O. Bratteli, P.E.T. Jorgensen, *Wavelets Through a Looking Glass: The World of the Spectrum* (Birkäuser, Boston, Basel, Berlin, 2002)
7. N. Brownlowe, Realising the C^*-algebra of a higher rank graph as an Exel crossed product. J. Oper. Theory **68**, 101–130 (2012)

8. T. Carlsen, S. Kang, J. Shotwell, A. Sims, The primitive ideals of the Cuntz-Krieger algebra of a row-finite higher-rank graph with no sources. J. Funct. Anal. **266**, 2570–2589 (2014)
9. F.K. Chung, *Spectral Graph Theory*. CBMS Reg. Conf. Ser. Math., vol. 92 (American Mathematical Society, Providence, RI, 1997), xii + 207 pp.
10. F.K. Chung, Laplacians and the Cheeger inequality for directed graphs. Ann. Comb. **9**, 1–19 (2005)
11. M. Crovella, E. Kolaczyk, Graph wavelets for spatial traffic analysis, in *Proceedings of IEEE Infocom 2003*, San Francisco, CA, USA, pp. 1848–1857 (2003)
12. J. D'Andrea, K. Merrill, J. Packer, Fractal wavelets of Dutkay-Jorgensen type for the Sierpinski gasket space, in *Frames and Operator Theory in Analysis and Signal Processing*. Contemp. Math., vol. 451 (Amer. Math. Soc., Providence, RI, 2008), pp. 69–88
13. K.R. Davidson, S.C. Power, D. Yang, Atomic representations of Rank 2 Graph Algebras. J. Funct. Anal. **255**, 819–853 (2008)
14. K.R. Davidson, D. Yang, Periodicity in Rank 2 Graph Algebras. Can. J. Math. **61**, 1239–1261 (2009)
15. D.E. Dutkay, P.E.T. Jorgensen, Wavelets on fractals. Rev. Mat. Iberoam. **22**, 131–180 (2006)
16. D.E. Dutkay, P.E.T. Jorgensen, Monic representations of the Cuntz algebra and Markov measure. J. Funct. Anal. **267**, 1011–1034 (2014)
17. D. Dutkay, G. Picioroaga, M.-S. Song, Orthonormal bases generated by Cuntz algebras. J. Math. Anal. Appl. **409**, 1128–1139 (2014)
18. D.G. Evans, On the K-theory of higher rank graph C^*-algebras. N. Y. J. Math. **14**, 1–31 (2008)
19. R. Exel. Inverse semigroups and combinatorial C*-algebras. Bull. Braz. Math. Soc. (N.S.) **39**, 191–313 (2008)
20. C. Farsi, E. Gillaspy, S. Kang, J. Packer, Separable representations, KMS states, and wavelets for higher-rank graphs. J. Math. Anal. Appl. **425**, 241–270 (2015)
21. C. Farthing, D. Pask, A. Sims, Crossed products of k-graph C^*-algebras by $\mathbb{Z}^l$. Houst. J. Math. **35**, 903–933 (2009)
22. D. Hammond, P. Vandergheynst, R. Gribonval, Wavelets on graphs via spectral graph theory. Appl. Comput. Harmon. Anal. **30**, 129–150 (2011)
23. A. an Huef, S. Kang, I. Raeburn, Spatial realisation of KMS states on the C^*-algebras of finite higher-rank graphs. J. Math. Anal. Appl. **427**, 977–1003 (2015)
24. A. an Huef, M. Laca, I. Raeburn, A. Sims, KMS states on the C^*-algebras of finite graphs. J. Math. Anal. Appl. **405**, 388–399 (2013)
25. A. an Huef, M. Laca, I. Raeburn, A. Sims, KMS states on C^*-algebras associated to higher-rank graphs. J. Funct. Anal. **266**, 265–283 (2014)
26. A. an Huef, M. Laca, I. Raeburn, A. Sims, KMS states on the C^*-algebra of a higher-rank graph and periodicity in the path space. J. Funct. Anal. **268**, 1840–1875 (2015)
27. A. an Huef, M. Laca, M.I. Raeburn, A. Sims, KMS states on the C^*-algebras of reducible graphs. Ergodic Theory Dyn. Syst. **35**, 2535–2558 (2015)
28. J.E. Hutchinson, Fractals and self-similarity. Indiana Univ. Math. J. **30**, 713–747 (1981)
29. D. Hammond, P. Vandergheynst, R. Gribonval, Wavelets on graphs via spectral graph theory. Appl. Comput. Harmon. Anal. **30**, 129–150 (2011)
30. A. Jonsson, Wavelets on fractals and Besov spaces. J. Fourier Anal. Appl. **4**, 329–340 (1998)
31. P.E.T. Jorgensen, E.P.J. Pearse, Spectral comparisons between networks with different conductance functions. *Random Walks, Boundaries and Spectra*. Progr. Probab., vol. 64 (Birkhauser/Springer Basel AG, Basel, 2011), pp. 111–142
32. S. Kang, D. Pask, Aperiodicity and primitive ideals of row-finite k-graphs. Int. J. Math. **25**, 1450022, 25 pp. (2014)
33. D.W. Kribs, S.C. Power, Analytic algebras of higher rank graphs. Math. Proc. Roy. Irish Acad. **106**, 199–218 (2006)
34. K. Kawamura, The Perron-Frobenius operators, invariant measures and representations of the Cuntz-Krieger algebras. J. Math. Phys. **46**, 083514, 6 pp. (2005)
35. A. Kumjian, D. Pask, Higher-rank graph C^*-algebras. N. Y. J. Math. **6**, 1–20 (2000)

36. A. Kumjian, D. Pask, A. Sims, Homology for higher-rank graphs and twisted C^*-algebras. J. Funct. Anal. **263**, 1539–1574 (2012)
37. A. Kumjian, D. Pask, A. Sims, On twisted higher-rank graph C^*-algebras. Trans. Am. Math. Soc. **367**, 5177–5216 (2015)
38. A. Kumjian, D. Pask, A. Sims, Twisted k-graph algebras associated to Bratteli diagrams. Integr. Equ. Oper. Theory **81**, 375–408 (2015)
39. A. Kumjian, D. Pask, A. Sims, On the K-theory of twisted higher-rank-graph C^*-algebras. J. Math. Anal. Appl. **401**, 104–113 (2013)
40. S. Mallat, Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. Trans. Am. Math. Soc. **315**, 69–87 (1989)
41. M. Marcolli, A.M. Paolucci, Cuntz-Krieger algebras and wavelets on fractals. Complex Anal. Oper. Theory **5**, 41–81 (2011)
42. D. Pask, I. Raeburn, N.A. Weaver, Periodic 2-graphs arising from subshifts. Bull. Aust. Math. Soc. **82**, 120–138 (2010)
43. D. Pask, I. Raeburn, N.A. Weaver, A family of 2-graphs arising from two-dimensional subshifts. Ergodic Theory Dyn. Syst. **29**, 1613–1639 (2009)
44. D. Pask, I. Raeburn, M. Rørdam, A. Sims, Rank-two graphs whose C^*-algebras are direct limits of circle algebras. J. Funct. Anal. **239**, 137–178 (2006)
45. D. Pask, A. Rennie, The noncommutative geometry of graph C^*-algebras. I. The index theorem. J. Funct. Anal. **233**, 92–134 (2006)
46. D. Pask, A. Rennie, A. Sims, The noncommutative geometry of k-graph C^*-algebras. J. K-theory **1**, 259–304 (2008)
47. S. Power, Classifying higher rank analytic Toeplitz algebras. N. Y. J. Math. **13**, 271–298 (2007)
48. I. Raeburn, *Graph Algebras*. CBMS Reg. Conf. Series in Math., vol. 103 (American Mathematical Society, Providence, RI, 2005), vii+113 pp.
49. D.I. Robertson, A. Sims, Simplicity of C^*-algebras associated to higher-rank graphs. Bull. Lond. Math. Soc. **39**, 337–344 (2007)
50. D.I. Robertson, A. Sims, Simplicity of C^*-algebras associated to row-finite locally convex higher-rank graphs. Isr. J. Math. **172**, 171–192 (2009)
51. G. Robertson, T. Steger, C^*-algebras arising from group actions on the boundary of a triangle building. Proc. Lond. Math. Soc. **72**, 613–637 (1996)
52. G. Robertson, T. Steger, Affine buildings, tiling systems and higher rank Cuntz-Krieger algebras. J. Reine Angew. Math. **513**, 115–144 (1999)
53. I. Raeburn, A. Sims, T. Yeend, Higher-rank graphs and their C^*-algebras. Proc. Edinb. Math. Soc. **46**, 99–115 (2003)
54. A. Sims, Gauge-invariant ideals in the C^*-algebras of finitely aligned higher-rank graphs. Can. J. Math. **58**, 1268–1290 (2006)
55. A. Sims, B. Whitehead, M.F. Whittaker, Twisted C^*-algebras associated to finitely aligned higher-rank graphs. Doc. Math. **19**, 831–866 (2014)
56. A. Skalski, J. Zacharias, Entropy of shifts on higher-rank graph C^*-algebras. Houst. J. Math. **34**, 269–282 (2008)
57. J. Spielberg, Graph-based models for Kirchberg algebras. J. Oper. Theory **57**, 347–374 (2007)
58. R. Strichartz, Construction of orthonormal wavelets, in "*Wavelets: Mathematics and Applications*". Stud. Adv. Math. (CRC Press, Boca Raton, FL, 1994), pp. 23–50
59. R. Strichartz, Piecewise linear wavelets on Sierpinski gasket type fractals. J. Fourier Anal. Appl. **3**, 387–416 (1997)

Part XVIII
Image and Signal Processing

Harmonic analysis has always been among the most important tools of image and signal processing. Despite all the recent developments in machine learning and neural networks, this is still the case and well illustrated by the present chapter.

We begin this chapter with a beautiful application of harmonic spectral methods to cancer research, as proposed by Mark Kon and Louise Raphael. In their paper, they adapt novel machine learning methods to regularize noisy and incomplete information. This is an important problem, studied widely in the area of supervised learning. To obtain their major results, Kon and Raphael utilize two types of techniques: local averaging of feature vectors on graphs, and support vector regression. Results are stated in the form of four main theorems, which present the pattern of bias-variance trade-off and the existence of a unique minimum for the estimation error in the regularization parameter. The authors also analyze the reconstruction accuracy for functions on graphs. They illustrate the strength of their approach on a case study in cancer genetics, with the aim of obtaining a novel prediction of cancer metastasis.

Robert S. Rand, Ronald G. Resmini, and David W. Allen present another fundamental example of the importance of methods arising in the context of harmonic analysis for applications in image and signal processing. In their paper, they analyze the intimate mixing phenomenon, which is a non-linear combination of endmember spectra. The traditional physics-based approach is augmented here by the use of generalized kernel fully constrained least squares optimization problems. The result of this approach is a novel algorithm which provides a way to adaptively estimate the mixture model most appropriate to the degree of non-linearity occurring at a given location in a given scene. The strength of this approach is validated with a dedicated laboratory experiment on hypespectral microscope imagery data.

In the last paper in this chapter, David A. Schug, Glenn R. Easley, and Dianne P. O'Leary present an important application of directional representations, such as those arising in the context of curvelets and composite wavelets, to problems in photogrammetry and tracking. In their work, they present a novel and promising approach to solve these problems, based on the use of wavelet and shearlet inspired edge detection algorithms for 3-dimensional imagery data formed. The edge

detection is then employed in the tracking methodology proposed by the authors and described in detail in this paper. The resulting techniques are well adapted to particular applications involving rigid motions and flat backgrounds, and perform well under such challenges as changing light conditions.

Precise State Tracking Using Three-Dimensional Edge Detection

David A. Schug, Glenn R. Easley, and Dianne P. O'Leary

Abstract An important goal in applications such as photogrammetry is precise kinematic state estimation (position, orientation, and velocity) of complex moving objects, given a sequence of images. Currently, no method achieves the precision and accuracy of manual tracking under difficult real-world conditions. In this work, we describe a promising new direction of research that processes the 3D datacube formed from the sequence of images and uses edge detectors to validate position hypotheses. We propose a variety of new 3D edge/surface detectors, including new variants of wavelet- and shearlet-based detectors and hybrid 3D detectors that provide computational efficiency. The edge detectors tend to produce broad edges, increasing the uncertainty in the state estimates. We overcome this limitation by finding the best match of the edge image from the 3D data to edge images derived from different state hypotheses. We demonstrate that our new 3D state trackers outperform those that only use 2D information, even under the challenge of changing lighting conditions.

Keywords State tracking • Edge detection • Kinematic state estimation from video • Surface detectors • Wavelet edge detectors • Shearlet edge detectors

1 Introduction

When cameras record a sequence of observations of an object moving in the field of view, we can try to track that object precisely. Complex object motion and complicated shape increase the difficulty of this type of tracking. Multiple target

D.A. Schug
NAWCAD, Patuxent River, MD, USA
e-mail: david.schug@navy.mil

G.R. Easley
MITRE, 7515 Colshire Drive, McLean, VA 22102, USA
e-mail: geasley@mitre.org

D.P. O'Leary (✉)
Computer Science Department and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA
e-mail: oleary@cs.umd.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4_4

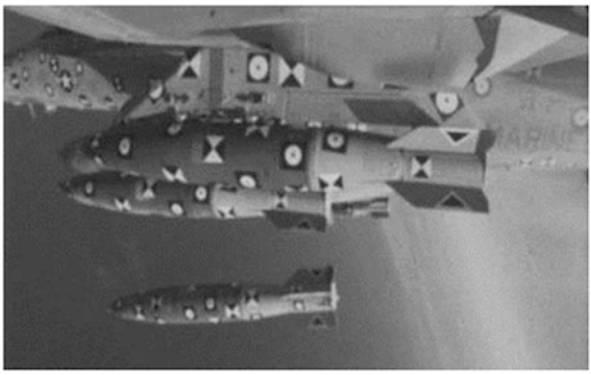


Fig. 1 Circular and bowtie-shaped features on objects to be tracked.

features on the object can be selected, as shown in Figure 1, in order to accurately fit the geometry and represent motion through space. It is important to have enough features to reliably represent the object and the motion, which means, for example, additional features if the object has fins or sharp curves. Other factors, including noise, clutter, and illumination variations, also make tracking difficult, since these change image intensities in complicated ways. In addition, low contrast image intensities make it difficult to discern an image feature from the background.

Very precise tracking is needed in photogrammetry, where the goal is to estimate the three-dimensional rigid-body kinematics of objects. With precision, we are concerned with a deviation from the target center that is relatively constant from frame to frame. A nearly constant deviation is more precise. Estimation accuracy will refer to the particular magnitude of deviation from a chosen standard. For real-world data, the chosen standard will depend on how well a human can estimate the image feature's center point. Typical tolerances require the object state parameters to be measurable to within one inch for Cartesian position and within one degree for Euler rotations. Tracking in 2D image space must therefore be accurate to within 2 pixels on average. Tracking accuracy will also depend on the camera's resolution influenced by the distance and orientation of the image feature with respect to the camera's field of view, and the camera's specific performance characteristics.

1.1 Previous Work in Tracking

Most methods that aim to estimate the kinematic state are feature-based searches that minimize a cost function that generates the sum of squared distances between chosen projected points and their corresponding observations from a particular image sequence (see [24] for more details). In practice, a two-dimensional feature tracker such as the Kanade-Lucas tracker [16] or other correlation based methods are used to collect observations while maintaining required correspondences with the chosen locations on the three-dimensional object.

Good techniques for general tracking include those provided by Lee et al. [14] and subspace methods for face tracking such as those in [1, 4] and [28]. Video-based tracking methods such as those provided in [11, 32] and [15] are also effective. Despite substantial progress in tracking algorithms, however, no single method has achieved the precision of manual tracking in photogrammetry. Most approaches to this particular kind of tracking problem have made use of edge detectors. Meaningful changes in edges are the fundamental criteria for distinguishing image features from the background. This is because the tracker can use the boundary of the feature to precisely register its orientation and position. At first glance, this may seem to be a complete solution to this tracking problem, but edge detectors can produce an estimated edge that is nonuniform in thickness, making it difficult to estimate critical attributes of the feature, such as its center and its velocity. In this work we use the edge detector to *validate* position estimates rather than to infer them.

1.2 Previous Work in Edge Detection

Traditional edge detection is performed with a single image I recorded at positions $\mathcal{P} = \{(i,j), i = 1,\ldots,m,\ j = 1,\ldots,n\}$. Given a threshold $h > 0$, the output of the edge detection process is a set of edge locations $\mathbf{P_E} \subseteq \mathcal{P}$ and an edge image $\mathbf{I_E}$ defined by

$$\mathbf{I_E}(i,j) = \begin{cases} 1 & (i,j) \in \mathbf{P}_E, \\ 0 & (i,j) \in \mathcal{P} \backslash \mathbf{P}_E. \end{cases} \tag{1}$$

Mathematically, we define

$$\mathbf{P}_E = \{\mathbf{p} \in \mathcal{P} : |\nabla I(\mathbf{p})| \geq h\}. \tag{2}$$

The image gradient $\nabla I(\mathbf{p})$ cannot be computed exactly, and edge detection algorithms differ in how they approximate this quantity.

There are many approaches to edge detection, including one based on computing eigenvalues of an autocovariance matrix to discover when the power spectrum of a function is slowly decreasing [5]. Perhaps the most widely used 2D edge detector is that due to Canny [3]. It is based on approximating the image gradient through convolution of the image with a filter formed from the derivative of a Gaussian. The algorithm depends on a single parameter σ_c, the standard deviation of the Gaussian

$$g_{\sigma_c}(x,y) = \exp\left(-(x^2+y^2)/(2\sigma_c^2)\right), \tag{3}$$

which effectively sets the scale for the smoothing of the image. The Canny algorithm is quite fast and is effective for simple shapes if σ_c is tuned to the image and the noise conditions, but this is not always possible. Figure 2 shows the results of applying MATLAB's `edge` algorithm using the Canny edge detector with various choices of

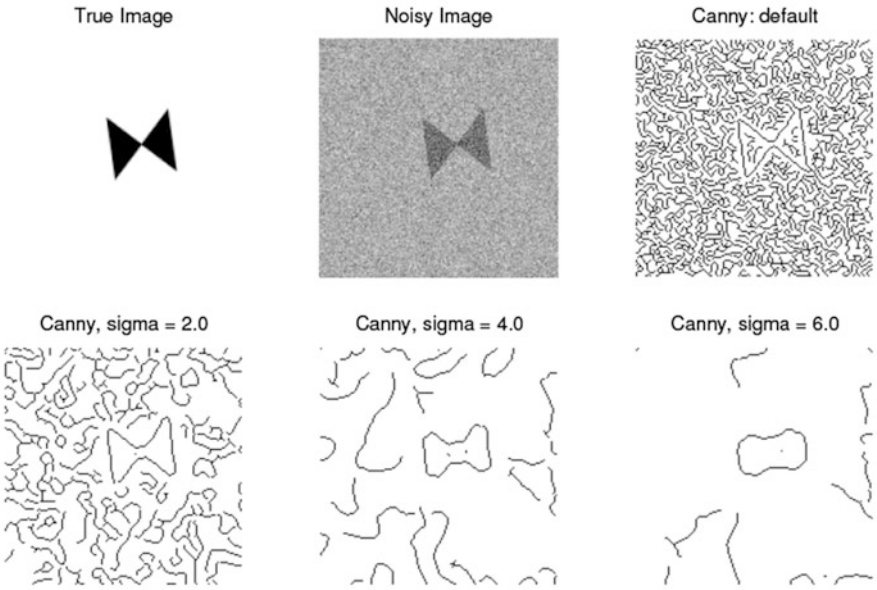


Fig. 2 Results of MATLAB's `edge` algorithm using the Canny edge detector with various choices of `sigma`. The pixel values in the true image are in the range [0, 1], and the standard deviation of the added noise is 0.4.

σ, to the noisy image shown in the top center. Smoothing at a single scale, i.e., with a single value of σ, makes it difficult to balance noise suppression with feature loss.

A further complication comes from choosing the correct thresholds that separate image features from the background or noise. Often one must manually adjust parameters until the method produces helpful results. The end result can be acceptable, but the level of human intervention is not satisfactory for automatic tracking applications.

A 2D wavelet edge detector [18, 19] overcomes the single-scale smoothing limitation of the Canny algorithm. The wavelet transform [29] provides a representation of an image using a set of basis functions that resolve detail at multiple scales and thus can reveal edge information at various levels of smoothing. Very efficient numerical implementations of the wavelet transform [17] make it practical. It can be much more effective than the Canny algorithm for complicated shapes, but because the wavelets are isotropic, it still has difficulty distinguishing close edges, crossing edges, and sharp changes in edge curvature. To resolve such cases, basis functions that have a sharp directional orientation are needed. One solution is to replace the scalable collection of isotropic Gaussian filters g_{σ_c} used in the Canny algorithm with a family of steerable and scalable anisotropic Gaussian filters [8]

$$G_{a_1,a_2,\theta}(x_1, x_2) = a_1^{-1/2} a_2^{-1/2} R_\theta \, G(a_1^{-1} x_1, a_2^{-1} x_2),$$

where $a_1, a_2 > 0$, R_θ is the rotation matrix for angle θ, and G is a Gaussian basis function parametrized by $a_1^{-1}x_1$ and $a_2^{-1}x_2$. The design and implementation of such filters is computationally involved, and the justification is essentially intuitive, lacking theory to prescribe parameters that best capture edges.

The shearlet transform provides this theory [13]. It provides a sparse directional representation [13, 30] key to obtaining a good edge detector. It is an invertible transform that relies on a set of analyzing functions (directionally oriented filters) that partition the frequency space at different scales and orientations. Knowing from mathematical analysis how the magnitudes at the edges change in this representation with respect to scale and shear (see [12] and [10]), a method for detecting edges and their orientation was given in [30]. The result is an improved capability to successfully detect subtle intensity differences and complicated object shapes.

All of these 2D methods have 3D counterparts that can be used for movies, sequences of images. Monga and Deriche [21] developed one of the first 3D Canny detectors, using separable Gaussians for smoothing. They applied their method to magnetic resonance and echographic volumetric data. Monga and Benayoun [20] extended the state of the art mathematically by using partial derivatives to treat the 3D image as a hypersurface in 4-dimensional space. They computed the curvatures at designated edge points using the partial derivatives of the image but did not obtain directional information. Brejl and Sonka [2] designed a directional 3D edge detector. Weiping and Hauzhong [27] used 3D wavelets to detect cerebral vessels in magnetic resonance angiograms (MRA). In [25] and [23], we proposed using the 3D shearlet transform for edge/surface detection and demonstrated its advantages over other methods in distinguishing high-curvature edges in low SNR conditions. It should be pointed out that this early 3D shearlet-based edge detector [25] is different from the one proposed in this work, which makes more explicit use of the analytic properties of the shearlet transform at surfaces and edges.

We believe that edge detection is much improved by including the time dimension, so we use 3D detectors. Previously, 3D detectors for tracking have been investigated in [6] and [31]. However, these are only for point features and are not suitable for our purpose. Another breakthrough in our approach is to make use of robust sparse and directional filtering concepts such as those provided in [6, 24], and [31].

1.3 Outline and Contributions

In Section 2, we discuss the tracking problem and the data. We then make several contributions.

- Section 3 focusses on new 3D edge detectors.
  - We develop new variants of 3D wavelet- and shearlet-based edge detection algorithms. We present a new 3D shearlet transform algorithm that is better suited to edge/surface detection than the version developed in [22]. This new algorithm exploits the theory provided in [9] and contains extensions as well

as important improvements over the algorithms developed in [30] for the purpose of feature tracking.
  - With efficiency in mind, we devise inexpensive but effective hybrids, combining results of 2D wavelet, shearlet, or Canny edge detectors on $x-y$, $y-t$, and $x-t$ slices, where x and y are spacial dimensions and t denotes time.
  - We demonstrate the effectiveness of these edge detection algorithms, but show that they are not adequate for precise tracking.
- In Section 4 we propose a totally new approach to tracking, using edge detectors to *validate* rather than *generate* state hypotheses, thus avoiding the uncertainty imposed by broad estimates of edges, and in Section 5 we demonstrate the effectiveness of our tracking ideas.

We draw conclusions in Section 6.

The edge detection algorithms described here were used by us in [26]. MATLAB software implementing these algorithms is available at http://www.cs.umd.edu/users/oleary/software.

2 The Data

The tracking problem starts with an observed image sequence (movie) that captures 2D snapshot information about 3D objects moving in the camera's field of view. We present two test problems that illustrate some of the difficulties in tracking a single feature from Figure 1. Tracking of multiple features can be done in parallel.

For our first test problem, a camera records a movie of a 3D ball of radius r whose position in the sequence of 2D images describes a circular orbit at radius R about the center pixel. Each image frame in the movie looks like a white disk on a black background, moved via translation, with the center (x_j, y_j) of the disk at time t_j, relative to the center of the image, given by

$$x_j = \lfloor R\cos(\alpha(t_j))\rfloor, \tag{4}$$

$$y_j = \lfloor R\sin(\alpha(t_j)\rfloor, \tag{5}$$

where $\alpha(t_j)$ is the angle defining the position of the object at time t.

Three frames from the resulting orbiting ball movie are shown in Figure 3. We chose the diameter of the disk to be an odd number of pixels so that in generating the data we can center it on the nearest pixel. The movie $\tilde{\mathbf{I}}$ is stored in an $m \times m \times \ell$ array, with $m^2 = 157^2$ pixels per frame and $\ell = 30$ frames. We generate the frames

Fig. 3 Three frames from the orbiting ball movie.

Fig. 4 Patch containing the disk (left) is inserted into a black background to create a noise-free frame of the movie (right).

Fig. 5 Three frames from an orbiting bow-tie movie with rotation and illumination changes.

Fig. 6 Patches containing a bow-tie (left), a rotated bow-tie (middle), and a shaded rotated bow-tie (right).

of the movie by inserting a $(2r + 3) \times (2r + 3)$ patch of pixels containing the disk into a black (zero) frame of size $m \times m$, as shown in Figure 4.

Our second test problem, with sample frames shown in Figure 5, is generated in a similar way, but uses a bow-tie patch, shown in Figure 6 (left), that orbits about the center of the frame but also rotates about its own center point, as illustrated in Figure 6 (middle). To perform rotation, we remap each pixel in the patch to its rotated position using bilinear interpolation. We also use this example to investigate changes in illumination. This is accomplished by generating a row vector $\mathbf{g}$ of increasing values in the range $[0.05, 2]$ with dimension equal to that of the patch. The illumination matrix is then defined as $\mathbf{L} = \mathbf{g}^T\mathbf{g}$. The shaded object is obtained by elementwise multiplication of the patch $\mathbf{P}$ by $\mathbf{L}$:

$$\mathbf{S} = \mathbf{P}. * \mathbf{L} \ . \tag{6}$$

This is performed after rotation and produces a result like that shown on the right in Figure 6.

We assume that we know the position of the object in the first frame of the movie. To study the robustness of our algorithms, we add white noise (independent normally distributed samples for each pixel) to the frames. We use our methods to estimate the position of the center of the ball or bow-tie and, for the bow-tie, its rotation angle, as a function of time. It is useful (but more difficult) to estimate the velocity of the object, too. In the next section we introduce 3D edge detectors that provide an important tool in making these estimates.

3 3D Edge Detectors

In this section we provide a brief description of how the 3D Canny, 3D wavelet, new 3D shearlet, and the new hybrid edge detectors identify edges.

The edge estimates produced by any of these algorithms can be refined by two well-known methods discussed in standard texts. *Nonmaximal suppression* labels a voxel as an element of the edge surface if its estimated gradient magnitude is at least h and if the magnitude is greater than at least one of its neighboring pairs. This can be applied in 3D or, to save time, in 2D, comparing each pixel to its neighboring pairs in the compass directions N-S, E-W, NE-SW, and NW-SE. *Hysteresis thresholding* identifies a voxel as a *strong* edge voxel if its gradient magnitude is greater than a threshold h. It is also identified as an edge voxel if it is connected to a strong edge and its gradient magnitude is larger than a threshold h_{low} and larger than the magnitude of each of its two neighbors in at least one of the compass directions. This, too, can be applied in 2D or 3D.

3.1 3D Canny Edge Detection

The 3D Canny algorithm (Algorithm 1) makes use of the 3D Gaussian low pass filter

$$\mathbf{g}_{\sigma}^{3D} = \exp\left(-(x^2 + y^2 + t^2)/(2\sigma^2)\right), \tag{7}$$

where σ is the standard deviation for the Gaussian. After convolution with this filter, it then uses convolution with a discretization of the partial derivatives $\mathbf{d}g_{\sigma,x}^{3D}$, $\mathbf{d}g_{\sigma,y}^{3D}$, and $\mathbf{d}g_{\sigma,t}^{3D}$ to estimate derivatives of the smoothed image sequence. It operates at a single scale, determined by the choice of σ, and produces estimates of the derivatives at the voxel centers.

Algorithm 1 The 3D Canny Edge Detection Algorithm.

1: **Input**: Raw image sequence $\widetilde{\mathbf{I}}$ and parameter σ.
2: **Output**: $(\nabla\widetilde{\mathbf{I}}_x, \nabla\widetilde{\mathbf{I}}_y, \nabla\widetilde{\mathbf{I}}_t)$, estimates of the gradient for each pixel in the input.
3: Compute the smoothed sequence $\widetilde{\mathbf{I}}_s = \widetilde{\mathbf{I}} * \mathbf{g}_{\sigma}^{3D}$.
4: Compute horizontal derivative estimate $\nabla\widetilde{\mathbf{I}}_x = \widetilde{\mathbf{I}}_s * \mathbf{d}g_{\sigma,x}^{3D}$.
5: Compute vertical derivative estimate $\nabla\widetilde{\mathbf{I}}_y = \widetilde{\mathbf{I}}_s * \mathbf{d}\mathbf{g}_{\sigma,y}^{3D}$.
6: Compute time derivative estimate $\nabla\widetilde{\mathbf{I}}_t = \widetilde{\mathbf{I}}_s * \mathbf{d}\mathbf{g}_{\sigma,t}^{3D}$.

3.2 3D Wavelet Edge Detection

A wavelet representation is a multiscale representation that allows us to overcome the problem of choosing an appropriate scale parameter σ. Given a function f in $L^2(\mathbb{R}^3)$ and an appropriate well-localized *mother wavelet* function $\psi \in L^2(\mathbb{R}^3)$, we define the continuous wavelet transform of f to be

$$\mathcal{W}_\psi f(a,t) = a^{-1} \int_{\mathbb{R}^3} f(x)\, \psi\left(a^{-1}(x-t)\right) dx,$$

where $a > 0$ and $t \in \mathbb{R}^3$. The analysis functions (wavelets) are $\psi_{a,t}(x) = a^{-1}\psi\left(a^{-1}(x-t)\right)$.

If f on $\mathbb{R}^3$ is smooth except for a discontinuity at $x_0 \in \mathbb{R}^3$, the wavelet transform $\mathcal{W}_\psi f(a,t)$ decays rapidly as $a \to 0$ everywhere, except where t is near x_0. Hence, the wavelet transform is able to signal the location of the singularity of f through its asymptotic decay at fine scales.

We discretize the wavelet transform and write $\tilde{\psi}_a(x) = a^{-1}\psi(-x/a)$, so that the wavelet transform can be expressed as the convolution product $\mathcal{W}_\psi f(a,t) = f * \tilde{\psi}_a(t)$. As a mother wavelet, we use a Sobel-like filter (instead of the Gaussian derivative used in Canny) to estimate the gradient and repeatedly apply a smoothing matrix to effectively dilate $\tilde{\psi}$ by an amount dependent on the number of iterations ℓ, obtaining $\tilde{\psi}_\ell$. This means we can concisely write the implementation as $f * \tilde{\psi}_\ell$ for $\ell = 1, 2, \ldots, n_\ell$.

Stacking each plane of the filter side-by-side reveals the contents for the horizontal, vertical, and time filters:

$$\mathbf{G}_x^{3D} = \left[\begin{array}{ccc|ccc|ccc} 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 2 & 0 & -2 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \end{array}\right], \tag{8}$$

$$\mathbf{G}_y^{3D} = \left[\begin{array}{ccc|ccc|ccc} 0 & 1 & 0 & 1 & 2 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & -2 & -1 & 0 & -1 & 0 \end{array}\right], \tag{9}$$

$$\mathbf{G}_t^{3D} = \left[\begin{array}{ccc|ccc|ccc} 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 & -1 & -2 & -1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{array}\right]. \tag{10}$$

After the image gradient is estimated using these filters, the wavelet edge detection algorithm repeatedly rescales by convolving the gradient components with a weighted average filter

$$\mathbf{G}_a^{3D} = \left[\begin{array}{c|c|c} 1\ 1\ 1 & 1\ 2\ 1 & 1\ 1\ 1 \\ 1\ 2\ 1 & 2\ 4\ 2 & 1\ 2\ 1 \\ 1\ 1\ 1 & 1\ 2\ 1 & 1\ 1\ 1 \end{array}\right], \tag{11}$$

choosing to save either the previous estimate or the new one, depending on which has smaller magnitude. This is summarized in Algorithm 2. Note that notation in the description of the algorithm emphasizes the approximate gradient aspects rather than the wavelet aspects.

The conditional re-enforcement in steps 7 and 8 emphasizes edge locations based on the change between adjacent scales. For both wavelets and shearlets, the local Lipschitz regularity of a point determines the decay of the magnitude of the response as a function of scale [7, 17]. The coefficients are likely to increase slowly with ℓ when the Lipschitz regularity is positive (i.e., an edge point) and increase rapidly when the Lipschitz regularity is negative (i.e., a noise point). Thus the choice is meant to strengthen the influence of edges and weaken the influence of noise.

Algorithm 2 The 3D Wavelet Edge Detection Algorithm.

1: **Input**: Raw image sequence $\widetilde{\mathbf{I}}$ and number of levels n_ℓ.
2: **Output**: $(\nabla\widetilde{\mathbf{I}}_x^{(n_\ell)}, \nabla\widetilde{\mathbf{I}}_y^{(n_\ell)}, \nabla\widetilde{\mathbf{I}}_t^{(n_\ell)})$, estimates of the gradient for each pixel in the input.
3: Compute the basic horizontal derivative estimate $\nabla\widetilde{\mathbf{I}}_x^{(0)} = \widetilde{\mathbf{I}} * \mathbf{G}_x^{3D}$, the basic vertical derivative estimate $\nabla\widetilde{\mathbf{I}}_y^{(0)} = \widetilde{\mathbf{I}} * \mathbf{G}_y^{3D}$ and the time derivative estimate $\nabla\widetilde{\mathbf{I}}_t^{(0)} = \widetilde{\mathbf{I}} * \mathbf{G}_t^{3D}$.
4: **for** level $\ell = 1, \ldots, n_\ell$ **do**
5: Compute the horizontal derivative estimate $\nabla\widetilde{\mathbf{I}}_x^{(\ell)} = \nabla\widetilde{\mathbf{I}}_x^{(\ell-1)} * \mathbf{G}_a^{3D}$.
6: Similarly, compute the vertical and time derivative estimates $\nabla\widetilde{\mathbf{I}}_y^{(\ell)}$ and $\nabla\widetilde{\mathbf{I}}_t^{(\ell)}$.
7: Modify the horizontal derivative estimate $\nabla\widetilde{\mathbf{I}}_x^{(\ell)}$ by choosing (for each point in $\mathcal{P}$) the minimum magnitude component from either $\nabla\widetilde{\mathbf{I}}_x^{(\ell-1)}$ or from the smoothed estimate $\nabla\widetilde{\mathbf{I}}_x^{(\ell)}$.
8: Similarly, compute the vertical and time derivative estimates $\nabla\widetilde{\mathbf{I}}_y^{(\ell)}$ and $\nabla\widetilde{\mathbf{I}}_t^{(\ell)}$.
9: **end for**

3.3 3D Shearlet Edge Detector

In shearlet analysis, we refine the wavelet analysis by, at each level, identifying components corresponding to different regions in frequency space.

The 3D shearlet transform implementation we have developed, like the 3D wavelet transform developed here, repeatedly rescales the gradient components, but at each scale it also partitions the frequency domain into a number of subdomains

$$\left\{(\eta_1,\eta_2,\eta_3) \mid \eta_1 \in \left[-\frac{2}{a}, -\frac{1}{2a}\right] \cup \left[\frac{1}{2a}, \frac{2}{a}\right], \right.$$
$$\left. \left|\frac{\eta_2}{\eta_1} - s_1\right| \le \frac{\sqrt{2a}}{4}, \ \left|\frac{\eta_3}{\eta_1} - s_2\right| \le \frac{\sqrt{2a}}{4}\right\}. \tag{12}$$

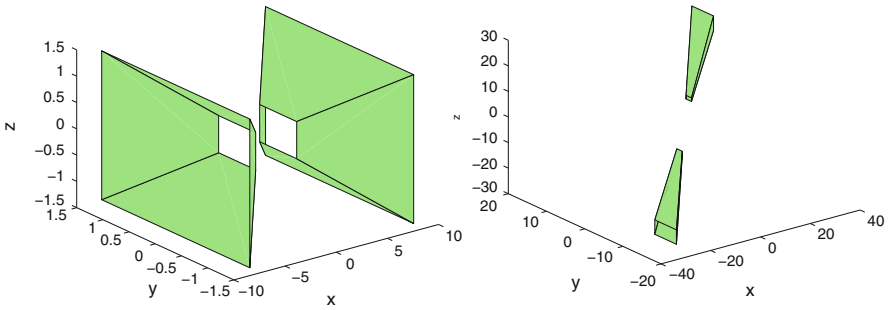


Fig. 7 The support of a 3D shearlet in the frequency domain with $a = 1/4$ and $s_1 = s_2 = 0$ (left) and $a = 1/16$, $s_1 = 0.5$, and $s_2 = 0.7$ (right).

Each subdomain, illustrated in Figure 7, is a pair of hyper-trapezoids, symmetric with respect to the origin, oriented according to the slope parameters s_1 and s_2, and more elongated as $a \to 0$.

Shearlet analyzing functions are defined as

$$\psi_{a,s_1 s_,2,t}(x) = |\det M_{as_1 s_2}|^{-\frac{1}{2}} \psi(M^{-1}_{as_1 s_2}(x-t))$$

where

$$M_{as_1 s_2} = \begin{pmatrix} a & -a^{1/2} s_1 & -a^{-1/2} s_2 \\ 0 & a^{1/2} & 0 \\ 0 & 0 & a^{1/2} \end{pmatrix},$$

and the mapping

$$\mathcal{SH}_\psi f(a, s_1, s_2, t) = \langle f, \psi_{a,s_1 s_,2,t} \rangle$$

defines the *continuous shearlet transform* of f for $a > 0$ and $t \in R^3$. (See [9] for a complete description as the technical issues in the construction are extensive.) The matrix $M_{as_1 s_2}$ is a product of a dilation matrix dependent on a and shearing matrices. Creating filters $w^{3D}_{d,\ell}$ whose frequency response produces the appropriate hyper-trapezoidal restrictions when combined with a wavelet filtering is done by extending the corresponding 2D filters $w_{d,\ell}$ constructed in [30]. The subscript d is an index used to replace the dependency of the window function on s_1 and s_2. The integer d ranges between 1 and n_d, where n_d indicates the total number of directional components for a scale we index by ℓ. An additional weighting correction is applied to each $w^{3D}_{d,\ell}$ to guarantee that the summation of all n_d components is a delta function. The continuous shearlet transform of f can then essentially be calculated as $f * (\tilde{\psi}_\ell * w^{3D}_{d,\ell})$ for $\ell = 1, \ldots, n_\ell$ and $d = 1, \ldots, n_d$. Thus, we can extend our wavelet transform algorithm to be a shearlet transform algorithm by doing an additional loop with a convolution dependent on $w^{3D}_{d,\ell}$.

Let Ω be a region in $\mathbb{R}^3$ with boundary denoted by $\partial\Omega$ and let $\gamma_j, j = 1, 2, \ldots, m$ be the smooth boundary segments of $\partial\Omega$, assumed to have positive Gaussian curvature at every point. If B is a function that is one for every point in Ω and zero elsewhere, then we know from [9] that:

- If $t \notin \partial\Omega$, then

$$\lim_{a\to 0^+} a^{-N}\mathcal{SH}_\psi B(a, s_1, s_2, s_2, t) = 0 \quad \text{for all } N > 0.$$

- If $t \in \partial\Omega \setminus \cup_{j=1}^m \gamma_j$ and (s_1, s_2) does not correspond to the normal direction of $\partial\Omega$ at t, then

$$\lim_{a\to 0^+} a^{-N}\mathcal{SH}_\psi B(a, s_1, s_2, s_2, t) = 0 \quad \text{for all } N > 0.$$

- If $t \in \partial\Omega \setminus \cup_{j=1}^m \gamma_j$ and $(s_1, s_2) = (\bar{s}_1, \bar{s}_2)$ corresponds to the normal direction of $\partial\Omega$ at t or $t \in \cup_{j=1}^m \gamma_j$ and (s_1, s_2) corresponds to one of the two normal directions of $\partial\Omega$ at t, then

$$\lim_{a\to 0^+} a^{-1}\mathcal{SH}_\psi B(a, s_1, s_2, t) \neq 0.$$

- If $p \in \gamma_j$ and (s_1, s_2) does not correspond to the normal directions of $\partial\Omega$ at t, then

$$|\mathcal{SH}_\psi B(a, s_1, s_2, t)| \leq Ca^{3/2}.$$

The above result essentially says that the continuous shearlet transform of a bounded region with piecewise smooth boundary has rapid decay everywhere, except when the location variable t is on the surface and the shearing variables correspond to the normal orientation, in which case it decays like $O(a)$ as $a \to 0$.

Our idea is that, at each level, for each shearlet region, we reinforce components that seem to be decaying at the proper rate by checking if the magnitude at a given position is between α and α^{-1} times the previous value. This comparison method is a significant improvement over the simple comparison concept originally conceived in [30] as this reduces thickening of the nominated edge points and increases efficiency.

Our shearlet edge detection algorithm, summarized in Algorithm 3, resembles the wavelet algorithm, but there is an inner loop at each scale that enhances the estimated derivative magnitude when an edge aligns with a shearlet filter $w_{d,\ell}^{3D}$.

3.4 3D Hybrid Wavelet and Shearlet Edge Detectors

Our experimental results show that the 3D wavelet and 3D shearlet edge detectors are quite effective but quite expensive. Therefore, we developed 3D hybrid wavelet-Canny and 3D hybrid shearlet-Canny edge detectors. These methods use 2D wavelet

or 2D shearlet methods to process each image, producing estimates of the x and y derivatives. The time derivative estimate is then taken from the 3D Canny algorithm. We summarize this process in Algorithm 4. There may be an advantage in averaging the Canny estimate for each horizontal and vertical derivative with the wavelet or shearlet estimate; if not, then the algorithm can be made more efficient by omitting the horizontal and vertical computations in the 3D Canny step.

Similarly, we developed edge detectors based on repeated use of 2D wavelet (or shearlet) edge detectors to compute 3D estimates as shown in Algorithm 5.

All of these algorithms are economical, and some proved quite effective.

Algorithm 3 The 3D Shearlet Edge Detection Algorithm.

Input: Raw image sequence $\widetilde{\mathbf{I}}$ and number of levels n_ℓ and parameter $\alpha \in [0, 1]$.
Output: $(\nabla\widetilde{\mathbf{I}}_x, \nabla\widetilde{\mathbf{I}}_y, \nabla\widetilde{\mathbf{I}}_t)$, estimates of the gradient for each pixel in the input.
Compute the basic horizontal derivative estimate $\nabla\widetilde{\mathbf{I}}_x = \widetilde{\mathbf{I}} * \mathbf{G}_x^{3D}$.
Similarly, compute the basic vertical and time derivative estimates $\nabla\widetilde{\mathbf{I}}_y = \widetilde{\mathbf{I}} * \mathbf{G}_y^{3D}$ and $\nabla\widetilde{\mathbf{I}}_t = \widetilde{\mathbf{I}} * \mathbf{G}_t^{3D}$.
Precompute the shearing filters $w_{d,\ell}^{3D}$.
for level $\ell = 1, \ldots, n_\ell$ **do**
    Compute the horizontal derivative estimate $\nabla\widetilde{\mathbf{I}}_x^{(+1)} = \nabla\widetilde{\mathbf{I}}_x * \mathbf{G}_a^{3D}$.
    Similarly, compute the vertical and time derivative estimates $\nabla\widetilde{\mathbf{I}}_y^{(+1)}$ and $\nabla\widetilde{\mathbf{I}}_t^{(+1)}$.
    Initialize $\Delta_x = \Delta_y = \Delta_t = \mathbf{0}$.
    for direction $d = 1, \ldots, n_d$ **do**
        Add into Δ_x any element of $\nabla\widetilde{\mathbf{I}}_x * w_{d,\ell}^{3D}$ whose magnitude is between α and α^{-1} times $\nabla\widetilde{\mathbf{I}}_x^{(+1)} * w_{d,\ell}^{3D}$.
        Update Δ_y and Δ_t similarly.
    end for
    Add Δ_x to $\nabla\widetilde{\mathbf{I}}_x$.
    Update $\nabla\widetilde{\mathbf{I}}_y$ and $\nabla\widetilde{\mathbf{I}}_t$ similarly.
end for

Algorithm 4 The 3D Hybrid Wavelet-Canny (or Shearlet-Canny) Edge Detection Algorithm.

Input: Raw image sequence $\widetilde{\mathbf{I}}$ and number of levels n_ℓ.
Output: $(\nabla\widetilde{\mathbf{I}}_x, \nabla\widetilde{\mathbf{I}}_y, \nabla\widetilde{\mathbf{I}}_t)$, estimates of the gradient for each pixel in the input.
Compute horizontal and vertical derivative estimates $\nabla\widetilde{\mathbf{I}}_x$ and $\nabla\widetilde{\mathbf{I}}_y$ by applying the 2D wavelet (or 2D shearlet) edge detector to each frame in the sequence $\widetilde{\mathbf{I}}$ using n_ℓ levels.
Compute a time derivative estimate $\nabla\widetilde{\mathbf{I}}_t$ by applying the 3D Canny edge detection algorithm to $\widetilde{\mathbf{I}}$.

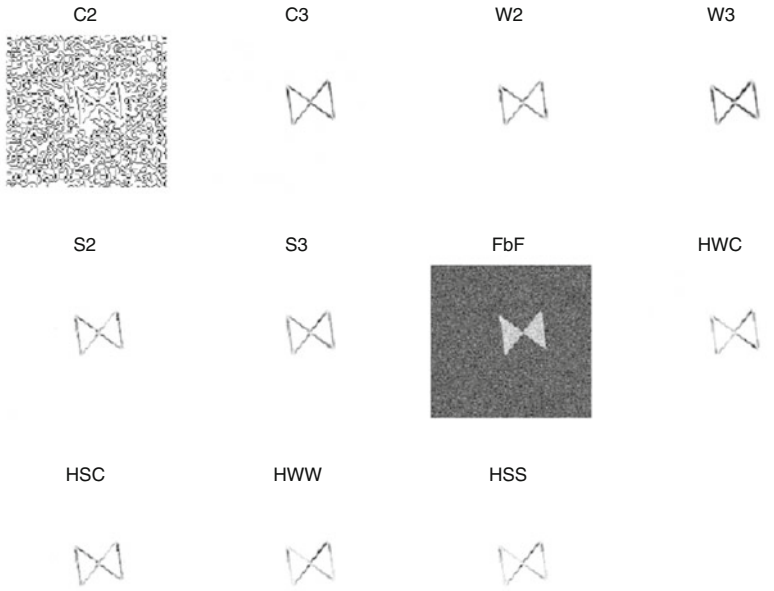


Fig. 8 Results of edge detectors on the rotating bow-tie movie, with standard deviation of noise 0.2. The methods are denoted as Canny (C), Wavelet (W), and Shearlet (S), 2D and 3D, the original frame (FbF), and Hybrid (H).

3.5 Performance of the Edge Detectors

To illustrate the potential of these edge detectors for tracking moving objects, we do a simple comparison of their performance. Figs. 8, 9, and 10 show results for the fifth frame of the rotating bow-tie movie with various noise levels. The movie frame is labeled FbF (frame-by-frame), and the results from Canny-2D (C2), Canny-3D (C3), Wavelet (W2 and W3), Shearlet (S2 and S3), and the hybrid algorithms (HWC, HSC, HWW, HSS) are also shown. We see that all of the algorithms are reliable when the SNR is high, but for low SNR, the 3D wavelet and shearlet algorithms are most reliable. Unfortunately, the algorithms can produce very broad edge estimates and can lose detail at sharp corners. This makes it very difficult to determine the precise location of a feature from the raw output of the edge detectors, so next we consider how to use this output more effectively in tracking.

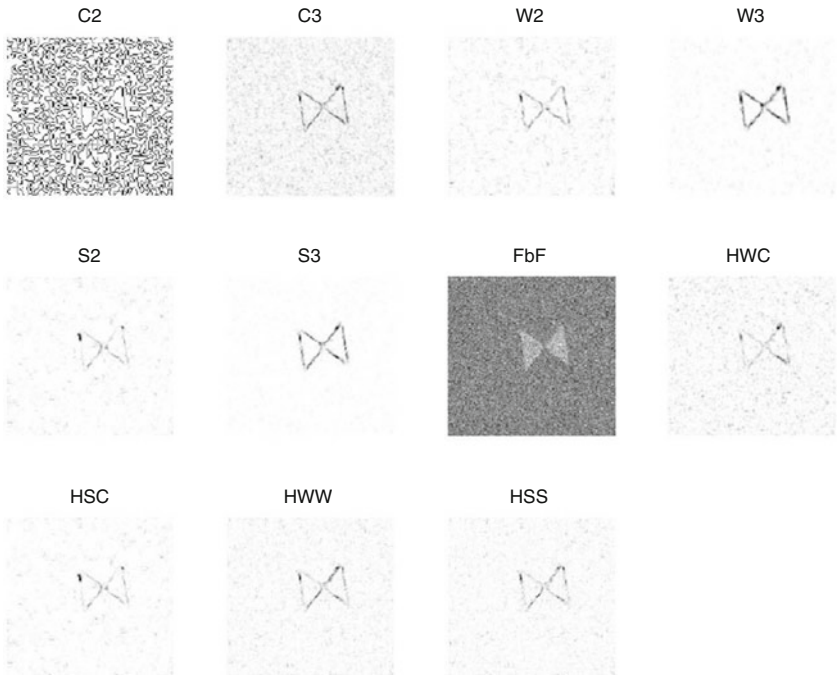


Fig. 9 Results of edge detectors on rotating the bow-tie movie, with standard deviation of noise 0.6.

Algorithm 5 The 3D Hybrid Wavelet-Wavelet (or Shearlet-Shearlet) Edge Detection Algorithm.

1: **Input**: Raw image sequence $\widetilde{\mathbf{I}}$ and number of levels n_ℓ.
2: **Output**: $(\nabla\widetilde{\mathbf{I}}_x, \nabla\widetilde{\mathbf{I}}_y, \nabla\widetilde{\mathbf{I}}_t)$, estimates of the gradient for each pixel in the input.
3: Apply the 2D wavelet (or 2D shearlet) edge detector to each frame (xy slice) in the sequence $\widetilde{\mathbf{I}}$ using n_ℓ levels to obtain horizontal and vertical derivative estimates $\boldsymbol{h}_{1x} = \nabla\widetilde{\mathbf{I}}_x$ and $\boldsymbol{h}_{1y} = \nabla\widetilde{\mathbf{I}}_y$.
4: Similarly, apply the 2D wavelet (or 2D shearlet) edge detector to each xt slice in the sequence to obtain horizontal and time derivative estimates $\boldsymbol{h}_{2x} = \nabla\widetilde{\mathbf{I}}_x$ and $\boldsymbol{h}_{2t} = \nabla\widetilde{\mathbf{I}}_t$.
5: Apply the 2D wavelet (or 2D shearlet) edge detector to each yt slice in the sequence to obtain vertical and time derivative estimates $\boldsymbol{h}_{3y} = \nabla\widetilde{\mathbf{I}}_y$ and $\boldsymbol{h}_{3t} = \nabla\widetilde{\mathbf{I}}_t$.
6: Compute derivative estimates $\nabla\widetilde{\mathbf{I}}_x = \frac{1}{2}(\boldsymbol{h}_{1x} + \boldsymbol{h}_{2x})$, $\nabla\widetilde{\mathbf{I}}_y = \frac{1}{2}(\boldsymbol{h}_{1y} + \boldsymbol{h}_{3y})$, and $\nabla\widetilde{\mathbf{I}}_t = \frac{1}{2}(\boldsymbol{h}_{2t} + \boldsymbol{h}_{3t})$.

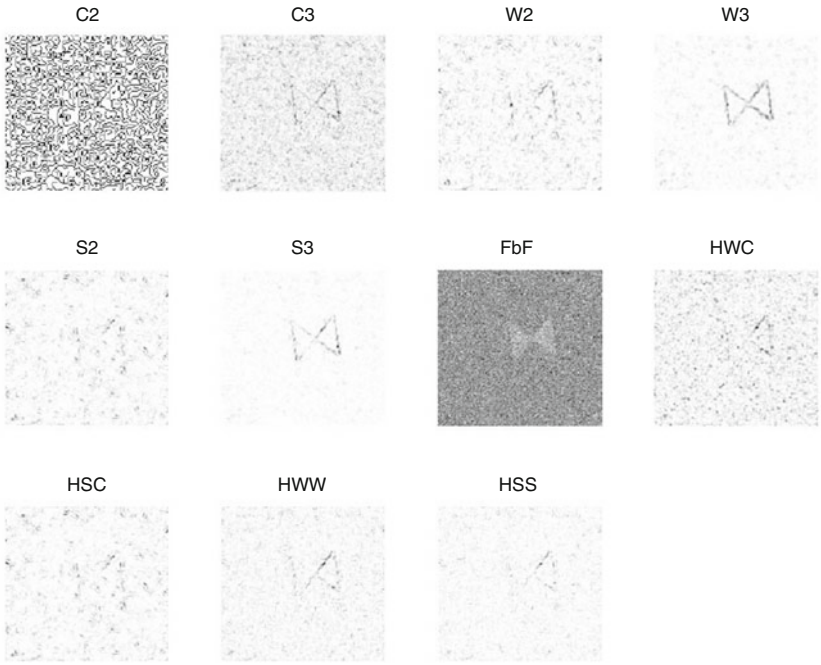


Fig. 10 Results of edge detectors on the rotating bow-tie movie, with standard deviation of noise 1.0.

4 From Edge Detection to Tracking

A tracking algorithm must determine the trajectory of the track object as it moves from frame to frame. For simplicity, we consider translation first and discuss object rotation later.

We tailor our algorithm to our photogrammetric application; since we have an image of the object to be tracked, we can make use of this information to avoid any need to identify features, and this is a significant advantage.

Assume that we are trying to determine the movement of the object between two particular frames, frame $k-1$ and frame k. Assume that the center of the object in frame $k-1$ is (i,j). We denote the displacement as Δx_k in the horizontal direction and Δy_k in the vertical direction and drop the subscript k when it is clear from context. Our first approach is to perform an exhaustive search for Δx and Δy by considering all possible positions of the patch of pixels defining the object, and testing to see which trial position best matches the data from the movie. In practice, velocity bounds can be used to limit the search, and in this study we only test integer displacement values between -2 and 2 for i and j, giving 25 possible positions.

For each trial position, we have two sets of data: $D(\tilde{\mathbf{I}})$, which is the data from the original movie $\tilde{\mathbf{I}}$, and $D(\tilde{\mathbf{I}}_p)$, where $\tilde{\mathbf{I}}_p$ is the movie $\tilde{\mathbf{I}}$ with the kth frame replaced

by one with the patch in its trial position. To find the correct position of the feature, we want to minimize the difference between the two sets of data, so we use a cost function

$$f(\tilde{\mathbf{I}}_p) = \|D(\tilde{\mathbf{I}}) - D(\tilde{\mathbf{I}}_p)\|. \tag{13}$$

A natural choice of norm is the square root of the sum of squares of the elements, but other choices are possible.

We also have a choice of the function D. The most obvious choice is $D(\tilde{\mathbf{I}}) = \tilde{\mathbf{I}}$. In this case f measures how the pixel values change when we replace the kth frame by the patched frame. Preservation of (noisy) pixel values is not our objective, however; we want to preserve edges. We propose, therefore, that D denote the edge image sequence produced by one of our edge detectors. In this case, f measures how much the edges change between the original movie and the patched movie. We generate the patched frame in the same way we generate our test examples, by overwriting pixels in the kth frame by the patch positioned at $(i + \Delta x, j + \Delta y)$.

If the object is also rotating, then we need to measure the cost function at various values of $\Delta\theta$, the change in rotation angle since the previous frame, as well as Δx and Δy. In our experiments, we tested values $\Delta\theta = -2, -1, 0, 1, 2$ degrees, making a total of 125 possible positions and rotations per frame. We found that our methods worked better if we added noise to the patch, comparable to that in the original movie, before inserting it into the kth frame of the movie.

We summarize our tracking method in Algorithm 6. There are two important observations to be made concerning the cost of the algorithm.

First, increasing the number of possible values of the Δ quantities quickly raises the expense of the exhaustive search algorithm. More sophisticated numerical optimization algorithms (steepest descent, Newton-like methods) can be used, but since our functions are non-differentiable and highly nonconvex, we did not have much success with them. One advantage of our admittedly primitive optimization approach is that it is quite easy to parallelize.

Second, there is a very important cost savings to be made. Rather than running the edge detector on the full image sequence, we can use a smaller *submovie* formed from a limited number of frames around frame k and a limited number of pixels within each frame, those near (i, j), since the effects on the edge detectors due to introducing the patch are primarily local. Making use of the submovie greatly reduces the cost of each trial.

From the computed Δ values, we can compute the magnitude of the planar velocity of the object at frame k,

Algorithm 6 Tracking Using Edge Detection.

1: D denotes the output of one of our edge detectors.
2: **Input**: Image sequence $\widetilde{\mathbf{I}}$ with ℓ frames, noise estimate, patch $\mathbf{P}$, shading vector $\mathbf{g}$, and initial patch location.
3: **Output**: Estimates of patch motion Δx, Δy, and $\Delta\theta$ for each frame.
4: Initialize $\Delta x_1 = \Delta y_1 = \Delta\theta_1 = 0$.
5: Add noise to the patch $\mathbf{P}$.
6: **for** $k = 2 : l$ **do**
7: Record $\Delta x_k = \Delta y_k = \Delta\theta_k = 0$ as the best guess so far.
8: **for** $d\theta = -2 : 1 : 2$ **do**
9: Rotate the patch by angle $d\theta$: $\mathbf{P}_r = \texttt{imrotate}(\mathbf{P}, d\theta)$.
10: Compute shaded patch $\mathbf{S} = \mathbf{P}_r . * (\mathbf{g}^T\mathbf{g})$.
11: **for** $dx = -2 : 1 : 2$ **do**
12: **for** $dy = -2 : 1 : 2$ **do**
13: The current trial location is the patch location at frame $k - 1$ plus (dx, dy).
14: Replace frame k of the image sequence $\widetilde{\mathbf{I}}$ with a frame containing the patch $\mathbf{S}$ at the trial location, obtaining $\widetilde{\mathbf{I}}_p$.
15: **if** $\|D(\widetilde{\mathbf{I}}) - D(\widetilde{\mathbf{I}}_p)\|$ (calculated using the relevant subimage sequence) is smaller than all previous values for frame k **then**
16: Set $\Delta x_k = dx$, $\Delta y_k = dy$, and $\Delta\theta_k = d\theta$.
17: **end if**
18: **end for**
19: **end for**
20: **end for**
21: Replace the patch $\mathbf{P}$ by rotating it by $\Delta\theta_k$.
22: **end for**

$$|v_k| = \sqrt{\Delta x_k^2 + \Delta y_k^2}, \tag{14}$$

and the direction of the planar velocity,

$$\phi_k = \arctan\left(\frac{\Delta y_k}{\Delta x_k}\right). \tag{15}$$

However, ϕ_k is quite sensitive to errors in Δx_k and Δy_k.

5 Experimental Results

Experiments were conducted to help understand and characterize how well our edge detectors work in tracking.

We used the difficult case of a spiraling shaded bow-tie with various amounts of noise added. For each standard deviation of the noise, we replicated the experiment 4 times, measuring the average error in single-frame tracking for frames 2 through

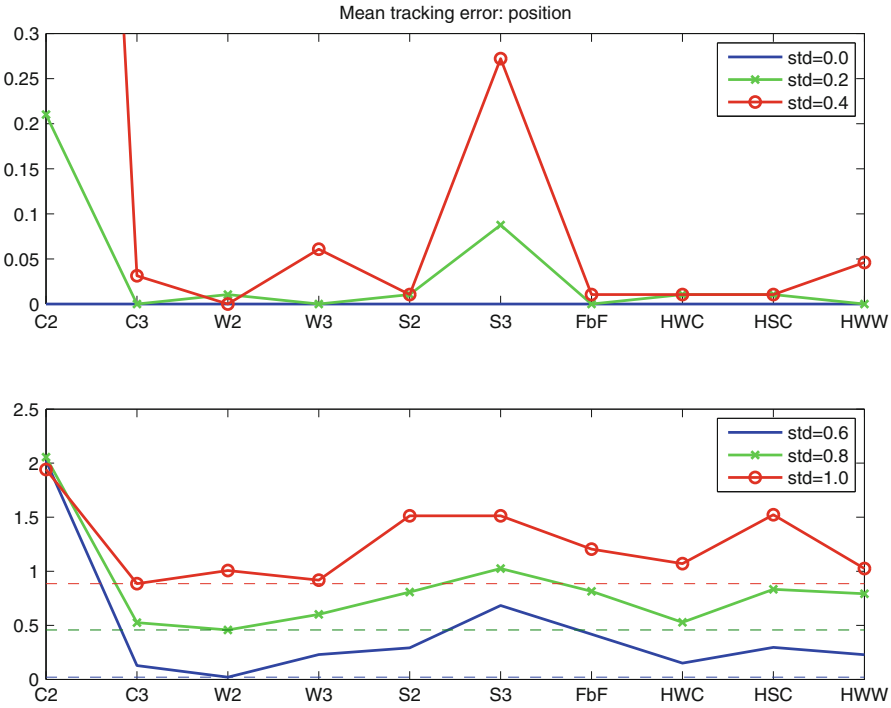


Fig. 11 Average error in position estimate (unit = pixel) for tracking a spiraling, rotating, shaded bowtie with noise. The dotted lines indicate the minimum error for each noise level.

26. The results are shown in Figs. 11, 12, and 13. If the standard deviation of the noise is 0.0, all of the edge detectors yielded perfect tracking. As the noise level increased, all algorithms except Canny-2D performed quite well, but the most reliable algorithms were Canny-3D, Wavelet-3D, Hybrid Wavelet-Canny, and Hybrid Wavelet-Wavelet for our particular set of experiments.

6 Conclusions

We have developed new variants of 3D wavelet- and shearlet-based edge detectors and new hybrid detectors that provide 3D information using only 2D wavelet or shearlet transforms. We demonstrated the effectiveness of these algorithms for edge detection. Our wavelet edge detectors try to filter noise from the gradient estimates, while our shearlet detectors reinforce gradients that change with scale at the expected rate. A variety of other implementations are possible, and some may perform better than these. All of these methods could be improved by tuning parameters and by applying standard post-processing techniques such as nonmaximal suppression or hysteresis thresholding.

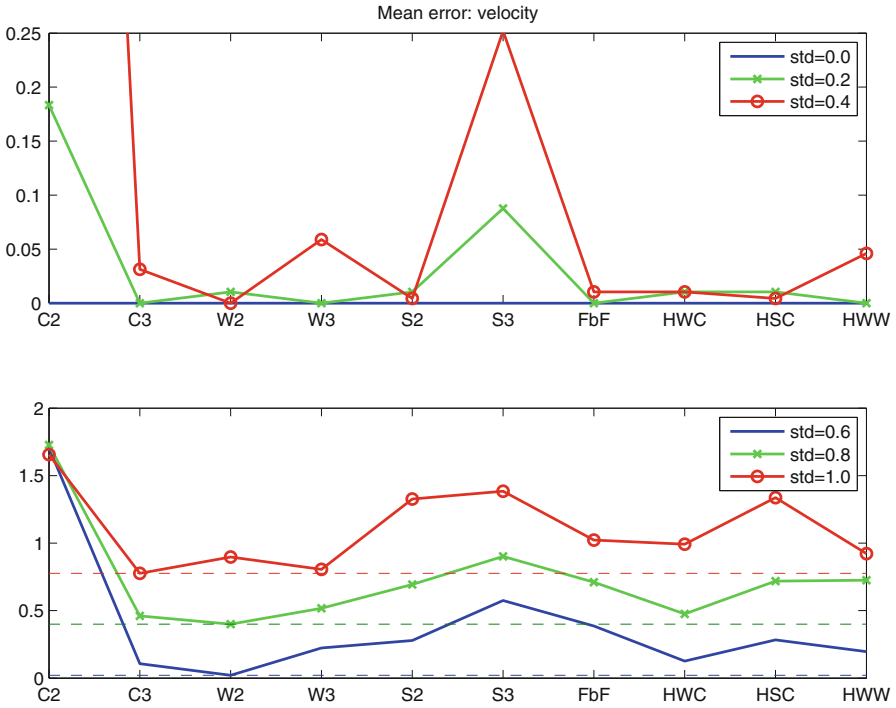


Fig. 12 Average error in velocity estimate (unit = pixel / frame) for tracking a spiraling, rotating, shaded bowtie with noise.

We then developed algorithms for tracking objects moving under translation and rotation, using edge detectors to validate position estimates. All of the methods tested, except Canny-2D, give low error in position, velocity, and rotation angle estimates in moderate noise. These methods are well adapted to particular applications involving rigid motion and flat backgrounds.

Transformations other than translation and rotation could be included in future work. Expansions and contractions of the patch would account for movement toward and away from the camera. We could also allow for roll and yaw of a 3D feature with known shape. Also, by fitting the patch to the object, our assumption of flat background could be removed.

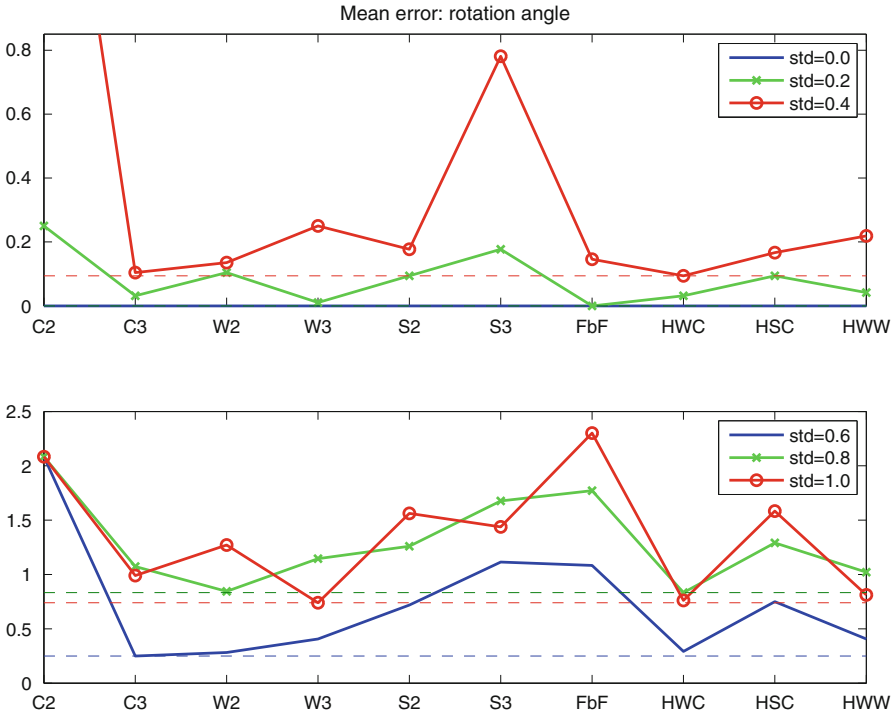


Fig. 13 Average error in rotation angle estimate (unit = degree) for tracking a spiraling, rotating, shaded bowtie with noise.

Acknowledgements D.P.O acknowledges partial support from the National Science Foundation under grant NSF DMS 1016266.

References

1. S. Birchfield, Elliptical head tracking using intensity gradients and color histograms, in *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE, Santa Barbara, 1998), pp. 232–237
2. M. Brejl, M. Sonka, Directional 3D edge detection in anisotropic data: detector design and performance assessment. Comput. Vis. Image Underst. **77**, 84–110 (1999)
3. J.F. Canny, Finding edges and lines in images. Master's thesis, MIT (1983)
4. Y. Chen, Y. Rui, T.S Huang, JPDAF based HMM for real-time contour tracking, in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001*, vol. 1 (IEEE, Kauai, 2001), pp. I–543
5. W. Czaja, M.V. Wickerhauser, Singularity detection in images using dual local autocovariance. Appl. Comput. Harmon. Anal. **13**(1), 77–88 (2002)
6. D. Davies, P. Palmer, M. Mirmehdi, Detection and tracking of very small low contrast objects, in *Proceedings of the British Machine Vision Conference* (BMVA Press, Surrey, 1998), pp. 60.1–60.10. doi:10.5244/C.12.60

7. G. Easley, K. Guo, D. Labate, et al., Analysis of singularities and edge detection using the shearlet transform, in *SAMPTA'09, International Conference on Sampling Theory and Applications* (2009)
8. J. Geusebroek, A.W.M. Smeulders, J. van de Weijer, Fast anisotropic Gauss filtering. IEEE Trans. Image Process. **8**, 938–943 (2003)
9. K. Guo, D. Labate, Analysis and detection of surface discontinuities using the 3D continuous shearlet transform. Appl. Comput. Harmon. Anal. **30**(2), 231–242 (2011)
10. K. Guo, D. Labate, W.-Q. Lim, Edge analysis and identification using the continuous shearlet transform. Appl. Comput. Harmon. Anal. **27**(1), 24–46 (2009)
11. V. Kruğer, S. Zhou, Exemplar-based face recognition from video, in *Computer Vision – ECCV 2002*, ed. by A. Heyden, G. Sparr, M. Nielsen, P. Johansen. Lecture Notes in Computer Science, vol. 2353 (Springer, Berlin, 2002)
12. G. Kutyniok, D. Labate, Resolution of the wavefront set using continuous shearlets. Trans. Am. Math. Soc. **361**(5), 2719–2754 (2009)
13. D. Labate, W.-Q. Lim, G. Kutyniok, G. Weiss, Sparse multidimensional representation using shearlets, in *Wavelets XI*. SPIE Proc., vol. 5914, Bellingham, WA (2005), pp. 254–262
14. K.-C. Lee, J. Ho, M.-H. Yang, D. Kriegman, Visual tracking and recognition using probabilistic appearance manifolds. Comput. Vis. Image Underst. **99**(3), 303–331 (2005)
15. X. Liu, T. Cheng, Video-based face recognition using adaptive hidden Markov models, in *Proceedings. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1 (IEEE, Madison, 2003), pp. I–340
16. B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in *Proceedings of the 7th International Joint Conference on Artificial Intelligence* (1981)
17. S.A. Mallat, *A Wavelet Tour of Signal Processing* (Academic, San Diego, 1998)
18. S.A. Mallat, W.L. Hwang, Singularity detection and processing with wavelets. IEEE Trans. Inf. Theory **38**(2), 617–643 (1992)
19. S.A. Mallat, S. Zhong, Characterization of signals from multiscale edges. IEEE Trans. Pattern Anal. Mach. Intell. **14**(7), 710–732 (1992)
20. O. Monga, S. Benayoun, Using partial derivatives of 3D images to extract typical surface features. Comput. Vis. Underst. **61**(2), 171–189 (1995)
21. O. Monga, R. Deriche, 3D edge detection using recursive filtering: application to scanner images. CVGIP Image Underst. **53**(1), 76–87 (1991)
22. P.S. Negi, D. Labate, 3-D discrete shearlet transform and video processing. IEEE Trans. Image Process. **21**(6), 2944–2954 (2012)
23. D.A. Schug, Three dimensional edge detection using wavelet and shearlet analysis. PhD thesis, Applied Mathematics and Statistics and Scientific Computing Program, University of Maryland (2012)
24. D.A. Schug, G.R. Easley, Three dimensional Bayesian state estimation using shearlet edge analysis and detection, in *2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)* (IEEE, Limassol, 2010), pp. 1–4
25. D.A. Schug, G.R. Easley, D.P. O'Leary, Three-dimensional shearlet edge analysis, in *SPIE: Defense, Security, and Sensing*, Orlando, FL, 25–29 April 2011
26. D.A. Schug, G.R. Easley, D.P. O'Leary, Wavelet–shearlet edge detection and thresholding methods in 3D, in *Excursions in Harmonic Analysis, Volume 3*, ed. by R. Balan, M.J. Begué, J.J. Benedetto, W. Czaja, K.A. Okoudjou (Springer, Cham, 2015), pp. 87–104
27. Z. Weiping, S. Hauzhong, Detection of cerebral vessels in MRA using 3D steerable filters, in *Engineering in Medicine and Biology 27 Annual Conference*, Shanghai, 1–4 September 2005
28. Y. Wu, T.S Huang, A co-inference approach to robust visual tracking, in *Proceedings. Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001*, vol. 2 (IEEE, Madison, 2001), pp. 26–33
29. S. Yi, D. Labate, G.R. Easley, H. Krim, Edge detection and processing using shearlets, in *Proceedings IEEE International Conference on Image Processing*, San Diego, CA, 12–15 October 2008

30. S. Yi, D. Labate, G.R. Easley, H. Krim, A shearlet approach to edge analysis and detection. IEEE Trans. Image Process. **18**(5), 929–941 (2009)
31. A. Yilmaz, K. Shafique, M. Shah, Target tracking in airborne forward looking infrared imagery. Image Vis. Comput. **21**(7), 623–635 (2003)
32. S. Zhou, V. Krueger, R. Chellappa, Probabilistic recognition of human faces from video. Comput. Vis. Image Underst. **91**(1), 214–245 (2003)

Approaches for Characterizing Nonlinear Mixtures in Hyperspectral Imagery

Robert S. Rand, Ronald G. Resmini, and David W. Allen

Abstract This study considers a physics-based and a kernel-based approach for characterizing pixels in a scene that may be linear (areal mixed) or nonlinear (intimately mixed). The physics-based method is based on earlier studies that indicate nonlinear mixtures in reflectance space are approximately linear in albedo space. The approach converts reflectance to single scattering albedo (SSA) according to Hapke theory assuming bidirectional scattering at nadir look angles and uses a constrained linear model on the computed albedo values. The kernel-based method is motivated by the same idea, but uses a kernel that seeks to capture the linear behavior of albedo in nonlinear mixtures of materials. The behavior of the kernel method is dependent on the value of a parameter, gamma. Validation of the two approaches is performed using laboratory data.

Keywords Nonlinear mixtures • Kernel-based methods • Single scattering albedo • Hapke theory • Hyperspectral

1 Introduction

Much consideration has been given in the past to linear mixing models, which are appropriate in cases where materials are presumed to be non-overlapping (areal) and can be mathematically expressed as a linear combination spectral endmembers, where the weights in the combination are associated with the abundances of each material. The endmembers are spectra (hopefully) representing unique materials in a given image such as water, soil, and vegetation. Abundances are the percentage of each endmember within a given pixel. However, there is no reason to presume that

R.S. Rand (✉)
National Geospatial-Intelligence Agency (NGA), Springfield, VA, 22150, USA
e-mail: rsrand7b@gmail.com

R.G. Resmini
The MITRE Corporation, McLean, VA, 22102, USA

D.W. Allen
National Institute of Standards and Technology, Gaithersburg, MD, 20899, USA

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4_5

such a model is appropriate in the case of intimately mixed materials that are likely to exhibit nonlinear spectral mixing behavior. For example, granular materials, such as soils, are often intimate mixtures of numerous different inorganic and organic substances, where the scattering of light and other nonlinear processes occur.

The linear model is constructed as linear combination of the spectra from multiple endmembers. It is basically a statistical linear regression model that has been posed in a number of ways with variations that impose either physical or sparseness constraints [1–7]. Image pixels are labeled as containing one, two, or perhaps many endmembers. Many approaches begin with a full linear model containing all possible variables (endmembers) and subsequently eliminate variables that do not contribute to the statistical significance (e.g., using an F-statistic) of the model [8]. Other approaches are basically step-wise regression, where the process begins with a pair of variables (endmembers) and adds variables if they contribute significantly to the model [6, 8].

The intimate mixing phenomenon is nonlinear. In reflectance space it involves a nonlinear combination of spectra from multiple endmembers. An intimate mixture model can be described by nonlinear functions, which are justified by Hapke scattering theory [9] and photometric phase functions [10]. This approach can be used to convert reflectance to Single Scattering Albedo (SSA). Prior results have shown up to 30% improvement in measurements over the linear mixing model when intimate mixtures are present [11]. Success was also achieved in efforts using a Constrained Energy Minimization (CEM) method and other linear methods applied to SSA data [12–14].

Kernel functions have also been introduced as a way to generalize linear algorithms to nonlinear data [15, 16]. In the case of detection and classification applications, kernel functions can induce high dimensional feature spaces. In these spaces, previously non-separable classes can be made linearly separable. Thus, linear methods can be applied in this new feature space that provides nonlinear boundaries back in the original data space. Another example is the kernel Principal Component Analysis (PCA) method [17]. The kernel, in this case, is not used to induce a high dimensional space, but is used to better match the data structure through nonlinear mappings. It is in this mode that kernels can be used to produce nonlinear mixing results while essentially using a linear mixture model. What is more appealing is that the physics suggests that such a method is ideal if one can model the kernel correctly.

The drawback with the earlier kernel algorithms for classification and detection is that they produced abundance estimates that do not meet the non-negativity and sum-to-one constraints. This was solved by the development of a Kernel Fully Constrained Least Squares (KFCLS) which computes kernel based abundance estimates to meet the physical abundance constraints [18]. Further investigation of the KFCLS method has resulted in (1) the development of a generalized kernel for areal (linear) and intimate (nonlinear) mixtures [19] and (2) an adaptive kernel-based technique for mapping areal and intimate mixtures [20]. The generalized kernel and adaptive technique provides a way to adaptively estimate a mixture model

suitable to the degree of nonlinearity that may be occurring at each pixel in a scene. This is important because a scene may contain both areal and intimate mixtures and we don't always know a priori which model is appropriate on a pixel-by-pixel basis. This situation was investigated further by Broadwater and Banerjee [21]. Building upon this work, a study investigating the behavior of the generalized KFCLS and adaptive kernel-based techniques was performed using both user-defined and SVDD automatically generated endmembers [22].

Research using laboratory data recently compared the performance of the generalized KFCLS applied to reflectance spectra with the Fully Constrained Least Squares (FCLS) method applied to spectra converted to SSA [23]. One of the conclusions of this effort was that similar accuracy in abundance estimates can be achieved using the SSA-based method, but with much faster computation time.

In the current study, we further this understanding by investigating both phenomenology-based SSA and mathematical-based kernel methods focused on laboratory data. The laboratory experiment is performed on highly controlled data containing pre-determined nonlinear mixtures of two materials.

2 Methodology

2.1 Fully Constrained Least Squares

The Fully Constrained Least Squares (FCLS) [7] mixing model for spectral mixtures can be written as

$$\mathbf{x} = \mathbf{E}\mathbf{a}, \quad \text{with two constraints: } a_i \geq 0 \text{ and } \sum_{i=1}^{N} a_i = 1 \tag{1}$$

where $\mathbf{x}$ is an $L \times 1$ vector containing the spectral signature of the current image pixel, $\mathbf{a}$ is an $N \times 1$ vector containing the estimated abundances (the ith entry represents the abundance value a_i), and $\mathbf{E}$ is an $L \times N$ matrix containing the endmember signatures (the ith column contains the ith endmember spectrum). The number of endmembers in the model is denoted by N and the number of bands in the spectra is denoted by L.

The FCLS method has been quite successful in the past for modeling linear mixing phenomenology. For our purposes, we will be using FCLS in two ways: The method will be used as a benchmark to compare with the proposed nonlinear methods; and the method will also be used in one of the nonlinear approaches, where we will apply the FCLS to spectra that has been converted to SSA, as discussed immediately, below.

2.2 *Proposed Method 1: Fully Constrained Least Squares (FCLS) Applied to Single Scattering Albedo Spectra*

As just mentioned, previous studies indicate that intimate (nonlinear) mixtures in reflectance space are approximately linear in albedo space. Accordingly, we investigate this behavior by applying a linear mixing method on albedo; specifically, by applying the FCLS method on data that's been converted to SSA. Conversion to SSA is described in Resmini et al. (1996) [12] and Resmini (1997) [13] (both studies following Hapke (1993) [9]; and Mustard and Pieters (1987) [10]) assuming the reflectance spectra are bidirectional. SSA spectra were also generated assuming the input reflectance spectra are hemispherical-directional. The expressions to transform reflectance spectra to SSA are given by Eqs. (2) and (3) for bi-directional (bd) reflectance and for hemispherical-directional (hd) reflectance, respectively. In the derivation of both expressions, phase angle is large enough that the opposition effect is assumed negligible.

$$\overline{\omega} = 1.0 \; - \left(\frac{\left[(\mu_0 + \mu)^2 \Gamma^2 + (1.0 + 4.0 \mu \mu_0 \Gamma)\,(1.0 - \Gamma) \right]^{0.5} - (\mu_0 + \mu)\,\Gamma}{(1.0 + 4.0 \mu \mu_0 \Gamma)} \right)^2 \tag{2}$$

$$\overline{\omega} = 1.0 - \left(\frac{1.0 - \Gamma}{(1.0 + 2.0 \mu \Gamma)} \right)^2 \tag{3}$$

In (2) and (3), ϖ is the single scattering albedo; Γ is the reflectance factor (see Section 2.1, Hapke [9]), μ_0 is the cosine of the angle of incidence of the illumination, and μ is the cosine of the viewing angle. Note that one reflectance is calculated as described previously in Section 2.1 though two different equations are used to generate the two sets of SSA spectra.

Subsequently, we refer to this approach as simply the "SSA method." The conversion of reflectance spectra to SSA using (2) and (3) is very fast as compared to the Generalized Kernel Least Squares (GKLS) method.

2.3 *Proposed Method 2: Generalized Kernel Fully Constrained Least Squares*

In previous work, a kernel-based mixing model was developed by Broadwater, Chellappa, and Banerjee [18], where the method estimates the abundances of a mixture using the expression

$$\widehat{\mathbf{a}} = \arg\min_{a} \frac{1}{2} \left(K(\mathbf{x}, \mathbf{x}) - 2\widehat{\mathbf{a}}^t K(\mathbf{E}, \mathbf{x}) + \widehat{\mathbf{a}}^t K(\mathbf{E}, \mathbf{E})\ \widehat{\mathbf{a}} \right), \quad s.t. \quad a_i \geq 0, \quad \forall_i \tag{4}$$

where $\widehat{\mathbf{a}}$ is the estimator for the abundance vector and $\mathbf{E}$ is the matrix of endmembers and, discussed above for (1). A quadratic programming method is used to calculate the abundance estimates and enforce the non-negativity constraint. The choice of kernel determines how well this method will respond to different types of mixing. Choosing the linear kernel $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^t\mathbf{y}$ is ideal for modeling linear mixtures; however, it is not a suitable kernel for intimate mixtures. A physics-inspired kernel has also been proposed and was shown to provide significantly improved behavior to model nonlinear mixtures [21]. The study concluded that although each kernel provides good results for the type of mixing intended, only one kernel or the other could be used, for either areal mixtures or intimate mixtures, but not both.

Broadwater and Banerjee further developed this approach into a generalized method for adaptive areal and intimate mixtures [19, 20]. They attempt to simulate Hapke theory for SSA, making use of the kernel:

$$K_\gamma(\mathbf{x}, \mathbf{y}) = \left(1 - e^{-\gamma \mathbf{x}}\right)^t \left(1 - e^{-\gamma \mathbf{y}}\right) \tag{5}$$

The kernel in (5) can be used for either areal or intimate mixtures through use of the appropriate γ. $K_\gamma(\mathbf{x}, \mathbf{y})$ approximates linear mixing whenever γ is very small. If γ is large, then $K_\gamma(\mathbf{x}, \mathbf{y})$ approximates intimate mixing in cases when the reflectance occurring from intimate mixing is modeled as

$$\mathbf{r} = \frac{\mathbf{w}}{4(\mu + \mu_0)} \left[H(\mathbf{w}, \mu)\ H(\mathbf{w}, \mu_0)\right] \tag{6}$$

where $\mathbf{r}$ is the reflectance vector, H is Chandrasekhar's function for isotropic scattering, $\mathbf{w}$ is the average single-scattering albedo vector, μ_0 is the cosine of the angle of incidence, and μ is the cosine of the angle of emergence [10].

The computation is similar in form to (4) except the minimization is done according to

$$\widehat{\gamma} = \arg\min_{\gamma} \frac{1}{2}\left(K_\gamma(\mathbf{x}, \mathbf{x}) - 2\widehat{\mathbf{a}}_\gamma{}^t K_\gamma(\mathbf{E}, \mathbf{x}) + \widehat{\mathbf{a}}_\gamma{}^t K(\mathbf{E}, \mathbf{E})\ \widehat{\mathbf{a}}_\gamma\right), \quad s.t. \quad a_i \geq 0, \quad \forall_i \tag{7}$$

where $\widehat{\mathbf{a}}_\gamma$ is the abundance estimate and $K_\gamma(\mathbf{x}, \mathbf{y})$ is the kernel evaluated with the parameter γ value. A numerical optimization based on the golden search method is used to minimize (7) [24]. An implementation of this generalized method is investigated, which we refer to as Generalized Kernel Least Squares (GKLS).

The GKLS method described by (7) at least theoretically has the ability to respond differently to differing degrees of nonlinearity. It attempts to automate the selection of γ, seeking to minimize the model's Root Mean Square Error (RMSE) to select the best gamma and compute more precise estimates of abundance.

The GKLS is also very compute intensive; therefore, as an alternative to automating the selection of γ, we also investigate a fixed-gamma GKLS implementation,

where the γ in (5) is chosen manually. This approach is much faster to compute; however, it comes at the disadvantage of losing flexibility to respond to different degrees of nonlinearity on a pixel-by-pixel basis.

3 Description of Experiment

A laboratory experiment is performed in this study to validate the proposed approach. Two granular materials were custom fabricated and mechanically mixed to form intimate mixtures. The materials are spherical beads of didymium glass and soda-lime glass, both ranging in particle size from 63 μm to 125 μm. The mixtures, which exhibit largely nonlinear spectral mixing, were then observed with a visible/near-infrared (VNIR; 400–900 nm) hyperspectral imaging (HSI) microscope.

In a configuration as shown in Figure 1, the glass bead mixtures were measured using the Resonon Pika II imaging spectrometer with a Xenoplan 1.4/23-0902 objective lens [26–28]. The device is a pushbroom sensor with a slit aperture, thus the need for a translation table to move the sample to facilitate hyperspectral image cube formation. Though capable of acquiring 240 bands from 400 to 900 nm, the sensor was configured to acquire 80 bands by binning (spectrally by three) resulting in a sampling interval of 6.25 nm and high signal-to-noise ratio spectra. The instrument is mounted nadir-looking at a mechanical translation table on which the sample to be imaged is placed. The height of the sensor above the table is user selectable; a height was chosen such that all mixtures are captured in the same scene thus the data have a ground sample distance of 75 μm/pixel. Four quartz-tungsten-halogen (QTH) lamps are used for illumination approximating a hemispherical-directional illumination/viewing geometry. Sensor and translation table operation, data acquisition, and data calibration are achieved by software that runs on a laptop computer. Calibration consists of a measurement of dark frame data (i.e., acquiring a cube with the lens cap on) and a measurement of a polytetrafluoroethylene (PTFE) reference plaque (large enough to entirely fill the field-of-view). Then, for each HSI cube measured, the sensor's software first subtracts the dark data and then uses the PTFE data (also dark subtracted) to ratio the spectral measurements to give relative reflectance (also known as reflectance factor: Hapke, [9] Schott [29]).

The mixtures are prepared and measured according to volume. Three binary mixtures (and the two endmembers) are constructed and emplaced in the wells of a 96-well sample plate: 0/100%, 25/75%, 50/50%, 80/20%, and 100/0% of didymium/soda-lime (percentages by volume). This was done as follows: Five cells of a 96-well sample plate, spray-painted flat black, were filled with the various glass bead mixtures; this is shown in Figure 2. The volume of each cell is 330 μL (0.33 mL). This is a data set with only nonlinear spectral mixing; the glass beads,

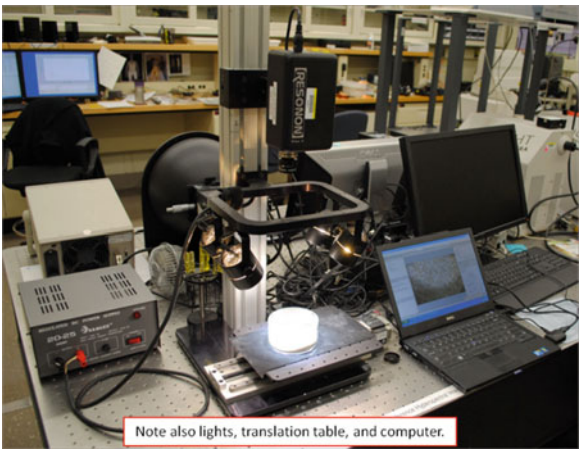


Fig. 1 Photograph of the VNIR HSI microscope. The Resonon Pika II is shown with the Xenoplan 1.4/23-0902 objective lens [25].

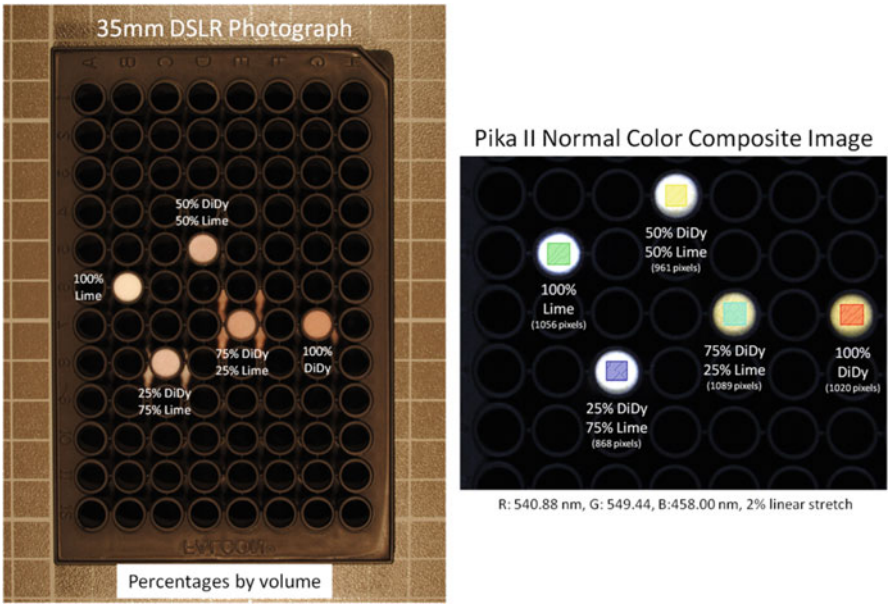


Fig. 2 (*Left*) 35 mm digital single-lens reflex (DSLR) camera photograph of the 96-well plate containing the glass beads. (*Right*) A Pika II normal color composite image (2% linear stretch of the bands used in the red-green-blue [RGB] image). All percentages are by volume of glass bead type indicated. Spillage onto the plate is evident in the photo on the *left* but not in the image on the *right*. Changing the stretch of the Pika II imagery will reveal the spillage [25].

didymium, and soda-lime are translucent. Their chemical composition, densities, and particle size range are well known. Note that the glass bead particle size range is much larger than the VNIR wavelengths used in this analysis. The glass beads and their mixtures display subtle, though interesting, gonioapparent changes in color.

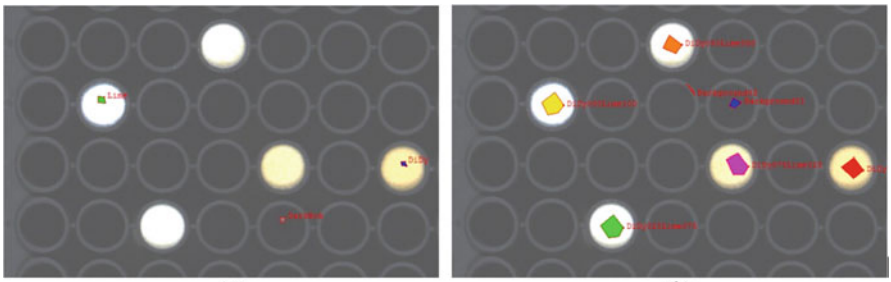


Fig. 3 RGB composite images of the hypercube used in the experiment trials showing in (**a**) the training regions and in (**b**) the test regions.

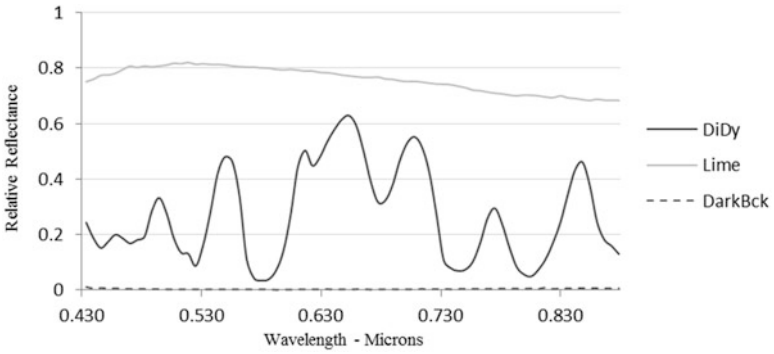


Fig. 4 Mean spectra for the three endmembers (DiDy, Lime, and Dark-Background) are shown.

Although many data cubes were acquired, we focus here on the analysis of one cube comprised of 640 samples, 500 lines, and 75 bands ranging from 434.0 nm to 885.0 nm. Of the 80 bands acquired the first 5 were discarded due to noise content.

Training and test data were extracted from the selected hyperspectral cube. Figure 3 shows polygons defining the training and test regions drawn on top of a Red-Green-Blue (RGB) color composite of this cube. For training, three training endmembers are defined: DiDy (100%), Lime (100%), and Background. The spectra within the small training polygon regions were extracted. Averages of the spectra in the regions were used as endmember spectra for the three methods under investigation. For purposes of testing the performance of the algorithm, five regions were extracted, corresponding to the five mixtures 100% DiDy, 75/25% DiDy/Lime, 50/50% DiDy/Lime, 25/75% DiDy/Lime, and 0/100% DiDy/Lime. Two additional test regions of background spectra were extracted. The training regions are shown in Figure 3a and the test regions are shown in Figure 3b. Note that none of these test regions overlapped the training regions. The spectra of the training endmembers are shown in Figure 4.

The image-derived training endmembers, as just described, are used to investigate the three methods described in Section 2: (1) FCLS applied in reflectance space; (2) GKLS applied in reflectance space; and (3) FCLS applied in SSA space. For the purposes of conciseness in reporting results, these methods henceforth will be referred to as the FCLS, GKLS, and SSA methods, respectively.

Numerous factors affect the performance of the methods. Three factors affecting the performance of all the methods are (1) the number the endmembers used in a model; (2) how well these endmembers span the space in which the mixing occurs; and (3) the Root Mean Square Error (RMSE) threshold for eliminating bad fits between the observed and model-estimated spectra. In addition, the GKLS method uses a kernel parameter "G" for γ that determines the nonlinear behavior introduced by the model's kernel. We test the GKLS method at fixed values of γ: G = 0.1, 0.5, 1.0, 2.0, 3.0, 4.0, 5.0, and 6.0, as well as the automated GKLS. For the SSA method, performance might be affected by the type SSA conversion: Hemispherical or Bi-Directional. If Bi-Directional, the input and output angles are other factors. In our case, we report on the results for an SSA conversion made assuming bi-directional reflectance with nadir input and output angles. We also tried the SSA conversions at other angles, but we didn't notice any noteworthy difference.

The experiment trials were made on the entire scene. Both qualitative results (shown by pictures of the entire scene) and quantitative results (applied in the test regions) are given.

4 Results

The results for the experiment are shown in Figs. 5, 6, and 7, as well as Tables 1 and 2. Figure 5 shows RGB Color Abundance Maps for four of the trials (Red = DiDy, Green = Lime, Blue = Background). Qualitatively, Figure 5a, b shows poor correspondence to the known mixtures shown in Figure 2. Figure 5c, d shows much better correspondence to the known mixtures shown in Figure 2. The variations in color within the discs containing the three mixtures (75/25, 50/50/ and 25/75) is noteworthy. These variations indicate the methods are detecting notable variance in the abundance proportions. This is in spite of the experiment goals to prepare mixture proportions that are as uniform as possible. There is little reason to doubt that these variations are real and that the methods (particularly, GKLS and SSA) are responding correctly. Static clinging and other inter-particle interactions can easily account for the clumping and variations observed.

Table 1 lists the average estimated abundances for the FCLS, GKLS, and SSA methods in the five test regions. Figure 6 shows these results graphically. The "truth" (actual physically measured) percent by volume of DiDy for these regions varied slightly from the goal of 100%, 75%, 50%, 25%, and 0%. In reality, these were measured as 100%, 78.8%, 50.5%, 24.2%, and 0% for DiDy100, DiD075, DiDy050, DiDy025, and DiDy000, respectively. Results for the GKLS method using a fixed γ parameter of G = 0.1, G = 0.5, G = 1.0, G = 2.0, G = 3.0, G = 4.0, G = 5.0, and

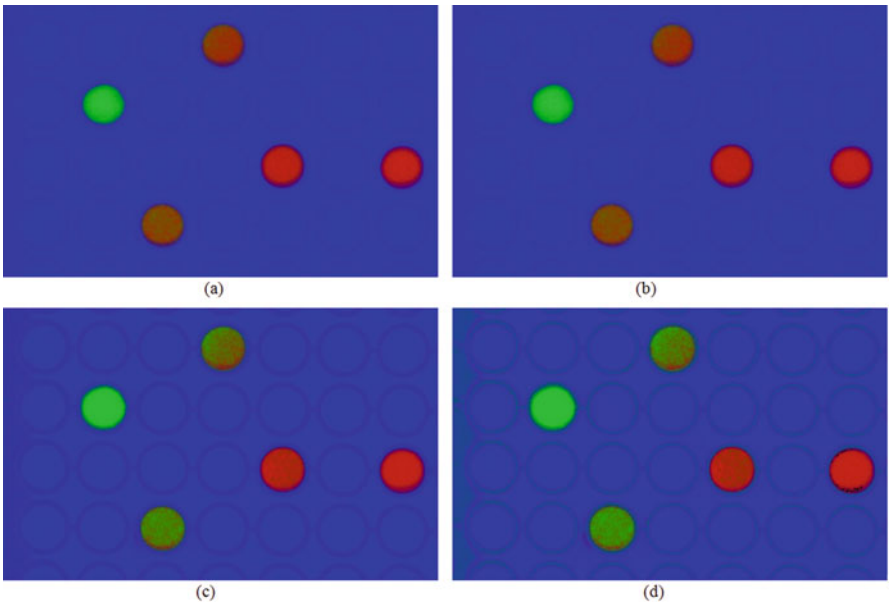


Fig. 5 RGB composite images showing Color Abundance Maps for four of the trials (*Red* = DiDy, *Green* = Lime, *Blue* = Background). (**a**) FCLS method (Linear) and (**b**) GKLS at γ-parameter G = 0.1 (Linear) show poor correspondence to the known mixtures shown in Figure 2. (**c**) GKLS at γ-parameter G = 5.0 (Nonlinear) and (**d**) SSA method (Nonlinear) show much better correspondence to the known mixtures.

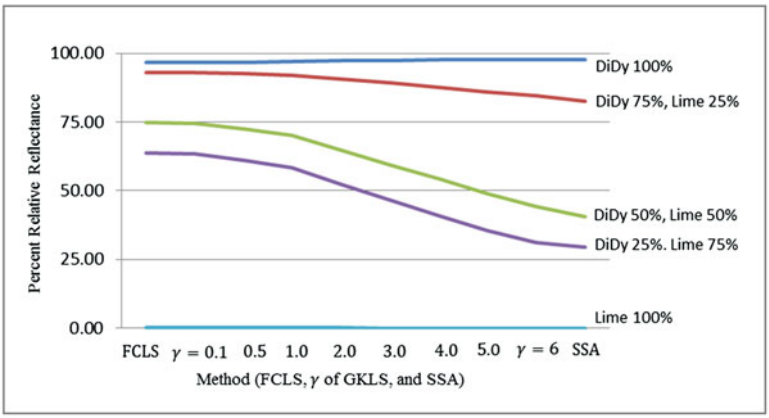


Fig. 6 A graph of the results listed in Table 1 for the FCLS, GKLS, and SSA methods is displayed.

G = 6.0 are given. This table shows FCLS to be poor at predicting the abundances for DiDy075 (93.11% vs. 78.8%) and DiDy050 (74.97% vs. 50.5%), as well as very poor at predicting DiDy025 (63.88% vs. 24.2%). The results of GKLS for small gamma agree with theoretical expectations of approximately a linear model.

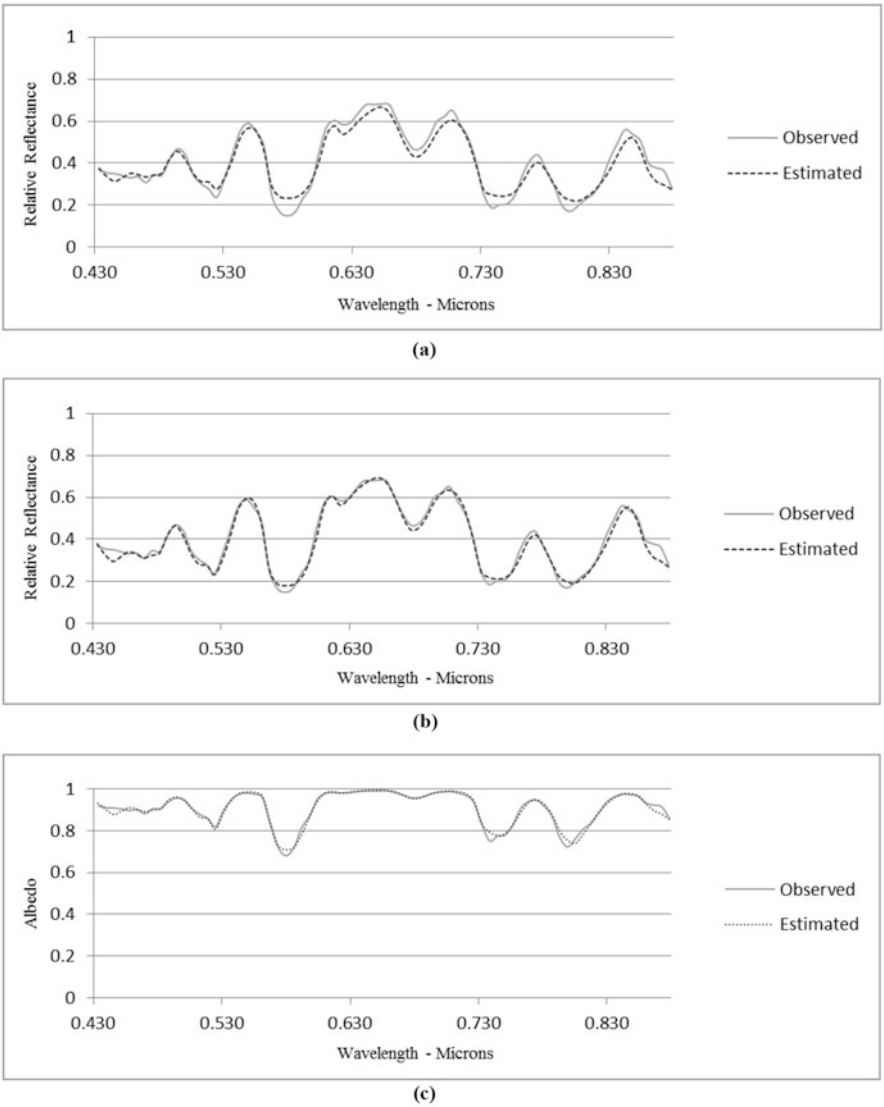


Fig. 7 The observed and estimated mixture spectra (averaged) of the 50/50% region using the FCLS, GKLS, and SSA methods. The *y*-axis of (**a**) FCLS method and (**b**) GKLS method with γ-parameter G = 5 is in reflectance units with a range 0.0–1.0; the *y*-axis of (**c**) SSA method is in albedo units with a range 0.0–1.0.

Specifically, the prediction of GKLS at G = 0.1 is almost exactly the same as the FCLS method. Out of the eight gamma values tested, GKLS at G = 5.0 provides the closest prediction for DiDy50 (49.08% vs. 50.5%) and is only slightly worse at predicting the correct abundance than GKLS at G = 6.0.

Table 1 Abundance results: The average estimated abundances are listed for the FCLS, GKLS, and SSA methods in the five test regions.

	FCLS	G = 0.1	G = 0.5	G = 1	G = 2	G = 3	G = 4	G = 5	G = 6	G = Auto	SSA
DiDy100	0.97	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.97	0.98
DiDy075	0.93	0.93	0.93	0.92	0.91	0.89	0.88	0.86	0.84	0.90	0.82
DiDy050	0.75	0.75	0.73	0.70	0.65	0.59	0.54	0.49	0.44	0.55	0.40
DiDy025	0.64	0.63	0.61	0.58	0.52	0.47	0.41	0.36	0.31	0.33	0.30
DiDy000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

The "truth" for these regions (actual physically measured proportions by volume) is 1.0, 0.788, 0.505, 0.242, and 0.0 for DiDy100, DiD075, DiDy050, DiDy025, and DiDy000, respectively. Results for the GKLS method using a fixed gamma of G = 0.1, G = 0.5, G = 1.0, G = 2.0, G = 3.0, G = 4.0, G = 5.0, and G = 6.0 are given, as well as for the automated GKLS

Table 2 Model diagnostics: The Root Mean Square Error (RMSE) results of the fit between the estimated and observed spectral mixtures are listed for selected points in the scene.

(x, y) DiDy%	FCLS	GKLS G = 5	GKLS G = 6	SSA
(585,248) 100%	**0.0141**	0.0293	0.0369	0.0315
(400,228) 75%	0.0245	**0.0244**	0.0248	0.0351
(418,239) 75%	0.0183	**0.0177**	0.0324	0.0262
(405,251) 75%	**0.0179**	0.0196	0.0205	0.0255
(324,049) 50%	0.0357	0.0225	0.0222	**0.0129**
(313, 062) 50%	0.0359	0.0318	0.0390	**0.0118**
(328, 067) 50%	0.0333	0.0294	0.0533	**0.0223**
(223,314) 25%	0.0526	0.0380	0.0379	**0.0140**
(224,327) 25%	0.0495	0.0383	0.0384	**0.0186**
(246,330) 25%	0.0420	0.0289	0.0636	**0.0121**

The first column lists the location and planned percentage mix of DiDy for these points. The actual physically measured mixes were 100%, 78.8%, 50.5%, 24.2%, and 0.0%
Bold values indicate the minimum RMSE value for each of the designated pixel locations

Figure 7 shows the observed and estimated mixture (averaged) spectra of the 50/50% region using the FCLS, GKLS, and SSA methods. Visually, we can see both the GKLS (G = 5) and SSA methods provide a better fit as compared to the FCLS method.

Table 2 lists the Root Mean Square Error (RMSE) of the fit between the estimated and observed spectral mixtures for selected points in the scene. Except for DiDy at 25% (0.242), the RMSE errors for FCLS were not considerably larger than the errors for the other methods. Yet we know FCLS is poorly predicting the known abundances for these samples. We conclude RMSE is not necessarily a good indicator of a method's accuracy to predict abundance. As far as the GKLS method is concerned, in most cases, a γ-parameter of G = 5.0 provides a better fit than G = 6.0. Noting that G = 5 provides a better prediction of abundance as compared to GKLS at the other values of γ and also provides a smaller RMSE as compared to GKLS at G = 6.0, we henceforth consider G = 5.0 to provide the best GKLS result.

In Tables 1 and 2, the results also show (unfortunately) the automated implementation of the GKLS method, which attempts to select the most appropriate gamma based on achieving a minimum of the model's RMSE, was not as successful as the fixed gamma GKLS (G = 3, 4, 5, or 6) for estimating the correct abundance. This automated GKLS method attempts to select the most appropriate gamma based on achieving a minimum of the model's RMSE. We conclude the RMSE metric seems to respond to a mixture being linear or nonlinear, but unfortunately, it is not a reliable metric to determine the degree of nonlinear behavior. RMSE could not be used to achieve the most accurate estimate of abundance. Consequently, RMSE cannot be considered effective for implementing an automated GKLS.

5 Concluding Remarks

This study has investigated the use of a Generalized Kernel Least Squares (GKLS) method applied to reflectance data, and a Single Scattering Albedo (SSA) method for nonlinear mixture analysis. In the case of the SSA method, a Fully Constrained Least Squares (FCLS) method is applied to data that has been converted from reflectance space to SSA space. Our baseline method was the FCLS method applied to reflectance data. Our hypothesis is that, for intimate (nonlinear) mixtures, both of these methods will provide improved modeling and abundance estimates as compared to the baseline FCLS method.

Overall the results for our laboratory experiment indicate the FCLS method has a poor capability for modeling intimate mixtures. In contrast, both the GKLS and SSA methods do a much better job. Whether or not one is better than the other is not conclusive. However, we conclude that our hypothesis is confirmed and that both of these methods provide a better estimate of abundance for mixtures exhibiting nonlinearity. For the laboratory experiment of known abundance quantities, the SSA and GKLS methods responded well to the nonlinearity present in a mixture of materials and provided better estimates of abundance than the linear FCLS method for the DiDy and Lime materials.

The GKLS parameter "gamma" determines the degree of nonlinear behavior exhibited by the GKLS method and affects its accuracy for estimating abundances. The automated GKLS method attempts to select the most appropriate gamma based on achieving a minimum of the model's RMSE. We conclude the RMSE metric seems to respond to a mixture being linear and nonlinear, but unfortunately it is not a reliable metric to determine the degree of nonlinear behavior. It could not be used to achieve the most accurate estimate of abundance and it is not recommended as a metric to automate the GKLS method.

For mixtures known to be nonlinear: A fixed gamma implementation of GKLS with $G = 5$ or 6 provides a good estimate of abundance. The fixed gamma GKLS and the SSA methods can be computed in approximately the same amount of time and provide approximately the same accuracy for estimating abundances. The automated GKLS was much slower to compute and did not achieve better accuracy. Further work has since been performed by the authors that elaborate on these results in an expanded study with additional experiments [30].

Acknowledgements The MITRE Innovation Program (MIP) is gratefully acknowledged for funding the HSI Microscopy aspect of the project in which the study presented here was conducted.

This book chapter has been approved for public release by NGA (Case Number 16-216).

References

1. J. Adams, M. Smith, P. Johnson, Spectral mixture modeling: a new analysis of rock and soil types at the Viking Lander 1 Site. J. Geophys. Res. **91**(B8), 8098–8112 (1986)
2. J. Boardman, in *Automating linear mixture analysis of imaging spectrometry data*. Proceedings of the International Symposium on Spectral Sensing Research (ISSSR), San Diego, CA (1994)
3. R.S. Rand, in *A physically-constrained localized linear mixing model for TERCAT applications*. Proceedings of the SPIE Aerosense, Orlando, FL (2003)
4. R.S. Rand, in *Automated classification of built-up areas using neural networks and subpixel demixing methods on multispectral/hyperspectral data*. Proceedings of the 23rd Annual Conference of the Remote Sensing Society (RSS97), Reading, United Kingdom (1997)
5. R.S. Rand, in *Exploitation of hyperspectral data using discriminants and constrained linear subpixel demixing to perform automated material identification*. Proceedings of the International Symposium on Spectral Sensing Research (ISSSR), Melbourne, Australia (1995)
6. R.S. Rand, D.M. Keenan, A spectral mixture process conditioned by Gibbs-based partitioning. IEEE Trans. Geosci. Remote Sens. **39**(7), 1421–1434 (2001)
7. D.C. Heinz, C.-I. Chang, Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **39**(3), 529–545 (2001)
8. D. Montgomery, E. Peck, *Introduction to Linear Regression Analysis*, Wiley Series in Probability and Mathematical Statistics, 2nd edn. (Wiley, New York, NY, 1992)
9. B. Hapke, *Theory of Reflectance and Emittance Spectroscopy* (Cambridge University Press, Cambridge, 1993.) 455 p.
10. J.F. Mustard, C.M. Pieters, Photometric phase functions of common geologic minerals and application to quantitative analysis of mineral mixture reflectance spectra. J. Geophys. Res. **94**, 13619–13634 (1989)
11. S.G. Herzog, J.F. Mustard, Reflectance spectra of five component mineral mixtures: implications for mixture modeling. Lunar Planet. Sci. XXVII **27**, 535–536 (1996)
12. R.G. Resmini, W.R. Graver, M.E. Kappus, M.E. Anderson, in *Constrained energy minimization applied to apparent reflectance and single-scattering albedo spectra: a comparison*, ed. By S. Shen Sylvia. Proceedings of the SPIE: Hyperspectral Remote Sensing and Applications, vol. 2821 (1996), pp. 3–13, Denver, Colo., August 5–6, doi: 10.1117/12.257168
13. R.G. Resmini, in *Enhanced detection of objects in shade using a single-scattering albedo transformation applied to airborne imaging spectrometer data*. The International Symposium on Spectral Sensing Research, San Diego, California, CD-ROM (1997), 7 p.
14. J.M.P. Nascimento, J.M. Bioucas-Dias. Unmixing hyperspectral intimate mixtures. Proc. SPIE. **7830**, 8 (2010). doi: 10.1117/12.8651188
15. H. Kwon, N.M. Nasrabadi, Kernel matched subspace detectors for hyperspectral target detection. IEEE Trans. Pattern Anal. Mach. Intell. **28**(2), 178–194 (2006)
16. G. Camps-Valls, L. Bruzzone, Kernel-based methods for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **43**(6), 1351–1362 (2005)
17. B. Scholkopf, A.J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (The MIT Press, Cambridge, MA, 2002)
18. J.B. Broadwater, R. Chellappa, A. Banerjee, P. Burlina, in *Kernel fully constrained least squares abundance estimates*. Proceedings of the IEEE International Geoscience and Remote Symposium (IGARSS 2007), Barcelona, Spain (2007), pp. 4091–4044
19. J.B. Broadwater, A. Banerjee, in *A generalized kernel for areal and intimate mixtures*. Proceedings of the IEEE WHISPERS '10, Reykjavik, Iceland (2010)
20. J.B. Broadwater, A. Banerjee, in *Mapping intimate mixtures using an adaptive kernel-based technique*. Proceedings of the IEEE WHISPERS '11, Lisbon, Portugal (2011)
21. J.B. Broadwater, A. Banerjee, in *A comparison of kernel functions for intimate mixture models*. Proceedings of the IEEE WHISPERS '09, Grenoble, France (2009)

22. R.S. Rand, A. Banerjee, J. Broadwater, in *Automated endmember determination and adaptive spectral mixture analysis using kernel methods*. Proceedings of SPIE, Optics and Photonics, San Diego, CA, August (2013)
23. R.G. Resmini, R.S. Rand, D.W. Allen, C.J. Deloy, in *An analysis of the nonlinear spectral mixing of didymium and soda lime glass beads using hyperspectral imagery (HSI) microscopy*. Proceedings of SPIE 9088, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XX, 9088OZ (2014), 15p. doi: 10.1117/12.2051434
24. R.P. Brent, *Algorithms for Minimization Without Derivatives* (Prentice-Hall, Englewood Cliffs, 1973)
25. Photographs in Figures 1 and 2 taken by the co-author Dr. David W. Allen of NIST and owned by the U.S. Government.
26. http://www.resonon.com/imagers_pika_iii.html (last Accessed on 3 Dec 2013)
27. We have also used an Edmund Optics Gold Series 1.0X telecentric lens that gives 8 μm/pixel. However, data at such a high spatial resolution were not required for the analyses reported upon here
28. Note: References are made to certain commercially available products in this paper to adequately specify the experimental procedures involved. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that these products are the best for the purpose specified
29. J.R. Schott, *Remote Sensing: The Image Chain Approach*, 2nd edn. (Oxford University Press, New York, NY, 2007.) 688 p.
30. R.S. Rand, R.G. Resmini, D.W. Allen, Modeling linear and intimate mixtures of materials in hyperspectral imagery with single—scattering Albedo and Kernel approaches. J. Appl. Remote Sens. **11**(1), 016005 (2017). doi:10.1117/1.JRS.11.016005

An Application of Spectral Regularization to Machine Learning and Cancer Classification

Mark Kon and Louise A. Raphael

Abstract We adapt supervised statistical machine learning methods to regularize noisy unsupervised feature vectors. Theorems on two graph-based denoising approaches taken from numerical analysis and harmonic spectral methods are discussed. A feature vector $\mathbf{x} = (x_1, \ldots, x_p) = \{x_q\}_{q=1}^p$ is viewed as a function $f(q)$ on its index set. This function can be regularized or smoothed using a graph/metric structure on the index set. This smoothing can involve a penalty functional on feature vectors analogous to those in statistical learning. Our regularization of feature vectors is independent of their role in subsequent supervised learning tasks. An application is given to cancer prediction/classification in computational biology.

Keywords statistical learning • kernel methods • regularization • cancer classification

1 Introduction

Regularization of noisy and partial information is an important problem in machine learning, studied widely in the area of supervised learning. We will use methods parallel to such supervised regularization, to denoise (unsupervised) input feature (data) vectors $\mathbf{x} = (x_1, \ldots, x_p)$ using prior information. We will illustrate this by regularizing feature vectors, using adaptations of two standard function denoising methods, local averaging (from numerical analysis) and support vector regression (from the theory of Tikhonov regularization).

In general the set of indices $\{1, \ldots, p\}$ (the *index space*) of a feature vector $\mathbf{x} = (x_1, \ldots, x_p)$ is an ordered set without additional structure. However, in high dimension ($p >> 1$) there are often helpful prior structures. If the index space is discrete with a notion of proximity (e.g., with a metric or a graph/network structure),

M. Kon (✉)
Department of Mathematics and Statistics, Boston University, Boston, MA 02215, USA
e-mail: mkon@bu.edu

L.A. Raphael
Department of Mathematics, Howard University, Washington, DC 20059, USA
e-mail: LRaphael@howard.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4_6

feature vectors become functions on a graph or metric structure G, often satisfying some notions of continuity. This viewpoint can improve denoising feature vectors and thus their subsequent classification. We give an example in gene expression analysis for cancer classification using the set of human genes (genome) as the index space for feature vectors. There the network structure is based on prior knowledge of interactions of genes via their protein products' interactions.

1.1 Machine Learning

In machine learning (ML) classification, *feature vectors* $\mathbf{x} = (x_1, \ldots, x_p)$ (numbers characterizing an object, e.g. a cancer sample) are mapped by a *classification function* $f(\mathbf{x})$ into appropriate classes $y = 1, \ldots, k$ (e.g., cancer subtypes). The function f is learned from a *training set* $T = (\mathbf{x}_i, y_i)_{i=1}^n$ of sample feature vectors $\mathbf{x}_i$ and their (correct) classes y_i. This function can ultimately be *tested* when it is applied to new feature vectors $\mathbf{x}$, and its reliability (e.g., percentage correct) in predicting their classes y is ascertained.

Constructing a classification function $y \approx f(\mathbf{x})$ for noisy and partial information is a central problem in ML. There are two major branches: *supervised learning* involves learning a class predictor function $f(\mathbf{x})$ from the examples in T above. Regularization in supervised learning chooses better $f(\mathbf{x})$ by combining the information in T with additional prior (supervised) knowledge (e.g., that f is smooth). When a penalty for non-adherence to this prior information (e.g., smoothness) is involved, this is known as *Tikhonov regularization* [26]. Local averaging methods to enforce continuity/smoothness of functions include kernel smoothing and local averaging

Unsupervised learning finds structure in the (unclassified) training inputs $\{\mathbf{x}_i\}_{i=1}^n$ themselves, with no information on their classes y_i. For high dimension p such structures can be complicated. In computational biology it is not unusual for $p = 10^5$ indices to exist, each index representing a gene, location in the genome, or protein.

Feature data $\mathbf{x}$ (e.g., gene expression level vectors) are often unreliable and noisy, and ML classification methods have often reached limiting accuracies on some widely studied benchmark ML datasets. We propose the use of *unsupervised regularization*, incorporation of prior structural information on feature vectors $\mathbf{x}$ without reference to their classes y. Some potentially useful applications involve denoising of feature vectors $\mathbf{x}$ using Tikhonov and other regularization methods, kernel smoothing, and local averaging methods adapted from those used to regularize functions $f : \mathbb{R}^p \to \mathbb{R}$ in supervised machine learning [9].

This approach can be used to adapt Lagrangian optimization functionals and other inference methods from supervised learning, as well as denoising methods in functional/numerical analysis for functions on $\mathbb{R}^p$. It treats feature vectors $\{x_q\}_{q=1}^n$ as functions of their indices q, and imposes continuity and other regularity constraints with respect to graph or other proximity measures in q. Two standard regularization methods on $\mathbb{R}^p$, local averaging and support vector regression, are

adapted here to unsupervised methods. This produces improved feature vectors and consequently better classification/regression using these in supervised classification tasks. We finish with an example in gene expression analysis for cancer classification with the genome as index space for feature vectors representing gene expression. We view noise in data as a source of complexity – feature vector regularization denoises by seeking less complex data forms.

1.2 Approach

Our approach to unsupervised regularization of noisy feature vectors $\mathbf{x} = (x_1, \ldots, x_p)$ typically assumes high data dimensionality p. We assume a prior graph (network) structure G with nodes formed by the (fixed) feature index set $V = \{1, \ldots, p\}$. The edge weights $\{w_{qr}\}_{q,r \in V}$ are based on prior information about the objects (e.g., genes) representing the feature index values $q \in V$, and their mutual relations. Our unsupervised regularization is a pre-processing step preparing feature vectors that are de-noised for training in subsequent ML classification tasks (e.g., identifying/predicting cancer subtypes). The regularization quality can be benchmarked by accuracy of such subsequent tasks.

1.3 Prior Work

Our theorems and application are motivated by the research of a number of mathematicians and computational biologists.

Local averaging over adjacent data locations has often been used to maximally quench noise (variance) and minimally add bias (systematic error) to data vectors. It has been used by computational biologists Ideker et al. [2, 12, 15] and Kasif et al. [13] on gene expression data – see also Section 3 below.

The approach of Ideker et al. [2, 12] uses supervised methods (with known cancer classes y combined with measured gene expression patterns $\mathbf{x}$) to identify groups of genes over which it helps to average gene expression signals x_q. This involves inputting a full training dataset (including class information) rather than just feature vectors. Local cluster averaging (i.e., combining and averaging gene expressions) is done using the protein-protein interaction (PPI) network. This network is a structure with the human genes as nodes, based on chemical interactions of their protein products. This method shows effectiveness of imposing closeness structures on feature indices (e.g., genes), here with supervised methods. The supervised index clusters are based on knowing classifications of tumor samples as metastatic/nonmetastatic so as to maximally differentiate feature vectors $\mathbf{x}$ in the two classes for later testing. The methods of Kasif et al. [13] used averaging based on biochemical pathway membership of genes.

A *spectral approach* to denoising gene expression was used by Rapaport and Vert [18, 29], and Belkin [1]. The approach of [18, 29] imposed structural constraints on gene expression feature vectors using prior known similarities in expressions among genes, to form a graph/network structure on them. This denoised the vectors based on smoothness constraints using spectral projections of the graph Laplacian. We will extend this type of smoothing to other adaptations of supervised function denoising methods.

Regularization of noisy high dimensional data has been studied widely in the functional and numerical analysis literature; see Tikhonov [26], Vapnik [28], Hastie and Tibshirani [9]. Kernel smoothing is standard for estimating or denoising real valued functions f, either fully or partially defined. The estimated function's required level of smoothness is usually determined by a single parameter. Support vector regression and some types of Tikhonov regularization use kernel smoothing in a principled optimization procedure, using the kernel function $K(\mathbf{x}, \mathbf{y})$ of a reproducing kernel Hilbert space.

Tikhonov regularization is important in supervised statistical machine learning. When partial information about a function $f(\mathbf{x})$ is known, the ill-posed problem of inverting this to a unique estimate of f is solved by adding a regularization requirement minimizing an expression in f, usually involving a norm (Vapnik [28]). The ideas of Tikhonov were extended by Krukovskii [11], who showed that if $f(\mathbf{x})$ is *fully* measured after it is perturbed into a noisy version $f_1(\mathbf{x})$, then regularizing f_1 can largely recover $f(\mathbf{x})$, reducing error $f_1 - f$. This was made precise as an asymptotic statement illustrating the proper scaling of the regularization with the size of the error.

Tikhonov's work on regularizing ill-posed problems was seminal and related to both of the above methods (averaging and support vector regression). More recent work in regularization methods based on this includes work ranging from Nashed and Wahba [16], Cuker and Smale [4], Vapnik [28], Hastie and Tibshirani [9, 25], Scholkopf and Smola [22], Smola and Kondor [23] to DX Zhou [31].

Our work extends some of the above seminal ideas. We illustrate the methods on the benchmark cancer metastasis data sets of Wang et al. [30] and van de Vijver et al. [27]. These are widely studied in computational biology as a context for supervised learning related to cancer prediction and classification. Our example of denoising their gene expression is based on regularization using a protein-protein interaction (PPI) network.

1.4 Paper Contents

In Section 2 we state four major theorems on two regularization approaches (*local averaging and support vector regularization*). These indicate conditions where regularization improves denoising accuracy. We show that as the regularization parameter (say α) increases, accuracy generically improves and then decreases. For the second (support vector regression) method we also give "smoothness"

conditions on the underlying (pre-noisy) feature vector f for it to be amenable to regularization when contaminated by noise $\epsilon g(q)$. The regularization process effectively diminishes high frequency components of the true signal f (increasing estimation bias error) in exchange for a reduction in noise (decrease in variance error). Proofs of two theorems are sketched.

In Section 3 we show this method for pre-processing feature vectors improves classification accuracy in subsequent supervised learning tasks. The task is predicting cancer metastasis using gene expression feature vectors, with a graph structure on genes based on PPI. The data are from the above breast cancer studies ([27, 30]).

2 Denoising Theorems

Within the larger class of regularization methods, we will discuss here two *graph-based adaptations* of standard regularization approaches for denoising functions $f : \mathbb{R}^p \to \mathbb{R}$.

A standard numerical analysis method for denoising a (fully or partially) measured function $f(\mathbf{x})$ ($\mathbf{x} \in \mathbb{R}^p$) is to average (smooth) f over adjacent locations. This is done, for example, in regularization of noisy photo images using blurring (e.g., Gaussian convolution) or penalty functional (Tikhonov) regularizations. Convolution-based regularization of functions on $\mathbb{R}^p$ is standard in image processing, where blurring (softening) an image improves visual information. See S. Geman et al [7] for regularization of images and also Coifman and Donoho [3], Coifman [24].

On a general graph G (analogous to definitions on $\mathbb{R}^p$), let $f(q)$ be a function on the graph G (i.e., on its vertices $q \in V$), forming a feature vector $x_q = f(q)$. We assume the measured values $f_1(q) = f(q) + \eta(q)$ of f are contaminated by variability $\eta(q)$, which can represent noise or measurement inconsistencies. Thus $f(q)$ is a underlying (true) signal, with $f_1(q)$ a measured approximation. Our goal is a regularization $R_\alpha(f_1)$ of the perturbed f_1 (with regularization parameter α), to optimally recover $f(q)$. Thus we want $R_\alpha(f_1) = R_\alpha(f) + R_\alpha(\eta)$ to quench error η by minimizing $R_\alpha(\eta)$ (reducing variance) and minimally bias (systematically change) the original f in the regularized signal $R_\alpha(f)$.

In this type of bias-variance tradeoff, the error is often U-shaped in the regularization parameter α. For small α, a learning algorithm learns too many nuances from its data. It will typically overfit and be thrown off by noise (leading to high variance error in estimates), interpreting noise as a function signal. However it will typically have low systematic *averaged* error (low bias). With large regularization α a learning algorithm imports high prior information on structure of the underlying function $f(x)$, and is less sensitive to noise (thus low variance) but may impose systematic error (high bias) on resulting estimates. A good supervised learning method properly balances this tradeoff between bias and variance, on $\mathbb{R}^p$ or any proximity structure. We will discuss theorems (analogous to those for supervised denoising on $\mathbb{R}^p$) on unsupervised feature vector regularization, here assuming graph structures on indices.

Our theorems will present the pattern of bias-variance tradeoff, and existence of a unique minimum for L^2 estimation error in the regularization parameter. This is done for two denoising methods: local averaging and support vector regression (SVR), assuming graph structures on feature vector indices. We also show that for appropriate prior knowledge on f (here smoothness with respect to the prior graph structure) both methods give good results.

We show that reconstruction accuracy for functions on graphs is non-monotonic, first increasing and then decreasing in α, with best accuracy at an intermediate regularization. Feature vector regularization inherits from Tikhonov denoising on $\mathbb{R}^p$, the same non-monotonicity of reconstruction accuracy - the best is at a finite positive value α. We view noise as a source of complexity, and regularization denoises by transforming feature vectors into more useful ones.

Local averaging of feature vector entries $x_q = f(q)$ with their (graph) neighboring entries cancels out noise, thereby reinforcing similarity of neighbors and canceling individually high errors. When the feature vector forms a visual image this can be done assuming adjacent pixels (feature indices) have close illumination levels (feature values). The bias from blurring is more than offset by the variance reduction (noise averaging). There is an optimal amount of blurring (too little or too much will reduce recovery). On $\mathbb{R}^p$ this is the foundation of Haar wavelet methods for piecewise constant function approximation (e.g. [14]). Our second method, support vector regression (SVR) on feature vectors, uses a penalty functional like those in statistical learning for supervised functions $f : \mathbb{R}^p \to \mathbb{R}$, based on the index set graph structure. SVR in the graph case regularizes feature vectors using the same Tikhonov-type regularization used for $\mathbb{R}^p$ functions.

Note the accuracy of SVR regularization/denoising of feature vectors cannot be measured directly, since noise is unknown. One way to test effectiveness is measuring performance of the regularized data when they are used in the training and testing of subsequent supervised learning tasks (Section 3).

2.1 Statements of Theorems

We consider regularization of feature vectors $\mathbf{x} = (x_1, \ldots, x_p)$ on finite graph structures G on (large) index sets $V = \{1, \ldots, p\}$. The $L^2(V)$ norm will measure error between the underlying $f(q)$ and recovery $R_\alpha(f_1)$ of its perturbation $f_1 = f + \eta$. The theorems below will assume noise $\eta(q) = \epsilon g(q)$, where $g(q)$ are standard $N(0, 1)$ normal random variables that are independent and identically distributed (iid).

The graph structure on the index set represents prior expected similarity between features (e.g., gene-gene similarity). We will show good graph structures yield good regularized recoveries of feature vectors $\mathbf{x} = (x_1, \ldots, x_p)$, with $x_q = f(q)$. As the regularization level increases, the error of recovering the underlying (pre-noisy) feature vector f from regularizing the perturbed vector f_1 first decreases, as variance is reduced while bias remains controlled. The error then increases, as bias increases beyond benefits of variance reduction. Our theorems show that

regularization/denoising approaches can be applied just as well to unsupervised data using prior graph structure data. The two regularization methods considered here are local averaging and support vector regression, modified to accommodate graph structures for feature vectors.

2.1.1 Method 1: Local averaging on a graph

We consider local averaging as a regularization approach. A graph $G = (V, E)$ consists of vertices $v \in V$ together with edges $e_{ij} \in E$, with e_{ij} connecting vertices i and j. The edges are associated with non-negative weights w_{ij} representing vertex similarity. In some cases we denote the vertex set as G when there is no ambiguity. We view feature vectors $\mathbf{x} = (x_1, \ldots, x_p)$ that are indexed by the graph G, i.e., by its vertices $V = 1, \ldots, p$, as functions $\{f(q) = x_q\}_{q \in G}$. The norm is $||f||^2 = \sum_{q \in G} f^2(q)$ unless stated otherwise. We consider the underlying (noise-free) feature vector $f(q)$ perturbed by noise or other effects, giving measured values $f(q) + \epsilon g(q) = f_1(q)$, with noise $\eta(q) = \epsilon g(q)$. We seek a regularization operation R_α (with α the *regularization parameter*) such that the regularized feature vector $R_\alpha f_1(q)$ approximately recovers f.

Thus $f(q) = x_q$ is a real-valued function of $q \in G$. The perturbation $\epsilon g(q)$ is generated from an independent standard $N(0, 1)$ Gaussian $g(q)$ for each $q \in G$. We will cluster features x_q (more properly the indices q) into a collection $\mathrm{Cl}_t = \{a_{ti}\}_{i=1}^{k_t}$ of clusters forming a partition $\bigcup_{i=1}^{k_t} a_{ti} = G$. The partitions are hierarchical, so that each for each $t' < t$, $a_{t'i}$ is a union of sets of the form a_{tj}. Equivalently let $\mathscr{F}_t$ be the finite σ-field of sets generated by Cl_t above. Then for $t' > t$, we have $\mathscr{F}_t \subset \mathscr{F}_{t'}$.

Here the discrete parameter $t = 1, \ldots T$ is a regularization parameter, corresponding to the (hierarchical) level of clustering. We will cluster-average $f_1(q)$ to obtain the averaged function $R_t f_1 = f_{1t}(q) = \mathbb{E}(f_1(q)|\mathscr{F}_t)$. The latter is the (probabilistic) conditional expectation of f_1 with respect to σ-field $\mathscr{F}_t$. We will always assume that Cl_t is a (proper) subpartition of Cl_{t-1}. A sequence of σ-fields $\{\mathscr{F}_t\}_t$ in which $\mathscr{F}_t$ is a refinement of $\mathscr{F}_{t-1}$ is known as a *filtration*.

We note that *decreasing* t represents fewer (larger) clusters and greater regularization. The increasing number k_t of clusters with t represents *less* regularization for larger t. The *highest* regularization is at $t = 1$ (assumed to have $k_1 = 1$ clusters). The *lowest* regularization is at $t = T$ (with $k_T = p$ clusters, i.e., one cluster per feature).

We define $f_t = \mathbb{E}(f|\mathscr{F}_t)$, and $g_t = \mathbb{E}(g|\mathscr{F}_t)$. We form the regularization of the noisy $f_1 = f(q) + \epsilon g(q)$ by cluster-averaging it to obtain

$$R_t(f_1) = \mathbb{E}(f_1|\mathscr{F}_t)$$

Below and henceforth $\mathbb{E}$ represents ordinary expectation (with respect to the random family $g(q)$) (note f is not random) while $\mathbb{E}(\cdot|\mathscr{F}_t)$ represents a conditional

expectation (cluster-average) of the argument function, defined above. By convention $\mathbb{E}\|\cdot\|^2 = \mathbb{E}\left(\|\cdot\|^2\right)$, and the norm $\|\cdot\| = \|\cdot\|_2$ denotes L^2 norm unless otherwise specified.

To study the relationship between cluster size and regularization, we need a basic

Lemma 1 *The error e of regularization of the level t cluster approximation satisfies*

$$e(Cl,t) \equiv \|f_{1t} - f\|^2 = \mathbb{E}\left(\|\mathbb{E}(f(q) + \epsilon g(q))|\mathscr{F}_t) - f\|^2\right)$$

$$\equiv \mathbb{E}\left(\|f_t(q) + \epsilon g_t(q) - f\|^2\right) = \|f_t - f\|^2 + \epsilon^2 \mathbb{E}\left(\|g_t(q)\|^2\right). \tag{1}$$

This lemma separates expected error into bias error $\|f_t - f\|^2$ and variance error $\epsilon^2\mathbb{E}\|g_t(q)\|^2$. Note here and below that when clustering level t is fixed we will denote the clusters a_{ti} as a_i.

Theorem 1 (Graph structures for feature vector averaging) *Let $x_q = f_1(q) = f(q) + \epsilon g(q)$ be a noisy feature vector associated with signal $f(q)$, with $g(q) \sim N(0,1)$ independent standard Gaussian noise for each q in a graph G. For fixed clustering level t and signal f, let $\mathrm{Cl}_t = \{a_i\}_{i=1}^k$ denote a clustering of G yielding a regularization $f_{1t} \equiv \mathbb{E}(f_1|\mathscr{F}_t)$ of f_1.*

Then the estimation error is

$$\mathbb{E}(||f_{1t} - f||^2) = \mathbb{E}(\mathbb{V}(f|\mathscr{F}_t)) + k_t\epsilon^2, \tag{2}$$

where the conditional variance $\mathbb{V}(f|\mathscr{F}_t)$ is the size-averaged variance of $f(q)$ over the family Cl_t of clusters $\{a_i\}_{i=1}^{k_t}$:

$$\mathbb{V}(f|\mathscr{F}_t) = \sum_{i=1}^{k_t} \mathbb{V}(\{f(q)\}_{q\in a_i})|a_i| \tag{3}$$

with $k_t = |\mathrm{Cl}_t|$ the number of clusters.

The conditional variance is equivalent to the weighted sum of the variances $x_q = f(q)$ of features in the feature vector within each cluster a_i of features, weighted by the number $|a_i|$ of features.

The corollaries below refer to a *fixed feature vector f* (thus a fixed index set), and a fixed regularization level t (number of clusters k_t).

Corollary 1 *Assume a fixed feature vector $f(q)$ and regularization level t (i.e., fixed number k_t of feature clusters) on the index space G. The error $\mathbb{E}(||f_{1t} - f||^2)$ is then reduced by any clustering Cl_t that is improved, as measured by reduced conditional variability $\mathbb{V}(f|\mathscr{F}_t)$ of f (with respect to the corresponding finite σ-field $\mathscr{F}_t$.)*

Corollary 2 *For fixed f and regularization level t, optimal (minimal) error clustering divides the features into k_t groups such that the weighted variances $\mathbb{V}\{x_q\}_{q\in a_i}$ of individual groups (weighted by the cluster sizes $|a_i|$) sum to the smallest total.*

Corollary 3 *Assume a fixed underlying feature vector $f(q)$, regularization level t, and a fixed mapping of graph structure G (weight matrix $W = \{w_{ij}\}$) to clustering $\mathcal{F}_{G,t}$. As graph structure G varies, regularization error $\|f_{1t} - f\|^2$ is monotone decreasing in $\mathbb{V}(f) - \mathbb{E}(\mathbb{V}(f|\mathcal{F}_{G,t}))$, i.e., the relative regularity provided to the underlying f by G, where $\mathbb{V}(f) = \mathbb{V}(f|\mathcal{F}_{G,T})$ is the variance of $f(q)$.*

(Recall final level T clustering has clusters all of size 1). We now focus on variable sizes and numbers k_t of clusters, or equivalently variable regularization levels t. We will show that, for a variety of potential graph structures G imposed on a feature vector f (with G not necessarily tuned to make f smooth), clustering regularization can nevertheless help recover f from its noisy perturbation f_1. However this must be done with a level of regularization that will cancel enough noise without introducing too much bias.

As the clustering parameter t *decreases* (smaller cluster count k_t), regularization *increases*, and at some point $k_t = k_{t_{\min}}$ is optimal. At this regularization level the bias $||f - f_t||^2$ cancels the variance $\epsilon^2||g_t||^2$ optimally. (Here $f_t = \mathbb{E}(f|\mathcal{F}_t)$ is the averaged version of the underlying signal f, while g_t is the locally averaged noise.) The level $t_{\min}$ is the point where error is minimized. Decreasing t (so $t < t_{\min}$ and $k_t \leq k_{t_{\min}}$), allows increased bias to take over, and error again begins to increase. Generically then, in terms of increasing t, total error $||f_{1t} - f||$ decreases monotonically for $t < t_{\min}$, and increases for $t > t_{\min}$. This is made precise in

Theorem 2 (Optimal regularization for feature vector averaging) *Consider all underlying feature vectors (feature functions) $x_q = f(q)$ whose index values q form finite graphs G, together with filtrations $\{\mathcal{F}_t\}_{t=1}^T$ on each graph. Assume also that the filtrations $\mathcal{F}_t$ are chosen so that the conditional expectations $f_t = \mathbb{E}(f|\mathcal{F}_t)$ satisfy the uniformity condition*

$$||f_t - f|| - ||f_{t+1} - f|| \geq K(t) \quad (t = 0, 1, \ldots, T) \tag{4}$$

on their errors $||f_t - f||$, with $K(t)$ a fixed positive function. Assume also that the hierarchical clusterings $\{\mathrm{Cl}_t\}_{0\leq t\leq T}$ generating each filtration $\{\mathcal{F}_t\}_{0\leq t\leq T}$ have cardinalities satisfying

$$|\mathrm{Cl}_{t+1}| \geq d|\mathrm{Cl}_t| \tag{5}$$

for some $d > 1$.

For such f, let $f_1 = f(q) + \epsilon g(q)$ be perturbations (with $g(q)$ the above Gaussian noise), and $f_{1t} = \mathbb{E}(f_1|\mathcal{F}_t)$ be the cluster-averaged sequence (in regularization parameter t) for recovering $f(q)$. If $1/\epsilon$ and T (and hence $|G|$) are sufficiently large, then with probability p arbitrarily close to 1, the approximation error $||f_{1t} - f||$ of the regularized feature vector decreases for small t and increases for large t.

Thus the minimum error is achieved for a positive value of the regularization parameter t, with the same probability p approaching 1.

Proof (Sketch)
Recall we define the noise $\eta(q) = \epsilon g(q)$. The proof of this theorem requires a

Lemma 2 *On any graph G, let* $\{\mathscr{F}_t\}_{1\le t\le T}$ *be a set filtration with corresponding clustering* Cl_t *for each t, with* $|\mathrm{Cl}_t|$ *the number of clusters at level t. Assume* $|\mathrm{Cl}_{t+1}| \ge d\,|\mathrm{Cl}_t|$ *for some fixed* $d > 1$. *Letting* $\eta_t(q) = \mathbb{E}(\eta(q)|\mathscr{F}_t)$, *then uniformly over all such graphs G and filtrations (and over* $|\mathrm{Cl}_t|$*)*

$$\frac{1}{\epsilon}\mathbb{E}(\|\eta_t\|) = \sqrt{|\mathrm{Cl}_t|} + O\left(1/\sqrt{|\mathrm{Cl}_t|}\right) \quad (|\mathrm{Cl}_t| \to \infty),$$

and

$$\frac{1}{\epsilon^2}\mathbb{V}(\|\eta_t\|) = O(1). \tag{6}$$

Recall $\|\cdot\| = \|\cdot\|_2$ is the L^2 norm unless otherwise specified. The Lemma's proof uses properties of the chi squared distribution, probabilistic bounds on expectation and variance, and some gamma function identities of Graham et al. [8].

A sketch of the remainder of the proof of Theorem 2 involves the following identity. With the same definitions as in Lemma 2,

$$\begin{aligned}\|\eta_{t+1}\| - \|\eta_t\| - \|f_{t+1} - f\| - \|f_t - f\| \le \|f_{1(t+1)} - f\| - \|f_{1t} - f\| \\ \le \|f_{t+1} - f\| - \|f_t - f\| + \|\eta_{t+1}\| + \|\eta_t\|,\end{aligned} \tag{7}$$

which follows by writing $f_{1t} - f = f_t + \eta_t - f$.

We will use the first inequality in (7) to show that $\|f_{1(t+1)} - f\| - \|f_{1t} - f\|$ is increasing for large t. By Lemma 2, it can be shown that (for d as in (5)),

$$\frac{1}{\epsilon}\left[\mathbb{E}\left(\|\eta_{t+1}\|\right) - \mathbb{E}\left(\|\eta_t\|\right)\right] = \left(\sqrt{d} - 1\right)\sqrt{|\mathrm{Cl}_t|} + O\left(\frac{1}{\sqrt{|\mathrm{Cl}_t|}}\right)$$

and

$$\frac{1}{\epsilon^2}\mathbb{V}(\|\eta_{t+1}\| - \|\eta_t\|) \le \frac{2}{\epsilon^2}\mathbb{V}(\|\eta_{t+1}\|) + \frac{2}{\epsilon^2}\mathbb{V}(\|\eta_t\|) = O(1) \quad (|\mathrm{Cl}_t| \to \infty).$$

From this some calculations show

$$P(\|f_{1(t+1)} - f\| - \|f_{1t} - f\| < 0) \le P(\|\eta_{t+1}\| - \|\eta_t\| - 2\|f\| < 0) = P(B_t < 0).$$

where

$$B_t \equiv \|\eta_{t+1}\| - \|\eta_t\| - 2\|f\|.$$

However, bounding $P(B_t \leq 0)$ shows that given a t_1 sufficiently large,

$$P(B_t \leq 0 \text{ for some } t \geq t_1) \leq \sum_{t \geq t_1} P(B_t \leq 0) \leq \sum_{t \geq t_1} \frac{2L}{\left(\sqrt{d}-1\right)^2 |\mathrm{Cl}_t|}, \tag{8}$$

for a universal constant L uniform over all graphs G and cluster sizes k_t. This is sufficient to prove that the error increases for large t with high probability as desired.

For small $t = 0, 1, 2, \ldots$ we combine the bound (4) with the second inequality in (7), giving

$$\begin{aligned} \|f_{1(t+1)} - f\| - \|f_{1t} - f\| &\leq \|f_{t+1} - f\| - \|f_t - f\| + \|\eta_{t+1}\| + \|\eta_t\| \\ &\leq \|\eta_{t+1}\| + \|\eta_t\| - K(t), \end{aligned} \tag{9}$$

which can be used to prove that error is decreasing for sufficiently small $t \leq t_2$, if ϵ is sufficiently small (recall $\eta = \epsilon g$). This together with the result for large t completes a sketch of the proof.

2.1.2 Method 2: Support vector regression/regularization on a graph

Let f be a real-valued function on a domain $D \subset \mathbb{R}^p$. Let the operator A sample f, so that $Af = \mathbf{y} = (y_1, \ldots, y_N)$ where $y_i = f(\mathbf{x}_i) + \epsilon_i$ are perturbed values of f at a fixed finite subset $\{\mathbf{x}_i\}_{i=1}^N \subset D$, with errors ϵ_i. The vector $\mathbf{y}$ is noisy partial information about f, and the problem of recovering f from $\mathbf{y}$ is ill-posed. This can be solved using Tikhonov regularization [26]. The given data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are fitted, typically in the least squares sense by f, and at the same time solutions with large norms $\|f\|_K^2$ are penalized. The norm $\|\cdot\|_K = \|\cdot\|_H$ is taken in a Hilbert space H of functions on domain D. It is assumed H is a reproducing kernel Hilbert space (RKHS), i.e., that there is a unique kernel function $K(\mathbf{x}, \mathbf{y})$, $(\mathbf{x}, \mathbf{y} \in D)$ with the reproducing property that for all $f \in H$ and all fixed $x \in D$,

$$f(\mathbf{x}) = \langle K(\mathbf{x}, \cdot), f(\cdot) \rangle_H.$$

Above the dot represents the active variable in the inner product. It is also required that K be positive definite, i.e., that for any fixed finite set $\{\mathbf{x}_i\}_i \subset D$, $K_{ij} \equiv K(\mathbf{x}_i, \mathbf{x}_j)$ is a positive matrix. We denote the regular L^2 norm as $\|\cdot\| = \|\cdot\|_{L^2}$. Tikhonov regularization solves the penalized minimization problem

$$\hat{f} = \inf_{f \in H} \mathscr{L}(f) = \inf_{f \in H} \left(\sum_{i=1}^N (y_i - f(\mathbf{x}_i))^2 + \lambda ||f||_K^2 \right). \tag{10}$$

The norm $\|\cdot\|_K$ is an RKHS norm typically based on a Sobolev or Gaussian kernel K (see below). The kernel can be selected to penalize large oscillations

in f. The regularization in $\hat{f}$ can serve to "fill in" missing information (the given measurements $y_i = f(\mathbf{x}_i)$ are not taken at all points $\mathbf{x} \in D$) or regularize against additive noise $\epsilon g(\mathbf{x})$.

To develop an analogous regularization for feature vectors $x_q = f_1(q)$ (as functions on a graph-structured index set $G = \{1, \ldots, p\}$) we adapt the above Tikhonov functional. Again the feature vector $\mathbf{x} = (x_1, \ldots, x_p)$ is a function $x_q = f(q)$. We want to incorporate prior information on which pairs of features q and r are similar, i.e., when $f(q)$ and $f(r)$ should be close. With the same function and noise model as earlier, the measured feature is $x_q = f_1(q) = f(q) + \epsilon g(q)$.

The previously mentioned RKHS formulation carries over fully to functions like $f_1(q)$ on graphs. The regularization serves to diminish the noise in f_1 on the graph. Note $f(q)$ is known for all q (there are no missing values), which does not change the problem in principle from that on $\mathbb{R}^p$.

We "regularize out" noise $\epsilon g(q)$ by minimizing penalized error (10), obtaining a support vector regression estimate $\hat{f}$, defined by

$$\hat{f} \equiv R_\lambda f = \arg\min_{h \in H} \mathscr{L}(h) = \arg\min_{h \in H} \left\{ \sum_{q \in G} (f_1(q) - h(q))^2 + \lambda ||h||_H^2 \right\}$$
$$= \arg\min_{h \in H} \left\{ ||f_1 - h||^2 + \lambda ||h||_H^2 \right\}. \tag{11}$$

Define the canonical operator K with kernel $K(q, r)$ on G, so that for any function $h(q)$ on G, $(Kh)(q) = \sum_{r \in G} K(q, r)h(r)$. As an operator the reproducing kernel can be $K = e^{-t\Delta}$ (Gaussian kernel) or $K = (1 + \Delta)^{-s/2}$ (Sobolev kernel) among others, with Δ the graph Laplacian on G. Note the reproducing kernel K and inner product $\langle \cdot, \cdot \rangle_H \equiv \langle \cdot, \cdot \rangle_K$ are chosen to have the same reproducing property as on $\mathbb{R}^p$, namely that for any graph function $h(q)$,

$$h(q) = \langle K(q, \cdot), h(\cdot) \rangle_H \equiv \sum_{r \in G} K(q, r)h(r).$$

The graph Laplacian is defined as $\Delta = D - W$. The adjacency matrix $W = (w_{ij})$ has entries w_{ij} equal to the edge weight between indices $i, j \in G$. Additionally $D = \text{diag}(d_i)$ is the diagonal matrix with entry d_i equal to the weighted degree of vertex i, i.e., $d_i = \sum_j w_{ij}$. By assumption $w_{ii} = 0$.

Both of the above kernels K are smoothness-enforcing. Specifically the norms (penalties) $\|f\|_H = \|f\|_K$ they induce are large for "non-smooth" f on G. Since graph size $|G|$ is finite, H will include all real functions on G (as with any norm on a space of finite cardinality); see [21]. For the following theorem we assume either of the above (Gaussian or Sobolev) kernels K, and more generally any kernel of the form $K = \mu(\Delta)$, where μ is a real-valued non-increasing function. The matrix operator $\mu(\Delta)$ is defined by the matrix operator calculus, so if Δ is defined by the eigenvalue-eigenvector pairs $\Delta \sim \{(v_i, u_i)\}_{i=1}^p$, then $\mu(\Delta) \sim \{(\mu(v_i), u_i)\}_{i=1}^p$. Thus

if ν_i are the eigenvalues of Δ, then $\mu_i = \mu(\nu_i)$ are the eigenvalues of K. We define $\omega_i = 1/\mu_i$ to be the eigenvalues of K^{-1} (so that ω_i increases with the frequency ν_i of the Laplacian), and the operator

$$L_\lambda = (K + \lambda)^{-1} K. \tag{12}$$

Below we will show that $h = L_\lambda f_1$ minimizes (11) and thus is the recovery approximation of the underlying feature vector f. We write $L_{\lambda,W} = L_\lambda$ to make dependence of the smoothing operator L_λ on the graph structure W explicit. We will view the approximation error $\|L_{\lambda,W} f_1 - f\|_{L^2}$ as a function of the prior graph structure W for *fixed* feature vector f plus noise.

An optimized graph structure (weight matrix) W for graph G will connect node pairs q, r such that $x_q = f(q)$ and $x_r = f(r)$ tend to have similar values. For example, if q, r represent genes, this might be known because their expressions $f(q)$ and $f(r)$ are correlated based on prior measurements, or from tables of known interactions between their protein products.

Equivalently, the graph structure W is optimized if f tends to be smooth with respect to this structure on G. In this case $f(q)$ has primarily low Laplacian frequencies (with respect to G), i.e. its primary eigenfunction components have low eigenvalues. On the other hand, since the noise ϵg has no structure with respect to G, it will have higher frequency components. Thus the signal $f(q)$ and the noise $\epsilon g(q)$, being in primarily different frequency bands, can be teased apart using spectral projections of the Laplacian.

In the theorem below note the underlying (true) feature vector $f(q)$ is fixed, since it arises from a measurement. We wish to adjust the graph structure (weight matrix W) on G so as to optimally recover f.

Since the operator L_λ on $L^2(G)$ depends on the weight matrix, we will write $L_\lambda = L_{W,\lambda}$. Below we will define a useful measure of the *low frequency content* of the feature function $f(q)$ with respect to G to be $\|f\|^2 - \|(1 - L_{W,\lambda})f\|^2$ (see (16) and (17) and the discussion following). In the Theorem and a sketch of a proof below we will need to distinguish this G-based low frequency content of f, against the *intrinsic* low frequency content of G itself. The latter (see below) will be defined as $\|L_{\lambda,W} U(q)\|^2$, where $U(q) = \sum_i u_i(q)$ is the (finite) sum of all eigenfunctions of Δ. The proof below defines the terminology of low and high frequency content more carefully.

We have

Theorem 3 (Graph structures for SVR feature vector regularization) *Consider a noisy feature vector* $\mathbf{x}$ *having components* $x_q = f_1(q) = f(q) + \epsilon g(q)$, *with the last two terms representing signal and noise, and the index values* q *forming a graph* G. *Assume* $g(q)$ *are iid and* $N(0, 1)$, *and that* f *is fixed with* $\|f\|_2 = 1$. *On* G, *assume a fixed reproducing kernel Hilbert space* H *of real-valued functions with kernel* $K = \mu(\Delta)$, *with* μ *a decreasing function such as*

$K = e^{-\beta\Delta}$ (Gaussian) or $K = (1+\Delta)^{-s/2}$ (Sobolev). Let $f_{1\lambda}(q)$ be the support vector regression approximation of f, minimizing (11), and assume λ remains fixed.

As the selected weighted graph structure W on G varies, the approximation error is a decreasing function of the smoothness (in terms of the variable W) of the fixed underlying feature vector f. Specifically:

(a) for a measured $\mathbf{x} = f_1(q)$ on G, the expected feature vector regularization error $\mathbb{E}\left(\|L_{\lambda,W}f_1 - f\|^2\right)$ is decreasing in the smoothness $\|f\|^2 - \|(1-L_{\lambda,W})f\|^2$ (i.e., low frequency content) of f, if we vary the graph structure W without changing (intrinsic) graph low frequency content. *The latter is defined as $\|L_{\lambda,W}U(q)\|^2$, where $U(q) = \sum_i u_i(q)$ is the (finite) sum of the orthonormal eigenfunctions of Δ.*

(b) the graph low frequency content is also given by the trace

$$||L_\lambda U||^2 = \mathrm{tr}(1+\lambda/\mu(\Delta))^{-2} \tag{13}$$

with $K = \mu(\Delta) = e^{-t\Delta}$ for the Gaussian kernel and $\mu(\Delta) = (1+\Delta)^{-s/2}$ for the Sobolev kernel of order $s/2$.

Note above we have defined $\lambda/\mu(\Delta) = \lambda\mu(\Delta)^{-1}$. Note also that the relationship between the L^2 and K norms can be summarized by $\|f\|_K = \|K^{-1/2}f\|_{L^2}$. Hence when $K = \mu(\Delta) = (1+\Delta)^{-s/2}$, we have $\|f\|_K = \|((1+\Delta)^{s/4}f\|$, which is a Sobolev norm of order $s/2$.

Proof (Sketch)
The proof of part (a) adapts methods from Tikhonov regularization on $\mathbb{R}^p$ to functions on the index graph G, using the regularization functional (11).

Let $f(q)$ be the unperturbed (true) feature vector. We are assuming iid standard Gaussian noise $g(q)$ for each $q \in G$, with perturbed function $f(q) + \epsilon g(q) = x_q$, and ϵ the noise intensity. The graph Tikhonov functional can be optimized using the same methods as for functions on $\mathbb{R}^p$ (see [9], Section 5.8) yielding

$$\hat{f} = L_\lambda f_1 = f_{1\lambda}(q) = \sum_{r\in V} a(r)K(q,r),$$

with

$$a = (K+\lambda)^{-1}f_1.$$

Putting these together gives

$$\hat{f} = Ka = (K+\lambda)^{-1}Kf_1 \equiv L_\lambda f_1; \tag{14}$$

see (12). As above $K(q,r)$ is the reproducing kernel defining the Hilbert norm $\|\cdot\|_H = \|\cdot\|_K$ as in the regularization functional (11) (see [21]). The expected error (averaged over noise $\epsilon g(q)$) of the regularized noisy feature vector $f_{1\lambda} = L_\lambda f_1(q)$ is (see (1))

$$\mathbb{E}\left(\|f_{1\lambda} - f\|_2^2\right) = \mathbb{E}\left(\|L_\lambda f + \epsilon L_\lambda g - f\|_2^2\right) = \|f - L_\lambda f\|^2 + \epsilon^2 \mathbb{E}\left(\|L_\lambda g\|^2\right). \quad (15)$$

As above we abbreviate $L_\lambda = L_{\lambda,W}$ when dependence on W is not important.

Assume that the eigenvalues and (orthonormal) eigenfunctions of K are $\mu_i = 1/\omega_i$ and $u_i(q)$, respectively. Let the coefficients f_i be defined from the orthonormal expansion $f(q) = \sum_i f_i u_i(q)$.

Then

$$f_{1\lambda} = L_\lambda f = (K + \lambda)^{-1} K f = \sum_i f_i \frac{\mu_i}{\lambda + \mu_i} u_i(q)$$

and

$$(1 - L_\lambda) f = \sum_i f_i \frac{\lambda}{\lambda + \mu_i} u_i(q).$$

Note that in fact the error in (15) can be parsed as the sum of bias and variance

$$\begin{aligned} h(\epsilon, \lambda) \equiv \|f_{1\lambda} - f\|_2^2 &= \|(1 - L_\lambda) f\|^2 + \epsilon^2 \mathbb{E}\left(\|L_\lambda g\|^2\right) \\ &= \sum_i \frac{f_i^2 (\omega_i \lambda)^2}{(\omega_i \lambda + 1)^2} + \epsilon^2 \sum_i \frac{1}{(\lambda \omega_i + 1)^2} \equiv E_B + E_V. \end{aligned}$$

The bias error

$$\begin{aligned} E_B = \|f - L_\lambda f\|^2 &= \|(1 - L_\lambda) f\|^2 \\ &= \sum_i \frac{f_i^2 (\omega_i \lambda)^2}{(\omega_i \lambda + 1)^2} = \sum_i f_i^2 r(\lambda \omega_i) \end{aligned} \quad (16)$$

with

$$r(\omega_i \lambda) = \frac{(\omega_i \lambda)^2}{(\omega_i \lambda + 1)^2} \approx \begin{cases} 1 & \text{if } \omega_i \gg 1/\lambda \\ 0 & \text{if } \omega_i \ll 1/\lambda \end{cases}. \quad (17)$$

Thus according to (16) and (17), the regularization parameter $1/\lambda$ is an approximate (smoothed) spectral cut-off for the sum defining $\|f - L_{\lambda,W} f\|^2$.

Above E_B is small if f is smooth with respect to G, i.e., if f has primarily low frequency components, frequencies ω_i satisfying $\omega_i \lambda \ll 1$, or $\omega_i \ll 1/\lambda$ for an appropriate choice of λ in (11). This occurs if the graph structure G (the weight matrix W) is appropriately well-matched to prior knowledge about f, i.e., with high edge weights w_{qr} for index pairs q, r for which $f(q) \approx f(r)$.

In addition the second (variance error) term

$$E_V = \epsilon^2 \mathbb{E}\left(\|L_\lambda g\|^2\right) = \epsilon^2 \sum_i \frac{1}{(\lambda\omega_i + 1)^2} \equiv \epsilon^2 \sum_i s(\lambda\omega_i) \tag{18}$$

with

$$s(\lambda\omega_i) = \frac{1}{(\lambda\omega_i + 1)^2} = \begin{cases} 1 & \text{if } \omega_i \ll 1/\lambda \\ 0 & \text{if } \omega_i \gg 1/\lambda \end{cases}. \tag{19}$$

Note the noise $g(q)$ will be largely diminished by the regularization operator if f is sufficiently smooth, so that $1/\lambda$ can be made small without increasing bias error E_B in (16). This is because for such λ, E_V will have more higher frequency components ω_i relative to the cut-off $1/\lambda$ ($\omega_i \gg 1/\lambda$), which will then be cut off due to the soft thresholding in (19). In fact the above-defined low frequency content $||L_\lambda U||^2$ of G exactly equals the variance error term $\mathbb{E}\left(\|L_\lambda g(q)\|^2\right) = E_V/\epsilon^2$, the expected low frequency component of $g(q)$.

The bias error term $E_B = \|f - L_{\lambda,W} f\|^2$ also depends on the graph structure of G, and it represents the high frequency content ($\omega_i \gg 1/\lambda$) of f with respect to this structure. Thus the full error $E_B + E_V$ is monotone decreasing in the smoothness $\|f\|^2 - \|f - L_{\lambda,W} f\|^2$, if W varies without changing the graph low frequency content, i.e., keeping E_V constant.

Part (b) follows directly from the definition of the regularization operator L_λ in terms of the kernel K in equation (12). Namely,

$$\begin{aligned} \|L_\lambda U\|^2 = \|\sum_i L_\lambda u_i(q)\|^2 &= \|\sum_i (\lambda + \mu_i)^{-1} \mu_i\, u_i(q)\|^2 \\ &= \sum_i \frac{\mu_i^2}{(\lambda + \mu_i)^2} \\ &= \text{tr}(1 + \lambda/\mu(\Delta))^{-2} \end{aligned} \tag{20}$$

since μ_i are eigenvalues of $\mu(\Delta)$. This completes a proof sketch.

The distinction between low frequency content of the graph G and of the function f (on G) can be clarified through equations (16) and (17). First consider the low frequency content of f, which by (16) is

$$\|f\|^2 - \|f - L_{\lambda,W} f\|^2 = \sum_i f_i^2 \left(1 - r(\lambda\omega_i)\right). \tag{21}$$

Since

$$1 - r(\lambda\omega_i) = \begin{cases} 1 & \text{if } \omega_i \ll 1/\lambda \\ 0 & \text{if } \omega_i \gg 1/\lambda \end{cases},$$

we see that only low frequency components of f (i.e. f_i for ω_i small) are contained in (21). On the other hand, the low frequency content of the graph G refers to the the number of eigenvalues ω_i (frequencies) that are small. Specifically (Theorem 3 part (b)), this equals

$$\|L_\lambda U\|^2 = \text{tr}(1 + \lambda/\mu(\Delta))^{-2} = \sum_i (1 + \lambda/\mu_i)^{-2} = \sum_i \frac{1}{(\lambda\omega_i + 1)^2} = \sum_i s(\lambda\omega_i),$$

with s as in (19). Thus only small eigenvalues $\omega_i \ll 1/\lambda$ are represented in the sum.

Theorem 3 thus states that among connection weight matrices W yielding a fixed trace above in (13) the error is a monotone decreasing function of

$$||(1 - L_{\lambda,W} f||^2 = \lambda^2 \sum_i \frac{f_i^2}{(\lambda + \mu_i)^2}.$$

(recall $\|f\|_2 = 1$). Since as shown above the regularizer L_λ is essentially a low pass filter, the quality of recovering f from $L_\lambda f_1$ requires the majority of the spectral content ω_i of f to lie below the spectral cut-off $1/\lambda$, while the majority of the noise $g(q)$ lies above this. This determines the optimal location of $1/\lambda$, i.e., so that bias error E_B is not too large, while variance error E_V is also controlled. The optimal λ minimizes

$$E_B + E_V = ||(1 - L_\lambda)f||^2 + \epsilon \mathbb{E}\left(||L_\lambda g||^2\right).$$

Bias as measured by loss of high frequency content

$$||(1 - L_\lambda)f||^2 = \lambda ||(K + \lambda)^{-1} f||^2$$

in f should be minimized conditioned on the smallest surviving low frequency content $\|\epsilon L_\lambda g\|^2$ of the noise (variance).

Finally we address a question parallel to the one in Theorem 2, regarding when a positive (nontrivial) regularization $\lambda > 0$ improves the estimate $L_\lambda f_1$ of f.

Theorem 4 (Optimal regularization for SVR feature vector regression) *For a fixed graph structure G and variable λ, the error $||(1 - L_\lambda)f||^2$*

(a) attains a minimum at a positive value $\lambda = \lambda_0$ of the regularization parameter λ;

(b) decreases for $\lambda > 0$ sufficiently small and increases for λ sufficiently large.

Proofs of above theorems will appear elsewhere [6].

3 Application: Using Prior Information to Form Graphs

3.1 Gene Expression

Our application involves denoising gene expression feature vectors $\mathbf{x} = (x_1, \ldots, x_p)$ subsequently used in (predictive) classification of tumors as metastatic or non-metastatic. See Table 1 for a sample of Wang et al.'s [30] unnormalized gene expression data. The quality of denoising is demonstrated by improvement of subsequent prediction of the metastatic/non-metastatic classes y in a test set. We exploit the underlying structure of the $p > 5,000$ feature indices i (genes) in feature vectors $\mathbf{x}$. These genes (and thus indices) form a graph structure with weights w_{ij} determined by whether the protein products of genes i and j interact chemically.

Genetics background. A strand of human DNA has 3 billion nucleotide bases consisting of the nucleotides A, C, G, and T. Genes are made of DNA and code for gene products, namely ribonucleic acid (*RNA*) and, downstream, proteins, which have specific biological functions. The *central dogma of biology* [17] summarizes the flow of information from DNA to RNA to proteins and their functions.

Humans have more than 20,000 protein-coding genes. Only about 1.5% of the genome (full DNA letter sequence) codes for proteins, while the rest consists of non-coding RNA genes, introns, and other DNA that we will not consider. *Gene expression* is the process by which gene information is used to make proteins. *Expression levels* for individual genes (measuring their RNA production) give their activity level in their translation to proteins. RNASeq gene expression technology is used to measure these levels; thousands of gene expression levels are taken at once. These so-called *high-throughput methods* allow collection of large amounts of data at relatively low cost. *Protein-protein interaction (PPI)* experiments measure when pairs of proteins tend to bind together to carry out their biological functions. This (prior) information is listed in a PPI database. This is used to produce a graph (network) on the set of gene indices so that indices (genes) i, j are connected if their protein products interact.

Cancer is a genetic disease caused primarily by DNA mutations. The purpose of the present application is to study effectiveness of predicting cancer outcomes (metastasis/no metastasis) based on gene expression feature vectors.

We have implemented the above models of feature vector regularization (denoising) to gene expression feature vectors $\mathbf{x} = (x_1, \ldots, x_p)$ taken from tumor samples in two studies. The denoising quality was tested by importing the denoised data into a training/testing algorithm for metastasis prediction and determining the quality of the prediction. The graph structure on the genome (set of human genes) based on PPI (see above) forms the index set graph for our feature vectors. Clustering on the network was accomplished by the *GraClus* [5] software, a computational graph clustering tool. This produced our increasing sequence of partitions of the genome. There are limitations to grouping genes via PPI interactions in this manner, as functionally related protein pairs sometimes correspond to gene pairs that do not have strictly correlated expressions.

Table 1 Gene expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer (sample from Wang dataset [30]). Unnormalized metastatic breast cancer dataset consisting of gene expression entries. The rows represent gene expressions and columns represent tumor samples (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse2034).

ID_REF	*GSM*36777	*GSM*36778	*GSM*36779	*GSM*36780	*GSM*36781	*GSM*36782	*GSM*36783	*GSM*36784
1007_s_at	3848.1	6520.9	5285.7	4043.7	4263.6	2949.8	5498.9	3863.1
1053_at	228.9	112.5	178.4	398.7	417.7	221.2	280.4	198.2
117_at	213.1	189.8	269.7	312.4	327.1	225	243.5	244.4
121_at	1009.4	2083.3	1203.4	1104.4	1043.3	1117.6	1085.4	1423.1
1255_g_at	31.8	145.8	42.5	108.2	69.2	47.4	84.3	102
1294_at	551.5	802.8	557.5	568.5	653.2	585	553.2	711.1
1316_at	176.7	278.4	183.3	187.7	185.8	166.6	92.5	259.3
1320_at	11.9	28.3	56.4	42.1	21.8	21.4	77.3	52.9
1405_i_at	309.3	449	101.9	899.1	3629.3	117.9	124.3	649.4
1431_at	49.9	122.9	85.9	90.7	96	148.1	52.6	126.3
1438_at	86.6	61.8	64.5	516.6	293.2	101.1	202.7	249
1487_at	452	660.7	687.6	667.8	994.6	474.4	829.5	747.9
1494_f_at	6589.1	604.7	12840.6	353.9	471.8	372.1	370	592.2
1598_g_at	3028.2	3625.1	2155.7	2399.1	1707.5	3431.3	1607.4	2731.6
160020_at	1135	1401.4	269.8	1165.1	893.2	940.4	521	1255.9
1729_at	759	1146.1	673.8	1066.7	605.5	543.4	363.6	511.8
1773_at	69.7	169.8	122.2	81.3	47.5	61.4	103.4	84.8
177_at	58.3	205	137.7	200.5	102.7	144.5	37.9	246.3
179_at	694.5	1007	724	846.7	969.5	641.4	715.8	730.4
1861_at	370.9	106.5	514.7	343.5	288.6	264.6	324.5	327.8

Total number of rows: 22283
Table truncated, full table size 36639 Kbytes.

One of the goals of the benchmark works of Wang et al. [30] and van de Vijver et al [27] on breast cancer was to predict metastatic breast cancer recurrence within a five year period, based on yes/no predictive classifications using gene expression. Of the 286 breast cancer cases in the Wang dataset, 93 metastasized, while, of the 295 patients in van de Vijver dataset, 79 metastasized. The PPI network for the genes we used was compiled from two databases, Reactome [10] and iRefIndex [19] (also see [20]). Our experiment used 5,747 genes in the Wang dataset (with 70,353 documented PPI interactions), and 5,310 genes (with 67,342 interactions) for the van de Vijver dataset.

We applied the local averaging and support vector regression denoising (pre-processing) methods to these cancer datasets. The numbers of gene clusters m in the local averaging study were 64, 128, 256, 512, 1024, and 2048 and *max*, the latter denoting the total number of genes, i.e., such that each gene forms its own cluster. The denoised datasets were subsequently used in training and testing machine learning (support vector machine, SVM) predictions. The predictions were of metastasis/no metastasis for both the Wang and the van de Vijver studies. We compared our results on the denoised training and test set against predictions obtained using the original data sets. The metric of prediction quality used was the area under the receiver operating characteristic (AUROC) curve. *AUROC* is a number between 0 and 1 measuring the accuracy of a binary classifier on a dataset. This is done in terms of numbers of correct/incorrect positive and negative predictions, as a function of a decision threshold and averaged over it. Our results showed that prediction of metastasis was improved when compared with the same methods using individual gene features (i.e., for which the number of clusters is greater than 5000).

In the training and testing with both original and denoised data, we used 5-fold cross-validation. Thus 1/5 of samples were randomly chosen and reserved as test data, while 4/5 formed training data. The classifier was a support vector machine (SVM), trained on the expression values as predictors of the known outcomes of the training set. The trained classifier was then tested for correctness of metastasis prediction on the remaining (test) samples. Standard deviations were calculated by repeating this cross-validation 200 times with different random 4/5-1/5 splits to produce separate training and test data.

Local averaging: The number of clusters m plays the role of a bandwidth parameter in smoothing. As m increased, the performance on both datasets first improved and then deteriorated. The optimal number of clusters was either 1024 or 2048. This means that the denoised (cluster-averaged) feature vectors represented data more accurately than raw expression features ($m = max$). In general the area under the ROC curve for classifiers predicting metastasis was improved by our clustering-based smoothing method from 53.4% to 73.0% and from 66.0% to 71.2% for the Wang and van de Vijver datasets, respectively. Details are listed in Table 2.

Support Vector Regression (SVR): For SVR we used the normalized diffusion (Gaussian) kernel K with three parameters. The first is the diffusion parameter β (so that the graph kernel $K = e^{-\beta\Delta}$). The regularization parameter C is defined

Table 2 Cluster-based averaging on Wang and van de Vijver breast cancer datasets. First column m is number of clusters. k_0 is the average size of each cluster, obtained by dividing total number of genes in each dataset by number of clusters. Numbers are mean values and those in parentheses are standard deviations based on 200 tests using 5-fold cross validation.

m	Wang		van de Vijver	
	k_0	AUROC	k_0	AUROC
64	89.8	0.658 (0.014)	83.0	0.687 (0.014)
128	44.9	0.680 (0.015)	41.5	0.705 (0.013)
256	22.4	0.692 (0.019)	20.7	0.689 (0.016)
512	11.2	0.684 (0.019)	10.4	0.686 (0.021)
1024	5.6	0.708 (0.019)	5.2	0.712 (0.019)
2048	2.8	0.730 (0.017)	2.6	0.500 (0.038)
MAX	1.0	0.534 (0.044)	1.0	0.660 (0.027)

as $\frac{1}{2\lambda n}$, with n the sample size and λ the Tikhonov regularization parameter. Our implementation of SVR (used on gene data) replaced the squared error loss $V(f,h) = \sum_{q\in G}(f_1(q) - h(q))^2$ in the Tikhonov regularization functional (11). The error measure used instead was the p-hinge loss function, given as

$$V_{\text{hin}}(f,h) = \sum_{q\in G}(|f(q) - h(q)| - p)_+$$

where $a_+ = \begin{cases} a & \text{if } a \geq 0 \\ 0 & \text{otherwise.} \end{cases}$ Here the parameter p controls the sensitivity of the loss. This loss function guarantees that prediction errors $f(q) - h(q)$ between $\pm p$ are fully tolerated.

Here both C and β are smoothness parameters: as C decreases (so λ increases) and as β increases, there is an added smoothness constraint on the optimizer $\hat{f}$ of the Tikhonov functional (11). In this simulation we have kept C (and thus λ) constant, and varied the bandwidth parameter β. As a local minimum in error was achieved when $C = 1$ and $p = .05$, we used these values in Table 3, where performance versus the smoothness constraint of the diffusion kernel bandwidth β is listed. The best performing classifier can obtain 74.2% and 74.1% AUROC for the Wang and van de Vijver datasets, respectively. Table 3 shows the performance of the SVR method.

We note that the improved classification performance here is based on an *unsupervised* method for processing (denoising) feature vectors $\mathbf{x}$. This is distinct, for example, from standard (supervised) feature selection methods, which depend on knowing the classes y_i of all feature vectors $\mathbf{x}_i$ in machine training. Unsupervised regularization of feature vectors is possible either before or after supervised feature selection, and can also be used without it. The method does not depend on the classes and is independent of the machine M later trained and used on the data. Since the method is useful independently of any dimensional reductions, these

Table 3 Performances of SVR smoothing on Wang and van de Vijver breast cancer datasets. First column β is the diffusion kernel bandwidth parameter. We use the parameters $C = 1$ and $p = 0.05$. Numbers are mean values, and numbers in parentheses are standard deviations using 200 tests with 5-fold cross validation. When $\beta = 0$, the diffusion kernel is the identity operator. As the kernel standard deviation β increases, the number of neighboring features in the averaging increases. Similarly to cluster-based averaging, denoised feature vectors then contain less noise but have a more biased signal.

β	**Wang**	**van de Vijver**
	AUROC	AUROC
0.01	0.735 (0.013)	0.738 (0.011)
0.05	0.736 (0.013))	0.738 (0.010)
0.1	0.738 (0.014)	0.737 (0.010)
0.5	0.742 (0.014)	0.737 (0.010)
1.0	0.741 (0.014)	0.740 (0.010)
2.0	0.735 (0.014)	0.738 (0.010)
0	0.534 (0.044)	0.660 (0.027)

improvements will supplement those of standard dimensional reduction methods. That is, our unsupervised regularization and subsequent feature selection are independent and deal with different parts of of classifier construction.

4 Conclusion

We have studied unsupervised smoothing/regularization methods for feature vectors in machine learning. These parallel the same methods in standard machine learning, Tikhonov regularization, and in function denoising methods such as local averaging using wavelets. This is done viewing feature vectors as functions on their indices, and adapting methods from real function regularization. We have illustrated this approach with two methods, using adaptations of local averaging and support vector regression, to regularize feature vectors. We apply these methods to cancer data regularization and their then to subsequent predictions of cancer metastasis/non-metastasis on such data. The improvement from regularization is accomplished entirely without knowing the cancer classes (metastatic/non-metastatic) of the training or test data.

Acknowledgements The authors thank Yue Fan for his expert computational contributions.The first author's research was partially supported by the US Air Force. The second author's research was partially supported by the National Science Foundation, HU AdvanceIT Grant No. 8538.

References

1. M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**(6), 1373–1396 (2003)
2. H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis. Mol. Syst. Biol. **3**(1), 1–10 (2007)
3. R.R. Coifman, D.L. Donoho, Translation-invariant de-noising, in *Wavelets and Statistics*. Lecture Notes in Statistics (Springer, Berlin, 1995)
4. F. Cucker, S. Smale, Best choices for regularization parameters in learning theory: on the bias-variance problem. Found. Comput. Math. **2**(4), 413–428 (2002)
5. I.S. Dhillon, Y. Guan, B. Kulis, Weighted graph cuts without eigenvectors: a multilevel approach. IEEE Trans. Pattern Anal. Mach. Intell. **29**(11), 1944–1957 (2007)
6. Y. Fan, M. Kon, L. Raphael, https://arxiv.org/abs/1212.4569 Feature vector regularization in machine learning. (2013)
7. S. Geman, E. Bienenstock, R. Doursat, Neural networks and the bias/variance dilemma. Neural Comput. **4**(1), 1–58 (1992)
8. R.L. Graham, D.E. Knuth, O. Patashnik, *Answer to Problem 9.60 in Concrete Mathematics: A Foundation for Computer Science* (Addison-Wesley, Boston, 1994)
9. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, vol. 2 (Springer, Berlin, 2009)
10. G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G.R. Gopinath, G.R. Wu, L. Matthews, Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. **33**(Suppl. 1), D428–D432 (2005)
11. N.M. Krukovskii, On the Tikhonov-stable summation of Fourier series with perturbed coefficients by some regular methods. Moscow Univ. Math. Bull. **28**(3), 7 (1973)
12. E. Lee, H.Y. Chuang, J.W. Kim, T. Ideker, D. Lee, Inferring pathway activity toward precise disease classification. PLoS Comput. Biol. **4**(11), e1000217 (2008)
13. M. Liu, A. Liberzon, S.W. Kong, W.R. Lai, P.J. Park, I.S. Kohane, S. Kasif, Network-based analysis of affected biological processes in type 2 diabetes models. PLoS Genet. **3**(6), e96 (2007)
14. S.G. Mallat, Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbf{R})$. Trans. AMS **315**(1), 69–87 (1989)
15. K. Mitra, A.-R. Carvunis, S.K. Ramesh, T. Ideker, Integrative approaches for finding modular structure in biological networks. Nat. Rev. Genet. **14**(10), 719–732 (2013)
16. M.Z. Nashed, G. Wahba, Regularization and approximation of linear operator equations in reproducing kernel spaces. Bull. AMS **80**(6), 1213–1218 (1974)
17. National Human Genome Research Institute NCBI. Central Dogma of Molecular Biology. https://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/central_dogma.html, (2017)
18. F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, J.-P. Vert, Classification of microarray data using gene networks. BMC Bioinf. **8**(1), 35 (2007)
19. S. Razick, G. Magklaras, I.M. Donaldson, iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinf. **9**(1), 1 (2008)
20. J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G.F. Berriz, F.D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D.S. Goldberg, L.V. Zhang, S.L. Wong, G. Franklin, S. Li, J.S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R.S. Sikorski, J. Vandenhaute, H.Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M.E. Cusick, D.E. Hill, F.P. Roth, M. Vidal, Towards a proteome-scale map of the human protein-protein interaction network. Nature **437**(7062), 1173–1178 (2005)
21. J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis* (Cambridge University Press, New York, NY, 2004)

22. A. Smola, R. Kondor, Kernels and regularization on graphs, in *Learning Theory and Kernel Machines* (Springer, New York, 2003), pp. 144–158
23. A.J. Smola, B. Scholkopf, A tutorial on support vector regression. Stat. Comput. **14**(3), 199–222 (2004)
24. A.D. Szlam, M. Maggioni, R.R. Coifman, Regularization on graphs with function-adapted diffusion processes. J. Mach. Learn. Res. **9**, 1711–1739 (2008)
25. R. Tibshirani, T. Hastie, B. Narasimhan, G.G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. **99**(10), 6567–6572 (2002)
26. A.N. Tikhonov, Stable methods for the summation of Fourier series. Soviet Math. Dokl. **5**, 4 (1964)
27. M.J. Van De Vijver, Y.D. He, L.J. van't Veer, H. Dai, A.A.M. Hart, D.W. Voskuil, G.J. Schreiber, J.L. Peterse, C. Roberts, M.J. Marton, A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. **347**(25), 1999–2009 (2002)
28. V. Vapnik, *Statistical Learning Theory*, vol. 1 (Wiley, New York, 1998)
29. J.-P. Vert, The optimal assignment kernel is not positive definite (2008). arXiv preprint. arXiv:0801.4061
30. Y. Wang, J.G.M. Klijn, Y. Zhang, A.M. Sieuwerts, M.P. Look, F. Yang, D. Talantov, M. Timmermans, M.E. Meijer-van Gelder, J. Yu, Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet **365**(9460), 671–679 (2005)
31. G.-B. Ye, D.-X. Zhou, Fully online classification by regularization. Appl. Comput. Harmon. Anal. **23**(2), 198–214 (2007)

Part XIX
Quantization

Analog-to-digital (A/D) is the model in which signals are represented by bit streams to provide effective storage, transmission, and processing. In the simplest terms it is a two-step process: linear sampling along the lines of the classical sampling theorem (Shannon's name is invoked, but it goes back to Cauchy in the 1840s) and quantization, a type of non-linear sampling, where the sampled data is assigned a value from a fixed finite *hard-wired* alphabet. There is also post-processing in the form of encoding and compression. In fact, the goal is reconstruction of the given signal from the quantized and compressed available data. An important quantization scheme, published in 1963 in the IEEE literature by Inose and Yasuda and now called $\Sigma\Delta$ quantization, invokes feedback mechanisms in order to mitigate various noises in many physical environments. One of the major influences, that escalated the mathematical influence in modern quantization theory, was the paper by Daubechies and DeVore in the Annals of Math. (2003).

The setting for Inose-Yasuda and Daubechies-DeVore was the space of band-limited functions. Independently, and for more traditional A/D, Goyal, Kovacević, and Vetterli introduced finite frame theory into the subject of quantization, because of the intrinsic noise reduction capability of frames and the inherent user-friendly computational advantage of finite tight frames. Then, it became natural to do $\Sigma\Delta$ quantization in the setting of finite frames. Further, compressive sensing or sampling became a staple in the subject for the several reasons, not the least of which was the aforementioned non-linear sampling step.

There has been an explosion of activity in this area in the past 10 years, and some of the leaders and deepest contributors are the authors of these three chapters in this section.

The chapter by Boufounos, Rane, and Mansour takes quantization as a starting point and generalizes the idea in the direction of embeddings of one signal space to another. The goal is to go beyond the raison d'être of approximating the actual signal in the encoding stage by preserving underlying information in the signal as it may be related to other signals affecting it. The technology is wonderful in that geometry and compressive sensing interleave necessarily and naturally. The exposition is a lucid presentation of what has been done and the starting point of their fascinating new theory.

The chapter by Chou and Güntürk is a sequel to the deep analysis and concept of distributed noise-shaping of their previous work in Constructive Approximation (2016). In this chapter they begin with a beautiful exposition of their published theory in 2016, which itself has Güntürk's profound work on beta encoding as background. Then, they provide the all important performance evaluation for the class of finite group frames, that includes finite Fourier frames and harmonic frames; and they conclude with analysis of the infinite dimensional case of band-limited functions. The interplay of topics is compelling and creative.

The chapter by Lee, Powell, and Whitehouse is a tour-de-force encompassing brilliant analytic technology and addressing fundamental problems of consistent reconstruction in quantization. The authors prove essential error bounds on error moments arising in consistent reconstruction, going beyond the mean-square theory they had already resolved. The proofs are not for the faint of heart, but the centrality of the problem they have solved provides hope for actual implementation at a very high level, as some of their implementations imply.

Embedding-Based Representation of Signal Geometry

Petros T. Boufounos, Shantanu Rane, and Hassan Mansour

Abstract Low-dimensional embeddings have emerged as a key component in modern signal processing theory and practice. In particular, embeddings transform signals in a way that preserves their geometric relationship but makes processing more convenient. The literature has, for the most part, focused on lowering the dimensionality of the signal space while preserving distances between signals. However, there has also been work exploring the effects of quantization, as well as on transforming geometric quantities, such as distances and inner products, to metrics easier to compute on modern computers, such as the Hamming distance.

Embeddings are particularly suited for modern signal processing applications, in which the fidelity of information represented by the signals is of interest, instead of the fidelity of the signal itself. Most typically, this information is encoded in the relationship of the signal to other signals and templates, as encapsulated in the geometry of the signal space. Thus, embeddings are very good tools to capture the geometry, while reducing the processing burden.

In this chapter, we provide a concise overview of the area, including foundational results and recent developments. Our goal is to expose the field to a wider community, to provide, as much as possible, a unifying view of the literature, and to demonstrate the usefulness and applicability of the results.

Keywords Dimensionality reduction • Distance-preserving embeddings • Nearest neighbors

1 Introduction

Signal representation theory and practice has primarily focused on how to best represent or approximate signals, while incurring the smallest possible distortion. Advances such as frames, compressive sensing, and sparse approximations have all

P.T. Boufounos (✉) • H. Mansour
Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA 02139, USA
e-mail: petrosb@merl.com; mansour@merl.com

S. Rane
Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA
e-mail: srane@parc.com

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4_7

been applied in improving the representation accuracy or sampling complexity using a fidelity metric as the principal figure of merit. On the other hand, as computation becomes more prevalent, signal representations are increasingly important in inference and estimation applications. Such applications typically exploit the geometry of the signal space, usually captured mathematically by norms and inner products. In these cases, the representation should faithfully preserve the geometry of the signal space, but not necessarily the signals themselves.

This chapter explores embeddings as a signal representation mechanism that preserves the geometry of the signal space. Embeddings are transformations from one signal space to another—the embedding space—which exactly or approximately preserve signal geometry. The use of an embedding is beneficial if the transformation provides some convenience in its use. For example, the embedding space might have significantly lower dimensionality than the signal space, might allow for easier computation of certain quantities, or might enable efficient transmission by quantizing in the embedding space.

In this chapter, we explore several aspects of embedding design. We start with the foundational work by Johnson and Lindenstrauss [40], and continue with more recent developments. We describe embeddings that preserve distances, inner products, and angles between signals, while reducing the dimension and the bit-rate. We also describe embedding design strategies, both data-agnostic and universal, as well as learning-based and data-driven. Our discussion also explores the effect of quantization, which becomes necessary when the embeddings are used to reduce the bit-rate of the representation.

Our goal is to expose the field to a wide community and show that embeddings are essential data processing tools. In our exposition, we attempt to provide, as much as possible, a unifying view of the literature. However, we remark that recent advances have reinvigorated research in this area, often making such unification elusive.

1.1 Notation

In the remainder of the chapter, we use regular typeface, e.g., x and y, to denote scalar quantities. Lowercase boldface such as $\mathbf{x}$ denotes vectors and uppercase boldface such as $\mathbf{A}$ denotes matrices. The m^{th} element of vector $\mathbf{x}$ is denoted using x_m. Functions are denoted using regular lowercase typefaces, e.g., $g(\cdot)$. Unless explicitly noted, all functions are scalar functions of one variable. In abuse of notation, a vector input to such functions, e.g., $g(\mathbf{x})$ means that the function is applied element-wise to all the elements of $\mathbf{x}$. Sets and vector spaces are denoted using calligraphic fonts, e.g., $\mathscr{W}, \mathscr{S}$.

1.2 Outline

The next section describes distance-preserving embeddings. Starting with general definitions and foundational results, the section explores embedding design

strategies—both data-agnostic and data-driven—and discusses the nature of distance-preserving guarantees. Section 3 examines embeddings that preserve angles and inner products, including kernel inner products. Quantization strategies and the effects of quantization on the embedding guarantees are discussed in Sec. 4. Section 5 provides a higher-level discussion and concludes the chapter.

2 Preserving Distances

The best-known embeddings preserve the geometry of the space by preserving the distance between signals. In this section, we examine distance-preserving embeddings, and explore some ways to design their distance-preserving properties.

2.1 Randomized Linear Embeddings

An embedding is a transformation of a set of signals in a high-dimensional space to a (typically) lower-dimensional one such that some aspects of the geometry of the set are preserved, as depicted in Figure 1. Since the set geometry is preserved, distance computations can be performed directly on the low-dimensional—and often low bit-rate—embeddings, rather than the underlying signals. For the purposes of this chapter, we define an embedding as follows.

Definition 1 A function $f : \mathscr{S} \to \mathscr{W}$ is a (g, δ, ϵ) embedding of $\mathscr{S}$ into $\mathscr{W}$ if, for all $\mathbf{x}, \mathbf{x}' \in \mathscr{S}$, it satisfies

$$(1-\delta)g\left(d_{\mathscr{S}}(\mathbf{x},\mathbf{x}')\right) - \epsilon \leq d_{\mathscr{W}}\left(f(\mathbf{x}),f(\mathbf{x}')\right) \leq (1+\delta)g\left(d_{\mathscr{S}}(\mathbf{x},\mathbf{x}')\right) + \epsilon. \quad (1)$$

In this definition, $g : \mathbb{R} \to \mathbb{R}$ is an invertible function mapping distances in $\mathscr{S}$ to distances in $\mathscr{W}$ and δ and ϵ quantify, respectively, the multiplicative and the additive ambiguity of the mapping. We will often refer to $g(\cdot)$ as the distance map and to $f(\cdot)$ as the embedding map. In most known embeddings, such as the ones

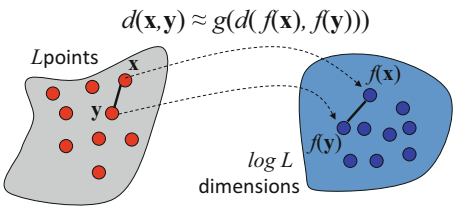


Fig. 1 Distance-preserving embeddings approximately preserve a function $g(\cdot)$ of the distance, allowing distances to be computed in a space that (typically) has fewer dimensions or has other desirable properties.

discussed in this section, the distance map is the identity $g(d) = d$ or a simple scaling. The similarity metrics $d_{\mathscr{S}}(\cdot,\cdot)$ and $d_{\mathscr{W}}(\cdot,\cdot)$ are typically distances, but could also be correlations, divergences, or other functions capturing signal geometry and similarity[1].

The best known embeddings are the Johnson-Lindenstrauss (JL) embeddings [40]. These are functions $f : \mathscr{S} \to \mathbb{R}^M$ from a finite set of signals $\mathscr{S} \subset \mathbb{R}^N$ to an M-dimensional vector space such that, given two signals $\mathbf{x}$ and $\mathbf{x}'$ in $\mathscr{S}$, their images satisfy:

$$(1-\delta)\|\mathbf{x}-\mathbf{x}'\|_2^2 \le \|f(\mathbf{x})-f(\mathbf{x}')\|_2^2 \le (1+\delta)\|\mathbf{x}-\mathbf{x}'\|_2^2. \tag{2}$$

In other words, these embeddings preserve Euclidean, i.e., ℓ_2, distances of point clouds within a small factor, measured by δ, and using the identity as a distance map.

In the context of Def. 1, a JL embedding is a $(g_I, \delta, 0)$ embedding of squared Euclidean distances—$d_{\mathscr{S}}(\mathbf{x},\mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|_2^2$ and $d_{\mathscr{W}}(f(\mathbf{x}),f(\mathbf{x}')) = \|f(\mathbf{x}) - f(\mathbf{x}')\|_2^2$—with an identity distance map $g_I(d) = d$. In this context, the JL theorem can be stated as:

Theorem 1 *Given $\delta \in (0,1)$ and a set $\mathscr{S} \subset \mathbb{R}^N$ of $\#(\mathscr{S}) = L$ points and $M = O(\delta^{-2}\ln L)$, there exists a Lipschitz map $f : \mathbb{R}^N \to \mathbb{R}^M$ that is a $(g_I, \delta, 0)$ embedding of $\mathscr{S}$, with $g_I(d) = d$, $d_{\mathscr{S}}(\mathbf{x},\mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|_2^2$ and $d_{\mathscr{W}}(f(\mathbf{x}),f(\mathbf{x}')) = \|f(\mathbf{x})-f(\mathbf{x}')\|_2^2$.*

Johnson and Lindenstrauss demonstrated that a distance-preserving embedding, as described above, exists in a space of dimension $M = O(\delta^{-2}\log L)$, where L is the number of signals in $\mathscr{S}$ (its cardinality) and δ the desired tolerance in the embedding. Remarkably, M is independent of N, the dimensionality of the signal set $\mathscr{S}$. Subsequent work showed that it is straightforward to compute such embeddings using a linear mapping. In particular, the function $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is an $M \times N$ matrix whose entries are drawn randomly from specific distributions, satisfies (2) for all $\mathbf{x},\mathbf{x}' \in \mathscr{S}$ with probability $1 - c_1 e^{\log L - c_2\delta^2 M}$, for some universal constants c_1, c_2, where the probability is with respect to the measure of $\mathbf{A}$. Commonly used distributions for the entries of $\mathbf{A}$ are i.i.d. Gaussian, i.i.d. Rademacher, or i.i.d. uniform [1, 25]. More recent work has shown that the embedding dimensionality $M = O(\delta^{-2}\log L)$ is also necessary, making these constructions tight [38].

Most proofs involve constructing a randomized map such that (1) holds with very high probability on a pair of points $\mathbf{x},\mathbf{x}' \in \mathscr{S}$. Using a concentration of measure argument, such as Hoeffding's inequality or a Chernoff bound, it can typically be shown that the guarantee fails with probability that decays exponentially with the number of measurements, i.e., with the dimensionality of the embedding space $M =$

[1]Technically, we could incorporate $g(\cdot)$ into $d_{\mathscr{S}}(\cdot,\cdot)$ and remove it from this definition. However, we choose to make it explicit here and consider it a distortion to be explicitly analyzed. In an abuse of nomenclature, we generally refer to $d(\cdot,\cdot)$ as distance, even if in some cases it is not strictly a distance metric but might be an inner product, or another geometric quantity of interest.

$\dim(\mathscr{W})$. In other words, the embedding fails on a pair of points with probability bounded by $\Omega(e^{-Mw(\delta,\epsilon)})$, where $w(\delta,\epsilon)$ is an increasing function of ϵ and δ that quantifies the concentration of measure exhibited by the randomized construction.

Once the embedding guarantee is established for a pair of signals, a union bound or chaining argument can be used to extend it to a finite set of signals. If the set $\mathscr{S}$ is finite, containing L points, then the probability that the embedding fails is upper bounded by $\Omega(L^2 e^{-Mw(\delta,\epsilon)}) = \Omega(e^{2\log L - Mw(\delta,\epsilon)})$, which decreases exponentially with M, as long as $M = O(\log L)$.

More recently, in the context of compressive sensing, such linear embeddings have been shown to embed infinite sets of signals. For example, the restricted isometry property (RIP) is an embedding of K-sparse signals and has been shown to be achievable with $M = O(K \log \frac{N}{K})$ [10, 23, 50]. A near equivalence of RIP with the JL lemma has also been established: an RIP matrix with its columns randomly multiplied with ± 1 will satisfy the JL lemma [41]. Similar properties have been shown for other signal set models, such as more general unions of subspaces and manifolds [9, 11, 12, 21, 28, 29, 50].

Typically, these generalizations are established by first proving that the embedding holds in a sufficiently dense point cloud on the signal set and exploiting linearity and smoothness to extend it to all the points of the set. The resulting guarantee uses the covering number of the set, i.e., its Kolmogorov complexity—instead of the number of points L—to measure the complexity of the set and determine the dimensionality required of the projection. A fairly general exposition of this approach, as well as generalizations for non-smooth embedding maps can be found in [20].

An alternative characterization of the complexity of $\mathscr{S}$ is its Gaussian width.

Definition 2 Given a set $\mathscr{S} \subseteq \mathbb{R}^N$, the quantity

$$W(\mathscr{S}) = \mathbb{E}\left\{ \sup_{\mathbf{x}\in\mathscr{S}} \mathbf{g}^T\mathbf{x} \right\}, \tag{3}$$

where the expectation is taken over $\mathbf{g} \sim \mathscr{N}(0, \mathbf{I})$ is called *the Gaussian width* of $\mathscr{S}$.

The Gaussian width of a set can sometimes be easier to characterize than its Kolmogorov complexity, although the latter can be bounded by the former [28].

Beyond the discussion above, in the remainder of this chapter, we defer on the rigorous development required to extend embedding guarantees to hold for infinite signal sets. Nevertheless, in many cases we will mention if such generalizations are possible or exist in the literature.

2.2 Embedding Map Design

One of the key elements in the embedding definition (1) is the embedding map $g(\cdot)$. The JL guarantee in (2) implies an embedding map $g(d) = d$, that does not distort the distance measure. However, it is often desirable to introduce such distortions

and understand their effect. For example, if the interest is in preserving only local distances, the distance map can be used to describe and characterize the distance preserving properties of the embedding [17, 19, 20].

A general approach to embedding design would use $g(\cdot)$ to derive an embedding function $f(\cdot)$, possibly randomized, that achieves (1) given sufficient dimensionality of the embedding space $\mathscr{W}$. Unfortunately, such a design is still an open problem. Furthermore, an arbitrary $g(\cdot)$ is not always possible. For example, any realizable $g(\cdot)$ satisfies a generalized subadditivity property [20].

Instead, [20] demonstrates a general probabilistic approach to designing the embedding function $f(\cdot)$ and deriving the embedding map. The mapping function takes the form $\mathbf{y} = f(\mathbf{x}) = h(\mathbf{A}\mathbf{x}+\mathbf{w})$, where the elements of $\mathbf{A}$ are randomly chosen from an i.i.d. distribution and the elements of the dither $\mathbf{w}$ are chosen from an i.i.d. distribution uniform in $[0, 1)$. The embedding is designed through $h(t)$, a bounded periodic scalar function with period 1, applied element-wise to its argument. The Fourier series coefficients of $h(\cdot)$ are denoted using H_k and $\bar{h} = \sup_t h(t) - \inf_t h(t)$.

Theorem 2 ([20], Thm. 4.1) *Consider a set $\mathscr{S}$ of Q points in $\mathbb{R}^N$, measured using $\mathbf{y} = h(\mathbf{A}\mathbf{x} + \mathbf{w})$, with $\mathbf{A}, \mathbf{w}$, and $h(t)$ as above. With probability greater than $1 - e^{2\log Q - 2M\frac{\epsilon^2}{\bar{h}^4}}$ the following holds*

$$g(d) - \epsilon \leq \frac{1}{M}\left\|\mathbf{y} - \mathbf{y}'\right\|_2^2 \leq g(d) + \epsilon \tag{4}$$

for all pairs $\mathbf{x}, \mathbf{x}' \in \mathscr{S}$ and their corresponding measurements $\mathbf{y}, \mathbf{y}'$, where

$$g(d) = 2\sum_k |H_k|^2(1 - \phi_l(2\pi k|d)) \tag{5}$$

defines the distance map of the embedding.

In the theorem above, $\phi_l(l|d)$ is a characteristic function depending on the density of $\mathbf{A}$. For example, if the elements of $\mathbf{A}$ are drawn from an i.i.d. Normal distribution, then the characteristic function is $\phi_l(\xi|d) = \phi_{\mathscr{N}(0,\sigma^2 d^2)}(\xi) = e^{-\frac{1}{2}(\sigma d\xi)^2}$ and the distance map becomes

$$g(d) = 2\sum_k |H_k|^2 \left(1 - e^{-2(\pi\sigma dk)^2}\right), \tag{6}$$

with d measuring the ℓ_2 distance.

If, instead, elements of $\mathbf{A}$ are drawn from an i.i.d. Cauchy distribution with zero location parameter and scale parameter γ, then the characteristic function is $\phi_l(\xi|d) = e^{-\gamma d|\xi|}$ and the corresponding distance map is

$$g(d) = 2\sum_k |H_k|^2(1 - e^{-2\pi\gamma dk}), \tag{7}$$

with d in this case measuring the ℓ_1 distance.

The guarantee in Thm. 2 is about embedding the ℓ_1 or ℓ_2 distance into ℓ_2^2. By taking the square root, the guarantee can be provided for embedding into ℓ_2 instead.

Corollary 1 ([20], Cor. 4.1) *Consider the signal set $\mathscr{S}$, defined and measured as in Thm. 2. With probability greater than $1 - e^{2\log Q - 2M\left(\frac{\epsilon}{h}\right)^4}$ the following holds*

$$\widetilde{g}(d) - \epsilon \leq \frac{1}{\sqrt{M}} \left\| \mathbf{y} - \mathbf{y}' \right\|_2 \leq \widetilde{g}(d) + \epsilon \tag{8}$$

for all pairs $\mathbf{x}, \mathbf{x}' \in \mathscr{S}$ and their corresponding measurements $\mathbf{y}, \mathbf{y}'$, where $\widetilde{g}(d) = \sqrt{g(d)}$.

2.3 Distance-preserving properties of the map

Typically, when designing a distance map, it is desirable to understand how accurate the embedding is in representing distances. In particular, embedding guarantees, as stated above and in the literature, bound how much the distance in the embedding space might deviate from the true distance between the signals.

However, in practice, embeddings are used as a proxy for the true distance of the signals. Given two signals, $\mathbf{x}$ and $\mathbf{x}'$, and their embedding distance, $d_{\mathscr{W}}(f(\mathbf{x}), f(\mathbf{x}'))$, a natural estimate of the true signal distance is [19, 20]

$$\widetilde{d}_{\mathscr{S}} = g^{-1}\left(d_{\mathscr{W}}\left(f(\mathbf{x}), f(\mathbf{x}')\right)\right), \tag{9}$$

assuming $g(\cdot)$ is differentiable. Thus, the approximation guarantee is often more useful when stated with respect to the estimate, $\widetilde{d}_{\mathscr{S}}$.

$$\left|\widetilde{d}_{\mathscr{S}} - d_{\mathscr{S}}(\mathbf{x}, \mathbf{x}')\right| \lesssim \frac{\epsilon + \delta d_{\mathscr{W}}\left(f(\mathbf{x}), f(\mathbf{x}')\right)}{g'\left(\widetilde{d}_{\mathscr{S}}\right)}. \tag{10}$$

An important component of this guarantee is its dependence on the gradient of the embedding map $g'(\cdot)$ around the distance of the signals. In regions where the embedding map is flatter, the ambiguity is higher. In hindsight, this is expected: estimates of a variable observed through a non-linear map and observation ambiguity are less accurate at regions of the map that are flatter.

Figure 2 demonstrates this effect using a $(g, 0, \epsilon)$ embedding as an example. The solid line in the left figure depicts the distance map $g(\cdot)$. The two dashed lines depict the upper and lower bounds of the guarantee, separated by ϵ above and below the distance map. In other words, the vertical ambiguity is constant across the range of $d_{\mathscr{S}}$. The figure also shows two example points on which the embedding distance is computed, $d_{\mathscr{W}}$ and $d'_{\mathscr{W}}$. The corresponding estimates of the true signal distance are $g^{-1}(d_{\mathscr{W}})$ and $g^{-1}(d'_{\mathscr{W}})$, respectively. However, the ambiguity of these estimates is

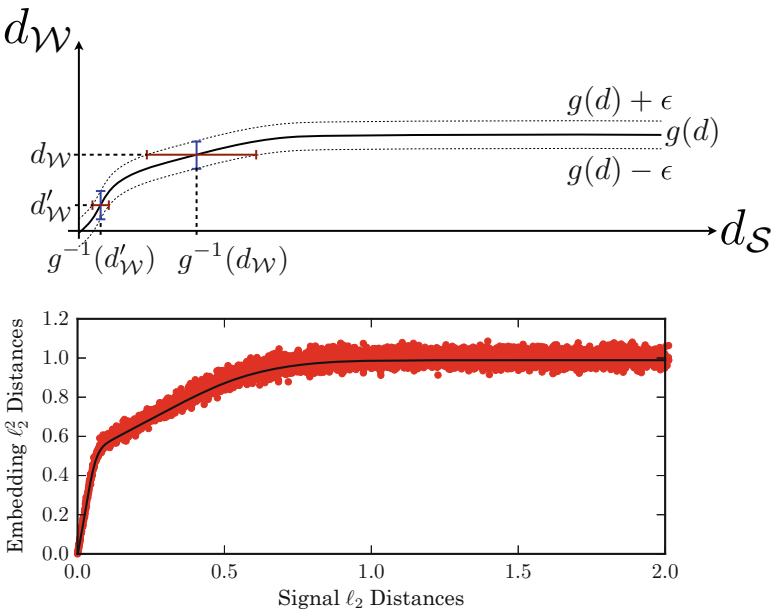


Fig. 2 Effect of the gradient of the distance map on the distance ambiguity of the embedding. (left) Even though the vertical ambiguity is constant across the distance map, the corresponding horizontal ambiguity varies significantly, depending on the slope of the map. (right) Example embedding exhibiting similar behavior as described by the map on the left.

significantly higher for $d_{\mathcal{W}}$ than for $d'_{\mathcal{W}}$, because of the difference in slope of $g(\cdot)$ at the corresponding points. Simulations using an actual embedding design exhibiting the same behavior are shown on the right-hand side.

Embedding maps designed using the approach in Sec. 2.2 eventually saturate and become flat beyond a certain signal distance. Thus, the ambiguity becomes infinite; the embedding does not preserve distances beyond a range. Given an embedding map $h(\cdot)$, this range can be controlled by the scaling parameters of the distribution of $\mathbf{A}$, such as σ and γ in (6) and (7), respectively. The same parameters also scale the gradient of the embedding, thus controlling the ambiguity, as described in (10). In other words, varying the scale parameters is equivalent to navigating a trade-off between smaller ambiguity while representing a smaller range of distances, and greater ambiguity while representing a larger range of distances. In fact, similar trade-offs are possible with any embedding function, simply by scaling the argument and replacing $f(\mathbf{x})$ with $f(a\mathbf{x})$ for any $a > 0$.

The distance preserving ambiguity described above characterizes distance preservation through $g(\cdot)$ along a full range of distances. However, it is often sufficient to only guarantee the locality of the embedding, i.e., that small distances remain small and larger distances do not become too small. Recent work has attempted to define locality in the context of binary embeddings [46, Def. 2.3], as well as,

implicitly, in the context of learning an embedding for classification [31, Eq. (6)]. In the same spirit, guarantees on using JL embeddings for classification have been recently established, assuming specific signal models. In particular, in [7] it is shown that separated convex ellipsoids remain separated when randomly projected to a space with sufficiently high dimensionality. However, an appropriate and useful characterization of locality is still a pending question.

One important property of the embeddings described so far is their universality. Their randomized construction does not take the data into account. The guarantees hold with very high probability on any set of points $\mathscr{S}$ to be embedded, as long as the set complexity is known. Thus, there is no adversarial selection of the data for which the embedding will fail, assuming the data set is generated independently of the embedding. The next section explores embeddings designed while taking sample data into account, their advantages, as well as their disadvantages.

2.4 Learning the Embedding Map

A key advantage of the embeddings described above is their universality and the simplicity in computing them. However, it is often advantageous to tune the embedding to an application using available training data. The main assumption is that the training data is representative of the data to be observed by the application; tuning the embedding to the data should provide an embedding that performs well on all future data on which the embedding will be used.

Inspired by the JL lemma, recent work [31, 54] demonstrates that given a set of L points $\mathscr{S} = \{\mathbf{x}_i \in \mathbb{R}^N,\ i = 1, \ldots, L\}$ as training data, it is possible to formulate a convex optimization problem and determine a linear embedding map, $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$, that preserves the squared Euclidean distance. The resulting map provides a $(g_I, \delta, 0)$ embedding. The problem can be formulated to either minimize the dimensionality of the embedding space under a fixed multiplicative distortion δ or minimize the distortion given a fixed embedding dimensionality.

In formulating the problem, the objects of interest are not the signals $\mathbf{x}_i$ but their differences $\mathbf{x}_i - \mathbf{x}_j$. Thanks to the linearity of the map, to guarantee a $1 \pm \delta$ multiplicative ambiguity it is sufficient to guarantee a δ distortion of the normalized difference $\frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|_2}$. Thus, the formulation starts with the secant set

$$\mathscr{X} = \left\{ \mathbf{v}_{ij} = \frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|_2},\ \mathbf{x}_i, \mathbf{x}_j \in \mathscr{S},\ i \neq j \right\} \tag{11}$$

The map $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$ satisfies the guarantee for all $\mathbf{v}_{ij} \in \mathscr{X}$ if

$$\left| \|\mathbf{A}\mathbf{v}_{ij}\|_2^2 - \|\mathbf{v}_{ij}\|_2^2 \right| \leq \delta, \tag{12}$$

where, by construction, $\|\mathbf{v}_{ij}\|_2^2 = 1$ for all i, j.

The squared norm can be expressed as a quadratic form $\|\mathbf{A}\mathbf{v}_{ij}\|_2^2 = \mathbf{v}_{ij}^T\mathbf{A}^T\mathbf{A}\mathbf{v}_{ij}$ which is linear in $\mathbf{P} = \mathbf{A}^T\mathbf{A}$. Furthermore, if $\mathbf{A} \in \mathbb{R}^{M\times N}$, then $\mathbf{P}$, which is positive semidefinite, has $\text{rank}(\mathbf{P}) = M$. Thus, the $\mathbf{P}$ corresponding to the embedding that satisfies (12) for all pairs $i \neq j$ with the minimum number of measurements can be found using the following optimization [31]:

$$\widehat{\mathbf{P}} = \arg \min_{\mathbf{P}^T=\mathbf{P}\succeq 0} \text{rank}(\mathbf{P}) \tag{13}$$

$$\text{subject to } |\mathbf{v}_{ij}^T\mathbf{P}\mathbf{v}_{ij} - 1| \leq \delta \text{ for all } i \neq j.$$

This is a non-convex and combinatorially complex program. To solve it, [31] proposes the relaxation of the rank using the nuclear norm, which results in the following polynomial-time semidefinite program:

$$\widehat{\mathbf{P}} = \arg \min_{\mathbf{P}^T=\mathbf{P}\succeq 0} \|\mathbf{P}\|_* \tag{14}$$

$$\text{subject to } |\mathbf{v}_{ij}^T\mathbf{P}\mathbf{v}_{ij} - 1| \leq \delta \text{ for all } i \neq j.$$

Alternatively, [54] modifies the formulation to determine the optimal δ using a fixed number of measurements M, also adding an energy constraint on the coefficients of the matrix $\mathbf{A}$. The resulting problem constrains both the rank and the trace norm of $\mathbf{P}$.

$$\widehat{\mathbf{P}} = \arg \min_{\mathbf{P}^T=\mathbf{P}\succeq 0} \max_{i\neq j} |\mathbf{v}_{ij}^T\mathbf{P}\mathbf{v}_{ij} - 1| \tag{15}$$

$$\text{subject to } \text{rank}(\mathbf{P}) \leq M \text{ and } \|\mathbf{P}\|_* \leq b, \tag{16}$$

where b is the energy constraint. Using a game-theoretic formulation, [54] also derives an algorithm to solve (16) with performance guarantees. It is also shown that the performance of the embedding can be guaranteed on new data, similar to the training set, using a continuity argument similar to the one in [10].

As mentioned in Sec. 2.3, a notion of semantic locality is also introduced in [31], in the context of classification. In particular, for elements i and j from the training data that belong in the same class, the embedding should guarantee that their distances do not increase significantly but does not need to limit how much they may shrink. On the other hand, if elements i and j belong to different classes, the embedding should guarantee that their distances do not shrink significantly but may allow them to grow unconstrained. Under those conditions, the embedding guarantees that each cluster stays together, even though two different clusters may separate from each other. Thus, classification is still possible in the embedded data. The resulting optimization is less constrained than (15).

$$\widehat{\mathbf{P}} = \arg \min_{\mathbf{P}^T = \mathbf{P} \succeq 0} \|\mathbf{P}\|_* \tag{17}$$

$$\text{subject to } \mathbf{v}_{ij}^T \mathbf{P} \mathbf{v}_{ij} \geq 1 - \delta \text{ for all } i \neq j \text{ in different classes.}$$

$$\mathbf{v}_{ij}^T \mathbf{P} \mathbf{v}_{ij} \leq 1 + \delta \text{ for all } i \neq j \text{ in the same class.}$$

In all the formulations above, $\mathbf{A}$ can be determined from $\widehat{\mathbf{P}}$ using a simple factorization. For example, the economy-sized singular value decomposition is $\widehat{\mathbf{P}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T$, where $\mathbf{U} \in \mathbb{R}^{N \times M}$ has orthonormal columns and $\boldsymbol{\Sigma} \in \mathbb{R}^{M \times M}$ is diagonal. The embedding can be computed using $\widehat{\mathbf{A}} = \boldsymbol{\Sigma}^{1/2}\mathbf{U}^T$.

3 Preserving Inner Products, Angles, and Correlations

The embeddings discussed in the previous section are designed to preserve distances between signals in the embedding space. However, in a number of problems, inner products and correlations should be preserved instead. In this section we consider how distance embeddings can be used to preserve regular inner products and kernel inner products, as well as how binary and phase embeddings can be used to preserve normalized correlations, i.e., angles, without preserving distances.

3.1 Inner Product Embeddings

When the signal and the embedding spaces are inner product spaces, then the inner product can be determined using the signal distances that are preserved. The inner product of the measurements $\langle \mathbf{y}, \mathbf{y}' \rangle$ can be derived from the ℓ_2^2 difference of the measurements, $\|\mathbf{y} - \mathbf{y}'\|_2^2$. Specifically,

$$\|\mathbf{y} - \mathbf{y}'\|_2^2 = \|\mathbf{y}\|_2^2 + \|\mathbf{y}'\|_2^2 - 2\langle \mathbf{y}, \mathbf{y}' \rangle \Longrightarrow \langle \mathbf{y}, \mathbf{y}' \rangle = \frac{\|\mathbf{y}\|_2^2 + \|\mathbf{y}'\|_2^2 - \|\mathbf{y} - \mathbf{y}'\|_2^2}{2}. \tag{18}$$

When all these norms are preserved by the embedding, it is straightforward to show that JL-type random projections satisfy [2]

$$\left| \langle \mathbf{y}, \mathbf{y}' \rangle - \langle \mathbf{x}, \mathbf{x}' \rangle \right| \leq \delta(\|\mathbf{x}\|_2^2 + \|\mathbf{x}'\|_2^2) \tag{19}$$

With a little more care, exploiting the linearity of the embedding, a tighter bound can be derived [27]

$$\left| \langle \mathbf{y}, \mathbf{y}' \rangle - \langle \mathbf{y}, \mathbf{y}' \rangle \right| \leq \delta \|\mathbf{x}\|_2 \|\mathbf{x}'\|_2 \tag{20}$$

In addition to standard inner products, appropriately designed embeddings can also be used to approximate kernel inner products. Kernel inner product embeddings were first introduced in [51] and significantly generalized in [17, 20]. Common kernels include the Gaussian $K(\mathbf{x},\mathbf{x}') = e^{\|\mathbf{x}-\mathbf{x}'\|_2^2/\sigma^2}$ and the Laplacian $K(\mathbf{x},\mathbf{x}') = e^{\gamma\|\mathbf{x}-\mathbf{x}'\|_1}$. Since computing those kernels relies on computing distances, the development in Sec. 1 could be used to directly estimate the distance and compute the kernel. However, the resulting ambiguity would manifest itself in the exponent, making it difficult to characterize and control.

Instead, guarantees based on computing the inner product in the embedding domain can be derived, exploiting the design approach in Sec. 2.2. Similarly to standard inner products, establishing the guarantees relies on (18). However, the difficulty lies in bounding $\|\mathbf{y}\|_2^2$ which is necessary, in addition to the distance between $\mathbf{y}$ and $\mathbf{y}'$. When using the embedding design in Thm. 2, it is straightforward to show that, in the embedding space,

$$\sum_k |H_k|^2 - \epsilon \le \frac{1}{M}\|\mathbf{y}\|_2^2 \le \sum_k |H_k|^2 + \epsilon, \tag{21}$$

with probability greater than $1 - 2e^{\log Q - 2M\frac{\epsilon^2}{h}}$. Thus, if $d_{\mathscr{W}}(\mathbf{y},\mathbf{y}') = \|\mathbf{y}-\mathbf{y}'\|_2^2$ in Def. 1, and substituting (4) and (21) in (18), we can show that the embedding can be designed to approximate a kernel.

Theorem 3 (Thm. 4.4 in [20]) *Consider a set $\mathscr{S}$ of Q points in $\mathbb{R}^N$, measured using $\mathbf{y} = h(\mathbf{A}\mathbf{x}+\mathbf{w})$, with $\mathbf{A},\mathbf{w}$, and $h(t)$ as in Thm. 2. With probability greater than $1 - e^{2\log Q - \frac{8}{9}M\frac{\epsilon^2}{h^4}}$ the following holds*

$$K(d) - \epsilon \le \frac{1}{M}\langle \mathbf{y},\mathbf{y}'\rangle \le K(d) + \epsilon \tag{22}$$

for all pairs $\mathbf{x},\mathbf{x}' \in \mathscr{S}$ and their corresponding measurements $\mathbf{y},\mathbf{y}'$, where

$$K(d) = \sum_k |H_k|^2 \phi_l(k|d) \tag{23}$$

defines the kernel of the embedding.

Thus, to embed a Gaussian kernel and linear combinations of it, it suffices to draw the elements of $\mathbf{A}$ from an i.i.d. Gaussian distribution. Alternatively, to embed a Laplacian kernel and linear combinations of it, the elements of $\mathbf{A}$ should be drawn from a Cauchy distribution. The resulting kernels will be described by plugging the corresponding $\phi(\cdot|d)$ in (23), in a similar manner as in (6) and (7):

$$K(d) = \sum_k |H_k|^2 e^{-2(\pi\sigma dk)^2} \tag{24}$$

$$K(d) = \sum_k |H_k|^2 e^{-2\pi\gamma dk} \tag{25}$$

where $\mathbf{A}$ is generated using an i.i.d. zero-mean Gaussian distribution with variance σ^2 or an i.i.d. Cauchy distribution with scale parameter γ and d is the ℓ_2 or the ℓ_1 distance between signals, respectively.

3.2 Angle Embeddings

Another geometric quantity of interest in a number of applications is the angle between signals.

$$d_{\angle}(\mathbf{x}, \mathbf{x}') = \frac{1}{\pi} \arccos \frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\|\mathbf{x}\|_2 \|\mathbf{x}'\|_2} \tag{26}$$

The cosine of the angle is the correlation coefficient of the signals, i.e., their inner product normalized by their respective norms.

Since JL-type embeddings preserve distances and inner products, it is expected that they should preserve angles as well. A tighter bound than a naive application of the definition and the bounds of the previous section was shown in [30] in the context of sparse signals and the RIP. Specifically,

Theorem 4 (Adapted from Thm. 1 and Remark 1 in [30]) *Consider an embedding satisfying the RIP for K-sparse vectors with RIP constant $\delta \leq 1/3$. For any K-sparse $\mathbf{x}$ and $\mathbf{x}'$ with the same support, such that $d_{\angle}(\mathbf{x}, \mathbf{x}') \leq 1/2$ then the angle between the embedded vectors $\mathbf{y}$, and $\mathbf{y}'$ satisfies*

$$-\sqrt{3\delta} \leq d_{\angle}(\mathbf{x}, \mathbf{x}') - d_{\angle}(\mathbf{y}, \mathbf{y}') \leq 3\delta \tag{27}$$

$$\implies \left| d_{\angle}(\mathbf{x}, \mathbf{x}') - d_{\angle}(\mathbf{y}, \mathbf{y}') \right| \leq \sqrt{3\delta}. \tag{28}$$

This result can be used to derive a generalized notion of the RIP, linking the inner product of the embeddings with the geometry of the signals in the signal space [30, Cor. 1].

More recently, an embedding was derived in the context of 1-bit CS, explicitly preserving only angles of signals, not their inner products or magnitudes [37]. In particular, the embedding map

$$\mathbf{y} = f(\mathbf{x}) = \operatorname{sign}(\mathbf{A}\mathbf{x}), \tag{29}$$

where $\mathbf{A}$ has i.i.d. Normally distributed, entries maps the signals to an M-dimensional binary space, denoted $\mathscr{B}^M$, in which the normalized Hamming distance, defined as $d_H(\mathbf{y}, \mathbf{y}') = (\sum_m y_m \oplus y'_m)/M$, is the natural metric.

In [37] it is shown that (29) preserves the angle between signals in the normalized Hamming distance between the measurements, making it a Binary ϵ-stable embedding

Definition 3 Let $\epsilon \in [0, 1)$ a mapping $f : \mathscr{S} \to \mathscr{B}^M$ is a Binary ϵ-stable embedding (BϵSE) of $\mathscr{S}$ if for all $\mathbf{x}, \mathbf{x}' \in \mathscr{S}$,

$$d_{\angle}(\mathbf{x}, \mathbf{x}') - \epsilon \le d_H(\mathbf{y}, \mathbf{y}') \le d_{\angle}(\mathbf{x}, \mathbf{x}') + \epsilon. \tag{30}$$

In other words, a BϵSE is a $(g_I, 0, \epsilon)$ embedding according to Def. 1, with $d_{\mathscr{S}} = d_{\angle}$ and $d_{\mathscr{W}} = d_H$. While the result has been developed for K-sparse vectors, it is straightforward to show that it holds for finite sets of L points using $M = O(\epsilon^{-2} \log L)$.

Theorem 5 (Adapted from Thm. 3 in [37]) *Let $\mathbf{A} \in \mathbb{R}^{M \times N}$ be a matrix generated from an i.i.d. Normal distribution and $\mathscr{S}$ be a set of L points. The map (29) is a BϵSE of $\mathscr{S}$ with probability greater than $1 - 2e^{2(\log P - \epsilon^2 M)}$.*

Subsequent work [3, 46–49] demonstrated variations of this result for infinite signal sets, as a function of their mean width, with varying dependence on ϵ. Furthermore, with some constraints on the signals, it can also be shown for more general matrix ensembles, with elements drawn from subgaussian distributions.

The generalization of the sign function to complex numbers is the phase. As expected in hindsight, similar to sign measurements, phase measurements of the form

$$\mathbf{y} = \angle(\mathbf{A}\mathbf{x}) \tag{31}$$

can also provide stable angle embeddings [14–16]. In particular, if two signals $\mathbf{x}, \mathbf{x}'$ in a finite set $\mathscr{W}$ of size L are measured with a complex random Gaussian matrix, the expected value of the m^{th} element of the measured phase difference is equal to

$$E\left\{\left|\angle\left(e^{i(y_m - y'_m)}\right)\right|\right\} = \pi d_{\angle}(\mathbf{x}, \mathbf{x}'), \tag{32}$$

Note that this way of calculating the phase difference naturally takes phase wrapping into account.

Similarly to the concentration of measure proofs so far, Hoeffding's inequality bounds the probability that the average of M random variables $|\angle(e^{i(y_m - y'_m)})|$ deviates from (32). A natural distance metric in the embedding space is

$$d_{\text{phase}}(\mathbf{y}, \mathbf{y}') = \frac{1}{M} \sum_m \left| \frac{1}{\pi} \angle\left(e^{i(y_m - y'_m)}\right) \right| \tag{33}$$

Using the union bound on L^2 point pairs, a stable embedding guarantee follows

Theorem 6 ([16]) *Consider a finite set $\mathscr{S}$ of L points measured using (31), with $\mathbf{A} \in \mathbb{C}^{M \times N}$ consisting of i.i.d elements drawn from the standard complex normal distribution. With probability greater than $1 - 2e^{2\log L - 2\epsilon^2 M}$ the following holds for all $\mathbf{x}, \mathbf{x}' \in \mathscr{S}$ and corresponding measurements $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^M$.*

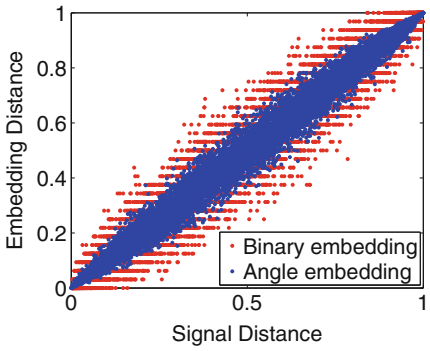


Fig. 3 Comparison of BϵSE (red) with continuous angle embedding (blue) for the same number of measurements. The continuous embedding becomes tighter as signals become more similar. As expected, the binary embedding has higher ambiguity for the same number of measurements.

$$\left| d_{\text{phase}}(\mathbf{y}, \mathbf{y}') - d_{\angle}(\mathbf{x}, \mathbf{x}') \right| \leq \epsilon \tag{34}$$

A complex-valued measurement matrix $\mathbf{A}$ is necessary here. If $\mathbf{A}$ only contains real elements, the information in $\mathbf{y}$ is essentially the sign of the measurement—0 and π for positive and negative measurements, respectively. In that case, the embedding becomes a BϵSE. Furthermore, even though the embedding has an additive ambiguity—i.e., is a $(g_I, 0, \epsilon)$ embedding—it is conjectured that a multiplicative ambiguity guarantee should be possible to derive—i.e., that it is, in fact, a $(g_I, \delta, 0)$ embedding [16].

Figure 3 compares the performance of this embedding with the BϵSE, and demonstrates that, as expected, it exhibits lower ambiguity for the same number of measurements M. Furthermore, it shows that the becomes tighter as signals become similar, supporting the conjecture that a multiplicative-only ambiguity exists.

4 Quantized Embeddings

Quite frequently, the embedding is performed not simply as a dimensionality reduction, but as a compression method. In those cases, the quantity of interest is not the embedding dimensionality, but the number of bits it uses. Therefore, it is necessary to understand how quantization affects the embedding performance, and what the quantizer design trade-offs are.

4.1 Quantization of Continuous Embeddings

Although quantization of some embeddings can be analyzed using the periodic embedding framework we describe above, it is often more convenient, especially in the case of high-rate quantization, to consider it separately, as an additional step after the projection. The following development closely follows [20] and the references within.

In particular, consider a (g, δ, ϵ) embedding which is subsequently quantized using an M-dimensional vector quantizer $Q(\cdot)$. We assume the quantization error is bounded, i.e., $d(Q(\mathbf{x}), \mathbf{x}) \leq E_Q$. The triangle inequality, $|d_{\mathscr{W}}(f(\mathbf{x}), f(\mathbf{w})) - d_{\mathscr{W}}(Q(f(\mathbf{x})), Q(f(\mathbf{w})))| \leq 2E_Q$, implies that the quantized embedding guarantee becomes a $(g, \delta, \epsilon + 2E_Q)$ embedding, with guarantee

$$(1-\delta)g\left(d_{\mathscr{S}}(\mathbf{x},\mathbf{y})\right) - \epsilon - 2E_Q \leq d_{\mathscr{W}}\left(Q(f(\mathbf{x})), Q(f(\mathbf{y}))\right) \leq (1+\delta)g\left(d_{\mathscr{S}}(\mathbf{x},\mathbf{y})\right) + \epsilon + 2E_Q. \quad (35)$$

Theorem 7 (Thm. 3.3 in [20]) *Consider a (g, δ, ϵ) embedding $f(\cdot)$ and a quantizer $Q(\cdot)$ with worst case quantization error E_Q, then the quantized embedding, $Q(f(\cdot))$, is a $(g, \delta, \epsilon + 2E_Q)$ embedding.*

In the specific case of a uniform scalar quantizer with quantization interval Δ, the M-dimensional quantization ℓ_2 error is bounded by $E_Q \leq \sqrt{M}\Delta/2$, assuming the quantizer is designed such that it does not saturate or such that the saturation error is negligible. The interval of the quantizer is a function of the number of bits B used per coefficient $\Delta = 2^{-B+1}S$, where S is the saturation level of the quantizer. Given a fixed rate to be used by the embedding, $R = MB$, the guarantee becomes

$$(1-\delta)g\left(d_{\mathscr{S}}(\mathbf{x},\mathbf{y})\right) - \epsilon - 2^{-\frac{R}{M}+1}\sqrt{M}S \leq \|Q(f(\mathbf{x})) - Q(f(\mathbf{y}))\|_2 \leq (1+\delta)g\left(d_{\mathscr{S}}(\mathbf{x},\mathbf{y})\right) + \epsilon + 2^{-\frac{R}{M}+1}\sqrt{M}S. \quad (36)$$

Note that the $\sqrt{M}$ factor can often be removed, depending on the normalization of the embedding.

Of course, ℓ_2 is not always the appropriate fidelity metric. If the $d_{\mathscr{S}}(\cdot,\cdot)$ corresponds to the ℓ_1 distance, the quantization error is bounded by $E_Q \leq M\Delta/2$. Again, with care in the normalization, the M factor can be removed. If, instead, the ℓ_∞ norm is desired, the quantization error is bounded by $E_Q \leq \Delta/2$.

One of the issues to consider in designing quantized embeddings using a uniform scalar quantizer is the trade-off between the number of bits per dimension and the total number of dimensions used. Since $R = MB$, increasing the number of bits per dimension B under a fixed bit budget R requires decreasing the number of dimensions M. While the former reduces the error due to quantization, the latter will typically increase the uncertainty in the embedding by increasing δ and ϵ.

In the case of randomized embeddings, this trade-off can be quantified through the function $w(\epsilon, \delta)$. Given a fixed probability lower bound to guarantee the embedding, then $M = \Omega(1/w(\epsilon, \delta))$. Since $w(\cdot,\cdot)$ is an increasing function of ϵ and δ, which quantify the ambiguity of the embedding, reducing M increases this ambiguity. On the other hand, the quantization ambiguity, given by $2^{-\frac{R}{M}+2}S\sqrt{M}$ decreases with M.

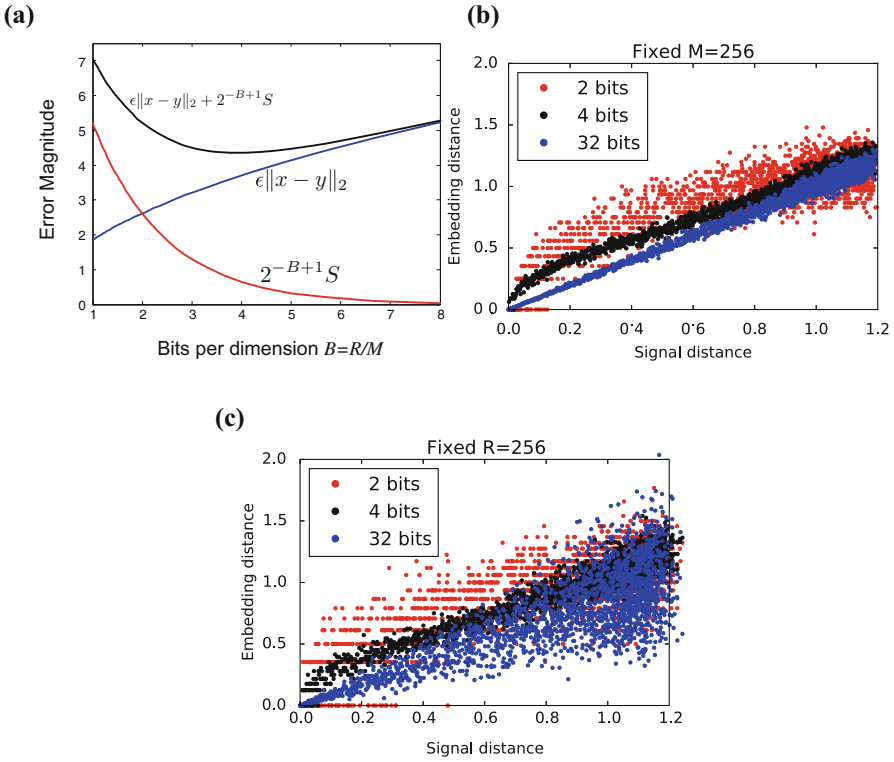


Fig. 4 Illustration of the bits vs. measurements trade-off for quantized JL embeddings. (a) A sketch of the trade-off between bits per coefficient and embedding dimension given a fixed bit-rate for quantized JL embeddings. The error due to the JL ambiguity δ also depends on the norm of the signals being compared, thus affecting the true optimum in practice. Constants were arbitrarily selected for illustration purposes; the true optimum also depends on the true value of the constants. (b) Three different simulation examples using the same $M = 256$, quantized at 2, 4, and 32 bits per dimension, consuming $R = 512,\ 1024$, and 8192 bits, respectively. As expected, the 32 bit embedding performs best, but at a significant rate penalty. (c) Three simulation examples using rate $R = 256$, quantized at 2, 4, and 32 bits per dimension, requiring $M = 128,\ 64$, and 8 dimensions, respectively. As evident, quantizing at 32 bits per coefficient is now suboptimal; the JL-type error due to δ dominates. In this example, 4 bits per coefficient quantization seems to provide the best trade-off overall.

This trade-off is explored, for example, in the context of quantized JL embeddings in [43, 53]. In particular, randomly generated JL embeddings exhibit ambiguity $\delta \sim 1/\sqrt{M}$. On the other hand, the quantization error scales as $E_Q \sim 2^{-B} \sim 2^{-1/M}$. An illustrative example is shown in Figure 4(a): as more bits are used per measurement the ambiguity due to quantization decreases; since fewer measurements are used, the ambiguity due to the embedding's δ increases. Figure 4(b) and (c) further demonstrates this using a simulation experiment. In practice, the optimum depends on assumptions on the signal distance and assumptions about the

constants of proportionality. The same issue exists for non-uniform quantizers and for vector quantizers, manifested with different constants but with the same order of magnitude effects (e.g., see [36]), as well as other embeddings, such as phase embeddings [14]. Unfortunately, other than experimentation with sample data, there is no known principled way to determine the optimal point in the trade-off.

In addition to the generic guarantees above, it is often possible to provide more explicit guarantees under certain conditions. For example, the 1-bit embedding guarantees in Sec. 3.2 were explicitly established from the embedding map. More recently, [34] draws similarities with the Buffon's needle problem to provide a tighter bound on the ℓ_1 embedding distance of quantized dithered JL-type embeddings

Theorem 8 (Adapted from Prop. 2 in [34]) *Let $\mathscr{S} \subset \mathbb{R}^N$ be a set of L points. Consider the map*

$$\mathbf{y} = f(\mathbf{x}) = Q_\epsilon(\mathbf{A}\mathbf{x} + \mathbf{w}), \tag{37}$$

where $\mathbf{A} \in \mathbb{R}^{M\times N}$ has elements drawn from an i.i.d., standard Normal distribution, the dither $\mathbf{w} \in \mathbb{R}^M$ has elements drawn from an i.i.d. distribution, uniform in $[0, \epsilon]$, and $Q_\epsilon(\cdot)$ is an infinite uniform quantizer with interval ϵ.

Given $0 < \delta < 1$, $\epsilon > 0$, and $M > C\delta^{-2}L$, then, with probability greater than $1 - e^{c''\epsilon^2 M}$, the map (37) satisfies

$$(1-\delta)\|\mathbf{x}-\mathbf{x}'\|_2 - c\epsilon\delta \le \frac{c'}{M}\|\mathbf{y}-\mathbf{y}'\|_1 \le (1+\delta)\|\mathbf{x}-\mathbf{x}'\|_2 + c\epsilon\delta \tag{38}$$

for all pairs of points $\mathbf{x}, \mathbf{x}' \in \mathscr{S}$.

A key insight in this result is the switch to the ℓ_1 norm in the embedding space, instead of the ℓ_2 norm used by the JL lemma and earlier results.

4.2 Universal Quantization and Embeddings

In contrast to conventional quantization analysis, universal scalar quantization, first introduced in [13], fundamentally revisits scalar quantization and redesigns the quantizer to have non-contiguous quantization regions. Unfortunately the discontinuous quantization regions render some of the tools introduced in Sec. 4 impractical. Fortunately, analysis based on the design described in Sec. 2.2 can be used instead.

A universal embedding also relies on a JL-style projection, followed by scaling, dithering, and scalar quantization:

$$\mathbf{y} = f(\mathbf{x}) = Q(\Delta^{-1}(\mathbf{A}\mathbf{x} + \mathbf{w})), \tag{39}$$

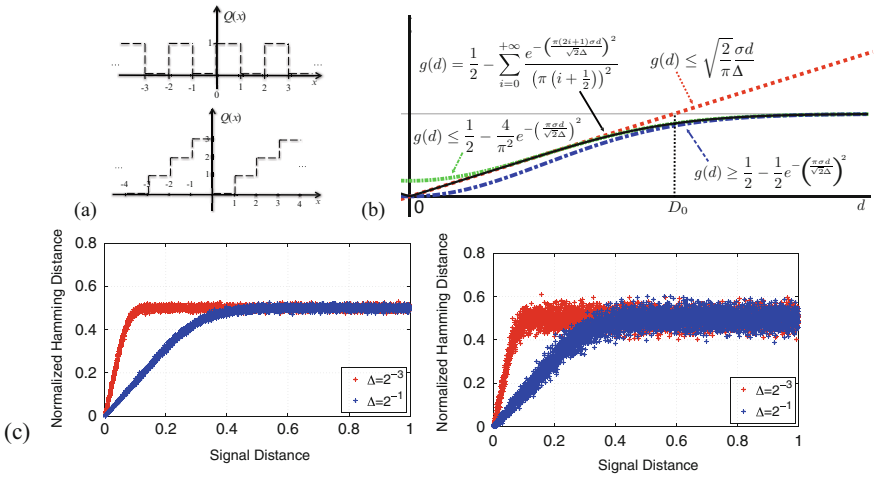


Fig. 5 (a) This non-monotonic quantization function $Q(\cdot)$ allows for universal rate-efficient scalar quantization. This function is equivalent to using a classical multibit scalar quantizer, and preserving only the least significant bits while discarding all other bits. 1-bit shown on top, multi-bit shown on bottom (b) The embedding map $g(d)$ and its bounds produced by the 1-bit quantization function in (a). (c) Experimental verification of the embedding for small and large Δ at high (left) and low (right) bit-rates.

where $\mathbf{A}$ is a $M \times N$ random matrix with $\mathcal{N}(0, \sigma^2)$-distributed, i.i.d. elements, Δ^{-1} a scaling factor, $\mathbf{w}$ a length-M dither vector with i.i.d. elements, uniformly distributed in $[0, 2^B \Delta]$, and $Q(\cdot)$ a B-bit scalar quantizer operating element-wise on its input.

The key component is a modified B-bit scalar quantizer. Fitting the analysis of Thm. 2, the quantizer is designed to be a periodic function with non-contiguous quantization intervals, as shown in Figure 5(a) for $B = 1$ and 2. The quantizer can be thought of as a regular uniform quantizer, computing a multi-bit representation of a signal and preserving only the least significant bits (LSB) of the representation. For example, for a 1-bit quantizer, scalar values in $[2l, 2l + 1)$ quantize to 1 and scalar values in $[2l + 1, 2(l + 1))$, for any integer l, quantize to 0. If $Q(\cdot)$ is a 1-bit quantizer, this method encodes using as many bits as the rows of $\mathbf{A}$, i.e., M bits.

This form of quantization, first proposed in [13] in the context of frame expansions and first used in an embedding in [18] is extensively analyzed in [20].

Theorem 9 (Adapted from Thm. 3.2 in [18]) *Consider a set $\mathcal{S} \subset \mathbb{R}^N$ with L points embedded using (39), as described above. For all $\mathbf{x}, \mathbf{x}' \in \mathcal{S}$, the embedding satisfies*

$$g\left(\|\mathbf{x} - \mathbf{y}\|_2\right) - \epsilon \le d_H\left(\mathbf{y}, \mathbf{y}'\right) \le g\left(\|\mathbf{x} - \mathbf{y}\|_2\right) + \epsilon, \tag{40}$$

with probability $1 - 2e^{2\log L - 2\epsilon^2 M}$ with respect to the measure of $\mathbf{A}$ and $\mathbf{w}$. In (40), $d_H(\cdot, \cdot)$ is the Hamming distance of the embedded signals, the function $f(\cdot)$ is as

specified in (39), and $g(d)$ is the map

$$g(d) = \frac{1}{2} - \sum_{i=0}^{+\infty} \frac{e^{-\left(\frac{\pi(2i+1)\sigma d}{\sqrt{2}\Delta}\right)^2}}{(\pi(i+1/2))^2}, \tag{41}$$

Furthermore, the distance map $g(d)$ can be bounded using

$$g(d) \geq \frac{1}{2} - \frac{1}{2} e^{-\left(\frac{\pi\sigma d}{\sqrt{2}\Delta}\right)^2}, \tag{42}$$

$$g(d) \leq \frac{1}{2} - \frac{4}{\pi^2} e^{-\left(\frac{\pi\sigma d}{\sqrt{2}\Delta}\right)^2}, \tag{43}$$

$$g(d) \leq \min\left(\sqrt{\frac{2}{\pi}}\frac{\sigma d}{\Delta}, \frac{1}{2}\right), \tag{44}$$

as shown in Figure 5(b).

The upper bound (44) also provides a very good approximation of the embedding, as also evident in the figure. The map is approximately linear for small d and becomes constant, equal to $1/2$, exponentially fast as d exceeds a threshold D_0. The slope of the linear section is determined by the parameter ratio σ/Δ, thus specifying the distance threshold $D_0 \approx \Delta\sqrt{\pi}/2\sqrt{2}\sigma$. In other words, the embedding ensures that the Hamming distance of the embedded signals is approximately proportional to the ℓ_2 distance between the original signals, as long as that ℓ_2 distance was smaller than D_0. Distances greater than D_0 are shrunk to Hamming distance $\approx 1/2$. In other words, the embedding can only reveal that the distance is greater than approximately D_0 but not how much greater.

This embedding enables a trade-off between the threshold D_0 and the slope of the linear part, which determines its ambiguity through (10). Assuming the linear approximation in (44), it is straightforward to show that the ratio of the range of distances preserved, as measured through D_0, to the ambiguity in preserving distances in the linear part, as measured through (44) remains constant as the embedding parameters Δ and σ change keeping a fixed embedding dimension M, and, therefore, a fixed rate $R = M$. In contrast to the trade-off depicted in Figure 4, both D_0 and the slope of the linear part are straightforward to compute and do not depend on difficult-to-characterize constants.

Figure 5(c) illustrates how the embedding behaves in simulations for smaller (red) and larger (blue) Δ and for higher (left) and lower (right) bit-rates. The figure plots the embedding (Hamming) distance as a function of the signal distance for randomly generated pairs of signals. The thickness of the curve is quantified by ϵ, whereas the slope of the upward sloping part is quantified by Δ.

In addition to 1-bit universal embedding for finite signal sets, [20] generalizes the guarantees to infinite sets and to multi-bit embeddings. Of course, multibit

embeddings re-introduce a similar trade-off as in Figure 4, which has not been explored in the literature.

In addition to the embedding properties, information-theoretic arguments can be used to guarantee that universal embeddings can preserve the query privacy [18, 39]. This can be a very useful property in implementing secure protocols for signal-based querying and retrieval in privacy-sensitive applications [52].

5 Discussion

As evident from the discussion above, embeddings can play a significant role in modern data processing systems. In this chapter, we have only presented a selective overview of the area, some important results, and pointers for further reading. However, increasing demand for efficient data processing has reinvigorated the field, leading to a flurry of new results in a number of interesting directions.

While we have only discussed ℓ_p distance and angle embeddings in their various forms, there exist embeddings for more exotic distance metrics, such as the edit distance [6, 8, 42, 44]. Furthermore, while JL embeddings and the RIP preserve ℓ_2 distances, there is a large body of work in preserving other similarity measurements, such as ℓ_p distances for various p's [32, 35, 36, 45, 46]. It should be noted that in some cases, such as embedding the ℓ_1 distance into a smaller ℓ_1 space, such embeddings have been proven impossible [22]. Still, even in such cases, embeddings have been developed that hold with high, but not exponentially decreasing, probability [32].

A principal motivation for dimensionality reduction is often a reduction in computational complexity. However, the cost of storing and using a dense, fully randomized, embedding matrix can often be prohibitive. Fast transforms have been developed in a number of cases [4, 24, 57], enabling efficient computation of the transform, often without explicitly storing the matrix but using an algorithm, such as the fast Fourier transform (FFT). Still, even when the computation is efficient and the cost of storing the matrix is mitigated, the complexity of using the embedding for very large data retrieval can sometimes be daunting. While the dimensionality reduction definitely helps, the amount of the data, i.e., the number of data points, can be such that search is impossible even if the complexity is linear in the amount of data.

In such cases, locality-sensitive hashing (LSH) methods—which significantly reduce the computational complexity of near-neighbor computation—can be very helpful [5, 26, 33]. These methods are intimately connected to randomized embeddings. The LSH literature shares a lot of the tools, especially with quantized embeddings, such as randomized projections, dithering, and quantization. The goal, however, is different. Given a query point, LSH will return its near neighbors very efficiently, using $O(1)$ computation. This efficiency comes at a cost: no attempt is made to represent the distances between neighbors. When LSH is used to compare signals it only provides a binary decision, namely whether the distance between the signals is smaller than a radius or not. There is no guarantee that further information

will be preserved. Thus, LSH may not be suitable for applications that require more accurate distance information. Still, the similarity of the methods suggests that some of the quantized embedding designs can be used as LSH functions. While there are examples of such use, [39], this is still an underdeveloped connection, especially for recent embedding designs. Techniques that learn a hash, such as spectral hashing [56] and LDAHash[55], also have strong similarities with embeddings and embedding learning.

Of course, this is a rich topic and it is impossible to exhaustively cover in this chapter. Our hope is that our development exposes the basic principles, some of the foundational work, and some interesting recent developments. Our goal is to expose embeddings to a wider community, establishing them as an important tool, essential in the belt of any data scientist.

References

1. D. Achlioptas, Database-friendly random projections: Johnson-Lindenstrauss with binary coins. J. Comput. Syst. Sci. **66**, 671–687 (2003)
2. D. Achlioptas, F. Mcsherry, B. Schölkopf, Sampling techniques for kernel methods, in *Advances in Neural Information Processing Systems* (2002), pp. 335–342
3. A. Ai, A. Lapanowski, Y. Plan, R. Vershynin, One-bit compressed sensing with non-gaussian measurements. Linear Algebra Appl. **441**, 222–239 (2014)
4. N. Ailon, B. Chazelle, Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform, in *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing* (2006), pp. 557–563
5. A. Andoni, P. Indyk, Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Commun. ACM **51**(1), 117–122 (2008). DOI:10.1145/1327452.1327494
6. A. Andoni, M. Deza, A. Gupta, P. Indyk, S. Raskhodnikova, Lower bounds for embedding edit distance into normed spaces, in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (2003), pp. 523–526
7. A.S. Bandeira, D.G. Mixon, B. Recht, Compressive classification and the rare eclipse problem (2014), arXiv preprint arXiv:1404.3203
8. Z. Bar-Yossef, T. Jayram, R. Krauthgamer, R. Kumar, Approximating edit distance efficiently, in *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science, 2004* (IEEE, Los Alamitos, 2004), pp. 550–559
9. R. Baraniuk, M. Wakin, Random projections of smooth manifolds. Found. Comput. Math. **9**(1), 51–77 (2009)
10. R. Baraniuk, M. Davenport, R. DeVore, M. Wakin, A simple proof of the restricted isometry property for random matrices. Const. Approx. **28**(3), 253–263 (2008)
11. R. Baraniuk, V. Cevher, M. Duarte, C. Hegde, Model-based compressive sensing. IEEE Trans. Inf. Theory **56**(4), 1982–2001 (2010)
12. T. Blumensath, M. Davies, Sampling theorems for signals from the union of finite-dimensional linear subspaces. IEEE Trans. Inf. Theory **55**(4), 1872–1882 (2009)
13. P.T. Boufounos, Universal rate-efficient scalar quantization. IEEE Trans. Inf. Theory **58**(3), 1861–1872 (2012). DOI:10.1109/TIT.2011.2173899
14. P.T. Boufounos, Angle-preserving quantized phase embeddings, in *Proceedings of SPIE Wavelets and Sparsity XV*, San Diego, CA (2013)
15. P.T. Boufounos, On embedding the angles between signals, in *Signal Processing with Adaptive Sparse Structured Representations*, Lausanne, Switzerland (2013)

16. P.T. Boufounos, Sparse signal reconstruction from phase-only measurements, in *Proceedings of International Conference on Sampling Theory and Applications*, Bremen, Germany (2013)
17. P.T. Boufounos, H. Mansour, Universal embeddings for kernel machine classification, in *Proceedings of Sampling Theory and Applications*, Washington, DC (2015)
18. P.T. Boufounos, S. Rane, Secure binary embeddings for privacy preserving nearest neighbors, in *Proceedings of the IEEE Workshop on Information Forensics and Security*, Foz do Iguau, Brazil (2011). DOI:10.1109/WIFS.2011.6123149
19. P.T. Boufounos, S. Rane, Efficient coding of signal distances using universal quantized embeddings, in *Proceedings of Data Compression Conference*, Snowbird, UT (2013)
20. P.T. Boufounos, S. Rane, H. Mansour, Representation and coding of signal geometry (2015), arXiv preprint arXiv:1512.07636
21. J. Bourgain, S. Dirksen, J. Nelson, Toward a unified theory of sparse dimensionality reduction in euclidean space. Geom. Funct. Anal. **25**(4), 1009–1088 (2015)
22. B. Brinkman, M. Charikar, On the impossibility of dimension reduction in l 1. J. ACM **52**(5), 766–788 (2005)
23. E. Candès, The restricted isometry property and its implications for compressed sensing. C. R. Acad. Sci. I **346**(9–10), 589–592 (2008)
24. E.J. Candes, J.K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
25. S. Dasgupta, A. Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algoritm. **22**(1), 60–65 (2003)
26. M. Datar, N. Immorlica, P. Indyk, V.S. Mirrokni, Locality-sensitive hashing scheme based on p-stable distributions, in *Proceedings of the Twentieth Annual Symposium on Computational Geometry* (ACM, New York, 2004), pp. 253–262
27. M.A. Davenport, P.T. Boufounos, M.B. Wakin, R.G. Baraniuk, Signal processing with compressive measurements. IEEE J. Sel. Top. Sign. Proces. **4**(2), 445–460 (2010). DOI:10.1109/JSTSP.2009.2039178. http://dx.doi.org/10.1109/JSTSP.2009.2039178
28. S. Dirksen, Dimensionality reduction with subgaussian matrices: a unified theory. Found. Comput. Math. **16**(5), 1367–1396
29. Y. Eldar, M. Mishali, Robust recovery of signals from a structured union of subspaces. IEEE Trans. Inf. Theory **55**(11), 5302–5316 (2009)
30. J. Haupt, R. Nowak, A generalized restricted isometry property. Tech. rep., University of Wisconsin-Madison (2007)
31. C. Hegde, A. Sankaranarayanan, W. Yin, R. Baraniuk, NuMax: a convex approach for learning near-isometric linear embeddings. IEEE Trans. Signal Process. **63**(22), 6109–6121 (2015). DOI:10.1109/TSP.2015.2452228
32. P. Indyk, Stable distributions, pseudorandom generators, embeddings, and data stream computation. J. ACM **53**(3), 307–323 (2006)
33. P. Indyk, R. Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality, in *ACM Symposium on Theory of computing* (1998), pp. 604–613
34. L. Jacques, A quantized Johnson-Lindenstrauss lemma: the finding of buffon's needle. IEEE Trans. Inf. Theory **61**(9), 5012–5027 (2015). DOI:10.1109/TIT.2015.2453355
35. L. Jacques, D.K. Hammond, J.M. Fadili, Dequantizing compressed sensing: when oversampling and non-gaussian constraints combine. IEEE Trans. Inf. Theory **57**(1), 559–571 (2011)
36. L. Jacques, D.K. Hammond, J.M. Fadili, Stabilizing nonuniformly quantized compressed sensing with scalar companders. IEEE Trans. Inf. Theory **59**(12), 7969–7984 (2013)
37. L. Jacques, J.N. Laska, P.T. Boufounos, R.G. Baraniuk, Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. IEEE Trans. Inf. Theory **59**(4) (2013). DOI:10.1109/TIT.2012.2234823. http://dx.doi.org/10.1109/TIT.2012.2234823
38. T. Jayram, D.P. Woodruff, Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. ACM Trans. Algoritm. **9**(3), 26 (2013)
39. A. Jimenez, B. Raj, J. Portelo, I. Trancoso, Secure modular hashing, in *IEEE International Workshop on Information Forensics and Security* (IEEE, Piscataway, 2015), pp. 1–6

40. W. Johnson, J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space. Contemp. Math. **26**, 189–206 (1984)
41. F. Krahmer, R. Ward, New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. SIAM J. Math. Anal. **43**(3), 1269–1281 (2011)
42. V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals. Sov. Phys. Dokl. **10**(8), 707–710 (1966)
43. M. Li, S. Rane, P.T. Boufounos, Quantized embeddings of scale-invariant image features for mobile augmented reality, in *Processing of the IEEE International Workshop on Multimedia Signal Processing*, Banff, Canada (2012)
44. R. Ostrovsky, Y. Rabani, Low distortion embeddings for edit distance. J. ACM **54**(5), 23 (2007)
45. D. Otero, G.R. Arce, Generalized restricted isometry property for alpha-stable random projections, in *IEEE International Conference on Acoustics, Speech and Signal Processing* (2011), pp. 3676–3679
46. S. Oymak, B. Recht, Near-optimal bounds for binary embeddings of arbitrary sets (2015), arXiv preprint arXiv:1512.04433
47. Y. Plan, R. Vershynin, One-bit compressed sensing by linear programming. Commun. Pure Appl. Math. **66**(8), 1275–1297 (2013)
48. Y. Plan, R. Vershynin, Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. IEEE Trans. Inf. Theory **59**(1), 482–494 (2013)
49. Y. Plan, R. Vershynin, Dimension reduction by random hyperplane tessellations. Discret. Comput. Geom. **51**(2), 438–461 (2014)
50. G. Puy, M. Davies, R. Gribonval, Recipes for stable linear embeddings from Hilbert spaces to $\mathbb{R}^m$ (2015), arXiv preprint arXiv:1509.06947
51. A. Rahimi, B. Recht, Random features for large-scale kernel machines, in *Advances in Neural Information Processing Systems* (2007), pp. 1177–1184
52. S. Rane, P.T. Boufounos, Privacy-preserving nearest neighbor methods: comparing signals without revealing them, in *IEEE Signal Processing Magazine* (2013). DOI:10.1109/MSP.2012.2230221, http://dx.doi.org/10.1109/MSP.2012.2230221
53. S. Rane, P.T. Boufounos, A. Vetro, Quantized embeddings: an efficient and universal nearest neighbor method for cloud-based image retrieval, in *Proceedings of SPIE Applications of Digital Image Processing XXXVI*, San Diego, CA (2013)
54. A. Sadeghian, B. Bah, V. Cevher, Energy-aware adaptive bi-Lipschitz embeddings, in *Proceedings of International Conference on Sampling Theory and Applications*, Bremen, Germany (2013)
55. C. Strecha, A. Bronstein, M. Bronstein, P. Fua, LDAHash: improved matching with smaller descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **34**(1), 66–78 (2012). DOI:10.1109/TPAMI.2011.103
56. Y. Weiss, A. Torralba, R. Fergus, Spectral hashing, in *Advances in Neural Information Processing Systems 21* (MIT, London, 2009), pp. 1753–1760
57. X. Yi, C. Caramanis, E. Price, Binary embedding: fundamental limits and fast algorithm, in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, vol. 37 (2015), pp. 2162–2170

Distributed Noise-Shaping Quantization: II. Classical Frames

Evan Chou and C. Sinan Güntürk

Abstract This chapter constitutes the second part in a series of papers on distributed noise-shaping quantization. In the first part, the main concept of distributed noise shaping was introduced and the performance of distributed beta encoding coupled with reconstruction via beta duals was analyzed for random frames (Chou and Güntürk, Constr Approx 44(1):1–22, 2016). In this second part, the performance of the same method is analyzed for several classical examples of deterministic frames. Particular consideration is given to Fourier frames and frames used in analog-to-digital conversion. It is shown in all these examples that entropic rate-distortion performance is achievable.

Keywords Finite frames • quantization • A/D conversion • noise shaping • beta encoding • beta duals.

1 Introduction

The "analysis formulation" for the quantization problem (in short, the *analysis problem*) associated to any given frame seeks to find out how well signals can be approximated after quantizing signal measurements that are taken using this frame (see, e.g., [9]). More concretely, let $\Phi := (\varphi_\alpha)_{\alpha\in I}$ be a (finite) frame in a real or complex (finite dimensional) Hilbert space $\mathscr{H}$ with inner-product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$, and $L \geq 2$ be a given integer representing the number of quantization levels to be used. The *analysis distortion* $\mathscr{D}_{\mathrm{a}}(\Phi, L)$ (see [6]) is formally defined by the quantity

$$\inf\left\{\sup_{\|x\|\leq 1}\inf_{q\in\mathscr{A}^I}\left\|x-\sum_{\alpha\in I}q_\alpha\psi_\alpha\right\| : (\psi_\alpha) \text{ is any dual frame of } \Phi \text{ and } |\mathscr{A}| = L\right\}.$$

E. Chou
Google, New York, NY, USA
e-mail: chou@cims.nyu.edu

C.S. Güntürk (✉)
Courant Institute, NYU, 251 Mercer Street, New York, NY 10012, USA
e-mail: gunturk@cims.nyu.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4_8

Here $\mathscr{A}$ stands for the quantization alphabet, i.e. any subset of the underlying field $\mathbb{F}$ (which equals $\mathbb{R}$ or $\mathbb{C}$) of L elements.

As it was described in [6] (albeit with slightly differing notation), the analysis distortion corresponds to a practical encoding-decoding scenario: The encoder chooses $\mathscr{A}$ and quantizes the signal measurements $(\langle x, \varphi_\alpha \rangle)_{\alpha \in I}$ to generate the discrete output $(q_\alpha)_{\alpha \in I}$ in $\mathscr{A}$, knowing that the decoder will produce the approximation $\sum q_\alpha \psi_\alpha$ where $\Psi := (\psi_\alpha)_{\alpha \in I}$ is some dual frame of Φ. In this sense, the quantization alphabet $\mathscr{A}$ and the dual frame Ψ are available to both the encoder and the decoder. $\mathscr{A}$ and Ψ should be seen as system parameters which can be optimized but must remain fixed for all signals x in the unit ball of $\mathscr{H}$. The analysis distortion then measures the best achievable reconstruction error bound (over all $\mathscr{A}$ and Ψ) that is valid uniformly for all x.

It is easy to see that the analysis distortion is invariant under scaling and unitary transformations. More precisely, given any frame $\Phi := (\varphi_\alpha)_{\alpha \in I}$ in $\mathscr{H}_1$, unitary transformation $U : \mathscr{H}_1 \to \mathscr{H}_2$, and nonzero scalar $c \in \mathbb{F}$, we have

$$\mathscr{D}_{\mathrm{a}}(cU\Phi, L) = \mathscr{D}_{\mathrm{a}}(\Phi, L)$$

where $cU\Phi$ stands for the frame $(cU\varphi_\alpha)_{\alpha \in I}$ in $\mathscr{H}_2$. Hence it is always possible to reduce the discussion of the analysis distortion of frames to that of matrices (finite or infinite) as it was done in [6] which focused on random matrices. In this paper it will be more convenient for us to maintain the general framework of Hilbert spaces to allow for the possibility of working with examples of frames that are not naturally presented as matrices.

The rate-distortion performance of any quantization method is constrained by universal entropic (or volumetric) bounds. For the analysis distortion, we have (see, e.g., [6])

$$\mathscr{D}_{\mathrm{a}}(\Phi, L) \geq L^{-N/d} \tag{1}$$

for all frames Φ in $\mathbb{R}^d$ of size $|I| =: N$, and all L. One of the main results of [6] is that if the φ_α are chosen independently from the standard Gaussian distribution on $\mathbb{R}^d$, then for any $\eta > 0$, the event

$$\left\{ \mathscr{D}_{\mathrm{a}}(\Phi, L) \leq \sqrt{d} L^{-(1-\eta)N/d} \text{ for all } L \geq 2 \right\} \tag{2}$$

holds with probability at least $1 - \exp(-c\eta^2 N)$, provided d and N/d are sufficiently large (depending only on η). Of course, with the observation made in the previous paragraph concerning unitary invariance, (1) and (2) continue to hold in any d-dimensional real Hilbert space $\mathscr{H}$ where the standard Gaussian distribution may be defined by means of any orthonormal basis of $\mathscr{H}$.

Complex Hilbert spaces were not studied in [6] but can be handled with relatively straightforward modifications which we will introduce in this paper (see Section 2 and the Appendix). Note, in particular, that the universal lower bound (1) needs to be

replaced by $L^{-N/2d}$ for the complex case; this can be seen by porting the Lebesgue measure on $\mathbb{R}^{2d}$ on to $\mathbb{C}^d$ and repeating the volume-covering argument given in [6].

1.1 Statement of the Main Results

This is the second part in an ongoing series of works on *distributed noise-shaping quantization*. In the first paper [6], the analysis distortion bound in (2) was achieved by means of a general algorithmic framework called *distributed noise-shaping*, and in particular, using the method of *distributed beta encoding* coupled with reconstruction via *beta duals*. In this second paper we will apply this method to some classical examples of deterministic frames.

The frames that we will consider in this paper fall into a general category we call *unitarily generated frames*. In essence, this means that the index set I can be chosen as $\mathbb{Z}_N$ or $\mathbb{Z}$ depending on the size of the frame, and there exists a unitary operator U on $\mathscr{H}$ such that

$$\varphi_n = U\varphi_{n-1} \tag{3}$$

for all $n \in I$. (See Section 4 for the technical definition.) Well-known examples that fall into this category include Fourier frames, real harmonic frames, and frames of (uniform) translates.

The main result of this paper in the case of unitarily generated frames of size N in d dimensions, assuming N is a multiple of d and a certain technical condition satisfied by Fourier frames, is that

$$\mathscr{D}_{\mathrm{a}}(\Phi, L) \lesssim c(\varphi_0) N d^{-1} \cdot \begin{cases} L^{-N/d}, & \text{if } \mathbb{F} = \mathbb{R} \text{ and } L \geq 2, \\ \lfloor \sqrt{L} \rfloor^{-N/d}, & \text{if } \mathbb{F} = \mathbb{C} \text{ and } L \geq 4, \end{cases} \tag{4}$$

where $c(\varphi_0)$ is a constant that is independent of N and L (see Theorem 2). Generically, $c(\varphi_0)$ is of order $\sqrt{d}$. Note that the bound in (4) behaves better than the one in (2), and considering (1), it is essentially optimal.

The case of infinite dimensional Hilbert spaces requires some modifications and we only consider the classical problem of analog-to-digital conversion of bandlimited functions via uniform sampling and reconstruction by interpolation. With the help of the beta dual machinery, first we establish a new sampling theorem, and then we show that for uniform sampling of real-valued bandlimited functions with oversampling ratio λ, the analysis distortion can be bounded by $C\lambda L^{-\lambda+1}$ which is the infinite dimensional analog of (4) (see Section 5).

2 Background and Review of Methodology

In this section we will review the general theory of noise-shaping quantizers as well as the particular method of distributed beta encoding and beta duals. Further details on the methodology can be found in [6].

2.1 Basics of Noise Shaping for Frames

The main principle of noise-shaping quantization is to arrange for the quantization error (the quantization "noise") to be close to the kernel of the reconstruction operator. For concreteness we assume here that I is a finite index set, but the principle extends to infinite dimensional cases with suitable modifications. Given the measurements $y_\alpha := \langle x, \varphi_\alpha \rangle$, $\alpha \in I$, of a signal $x \in \mathscr{H}$ using a frame $\Phi := (\varphi_\alpha)_{\alpha \in I}$, a noise-shaping quantizer seeks to find a solution (u, q) to the equation

$$y - q = Hu \tag{5}$$

where $y := (y_\alpha) \in \mathbb{F}^I$, $q := (q_\alpha) \in \mathscr{A}^I$, $H : \mathbb{F}^I \to \mathbb{F}^I$ is a linear operator called the "noise transfer operator" of the noise-shaping quantizer, and $u \in \mathbb{F}^I$ is an auxiliary variable, often called the "state vector." Sigma-delta ($\Sigma\Delta$) modulators constitute the most important example of traditional noise-shaping quantizers (see [10] for an engineering perspective, [7, 8] for mathematical expositions, and [2, 9] for applications to finite frames).

Given any dual frame $\Psi := (\psi_\alpha)$ of Φ, we then have

$$x - \sum_{\alpha \in I} q_\alpha \psi_\alpha = \sum_{\alpha \in I} (Hu)_\alpha \psi_\alpha = \sum_{\alpha' \in I} u_{\alpha'} \psi^H_{\alpha'} \tag{6}$$

where

$$\psi^H_{\alpha'} := \sum_{\alpha \in I} H_{\alpha,\alpha'} \psi_\alpha$$

and H has the matrix representation $(H_{\alpha,\alpha'})$. Noise-shaping quantizers are typically designed to keep $\|u\|_\infty$ small. Ideally $\|u\|_\infty$ should be controlled independently of $|I|$; such a scheme is called *stable*. With stability, the error representation (6) results in the effective bound

$$\left\| x - \sum_{\alpha \in I} q_\alpha \psi_\alpha \right\| \le \|u\|_\infty \|\Psi^H\|_{\ell^\infty(I) \to \mathscr{H}} \tag{7}$$

where $\Psi^H : \ell^\infty(I) \to \mathscr{H}$ is the operator given by

$$\Psi^H(u) := \sum_{\alpha \in I} u_\alpha \psi_\alpha^H.$$

Picking an orthogonal basis for $\mathscr{H}$, we may identify the frame Ψ with a matrix (which we may also denote by Ψ) whose columns consist of the coefficients of ψ_α in this basis. Then we have $\Psi^H = \Psi H$ and the operator norm $\|\Psi^H\|_{\ell^\infty(I)\to\mathscr{H}}$ equals the matrix norm $\|\Psi H\|_{\infty\to 2}$.

With the objective of minimizing the error bound (7), the main question is then how to choose H and the dual frame Ψ while ensuring stability of u. In the next subsection we will review a particular choice of H and Ψ that was proposed in [6], namely the noise transfer operator of *distributed beta encoding* and the *beta dual* of Φ, respectively. To ensure stability, we will employ the common toolkit known as the *greedy quantizer*, which was also used in [6]. The small but necessary modifications for complex-valued measurements are explained in the Appendix where a general form of the greedy quantizer which results in some additional improvements is also given.

2.2 Distributed Beta Encoding and Beta Duals of Frames

For any given frame $\Phi := (\varphi_\alpha)_{\alpha\in I}$ in $\mathscr{H}$, pick a partition $\Pi := (I_0, \ldots, I_{p-1})$ of I where $N := |I| \geq p \geq d := \dim(\mathscr{H})$, and for each j in

$$[p] := \{0, \ldots, p-1\},$$

pick a scalar $\beta_j \in \mathbb{F}$ with magnitude at least 1 and a bijection $\sigma_j : [N_j] \to I_j$ where N_j denotes $|I_j|$. Define

$$\zeta_j := \sum_{n\in[N_j]} \bar{\beta}_j^{-n} \varphi_{\sigma_j(n)};\ j \in [p].$$

Suppose $(\zeta_j)_0^{p-1}$ is itself a frame for $\mathscr{H}$. Let $(\eta_j)_0^{p-1}$ be any dual frame of $(\zeta_j)_0^{p-1}$ and define a new collection of vectors $\Psi := (\psi_\alpha)_{\alpha\in I}$ via

$$\psi_{\sigma_j(n)} := \beta_j^{-n}\eta_j;\ n \in [N_j],\ j \in [p].$$

Then Ψ is a dual of Φ because

$$\sum_{\alpha\in I} \langle x, \varphi_\alpha\rangle \psi_\alpha = \sum_{j\in[p]} \sum_{n\in[N_j]} \langle x, \varphi_{\sigma_j(n)}\rangle \psi_{\sigma_j(n)}$$

$$
\begin{aligned}
&= \sum_{j\in[p]} \sum_{n\in[N_j]} \langle x, \bar{\beta}_j^{-n} \varphi_{\sigma_j(n)} \rangle \eta_j \\
&= \sum_{j\in[p]} \langle x, \zeta_j \rangle \eta_j \\
&= x.
\end{aligned}
$$

We assume that $(\eta_j)_0^{p-1}$ is chosen to be the canonical dual of $(\zeta_j)_0^{p-1}$, denoted by $(\widetilde{\zeta}_j)_0^{p-1}$. Suppressing the underlying partition Π, the bijections (σ_j), and the values (β_j), we then call Ψ the *beta dual* of Φ. We also say that $(\zeta_j)_0^{p-1}$ is the *beta condensation* of Φ. (See [6] for a more general definition of condensation of frames.)

The concept of beta duals is inherently tied with what we call *distributed beta encoding*. This is a noise-shaping quantization method which is carried out via the system of difference equations

$$
y_{\sigma_j(n)} - q_{\sigma_j(n)} = u_{\sigma_j(n)} - \beta_j u_{\sigma_j(n-1)}, \ n \in [N_j], \ j \in [p], \tag{8}
$$

where for notational convenience we set $u_{\sigma_j(-1)} := 0$. In other words, the noise-transfer operator H has a block-diagonal matrix representation (see [6]).

The significance of distributed beta encoding coupled with beta duals for reconstruction lies in the following calculation:

$$
\begin{aligned}
x - \sum_{\alpha\in I} q_\alpha \psi_\alpha &= \sum_{j\in[p]} \sum_{n\in[N_j]} (y_{\sigma_j(n)} - q_{\sigma_j(n)}) \psi_{\sigma_j(n)} \\
&= \sum_{j\in[p]} \left(\sum_{n\in[N_j]} (u_{\sigma_j(n)} - \beta_j u_{\sigma_j(n-1)}) \beta_j^{-n} \right) \widetilde{\zeta}_j \\
&= \sum_{j\in[p]} u_{\sigma_j(N_j-1)} \beta_j^{-N_j+1} \widetilde{\zeta}_j.
\end{aligned} \tag{9}
$$

Let A_ζ be the lower frame bound of $(\zeta_j)_0^{p-1}$, that is,

$$
\sum_{j\in[p]} |\langle x, \zeta_j \rangle|^2 \geq A_\zeta \|x\|^2 \ \text{ for all } x \in \mathscr{H}.
$$

Then, as is well known in frame theory, we have

$$
\left\| \sum_{j\in[p]} a_j \widetilde{\zeta}_j \right\| \leq \|a\|_2 / \sqrt{A_\zeta} \ \text{ for all } a \in \mathbb{F}^p. \tag{10}
$$

(In frame theory terminology, this result is a consequence of the fact that if T, T^*, and $S := T^*T$ denote the analysis, the synthesis, and the frame operators for $(\zeta_j)_0^{p-1}$, respectively, then $S^{-1}T^*$ is the synthesis operator for $(\widetilde{\zeta}_j)_0^{p-1}$ with norm equal to $\|S^{-1/2}\| = 1/\sqrt{A_\zeta}$.) Combining (9) and (10), it follows that

$$\left\| x - \sum_{\alpha\in I} q_\alpha \psi_\alpha \right\| \le \frac{1}{\sqrt{A_\zeta}} \left\| \left(u_{\sigma_j(N_j-1)} \beta_j^{-N_j+1} \right)_0^{p-1} \right\|_2 \le \|u\|_\infty \beta_*^{-N_*+1} \sqrt{\frac{p}{A_\zeta}}, \quad (11)$$

where $\beta_* := \min_j |\beta_j|$, $N_* := \min_j N_j$. Note that $N_* \le N/p$ but there always exists a partition Π that achieves $N_* = \lfloor N/p \rfloor$. As in [6] we will assume this is the case. In fact, in all of the examples considered in this paper p will divide N and all the β_j will be equal to a common positive real number that we will call β.

We show in the Appendix that

- for $\mathbb{F} = \mathbb{R}$, the condition $\beta + \|y\|_\infty/\delta \le L$ is sufficient to guarantee that (8) is solvable with $\|u\|_\infty \le \delta$ and $q \in \mathscr{A}^I$ for some $\mathscr{A} \subset \mathbb{R}$, $|\mathscr{A}| = L$, and
- for $\mathbb{F} = \mathbb{C}$, the condition $\beta + \|y\|_\infty/\delta \le \lfloor \sqrt{L} \rfloor$ is sufficient to guarantee that (8) is solvable with $\|u\|_\infty \le \sqrt{2}\delta$ and $q \in \mathscr{A}^I$ for some $\mathscr{A} \subset \mathbb{C}$, $|\mathscr{A}| = L$.

Since $\beta \ge 1$, the above sufficient condition for the complex case can only be invoked if $L \ge 4$. However, $L = 3$ can also be employed using a different quantizer. We show in the Appendix that (8) is solvable for any $\beta < 4/3$. Note that $\beta \le \sqrt{L}$ is a necessary condition for the complex case due to the entropic lower bound $L^{-N/2d}$ for the analysis distortion. Currently we do not know if the gap from $4/3$ to $\sqrt{3}$ can be closed for $L = 3$. Also see [1] where the case $L = 3$ appears for $\beta = 1$.

In order to bound $\mathscr{D}_{\mathrm{a}}(\Phi, L)$ via (11), a two-level strategy can be executed: At the basic level, the system parameters β and δ should be chosen optimally, i.e. so as to minimize $\delta\beta^{-N_*+1}$, subject to one of the sufficient stability conditions above. At the more advanced level, the partition Π and the bijections $(\sigma_j)_0^{p-1}$ should also be seen as system parameters that can be chosen optimally so as to minimize $1/\sqrt{A_\zeta}$. In other words, the beta condensation frame $(\zeta_j)_0^{p-1}$ should be made as tight as possible. This second stage of optimization was not invoked for random frames in [6] (except for the value of p) and it will not be invoked for the classical examples considered in this paper either because natural partition choices will work near optimally; however, in other specific examples there may be need to consider it. Here note that A_ζ implicitly depends on β too, but for the examples we will study in this paper this dependence will not play a critical role.

It is worth noting that the case $\beta = 1$ with $p = 1$ corresponds to first-order $\Sigma\Delta$ quantization which has been studied in depth for finite frames [2]. The second level of optimization that arises in this case has been found to relate to the traveling salesman problem [11]. Higher-order $\Sigma\Delta$ schemes perform better but they remain sub-optimal in the rate-distortion sense.

3 Warm up: Beta Duals of Finite Fourier Frames

Let $\mathscr{H} := \mathbb{C}^d$ be equipped with the Euclidean inner-product. For any $N \geq d$, the standard finite Fourier frame $\mathscr{F}_{N,d} := (\varphi_n)_0^{N-1}$ of size N is given in Cartesian coordinates by

$$\varphi_{n,k} := \frac{1}{\sqrt{d}} \mathrm{e}^{2\pi \mathrm{i} nk/N};\ n \in [N];\ k \in [d].$$

For simplicity, we assume in this paper that N is a multiple of d. With this assumption, we set $p := d, N_j := N_* := N/d$ for all $j \in [d]$, and

$$\sigma_j(n) := jN_* + n;\ j \in [d];\ n \in [N_*].$$

Also we set $\beta_j = \beta$ for all $j \in [p]$, where β is a real number greater than 1 to be determined later. Then the beta condensation of $\mathscr{F}_{N,d}$ is computed explicitly to be

$$\zeta_{j,k} = \sum_{n \in [N_*]} \beta^{-n} \varphi_{\sigma_j(n),k} = \frac{1}{\sqrt{d}} w_k \mathrm{e}^{2\pi \mathrm{i} jk/d};\ j \in [d];\ k \in [d],$$

where

$$w_k := \sum_{n \in [N_*]} \left(\beta^{-1} \mathrm{e}^{2\pi \mathrm{i} k/N}\right)^n;\ k \in [d].$$

This formula shows that the beta condensation of a finite Fourier frame (with the parameters we have used) is actually a weighted discrete Fourier system (which is a basis if and only if all w_k are nonzero). It is now straightforward to compute the frame bounds. Indeed $\langle x, \zeta_j \rangle$ can be seen as the jth Discrete Fourier Transform (DFT) coefficient of $(x_k \overline{w}_k)_0^{d-1}$ so that (either by Parseval's identity or by explicit calculation) we have

$$\sum_{j \in [d]} |\langle x, \zeta_j \rangle|^2 = \sum_{k \in [d]} |x_k|^2 |w_k|^2.$$

Note that for any complex number $|z| < 1$ and any $m \geq 1$, we have

$$|1 + z + \cdots + z^{m-1}| = \left| \frac{1 - z^m}{1 - z} \right| \geq \frac{1 - |z|}{1 + |z|} \tag{12}$$

so that

$$\min_{k \in [d]} |w_k| \geq \frac{1 - \beta^{-1}}{1 + \beta^{-1}} =: C_\beta. \tag{13}$$

Hence the lower frame bound A_ζ of $(\zeta_j)_0^{d-1}$ satisfies

$$A_\zeta \geq C_\beta^2. \tag{14}$$

In light of the discussion of the previous section, we can now proceed with the optimization of system parameters. For all $x \in \mathbb{C}^d$ such that $\|x\|_2 \leq 1$, we have $\|y\|_\infty \leq 1$ so that for any $L \geq 4$ we can employ a quantization alphabet $\mathscr{A} \subset \mathbb{C}$ with at most L elements, guaranteeing $\|u\|_\infty \leq \sqrt{2}\delta$, where β and δ must satisfy the condition $\beta + 1/\delta \leq \lfloor \sqrt{L} \rfloor$. For any such β and δ, it follows from (11) that

$$\mathscr{D}_{\mathrm{a}}(\mathscr{F}_{N,d}, L) \leq \sqrt{2d} C_\beta^{-1} \delta \beta^{-\frac{N}{d}+1}. \tag{15}$$

In order to choose the special values of δ and β, we employ the following elementary lemma whose proof we leave as an exercise (for a nearly identical version, see [6, Lemma 3.2]):

Lemma 1 *For any $K \geq 2$ and $\alpha \geq 1$, let $\beta := K(\alpha+1)/(\alpha+2)$ and $\delta := (\alpha+2)/K$. Then $\beta \geq 4/3$, $\beta + 1/\delta = K$, and*

$$\delta \beta^{-\alpha+1} < \mathrm{e}(\alpha+1) K^{-\alpha}. \tag{16}$$

Furthermore, C_β as defined by (13) satisfies $C_\beta^{-1} \leq 7$.

We use this lemma for $K := \lfloor \sqrt{L} \rfloor$ and $\alpha := N/d$. Injecting the resulting bound (16) and the bound $C_\beta^{-1} \leq 7$ in (15), we arrive at the following near-optimal result:

Theorem 1 *Suppose N is a multiple of d. Then for any number of quantization levels $L \geq 4$, the analysis distortion of the finite Fourier frame of N elements in $\mathbb{C}^d$ satisfies*

$$\mathscr{D}_{\mathrm{a}}(\mathscr{F}_{N,d}, L) < 7\mathrm{e}\sqrt{2d} \left(\frac{N}{d} + 1 \right) \lfloor \sqrt{L} \rfloor^{-N/d}.$$

Remark 1 The above theorem is actually still valid for $L \leq 3$ but it does not offer a useful bound since then we have $\lfloor \sqrt{L} \rfloor = 1$. For $L = 3$ we may instead invoke the triangular alphabet $\mathscr{A}$ and the associated quantization rule described in the Appendix for which we may set $\beta := \left(\frac{4}{3}\right)^{1-\varepsilon}$ for any $\varepsilon \in (0,1)$. It can then be checked that

$$\mathscr{D}_{\mathrm{a}}(\mathscr{F}_{N,d}, 3) \lesssim_\varepsilon \sqrt{d} \left(\frac{4}{3} \right)^{-(1-\varepsilon)N/d}.$$

We omit the details. This is the only upper bound we know for $L = 3$ that is exponentially small in N/d; however, it does not match the entropic lower bound of $3^{-N/2d}$.

4 Generalization: Unitarily Generated Frames

A general method of constructing uniform tight frames based on frame paths was introduced in [4] in connection with analyzing the performance of $\Sigma\Delta$ quantization for finite frames. In this section, first we will slightly extend this frame construction method to include a larger class of frames, and then bound the analysis distortion of these frames using distributed beta encoding.

4.1 Unitary frame paths

Let $\mathscr{H} := \mathbb{C}^d$ be equipped with the Euclidean inner-product and Ω be a $d \times d$ Hermitian matrix. Consider the 1-parameter group of unitary operators on $\mathscr{H}$ given by

$$U_t := \mathrm{e}^{2\pi \mathrm{i}\Omega t};\ t \in \mathbb{R},$$

and for any $\varphi_0 \in \mathbb{C}^d$ of unit norm, let

$$\varphi_n := U_{\frac{n}{N}}\varphi_0;\ n = 0, \ldots, N-1.$$

The curve $\{t \mapsto U_t\varphi_0 : t \in [0,1]\}$ is called a *unitary frame path* if $\Phi := (\varphi_n)_0^{N-1}$ yields a frame for infinitely many $N \geq d$. We also say that Φ is *unitarily generated*.

Assume Ω has d distinct integer eigenvalues $\lambda_0, \ldots, \lambda_{d-1}$ which are also distinct modulo N. Let us denote the corresponding normalized eigenvectors of Ω by $v_0, \ldots, v_{d-1}$. This collection gives us an orthogonal basis of $\mathscr{H}$. Now note that

$$\langle v_k, \varphi_n\rangle = \langle \mathrm{e}^{-2\pi \mathrm{i}\Omega n/N} v_k, \varphi_0\rangle = \mathrm{e}^{-2\pi \mathrm{i}\lambda_k n/N}\langle v_k, \varphi_0\rangle;\ n \in [N],\ k \in [d],$$

so that

$$\begin{aligned}\sum_{n\in[N]} |\langle x, \varphi_n\rangle|^2 &= \sum_{n\in[N]} \left|\sum_{k\in[d]} \langle x, v_k\rangle\langle v_k, \varphi_n\rangle\right|^2 \\ &= \sum_{k\in[d]}\sum_{l\in[d]} \langle x, v_k\rangle\langle v_l, x\rangle\langle v_k, \varphi_0\rangle\langle \varphi_0, v_l\rangle \sum_{n\in[N]} \mathrm{e}^{2\pi \mathrm{i}(\lambda_l - \lambda_k)n/N} \\ &= N\sum_{k\in[d]} |\langle x, v_k\rangle|^2 |\langle \varphi_0, v_k\rangle|^2,\end{aligned}$$

where in the last equality we have used the assumption that $\lambda_0, \ldots, \lambda_{d-1}$ are distinct modulo N.

With this identity, it now follows that

$$N\left(\min_{k\in[d]}|\langle\varphi_0,v_k\rangle|^2\right)\|x\|^2\le\sum_{n\in[N]}|\langle x,\varphi_n\rangle|^2\le N\left(\max_{k\in[d]}|\langle\varphi_0,v_k\rangle|^2\right)\|x\|^2. \quad (17)$$

Hence we see that $(\varphi_n)_0^{N-1}$ is a frame if and only if $\langle\varphi_0,v_k\rangle\neq 0$ for all $k\in[d]$. We also see (as in [4]) that $(\varphi_n)_0^{N-1}$ is a unit-norm tight frame if and only $|\langle\varphi_0,v_k\rangle|=1/\sqrt{d}$ for all k. Note that the frame condition is generic, i.e. the set of φ_0 which yield a frame is an open dense subset of $\mathscr{H}$. In contrast, the condition for tightness of the frame is quite strict, corresponding to a nowhere dense set of φ_0.

Remark 2 The above argument continues to hold under the weaker assumption that all pairwise differences $\lambda_l-\lambda_k$ are integers and are nonzero modulo N if $l\neq k$. In other words, it is possible to shift all the eigenvalues by a common real value without changing the frame property. Note that (U_t) is 1-periodic in t if and only if all the eigenvalues are integers in which case the frame path is a closed curve.

Remark 3 Note that the finite Fourier frame of the previous section corresponds to the case when Ω is the diagonal matrix with the diagonal entries $0,\dots,d-1$ and $\varphi_0=(1,\dots,1)/\sqrt{d}$.

More generally, we may pick any $J\subset[N]$ of cardinality d to form the diagonal entries $\lambda_0,\dots,\lambda_{d-1}$ (in increasing order) of a diagonal matrix Ω. The resulting tight frame can be characterized equivalently as the restriction of the finite Fourier basis of $\mathbb{C}^N$ to the space of *timelimited* vectors $\mathscr{H}:=\{x\in\mathbb{C}^N:\operatorname{supp}(x)\subset J\}$.

By duality we can also consider the space of discrete *bandlimited* vectors $\mathscr{B}_J:=\{x\in L^2(\mathbb{Z}_N):\operatorname{supp}(\widehat{x})\subset J\}$. For any φ_0 such that $\operatorname{supp}(\widehat{\varphi_0})=J$, the system $(\varphi_n)_{n\in\mathbb{Z}_N}$ defined via translating φ_0, i.e. by setting $\varphi_{n,k}:=\varphi_0(k-n)$, $k,n\in\mathbb{Z}_N$, constitute a unitarily generated frame for $\mathscr{B}_J$.

Unitarily generated frames in $\mathbb{R}^d$. If the Hermitian Ω is such that all of its entries are purely imaginary, i.e., $\mathrm{i}\Omega$ is a real, skew-symmetric matrix, then (U_t) reduces to a group of real, orthogonal matrices. Then $(\varphi_n)_0^{N-1}$ is a unitarily generated frame in $\mathbb{R}^d$ provided $\varphi_0\in\mathbb{R}^d$ and $\langle\varphi_0,v_k\rangle\neq 0$ for all $k\in[d]$. Note that the eigenvectors (v_k) would still need to be considered as vectors in $\mathbb{C}^d$.

The simplest nontrivial example is in $\mathbb{R}^2$. Two examples are worth mentioning: First, we may consider $\Omega:=B:=\left(\begin{smallmatrix}0&\mathrm{i}\\-\mathrm{i}&0\end{smallmatrix}\right)$. Here the eigenvalues 1 and -1 of Ω are distinct modulo N if only if $N\ge 3$. We may also consider $\Omega:=B/2=\left(\begin{smallmatrix}0&\mathrm{i}/2\\-\mathrm{i}/2&0\end{smallmatrix}\right)$ for which the condition in Remark 2 is satisfied for all $N\ge 2$. This frame is actually the *semicircle frame* in $\mathbb{R}^2$ (see [4, 6]).

The *harmonic frames* in $\mathbb{R}^d$ for $d=2m$ are obtained by setting Ω to be the block diagonal matrix with the blocks $B,2B,\dots,mB$, and eigenvalues $\{\pm1,\dots,\pm m\}$. Again we require $N\ge d+1$ as a frame condition. For $d=2m+1$, a 1×1 "0 block" is added resulting in the eigenvalues $\{0,\pm1,\dots,\pm k\}$. See [4] for additional information.

4.2 Beta Duals of Unitarily Generated Frames

Let $(\varphi_n)_0^{N-1}$ be a unitarily generated frame in $\mathbb{F}^d$ as described in Section 4.1 where Ω, $(\lambda_k)_0^{d-1}$, and $(v_k)_0^{d-1}$ have the same meaning as before. Let $d \le p \le N$. For simplicity, we assume that N is a multiple of p, and set $N_* := N_j := N/p$ for all $j \in [p]$. As in Section 3, we set $\sigma_j(n) := jN_* + n$, $n \in [N_*]$, and $\beta_j = \beta > 1$ for all $j \in [p]$. Then the beta condensation of the frame $(\varphi_n)_0^{N-1}$ is given by

$$\zeta_j := \sum_{n\in[N_*]} \beta^{-n}\varphi_{\sigma_j(n)} = \sum_{k\in[d]} w_k \mathrm{e}^{2\pi \mathrm{i} j\lambda_k/p}\langle \varphi_0, v_k\rangle v_k;\ j \in [p],$$

where

$$w_k := \sum_{n\in N_*} \left(\beta^{-1}\mathrm{e}^{2\pi \mathrm{i}\lambda_k/N}\right)^n;\ k \in [d].$$

Assuming the stronger hypothesis that $\lambda_0, \ldots, \lambda_{d-1}$ are distinct modulo p, or more generally, that

$$\lambda_l - \lambda_k \text{ are integers and nonzero modulo } p \text{ if } l \ne k, \tag{18}$$

we have

$$\begin{aligned}\sum_{j\in[p]} |\langle x, \zeta_j\rangle|^2 &= \sum_{j\in[p]} \left| \sum_{k\in[d]} w_k \mathrm{e}^{2\pi \mathrm{i} j\lambda_k/p}\langle \varphi_0, v_k\rangle\langle x, v_k\rangle \right|^2 \\ &= p \sum_{k\in[d]} |\langle x, v_k\rangle|^2 |\langle \varphi_0, v_k\rangle|^2 |w_k|^2.\end{aligned}$$

Using (12), we have $|w_k| \ge C_\beta = (1-\beta^{-1})/(1+\beta^{-1})$ as before, so we find that the lower frame bound A_ζ of $(\zeta_j)_0^{p-1}$ satisfies

$$A_\zeta \ge pC_\beta^2 \left(\min_{k\in[d]} |\langle \varphi_0, v_k\rangle|^2 \right).$$

The rest of the discussion where we bound the analysis distortion of Φ is the same as before. Namely, we invoke (11) and follow the same procedure as in the case of finite Fourier frames of Section 3, starting from (14). In addition, for $\mathbb{F} = \mathbb{R}$ we may employ a quantizer in $\mathbb{R}$ for all $L \ge 2$ where we can set $K = L$ and the state vector satisfies $\|u\|_\infty \le \delta$. The result is summarized in the following theorem:

Theorem 2 *Suppose N is a multiple of p where $p \ge d$, and Φ is a unitarily generated frame in $\mathbb{F}^d$ such that (18) holds. Then we have*

$$\mathscr{D}_{\mathrm{a}}(\Phi, L) < 7\mathrm{e}\left(\frac{N}{p}+1\right) c(\varphi_0) \cdot \begin{cases} \sqrt{2}\lfloor\sqrt{L}\rfloor^{-N/p}, & \text{if } \mathbb{F}=\mathbb{C} \text{ and } L \geq 4, \\ L^{-N/p}, & \text{if } \mathbb{F}=\mathbb{R} \text{ and } L \geq 2, \end{cases}$$

where

$$c(\varphi_0) := \left(\min_{1\leq k\leq d} |\langle \varphi_0, v_k\rangle|\right)^{-1}.$$

Of course, a bound for the case $L = 3$ can also be given as in the previous section.

5 An Infinite-Dimensional Case: Bandlimited Functions on $\mathbb{R}$

Discussing quantized frame representations in infinite dimensional Hilbert spaces requires special care due to the fact that the coefficient sequence $(q_\alpha)_{\alpha\in I}$ is not in $\ell^2(I)$ (and therefore the reconstruction is not guaranteed to be of finite norm) unless $q_\alpha = 0$ for all but finitely many α. Of course, for this to happen, 0 would need to be a permissible quantization level in $\mathscr{A}$ in the first place. Then the problem becomes similar to a finite dimensional one with one main difference: the finite dimensional subspace from which a quantized approximation is sought would need to be either specified *a priori*, or determined *a posteriori* by means of the quantization algorithm itself.

Another approach is to relax the Hilbertian frame setting and consider frame-like representations in other suitable normed spaces and with a different sense of convergence, as well as the possibility of approximation by quantized representations from outside these spaces. Indeed this is the sense in which the classical oversampled quantization problem of bandlimited functions on $\mathbb{R}$ has been studied mathematically [7, 8]. A general (and highly nontrivial) theory for quantization for frames in Banach spaces was also developed in [5]. In this short section we will only be concerned with the case of uniform sampling of bandlimited functions where it will be possible for us to work from scratch.

The analysis distortion of sampling. Let $\mathscr{B}_\Omega$ be the space of bounded continuous functions x on $\mathbb{R}$ for which the (distributional) Fourier transform $\widehat{x}$ is supported in $[-\Omega, \Omega]$. This space contains the classical Paley-Wiener space PW_Ω which comes with an additional square-integrability constraint. (PW_Ω is therefore a Hilbert space with respect to the standard inner-product on $L^2(\mathbb{R})$.) We equip $\mathscr{B}_\Omega$ with the L^∞-norm which is more suitable for quantization. The celebrated Shannon-Nyquist sampling theorem (in the context of $\mathscr{B}_\Omega$) says that any $x \in \mathscr{B}_\Omega$ can be recovered perfectly from its samples $(x(k\tau))_{k\in\mathbb{Z}}$ via a pointwise absolutely convergent expansion

$$x(t) = \tau \sum_{k\in\mathbb{Z}} x(k\tau)\psi(t-k\tau), \tag{19}$$

where $\tau < \tau_{\text{crit}} := \frac{1}{2\Omega}$ and ψ is any function of rapid decay on $\mathbb{R}$ such that

$$\widehat{\psi}(\xi) = \begin{cases} 1, & |\xi| \le \Omega, \\ 0, & |\xi| \ge \frac{1}{2\tau}. \end{cases} \tag{20}$$

We will say that such a ψ is (Ω, τ)-*admissible*. The value $\rho := 1/\tau$ is called the sampling rate, and $\rho_{\text{crit}} := 1/\tau_{\text{crit}} = 2\Omega$ is called the critical (or Nyquist) sampling rate. The *oversampling ratio* given by

$$\lambda := \frac{\rho}{\rho_{\text{crit}}} = \frac{\tau_{\text{crit}}}{\tau} \tag{21}$$

corresponds to the "redundancy" of the sampling operator $\Phi_\tau : \mathscr{B}_\Omega \to \ell^\infty(\mathbb{Z})$ where

$$(\Phi_\tau x)_k := x(k\tau), \ k \in \mathbb{Z}.$$

Let us say that a collection of bounded continuous functions $\Psi := (\psi_k)_{k\in\mathbb{Z}}$ on $\mathbb{R}$ is *quantization admissible* if $\sum c_k \psi_k$ converges (pointwise absolutely) to a bounded function whenever $c \in \ell^\infty(\mathbb{Z})$. Let us also say that Ψ is *dual* to Φ_τ on $\mathscr{B}_\Omega$ if, in addition, we have

$$x = \sum_{k\in\mathbb{Z}} (\Phi_\tau x)_k \psi_k = \sum_{k\in\mathbb{Z}} x(k\tau)\psi_k \ \text{ for all } x \in \mathscr{B}_\Omega, \tag{22}$$

where again the convergence is understood to be pointwise and absolute. This equation generalizes the concept of frame and the classical sampling formula (19) where the ψ_k are $\tau\mathbb{Z}$-translations of a fixed function. The analog of analysis distortion associated to Φ_τ on $\mathscr{B}_\Omega$ for L levels of quantization, now denoted by $\mathscr{D}_{\text{a}}(\Phi_\tau|\mathscr{B}_\Omega, L)$ is then naturally defined to be

$$\inf\left\{ \sup_{\|x\|_\infty \le 1} \inf_{q\in\mathscr{A}^{\mathbb{Z}}} \left\| x - \sum_{k\in\mathbb{Z}} q_k \psi_k \right\|_\infty : \Psi \text{ is dual to } \Phi_\tau \text{ on } \mathscr{B}_\Omega \text{ and } |\mathscr{A}| = L \right\}.$$

Beta dual of the sampling operator. Note that in the context of PW_Ω this sampling operator can be realized in terms of the unitarily generated frame consisting of the $\tau\mathbb{Z}$-translations of a fixed sinc kernel. The following construction mimicks the beta dual machinery of Sections 2 and 4. Since our setup is not Hilbertian, we will take a direct approach in our construction.

Given $\tau < \tau_{\text{crit}}$, let $\lambda_* := \lceil\lambda\rceil - 1 = \lceil\tau_{\text{crit}}/\tau\rceil - 1$ and $\tau_* := \lambda_*\tau$. Note that $\lambda > \lambda_* \ge 1$ and $\tau \le \tau_* < \tau_{\text{crit}}$. For any given $\beta > 1$, consider the operators

$$Tf := \sum_{n\in[\lambda_*]} \beta^{-n} f(\cdot + n\tau),$$

$$Sf := f - \beta^{-1} f(\cdot + \tau), \quad \text{and}$$
$$Rf := \sum_{n=0}^{\infty} \beta^{-\lambda_* n} f(\cdot + n\tau_*)$$

on $L^\infty(\mathbb{R})$ where they are also clearly bounded. All three operators represent convolution operators with distributional kernels and it is evident after inspecting their Fourier multipliers that RS inverts T. Avoiding distributions, we can check this fact directly. Indeed, for any $f \in L^\infty(\mathbb{R})$, we have

$$RSTf = R(f - \beta^{-\lambda_*} f(\cdot + \tau_*)) = f.$$

All three operators enjoy a crucial property which is stronger than their continuity on $L^\infty(\mathbb{R})$: Whenever a function series $\sum f_k$ converges pointwise absolutely (but not necessarily uniformly) to a bounded function, we have

$$R \sum f_k = \sum Rf_k \text{ (similarly for } S \text{ and } T), \tag{23}$$

where the latter series also converges pointwise absolutely. To see this, simply note that the iterated series

$$\sum_{n=0}^{\infty} \beta^{-\lambda_* n} \sum_k |f_k(t + n\tau_*)|$$

is convergent for all t; hence, it is justified to change the order of summation that is required to prove (23).

Let ψ_* be (Ω, τ_*)-admissible. Given any $x \in \mathscr{B}_\Omega$, it is clear that $Tx \in \mathscr{B}_\Omega$ as well, and we can apply Shannon's sampling theorem to Tx with the reconstruction filter ψ_* on the sampling grid $\tau_* \mathbb{Z}$ to obtain

$$Tx = \tau_* \sum_{j \in \mathbb{Z}} \left(\sum_{n \in [\lambda_*]} \beta^{-n} x(j\tau_* + n\tau) \right) \psi_*(\cdot - j\tau_*).$$

We apply RS to both sides of this equation. Using (23) and noting translation invariance of RS, we obtain

$$x = \sum_{j \in \mathbb{Z}} \sum_{n \in [\lambda_*]} x\left((\lambda_* j + n)\tau\right) \tau_* \beta^{-n} (RS\psi_*)(\cdot - j\tau_*). \tag{24}$$

We now set $\psi := RS\psi_*$, and define $\Psi := (\psi_k)_{k \in \mathbb{Z}}$ by

$$\psi_{\lambda_* j + n} := \tau_* \beta^{-n} \psi(\cdot - j\tau_*);\ j \in \mathbb{Z},\ n \in [\lambda_*].$$

Then (24) says nothing but that Ψ is dual to Φ_τ on $\mathscr{B}_\Omega$. It is easy to see that ψ also has rapid decay so that Ψ is a quantization-admissible dual.

For the quantization process, we employ the same distributed beta encoding approach as before, and this time, set $y_k := x(k\tau)$, $\sigma_j(n) := \lambda_* j + n$,

$$y_{\sigma_j(n)} - q_{\sigma_j(n)} = u_{\sigma_j(n)} - \beta u_{\sigma_j(n-1)};\ j \in \mathbb{Z},\ n \in [\lambda_*],$$

with $\sigma_j(-1) := 0$ so that

$$x - \sum_{j\in\mathbb{Z}} \sum_{n\in[\lambda_*]} q_{\lambda_* j+n} \tau_* \beta^{-n} \psi(t - j\tau_*) = \beta^{-\lambda_*+1} \sum_{j\in\mathbb{Z}} u_{\sigma_j(\lambda_*-1)} \tau_* \psi(t - j\tau_*),$$

and therefore

$$\left\| x - \sum_{k\in\mathbb{Z}} q_k \psi_k \right\|_\infty \le \beta^{-\lambda_*+1} \|u\|_\infty C(\psi), \tag{25}$$

where

$$\begin{aligned} C(\psi) &:= \left\| \sum_{j\in\mathbb{Z}} \tau_* |\psi(\cdot - j\tau_*)| \right\|_\infty \\ &\le \left\| \sum_{j\in\mathbb{Z}} \tau_* RS |\psi_*(\cdot - j\tau_*)| \right\|_\infty \\ &\le \|RS\|_{\infty\to\infty} \left\| \sum_{j\in\mathbb{Z}} \tau_* |\psi_*(\cdot - j\tau_*)| \right\|_\infty \\ &\le \frac{\beta+1}{\beta-1} C(\psi_*). \end{aligned} \tag{26}$$

Note that τ_* is near τ_{crit} in a uniform manner; for example, it is easy to show that we have $\tau_* \in [\tau_{\text{crit}}/2, \tau_{\text{crit}})$. This allows us to choose ψ_* purely as a function of Ω via $\psi_*(t) := \Omega \psi_{*,0}(\Omega t)$ for a fixed $\psi_{*,0}$. Consequently, we may replace $C(\psi_*)$ by a universal constant independent of Ω.

Assuming we are only concerned with real-valued functions and employing a quantization alphabet of L levels requiring $\beta + 1/\delta \le L$, we may set β and δ in (25) as indicated by Lemma 1. The end result now reads

$$\mathscr{D}_{\text{a}}(\Phi_\tau | \mathscr{B}_\Omega, L) \lesssim \lambda L^{-\lceil\lambda\rceil+1}.$$

The modifications for complex-valued bandlimited functions would be the same as before.

6 Concluding Remarks

We have not touched upon many classical frames that are popular in theory and practice, such as non-harmonic Fourier frames, frames of irregular sampling and interleaved sampling, Gabor frames, and filter bank frames. Gabor frames are generated by two unitary transformations, modulation and translation, which do not commute. Sub-optimal results can be obtained in a straightforward manner by focusing on only one of the generators and applying the basic beta dual machinery. However, additional work (e.g., on the noise transfer operator) may be necessary in order to exploit all of the redundancy present in a Gabor frame. Similar comments are applicable for filter bank frames as well.

Appendix: Greedy Quantizer for Complex Measurements

In this section we will provide a generalization of the complex-valued $\Sigma\Delta$ quantization algorithm given in [3, Proposition 3.1] and the greedy noise-shaping quantization algorithm given in [6, Theorem 2.1]. The result, which is applicable to both real and complex quantization alphabets, offers nontrivial improvements in the complex case, thanks to the use of general semi-norms to measure closeness.

Lemma 2 *Let $\mathscr{A}$ be a quantization alphabet in $\mathbb{C}$, B_* be the closed unit ball of a semi-norm $|\cdot|_*$ on $\mathbb{C}$ treated as a vector space over $\mathbb{R}$, and $H := (H_{n,m})_{n,m\in[N]}$ be an $N \times N$ real-valued lower-triangular matrix with unit diagonal. Suppose there exist positive real numbers μ, δ, γ such that*

$$\delta B_* + \mathscr{A} \supset \gamma B_* \tag{27}$$

and

$$\mu + \delta \max_{n\in[N]} \sum_{m<n} |H_{n,m}| \leq \gamma. \tag{28}$$

Then for any $y \in \mathbb{C}^N$ such that $|y_n|_* \leq \mu$ for all $n \in [N]$, there exist $q \in \mathscr{A}^N$ and $u \in \mathbb{C}^N$ such that

$$y - q = Hu$$

where $|u_n|_* \leq \delta$ for all $n \in [N]$.

Proof The proof of this result is yet another adaptation of a well-known induction argument. By our assumption on H, we are seeking to satisfy the equations

$$u_n = \left(y_n - \sum_{m<n} H_{n,m} u_m \right) - q_n \tag{29}$$

for all $n \in [N]$.

Since $|y_0|_* \leq \mu \leq \gamma$, (27) implies that there exist $q_0 \in \mathscr{A}$ and $u_0 \in \delta B_*$ such that $u_0 + q_0 = y_0$. Hence (29) is satisfied for $n = 0$ and $|u_0|_* \leq \delta$.

For the induction step, assume that $|u_m|_* \leq \delta$ for all $m < n$, and let

$$w_n := y_n - \sum_{m<n} H_{n,m} u_m.$$

Using sub-additivity and homogeneity of $|\cdot|_*$ followed by the condition given in (28), we get

$$|w_n|_* \leq \mu + \delta \sum_{m<n} |H_{n,m}| \leq \gamma;$$

hence, because of (27) again, there exist $q_n \in \mathscr{A}$ and $u_n \in \delta B_*$ such that $u_n + q_n = w_n$, i.e. (29) holds. □

Special known cases. There are certainly many ways to choose $\mathscr{A}$ and $|\cdot|_*$. We first note two important special cases of practical importance. Here L denotes $|\mathscr{A}|$.

($\mathbb{R}$) *Real arithmetic progression*
This quantizer uses $\mathscr{A} := \mathscr{A}_{L,\delta} := \{(-L + 2l - 1)\delta : 1 \leq l \leq L\} \subset \mathbb{R}$, i.e. the origin-symmetric arithmetic progression of length L and spacing 2δ along with $|z|_* := |\Re(z)|$. Then B_* is the infinite vertical strip $\{z : |\Re(z)| \leq 1\}$ and (27) holds for $\gamma := L\delta$. Using the algorithm in Lemma 2, $y \in \mathbb{R}^N$ results in $u \in \mathbb{R}^N$, and $\|y\|_\infty \leq \mu$ implies $\|u\|_\infty \leq \delta$ so that the setup becomes identical to that of [6].

($\mathbb{C}$) *Complex square lattice quantizer*
This quantizer assumes $L = K^2$ for some positive integer K and sets $\mathscr{A} := \mathscr{A}_{K,\delta} + i\mathscr{A}_{K,\delta} \subset \mathbb{C}$ along with $|z|_* := \max(|\Re(z)|, |\Im(z)|)$. B_* can be identified with $[-1, 1]^2$ (as a subset $\mathbb{R}^2$) so that (27) is valid for $\gamma := K\delta$. Since $|z|_* \leq |z| \leq \sqrt{2}|z|_*$ for any $z \in \mathbb{C}$, $\|y\|_\infty \leq \mu$ implies $|y_n|_* \leq \mu$ for all n and Lemma 2 then yields $\|u\|_\infty \leq \sqrt{2}\delta$.
When K is even, the resulting $\mathscr{A}$ has no real points and it may be desirable to require that $y \in \mathbb{R}^N$ always yields $q \in \mathbb{R}^N$. In this case, we may instead use the slightly larger alphabet $\mathscr{A} := \mathscr{A}_{K,\delta} + i\mathscr{A}_{K+1,\delta}$ for which $L = K(K + 1)$. This choice indeed corresponds to the one made in [3]. Another natural possibility in this case is to use the 1-norm in $\mathbb{R}^2$ coupled with the diamond lattice as shown in Figure 1 for $K = 2$.
Note that for the square (or diamond) lattice quantizer of K^2 levels, using the Euclidean norm $|\cdot|$ on $\mathbb{C}$ would be sub-optimal. Indeed, the largest value of γ that can be used in (27) is $\gamma = \frac{K}{\sqrt{2}}\delta$.

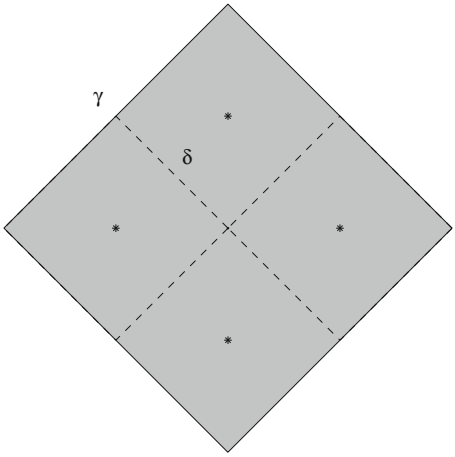


Fig. 1 Lattice covering for the 1-norm in $\mathbb{R}^2$ where $L = 4$ and $\gamma = 2\delta$.

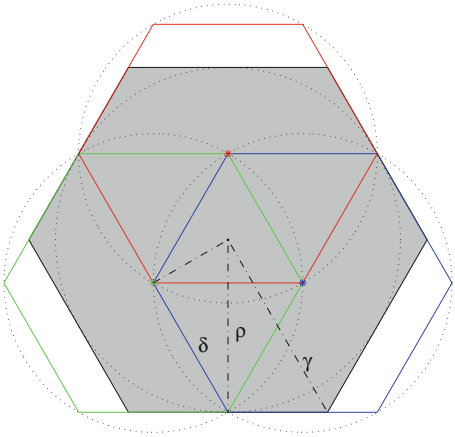


Fig. 2 Three identical hexagons at scale δ covering a larger hexagon at scale $\gamma = \frac{4}{3}\delta$ compared to three identical circular discs at scale δ covering a larger circular disc at scale $\rho = \frac{2}{\sqrt{3}}\delta$.

Hexagonal norm for a tri-level complex alphabet. It is natural to ask if a complex quantization alphabet $\mathscr{A}$ with fewer than 4 levels can be used in connection with the noise-shaping quantization algorithm of Lemma 2. For $L = 3$, we may set $\mathscr{A}$ to be the vertices of an equilateral triangle in $\mathbb{C}$ centered at the origin. If the Euclidean norm is used, then it is not difficult to prove that the largest value of γ that can be used in (27) is $\gamma = \frac{2}{\sqrt{3}}\delta$ (see Figure 2 for a demonstration of this covering). In this case, $\|y\|_\infty \leq \mu$ yields $\|u\|_\infty \leq \delta$.

An alternative we have found useful is to employ the norm $|\cdot|_*$ induced by a regular hexagonal body whose sides are aligned with the sides of the triangle. Then, as shown in Figure 2, we can attain $\gamma = \frac{4}{3}\delta$. By choosing the scale of the hexagonal body suitably, we can ensure $|z|_* \leq |z| \leq \frac{2}{\sqrt{3}}|z|_*$ so that $\|y\|_\infty \leq \mu$ implies $|y_n|_* \leq \mu$ for all n, and therefore Lemma 2 yields $\|u\|_\infty \leq \frac{2}{\sqrt{3}}\delta$. Despite the increase in the bound for $\|u\|_\infty$, there is a sizable gain in the "expansion factor"

γ/δ from $\frac{2}{\sqrt{3}}$ to $\frac{4}{3}$. This gain is crucial for beta encoding because any β up to this expansion factor is admissible for stability via Lemma 2 provided $\mathscr{A}$, γ, and δ are suitably scaled to meet (27) and (28) simultaneously.

References

1. R. Adler, T. Nowicki, G. Świrszcz, C. Tresser, Convex dynamics with constant input. Ergodic Theory Dyn. Syst. **30**, 957–972 (2010)
2. J.J. Benedetto, A.M. Powell, Ö. Yılmaz, Sigma-delta quantization and finite frames. IEEE Trans. Inf. Theory **52**(5), 1990–2005 (2006)
3. J.J. Benedetto, O. Oktay, A. Tangboondouangjit, Complex sigma-delta quantization algorithms for finite frames, in *Radon Transforms, Geometry, and Wavelets*. Contemporary Mathematics, vol. 464 (American Mathematical Society, Providence, RI, 2008), pp. 27–49
4. B.G. Bodmann, V.I. Paulsen, Frame paths and error bounds for sigma-delta quantization. Appl. Comput. Harmon. Anal. **22**(2), 176–197 (2007)
5. P.G. Casazza, S.J. Dilworth, E. Odell, Th. Schlumprecht, A. Zsk, Coefficient quantization for frames in Banach spaces. J. Math. Anal. Appl. **348**(1), 66–86 (2008)
6. E. Chou, C.S. Güntürk, Distributed noise-shaping quantization: I. Beta duals of finite frames and near-optimal quantization of random measurements. Constr. Approx. **44**(1), 1–22 (2016)
7. I. Daubechies, R. DeVore, Approximating a bandlimited function using very coarsely quantized data: a family of stable sigma-delta modulators of arbitrary order. Ann. Math. **158**(2), 679–710 (2003)
8. C.S. Güntürk, Mathematics of analog-to-digital conversion. Comm. Pure Appl. Math. **65**(12), 1671–1696 (2012)
9. A.M. Powell, R. Saab, Ö. Yılmaz, Quantization and finite frames, in *Finite Frames: Theory and Applications*, ed. by G.P. Casazza, G. Kutyniok. Applied and Numerical Harmonic Analysis (Birkhäuser/Springer, New York, 2013), pp. 267–302
10. R. Schreier, G.C. Temes, *Understanding Delta-Sigma Data Converters* (Wiley/IEEE Press, Chichester, 2004)
11. Y. Wang, Sigma-delta quantization errors and the traveling salesman problem. Adv. Comput. Math. **28**(2), 101–118 (2008)

Consistent Reconstruction: Error Moments and Sampling Distributions

Chang-Hsin Lee, Alexander M. Powell, and J. Tyler Whitehouse

Abstract Consistent reconstruction is a method for estimating a signal from a collection of linear measurements that have been corrupted by uniform noise. We prove upper bounds on general error moments for consistent reconstruction, and we establish general admissibility conditions on the sampling distributions used for consistent reconstruction. This extends previous work in Powell and Whitehouse (Found Comput Math 16:395–423, 2016) that addressed mean squared error in the setting of unit-norm sampling distributions.

Keywords Consistent reconstruction • Estimation with uniform noise

1 Introduction

Consistent reconstruction is a method for estimating a signal $x \in \mathbb{R}^d$ from a collection of linear measurements that have been corrupted by uniform noise or, more generally, bounded noise. Estimation with uniform noise arises naturally in quantization problems in signal processing, especially in connection with dithering and the uniform noise model [7, 11]. Consistent reconstruction has been used as a signal recovery method for memoryless scalar quantization [1, 2, 4, 11, 13], Sigma-Delta quantization [12], and compressed sensing [5, 6, 9]. See [10] for background and motivation on consistent reconstruction and estimation with uniform noise.

Let $x \in \mathbb{R}^d$ be an unknown signal and let $\{\varphi_n\}_{n=1}^N \subset \mathbb{R}^d$ be a given spanning set for $\mathbb{R}^d$ that is used to make linear measurements $\langle x, \varphi_n \rangle$ of x. We consider the problem of recovering an estimate for x from the noisy measurements

$$q_n = \langle x, \varphi_n \rangle + \epsilon_n, \quad 1 \leq n \leq N, \tag{1}$$

C.-H. Lee (✉) • A.M. Powell
Department of Mathematics, Vanderbilt University, Nashville, TN 37240, USA
e-mail: chang-hsin.lee@vanderbilt.edu; alexander.m.powell@vanderbilt.edu

J.T. Whitehouse
Quantitative Scientific Solutions, LLC, Arlington, VA, USA
e-mail: tyler.whitehouse@qs-2.com

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4_9

where $\{\epsilon_n\}_{n=1}^N$ are independent uniform random variables on $[-\delta, \delta]$. For the setting of this chapter, the collection $\{\varphi_n\}_{n=1}^N$ is known but randomly generated, the noise level $\delta > 0$ is fixed and known, whereas x and the noise $\{\epsilon_n\}_{n=1}^N$ are both unknown. We focus on the situation when $\{\varphi_n\}_{n=1}^N$ are independent versions of a random vector $\varphi \in \mathbb{R}^d$ whose distribution we refer to as the sampling distribution.

Consistent reconstruction seeks an estimate $\widetilde{x}$ for the unknown signal x that is consistent with the knowledge that the noise is bounded in $[-\delta, \delta]$. Specifically, consistent reconstruction produces an estimate $\widetilde{x} \in \mathbb{R}^d$ for x by selecting any solution of the linear feasibility problem

$$|\langle \widetilde{x}, \varphi_n \rangle - q_n| \leq \delta, \quad 1 \leq n \leq N. \tag{2}$$

There are generally infinitely many solutions to this feasibility problem. In this chapter, we mainly focus on the worst case error associated to consistent reconstruction.

1.1 Worst case error

To describe the worst case error of consistent reconstruction, note that if $\widetilde{x}$ is any solution to (2), then the error $(\widetilde{x} - x)$ lies in each of the closed convex sets

$$E_n = \left\{ u \in \mathbb{R}^d : |\langle u, \varphi_n \rangle - \epsilon_n| \leq \delta \right\}. \tag{3}$$

The intersection of the sets E_n forms the following error polytope:

$$P_N = \bigcap_{n=1}^N E_n, \tag{4}$$

which is the set of all possible errors associated to consistent reconstruction (2). The worst case error W_N associated to consistent reconstruction is thus defined by

$$W_N = \max \{ \|u\| : u \in P_N \}, \tag{5}$$

where $\| \cdot \|$ denotes the Euclidean norm on $\mathbb{R}^d$.

1.2 Background

The main results in [10] proved error bounds for the expected worst case error squared $\mathbb{E}[(W_N)^2]$ of consistent reconstruction when the sampling vectors $\{\varphi_n\}_{n=1}^N$ are drawn at random from a suitable probability distribution on the unit sphere $\mathbb{S}^{d-1}$.

The work in [10] considered sampling vectors $\{\varphi_n\}_{n=1}^N \subset \mathbb{S}^{d-1}$ that are independently drawn instances of a unit-norm random vector φ that satisfies the following admissibility condition:

$$\exists\, \alpha \geq 1, \exists\, 0 < s \leq 1, \quad \forall\, 0 \leq t \leq 1, \quad \forall x \in \mathbb{S}^{d-1}, \qquad \Pr[\,|\langle x, \varphi\rangle| \leq t] \leq \alpha t^s. \tag{6}$$

See Section 5 of [10] for further discussion of the admissibility condition (6). For example, if φ is uniformly distributed on $\mathbb{S}^{d-1}$, then φ satisfies (6) with $s = 1$ and $\alpha = \frac{2\Gamma(\frac{d}{2})}{\sqrt{\pi}\,\Gamma(\frac{d-1}{2})}$. On the other hand, if φ has a point mass, then φ does not satisfy (6).

Suppose that $\{\varphi_n\}_{n=1}^N \subset \mathbb{S}^{d-1}$ are independently drawn at random according to a distribution that satisfies the admissibility condition (6). Theorem 5.5 and Corollary 5.6 in [10] prove that there exist absolute constants $c_1, c_2 > 0$ such that if

$$N \geq c_2 d \ln(32(2\alpha)^{1/s}),$$

then the expected worst case error squared for consistent reconstruction satisfies

$$\mathbb{E}[(W_N)^2] \leq \frac{c_1 \delta^2 d^2 (2\alpha)^{1/s} \ln^2(16(2\alpha)^{1/s})}{(N+1)(N+2)}.$$

Moreover, in the special case when $\{\varphi_N\}_{n=1}^N$ are drawn independently at random according to the uniform distribution on $\mathbb{S}^{d-1}$, Theorem 6.1 and Corollary 6.2 in [10] proved a refined error bound with a constant that has cubic dependence on the dimension

$$\mathbb{E}[(W_N)^2] \leq \frac{c\delta^2 d^3}{N^2}.$$

For perspective, it is known that mean squared error rates of order $1/N^2$ are generally optimal for estimation with uniform noise, see [11].

1.3 Overview and main results

The error bounds for consistent reconstruction in [10] only considered the mean squared error $\mathbb{E}[(W_N)^2]$ and only considered the admissibility condition (6) in the setting of unit-norm random vectors (for example, this excludes the case of Gaussian random vectors). The main contributions of this chapter are two-fold:

1. We prove bounds on general error moments $\mathbb{E}[(W_N)^p]$ for consistent reconstruction. Our main results show that the error decreases like $\mathbb{E}[(W_N)^p] \lesssim 1/N^p$, as the number of measurements N increases.
2. We establish a general admissibility condition on the sampling distribution that does not require φ to be unit-norm.

In Section 2, we prove our first main result, Theorem 1, which gives upper bounds on $\mathbb{E}[(W_N)^p]$ for unit-norm sampling distributions. Section 3 builds on Theorem 1 and proves our second main result, Theorem 2, for general sampling distributions that need not be unit-norm.

2 Error moments for consistent reconstruction: unit-norm distributions

In this section we prove our first main result, Theorem 1. Theorem 1 extends Theorem 5.5 in [10] to the setting of general error moments $\mathbb{E}[(W_N)^p]$. In this section, we assume that the sampling vectors $\{\varphi_n\}_{n=1}^N$ are unit-norm and satisfy the admissibility condition (6). We shall later remove the unit-norm requirement from the admissibility condition in Section 3.

2.1 Consistent reconstruction and coverage problems

We begin by recalling a useful connection between consistent reconstruction and a problem on covering the sphere by random sets.

Definition 1 Let $\{\varphi_n\}_{n=1}^N$ be a set of unit-norm vectors and let $\{\epsilon_n\}_{n=1}^N \subset [-\delta, \delta]$. For each $\lambda > 0$, define

$$B_n(\lambda) = B(\varphi_n, \epsilon_n, \lambda) = \left\{ u \in \mathbb{S}^{d-1} : \langle u, \varphi_n \rangle > \frac{\epsilon_n + \delta}{\lambda} \text{ or } \langle u, \varphi_n \rangle < \frac{\epsilon_n - \delta}{\lambda} \right\}$$
$$= \left\{ u \in \mathbb{S}^{d-1} : |\lambda \langle u, \varphi_n \rangle - \epsilon_n| > \delta \right\}. \tag{7}$$

In our setting, the sets $B_n(\lambda)$ are random subsets of $\mathbb{S}^{d-1}$ because $\{\varphi_n\}_{n=1}^N$ and $\{\epsilon_n\}_{n=1}^N$ are random.

Note that each $B_n(\lambda)$ can be expressed as a union of two (possibly empty) antipodal open spherical caps of different sizes

$$B_n(\lambda) = \mathrm{Cap}\left(\varphi_n, \theta_n^+\right) \cup \mathrm{Cap}\left(-\varphi_n, \theta_n^-\right), \tag{8}$$

where the angular radii θ_n^+ and θ_n^- are given by

$$\theta_n^+ = \begin{cases} \arccos\left(\frac{\delta + \epsilon_n}{\lambda}\right), & \text{if } \delta + \epsilon_n < \lambda, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\theta_n^- = \begin{cases} \arccos\left(\frac{\delta - \epsilon_n}{\lambda}\right), & \text{if } \delta - \epsilon_n < \lambda, \\ 0, & \text{otherwise.} \end{cases}$$

The following lemma shows a connection between consistent reconstruction and the problem of covering the unit sphere by the random sets $B_n(\lambda)$, see Lemma 4.1 in [10].

Lemma 1 *For all $\lambda > 0$, the worst case error satisfies*

$$\Pr\left[W_N > \lambda\right] \leq \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B_n\left(\lambda\right)\right]. \tag{9}$$

The following lemmas collect upper bounds on $\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B_n\left(\lambda\right)\right]$ that are spread out over various parts of [10].

Lemma 2 *If $\lambda \geq 4\delta$, then*

$$\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B_n\left(\lambda\right)\right] \leq 4^{d-1}(4^s\alpha)^N \left(\frac{\delta}{\lambda}\right)^{sN-d+1}. \tag{10}$$

Lemma 2 was shown in equation (5.9) in [10].

Lemma 3 *If $0 \leq \lambda \leq 4(2\alpha)^{1/s}\delta$, then*

$$\begin{aligned}&\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B_n\left(\lambda\right)\right] \\ &\leq \sum_{k=0}^{N} q(k, d-1, \alpha, s)\binom{N}{k}\left(1 - \frac{\lambda}{4\delta(2\alpha)^{1/s}}\right)^{N-k}\left(\frac{\lambda}{4\delta(2\alpha)^{1/s}}\right)^{k},\end{aligned} \tag{11}$$

where $q(k, d-1, \alpha, s)$ satisfies

$$q(k, d-1, \alpha, s) \leq 1, \tag{12}$$

and

$$k \geq \frac{2d\ln(16(2\alpha)^{1/s})}{\ln(4/3)} \quad \Longrightarrow \quad q(k, d-1, \alpha, s) \leq \left(\frac{3}{4}\right)^{k/2}. \tag{13}$$

The bound (11) appears in (5.12) in [10]. The bound (12) follows from (5.11) in [10], and the bound (13) appears in Step VI in the proof of Theorem 5.5 in [10].

2.2 Error moment bounds

We now prove our first main result that provides error moment bounds for consistent reconstruction.

Theorem 1 *Suppose that* $\{\varphi_n\}_{n=1}^N \subset \mathbb{S}^{d-1}$ *are independently drawn at random according to a distribution that satisfies the admissibility condition* (6) *with parameters* $\alpha \geq 1$ *and* $0 < s \leq 1$. *If* $p \in \mathbb{N}$ *and* $N \geq (d+p)/s$, *then the pth error moment for consistent reconstruction satisfies*

$$\mathbb{E}[(W_N)^p] \leq C' \delta^p \left(\prod_{j=1}^{p} (N+j) \right)^{-1} + C'' \delta^p \left(\frac{1}{2} \right)^N, \tag{14}$$

where

$$C' = C'_{p,\alpha,s} = 2p(4(2\alpha)^{1/s})^p \left(\frac{2d \ln(16(2\alpha)^{1/s})}{\ln(4/3)} + p \right)^p \left(\sum_{k=1}^{\infty} (k+1)^{p-1} (3/4)^{k/2} \right),$$

and

$$C'' = C''_{p,\alpha,s,d} = 2p(32(2\alpha)^{1/s})^{p+d-1}.$$

Proof We proceed by directly building on the proof of Theorem 5.5 in [10].

Step 1. We need to compute

$$\mathbb{E}[(W_N)^p] = p \int_0^\infty \lambda^{p-1} \Pr[W_N > \lambda] d\lambda. \tag{15}$$

By Lemma 1, we have

$$\mathbb{E}[(W_N)^p] \leq p \int_0^\infty \lambda^{p-1} \Pr \left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B_n(\lambda) \right] d\lambda. \tag{16}$$

Thus, it suffices to bound the integral on right side of (16).

Step 2. We shall bound the integral in (16) by breaking it up into three separate integrals. We begin by estimating the integral in the range $0 \leq \lambda \leq 4\delta(2\alpha)^{1/s}$.

Using (11) and a change of variables gives

$$p \int_0^{4\delta(2\alpha)^{1/s}} \lambda^{p-1} \Pr \left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B_n(\lambda) \right] d\lambda$$

$$\leq p \sum_{k=0}^{N} q(k, d-1, \alpha, s) \binom{N}{k} \int_0^{4\delta(2\alpha)^{1/s}} \lambda^{p-1} \left(1 - \frac{\lambda}{4\delta(2\alpha)^{1/s}}\right)^{N-k} \left(\frac{\lambda}{4\delta(2\alpha)^{1/s}}\right)^{k} d\lambda$$

$$= p \sum_{k=0}^{N} q(k, d-1, \alpha, s) \binom{N}{k} \left(4\delta(2\alpha)^{1/s}\right)^{p} \int_0^1 v^{k+p-1}(1-v)^{N-k} dv$$

$$= p \left(4\delta(2\alpha)^{1/s}\right)^{p} \sum_{k=0}^{N} q(k, d-1, \alpha, s) \binom{N}{k} \frac{(N-k)!(k+p-1)!}{(N+p)!}$$

$$= p \left(4\delta(2\alpha)^{1/s}\right)^{p} \left(\prod_{j=1}^{p} (N+j)\right)^{-1} \left[\sum_{k=0}^{N} \frac{(k+p-1)!}{k!} q(k, d-1, \alpha, s)\right]. \tag{17}$$

Here, we used the property of the beta function that

$$\int_0^1 v^{k+p-1}(1-v)^{N-k} dv = \frac{(N-k)!(k+p-1)!}{(N+p)!}. \tag{18}$$

It remains to bound the sum $\sum_{k=0}^{N} \frac{(k+p-1)!}{k!} q(k, d-1, \alpha, s)$ in (17). We will bound this sum by breaking it up into two separate sums, in an analogous manner to Step VI in the proof of Theorem 5.5 in [10]. Let

$$K = \left\lfloor \frac{2d \ln(16(2\alpha)^{1/s})}{\ln(4/3)} \right\rfloor. \tag{19}$$

Since $q(k, d-1, \alpha, s) \leq 1$, we have

$$\sum_{k=0}^{K} \frac{(k+p-1)!}{k!} q(k, d-1, \alpha, s) \leq \sum_{k=0}^{K} (K+p-1)^{p-1} \leq (K+p)^{p}. \tag{20}$$

Using (13) we have

$$\begin{aligned} \sum_{k=K+1}^{N} \frac{(k+p-1)!}{k!} q(k, d-1, \alpha, s) &\leq \sum_{k=K+1}^{\infty} \frac{(k+p-1)!}{k!} \left(\frac{3}{4}\right)^{k/2} \\ &\leq \sum_{k=K+1}^{\infty} (k+p-1)^{p-1} \left(\frac{3}{4}\right)^{k/2} \\ &= \sum_{k=1}^{\infty} (k+K+p-1)^{p-1} \left(\frac{3}{4}\right)^{(k+K)/2} \end{aligned}$$

$$\leq (K+p)^{p-1} \sum_{k=0}^{\infty} (k+1)^{p-1} \left(\frac{3}{4}\right)^{k/2}$$
$$= (K+p)^{p-1} S_p, \tag{21}$$

where $S_p = \sum_{k=1}^{\infty} (k+1)^{p-1} (3/4)^{k/2}$ satisfies $1 < S_p < \infty$.

By (20) and (21) we have

$$\sum_{k=0}^{N} \frac{(k+p-1)!}{k!} q(k, d-1, \alpha, s) \leq (K+p)^p (1+S_p) \leq 2(K+p)^p S_p. \tag{22}$$

Combining (17) and (22) yields

$$p \int_0^{4\delta(2\alpha)^{1/s}} \lambda^{p-1} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B_n(\lambda)\right] d\lambda$$
$$\leq 2p(4\delta(2\alpha)^{1/s})^p (K+p)^p S_p \left(\prod_{j=1}^{p} (N+j)\right)^{-1}. \tag{23}$$

Step 3. Next, we bound the integral (16) in the range $4\delta(2\alpha)^{1/s} \leq \lambda \leq 8\delta(2\alpha)^{1/s}$. By Lemma 2 we know that in this range of λ,

$$\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B_n(\lambda)\right] \leq (16(2\alpha)^{1/s})^{d-1} \left(\frac{1}{2}\right)^N.$$

Thus

$$p \int_{4\delta(2\alpha)^{1/s}}^{8\delta(2\alpha)^{1/s}} \lambda^{p-1} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B_n(\lambda)\right] d\lambda$$
$$\leq p(16(2\alpha)^{1/s})^{d-1} \left(\frac{1}{2}\right)^N \int_{4\delta(2\alpha)^{1/s}}^{8\delta(2\alpha)^{1/s}} \lambda^{p-1} d\lambda$$
$$\leq \delta^p (16(2\alpha)^{1/s})^{d+p-1} \left(\frac{1}{2}\right)^N. \tag{24}$$

Step 4. We next bound the integral (16) in the range $\lambda \geq 8\delta(2\alpha)^{1/s}$. By Lemma 2 we know that in this range of λ,

$$\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B_n(\lambda)\right] \leq 4^{d-1}(4^s\alpha)^N \left(\frac{\delta}{\lambda}\right)^{sN-d+1}.$$

It follows that when $N \geq (d+p)/s$,

$$\begin{aligned} p\int_{8\delta(2\alpha)^{1/s}}^{\infty} \lambda^{p-1} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B_n(\lambda)\right] d\lambda \\ &\leq p \cdot 4^{d-1}(4^s\alpha)^N \delta^{sN-d+1} \int_{8\delta(2\alpha)^{1/s}}^{\infty} \lambda^{p-sN+d-2} d\lambda \\ &= p \cdot 4^{d-1}(4^s\alpha)^N \delta^{sN-d+1} \left(\frac{(8\delta(2\alpha)^{1/s})^{p-sN+d-1}}{sN-p-d+1}\right) \\ &\leq p \cdot \delta^p (32(2\alpha)^{1/s})^{p+d-1} \left(\frac{1}{2}\right)^N. \end{aligned} \tag{25}$$

Combining (16), (23), (24), and (25) completes the proof.

Theorem 1 yields the following corollary.

Corollary 1 *Suppose that* $\{\varphi_n\}_{n=1}^N \subset \mathbb{S}^{d-1}$ *are independently drawn at random according to a distribution that satisfies the admissibility condition* (6) *with parameters* $\alpha \geq 1$ *and* $0 < s \leq 1$. *If* $p \in \mathbb{N}$ *and*

$$N \geq \max\left\{\frac{2}{\ln 2}\left[\ln\left(\frac{C''}{C'}\right) + 2p\ln\left(\frac{4p}{e\ln 2}\right)\right], \frac{d+p}{s}\right\}, \tag{26}$$

then

$$\mathbb{E}[(W_N)^p] \leq 2C'\delta^p \left(\prod_{j=1}^{p}(N+j)\right)^{-1}, \tag{27}$$

where C', C'' *are as in Theorem 1.*

Proof In view of Theorem 1, it suffices to show that if N satisfies (26) then

$$C''\left(\frac{1}{2}\right)^N \leq C'\left(\prod_{j=1}^{p}(N+j)\right)^{-1}.$$

Equivalently, it suffices to show

$$\ln\left(\frac{C''}{C'}\right) + \sum_{j=1}^{p} \ln(N+j) \le N \ln 2. \tag{28}$$

To begin, note that

$$\forall x > 0, \quad \ln(x) \le x - 1,$$

gives

$$\begin{aligned} \ln(N) &= \ln\left(\frac{N \ln 2}{4p}\right) + \ln\left(\frac{4p}{\ln 2}\right) \\ &\le \frac{N \ln 2}{4p} - 1 + \ln\left(\frac{4p}{\ln 2}\right) \\ &= \frac{N \ln 2}{4p} + \ln\left(\frac{4p}{e \ln 2}\right). \end{aligned} \tag{29}$$

Next, use (29) and $N \ge (d+p)/s \ge \max\{p, 2\}$ to obtain

$$\begin{aligned} \sum_{j=1}^{p} \ln(N+j) &= \sum_{j=1}^{p} \left[\ln(N) + \ln\left(1 + \frac{j}{N}\right)\right] \\ &\le p \ln(N) + p \ln 2 \\ &\le 2p \ln(N) \\ &\le \frac{N \ln 2}{2} + 2p \ln\left(\frac{4p}{e \ln 2}\right). \end{aligned} \tag{30}$$

In view of (30), to show (28) it suffices to have

$$\ln\left(\frac{C''}{C'}\right) + \frac{N \ln 2}{2} + 2p \ln\left(\frac{4p}{e \ln 2}\right) \le N \ln 2. \tag{31}$$

Since (31) holds by the assumption (26), this completes the proof.

We conclude this section with some perspective on the dimension dependence of the constant C' in Theorem 1 and Corollary 1. We consider the special case when φ is uniformly distributed on the unit-sphere $\mathbb{S}^{d-1}$ with $d \ge 3$. In this case, one may take $s = 1$ and $\alpha = \frac{2\Gamma(d/2)}{\sqrt{\pi}\Gamma((d-1)/2)}$ in (6), see Example 5.1 in [10], and the constant C' is of order $\left(d^{\frac{3}{2}} \ln d\right)^p$. Here, the logarithmic factor $\ln d$ is an artifact of the general setting of Theorem 1. In particular, for $p = 2$ the refined analysis in Theorem 6.1 and Corollary 6.2 of [10] shows that the factor $\ln d$ can be removed

when φ is uniformly distributed on the unit-sphere $\mathbb{S}^{d-1}$. A similar analysis extends to moments with general values of $p \in \mathbb{N}$ and shows that the factor $\ln d$ can be replaced by an absolute constant that is independent of d.

3 Error moments for consistent reconstruction: general distributions

In Section 2 we proved bounds on the pth error moment for consistent reconstruction when the measurements are made using i.i.d. copies of a unit-norm random vector $\varphi \in \mathbb{S}^{d-1}$. In this section, we relax the unit-norm constraint to accommodate more general distributions.

3.1 General admissibility condition

Definition 2 We shall say that a random vector $\varphi \in \mathbb{R}^d$ satisfies the general admissibility condition if the following conditions hold:

- $\varphi = a\psi$, where a is a non-negative random variable, ψ is a unit-norm random vector, and a and ψ are independent.
- ψ satisfies the admissibility condition (6).
- $\exists C > 0$ such that

$$\forall \lambda > 0, \quad \lambda \Pr[a\lambda \leq 1] \leq C. \tag{32}$$

- $r_a = \Pr[a > 1]$ satisfies $0 < r_a < 1$.

Example 1 A sufficient condition for the small-ball inequality (32) to hold is when a is an absolutely continuous random variable whose probability density function f is in $L^\infty(\mathbb{R})$. In this case, for each $\lambda > 0$,

$$\Pr[a\lambda \leq 1] = \Pr\left[a \leq \frac{1}{\lambda}\right] = \int_0^{1/\lambda} f(a)\, da \leq \frac{\|f\|_\infty}{\lambda}.$$

This shows that a large class of probability distributions satisfy the conditions in Definition 2. For example, if φ is a random vector whose entries are i.i.d zero mean Gaussian random variables, then φ satisfies the conditions in Definition 2.

In Definition 2, there would be no loss of generality if a were scaled differently so that $0 < \Pr[a > T] < 1$ for some $T > 0$. In particular, suppose that $\varphi_n = a_n\psi_n$ with $0 < \Pr[a_n > T] < 1$, and $q_n = \langle x, \varphi_n\rangle + \epsilon_n$ with ϵ_n uniformly distributed on $[-\delta, \delta]$. Then $\widetilde{x} \in \mathbb{R}^d$ satisfies

$$|\langle \widetilde{x}, \varphi_n \rangle - q_n| \le \delta \quad \text{if and only if} \quad |\langle \widetilde{x}, \varphi'_n \rangle - q'_n| \le \delta',$$

where $\varphi'_n = \varphi_n/T = a'_n \psi_n$ and $a'_n = a_n/T$ and $q'_n = \langle x, \varphi'_n \rangle + \epsilon'_n$, where $\epsilon'_n = \epsilon_n/T$ is uniformly distributed on $[-\delta', \delta']$ with $\delta' = \delta/T$.

3.2 Coverage problems revisited

Suppose that $\{\varphi_n\}_{n=1}^N$ are i.i.d. versions of a random vector φ that satisfies the conditions of Definition 2. In particular, $\varphi_n = a_n \psi_n$, where $\{a_n\}_{n=1}^N$ i.i.d. versions of a random variable a, and $\{\psi_n\}_{n=1}^N$ are i.i.d. versions of a random vector ψ. Similar to Lemma 1, the worst case error W_N for consistent reconstruction can be bounded by

$$\Pr[W_N > \lambda] \le \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^N B(\psi_n, \epsilon_n, a_n \lambda) \right], \tag{33}$$

where $B(\psi_n, \epsilon_n, a_n \lambda)$ is defined using (7).

3.2.1 Conditioning and a bound by caps with $a_n = 1$

The following lemma bounds (33) by coverage probabilities involving caps with $a_n = 1$.

Lemma 4 *Suppose $\{\varphi_n\}_{n=1}^N$, with $\varphi_n = a_n \psi_n$, are i.i.d. versions of a random vector φ that satisfies the conditions of Definition 2. Then*

$$\begin{aligned} &\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^N B(\psi_n, \epsilon_n, a_n \lambda) \right] \\ &\le \sum_{j=1}^N \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^j B(\psi_n, \epsilon_n, \lambda) \right] \mathrm{bino}(j, N, r) + (1-r)^N, \end{aligned} \tag{34}$$

where

$$\mathrm{bino}(j, N, r) = \binom{N}{j} r^j (1-r)^{N-j},$$

and $r = r_a = \Pr[a > 1]$ is as in Definition 2.

Proof Let $\mathscr{J}_{j,N}$ denote the event that exactly j elements of $\{a_n\}_{n=1}^N$ satisfy $a_n > 1$. Since the $\{a_n\}_{n=1}^N$ are independent versions of the random variable a,

$$\begin{aligned}\Pr[\mathscr{J}_{j,N}] &= \binom{N}{j}(\Pr[a > 1])^j(1 - \Pr[a > 1])^{N-j} \\ &= \binom{N}{j} r^j(1-r)^{N-j} = \mathrm{bino}(j, N, r).\end{aligned}$$

Thus,

$$\begin{aligned}&\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^N B(\psi_n, \epsilon_n, a_n\lambda)\right] \\ &= \sum_{j=0}^N \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^N B(\psi_n, \epsilon_n, a_n\lambda) \,\middle|\, \mathscr{J}_{j,N}\right] \Pr\left[\mathscr{J}_{j,N}\right] \\ &= \sum_{j=0}^N \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^N B(\psi_n, \epsilon_n, a_n\lambda) \,\middle|\, \mathscr{J}_{j,N}\right] \mathrm{bino}(j, N, r). \qquad (35)\end{aligned}$$

By (7), when $a_n > 1$ we have $B(\psi_n, \epsilon_n, a_n\lambda) \supset B(\psi_n, \epsilon_n, \lambda)$. Thus for $1 \le j \le N$,

$$\begin{aligned}\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^N B(\psi_n, \epsilon_n, a_n\lambda) \,\middle|\, \mathscr{J}_{j,N}\right] &\le \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{\{n:a_n>1\}} B(\psi_n, \epsilon_n, a_n\lambda) \,\middle|\, \mathscr{J}_{j,N}\right] \\ &\le \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{\{n:a_n>1\}} B(\psi_n, \epsilon_n, \lambda) \,\middle|\, \mathscr{J}_{j,N}\right] \\ &= \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^j B(\psi_n, \epsilon_n, \lambda)\right], \qquad (36)\end{aligned}$$

where the last equality holds because $\{a_n\}_{n=1}^N$ are i.i.d. random variables that are independent of the i.i.d. random vectors $\{\psi_n\}_{n=1}^N$. For $j = 0$, we use the trivial bound

$$\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{\{n:a_n>1\}} B(\psi_n, \epsilon_n, \lambda) \,\middle|\, \mathscr{J}_{j,N}\right] \le 1.$$

Combining (35) and (36) completes the proof.

To bound the binomial terms in Lemma 4 it will be useful to recall Hoeffding's inequality for Bernoulli random variables. If $0 < p < 1$ and $m \leq Np$, then

$$\sum_{j=0}^{m} \text{bino}(j, N, p) \leq \exp\left(-2\left(Np - m\right)^2 / N\right). \tag{37}$$

3.2.2 Covering and discretization

A useful technique for bounding coverage probabilities such as (33) is to discretize the problem by discretizing the sphere $\mathbb{S}^{d-1}$ with an ϵ-net, see [3]. In this section, we briefly recall necessary aspects of this discretization method as used in [10].

Recall that a set $\mathscr{N}_\epsilon \subset \mathbb{S}^{d-1}$ is a geodesic ϵ-net for $\mathbb{S}^{d-1}$ if

$$\forall x \in \mathbb{S}^{d-1},\ \exists\, z \in \mathscr{N}_\epsilon, \quad \text{such that} \quad \arccos(\langle x, z\rangle) \leq \epsilon.$$

For the remainder of this section, let $\mathscr{N}_\epsilon$ be a geodesic ϵ- net of cardinality

$$\#\left(\mathscr{N}_\epsilon\right) \leq \left(\frac{8}{\epsilon}\right)^{d-1}.$$

It is well known that geodesic ϵ-nets of such cardinality exist, e.g., see Lemma 13.1.1 in [8] or Section 2.2 in [10].

Recalling (8), define the shrunken bi-cap $T_\epsilon\left[B(\psi_n, \epsilon_n, a_n\lambda)\right]$ by

$$T_\epsilon\left[B(\psi_n, \epsilon_n, a_n\lambda)\right] = \text{Cap}\left(\psi_n, T_\epsilon(\theta_n^+)\right) \cup \text{Cap}\left(-\psi_n, T_\epsilon(\theta_n^-)\right),$$

where

$$T_\epsilon(\theta) = \begin{cases} \theta - \epsilon, & \text{if } \theta \geq \epsilon; \\ 0, & \text{if } 0 \leq \theta \leq \epsilon. \end{cases}$$

Similar to equations (5.4) and (5.5) in [10], the coverage probability (33) can be discretized as follows:

$$\begin{aligned} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B(\psi_n, \epsilon_n, a_n\lambda)\right] &\leq \Pr\left[\mathscr{N}_\epsilon \not\subset \bigcup_{n=1}^{N} T_\epsilon\left[B(\psi_n, \epsilon_n, a_n\lambda)\right]\right] \\ &\leq \left(\frac{8}{\epsilon}\right)^{d-1} \left(\sup_{z \in \mathbb{S}^{d-1}} \Pr\left[z \notin T_\epsilon\left[B(\psi_n, \epsilon_n, a_n\lambda)\right]\right]\right)^N. \end{aligned} \tag{38}$$

Similar to equation (5.6) in [10], one has that

$$B(\psi_n, \epsilon_n, a_n\lambda) \supset \left\{ u \in \mathbb{S}^{d-1} : |\langle u, \psi_n \rangle| > \frac{2\delta}{a_n\lambda} \right\}$$

and

$$T_\epsilon \left[B(\psi_n, \epsilon_n, a_n\lambda) \right] \supset \left\{ u \in \mathbb{S}^{d-1} : |\langle u, \psi_n \rangle| > \frac{2\delta}{a_n\lambda} + \epsilon \right\}.$$

This gives

$$\Pr\left[z \notin T_\epsilon \left[B(\psi_n, \epsilon_n, a_n\lambda) \right] \right] \leq \Pr\left[|\langle z, \psi_n \rangle| \leq \frac{2\delta}{a_n\lambda} + \epsilon \right]. \tag{39}$$

3.3 *Moment bounds for general distributions*

We now state our next main theorem.

Theorem 2 *Suppose that* $\{\varphi_n\}_{n=1}^N$ *are i.i.d. versions of a random vector* φ *that satisfies the conditions of Definition 2. Let* $r = r_a = \Pr[a > 1]$ *be as in Definition 2. If*

$$N \geq \frac{2(d+p)}{sr}, \tag{40}$$

then the pth error moment for consistent reconstruction satisfies

$$\mathbb{E}\left[(W_N)^p \right] \leq pC' \left(\frac{2\delta}{Nr} \right)^p + pC'' \delta^p \left(\frac{1}{2} \right)^{Nr/2} + \delta^p \Lambda^p e^{-Nr^2/2} + \delta^p C''' \left(\frac{1}{2} \right)^N,$$

where C', C'' *are as in Theorem 1,* Λ *is defined by* (42) *and* (57)*, and* C''' *is defined by* (60) *and* (57)*.*

Proof As in Theorem 1 we shall use (15). In view of (33), we need to estimate

$$\mathbb{E}[(W_N)^p] \leq p \int_0^\infty \lambda^{p-1} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^N B(\psi_n, \epsilon_n, a_n\lambda) \right] d\lambda. \tag{41}$$

Step 1. We begin by estimating the integral in (41) over the range $0 \le \lambda \le \Lambda\delta$, where

$$\Lambda = \max\{\Lambda_0, \Lambda_1\}, \quad \text{with} \quad \Lambda_0 = \frac{2^{s+3}C}{\alpha} \quad \text{and} \quad \Lambda_1 = 4\left(2K''\right)^{\frac{s+1}{s}}, \tag{42}$$

and K'' is defined in (57).

By Lemma 4 we have

$$\begin{aligned}
&p\int_0^{\Lambda\delta} \lambda^{p-1} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B(\psi_n, \epsilon_n, a_n\lambda)\right] d\lambda \\
&\le p\int_0^{\Lambda\delta} \lambda^{p-1} \sum_{j=0}^{N} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{j} B(\psi_n, \epsilon_n, \lambda)\right] \mathrm{bino}(j, N, r) d\lambda \\
&= p\int_0^{\Lambda\delta} \lambda^{p-1} \sum_{j=0}^{\lfloor Nr/2\rfloor} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{j} B(\psi_n, \epsilon_n, \lambda)\right] \mathrm{bino}(j, N, r) d\lambda \qquad (43) \\
&\quad + p\int_0^{\Lambda\delta} \lambda^{p-1} \sum_{j=\lceil Nr/2\rceil}^{N} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{j} B(\psi_n, \epsilon_n, \lambda)\right] \mathrm{bino}(j, N, r) d\lambda. \qquad (44)
\end{aligned}$$

Hoeffding's inequality and the trivial bound $\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{j} B(\psi_n, \epsilon_n, \lambda)\right] \le 1$ can be used to bound (43) as follows:

$$\begin{aligned}
&p\int_0^{\Lambda\delta} \lambda^{p-1} \sum_{j=0}^{\lfloor Nr/2\rfloor} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{j} B(\psi_n, \epsilon_n, \lambda)\right] \mathrm{bino}(j, N, r) d\lambda \\
&\qquad \le p\int_0^{\Lambda\delta} \lambda^{p-1} \left(\sum_{j=0}^{\lfloor Nr/2\rfloor} \mathrm{bino}(j, N, r)\right) d\lambda \\
&\qquad \le p\left(e^{-Nr^2/2}\right) \int_0^{\Lambda\delta} \lambda^{p-1} d\lambda \\
&\qquad = \delta^p \Lambda^p e^{-Nr^2/2}. \qquad (45)
\end{aligned}$$

To bound the integral in (44), recall (40) and note that if j satisfies $(d+p)/s \le \lceil Nr/2\rceil \le j \le N$, then the bounds on (16) obtained in the proof of Theorem 1 give that

$$p\int_0^{\Lambda\delta} \lambda^{p-1} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{j} B(\psi_n, \epsilon_n, \lambda)\right] d\lambda$$

$$\leq p \int_0^\infty \lambda^{p-1} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{j} B(\psi_n, \epsilon_n, \lambda)\right] d\lambda$$

$$\leq C' \delta^p \left(\prod_{l=1}^{p} (j+l)\right)^{-1} + C'' \delta^p \left(\frac{1}{2}\right)^j$$

$$\leq \frac{C' \delta^p}{j^p} + C'' \delta^p \left(\frac{1}{2}\right)^j, \tag{46}$$

where C' and C'' are as in Theorem 1.

Using (46), along with $\sum_{j=0}^{N} \text{bino}(j, N, r) = 1$, one may bound (44) as follows:

$$p \sum_{j=\lceil Nr/2 \rceil}^{N} \int_0^{\Lambda\delta} \lambda^{p-1} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{j} B(\psi_n, \epsilon_n, \lambda)\right] \text{bino}(j, N, r) d\lambda$$

$$\leq p \sum_{j=\lceil Nr/2 \rceil}^{N} \text{bino}(j, N, r) \left[\frac{C' \delta^p}{j^p} + C'' \delta^p \left(\frac{1}{2}\right)^j\right]$$

$$\leq p\delta^p \left[\frac{2^p C'}{(Nr)^p} + C'' \left(\frac{1}{2}\right)^{Nr/2}\right] \sum_{j=\lceil Nr/2 \rceil}^{N} \text{bino}(j, N, r)$$

$$\leq p\delta^p \left[C' \left(\frac{2}{Nr}\right)^p + C'' \left(\frac{1}{2}\right)^{Nr/2}\right]. \tag{47}$$

Applying the bounds (45) and (47) to (43) and (44) gives

$$p \int_0^{\Lambda\delta} \lambda^{p-1} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B(\psi_n, \epsilon_n, a_n\lambda)\right] d\lambda$$

$$\leq \delta^p \Lambda^p e^{-Nr^2/2} + pC' \left(\frac{2\delta}{Nr}\right)^p + pC'' \delta^p \left(\frac{1}{2}\right)^{Nr/2}. \tag{48}$$

Step 2. We next estimate the integral in (41) over the range $\lambda \geq \Lambda\delta$. By (38) and (39) we have

$$\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B(\psi_n, \epsilon_n, a_n\lambda)\right]$$

$$\leq \left(\frac{8}{\epsilon}\right)^{d-1} \left(\sup_{z \in \mathbb{S}^{d-1}} \Pr\left[|\langle z, \psi_n \rangle| \leq \frac{2\delta}{a_n\lambda} + \epsilon\right]\right)^N. \tag{49}$$

We therefore need to bound $\Pr[|\langle z, \psi_n\rangle| \le \frac{2\delta}{a_n\lambda} + \epsilon]$.

For the remainder of this step set

$$A = \left(\frac{\alpha}{C}\right)^{\frac{1}{s+1}} \left(\frac{4\delta}{\lambda}\right)^{\frac{s}{s+1}} \quad \text{and} \quad \epsilon = \frac{2\delta}{A\lambda} = \left(\frac{1}{2}\right)\left(\frac{4\delta C}{\lambda\alpha}\right)^{\frac{1}{s+1}}, \tag{50}$$

where C, α, s are the parameters in (6) and Definition (2). By (42), note that $\lambda \ge \Lambda\delta \ge \Lambda_0\,\delta$ implies that $0 < \epsilon \le 1/4$.

For any $z \in \mathbb{S}^{d-1}$ we have

$$\Pr\left[|\langle z, \psi_n\rangle| \le \frac{2\delta}{a_n\lambda} + \epsilon\right] = \Pr\left[|\langle z, \psi_n\rangle| \le \frac{2\delta}{a_n\lambda} + \epsilon \,\middle|\, a_n > A\right] \Pr[a_n > A] \tag{51}$$

$$+ \Pr\left[|\langle z, \psi_n\rangle| \le \frac{2\delta}{a_n\lambda} + \epsilon \,\middle|\, a_n \le A\right] \Pr[a_n \le A]. \tag{52}$$

We now bound the terms appearing in (51). Recall that $\lambda \ge \Lambda\delta$ implies that $4\delta/(A\lambda) = 2\epsilon \le 1/2$. By our choice of ϵ in (50), and using the admissibility assumption (6), for each $\lambda \ge \Lambda\delta$ one has

$$\begin{aligned}
&\Pr\left[|\langle z, \psi_n\rangle| \le \frac{2\delta}{a_n\lambda} + \epsilon \,\middle|\, a_n > A\right] \Pr[a_n > A] \\
&\le \Pr\left[|\langle z, \psi_n\rangle| \le \frac{2\delta}{A\lambda} + \epsilon \,\middle|\, a_n > A\right] \Pr[a_n > A] \\
&= \Pr\left[|\langle z, \psi_n\rangle| \le \frac{4\delta}{A\lambda} \,\middle|\, a_n > A\right] \Pr[a_n > A] \\
&\le \Pr\left[|\langle z, \psi_n\rangle| \le \frac{4\delta}{A\lambda}\right] \\
&\le \alpha\left(\frac{4\delta}{A\lambda}\right)^s.
\end{aligned} \tag{53}$$

To bound (52), note that by (32) one has $\Pr[a_n \le A] \le CA$, and thus

$$\Pr\left[|\langle z, \psi_n\rangle| \le \frac{2\delta}{a_n\lambda} + \epsilon \,\middle|\, a_n \le A\right] \Pr[a_n \le A] \le \Pr[a \le A] \le CA. \tag{54}$$

Using the bounds (53) and (54) in (51) and (52) gives

$$\Pr\left[|\langle z, \psi_n\rangle| \le \frac{2\delta}{a_n\lambda} + \epsilon\right] \le \alpha\left(\frac{4\delta}{A\lambda}\right)^s + CA. \tag{55}$$

Since our choice of A in (50) gives

$$\alpha\left(\frac{4\delta}{A\lambda}\right)^{s} = CA,$$

we have

$$\Pr\left[|\langle z,\psi_n\rangle| \le \frac{2\delta}{a_n\lambda} + \epsilon\right] \le 2CA = 2C\left(\frac{\alpha}{C}\right)^{\frac{1}{s+1}}\left(\frac{4\delta}{\lambda}\right)^{\frac{s}{s+1}}. \tag{56}$$

Thus, combining (49) and (56) gives

$$\begin{aligned}\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B(\psi_n,\epsilon_n,a_n\lambda)\right] &\le \left(\frac{8}{\epsilon}\right)^{d-1}\left[2C\left(\frac{\alpha}{C}\right)^{\frac{1}{s+1}}\left(\frac{4\delta}{\lambda}\right)^{\frac{s}{s+1}}\right]^{N}\\ &= \left(16\left(\frac{\alpha\lambda}{4\delta C}\right)^{\frac{1}{s+1}}\right)^{d-1}\left[2C\left(\frac{\alpha}{C}\right)^{\frac{1}{s+1}}\left(\frac{4\delta}{\lambda}\right)^{\frac{s}{s+1}}\right]^{N}.\end{aligned}$$

To simplify notation, let

$$K' = \left(16\left(\frac{\alpha}{C}\right)^{\frac{1}{s+1}}\right)^{d-1} \quad \text{and} \quad K'' = 2C\left(\frac{\alpha}{C}\right)^{\frac{1}{s+1}}, \tag{57}$$

so that

$$\begin{aligned}\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B(\psi_n,\epsilon_n,a_n\lambda)\right] &\le K'\left(\frac{\lambda}{4\delta}\right)^{\frac{d-1}{s+1}}\left[K''\left(\frac{4\delta}{\lambda}\right)^{\frac{s}{s+1}}\right]^{N}\\ &= K'\,(K'')^{N}\left(\frac{4\delta}{\lambda}\right)^{\left(\frac{sN-d+1}{s+1}\right)}.\end{aligned} \tag{58}$$

Since $0 < s \le 1$ and $0 < r < 1$, note that (40) implies $\left(\frac{sN-d+1}{s+1} - p + 1\right) \ge 2$. By (58) we have

$$\begin{aligned}&p\int_{\Lambda\delta}^{\infty}\lambda^{p-1}\Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B(\psi_n,\epsilon_n,a_n\lambda)\right]d\lambda\\ &\quad\le pK'(K'')^{N}\int_{\Lambda\delta}^{\infty}\lambda^{p-1}\left(\frac{4\delta}{\lambda}\right)^{\left(\frac{sN-d+1}{s+1}\right)}d\lambda\\ &\quad= pK'(K'')^{N}(4\delta)^{p-1}\int_{\Lambda\delta}^{\infty}\left(\frac{4\delta}{\lambda}\right)^{\left(\frac{sN-d+1}{s+1}-p+1\right)}d\lambda\end{aligned}$$

$$
\begin{aligned}
&= pK'(K'')^N(4\delta)^p \int_{\Lambda/4}^{\infty} \left(\frac{1}{\lambda}\right)^{\left(\frac{sN-d+1}{s+1}-p+1\right)} d\lambda \\
&= pK'(K'')^N(4\delta)^p \left(\frac{\Lambda}{4}\right)^{p-\frac{sN-d+1}{s+1}} \left(\frac{sN-d+1}{s+1}-p\right)^{-1}. \\
&\leq pK'(K'')^N(4\delta)^p \left(\frac{\Lambda}{4}\right)^{p-\frac{sN-d+1}{s+1}} \\
&= pK'(4\delta)^p \left(\frac{\Lambda}{4}\right)^{p+\frac{d-1}{s+1}} \left[K''\left(\frac{4}{\Lambda}\right)^{\frac{s}{s+1}}\right]^N.
\end{aligned}
$$

Since (42) implies that $K''\left(\frac{4}{\Lambda}\right)^{\frac{s}{s+1}} \leq 1/2$, it follows that

$$
p\int_{\Lambda\delta}^{\infty} \lambda^{p-1} \Pr\left[\mathbb{S}^{d-1} \not\subset \bigcup_{n=1}^{N} B(\psi_n, \epsilon_n, a_n\lambda)\right] d\lambda \leq \delta^p C'''\left(\frac{1}{2}\right)^N, \tag{59}
$$

where

$$
C''' = pK'4^p\left(\frac{\Lambda}{4}\right)^{p+\frac{d-1}{s+1}}. \tag{60}
$$

Combining (41), (48) and (59) completes the proof.

Similar to Corollary 1, the following corollary of Theorem 2 shows that $\mathbb{E}[(W_N)^p]$ is at most of order $1/N^p$ when N is sufficiently large.

Corollary 2 *Let $\{\varphi_n\}_{n=1}^N$ be as in Theorem 2. There exist constants $C_1, C_2 > 0$ such that*

$$
\forall N \geq C_1, \quad \mathbb{E}\left[(W_N)^p\right] \leq \frac{C_2\delta^p}{N^p}. \tag{61}
$$

The constants C_1, C_2 depend on α, s, C, p, d.

3.4 Numerical experiment

This section illustrates Theorem 2 with a numerical experiment.

Let $x = (2, \pi)$ and $\delta = \frac{1}{10}$. Given $N \geq 3$, let $\{e_n\}_{n=1}^N \subset \mathbb{R}^2$ be independent random vectors with i.i.d. $N(0, 1)$ entries. Let $\{q_n\}_{n=1}^N$ be defined as in (1). Since there infinitely many different solutions $\widetilde{x}$ to the consistent reconstruction condition (2), we select the minimal norm estimate by

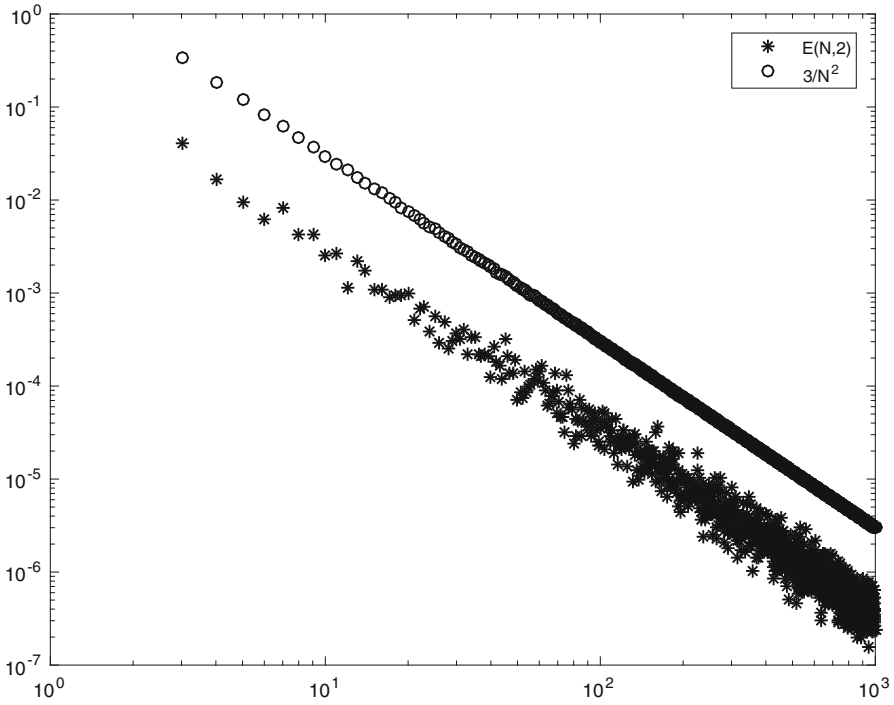


Fig. 1 Log-log plot of $E(N, 2)$ versus N, see Section 3.4.

$$\widetilde{x} = \text{ argmin}_{z\in\mathbb{R}^2} \|z\|^2 \quad \text{subject to} \quad |\langle z, \varphi_n\rangle - q_n| \leq \delta, \quad 1 \leq n \leq N. \tag{62}$$

We repeat this experiment 20 times and let $E(N, p)$ denote the average value of $\|\widetilde{x} - x\|^p$. Figures 1 and 2 show log-log plots of $E(N, p)$ versus N for $p = 2$ and $p = 5$. For comparison, these respective figures also show log-log plots of $3/N^2$ and $20/N^5$ versus N. In particular, $E(N, p)$ appears to decay like $1/N^p$, as predicted by the worst case error bounds in Theorem 2.

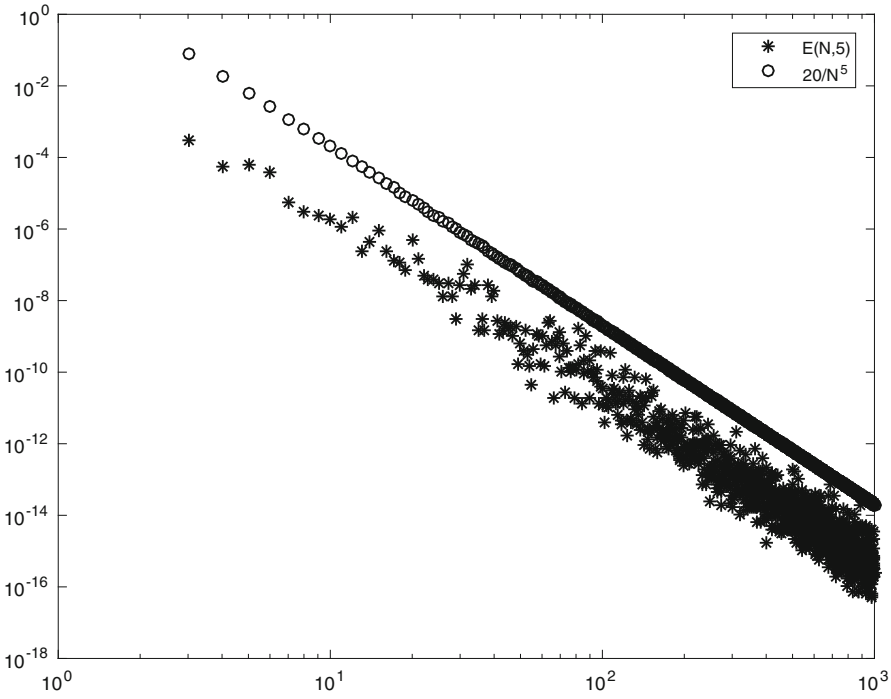


Fig. 2 Log-log plot of $E(N, 5)$ versus N, see Section 3.4.

Acknowledgements C.-H. Lee was partially supported by NSF DMS 1211687. A.M. Powell was partially supported by NSF DMS 1521749 and NSF DMS 1211687. A.M. Powell gratefully acknowledges the hospitality and support of the Academia Sinica Institute of Mathematics (Taipei, Taiwan).

References

1. Z. Cvetković, Resilience properties of redundant expansions under additive noise and quantization. IEEE Trans. Inf. Theory **49**, 644–656 (2003)
2. Z. Cvetković, M. Vetterli, On simple oversampled A/D conversion in $L^2(\mathbb{R})$. IEEE Trans. Inf. Theory **47**, 146–154 (2001)
3. N. Flatto, D.J. Newman, Random coverings. Acta Math. **128**, 241–264 (1977)
4. V.K. Goyal, M. Vetterli, N.T. Thao, Quantized overcomplete expansions in $\mathbb{R}^n$: analysis, synthesis, and algorithms. IEEE Trans. Inf. Theory **44**, 16–30 (1998)
5. L. Jacques, D.K. Hammond, J.M. Fadili, Dequantizing compressed sensing: when oversampling and non-Gaussian constraints combine. IEEE Trans. Inf. Theory **57**, 560–571 (2011)

6. L. Jacques, J.N. Laska, P.T. Boufounos, R.G. Baraniuk, Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. IEEE Trans. Inf. Theory **59**, 2082–2102 (2013)
7. D. Jimenez, L. Wang, Y. Wang, White noise hypothesis for uniform quantization errors. SIAM J. Math. Anal. 38, 2042–2056 (2007)
8. J. Matoušek, *Lectures on Discrete Geometry* (Springer, New York, 2002)
9. Y. Plan, R. Vershynin, One-bit compressed sensing by linear programming. Commun. Pure Appl. Math. **66**, 1275–1297 (2013)
10. A.M. Powell, J.T. Whitehouse, Error bounds for consistent reconstruction: random polytopes and coverage processes. Found. Comput. Math. **16**, 395–423 (2016)
11. S. Rangan, V.K. Goyal, Recursive consistent estimation with bounded noise. IEEE Trans. Inf. Theory **47**, 457–464 (2001)
12. N.T. Thao, M. Vetterli, Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates. IEEE Trans. Signal Process. **42**, 519–531 (1994)
13. N.T. Thao, M. Vetterli, Reduction of the MSE in R-times oversampled A/D conversion from $\mathcal{O}(R^{-1})$ to $\mathcal{O}(R^{-2})$. IEEE Trans. Signal Process. **42**, 200–203 (1994)

Part XX
Algorithms and Representations

The chapters in this part cover the design and use of efficient representation methods for signals. In addition, algorithms to obtain such representations are also proposed in some of the following chapters. In particular, the topics covered in this part include the application of frames in the processing of psychoacoustic signals, the construction of a special class of wavelets filter banks, and a Fast Fourier Transform for approximating fractal signals.

In the first chapter, Peter Balazs, Nicki Holihaus, Thibaud Necciari, and Diana Steova give a survey of finite frame theory and filter banks. Subsequently, they proceed to show how finite frames and filter banks provide a very flexible framework for the processing of psychoacoustics signals. At the same time the chapter offers applied harmonic analysts a brief introduction to psychoacoustics signal processing.

In the second chapter, Youngmi Hur gives an overview of an algebraic geometry method for the construction of wavelet filter banks. In particular, she shows how the Quillen-Suslin Theorem together with the polyphase representation of a filter bank can be used to construct (redundant) wavelet filter banks. In the process, she also offers some algorithms for the construction of these filter banks.

Chapter three by Jan Ernst proposes a generic scheme for constructing signal representations that are quasi-invariant to perturbations of the domain. A motivation for this construction can be found in the invariance of topological properties of sets under homeomorphisms. The chapter also includes computational methods for applications of this construction to problems in image processing and computer vision.

In the final chapter of this part, Calvin Hotchkiss and Eric S. Weber consider signals defined on finite approximations of a fractal generated by an iterated functions system. Using appropriately chosen sets of frequencies from a second iterated functions system, they obtain an orthonormal basis for signals defined on the finite approximations of the underlying fractal. They show that this orthonormal basis gives rise to a fractal analog of the classical Discrete Fourier Transform. As a result they develop a theory of a Fast Fourier Transform for signals defined on these finite approximations to the fractal set.

Frame Theory for Signal Processing in Psychoacoustics

Peter Balazs, Nicki Holighaus, Thibaud Necciari, and Diana Stoeva

Abstract This review chapter aims to strengthen the link between frame theory and signal processing tasks in psychoacoustics. On the one side, the basic concepts of frame theory are presented and some proofs are provided to explain those concepts in some detail. The goal is to reveal to hearing scientists how this mathematical theory could be relevant for their research. In particular, we focus on frame theory in a filter bank approach, which is probably the most relevant view point for audio signal processing. On the other side, basic psychoacoustic concepts are presented to stimulate mathematicians to apply their knowledge in this field.

Keywords ERB • Bark • Gammatone • Frame • Gabor frame • Masking Pattern • Amount of masking • Frame Operator • Perfect reconstruction • Large Time-Frequency Analysis Toolbox • (LTFAT) • Uniform filterbank • AUDlet • Irrelevance Filter • Frame multipliers • Alias • Analysis • Synthesis • Dual • Parseval frame • Z-transform • Impulse response

1 Introduction

In the fields of audio signal processing and hearing research, continuous research efforts are dedicated to the development of optimal representations of sound signals, suited for particular applications. However, each application and each of these two disciplines has specific requirements with respect to *optimality* of the transform.

For researchers in audio signal processing, an optimal signal representation should allow to extract, process, and re-synthesize relevant information, and avoid any useless inflation of the data, while at the same time being easily interpretable. In addition, although not a formal requirement, but being motivated by the fact that most audio signals are targeted at humans, the representation should take human auditory perception into account. Common tools used in signal processing are linear time-frequency analysis methods that are mostly implemented as filter banks.

P. Balazs (✉) • N. Holighaus • T. Necciari • D. Stoeva
Acoustics Research Institute, Austrian Academy of Sciences,
Wohllebengasse 12-14, 1040 Wien, Austria
e-mail: peter.balazs@oeaw.ac.at; nicki.holighaus@oeaw.ac.at; thibaud.necciari@oeaw.ac.at; diana.stoeva@oeaw.ac.at

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4_10

For hearing scientists, an optimal signal representation should allow to extract the perceptually relevant information in order to better understand sound perception. In other terms, the representation should reflect the peripheral "internal" representation of sounds in the human auditory system. The tools used in hearing research are computational models of the auditory system. Those models come in various flavors but their initial steps in the analysis process usually consist in several parallel bandpass filters followed by one or more nonlinear and signal-dependent processing stages. The first stage, implemented as a (linear) filter bank, aims to account for the spectro-temporal analysis performed in the cochlea. The subsequent nonlinear stages aim to account for the various nonlinearities that occur in the periphery (e.g., cochlear compression) and at more central processing stages of the nervous system (e.g., neural adaptation). A popular auditory model, for instance, is the compressive gammachirp filter bank (see Sec. 2.2). In this model, a linear prototype filter is followed by a nonlinear and level-dependent compensation filter to account for cochlear compression. Because auditory models are mostly intended as perceptual analysis tools, they do not feature a synthesis stage, i.e. they are not necessarily invertible. Note that a few models do allow for an approximate reconstruction, though.

It becomes clear that filter banks play a central role in hearing research and audio signal processing alike, although the requirements of the two disciplines differ. This divergence of the requirements, in particular the need for signal-dependent nonlinear processing in auditory models, may contrast with the needs of signal processing applications. But even within each of those fields, demands for the properties of transforms are diverse, as becoming evident by the many already existing methods. Therefore, it can be expected that the perfect signal representation, i.e. one that would have all desired properties for arbitrary applications in one or even both fields, does not exist.

This manuscript demonstrates how *frame theory* can be considered a particularly useful *conceptual* background for scientists in both hearing and audio processing, and presents some first motivating applications. Frames provide the following general properties: *perfect reconstruction*, *stability*, *redundancy*, and a *signal-independent, linear inversion procedure*. In particular, frame theory can be used to analyze any filter bank, thereby providing useful insight into its structure and properties. In practice, if a filter bank construction (i.e., including both the analysis and synthesis filter banks) satisfies the frame condition (see Sec. 4), it benefits from all the frame properties mentioned above. Why are those properties essential to researchers in audio signal processing and hearing science?

Perfect reconstruction property: With the possible exception of frequencies outside the audible range, a non-adaptive analysis filter bank, i.e. one that is general, not signal-dependent, has no means of determining and extracting exactly the perceptually relevant information. For such an extraction, signal-dependent information would be crucial. Therefore, the only way to ensure that a linear, signal-independent analysis stage[1], possibly followed by a nonlinear processing stage,

[1] As given by any fixed analysis filter bank.

captures all *perceptually relevant signal components* is to ensure that it does *not lose any* information at all. This, in fact, is *equivalent to being perfectly invertible*, i.e. having a perfect reconstruction property. Thus, this property benefits the user even when reconstruction is not intended per-se. Note that in general "being perfectly invertible" need not necessarily imply that a concrete inversion procedure is known. In the frame case, a constructive method exists, though.

Stability: For sound processing, stability is essential in the sense that, for the analysis stage, when two signals are similar (i.e., their difference is small), the difference between their corresponding analysis coefficients should also be small. For the synthesis stage, a signal reconstructed from slightly distorted coefficients should be relatively close to the original signal, that is the one reconstructed from undistorted coefficients. From an energy point of view, signals which are similar in energy should provide analysis coefficients whose energy is also similar. So the respective energies remain roughly proportional. In particular, considering a signal mixture, the combination of stability and linearity ensures that every signal component is represented and weighted according to its original energy. In other terms, individual signal components are represented proportional to their energy, which is very important for, e.g., visualization. Even in a perceptual analysis, where inaudible components should not be visualized equally to audible components having the same energy, this stability property is important. To illustrate this, recall that the nonlinear post-processing stages in auditory models are signal dependent. That is, also the inaudible information can be essential to properly characterize the nonlinearity. For instance, consider again the setup of the *compressive gammachirp* model where an intermediate representation is obtained through the application of a linear analysis filter bank to the input signal. The result of this linear transform determines the shape of the subsequent nonlinear compensation filter. Note that the *whole* intermediate representation is used. Consequently, the proper estimation of the nonlinearity crucially relies on the signal representation being accurate, i.e. *all* signal components being represented and appropriately weighted. This *accurateness* comes for free if the analysis filter bank forms a frame.

Signal-independent, linear inversion: A consistent (i.e., signal-independent) inversion procedure is of great benefit in signal processing applications. It implies that a single algorithm/implementation can perform all the necessary synthesis tasks. For nonlinear representations, finding a signal-independent procedure which provides a stable reconstruction is a highly nontrivial affair, if it is at all possible. With linear representations, such a procedure is easier to determine and this can be seen as an advantage of the linearity. The linearity provided by the reconstruction algorithm also significantly simplifies separation tasks. In a linear representation, a separation in the coefficient (time-frequency) domain, i.e. before synthesis, is equivalent to a separation in the signal domain. Such a property is highly relevant, for instance, to computational auditory scene analysis systems that, to some extent, are sound source separators (see Sec. 2.4).

Redundancy: Representations which are sampled at critical density are often unsuitable for visualization, since they lead to a low resolution, which may lead to many distinct signal components being integrated into a single coefficient of the transform. Thus, the individual coefficients may contain information from a

lot of different sources, which makes them hard to interpret. Still, the whole set of coefficients captures all the desired signal information if (and only if) the transform is invertible. Redundancy provides higher resolution and so components that are separated in time or in frequency can be separated in the transform domain. Furthermore, redundant representations are smoother and therefore easier to read than their critically sampled counterparts.

Moreover, redundant representations provide some resistance against noise and errors. This is in contrast to non-redundant systems, where distortions cannot be compensated for. This is used for de-noising approaches. In particular, if a signal is synthesized in a straightforward way from noisy (redundant) coefficients, the synthesis process has the tendency to reduce the energy of the noise, i.e. there is some noise cancellation.

Besides the above properties, which are direct consequences of the frame inequalities, the generality of frame theory enables the consideration of *additional important properties*. In the setting of perceptually motivated audio signal analysis and processing, these include:

Perceptual relevance: We have stressed that the only way to ensure that all perceptually relevant information is kept is to accurately capture all the information by using a stable and perfectly invertible system for analysis. However, in an auditory model or in perceptually motivated signal processing, perceptually irrelevant components should be discarded at some point. If only a linear signal processing framework is desired, this can be achieved by applying a perceptual weighting[2] and a masking model, see Sec. 2. If a nonlinear auditory model like the compressive gammachirp filter bank is used, recall that the nonlinear stage is mostly determined by the coefficients at the output of the linear stage. Therefore, all information should be kept up to the nonlinear stage. In other words, discarding information already in the analysis stage might falsify the estimation of the nonlinear stage, thereby resulting in an incorrect perceptual analysis. We want to stress here the importance of being able to *selectively* discard unnecessary information, in contrast to information being *involuntarily lost* during the analysis and/or synthesis procedures.

A flexible signal processing framework: All stable and invertible filter banks form a frame and therefore benefit from the frame properties discussed above. In addition, using filter banks that are frames allows for flexibility. For instance, one can gradually tune the signal representation such as the *time-frequency resolution*, analysis filters' *shape* and *bandwidth*, *frequency scale*, *sampling density* etc., while at the same time retaining the crucial frame properties. It can be tremendously useful to provide a single and adaptable framework that allows to switch model parameters and/or transition between them. By staying in the common general setting of filter bank frames, the linear filter bank analysis in an auditory model or signal processing scheme can be seen as an exchangeable, practically self-contained block in the scheme. Thus, the filter bank parameters, e.g. those mentioned before, can be tuned by scientists according to their preference, without the need to redesign the

[2]Different frequency ranges are given varying importance in the auditory system

remainder of the model/scheme. Such a common background leads to results being more comparable across research projects and thus benefits not only the individual researcher, but also the whole field. Two main advantages of a common background are the following: first, the properties and parameters of various models can be easily interpreted and compared across contributions; second, by the adaption of a linear model to obtain a nonlinear model the new model parameters remain interpretable.

Ease of integration: Filter banks are already a common tool in both hearing science and signal processing. Integrating a filter bank frame into an existing analysis/processing framework will often only require minor modifications of existing approaches. Thus, frames provide a theoretically sound foundation without the need to fundamentally re-design the remainder of your analysis (or processing) framework.

In some cases, you might already implicitly use frames without knowing it. In that case, we provide here the conceptual background necessary to unlock the full potential of your method.

The rest of this chapter is organized as follows: In Section 2, we provide basic information about the human auditory system and introduce some psychoacoustic concepts. In Section 3 we present the basics of frame theory providing the main definitions and a few crucial mathematical statements. In Section 4 we provide some details on filter bank frames. The chapter concludes with Section 5 where some examples are given for the application of frame theory to signal processing in psychoacoustics.

2 The auditory analysis of sounds

This section provides a brief introduction to the human auditory system. Important concepts that are relevant to the problems treated in this chapter are then introduced, namely auditory filtering and auditory masking. For a more complete description of the hearing organ, the interested reader is referred to, e.g., [32, 73].

2.1 Ear's anatomy

The human ear is a very sensitive and complex organ whose function is to transform pressure variations in the air into the percept of sound. To do so, sound waves must be converted into a form interpretable by the brain, specifically into neural action potentials. Figure 1 shows a simplified view of the ear's anatomy. Incoming sound waves are guided by the pinna into the ear canal and cause the eardrum to vibrate. Eardrum vibrations are then transmitted to the cochlea by three tiny bones that constitute the ossicular chain: the malleus, incus, and stapes. The ossicular chain acts as an impedance matcher. Its function is to ensure efficient transmission of pressure variations in the air into pressure variations in the fluids present in the cochlea. The cochlea is the most important part of the auditory system because it is where pressure variations are converted into neural action potentials.

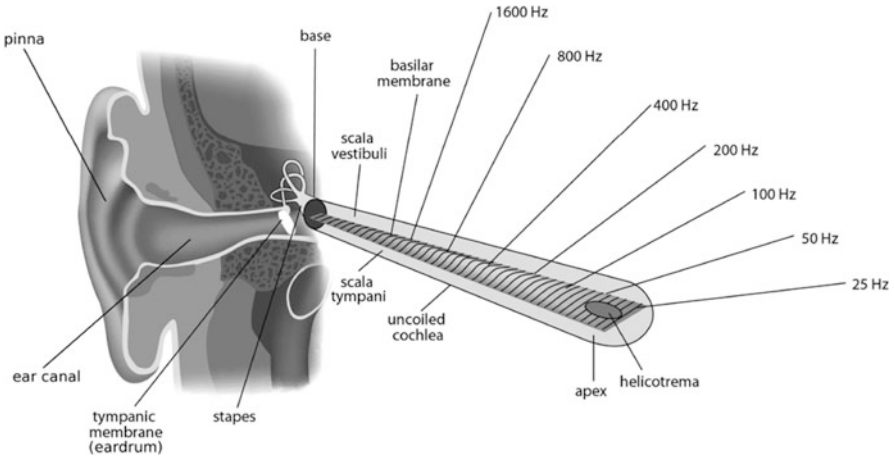


Fig. 1 Anatomy of the human ear with a schematic view of the unrolled cochlea. Adapted from [52].

The cochlea is a rolled-up tube filled with fluids and divided along its length by two membranes, the Reissner's membrane and basilar membrane (BM). A schematic view of the unrolled cochlea is shown in Figure 1 (the Reissner's membrane is not represented). It is the response of the BM to pressure variations transmitted through the ossicular chain that is of primary importance. Because the mechanical properties of the BM vary across its lengths (precisely, there is a gradation of stiffness from base to apex), BM stimulation results in a complex movement of the membrane. In case of a sinusoidal stimulation, this movement is described as a traveling wave. The position of the peak in the pattern of vibration depends on the frequency of the stimulation. High-frequency sounds produce maximum displacement of the BM near the base with little movement on the rest of the membrane. Low-frequency sounds rather produce a pattern of vibration which extends all the way along the BM but reaches a maximum before the apex. The frequency that gives the maximum response at a particular point on the BM is called the "characteristic frequency" (CF) of that point. In case of a broadband stimulation (e.g., an impulsive sound like a click), all points on the BM will oscillate. In short, the BM separates out the spectral components of a sound similar to a Fourier analyzer.

The last step of peripheral processing is the conversion of BM vibrations into neural action potentials. This is achieved by the inner hair cells that sit on top of the BM. There are about 3500 inner hair cells along the length of the cochlea ($\approx$35 mm in humans). The tip of each cell is covered with sensor hairs called stereocilia. The base of each cell directly connects to auditory nerve fibers. When the BM vibrates, the stereocilia are set in motion, which results in a bio-electrical process in the inner hair cells and, finally, in the initiation of action potentials in auditory nerve fibers. Those action potentials are then coded in the auditory nerve and conveyed to the central system where they are further processed to end up in a sound percept.

Because the response of auditory nerve fibers is also frequency specific and the action potentials vary over time, the "internal representation" of a sound signal in the auditory nerve can be likened to a time-frequency representation.

2.2 *The auditory filters concept*

Because of the frequency-to-place transformation (also called tonotopic organization) in the cochlea, and the transmission of time-dependent neural signals, the BM can be modeled in a first linear approximation as a bank of overlapping bandpass filters, named "critical bands" or "auditory filters." The center frequencies and bandwidth of the auditory filters, respectively, approximate the CF and width of excitation on the BM. Noteworthy, the width of excitation depends on level as well: patterns become wider and asymmetric as sound level increases (e.g., [37]). Several auditory filter models have been proposed based on the results from psychoacoustics experiments on masking (see, e.g., [59] and Sec. 2.3). A popular auditory filter model is the gammatone filter [71] (see Figure 2). Although gammatone filters

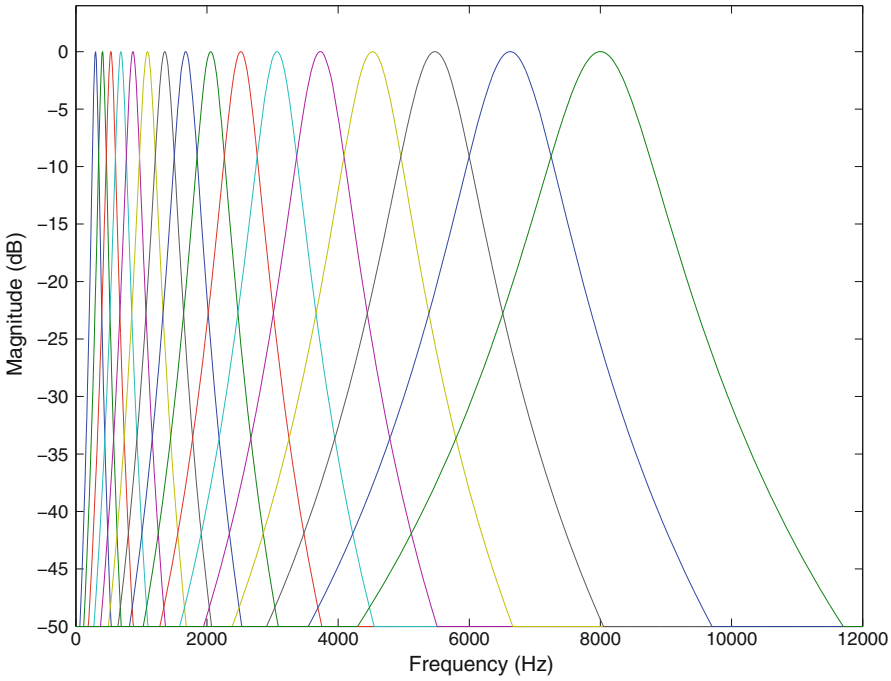


Fig. 2 A popular auditory filter model: the gammatone filter bank. The magnitude responses (in dB) of 16 gammatone filters in the frequency range 300–8000 Hz are represented on a linear frequency scale.

do not capture the level dependency of the actual auditory filters, their ease of implementation made them popular in audio signal processing (e.g., [90, 96]). More realistic auditory filter models are, for instance, the roex and gammachirp filters [37, 88]. Other level-dependent and more complex auditory filter banks include, for example, the dual resonance nonlinear filter bank [58] or the dynamic compressive gammachirp filter bank [49]. The two approaches in [49, 58] feature a linear filter bank followed by a signal-dependent nonlinear stage. As mentioned in the introduction, this is a particular way of describing a nonlinear system by modifying a linear system. Finally, it is worth noting that besides psychoacoustic-driven auditory models, mathematically founded models of the auditory periphery have been proposed. Those include, for instance, the wavelet auditory model [12] or the "EarWig" time-frequency distribution [67].

The bandwidth of the auditory filters has been determined based on psychoacoustic experiments. The estimation of bandwidth based on loudness perception experiments gave rise to the concept of Bark bandwidth defined by [98]

$$BW_{\text{Bark}} = 25 + 75\left(1 + 1.4 \times 10^{-6}\xi^2\right)^{0.69} \tag{1}$$

where ξ denotes the frequency and BW denotes the bandwidth, both in Hz. Another popular concept is the equivalent rectangular bandwidth (ERB), that is the bandwidth of a rectangular filter having the same peak output and energy as the auditory filter. The estimations of ERBs are based on masking experiments. The ERB is given by [37]

$$BW_{\text{ERB}} = 24.7 + \frac{\xi}{9.265}. \tag{2}$$

BW_{Bark} and BW_{ERB} are commonly used in psychoacoustics and signal processing to approximate the auditory spectral resolution at low to moderate sound pressure levels (i.e., 30–70 dB) where the auditory filters' shape remains symmetric and constant. See, for example, [37, 88] for the variation of BW_{ERB} with level.

Based on the concepts of Bark and ERB bandwidths, corresponding frequency scales have been proposed to represent and analyze data on a scale related to perception. To describe the different mappings between the linear frequency domain and the nonlinear perceptual domain we introduce the function $F_{\text{AUD}} : \xi \to \text{AUD}$ where AUD is an auditory unit that depends on the scale. The Bark scale is [98]

$$F_{\text{Bark}}(\xi) = 13\arctan(0.00076\xi) + 3.5\arctan(\xi/7500)^2 \tag{3}$$

and the ERB scale is [37]

$$F_{\text{ERB}}(\xi) = 9.265\ln\left(1 + \frac{\xi}{228.8455}\right). \tag{4}$$

Both auditory scales are connected to the ear's anatomy. One AUD unit indeed corresponds to a constant distance along the BM. 1 Bark corresponds to 1.3 mm [32] while 1 ERB corresponds to 0.9 mm [37, 38].

2.3 Auditory masking

The phenomenon of masking is highly related to the spectro-temporal resolution of the ear and has been the focus of many psychoacoustics studies over the last 70 years. Auditory masking refers to the increase in the detection threshold of a sound signal (referred to as the "target") due to the presence of another sound (the "masker"). Masking is quantified by measuring the detection thresholds of the target in presence and absence of the masker; the difference in thresholds (in dB) thus corresponds to the *amount of masking*. In the literature, masking has been extensively investigated in the spectral or temporal domain. The results were used to develop models of spectral or temporal masking that are currently implemented in audio applications like perceptual coding (e.g., [70, 76]) or sound processing (e.g., [7, 41]. Only a few studies investigated masking in the joint time-frequency domain. We present below some typical psychoacoustic results on spectral, temporal, and spectro-temporal masking. For more results and discussion on the origins of masking, the interested reader is referred to, e.g., [32, 62, 64].

In the following, we denote by $\xi_{\{M,T\}}$, $D_{\{M,T\}}$, and $L_{\{M,T\}}$ the frequency, duration, and level, respectively, of masker or target. Those signal parameters are fixed by the experimenter, i.e. they are known. The frequency shift between masker and target is $\Delta\xi = \xi_T - \xi_M$ and the time shift ΔT is defined as the onset delay between masker and target. Finally, AM denotes the amount of masking in dB.

2.3.1 Spectral masking

To study spectral masking, masker and target are presented simultaneously (since usually $D_M > D_T$, this is equivalent to saying that $0 \leq \Delta T < D_M - D_T$) and $\Delta\xi$ is varied. There are two ways to vary $\Delta\xi$, either fix ξ_T and vary ξ_M or vice versa. Similarly, one can fix L_M and vary L_T or vice versa. In short, various types of masking curves can be obtained depending on the signal parameters. A common spectral masking curve is a masking pattern that represents L_T or AM as a function of ξ_T or $\Delta\xi$ (see Figure 3). To measure masking patterns, ξ_M and L_M are fixed and AM is measured for various $\Delta\xi$. Under the assumption that $AM(\xi_T)$ corresponds to a certain ratio of masker-to-target energy at the output of the auditory filter centered at ξ_T, masking patterns measure the responses of the auditory filters centered at the individual ξ_Ts. Thus, masking patterns can be used as indicator of the *spectral spread of masking* of the masker or, in other terms, the spread of excitation of the masker on the BM. This spectral spread can in turn be used to derive a masking threshold, as used, for example, in audio codecs [70]. See also Sec. 5.2.

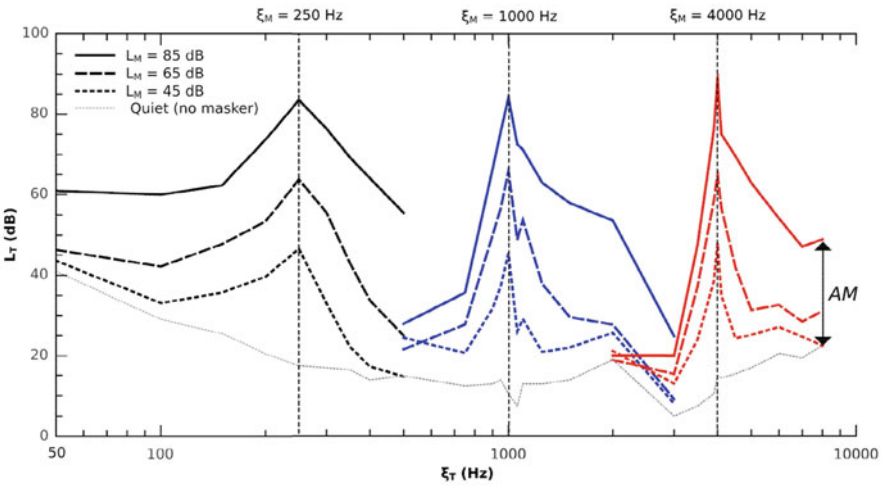


Fig. 3 Masking patterns for narrow-band noise maskers of different levels and frequencies. L_T (in dB SPL) is plotted as a function of ξ_T (in Hz) on a logarithmic scale. The gray dotted curve indicates the threshold in quiet. The difference between any of the colored curves and the gray curve thus corresponds to *AM*, as indicated by the arrow. Source: mean data for listeners JA and AO in [63, Experiment 3, Figs. 5–6].

Figure 3 shows typical masking patterns measured for narrow-band noise maskers of different levels ($L_M = 45$, 65, and 85 dB SPL, as indicated by the different lines) and frequencies ($\xi_M = 0.25$, 1, and 4 kHz, as indicated by the different vertical dashed lines). In this study, $D_M = D_T = 200$ ms. The masker was a 80-Hz-wide band of Gaussian noise centered at ξ_M. The target was also a 80-Hz band of noise centered at ξ_T. The main properties to be observed here are:

(i) For a given masker (i.e., a pair of ξ_M and L_M), *AM* is maximum for $\Delta\xi = 0$ and decreases as $|\Delta\xi|$ increases. This reflects the decay of masker excitation on the BM.
(ii) Masking patterns broaden with increasing level. This reflects the broadening of auditory filters with increasing level [37].
(iii) Masking patterns are broader at low than at high frequencies (see (1)–(2)). This reflects the fact that the density of auditory filters is higher at low than at high frequencies. Consequently, a masker with a given bandwidth will excite more auditory filters at low frequencies.

2.3.2 Temporal masking

By analogy with spectral masking, temporal masking is measured by setting $\Delta\xi = 0$ and varying ΔT. *Backward* masking is observed for $\Delta T < 0$, that is when the target precedes the masker in time. *Forward* masking is observed for $\Delta T \geq D_M$,

that is when the target follows the masker. Backward masking is hardly observed for ΔT < -20 ms and is mainly thought to result from attentional effects [32, 79]. In contrast, forward masking can be observed for $\Delta T \geq D_M + 200$ ms. Therefore, in the following we focus on forward masking.

Typical forward masking curves are represented in Figure 4. The left panel shows the effect of L_M for $\xi_M = \xi_T = 4$ kHz (mean data from [51]). In this study, masker and target were sinusoids ($D_M = 300$ ms, $D_T = 20$ ms). The main features to be observed here are (i) the temporal decay of forward masking is a linear function of $\log(\Delta T)$ and (ii) the rate of this decay strongly depends on L_M. The right panel shows the effect of D_M for $\xi_T = 2$ kHz and $L_M = 60$ dB SPL (mean data from [97]). In this study, the masker was a pulse of uniformly masking noise (i.e., a broad-band noise producing the same *AM* at all frequencies in the range 0–20 kHz, see [32]). The target was a sinusoid with $D_T = 5$ ms. It can be seen that the *AM* (i.e., the difference between the connected symbols and the star) at a given ΔT increases with increasing D_M, at least for $\Delta T - D_M < 100$ ms. Finally, a comparison of the two panels in Figure 4 for $L_M = 60$ dB indicates that, for $\Delta T - D_M \leq 50$ ms, the 300-ms sinusoidal masker (empty diamonds left) produces more masking than the 200-ms broad-band noise masker (empty squares right). Despite the difference in D_M, increasing the duration of the noise masker to 300 ms is not expected to account for the difference in *AM* of up to 20 dB observed here [32, 97].

2.3.3 Time-frequency masking

Only a few studies measured spectro-temporal masking patterns, that is ΔT and $\Delta\xi$ both systematically varied (e.g., [53, 79]). Those studies mostly involved long ($D_M \geq 100$ ms) sinusoidal maskers. In other words, those studies provide data on the time-frequency spread of masking for long and narrow-band maskers. In the context of time-frequency decompositions, a set of elementary functions, or "atoms," with good localization in the time-frequency domain (i.e., short and narrow-band) is usually chosen, see Sec. 3. To best predict masking in the time-frequency decompositions of sounds, it seems intuitive to have data on the time-frequency spread of masking for such elementary atoms, as this will provide a good match between the masking model and the sound decomposition. This has been investigated in [64]. Precisely, spectral, forward, and time-frequency masking have been measured using Gabor atoms of the form $s_i(t) = \sin(2\pi\xi_i t + \pi/4)e^{-\pi(\Gamma t)^2}$ with $\Gamma = 600$ s^{-1} as masker and target. According to the definition of Gabor atoms in (7), the masker was defined by $s_M(t) = \Im\{e^{i\pi/4} g_{\xi_M,0}\}$, where $\Im$ denotes the imaginary part, with a Gaussian window $\gamma(t) = e^{-\pi(\Gamma t)^2}$ and $\xi_M = 4$ kHz. The masker level was fixed at $L_M = 80$ dB. The target was defined by $s_T(t + \Delta T) = \Im\{e^{i(\pi/4+2\pi\xi_T\Delta T)}\gamma_{\xi_T,-\Delta T}\}$ with $\xi_T = \xi_M + \Delta\xi$. The set of time-frequency conditions measured in [64] is illustrated in Figure 5a. Because in this particular case we have $\xi_T\Delta T \in \mathbb{N}$, the target term reduces to $s_T(t + \Delta T) = \Im\{e^{i(\pi/4)}\gamma_{\xi_T,-\Delta T}\}$. The mean masking data are summarized in Figure 5b. These data, together with those collected by Laback

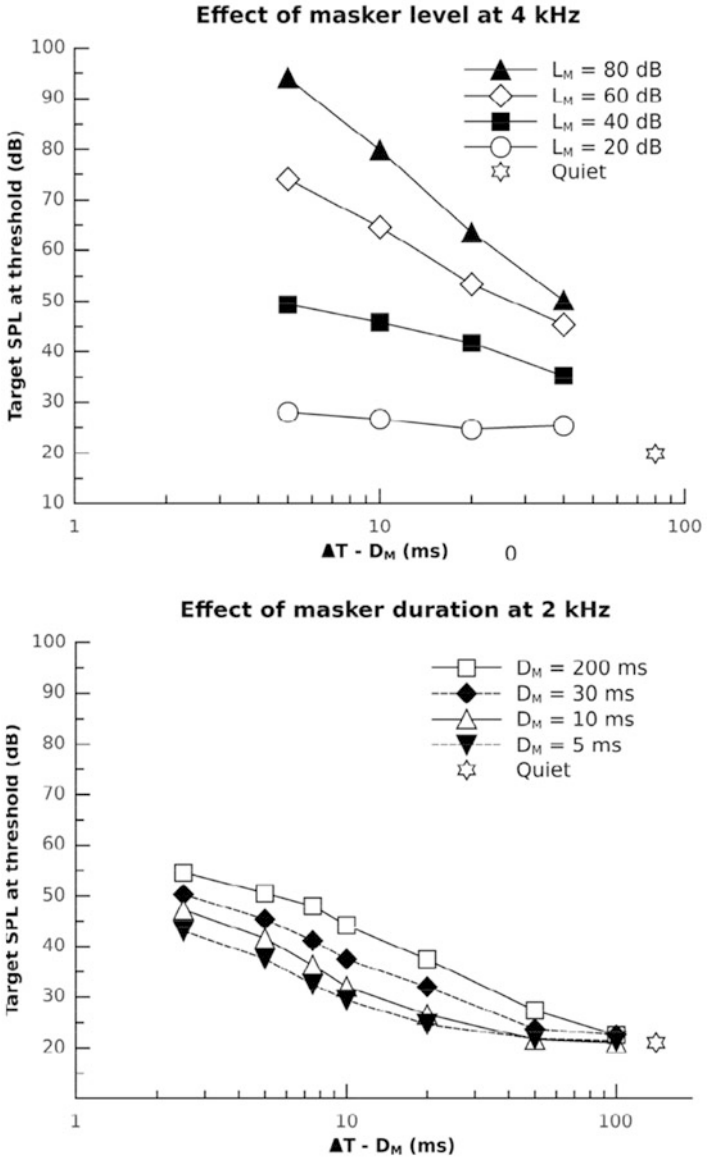


Fig. 4 Temporal (forward) masking curves for sinusoidal (left) and broadband noise maskers (right). L_T (in dB SPL) is plotted as a function of the temporal gap between masker offset and target onset, i.e. $\Delta T - D_M$ (in ms) on a logarithmic scale. Top panel: masking curves for various L_Ms and D_M = 300 ms (adapted from [51]). Bottom panel: masking curves for various D_Ms and L_M = 60 dB (adapted from [97]). Stars indicate the target thresholds in quiet.

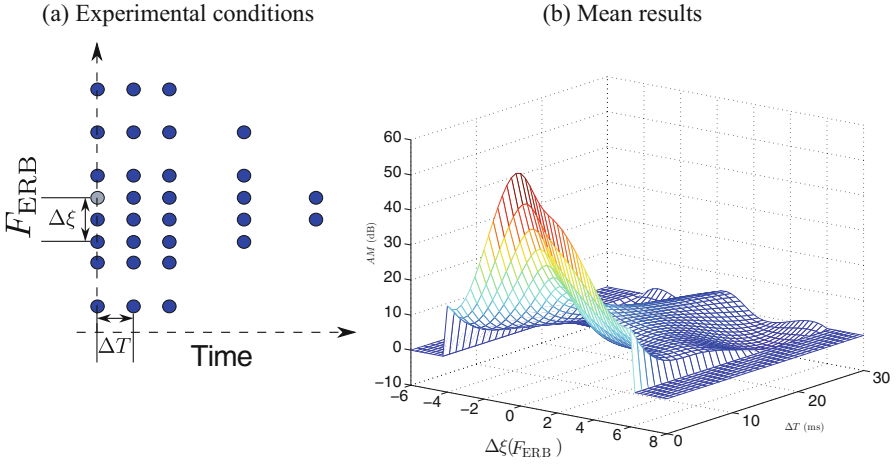


Fig. 5 (a) Conditions measured in [64] illustrated in the time-F_{ERB} plane. The gray circle symbolizes the masker atom $s_M(t)$. The blue circles symbolize the target atoms $s_T(t + \Delta T)$. The values of $\Delta\xi$ were -4, -2, -1, 0, +1, +2, +4, and +6 F_{ERB}. The values of ΔT were 0, 5, 10, 20, and 30 ms. (b) Mean data interpolated based on a cubic spline fit along the time-frequency plane. The ΔT axis was sampled at a step of 1 ms and the $\Delta\xi$ axis at a step of 0.25 F_{ERB}. For $\Delta\xi$ coordinates outside the range of measurements a value of $AM = 0$ was used.

et al on the additivity of spectral [56] and temporal masking [55] for the same Gabor atoms, constitute a crucial basis for the development of an accurate time-frequency masking model to be used in audio applications like audio coding or audio processing (see Sec. 5).

2.4 Computational auditory scene analysis

The term auditory scene analysis (ASA), introduced by Bregman [16], refers to the perceptual organization of auditory events into auditory streams. It is assumed that this perceptual organization constitutes the basis for the remarkable ability of the auditory system to separate sound sources, especially in noisy environments. A demonstration of this ability is the so-called cocktail party effect, i.e. when one is able to concentrate on and follow a single speaker in a highly competing background (e.g., many concurring speakers combined with cutlery and glass sounds). The term computational auditory scene analysis (CASA) thus refers to the study of ASA by computational means [92]. The CASA problem is closely related to the problem of source separation. Generally speaking, CASA systems can be considered as perceptually motivated sound source separators. The basic work flow of a CASA system is to first compute an auditory-based time-frequency transform (most systems use a gammatone filter bank, but any auditory representation that allows reconstruction

can be used, see Sec. 5.1). Second, some acoustic features like periodicity, pitch, amplitude, and frequency modulations are extracted so as to build the perceptive organization (i.e. constitute the streams). Then, stream separation is achieved using the so-called time-frequency masks. These masks are directly applied to the perceptual representation; they retain the "target" regions (mask = 1) and suppress the background (mask = 0). Those masks can be binary or real, see, e.g., [92, 96]. The target regions are then re-synthesized by applying the inverse transform to obtain the signal of interest. Noteworthy, a perfect reconstruction transform is of importance here. Furthermore, the linearity and stability of the transform allow a separation of the audio streams directly in the transform domain. Most gammatone filter banks implemented in CASA systems are only approximately invertible, though. This is due to the fact that such systems implement gammatone filters in the analysis stage and their time-reversed impulse responses in the synthesis stage. This setting implies that the frequency response of the gammatone filter bank has an all-pass characteristic and features no ripple (equivalently in the frame context, that the system is tight, see 4.3). In practice, however, gammatone filter banks usually consider only a limited range of frequencies (typically in the interval 0.1–4 kHz for speech processing) and the frequency response features ripples if the filters' density is not high enough. If a high density of filters is used, the audio quality of the reconstruction is rather good [85, 96]. Still, the quality could be perfect by using frame theory [66]. For instance, one could render the gammatone system tight (see Proposition 2) or use its dual frame (see Sec. 3.1.2).

The use of binary masks in CASA is directly motivated by the phenomenon of auditory masking explained above. However, time-frequency masking is hardly considered in CASA systems. As a final remark, an analogy can be established between the (binary) masks used in CASA and the concept of frame multipliers defined in Sec. 3.2. Specifically, the masks used in CASA systems correspond to the symbol m in (15). This analogy is not considered in most CASA studies, though, and offers the possibility for some future research connecting acoustics and frame multipliers.

3 Frame theory

What is an appropriate setting for the mathematical background of audio signal processing? Since real-world signals are usually considered to have finite energy and technically are represented as functions of some variable (e.g., time), it is natural to think about them as elements of the space $L^2(\mathbb{R})$. Roughly speaking, $L^2(\mathbb{R})$ contains all functions $x(t)$ with finite energy, i.e. with $\|x\|^2 = \int_{-\infty}^{+\infty} |x(t)|^2 \mathrm{dt} < \infty$. For working with sampled signals, the analogue appropriate space is $\ell^2(K)$ (K denoting a countable index set) which consists of the sequences $c = (c_k)_{k\in K}$ with finite energy, i.e. $\|c\|^2 = \sum_{k\in K} |c_k|^2 < \infty$.

Both spaces $L^2(\mathbb{R})$ and $\ell^2(K)$ are Hilbert spaces and one may use the rich theory ensured by the availability of an inner product, that serves as a measure of

correlation, and is used to define orthogonality, of elements in the Hilbert space. In particular, the inner product enables the representation of all functions in $\mathscr{H}$ in terms of their inner products with a set of reference functions: A standard approach for such representations uses orthonormal bases (ONBs), see, e.g., [42]. Every separable Hilbert space $\mathscr{H}$ has an ONB $(e_k)_{k\in K}$ and every element $x \in \mathscr{H}$ can be written as

$$x = \sum_{k\in K} \langle x, e_k\rangle e_k \tag{5}$$

with uniqueness of the coefficients $\langle x, e_k\rangle$, $k \in K$. The convenience of this approach is that there is a clear (and efficient) way for calculating the coefficients in the representations using the same orthonormal sequence. Even more, the energy in the coefficient domain (i.e., the square of the ℓ^2-*norm*) is exactly the energy of the element x:

$$\sum_{k\in K} |\langle x, e_k\rangle|^2 = \|x\|^2. \qquad \text{(Parseval equality)}$$

Furthermore, the representation (5) is stable: if the coefficients $(\langle x, e_k\rangle)_{k\in K}$ are slightly changed to $(a_k)_{k\in K} \in \ell^2$, one obtains an element $\widetilde{x} = \sum_{k\in K} a_k e_k$ close to the original one x.

However, the use of ONBs has several disadvantages. Often the construction of orthonormal bases with some given side constraints is difficult or even impossible (see below). "Small perturbation" of the orthonormal basis' elements may destroy the orthonormal structure [95]. Finally, the uniqueness of the coefficients in (5) leads to a lack of exact reconstruction when some of these coefficients are lost or disturbed during transmission.

This naturally leads to the question how the concept of ONBs could be generalized to overcome those disadvantages. As an extension of the above-mentioned Parseval equality for ONBs, one could consider inequalities instead of an equality, i.e. boundedness from above and below (see Def. 1). This leads to the concept of *frames*, which was introduced by Duffin and Schaeffer [29] in 1952. It took several decades for scientists to realize the importance and applicability of frames. Popularized around the 90s in the wake of wavelet theory [26, 27, 43], frames have seen increasing interest and extensive investigation by many researchers ever since. Frame theory is both a beautiful abstract mathematical theory and a concept applicable in many other disciplines, e.g., engineering, medicine, and psychoacoustics, see Sec. 5.

Via frames, one can avoid the restrictions of ONBs while keeping their important properties. Frames still allow perfect and stable reconstruction of all the elements of the space, though the representation-formulas in general are not as simple as the ones via an ONB (see Sec. 3.1.2). Compared to orthonormal bases, the frame property itself is much more stable under perturbations (see, e.g., [22, Sec. 15]). Also, in contrast to orthonormal bases, frames allow redundancy which is desirable,

e.g., in signal transmission, for reconstructing signals when some coefficients are lost, and for noise reduction. Via redundant frames one has multiple representations and this allows to choose appropriate coefficients fulfilling particular constraints, e.g. when aiming at sparse representations. Furthermore, frames can be easier and faster to construct than ONBs. Some advantageous side constraints can *only* be fulfilled for frames. For example, Gabor frames provide convenient and efficient signal processing tools, but good localization in both time and frequency can never be achieved if the Gabor frame is an ONB or even a Riesz basis (cf. Balian-Low Theorem, see, e.g., [22, Theor. 4.1.1]), while redundant Gabor frames for this purpose are easily constructed (for example, using the Gaussian function). See Sec. 2.3.3 on how good localization in time and frequency is important in masking experiments.

Some of the main properties of frames were already obtained in the first paper [29]. For extensive presentation on frame theory, we refer to [17, 22, 40, 42].

In this section we collect the basics of frame theory relevant to the topic of the current paper. All the statements presented here are well known. Proofs are given just to make the paper self-contained, for convenience of the readers, and to facilitate a better understanding of the mathematical concepts. They are mostly based on [22, 29, 40]. Throughout the rest of the section, $\mathscr{H}$ denotes a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$, $\mathrm{Id}_{\mathscr{H}}$ - the identity operator on $\mathscr{H}$, K - a countable index set, and Φ (resp. Ψ) - a sequence $(\phi_k)_{k\in K}$ (resp. $(\psi_k)_{k\in K}$) with elements from $\mathscr{H}$. The term *operator* is used for a linear mapping. Readers not familiar with Hilbert space theory can simply assume $\mathscr{H} = \mathbf{L}^2(\mathbb{R})$ for the remainder of this section.

3.1 Frames: A Mathematical viewpoint

The frame concept extends naturally the Parseval equality permitting inequalities, i.e., the ratio of the energy in the coefficient domain to the energy of the signal may be bounded from above and below instead of being necessarily one:

Definition 1 A countable sequence $\Phi = (\phi_k)_{k\in K}$ is called a *frame* for the Hilbert space $\mathscr{H}$ if there exist positive constants A and B such that

$$A \cdot \|x\|_{\mathscr{H}}^2 \leq \sum_{k\in K} |\langle x, \phi_k \rangle|^2 \leq B \cdot \|x\|_{\mathscr{H}}^2 , \ \forall\, x \in \mathscr{H}. \tag{6}$$

The constant A (resp. B) is called a *lower* (resp. *upper*) *frame bound* of Φ. A frame is called *tight with frame bound A* if A is both a lower and an upper frame bound. A tight frame with bound 1 is called a *Parseval frame*.

Clearly, every ONB is a frame, but not vice versa. Frames can naturally be split into two classes - the frames which still fulfill a basis-property, and the ones that do not:

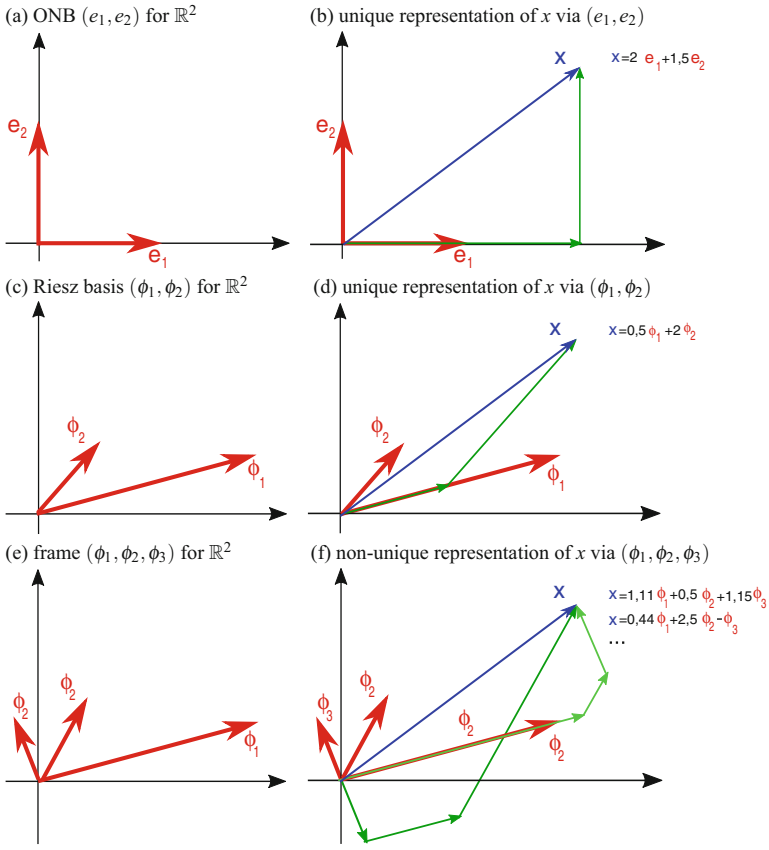


Fig. 6 Examples in $\mathbb{R}^2$: ONB (a,b), Riesz basis (c,d), frame (e,f)

Definition 2 A frame Φ for $\mathscr{H}$ which is a Schauder basis[3] for $\mathscr{H}$ is called a *Riesz basis* for $\mathscr{H}$. A frame for $\mathscr{H}$ which is not a Schauder basis for $\mathscr{H}$ is called *redundant* (also called *overcomplete*).

Note that Riesz bases were introduced by Bari [11] in different but equivalent ways. Riesz bases also extend ONBs, but contrary to frames, Riesz bases still have the disadvantages resulting from the basis-property, as they do not allow redundancy. For more on Riesz bases, see, e.g., [95]. As an illustration of the concepts of ONBs, Riesz bases, and redundant frames in a simple setting, consider examples in the Euclidean plane, see Figure 6.

[3] A sequence Φ is called a *Schauder basis* for $\mathscr{H}$ if every element $x \in \mathscr{H}$ can be written as $x = \sum_{k \in K} c_k \phi_k$ with unique coefficients $(c_k)_{k \in K}$.

Note that in a finite dimensional Hilbert space, considering only finite sequences, frames are precisely the complete sequences (see, e.g., [22, Sec. 1.1]), i.e., the sequences which span the whole space. However, this is not the case in infinite-dimensional Hilbert spaces - every frame is complete, but completeness is not sufficient to establish the frame property [29]. For results focused on frames in finite dimensional spaces, refer to [4, 19].

As non-trivial examples, let us mention a specific type of frames used often in signal processing applications, namely Gabor frames. A Gabor system is comprised of atoms of the form

$$g_{\omega,\tau}(t) = e^{2\pi \mathrm{i}\omega t} g(t-\tau), \tag{7}$$

with function $g \in L^2(\mathbb{R})$ called the *(generating) window* and with time- and frequency-shift $\tau, \omega \in \mathbb{R}$, respectively. To allow perfect and stable reconstruction, the Gabor system $(g_{\omega,\tau})_{(\omega,\tau)\in K(\subset\mathbb{R}^2)}$ is assumed to have the frame-property and in this case is called a *Gabor frame*. Note that the analysis operator of a Gabor frame corresponds to a *sampled Short-Time-Fourier transform* (see, e.g., [40]) also referred to as *Gabor transform*.

Most commonly, *regular Gabor frames* are used; these are frames of the form $(g_{k,l})_{k,l\in\mathbb{Z}} = \left(e^{2\pi \mathrm{i}kb\cdot} g(\cdot - la)\right)_{k,l\in\mathbb{Z}}$ for some positive a and b satisfying necessarily (but in general not sufficiently) $ab \leq 1$. To mention a concrete example - for the Gaussian $g(t) = e^{-t^2}$, the respective regular Gabor system $(g_{k,l})_{k,l\in\mathbb{Z}}$ is a frame for $L^2(\mathbb{R})$ if and only if $ab < 1$ (see, e.g., [40, Sec. 7.5] and the references therein).

Other possibilities include using alternative sampling structures, on subgroups [94] or irregular sets [18]. If the window is allowed to change with time (or frequency) one obtains the non-stationary Gabor transform [9]. There it becomes apparent that frames allow to create adaptive and adapted transforms [10], while still guaranteeing perfect reconstruction.

If not continuous but sampled signals are considered, Gabor theory works similarly. *Discrete Gabor frames* can be defined in an analogue way, namely, frames of the form $\left(e^{2\pi \mathrm{i}k/M\cdot} h[\cdot - la]\right)_{l\in\mathbb{Z},k=0,1,\ldots,M-1}$ for $h \in \ell^2(\mathbb{Z})$ with $a, M \in \mathbb{N}$, where $a/M \leq 1$ is necessary for the frame property. For readers interested in the theory of Gabor frames on $\ell^2(\mathbb{Z})$, see, e.g., [91]. For constructions of discrete Gabor frames from Gabor frames for $L^2(\mathbb{R})$ through sampling, refer to [50, 80].

3.1.1 Frame-related operators

Given a frame Φ for $\mathscr{H}$, consider the following linear mappings:

$$\begin{aligned}
&\textit{Analysis operator}: \quad \mathbf{C}_\Phi : \mathscr{H} \to l^2(K),\ \mathbf{C}_\Phi x := (\langle x, \phi_k\rangle)_{k\in K};\\
&\textit{Synthesis operator}: \mathbf{D}_\Phi : l^2(K) \to \mathscr{H},\ \mathbf{D}_\Phi (c_k)_{k\in K} := \textstyle\sum_{k\in K} c_k \phi_k;\\
&\textit{Frame operator}: \quad \mathbf{S}_\Phi : \mathscr{H} \to \mathscr{H}, \quad \mathbf{S}_\Phi x := \mathbf{D}_\Phi \mathbf{C}_\Phi x = \textstyle\sum_{k\in K} \langle x, \phi_k\rangle\, \phi_k.
\end{aligned} \tag{8}$$

These operators are tremendously important for the theoretical investigation of frames as well as for signal processing. As one can observe, the analysis (resp. synthesis, frame) operator corresponds to analyzing (resp. synthesizing, analyzing and re-synthesizing) a signal. In the following statement the main properties of the frame-related operators are listed.

Theorem 1 (e.g., [22, Sec. 5]) *Let Φ be a frame for $\mathscr{H}$ with frame bounds A and B ($A \leq B$). Then the following holds.*

(a) $\mathbf{C}_\Phi$ *is a bounded injective operator with bound* $\|\mathbf{C}_\Phi\| \leq \sqrt{B}$.
(b) $\mathbf{D}_\Phi$ *is a bounded surjective operator with bound* $\|\mathbf{D}_\Phi\| \leq \sqrt{B}$ *and* $\mathbf{D}_\Phi = \mathbf{C}_\Phi^*$.
(c) $\mathbf{S}_\Phi$ *is a bounded bijective positive self-adjoint operator with* $\|\mathbf{S}_\Phi\| \leq B$.
(d) $(\mathbf{S}_\Phi^{-1}\phi_k)_{k\in K}$ *is a frame for $\mathscr{H}$ with frame bounds* $1/B, 1/A$.

Proof (a) By the frame inequalities (6) we have $\sqrt{A}\|x\|_{\mathscr{H}} \leq \|\mathbf{C}_\Phi x\|_{\ell^2} \leq \sqrt{B}\|x\|_{\mathscr{H}}$ for every $x \in \mathscr{H}$; the upper inequality implies the boundedness and the lower one - the injectivity, i.e. the operator is one-to-one.

(b) First show that $\mathbf{D}_\Phi$ is well defined, i.e., that $\sum_{k\in K} c_k\phi_k$ converges for every $(c_k)_{k\in K} \in \ell^2(K)$. Without loss of generality, for simplicity of the writing, we may denote K as $\mathbb{N}$. Fix arbitrary $(c_k)_{k\in\mathbb{N}} \in \ell^2$. For every $p, q \in \mathbb{N}$, $p > q$,

$$\begin{aligned}
\|\sum_{k=1}^{p} c_k\phi_k - \sum_{k=1}^{q} c_k\phi_k\|_{\mathscr{H}} &= \sup_{x\in\mathscr{H},\|x\|_{\mathscr{H}}=1} |\langle \sum_{k=q+1}^{p} c_k\phi_k, x\rangle| \\
&\leq \sup_{x\in\mathscr{H},\|x\|_{\mathscr{H}}=1} (\sum_{k=q+1}^{p} |c_k|^2)^{1/2} (\sum_{k=q+1}^{p} |\langle\phi_k, x\rangle|^2)^{1/2} \\
&\leq \sqrt{B}\,(\sum_{k=q+1}^{p} |c_k|^2)^{1/2} \xrightarrow[p,q\to\infty]{} 0,
\end{aligned}$$

which implies that $\sum_{k=1}^{p} c_k\phi_k$ converges in $\mathscr{H}$ as $p \to \infty$. Using the adjoint of $\mathbf{C}_\Phi$, for every $(c_k)_{k=1}^{\infty} \in \ell^2$ and every $y \in \mathscr{H}$, one has that

$$\langle \mathbf{C}_\Phi^*(c_k)_{k=1}^{\infty}, y\rangle = \langle (c_k)_{k=1}^{\infty}, \mathbf{C}_\Phi y\rangle = \sum_{k=1}^{\infty} c_k\overline{\langle y, \phi_k\rangle} = \sum_{k=1}^{\infty} c_k\langle\phi_k, y\rangle = \langle\sum_{k=1}^{\infty} c_k\phi_k, y\rangle.$$

Therefore $\mathbf{D}_\Phi = \mathbf{C}_\Phi^*$, implying also the boundedness of D_Φ.

For every $x \in \mathscr{H}$, we have $\|\mathbf{D}_\Phi^* x\|_{\ell^2} = \|\mathbf{C}_\Phi x\|_{\ell^2} \geq \sqrt{A}\|x\|$, which implies (see, e.g., [78, Theorem 4.15]) that $\mathbf{D}_\Phi$ is surjective, i.e. it maps onto the whole space $\mathscr{H}$.

(c) The boundedness and self-adjointness of $\mathbf{S}_\Phi$ follow from (a) and (b). Since $\langle \mathbf{S}_\Phi x, x\rangle = \sum_{k\in K} |\langle x, \phi_k\rangle|^2$, $\mathbf{S}_\Phi$ is positive and the frame inequalities (6) mean that

$$A\|x\|_{\mathscr{H}}^2 \leq \langle \mathbf{S}_\Phi x, x\rangle \leq B\|x\|_{\mathscr{H}}^2, \forall x \in \mathscr{H}, \tag{9}$$

implying that $0 \leq \langle (\mathrm{Id}_{\mathscr{H}} - \frac{1}{B}\mathbf{S}_\Phi)x, x\rangle \leq \frac{B-A}{B}\|x\|^2_{\mathscr{H}}$ for all $x \in \mathscr{H}$. Then the norm of the bounded self-adjoint operator $\mathrm{Id}_{\mathscr{H}} - \frac{1}{B}\mathbf{S}_\Phi$ satisfies

$$\|\mathrm{Id}_{\mathscr{H}} - \frac{1}{B}\mathbf{S}_\Phi\| = \sup_{x\in\mathscr{H}, \|x\|_{\mathscr{H}}=1} \langle (\mathrm{Id}_{\mathscr{H}} - \frac{1}{B}\mathbf{S}_\Phi)x, x\rangle \leq \frac{B-A}{B} < 1,$$

which by the Neumann theorem (see, e.g., [45, Theor. 8.1]) implies that $\mathbf{S}_\Phi$ is bijective.

(d) As a consequence of (c), S_Φ^{-1} is bounded, self-adjoint, and positive. In the language of partial ordering of self-adjoint operators (see, e.g., [45, Sec. 68]), (9) can be written as

$$A \cdot \mathrm{Id}_{\mathscr{H}} \leq \mathbf{S}_\Phi \leq B \cdot \mathrm{Id}_{\mathscr{H}}. \tag{10}$$

Since $\mathbf{S}_\Phi^{-1}$ is positive and commutes with $\mathbf{S}_\Phi$ and $\mathrm{Id}_{\mathscr{H}}$, one can multiply the inequalities in (10) with $\mathbf{S}_\Phi^{-1}$ (see, e.g., [45, Prop. 68.9]) and obtain

$$\frac{1}{B}\mathrm{Id}_{\mathscr{H}} \leq \mathbf{S}_\Phi^{-1} \leq \frac{1}{A}\mathrm{Id}_{\mathscr{H}},$$

which means that

$$\frac{1}{B}\|x\|^2_{\mathscr{H}} \leq \langle \mathbf{S}_\Phi^{-1}x, x\rangle \leq \frac{1}{A}\|x\|^2_{\mathscr{H}}, \ \forall x \in \mathscr{H}. \tag{11}$$

For every $x \in \mathscr{H}$, denote $y_x = \mathbf{S}_\Phi^{-1}x$ and use the fact that $\mathbf{S}_\Phi^{-1}$ is self-adjoint to obtain

$$\sum_{k\in K} |\langle x, \mathbf{S}_\Phi^{-1}\phi_k\rangle|^2 = \sum_{k\in K} |\langle y_x, \phi_k\rangle|^2 = \langle y_x, \mathbf{S}_\Phi y_x\rangle = \langle \mathbf{S}_\Phi^{-1}x, x\rangle.$$

Now (11) completes the conclusion that $(\mathbf{S}_\Phi^{-1}\phi_k)_{k\in K}$ is a frame for $\mathscr{H}$ with frame bounds $1/B$, $1/A$. □

3.1.2 Perfect reconstruction via frames

Here we consider one of the most important properties of frames, namely, the possibility to have perfect reconstruction of all the elements in the space.

Theorem 2 (e.g., [40, Corol. 5.1.3]) *Let Φ be a frame for $\mathscr{H}$. Then there exists a frame Ψ for $\mathscr{H}$ such that*

$$x = \sum_{k\in K} \langle x, \psi_k\rangle \phi_k = \sum_{k\in K} \langle x, \phi_k\rangle \psi_k, \ \forall x \in \mathscr{H}. \tag{12}$$

Proof By Theorem 1(d), the sequence $(\mathbf{S}_\Phi^{-1}\phi_k)_{k\in K}$ is a frame for $\mathscr{H}$. Take $\Psi := (\mathbf{S}_\Phi^{-1}\phi_k)_{k\in K}$. Using the boundedness and the self-adjointness of S_Φ, for every $x \in \mathscr{H}$,

$$\sum_{k\in K}\langle x,\phi_k\rangle\psi_k = \sum_{k\in K}\langle x,\phi_k\rangle S_\Phi^{-1}\phi_k = S_\Phi^{-1}\sum_{k\in K}\langle x,\phi_k\rangle\phi_k = S_\Phi^{-1}S_\Phi x = x,$$

$$\sum_{k\in K}\langle x,\psi_k\rangle\phi_k = \sum_{k\in K}\langle x,S_\Phi^{-1}\phi_k\rangle\phi_k = \sum_{k\in K}\langle S_\Phi^{-1}x,\phi_k\rangle\phi_k = S_\Phi S_\Phi^{-1}x = x.$$

□

Let Φ be a frame for $\mathscr{H}$. Any frame Ψ for $\mathscr{H}$, which satisfies (12), is called a *dual frame* of Φ. By the above theorem, every frame has at least one dual frame, namely, the sequence

$$(S_\Phi^{-1}\phi_k)_{k\in K}, \tag{13}$$

called the *canonical dual of* Φ. When the frame is a Riesz basis, then the coefficient representation is unique and thus there is only one dual frame, the canonical dual. When the frame is redundant, then there are other dual frames different from the canonical dual (see, e.g., [22, Lemma 5.6.1]), even infinitely many. This provides multiple choices for the coefficients in the frame representations, which is desirable in some applications (see, e.g., [10]). The canonical dual has a minimizing property in the sense that the coefficients $(\langle x,S_\Phi^{-1}\phi_k\rangle)_{k\in K}$ in the representation $x = \sum_{k\in K}\langle x,S_\Phi^{-1}\phi_k\rangle\phi_k$ have the minimal ℓ^2-norm compared to the coefficients $(c_k)_{k\in K}$ in all other possible representations $x = \sum_{k\in K} c_k\phi_k$. However, for certain applications other constraints are of interest - e.g. sparsity, efficient algorithms for representations or particular shape restrictions on the dual window [72, 93]. The canonical dual is not always efficient to calculate nor does it always have the desired structure; in such cases other dual frames are of interest [15, 23, 57]. The particular case of tight frames is very convenient for efficient reconstructions, because the canonical dual is simple and does not require operator-inversion:

Corollary 1 (e.g. [22, Sec. 5.7]) *The canonical dual of a tight frame* $(\phi_k)_{k\in K}$ *with frame bound A is the sequence* $(\frac{1}{A}\phi_k)_{k\in K}$.

Proof Let Φ be a tight frame for $\mathscr{H}$ with frame bound A. It follows from (10) that $\mathbf{S}_\Phi = A\cdot \mathrm{Id}_{\mathscr{H}}$ and thus the canonical dual of Φ is $(\mathbf{S}_\Phi^{-1}\phi_k)_{k\in K} = (\frac{1}{A}\phi_k)_{k\in K}$. □

In acoustic applications, it can be of big advantage to not be forced to distinguish between analysis and synthesis atoms. So, one may aim to do analysis and synthesis with the same sequence as an analogue to the case with ONBs. However, such an analysis-synthesis strategy would perfectly reconstruct all the elements of the space if and only if this sequence is a Parseval frame:

Proposition 1 (e.g., [22, Lemma 5.7.1]) *The sequence Φ satisfies*

$$x = \sum_{k\in K} \langle x, \phi_k\rangle \phi_k, \ \forall x \in \mathscr{H}, \tag{14}$$

if and only it is a Parseval frame for $\mathscr{H}$.

Proof Let Φ be a Parseval frame for $\mathscr{H}$. By Corollary 1, the canonical dual of Φ is the same sequence Φ, which implies that (14) holds. Now assume that (14) holds. Then for every $x \in \mathscr{H}$,

$$\|x\|^2 = \langle \sum_{k\in K} \langle x, \phi_k\rangle \phi_k, x\rangle = \sum_{k\in K} \langle x, \phi_k\rangle \langle \phi_k, x\rangle = \sum_{k\in K} |\langle x, \phi_k\rangle|^2,$$

which means that Φ is a Parseval frame for $\mathscr{H}$. □

The above statement characterizes the sequences which provide reconstructions exactly like ONBs - these are precisely the Parseval frames. A trivial example of such a frame which is not an ONB is the sequence $(e_1, e_2/\sqrt{2}, e_2/\sqrt{2}, e_3/\sqrt{3}, e_3/\sqrt{3}, e_3/\sqrt{3}, \ldots)$, where $(e_k)_{k=1}^{\infty}$ denotes an ONB for $\mathscr{H}$. Clearly, any tight frame with frame bound A is easily converted into a Parseval frame by dividing the frame elements by the square root of A. Given any frame, one can always construct a Parseval frame as follows:

Proposition 2 (e.g. [22, Theor. 5.3.4]) *Let Φ be a frame for $\mathscr{H}$. Then $\mathbf{S}_\Phi^{-1}$ has a positive square root and $(\mathbf{S}_\Phi^{-1/2}\phi_k)_{k\in K}$ forms a Parseval frame for $\mathscr{H}$.*

Proof Since $\mathbf{S}_\Phi^{-1}$ is a bounded positive self-adjoint operator, there is a unique bounded positive self-adjoint operator, which is denoted by $\mathbf{S}_\Phi^{-1/2}$, with $\mathbf{S}_\Phi^{-1} = \mathbf{S}_\Phi^{-1/2}\mathbf{S}_\Phi^{-1/2}$. Furthermore, $\mathbf{S}_\Phi^{-1/2}$ commutes with $\mathbf{S}_\Phi$. For every $x \in \mathscr{H}$,

$$\sum_{k\in K} \langle x, \mathbf{S}_\Phi^{-1/2}\phi_k\rangle \mathbf{S}_\Phi^{-1/2}\phi_k = \mathbf{S}_\Phi^{-1/2} \sum_{k\in K} \langle \mathbf{S}_\Phi^{-1/2}x, \phi_k\rangle \phi_k = \mathbf{S}_\Phi^{-1/2}\mathbf{S}_\Phi\mathbf{S}_\Phi^{-1/2}x = \mathbf{S}_\Phi^{-1}\mathbf{S}_\Phi x = x.$$

By Proposition 1 this means that $(\mathbf{S}_\Phi^{-1/2}\phi_k)_{k\in K}$ is a Parseval frame for $\mathscr{H}$. □

Finally, note that frames guarantee stability. Let Φ be a frame for $\mathscr{H}$ with frame bounds A, B. Then $\sqrt{A}\|x-y\| \le \|(\langle x, \phi_k\rangle) - (\langle y, \phi_k\rangle)_{k\in K}\|_{\ell^2} \le \sqrt{B}\|x-y\|$ for $x, y \in \mathscr{H}$, which implies that close signals lead to close analysis coefficients and vice versa. Furthermore, the representations via Φ and a dual frame Ψ is stable. If a signal x is transmitted via the coefficients $(\langle x, \psi_k\rangle)_{k\in K}$ but, during transmission, the coefficients are slightly disturbed (i.e., modified to a sequence $(a_k)_{k\in K} \in \ell^2$ with small ℓ^2-difference), then by Theorem 1(b), the reconstructed signal $y = \sum_{k\in K} a_k\phi_k$ will be close to x: $\|x-y\| = \|\sum_{k\in K}(\langle x, \psi_k\rangle - a_k)_{k\in K}\phi_k\| \le \sqrt{B}\|(\langle x, \psi_k\rangle - a_k)_{k\in K}\|_{\ell^2}$.

3.2 Frame multipliers

Multipliers have been used implicitly for quite some time in applications, as time-variant filters, see, e.g., [60]. The first systematic theoretical development of Gabor multipliers appeared in [33]. An extension of the multiplier concept to general frames in Hilbert spaces was done in [3] and it can be derived as an easy consequence of Theorem 1:

Proposition 3 [3] Let Φ and Ψ be frames for $\mathscr{H}$ and let $m = (m_k)_{k\in K}$ be a complex scalar sequence in $\ell^\infty(K)$. Then the series $\sum_{k\in K} m_k \langle x, \psi_k\rangle \phi_k$ converges for every $x \in \mathscr{H}$ and determines a bounded operator on $\mathscr{H}$.

Proof For every $x \in \mathscr{H}$, Theorem 1(a) implies that $(\langle x, \psi_k\rangle)_{k\in K} \in \ell^2$ and thus $(m_k\langle x, \psi_k\rangle)_{k\in K} \in \ell^2$, which by Theorem 1(b) implies that the series $\sum_{k\in K} m_k \langle x, \psi_k\rangle \phi_k$ converges. Thus, the mapping $\mathbf{M}_{m,\Phi,\Psi}$ determined by $\mathbf{M}_{m,\Phi,\Psi} x := \sum_{k\in K} m_k \langle x, \psi_k\rangle \phi_k$ is well defined on $\mathscr{H}$ and furthermore linear. For every $x \in \mathscr{H}$,

$$\begin{aligned} \|\mathbf{M}_{m,\Phi,\Psi} x\|_{\mathscr{H}} &= \|\mathbf{D}_\Phi (m_k\langle x, \psi_k\rangle)_{k\in K}\|_{\mathscr{H}} \le \|\mathbf{D}_\Phi\| \cdot \|(m_k\langle x, \psi_k\rangle)_{k\in K}\|_{\ell^2} \\ &\le \|\mathbf{D}_\Phi\| \cdot \|m\|_\infty \cdot \|\mathbf{C}_\Psi\| \cdot \|x\|_{\mathscr{H}}, \end{aligned}$$

implying the boundedness of $\mathbf{M}_{m,\Phi,\Psi}$. □

Due to the above proposition, frame multipliers can be defined as follows:

Definition 3 Given frames Φ and Ψ for $\mathscr{H}$ and given complex scalar sequence $m = (m_k)_{k\in K} \in \ell^\infty(K)$, the operator $M_{m,\Phi,\Psi}$ determined by

$$\mathbf{M}_{m,\Phi,\Psi} x := \sum_{k\in K} m_k \langle x, \psi_k\rangle \phi_k, \; x \in \mathscr{H}, \tag{15}$$

is called a *frame multiplier* with a *symbol* m.

Thus, frame multipliers extend the frame operator, allowing different frames for the analysis and synthesis step, and modification in between (for an illustration, see Figure 7). However, in contrast to frame operators, multipliers in general lose the bijectivity (as well as self-adjointness and positivity). For some applications it might be necessary to invert multipliers, which brings the interest to bijective multipliers and formulas for their inverses - for interested readers, we refer to [5, 82–84] for some investigation in this direction.

In the language of signal processing, Gabor filters [61] are a particular way to do time-variant filtering. In fact, Gabor filters are nothing but frame multipliers associated to a Gabor frame. A signal x is transformed to the time-frequency domain (with a Gabor frame Φ), then modified there by point-wise multiplication with the symbol m, followed by re-synthesis via some Gabor frame Ψ providing amodified

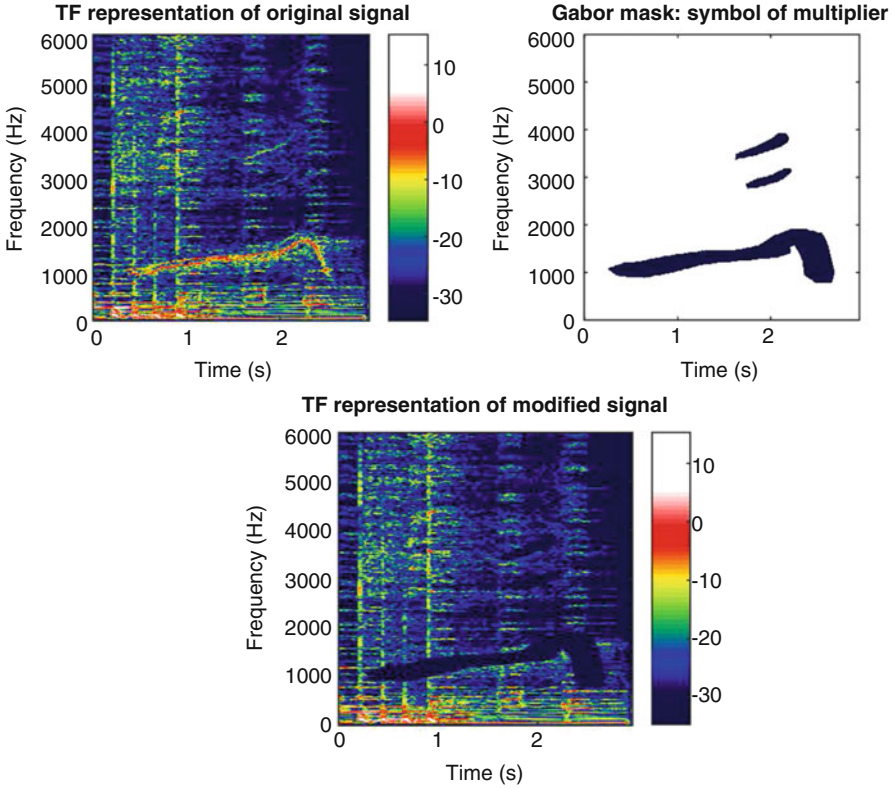


Fig. 7 *An illustrative example to visualize a multiplier (taken from [5]).* (TOP LEFT) The time-frequency representation of the music signal f. (TOP RIGHT) The symbol m, found by a (manual) estimation of the time-frequency region of the singer's voice. (BOTTOM) Time-frequency representation of $M_{m,\widetilde{\Psi},\Psi}f$.

signal. If some elements m_k of the symbol m are zero, the corresponding coefficients are removed, as sometimes used in applications like CASA or perceptual sparsity, see Secs. 2.4 and 5.2.

3.2.1 Implementation

In the finite-dimensional case, frames lend themselves easily to implementation in computer codes [4]. The Large Time-Frequency Analysis Toolbox (LTFAT) [81], see http://ltfat.github.io/, is an open-source Matlab/Octave toolbox intended for time-frequency analysis, synthesis and processing, including multipliers. It provides robust and efficient implementations for a variety of frame-related operators for generic frames and several special types, e.g. Gabor and filter bank frames.

In a recent release, reported in [74], a "frames framework" was implemented, which models the abstract frame concept in an object-oriented approach. In this setting any algorithm can be designed to use a general frame. If a structured frame, e.g. of Gabor or wavelet type, is used, more efficient algorithms are automatically selected.

4 Filter bank frames: a signal processing viewpoint

Linear time-invariant *filter banks (FB)* are a classical signal analysis and processing tool. Their general, potentially non-uniform structure provides the natural setting for the design of flexible, frequency-adaptive time-frequency signal representations [10]. In this section, we recall some basics of FB theory and consider the relation of perfect reconstruction FBs to certain frame systems.

4.1 Basics of filter banks

In the following, we consider discrete signals with finite energy ($x \in \ell^2(\mathbb{Z})$), interpreted as samples of a continuous signal, sampled at sampling frequency ξ_s, i.e. the signal was sampled every $1/\xi_s$ seconds. Bold italic letters indicate matrices (upper case), e.g. $\boldsymbol{G}$, and vectors (lower case), e.g. $\boldsymbol{h}$. We denote by $W_N = e^{2i\pi/N}$ the Nth root of unity and by $\delta_k = \delta_0[\cdot - k]$ the (discrete) Dirac symbol, with $\delta_k[n] = 1$ for $n = k$ and 0 otherwise. Observe that for $q = D/d$ we have

$$\sum_{l=0}^{q-1} W_D^{jld} = \sum_{l=0}^{q-1} e^{2\pi ijl/q} = \begin{cases} q \text{ if } j \text{ is a multiple of } q \\ 0 \text{ otherwise.} \end{cases} \tag{16}$$

The *z-transform* maps a *(discrete-)time domain* signal x to its *frequency domain* representation X by

$$\mathscr{Z} : x[n] \mapsto X(z) = \sum_{n\in\mathbb{Z}} x[n] z^n, \text{ for all } z \in \mathbb{C}.$$

By setting $z = e^{2\pi i\xi}$ for $\xi \in \mathbb{T}$, the z-transform equals the discrete-time Fourier transform (DTFT). Note that the z-transform is uniquely determined by its values on the complex unit circle [68]. It is easy to see that, $\mathscr{Z}(\delta_k) = z^k$, a property that we will use later on.

The application of a filter to a signal x is given by the convolution of x with the time domain representation, or *impulse response* $h \in \ell^2(\mathbb{Z})$ of the filter

$$y[n] = x * h[n] = \sum_{l\in\mathbb{Z}} x[l] h[n-l], \ \forall\, n \in \mathbb{Z}, \tag{17}$$

or equivalently by multiplication in the frequency domain $Y(z) = X(z)H(z)$, where $H(z)$ is the *transfer function*, or frequency domain representation, of the filter.

Furthermore define the *downsampling* and *upsampling* operators $\downarrow_d$, $\uparrow_d$ by

$$\downarrow_d \{x\}[n] = x[d \cdot n] \quad \text{and} \quad \uparrow_d \{x\}[n] = \begin{cases} x[n/d] & \text{if } n \in d\mathbb{Z}, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

Here, $d \in \mathbb{N}$ is called the *downsampling* or *upsampling factor*, respectively. In the frequency domain, the effect of down- and upsampling is the following [69]:

$$\mathscr{Z}(\downarrow_d \{x\})(z) = d^{-1} \sum_{j=0}^{d-1} X(W_d^j z^{1/d}) \quad \text{and} \quad \mathscr{Z}(\uparrow_d \{x\})(z) = X(z^d). \tag{19}$$

In words, downsampling a signal by d results in the dilation of its spectrum by d and the addition of $(d-1)$ copies of the dilated spectrum. These copies of the spectrum (the terms $X(W_d^j z^{1/d})$ for $j \neq 0$ in the sum above) are called *aliasing terms*. Conversely, upsampling a signal by d results in the contraction of its spectrum by d.

An FB is a collection of analysis filters $H_k(z)$, synthesis filters $G_k(z)$, and downsampling and upsampling factors d_k, $k \in \{0, \ldots, K\}$, see Figure 8. An FB is called *uniform*, if all filters have the same downsampling factor, i.e. $d_k = D$ for all k.

The sub-band components $y_k[n]$ of the system represented in Figure 8 are given in the time domain by

$$y_k[n] = \downarrow_{d_k} \{h_k * x\}[n] \tag{20}$$

The output signal is $\tilde{x}[n] = \sum_{k=0}^{K} (g_k * \uparrow_{d_k} \{y_k\})[n]$. When analyzing the properties of a filter (bank), it is often useful to transform the expression for $\tilde{x}$ to the frequency domain. First, apply the z-transform to the output of a single analysis/synthesis branch, obtaining

$$\mathscr{Z}(g_k * \uparrow_{d_k} \{y_k\})(z) = d_k^{-1} [X(W_{d_k}^0 z), \ldots, X(W_{d_k}^{d_k-1} z)] \begin{bmatrix} H_k(W_{d_k}^0 z) \\ \vdots \\ H_k(W_{d_k}^{d_k-1} z) \end{bmatrix} G_k(z), \tag{21}$$

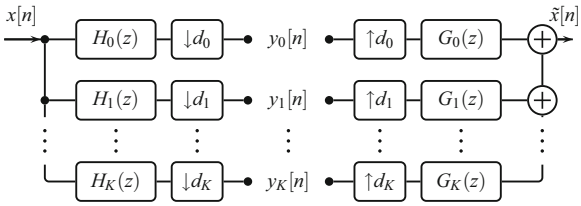


Fig. 8 General structure of a non-uniform analysis-synthesis FB.

where the down- and upsampling properties of the z-transform were applied, see Eq. (19). Now let $D = \operatorname{lcm}(d_0, \ldots, d_K)$, i.e. the least common multiple of the downsampling factors, and $D/d_k = q_k$. Then (21) gives

$$\mathscr{Z}\left(g_k * \uparrow_{d_k} \{y_k\}\right)(z) = D^{-1}[X(W_D^0 z), \ldots, X(W_D^{D-1} z)]\boldsymbol{h}_k(z)G_k(z), \tag{22}$$

where,

$$\boldsymbol{h}_k(z) = q_k \cdot \Big[H_k(z), \underbrace{0, \cdots, 0}_{q_k-1\text{zeros}}, H_k(W_D^{q_k} z), \underbrace{0, \cdots, 0}_{q_k-1\text{zeros}}, \cdots, H_k(W_D^{(d_k-1)q_k} z), \underbrace{0, \cdots, 0}_{q_k-1\text{zeros}}\Big]^T.$$

The relevance of this equality becomes clear if we use linearity of the z-transform to obtain a frequency domain representation of the full FB output, also called the *alias domain representation* [89]

$$\begin{aligned}
\tilde{X}(z) &= \sum_{k=0}^{K} \mathscr{Z}\left(g_k * \uparrow_{d_k} \{y_k\}\right)(z) \\
&= D^{-1}[X(W_D^0 z), \ldots, X(W_D^{D-1} z)] \left[\boldsymbol{h}_0(z), \ldots, \boldsymbol{h}_K(z)\right] \begin{bmatrix} G_0(z) \\ \vdots \\ G_K(z) \end{bmatrix} \\
&= D^{-1}[X(W_D^0 z), \ldots, X(W_D^{D-1} z)]\boldsymbol{H}(z)\boldsymbol{G}(z),
\end{aligned} \tag{23}$$

where $\boldsymbol{H}(z) = [\boldsymbol{h}_0(z), \ldots, \boldsymbol{h}_K(z)]$ is the $D \times (K+1)$ *alias component matrix* [89] and $\boldsymbol{G}(z) = [G_0(z), \ldots, G_K(z)]$.

An FB system is *undersampled, critically sampled, or oversampled*, if $R = \sum_{k=0}^{K} d_k^{-1}$ is smaller than, equal to, or larger than 1, respectively. Consequently, a uniform FB is critically sampled if it has exactly D subbands. For a deeper treatment of FBs, see, e.g., [54, 89].

Perfect reconstruction FBs: An FB is said to provide perfect reconstruction if $\tilde{x}[n] = x[n-l]$ for all $x \in \ell^2(\mathbb{Z})$ and some fixed $l \in \mathbb{Z}$. In the case when $l \neq 0$, the FB output is *delayed* by l. Using the alias domain representation of the FB, the *perfect reconstruction condition* can be expressed as

$$\boldsymbol{H}(z)\,\boldsymbol{G}(z) = z^l\,[D\,0 \cdots 0\,]^T, \tag{24}$$

for some $l \in \mathbb{Z}$, as this condition is equivalent to $\tilde{X}(z) = z^l X(z) = \mathscr{Z}(x * \delta_k)(z)$. From this vantage point the perfect reconstruction condition can be interpreted as all the alias components (i.e., from the 2nd to $D+1$-th) in $\boldsymbol{H}(z)$ being uniformly canceled over all $z \in \mathbb{C}$ by the synthesis filters $\boldsymbol{G}(z)$, while the first component of $\boldsymbol{H}(z)$ remains constant over all $z \in \mathbb{C}$ (up to a fixed power of z). The perfect reconstruction condition is of tremendous importance for determining whether an FB, including both analysis and synthesis steps, provides perfect reconstruction.

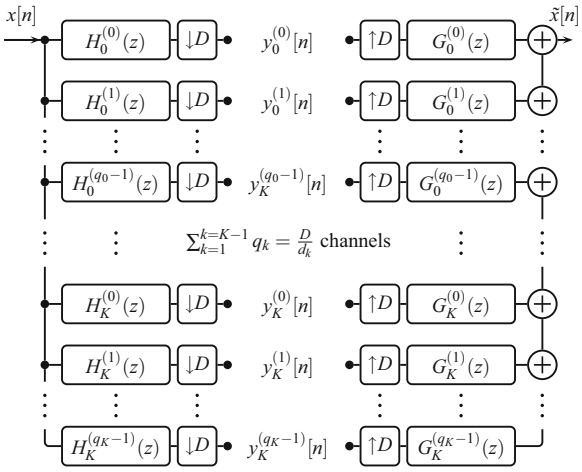


Fig. 9 The equivalent uniform FB [1] corresponding to the non-uniform FB in Figure 8. The terms $H_k^{(l)}$ and $G_k^{(l)}$ in (b) correspond to the z-transforms of the terms $h_k^{(l)}$ and $g_k^{(l)}$ defined in (25).

However, given a fixed analysis FB, the alias domain representation may fail to provide straightforward or efficient ways to find suitable synthesis filters that provide perfect reconstruction. It can sometimes be used to determine whether such a system can exist, although the process is far from intuitive [46]. Consequently, non-uniform perfect reconstruction FBs are still not completely investigated, and thus frame theory may provide valuable new insights. However, for uniform FBs the perfect reconstruction conditions have been largely treated in the literature [54, 89]. Therefore, before we indulge in the frame theory of FBs, we also show how a non-uniform FB can be decomposed into its equivalent uniform FB. Such a uniform equivalent of the FB always exists [1, 54] and can be obtained as shown in Figure 9 and described below.

4.2 The equivalent uniform filter bank

To construct the equivalent uniform FB to a general FB specified by analysis filters $H_k(z)$, synthesis filters $G_k(z)$, and downsampling and upsampling factors d_k, $k \in \{0, \ldots, K\}$, start by denoting again $D = \text{lcm}(d_0, \ldots, d_K)$. We first construct the desired uniform FB, before showing that it is in fact equivalent to the given non-uniform FB. For every filter h_k, g_k in the non-uniform FB, introduce $q_k = D/d_k$ filters, given by specific delayed versions of h_k, g_k:

$$h_k^{(l)}[n] = h_k * \delta_{ld_k} = h_k[n - ld_k] \quad \text{and} \quad g_k^{(l)}[n] = g_k * \delta_{-ld_k} = g_k[n + ld_k], \quad (25)$$

for $l = 0, \ldots, q_k - 1$. It is easily seen that convolution with δ_k equals translation by k samples by just checking the definition of the convolution operation (17). Consequently, the sub-band components are

$$y_k^{(l)}[n] = y_k[nq_k - l] = \downarrow_D \{\underbrace{h_k * \delta_{ld_k}}_{:=h_k^{(l)}} * x\}[n], \tag{26}$$

where y_k is the k-th sub-band component with respect to the non-uniform FB. Thus, by grouping the corresponding q_k sub-bands, we obtain

$$y_k[n] = \sum_{l=0}^{q_k-1} \uparrow_{q_k} \left\{y_k^{(l)}\right\}[n+l].$$

In the frequency domain, the filters $h_k^{(l)}, g_k^{(l)}$ are given by

$$H_k^{(l)}(z) = z^{ld_k} H_k(z) \quad \text{and} \quad G_k^{(l)}(z) = z^{-ld_k} G_k(z).$$

Similar to before, the output of the FB can be written as

$$\begin{aligned}\widetilde{X}(z) &= D^{-1} \sum_{k=0}^{K} \sum_{j=0}^{D-1} \sum_{l=0}^{q_k-1} G_k^{(l)}(z) H_k^{(l)}\left(W_D^j z\right) X\left(W_D^j z\right) \\ &= D^{-1} \sum_{k=0}^{K} \sum_{j=0}^{D-1} G_k(z) H_k\left(W_D^j z\right) X\left(W_D^j z\right) \sum_{l=0}^{q_k-1} W_D^{jld_k}\end{aligned} \tag{27}$$

To obtain the second equality, we have used that $G_k^{(l)}(z) H_k^{(l)}\left(W_D^j z\right) = W_D^{jld_k} G_k(z) H_k\left(W_D^{jld_k} z\right)$. Insert Eq. (16) into (27) to obtain

$$\begin{aligned}\widetilde{X}(z) &= D^{-1} \sum_{k=0}^{K} \sum_{j=0}^{d_k-1} q_k G_k(z) H_k\left(W_D^{jq_k} z\right) X\left(W_D^{jq_k} z\right) \\ &= D^{-1} \sum_{k=0}^{K} [X(W_D^0 z), \ldots, X(W_D^{D_1}(z)] \boldsymbol{h}_k(z) G_k(z) \\ &= D^{-1} [X(W_D^0 z), \ldots, X(W_D^{D-1} z)] \boldsymbol{H}(z) \boldsymbol{G}(z),\end{aligned} \tag{28}$$

which is exactly the output of the non-uniform FB specified by the h_k's, g_k's, and d_k's, see (23). Therefore, we see that an equivalent uniform FB for every non-uniform FB is obtained by decomposing each k-th channel of the non-uniform system into q_k channels. The uniform system then features $\sum_{k=0}^{K} q_k$ channels in total with the downsampling factor $D = \operatorname{lcm}(d_0, \ldots, d_K)$ in all channels.

4.3 Connection to Frame Theory

We will now describe in detail the connection between non-uniform FBs and frame theory. The main difference to previous work in this direction, cf. [14, 20, 25, 34], is that we do not restrict to the case of uniform FBs. The results in this section are not new, but this presentation is their first appearance in the context of non-uniform FBs. Besides using the equivalent uniform FB representation, see Figure 9, we transfer results previously obtained for *generalized shift-invariant systems* [44, 77] and nonstationary Gabor systems [9, 47, 48] to the non-uniform FB setting. For that purpose, we consider frames over the Hilbert space $\mathscr{H} = \ell^2(\mathbb{Z})$ of finite energy sequences. Moreover, we consider only FBs with a finite number $K+1 \in \mathbb{N}$ of channels, a setup naturally satisfied in every real-world application. The central observation linking FBs to frames is that the convolution can be expressed as an inner product:

$$y_k[n] = \downarrow_{d_k} \{h_k * x\}[n] = \langle x, \overline{h_k[nd_k - \cdot]} \rangle$$

where the bar denotes the complex conjugate. Hence, the sub-band components with respect to the filters h_k and downsampling factors d_k equal the frame coefficients of the system $\Phi = \left(\overline{h_k[nd_k - \cdot]}\right)_{k,n}$. Note that the upper frame inequality, see Eq. (6), is equivalent to the h_k's and d_k's defining a system where bounded energy of the input implies bounded energy of the output. We will investigate the frame properties of this system by transference to the Fourier domain [8]; we consider $\widehat{\Phi} = \left(\mathbf{E}_{-nd_k}\widehat{h_k}\right)_{k,n}$, where $\widehat{h_k}(\xi) = \overline{H_k(e^{2\pi i\xi})}$ denotes the Fourier transform of $\overline{h_k[-\cdot]}$ and the operator $\mathbf{E}_\omega$ denotes modulation, i.e. $\mathbf{E}_{-nd_k}\widehat{h_k}(\xi) = \widehat{h_k}(\xi)e^{-2\pi i n d_k \xi}$.

If Φ satisfies at least the upper frame inequality in Eq. (6), then the frame operators $\mathbf{S}_\Phi$ and $\mathbf{S}_{\widehat{\Phi}}$ are related by the matrix Fourier transform [2]:

$$\mathbf{S}_{\widehat{\Phi}} = \mathscr{F}_{DT}\, \mathbf{S}_\Phi\, \mathscr{F}_{DT}^{-1},$$

where $\mathscr{F}_{DT}$ denotes the discrete-time Fourier transform. Since the matrix Fourier transform is a unitary operation, the study of the frame properties of Φ reduces to the study of the operator $\mathbf{S}_{\widehat{\Phi}}$. In the context of FBs, the frame operator can be expressed as the action of an FB with analysis filters h_k's, downsampling and upsampling factors d_k's, and synthesis filters $\overline{h_k[-\cdot]}$. That is, the synthesis filters are given by the time-reversed, conjugate impulse responses of the analysis filters. This is a very common approach to FB synthesis. But note that it only gives perfect reconstruction if the system constitutes a Parseval frame, see Prop. 1. The z-transform of a time-reversed, conjugated signal is given by $\mathscr{Z}(\overline{h[-\cdot]})(z) = \overline{\mathscr{Z}(h)(1/\overline{z})}$. Inserting this into the alias domain representation of the FB (23) yields

$$\mathbf{S}_{\widehat{\Phi}} X(z) = \tfrac{1}{D}\left[X(W_D^0 z)\cdots X(W_D^{D-1} z)\right] \boldsymbol{H}(z) \begin{bmatrix} \overline{H_0(1/\overline{z})} \\ \vdots \\ \overline{H_K(1/\overline{z})} \end{bmatrix} \tag{29}$$

or, restricted to the Fourier domain

$$\mathbf{S}_{\widehat{\Phi}} X(e^{2\pi i\xi}) = [X(e^{2\pi i(\xi+0/D)})\cdots X(e^{2\pi i(\xi+(D-1)/D)})]\mathscr{H}(\xi), \tag{30}$$

with

$$\mathscr{H}(\xi) := [\mathscr{H}_0(\xi),\ldots,\mathscr{H}_{D-1}(\xi)]^T := \frac{1}{D}\boldsymbol{H}(e^{2\pi i\xi})\left[\overline{H_0(e^{2\pi i\xi})},\ldots,\overline{H_K(e^{2\pi i\xi})}\right]^T, \tag{31}$$

for $\xi \in \mathbb{T} = \mathbb{R}/\mathbb{Z}$. Here, we used $\overline{1/e^{2\pi i\omega}} = e^{2\pi i\omega}$ for all $\omega \in \mathbb{R}$. We call $\mathscr{H}_0$ the *frequency response* and $\mathscr{H}_n$, $n = 1, \cdot D-1$ the *alias components* of the FB.

Another way to derive Eq. (30) is by using the Walnut representation of the frame operator for the nonstationary Gabor frame $\widehat{\Phi} = \left(\mathbf{E}_{-nd_k}\widehat{h}_k\right)_{k,n}$, first introduced in [28] for the continuous case setting.

Proposition 4 *Let* $\widehat{\Phi} = \left(\mathbf{E}_{-nd_k}\widehat{h}_k\right)_{k\in\{0,\ldots,K\},n\in\mathbb{Z}}$, *with* $\widehat{h}_k \in L^2(\mathbb{T})$ *being (essentially) bounded and* $d_k \in \mathbb{N}$. *Then the frame operator* $\mathbf{S}_{\widehat{\Phi}}$ *admits the Walnut representation*

$$S_{\widehat{\Phi}}\widehat{x}(\xi) = \sum_{k=0}^{K}\sum_{n=0}^{d_k-1} d_k^{-1}\widehat{h}_k(\xi)\overline{\widehat{h}_k(\xi - nd_k^{-1})}\widehat{x}(\xi - nd_k^{-1}), \tag{32}$$

for almost every $\xi \in \mathbb{T}$ *and all* $\widehat{x} \in L^2(\mathbb{T})$.

Proof By the definition of the frame operator, see Eq. (8), we have

$$\mathbf{S}_{\widehat{\Phi}}\widehat{x}(\xi) = \sum_{k,n}\left\langle \widehat{x}, \widehat{h}_k e^{-2\pi i n d_k\xi}\right\rangle \widehat{h}_k(\xi)e^{-2\pi i n d_k\xi}.$$

Note that

$$\sum_{n\in\mathbb{Z}}\langle\widehat{x}, e^{-2\pi i\xi n d_k}\widehat{h}_k\rangle e^{-2\pi i\xi n d_k} = \sum_{n\in\mathbb{Z}}\mathscr{F}_{DT}^{-1}(\widehat{x}\overline{\widehat{h}_k})[nd_k]e^{-2\pi i\xi n d_k}.$$

to get the result by applying Poisson's summation formula, see, e.g., [40]. □

The sums in (32) can be reordered to obtain

$$\sum_{n=0}^{D-1}\widehat{x}(\xi - nD^{-1})\sum_{k\in K_n} d_k^{-1}\widehat{h}_k(\xi)\overline{\widehat{h}_k(\xi - nD^{-1})},$$

where $K_n = \{k \in \{0,\dots,K\} \,:\, nD^{-1} = jd_k^{-1} \text{ for some } j \in \mathbb{N}\}$. Inserting $\widehat{h_k}(\xi) = \overline{H_k(e^{2\pi i\xi})}$ and comparing the definition of $\mathscr{H}_n$ in (31), we can see that

$$\sum_{k\in K_n} \widehat{h_k}(\xi)\overline{\widehat{h_k}(\xi - nD^{-1})} = \sum_{k\in K_n} \overline{H_k(e^{2\pi i\xi})}H_k(e^{2\pi i(\xi - n/D^{-1})}) = \mathscr{H}_n(\xi)$$

for almost every $\xi \in \mathbb{T}$ and all $n = 0,\dots,D-1$. Hence, we recover the representation of the frame operator as per (30), as expected. What makes Proposition 4 so interesting is that it facilitates the derivation of some important sufficient frame conditions. The first is a generalization of the theory of painless non-orthogonal expansions by Daubechies et al. [27], see also [9] for a direct proof.

Corollary 2 *Let* $\widehat{\Phi} = \left(\mathbf{E}_{-nd_k}\widehat{h_k}\right)_{k\in\{0,\dots,K\},n\in\mathbb{Z}}$, *with* $\widehat{h_k} \in L^2(\mathbb{T})$ *and* $d_k \in \mathbb{N}$. *Assume for all* $0 \le k \le K$, *there is* $I_k \subseteq \mathbb{T}$ *with* $|I_k| \le d_k^{-1}$ *and* $\widehat{h_k}(\xi) = 0$ *for almost every* $\xi \in \mathbb{T} \setminus I_k$. *Then* $\widehat{\Phi}$ *is a frame if and only if there are* A, B *such that*

$$0 < A \le \sum_{k=0}^{K} d_k^{-1}|\widehat{h_k}|^2 = \mathscr{H}_0 \le B < \infty, \ a.e. \tag{33}$$

Moreover, a dual frame for $\widehat{\Phi}$ *is given by* $\widehat{\Psi} = (\mathbf{E}_{-nd_k}\widehat{g_k})_{k\in\{0,\dots,K\},n\in\mathbb{Z}}$, *where*

$$\widehat{g_k}(\xi) = \frac{\widehat{h_k}(\xi)}{\mathscr{H}_0(\xi)} \ a.e. \tag{34}$$

Proof First, note that the existence of the upper bound B is equivalent to $\widehat{h_k} \in L^\infty(\mathbb{T})$, for all $k = 0,\dots,K$. It is easy to see that under the assumptions given, Eq. (32) equals

$$\mathbf{S}_{\widehat{\Phi}}\widehat{x}(\xi) = \widehat{x}(\xi)\sum_{k=0}^{K} d_k^{-1}|\widehat{h_k}|^2(\xi) = \widehat{x}(\xi)\cdot\mathscr{H}_0(\xi).$$

Hence, $\mathbf{S}_{\widehat{\Phi}}$ is invertible if and only if $\mathscr{H}_0$ is bounded above and below, proving the first part. Moreover, $\mathbf{S}_{\widehat{\Phi}}^{-1}$ is given by pointwise multiplication with $1/\mathscr{H}_0$ and therefore, the elements of the canonical dual frame for $\widehat{\Phi}$, defined in Eq. (13), are given by

$$\mathbf{S}_{\widehat{\Phi}}^{-1}\mathbf{E}_{-nd_k}\widehat{h_k} = \frac{\mathbf{E}_{-nd_k}\widehat{h_k}}{\mathscr{H}_0} = \mathbf{E}_{-nd_k}\frac{\widehat{h_k}}{\mathscr{H}_0} = \widehat{g_k}.$$

□

In other words, recalling $\widehat{h_k}(\xi) = \overline{H_k(e^{2\pi i\xi})}$, if the filters h_k are strictly band-limited, the downsampling factors d_k are small and $0 < A \leq \mathscr{H}_0 \leq B < 0$ almost everywhere, then we obtain a perfect reconstruction system with synthesis filters g_k defined by

$$G_k(e^{2\pi i\xi}) = \frac{\overline{H_k(e^{2\pi i\xi})}}{\mathscr{H}_0(\xi)}.$$

The second, more general and more interesting condition can be likened to a diagonal dominance result, i.e. if the main term $\mathscr{H}_0$ is *stronger* than the sum of the magnitude of alias components $\mathscr{H}_n, n = 1, \ldots, D-1$, then the FB analysis provided by the filters h_k and downsampling factors d_k is invertible.

Proposition 5 *Let* $\widehat{\Phi} = \left(\mathbf{E}_{-nd_k}\widehat{h_k}\right)_{k\in\{0,\ldots,K\},n\in\mathbb{Z}}$, *with* $\widehat{h_k} \in L^2(\mathbb{T})$ *and* $d_k \in \mathbb{N}$. *If there are* $0 < A \leq B < \infty$ *with*

$$A \leq \sum_{k=0}^{K} d_k^{-1}|\widehat{h_k}|^2(\xi) \pm \sum_{k=0}^{K}\sum_{n=1}^{d_k-1} d_k^{-1}\left|\widehat{h_k}(\xi)\widehat{h_k}(\xi - nd_k^{-1})\right| \leq B, \tag{35}$$

for almost every $\xi \in \mathbb{T}$, *then* $\widehat{\Phi}$ *forms a frame with frame bounds* A, B.

Note that (35) impliest $\widehat{h_k} \in \mathbf{L}^\infty(\mathbb{R})$ for all $k \in \{0, \ldots, K\}$. Therefore, Proposition 4 applies for any FB that satisfies (35). The proof of Proposition 5 is somewhat lengthy and we omit it here. It is very similar to the proof of the analogous conditions for Gabor and wavelet frames that can be found in [26] for the continuous case. It can also be seen as a corollary of [24, Theorem 3.4], covering a more general setting. A few things should be noted regarding Proposition 5.

(a) As mentioned before, this is a sort of diagonal dominance result. While the sum $\sum_{k=0}^{K} d_k^{-1}|\widehat{h_k}|^2(\xi)$ corresponds to $\mathscr{H}_0$, we have

$$\sum_{k=0}^{K}\sum_{n=1}^{d_k-1} d_k^{-1}\left|\widehat{h_k}(\xi)\widehat{h_k}(\xi - nd_k^{-1})\right| = \sum_{n=1}^{D-1}|\mathscr{H}_n|(\xi).$$

Since, in fact, the finite number of channels guarantees the existence of B if and only if $\widehat{h_k} \in L^\infty(\mathbb{T})$, for all $k = 0, \ldots, K$, the result implies that the FB analysis provided by h_k's and d_k's is invertible, whenever

$$\mathscr{H}_0 - \sum_{n=1}^{D-1}|\mathscr{H}_n| \geq A > 0, \text{ almost everywhere.}$$

(b) No explicit dual frame is provided by Proposition 5. So, while we can determine invertibility quite easily, provided the Fourier transforms of the filters can be computed, the actual inversion process is still up in the air. In fact, it

is unclear whether there are synthesis filters g_k such that the h_k's and g_k's form a perfect reconstruction system with down-/upsampling factors d_k. We consider here two possible means of recovering the original signal X from the sub-band components Y_k.

First, the equivalent unform FB, comprised of the filters $h_k^{(l)}$, for $l \in \{0, \dots, q_k - 1\}$ and all $k \in \{0, \cdots, K\}$, with downsampling factor $D = \mathrm{lcm}(d_k : k \in \{0, \dots, K\})$ can be constructed. Since the non-uniform FB forms a frame, so does its uniform equivalent and hence the existence of a dual FB $g_k^{(l)}$, for $l \in \{0, \dots, q_k - 1\}$ and all $k \in \{0, \cdots, K\}$, is guaranteed. Note that the $g_k^{(l)}$ are not necessarily delayed versions of $g_k^{(0)}$, as it is the case for $h_k^{(l)}$. Then, the structure of the alias domain representation in (23) with $g_k = \overline{h_k[-\cdot]}$ can be exploited [14] to obtain perfect reconstruction synthesis. In the finite, discrete setting, i.e. when considering signals in $\mathbb{R}^L$ ($\mathbb{C}^L$), a dual FB can be computed explicitly and efficiently by a generalization of the methods presented by Strohmer [86], see also [75]. In practice, both the storage and time efficiency of computing the dual uniform FB rely crucially on $D = \mathrm{lcm}(d_k : k \, in \{0, \dots, K\})$ being small, i.e. $\sum_k q_k$ not being much larger than $K + 1$.

If that is not the case, the frame property of $\widehat{\Phi} = \left(\mathbf{E}_{-nd_k}\widehat{h_k}\right)_{k\in\{0,\dots,K\},n\in\mathbb{Z}}$ guarantees the convergence of the Neumann series

$$\mathbf{S}_{\widehat{\Phi}}^{-1} = \frac{2}{A_0 + B_0} \sum_{l=0}^{\infty} \left(\mathbf{I} - \frac{2}{A_0 + B_0}\mathbf{S}_{\widehat{\Phi}}\right)^l, \tag{36}$$

where $0 < A_0 \leq B_0 < \infty$ are the optimal frame bounds of $\widehat{\Phi}$. Instead of computing the elements of any dual frame explicitly, we can apply the inverse frame operator to the FB output

$$\tilde{X}(z) = \mathbf{S}_{\widehat{\Phi}} X(z) = \sum_{k=0}^{K} Y_k(z^{d_k}) H_k(z), \tag{37}$$

obtaining $\mathbf{S}_{\widehat{\Phi}}^{-1}\tilde{X} = X$. This can be implemented with the *frame algorithm* [29, 39]. However, any frame operator is positive definite and self-adjoint, allowing for extremely efficient implementation via the *conjugate gradients (CG)* [39, 87] algorithm. In addition to a significant boost in efficiency compared to the frame algorithm, the conjugate gradients algorithm does not require an estimate of the optimal frame bounds A_0, B_0 and convergence speed depends solely on the condition number of $\mathbf{S}_{\widehat{\Phi}}$. It provides guaranteed, exact convergence in L steps for signals in $\mathbb{C}^L$, where every step essentially comprises one analysis and one synthesis step with the filters h_k and $g_k = \overline{h_k[-\cdot]}$, respectively. If furthermore, $\mathscr{H}_0 \gg \sum_{n=1}^{D-1} |\mathscr{H}_n|$, then convergence speed can be further increased by preconditioning [6], considering instead the operator defined by

$$\widetilde{\mathbf{S}_{\widehat{\Phi}}} X(e^{2\pi i\xi}) = \mathscr{H}_0(\xi)^{-1}\mathbf{S}_{\widehat{\Phi}} X(e^{2\pi i\xi}).$$

More specifically, the CG algorithm is employed to solve the system $\mathbf{D}_\Phi c = \mathbf{S}_\Phi x$ for x, given the coefficients c. Recall the analysis/synthesis operators $\mathbf{C}_\Phi, \mathbf{D}_\Phi$ (see Sec. 3.1.1), associated to a frame Φ, which are equivalent to the analysis/synthesis stages of the FB. The preconditioned case can be implemented most efficiently by precomputing an approximate dual FB, defined by $G_k(e^{2\pi i\xi}) = \mathscr{H}_0(\xi)^{-1}H_k(e^{2\pi i\xi})$ and solving instead

$$\mathbf{D}_\Psi c = \mathscr{F}^{-1}\mathscr{H}_0(\xi)^{-1}\mathbf{S}_{\widehat{\Phi}}\mathscr{F}x = \mathbf{D}_\Psi \mathbf{C}_\Phi x, \text{ where} \Psi = \{\overline{g_k[nd_k - \cdot]}\}_{k,n},$$

for x, given the coefficients c. Algorithm 1 shows a pseudo-code implementation of such a preconditioned CG scheme, available in the LTFAT Toolbox as the routine `ifilterbankiter`.

5 Frame Theory: Psychoacoustics-motivated Applications

5.1 A perfectly invertible, perceptually motivated filter bank

The concept of auditory filters lends itself nicely to the implementation as an FB. As motivated in Sec. 1, it can be expected that many audio signal processing applications greatly benefit from an invertible FB representation adapted to the auditory time-frequency resolution. Despite the auditory system showing significant nonlinear behavior, the results obtained through a linear representation are desirable for being much more predictable than when accounting for nonlinear effects. We

Algorithm 1 Iterative synthesis: $\tilde{x} = \mathbf{FBSYN}^{it}(c, (h_k, g_k, d_k)_k, \lambda)$

Initialize $x_0 = 0$, $k = 0$
$b \leftarrow \mathbf{D}_\Psi c$
$r_0 \leftarrow b$
$h_0, p_0 \leftarrow r_0$
repeat
  $q_k = \mathbf{D}_\Psi(\mathbf{C}_\Phi p_0)$
  $\alpha_k \leftarrow \frac{\langle r_k, h_k\rangle}{\langle p_k, q_k\rangle}$
  $x_{k+1} \leftarrow x_k + \alpha_k p_k$
  $r_{k+1} \leftarrow r_k + \alpha_k q_k$
  $h_{k+1} \leftarrow r_{k+1}$
  $\beta_k \leftarrow \frac{\langle r_{k+1}, h_{k+1}\rangle}{\langle r_k, h_k\rangle}$
  $p_{k+1} \leftarrow h_{k+1} + \beta_k p_k$
  $k \leftarrow k + 1$
until $r_k \le \lambda$
$\tilde{x} \leftarrow x_k$

call such a system *perceptually motivated FB*, to distinguish from *auditory FBs* that attempt to mimic the nonlinearities in the auditory system. Note that, as mentioned in Section 2.2, the first step in many auditory FBs is the computation of a perceptually motivated FB, see, e.g., [49]. The *AUDlet FBs* we present here are a family of perceptually motivated FBs that satisfy a perfect reconstruction property, offer flexible redundancy, and enable efficient implementation. They were introduced in [65, 66] and an implementation is available in the LTFAT Toolbox [81].

The AUDlet FB has a general non-uniform structure as presented in Figure 8 with analysis filters $H_k(z)$, synthesis filters $G_k(z)$, and downsampling and upsampling factors d_k. Considering only real-valued signals allows us to deal with symmetric $\mathscr{F}_{DT}$s and process only the positive-frequency range. Therefore let K denote the number of filters in the frequency range $[f_{\min}, f_{\max}] \cap [0, f_s/2[$, where $f_{\min} \geq 0$ to $f_{\max} \leq f_s/2$ and $f_s/2$ is the Nyquist frequency, i.e. half the sampling frequency. If $f_{\min} > 0$, this range includes an additional filter at the zero frequency. Furthermore, another filter is always positioned at the Nyquist frequency to ensure that the full frequency range is covered. Thus, all FBs below feature $K+1$ filters in total and their redundancy is given by $R = d_0^{-1} + 2\sum_{k=1}^{K-1} d_k^{-1} + d_K^{-1}$, since coefficients in the 1st to $K-1$-th subbands are complex-valued.

The AUDlet filters H_k's, $k \in \{0, \ldots, K\}$ are constructed in the frequency domain by

$$H_k(e^{2i\pi\xi}) = \Gamma_k^{-\frac{1}{2}} w\left(\frac{f_s \cdot \xi - f_k}{\Gamma_k}\right) \tag{38}$$

where $w(\xi)$ is a prototype filter shape with bandwidth 1 and center frequency 0. Here, the shape factor Γ_k controls the effective bandwidth of H_k and f_k determines its center frequency. The factor $\Gamma_k^{-1/2}$ ensures that all filters (i.e., for all k) have the same energy. To obtain filters equidistantly spaced on a perceptual frequency scale, the sets $\{f_k\}$ and $\{\Gamma_k\}$ are calculated using the corresponding F_{AUD} and BW_{AUD} formulas, see Table 1 for more information on the AUDlet parameters and their

Table 1 Parameters of the perceptually motivated AUDlet FB

Parameter	Role	Information
$f_{\min}$	minimum frequency in Hz	$f_{\min} \in [0, f_s/2[,\ f_{\min} < f_{\max}$
$f_{\max}$	maximum frequency in Hz	$f_{\max} \in]0, f_s/2[,\ f_{\max} > f_{\min}$
f_k	center frequencies in Hz	$F_{\text{AUD}}^{-1}(F_{\text{AUD}}(f_0) + k/V)$
K	(essential) number of channels	$K = V\,(F_{\text{AUD}}(\xi_{\max}) - F_{\text{AUD}}(f_{\min})) + (1 - \delta_{0,f_{\min}})$
V	channels per scale unit	$V = \left(F_{\text{AUD}}(f_{k+1}) - F_{\text{AUD}}(f_k)\right)^{-1}, k \in [1, K-2]$
w	frequency domain filter prototype	$w \in L^2(\mathbb{T})$
Γ_k	dilation factors	$r_{bw} BW_{\text{AUD}}(f_k),\ r_{bw} > 0$ (default $= 1$)
H_k	filter transfer functions	$H_k(e^{2i\pi\xi}) = \Gamma_k^{-\frac{1}{2}} w\left(\frac{f_s \cdot \xi - f_k}{\Gamma_k}\right)$
d_k	downsampling factors	$r_d BW_{\text{AUD}}^{-1}(\xi_k),\ r_d > 0$ (default non-uniform = 1)
R	redundancy	$R = d_0^{-1} + 2\sum_{k=1}^{K-1} d_k^{-1} + d_K^{-1}$

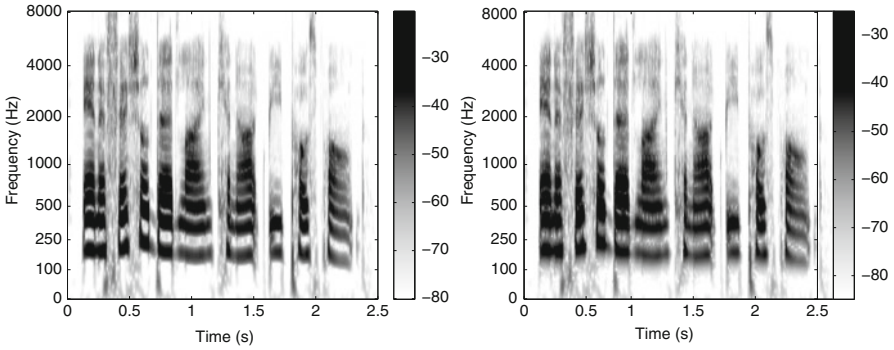


Fig. 10 Analyses of a female speech signal taken from the TIMIT database [36] by (left) the AUDlet FB and (right) the gammatone FB using $V = 6$ filters per ERB ($K = 201$). It can be seen that the two signal representations are very similar over the whole time-frequency plane.

relations. Since we emphasize inversion, the default analysis parameters are chosen such that the filters H_k and downsampling factors d_k form a frame. As an example, the AUDlet (a) and gammatone (b) analyses of a speech signal are represented in Figure 10 using AUD = ERB and $V = 6$ filters per ERB. The filter prototype w for the AUDlet was a Hann window. It can be seen that the two signal representations are very similar over the whole time-frequency plane. Since the gammatone filter is an acknowledged auditory filter model, this indicates that the time-frequency resolution of the AUDlet approximates well the auditory resolution.

5.2 *Perceptual Sparsity*

As discussed in Sec. 2.3 not all components of a sound are perceived. This effect can be described by masking models and naturally leads to the following question: Given a time-frequency representation or any representation linked to audio, how can we apply that knowledge to only include audible coefficients in the synthesis? In an attempt to answer this question, efforts were made to combine frame theory and masking models into a concept called the *Irrelevance Filter*. This concept is somehow linked to the currently very prominent sparsity and compressed sensing approach, see, e.g., [31, 35] for an overview. To reduce the amount of non-zero coefficients, the irrelevance filter uses a perceptual measure of sparsity, hence *perceptual sparsity*. Perceptual and compressed sparsity can certainly be combined, see e.g. [21]. Similar to the methods used in compressed sensing, a redundant representation offers an advantage for perceptual sparsity as well, since the same signal can be reconstructed from several sets of coefficients.

The concept of the irrelevance filter was first introduced in [30] and fully developed in [7]. It consists in removing the inaudible atoms in a Gabor transform while causing no audible difference to the original sound after re-synthesis. Precisely, an adaptive threshold function is calculated for each spectrum (i.e., at each time slice)

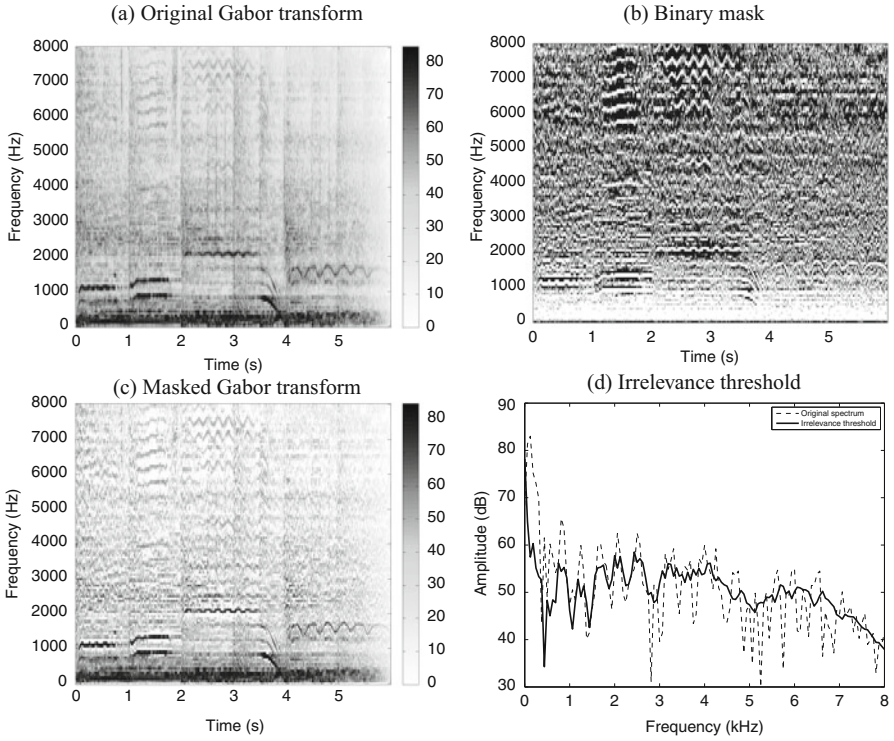


Fig. 11 Example application of the irrelevance filter as implemented in [7] to a music signal (excerpt from the song "Heart of Steel" from Manowar). (a) Squared magnitude of the Gabor transform (in dB). (b) Binary mask estimated from the irrelevance threshold. White = 1, black = 0. (c) Squared magnitude (in dB) of the masked Gabor transform, i.e. the result of the point-wise multiplication between the original transform and the binary mask. (d) Amplitudes (in dB) of the irrelevance threshold (bold straight line) and original spectrum (dashed line) at a given time slice.

of the Gabor transform using a simple model of spectral masking (see Sec. 2.3.1), resulting in the so-called irrelevance threshold. Then, the amplitudes of all atoms falling below the irrelevance threshold are set to zero and the inverse transform is applied to the set of modified Gabor coefficients. This corresponds to an adaptive *Gabor frame multiplier* with coefficients in $\{0, 1\}$. The application of the irrelevance filter to a musical signal sampled at 16 kHz is shown in Figure 11. A Matlab implementation of the algorithm proposed in [7] was used. All Gabor transform and filter parameters were identical to those mentioned in [7]. Noteworthy, the offset parameter o was set to -2.59 dB. In this particular example, about 48% components were removed without causing any audible difference to the original sound after re-synthesis (as judged by informal listening by the authors). A formal listening test performed in [7] with 36 normal-hearing listeners and various musical and speech signals indicated that, on average, 36% coefficients can be removed without causing any audible artifact in the re-synthesis.

The irrelevance filter as depicted here has shown very promising results but the approach could be improved. Specifically, the main limitations of the algorithm are the fixed resolution in the Gabor transform and the use of a simple spectral masking model to predict masking in the time-frequency domain. Combining an invertible perceptually motivated transform like the AUDlet FB (Sec. 5.1) with a model of time-frequency masking (Sec. 2.3.3) is expected to improve performance of the filter. This is work in progress. Potential applications of perceptual sparsity include, for instance:

1. Sound/Data Compression: For applications where perception is relevant, there is no need to encode perceptually irrelevant information. Data that cannot be heard should be simply omitted. A similar algorithm is, for example, used in the MP3 codec. If "over-masking" is used, i.e. the threshold is moved beyond the level of relevance, a higher compression rate can be reached [70].
2. Sound Design: For the visualization of sounds the perceptually irrelevant part can be disregarded. This is, for example, used for car sound design [13].

6 Conclusion

In this chapter, we have discussed some important concepts from hearing research and perceptual audio signal processing, such as auditory masking and auditory filter banks. Natural and important considerations served as a strong indicator that frame theory provides a solid foundation for the design of robust representations for perceptual signal analysis and processing. This connection was further reinforced by exposing the similarity between some concepts arising naturally in frame theory and signal processing, e.g. between frame multipliers and time-variant filters. Finally, we have shown how frame theory can be used to analyze and implement invertible filter banks, in a quite general setting where previous synthesis methods might fail or be highly inefficient. The codes for Matlab/Octave to reproduce the results presented in Secs. 3 and 5 in this chapter are available for download on the companion Webpage https://www.kfs.oeaw.ac.at/frames_for_psychoacoustics.

It is likely that readers of this contribution who are researchers in psychoacoustics or audio signal processing have already used frames without being aware of the fact. We hope that such readers will, to some extent, grasp the basic principles of the rich mathematical background provided by frame theory and its importance to fundamental issues of signal analysis and processing. With that knowledge, we believe, they will be able to better understand the signal analysis tools they use and might even be able to design new techniques that further elevate their research.

On the other hand, researchers in applied mathematics or signal processing have been supplied with basic knowledge of some central psychoacoustics concepts. We hope that our short excursion piqued their interest and will serve as a starting point for applying their knowledge in the rich and various fields of psychoacoustics or perceptual signal processing.

Acknowledgements The authors acknowledge support from the Austrian Science Fund (FWF) START-project FLAME ("Frames and Linear Operators for Acoustical Modeling and Parameter Estimation"; Y 551-N13) and the French-Austrian ANR-FWF project POTION ("Perceptual Optimization of Time-Frequency Representations and Audio Coding; I 1362-N30"). They thank B. Laback for discussions and W. Kreuzer for the help with a graphics software.

References

1. S. Akkarakaran, P. Vaidyanathan, Nonuniform filter banks: new results and open problems, in *Beyond Wavelets*. Studies in Computational Mathematics, vol. 10 (Elsevier, Amsterdam, 2003), pp. 259–301
2. P. Balazs, Regular and irregular Gabor multipliers with application to psychoacoustic masking. PhD thesis, University of Vienna (2005)
3. P. Balazs, Basic definition and properties of Bessel multipliers. J. Math. Anal. Appl. **325**(1), 571–585 (2007)
4. P. Balazs, Frames and finite dimensionality: frame transformation, classification and algorithms. Appl. Math. Sci. **2**(41–44), 2131–2144 (2008)
5. P. Balazs, D.T. Stoeva, Representation of the inverse of a frame multiplier. J. Math. Anal. Appl. **422**(2), 981–994 (2015)
6. P. Balazs, H.G. Feichtinger, M. Hampejs, G. Kracher, Double preconditioning for Gabor frames. IEEE Trans. Signal Process. **54**(12), 4597–4610 (2006)
7. P. Balazs, B. Laback, G. Eckel, W.A. Deutsch, Time-frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking. IEEE Trans. Audio Speech Lang. Process. **18**(1), 34–49 (2010)
8. P. Balazs, C. Cabrelli, S.B. Heineken, U. Molter, Frames by multiplication. Curr. Dev. Theory Appl. Wavelets **5**(2–3), 165–186 (2011)
9. P. Balazs, M. Dörfler, F. Jaillet, N. Holighaus, G.A. Velasco, Theory, implementation and applications of nonstationary Gabor frames. J. Comput. Appl. Math. **236**(6), 1481–1496 (2011)
10. P. Balazs, M. Dörfler, M. Kowalski, B. Torrésani, Adapted and adaptive linear time-frequency representations: a synthesis point of view. IEEE Signal Process. Mag. **30**(6), 20–31 (2013)
11. N.K. Bari, Biorthogonal systems and bases in Hilbert space. Uch. Zap. Mosk. Gos. Univ. **148**, 69–107 (1951)
12. J.J. Benedetto, A. Teolis, A wavelet auditory model and data compression. Appl. Comput. Harmon. Anal. **1**, 3–28 (1994)
13. M. Bézat, V. Roussarie, T. Voinier, R. Kronland-Martinet, S. Ystad, Car door closure sounds: characterization of perceptual properties through analysis-synthesis approach, in *Proceedings of the 19th International Congress on Acoustics (ICA)*, Madrid (2007)
14. H. Bölcskei, F. Hlawatsch, H. Feichtinger, Frame-theoretic analysis of oversampled filter banks. IEEE Trans. Signal Process. **46**(12), 3256–3268 (1998)
15. M. Bownik, J. Lemvig, The canonical and alternate duals of a wavelet frame. Appl. Comput. Harmon. Anal. **23**(2), 263–272 (2007)
16. A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA, 1990)
17. P.G. Casazza, The art of frame theory. Taiwan. J. Math. **4**(2), 129–201 (2000)
18. P.G. Casazza, O. Christensen, Gabor frames over irregular lattices. Adv. Comput. Math. **18**(2–4), 329–344 (2003)
19. P. Casazza, G. Kutyniok, *Finite Frames: Theory and Applications*. Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, 2012)
20. L. Chai, J. Zhang, C. Zhang, E. Mosca, Bound ratio minimization of filter bank frames. IEEE Trans. Signal Process. **58**(1), 209–220 (2010)

21. G. Chardon, T. Necciari, P. Balazs, Perceptual matching pursuit with Gabor dictionaries and time-frequency masking, in *Proceedings of the 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)* (2014)
22. O. Christensen, *An Introduction to Frames and Riesz Bases*. Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, 2003)
23. O. Christensen, Pairs of dual Gabor frame generators with compact support and desired frequency localization. Appl. Comput. Harmon. Anal. **20**(3), 403–410 (2006)
24. O. Christensen, S.S. Goh, Fourier-like frames on locally compact abelian groups. J. Approx. Theory **192**, 82–101 (2015)
25. Z. Cvetković, M. Vetterli, Oversampled filter banks. IEEE Trans. Signal Process. **46**(5), 1245–1255 (1998)
26. I. Daubechies, *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61 (SIAM, Philadelphia, PA, 1992)
27. I. Daubechies, A. Grossmann, Y. Meyer, Painless nonorthogonal expansions. J. Math. Phys. **27**(5), 1271–1283 (1986)
28. M. Dörfler, E. Matusiak, Nonstationary Gabor frames - existence and construction. Int. J. Wavelets Multiresolution Inf. Process. **12**(3) (2014)
29. R.J. Duffin, A.C. Schaeffer, A class of nonharmonic Fourier series. Trans. Am. Math. Soc. **72**, 341–366 (1952)
30. G. Eckel, Ein Modell der Mehrfachverdeckung für die Analyse musikalischerSchallsignale. PhD thesis, University of Vienna (1989)
31. M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing* (Springer, New York, 2010)
32. H. Fastl, E. Zwicker, *Psychoacoustics — Facts and Models*, 3rd edn. (Springer, Berlin, 2006)
33. H.G. Feichtinger, K. Nowak, A first survey of Gabor multipliers, in *Advances in Gabor Analysis*, ed. by H.G. Feichtinger, T. Strohmer. Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, 2003), pp. 99–128
34. M. Fickus, M.L. Massar, D.G. Mixon, Finite frames and filter banks, in *Finite Frames*. Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, 2013), pp. 337–379
35. M. Fornasier, *Theoretical Foundations and Numerical Methods for Sparse Recovery*. Radon Series on Computational and Applied Mathematics, vol. 9 (Walter de Gruyter, Berlin, 2010)
36. J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1* (Linguistic Data Consortium, Philadelphia, 1993)
37. B.R. Glasberg, B.C.J. Moore, Derivation of auditory filter shapes from notched-noise data. Hear. Res. **47**, 103–138 (1990)
38. D.D. Greenwood, A cochlear frequency-position function for several species—29 years later. J. Acoust. Soc. Am. **87**(6), 2592–2605 (1990)
39. K. Gröchenig, Acceleration of the frame algorithm. IEEE Trans. Signal Process. **41**(12), 3331–3340 (1993)
40. K. Gröchenig, *Foundations of Time-Frequency Analysis*. Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, MA, 2001)
41. T.S. Gunawan, E. Ambikairajah, J. Epps, Perceptual speech enhancement exploiting temporal masking properties of human auditory system. Speech Commun. **52**(5), 381–393 (2010)
42. C. Heil, *A Basis Theory Primer*, Expanded edn. Applied and Numerical Harmonic Analysis (Birkhäuser, Basel, 2011)
43. C. Heil, D.F. Walnut, Continuous and discrete wavelet transforms. SIAM Rev. **31**, 628–666 (1989)
44. E. Hernández, D. Labate, G. Weiss, A unified characterization of reproducing systems generated by a finite family. II. J. Geom. Anal. **12**(4), 615–662 (2002)
45. H.G. Heuser, *Functional Analysis*, Transl. by John Horvath (Wiley, Chichester, 1982), 408 pp
46. P.Q. Hoang, P.P. Vaidyanathan, Non-uniform multirate filter banks: theory and design, in *IEEE International Symposium on Circuits and Systems*, vol. 1 (1989), pp. 371–374

47. N. Holighaus, Structure of nonstationary Gabor frames and their dual systems. Appl. Comput. Harmon. Anal. **37**(3), 442–463 (2014)
48. N. Holighaus, M. Dörfler, G. Velasco, T. Grill, A framework for invertible, real-time constant-Q transforms. IEEE Audio Speech Language Process. **21**(4), 775–785 (2013)
49. T. Irino, R.D. Patterson, A dynamic compressive gammachirp auditory filterbank. IEEE Audio Speech Language Process. **14**(6), 2222–2232 (2006)
50. A. Janssen, From continuous to discrete Weyl-Heisenberg frames through sampling. J. Fourier Anal. Appl. **3**(5), 583–596 (1997)
51. W. Jesteadt, S.P. Bacon, J.R. Lehman, Forward masking as a function of frequency, masker level, and signal delay. J. Acoust. Soc. Am. **71**(4), 950–962 (1982)
52. A. Kern, C. Heid, W.-H. Steeb, N. Stoop, R. Stoop, Biophysical parameters modification could overcome essential hearing gaps. PLoS Comput. Biol. **4**(8), e1000161 (2008)
53. G. Kidd Jr., L.L. Feth, Patterns of residual masking. Hear. Res. **5**, 49–67 (1981)
54. J. Kovačević, M. Vetterli, Perfect reconstruction filter banks with rational sampling factors. IEEE Trans. Signal Process. **41**(6), 2047–2066 (1993)
55. B. Laback, P. Balazs, T. Necciari, S. Savel, S. Meunier, S. Ystad, R. Kronland-Martinet, Additivity of nonsimultaneous masking for short Gaussian-shaped sinusoids. J. Acoust. Soc. Am. **129**(2), 888–897 (2011)
56. B. Laback, T. Necciari, P. Balazs, S. Savel, S. Ystad, Simultaneous masking additivity for short Gaussian-shaped tones: spectral effects. J. Acoust. Soc. Am. **134**(2), 1160–1171 (2013)
57. J. Leng, D. Han, T. Huang, Optimal dual frames for communication coding with probabilistic erasures. IEEE Trans. Signal Process. **59**(11), 5380–5389 (2011)
58. E.A. Lopez-Poveda, R. Meddis, A human nonlinear filterbank. J. Acoust. Soc. Am. **110**(6), 3107–3118 (2001)
59. R. Lyon, A. Katsiamis, E. Drakakis, History and future of auditory filter models. in *Proceedings of ISCAS* (IEEE, Paris, 2010), pp. 3809–3812
60. G. Matz, F. Hlawatsch, Time-frequency transfer function calculus (symbolic calculus) of linear time-varying systems (linear operators) based on a generalized underspread theory. J. Math. Phys. **39**(8), 4041–4070 (1998)
61. G. Matz, F. Hlawatsch, Linear time-frequency filters: on-line algorithms and applications, Chap. 6, in *Application in Time-Frequency Signal Processing*, ed. by A. Papandreou-Suppappola (CRC Press, Boca Raton, FL, 2002), pp. 205–271
62. B.C.J. Moore, *An Introduction to the Psychology of Hearing*, 6th edn. (Emerald Group Publishing, Bingley, 2012)
63. B.C.J. Moore, J.I. Alcántara, T. Dau, Masking patterns for sinusoidal and narrow-band noise maskers. J. Acoust. Soc. Am. **104**(2), 1023–1038 (1998)
64. T. Necciari, Auditory time-frequency masking: psychoacoustical measures and application to the analysis-synthesis of sound signals. PhD thesis, Aix-Marseille University, France (2010)
65. T. Necciari, P. Balazs, N. Holighaus, and P. Søndergaard. The ERBlet transform: an auditory-based time-frequency representation with perfect reconstruction, in *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)* (2013), pp. 498–502
66. T. Necciari, N. Holighaus, P. Balazs, Z. Průša, P. Majdak, Frame-theoretic recipe for the construction of gammatone and perceptually motivated filter banks with perfect reconstruction, http://arxiv.org/abs/1601.06652
67. J.J. O'Donovan, D.J. Furlong, Perceptually motivated time-frequency analysis. J. Acoust. Soc. Am. **117**(1), 250–262 (2005)
68. A.V. Oppenheim, R.W. Schafer, *Discrete-time Signal Processing* (Prentice Hall, Englewood Cliffs, NJ, 1989)
69. A.V. Oppenheim, R.W. Schafer, J.R. Buck, et al., *Discrete-Time Signal Processing*, vol. 2 (Prentice Hall, Englewood Cliffs, NJ, 1989)
70. T. Painter, A. Spanias, Perceptual coding of digital audio. Proc. IEEE **88**, 451–515 (2000)
71. R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, M.H. Allerhand, Complex sounds and auditory images, in *Auditory Physiology and Perception, Proceedings of the 9th International Symposium on Hearing* (Pergamon, Oxford, 1992), pp. 429–446

72. N. Perraudin, N. Holighaus, P. Søndergaard, P. Balazs, Gabor dual windows using convex optimization, in *Proceedings of the 10th International Conference on Sampling Theory and Applications (SAMPTA 2013)* (2013)
73. C.J. Plack, *The Sense of Hearing*, 2nd edn. (Psychology Press, Oxon, 2013)
74. Z. Průša, P. Søndergaard, N. Holighaus, C. Wiesmeyr, P. Balazs, The large time-frequency analysis toolbox 2.0, in *Sound, Music, and Motion*, ed. by M. Aramaki, O. Derrien, R. Kronland-Martinet, S. Ystad. Lecture Notes in Computer Science (Springer, Berlin, 2014), pp. 419–442
75. Z. Průša, P. Søndergaard, P. Rajmic, Discrete wavelet transforms in the large time-frequency analysis toolbox for MATLAB/GNU octave. ACM Trans. Math. Softw. **42**(4), Article 32, 23 p. (2016)
76. E. Ravelli, G. Richard, L. Daudet, Union of MDCT bases for audio coding. IEEE Trans. Audio Speech Language Process. **16**(8), 1361–1372 (2008)
77. A. Ron, Z. Shen, Generalized shift-invariant systems. Constr. Approx. **22**, 1–45 (2005)
78. W. Rudin, *Functional Analysis.* McGraw-Hill Series in Higher Mathematics (McGraw-Hill, New York, 1973), 397 pp
79. D. Soderquist, A. Carstens, G. Frank, Backward, simultaneous, and forward masking as a function of signal delay and frequency. J. Aud. Res. **21**, 227–245 (1981)
80. P. Søndergaard, Gabor frames by sampling and periodization. Adv. Comput. Math. **27**(4), 355–373 (2007)
81. P. Søndergaard, B. Torrésani, P. Balazs, The linear time frequency analysis toolbox. Int. J. Wavelets Multiresolution Inf. Process. **10**(4), 1250032 (2012)
82. D.T. Stoeva, P. Balazs, Invertibility of multipliers. Appl. Comput. Harmon. Anal. **33**(2), 292–299 (2012)
83. D.T. Stoeva, P. Balazs, Canonical forms of unconditionally convergent multipliers. J. Math. Anal. Appl. **399**, 252–259 (2013)
84. D.T. Stoeva, P. Balazs, Riesz bases multipliers, in *Concrete Operators, Spectral Theory, Operators in Harmonic Analysis and Approximation*, ed. by M.C. Boiso, H. Hedenmalm, M.A. Kaashoek, A. Montes-Rodrguez, S. Treil. Operator Theory: Advances and Applications, vol. 236 (Birkhäuser/Springer, Basel, 2014), pp. 475–482
85. S. Strahl, A. Mertins, Analysis and design of gammatone signal models. J. Acoust. Soc. Am. **126**(5), 2379–2389 (2009)
86. T. Strohmer, Numerical algorithms for discrete Gabor expansions, in *Gabor Analysis and Algorithms: Theory and Applications*, ed. by H.G. Feichtinger, T. Strohmer. Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, 1998), pp. 267–294
87. L.N. Trefethen, D. Bau III, *Numerical Linear Algebra* (SIAM, Philadelphia, PA, 1997)
88. M. Unoki, T. Irino, B. Glasberg, B.C.J. Moore, R.D. Patterson, Comparison of the roex and gammachirp filters as representations of the auditory filter. J. Acoust. Soc. Am. **120**(3), 1474–1492 (2006)
89. P. Vaidyanathan, *Multirate Systems And Filter Banks*. Electrical Engineering. Electronic and Digital Design (Prentice Hall, Englewood Cliffs, NJ, 1993)
90. X. Valero, F. Alias, Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification. IEEE Trans. Multimedia **14**(6), 1684–1689 (2012)
91. M. Vetterli, J. Kovacević, *Wavelets and Subband Coding* (Prentice Hall, Englewood Cliffs, NJ, 1995)
92. D. Wang, G.J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley, Hoboken, 2006)
93. T. Werther, Y.C. Eldar, N.K. Subbana, Dual Gabor frames: theory and computational aspects. IEEE Trans. Signal Process. **53**(11), 4147–4158 (2005)
94. C. Wiesmeyr, N. Holighaus, P. Søndergaard, Efficient algorithms for discrete Gabor transforms on a nonseparable lattice. IEEE Trans. Signal Process. **61**(20), 5131–5142 (2013)
95. R.M. Young, *An Introduction to Nonharmonic Fourier Series, Pure and Applied Mathematics*, vol. 93 (Academic, New York, 1980)

96. X. Zhao, Y. Shao, D. Wang, Casa-based robust speaker identification. IEEE Trans. Audio Speech Language Process. **20**(5), 1608–1616 (2012)
97. E. Zwicker, Dependence of post-masking on masker duration and its relation to temporal effects in loudness. J. Acoust. Soc. Am. **75**(1), 219–223 (1984)
98. E. Zwicker, E. Terhardt, Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. J. Acoust. Soc. Am. **68**(5), 1523–1525 (1980)

A Flexible Scheme for Constructing (Quasi-)Invariant Signal Representations

Jan Ernst

Abstract We describe a generic scheme for constructing signal representations that are (quasi-)invariant to perturbations of the domain. It is motivated from first principles and based on the preservation of topology under homeomorphisms. Under certain assumptions the resulting models can be used as direct plug ins to render an existing signal processing algorithm invariant. We show one concretization of the general scheme and develop it into a computational procedure that leads to applications in image processing and computer vision. The latter factorizes the $n-$dimensional problem into an ensemble of one-dimensional problems, which in turn can be reduced to proving the existence of paths in a graph. We show empirical results on real-world data in two important problems in computer vision, template matching and online tracking.

Keywords Invariance • Quasi-invariance • Shape matching • Template matching • Tracking

1 Introduction

This work starts with a simple question:

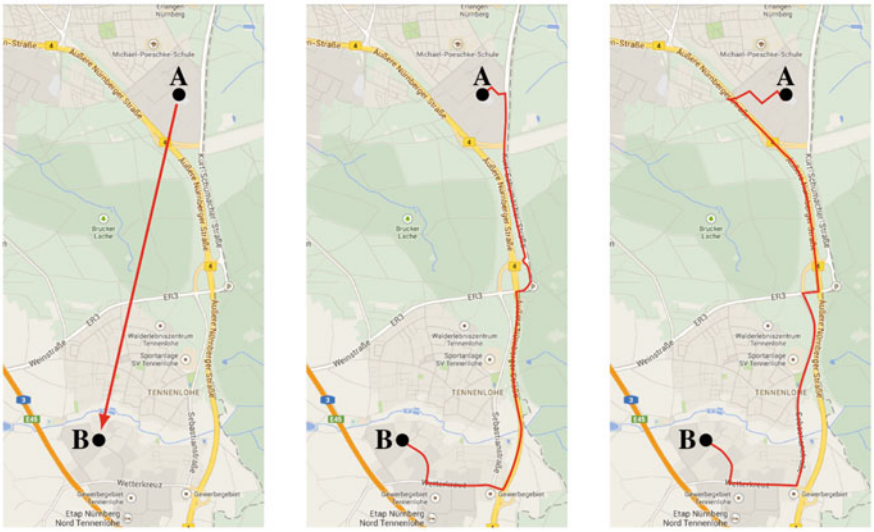
"How do we get from A to B?"

Let us assume for now that **A** and **B** are geographical locations, for instance in a city. Figure 1a shows a city map with two marked locations. Let us further assume that we are a tourist who just newly arrived in the city and we are presently at location **A** (e.g., the train station). We now would like to travel to location **B** (e.g., some attraction) and need directions. Asking bystanders, we may receive different answers, such as: "To get to location **B** from here, you have to "

1. "Move 3,142m towards bearing 193°"
2. "Follow this road until you see a three-story office building to your left, then take the third right, follow through until you see the roundabout, . . ."

J. Ernst (✉)
Siemens Corporate Technology, 755 College Road East, Princeton, NJ, USA
e-mail: jan.ernst@siemens.com

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4_11



(a) Euclidean: "3.0 km, Bearing 193°"

(b) Topological: "Follow this road . . ."

(c) Functional: "Take bus #30,. . ."

Fig. 1 Three answers to the question: "How does one get from *A* to *B*?"

3. "Walk to bus stop Erlangen South, take bus #30 to Thon, exit on Wetterkreuz, change to line #42, . . ."

The resulting paths on the map are illustrated in figures 1a to 1c. Although all three eventually lead to the same location **B**, they differ in the nature of their description. The first answer is *metric*: it describes the relative location of **A** and **B** by their Euclidean relation. The second answer has a *topological* aspect: it does not care about lengths and angles, but the topology of the underlying space as expressed by observable signal components such as street names and intersections. The third answer may be considered *functional*: it is formulated in terms of functional components on top of the signal space (e.g., bus and train routes). Signal processing algorithms often relate points in the signal space via descriptions of the first kind (e.g., the Euclidean or other metrics).

The central premise of this work is to *model the relation of two points based on the signal connectedness between them*, as in descriptions of the second kind. Under certain assumptions, these models can then replace metric point relations in an algorithm as a plug in. We will show that if this is done properly, one gains invariance to perturbations of the signal's domain.

The rest of the chapter is structured as follows: We will first informally study the properties of the Euclidean and topological descriptions in the next sections 1.1 to 1.2, with a brief excursion into the importance of invariance. This is followed in section 1.3 by a discussion of the role of pairwise relations in algorithms and how perturbations of the domain are commonly addressed. Section 2 formalizes the concepts of invariance, uniqueness, and completeness of a representation based on pairwise set representations. It can be skipped on first reading. Section 3 then introduces an example model for a restricted family of signals, while section 4 extends it to a significantly larger class of real-world signals. The resulting continuous model is then plugged into an existing algorithm in section 5 to demonstrate the plug-in characteristic, yielding an invariant version of the algorithm. As the model is formulated in the continuous domain, effort needs to be made in making it amenable to computation. One possible instantiation of a discrete implementation is sketched in Section 6, followed by results on two challenging real-world problems in section 7.

1.1 Path Descriptions under Deformations

One way to examine the difference in the above metric and topological representations is to expose the signal space to perturbations and see what happens. Figure 2 shows our example city map under various transformations, such as in-

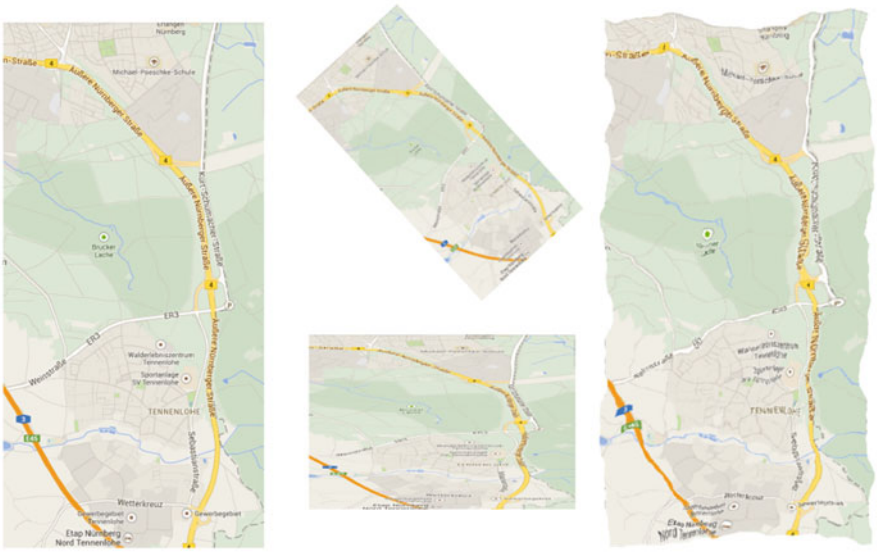


Fig. 2 Homeomorphic perturbations of the signal domain (starting left, clockwise): Original, in-plane rotation, generic local deformation, non-uniform scaling.

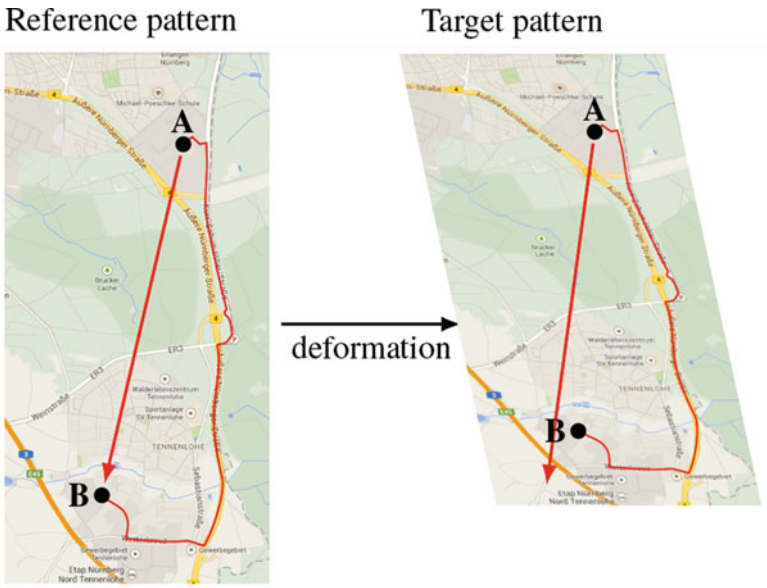


Fig. 3 The Euclidean and topological path descriptions under a homeomorphism.

plane rotation, local deformation, and non-uniform scaling. Now let's look at the paths under these transformation. Figure 3 illustrates the city map under a combined shear and non-uniform scaling (accelerated plate tectonics to stay with the example). The Euclidean description in the transformed frame now certainly leads to the wrong location, as following the instruction "move 3,142m towards bearing 193°" is not consistent with the new frame. The topological description, however, still leads to the correct location in the transformed frame. All we are doing is to follow landmarks according to a pre-defined sequence. If a particular road, for instance, is now twice as long until we reach a "T"-intersection, we still can recognize that we arrived at the intersection, although it takes twice as long. In other words, the descriptions may be *invariant* to certain transformations of the domain.

1.2 The Importance of Invariants

Invariance plays a central role in computer vision and many other domains. An invariant is a function of a signal that does not change its value when the signal undergoes a particular transformation. The transformations are often considered to be *nuisance perturbations*, i.e. they impede recovering the actual measurement of interest. Common examples in computer vision of such perturbations are the perspectivity in the formation of images by pinhole cameras, the rotation and translation of the domain of digital images and monotonic transforms of the image function due

to illumination changes. Examples of corresponding invariants to these particular perturbations are the cross ratio of points on the projective line under perspective transforms, the Euclidean distance under translation of the domain and the pairwise order relations of signal values under monotonic transforms of the signal's range. The importance of invariances amongst others lies in the fact that they provide a basis for building higher-level algorithms from invariant low-level representations. The idea is that the higher-level representation inherits the invariance properties from the low-level representation if constructed properly. In practice it is often challenging to find functions that are strictly invariant and practically useful at the same time. The use of quasi-invariants has been introduced in computer vision by [5, 6] to address this issue. While invariants have the strict requirement that they are constant over the entire range of the perturbation parameter, quasi-invariants are only expected to be close to constant for a limited range of the perturbation domain. No strict mathematical model for quasi-invariants is available and the theoretical characterization of quasi-invariants remains challenging [6, 24].

Examining the path descriptions, one notes that the Euclidean description is invariant to translations of the domain, but not rotations or scaling. The topological description is invariant at least to translations, rotations, scaling and shear (as illustrated in figure 3). As it is defined in terms of the topological structure of the signal space, it is precisely invariant to transformations that do not change the space's topology. The largest set of such transformations is the set of homeomorphisms, i.e. continuous bijections with continuous inverse.

1.3 Pairwise Relations in Signal Modeling

Pairwise point relations are at the heart of many algorithms that model properties of signals, e.g. in detection or recognition tasks in optical images, image sequences, or volumetric imaging. For instance in computer vision, one may want to model an object by its visual appearance. Figure 4 shows an example image of an object under a homeomorphism $\mathcal{H}$. An algorithm now may depend on many pairs of locations $(\mathbf{A}, \mathbf{B})$ on the object to be modeled. Under the perturbation $\mathcal{H}$, $\mathbf{A}$ and $\mathbf{B}$ are mapped into $\mathbf{A}' = \mathcal{H}(\mathbf{A})$ and $\mathbf{B}' = \mathcal{H}(\mathbf{B})$, respectively. Here we assume that $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{A}'$ are known. The exact $\mathcal{H}$ is usually unknown a priori however, making the determination of $\mathbf{B}'$ given $\mathbf{A}'$ challenging. Two common strategies to address this are:

Accept the uncertainty The idea is to treat $\mathcal{H}$ as a random variable and assume that there is prior knowledge, e.g. in the form of a probability distribution over the space of all possible $\mathcal{H}$. This prior is then used in defining a conditional probability for the location of $\mathbf{B}'$ of the form $P_{\mathcal{H}}(\mathbf{B}'|\mathbf{A}', \mathbf{A}, \mathbf{B})$ as shown in figure 5. Naturally, this introduces *uncertainty,* as the estimation of the position of $\mathbf{B}'$ now can assume a range of values. Whether this uncertainty is acceptable or not depends on many factors, e.g. if the uncertainty can be compensated in the later stages of the algorithm.

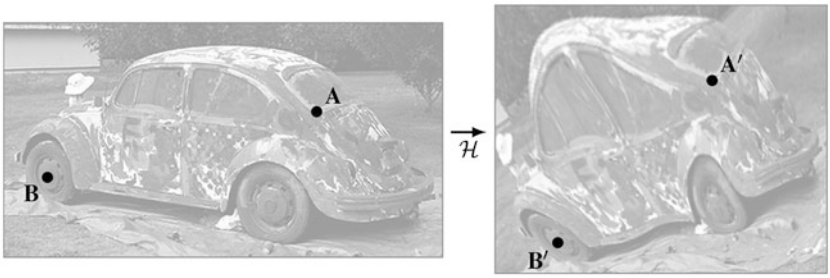


Fig. 4 An example object image under a homeomorphism $\mathcal{H}$.

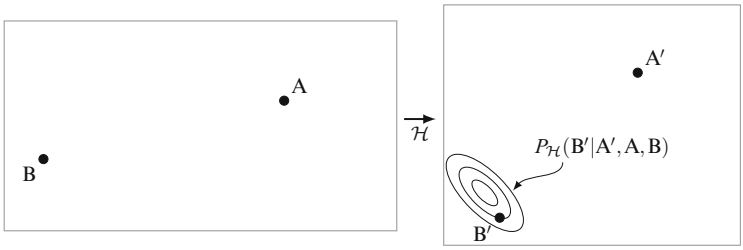


Fig. 5 Conditional probability of $\mathbf{B}'$ based on model assumptions.

Estimate $\mathcal{H}$ Another means of dealing with the perturbations introduced by $\mathcal{H}$ is to estimate them and then compensate for them. For instance, the popular concept of detecting key points in the scale space of a signal yields a local estimate for $\mathcal{H}$ [17, 18]. Low-level representations such as feature descriptors then compensate locally for the estimated $\mathcal{H}$ and compute the subsequent algorithmic steps in a frame that is normalized with respect to the estimated $\mathcal{H}$. This method supposes that the perturbations can be estimated, at least locally. A common strategy is to restrict $\mathcal{H}$ to a smaller class of parametric models, such as globally or locally affine transformations. This limits the applicability of the approach to either $\mathcal{H}$ that are globally affine (and thus ignoring the much larger class of homeomorphisms) or modelling only small neighborhoods, where the locally affine assumption works reasonably well. The latter effectively limits the scale at which models can be built for solving signal problems if $\mathcal{H}$ is not affine over longer ranges.

An alternative that avoids uncertainty as well as the need to estimate $\mathcal{H}$ explicitly is to relate $\mathbf{A}$ and $\mathbf{B}$ in a manner that is invariant to $\mathcal{H}$. If one is able to find an invariant way to get from $\mathbf{B}$ given $\mathbf{A}$, then the impact of $\mathcal{H}$ is compensated by virtue of the invariance. In the following we elaborate such invariant relations in general and subsequently derive a specific example for common multi-dimensional signals.

2 Invariance, Uniqueness, and Completeness

In the previous section we established that we may be able to relate locations in an invariant manner. This section introduces and formalizes the properties of invariance, completeness, and uniqueness of relations in the general case (i.e., with or without a signal function). The following section 3 then introduces a limited class of signals and studies invariant relations based on describing the space between locations based on topological connectedness.

2.1 Definitions and Examples

Let the description be from the set Λ in the sense that there are functions $\lambda \in \Lambda : \mathbb{R}^n \to \{x : x \in \mathbb{R}^n\}$ that map points into sets. Further let $\theta \in \Theta : \mathbb{R}^n \to \mathbb{R}^n$ be the family of perturbations that are under consideration (e.g., the set of all affinities or all homeomorphisms $\mathcal{H}$) and two points $\mathbf{A}, \mathbf{B} \in \mathbb{R}^n$. In the following we will call $\lambda(\mathbf{A})$ the *feasible set* of $\mathbf{A}$ given λ.

Definition 1 Λ is *forward invariant* under Θ if

$$\forall(\mathbf{A}, \mathbf{B}, \lambda, \theta) : \mathbf{B} \in \lambda(\mathbf{A}) \Rightarrow \theta(\mathbf{B}) \in \lambda(\theta(\mathbf{A})).$$

Definition 2 If all $\theta \in \Theta$ have an inverse, Λ is *backward invariant* under Θ if

$$\forall(\mathbf{A}, \mathbf{B}, \lambda, \theta) : \mathbf{B} \in \lambda(\theta(\mathbf{A})) \Rightarrow \theta^{-1}(\mathbf{B}) \in \lambda(\mathbf{A}).$$

Definition 3 Λ is *invariant* under Θ if it is forward invariant and backward invariant or alternatively if the descriptions λ and perturbations θ commute:

$$\forall(\mathbf{A}, \lambda, \theta) : \theta(\lambda(\mathbf{A})) = \lambda(\theta(\mathbf{A})).$$

In order to avoid trivially invariant solutions such as $\lambda(x) = \{\emptyset\}$, we consider *completeness:*

Definition 4 $\Lambda$is $S-$complete with respect to a set $S \subseteq \mathbb{R}^n$ if

$$\forall(\mathbf{A} \in S, \mathbf{B} \in S), \exists\lambda : \mathbf{B} \in \lambda(\mathbf{A}).$$

Definition 5 Λ is complete if it $S-$complete for $S = \mathbb{R}^n$.

In order to avoid trivially complete solutions such as $\lambda(x) = \mathbb{R}^n$, we consider *uniqueness:*

Definition 6 Λ is unique if

$$\forall \mathbf{A}, \lambda : |\lambda(\mathbf{A})| \leq 1$$

Note that the properties of completeness and uniqueness are independent of the set of perturbations Θ. Before we elaborate in more detail how Λ may look in practice, we have to realize that invariance comes at a cost: the properties of invariance, uniqueness, and completeness may be jointly incompatible for some Θ. As an example one can consider the Euclidean distance relation in this framework by letting

$$\Lambda^e : \{x \to x + \delta : \delta \in \mathbb{R}^n\}$$

with the elements $\lambda^e_\delta(x) := x+\delta$. This particular set Λ^e is unique as any x is mapped into $x+\delta$, a set of cardinality 1. It is also complete, which can be seen by considering that there is a description $\lambda^e_{\mathbf{B}-\mathbf{A}}$ for each point pair:

$$(\mathbf{A}, \mathbf{B}) \to \lambda^e_{\mathbf{B}-\mathbf{A}}(x) = x + \mathbf{B} - \mathbf{A}$$

When it comes to invariance, one has to choose a family of perturbations Θ. An example is the family of translations in $\mathbb{R}^n$:

$$\theta^t \in \Theta^t : \{x \to x + t : t \in \mathbb{R}^n\}.$$

Then for any $\mathbf{A}, \lambda^e_\delta, \theta^t$, there is exactly one element in $\lambda^e_\delta(\mathbf{A})$: $\mathbf{A} + \delta$. Furthermore,

$$\begin{aligned}\theta^t(\lambda^e_\delta(\mathbf{A})) &= \theta^t(\mathbf{A} + \delta) = \mathbf{A} + t + \delta \\ \lambda^e_\delta(\theta^t(\mathbf{A})) &= \lambda^e_\delta(\mathbf{A} + t) = \mathbf{A} + t + \delta.\end{aligned}$$

i.e., the condition in definition 1 is met and Λ^e is invariant to translations. Another example of perturbations is the family of uniform scalings:

$$\theta^s \in \Theta^s : \{x \to sx : s \in \mathbb{R}\}.$$

For any $\mathbf{A}, \lambda^e_\delta, \theta^s$, there is again exactly one element in $\lambda^e_\delta(\mathbf{A})$: $\mathbf{A} + \delta$. Furthermore,

$$\begin{aligned}\theta^s(\lambda^e_\delta(\mathbf{A})) &= \theta^s(\mathbf{A} + \delta) &= s(\mathbf{A} + \delta) \\ \lambda^e_\delta(\theta^s(\mathbf{A})) &= \lambda^e_\delta(s\mathbf{A}) &= s\mathbf{A} + \delta.\end{aligned}$$

Thus θ^s and λ^e_δ do not commute, violating the condition in definition 1 and Λ^e is not invariant to scale (and thus to homeomorphisms in general). In order to gain more insight, we can play around with Λ^e and see what happens. Let's modify Λ^e as such:

$$\Lambda^{\bar{e}} : \{x \to \{x + \bar{s}\delta : \bar{s} \in \mathbb{R}\} : \delta \in \mathbb{R}^n\}. \tag{1}$$

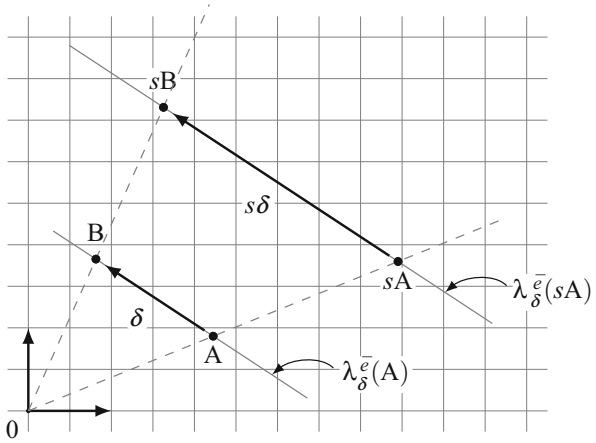


Fig. 6 Invariance of $\Lambda^{\bar{e}}$ under uniform scaling Θ^S. The feasible sets $\lambda^{\bar{e}}_{\delta}(\mathbf{A})$ and $\lambda^{\bar{e}}_{\delta}(s\mathbf{A})$ are shown as gray lines.

In other words, instead of a single location as in the Euclidean distance relation, we now allow the line through x along the direction δ in the feasible set. Now $\lambda^{\bar{e}}_{\delta}(\mathbf{A})$ is a set of cardinality larger than one, specifically

$$\lambda^{\bar{e}}_{\delta}(\mathbf{A}) = \{\mathbf{A} + \bar{s}\delta : \bar{s} \in \mathbb{R}\}$$

Figure 6 illustrates this situation. Under the family of perturbations Θ^s, the invariance of $\Lambda^{\bar{e}}$ can again be tested via

$$\begin{aligned} \theta^s(\lambda^{\bar{e}}_{\delta}(\mathbf{A})) &= \theta^s(\{\mathbf{A} + \bar{s}\delta : \bar{s} \in \mathbb{R}\}) = \{s\mathbf{A} + s\bar{s}\delta : \bar{s} \in \mathbb{R}\} \\ \lambda^{\bar{e}}_{\delta}(\theta^s(\mathbf{A})) &= \lambda^{\bar{e}}_{\delta}(s\mathbf{A}) = \{s\mathbf{A} + \bar{s}\delta : \bar{s} \in \mathbb{R}\} \end{aligned}$$

The resulting sets are equivalent and consequently the modified $\Lambda^{\bar{e}}$ is now invariant to scaling. $\Lambda^{\bar{e}}$ is also complete as for any $\mathbf{A}, \mathbf{B}$ there is a $\lambda^{\bar{e}}_{\delta}$ for which $\mathbf{B} \in \lambda^{\bar{e}}_{\delta}(\mathbf{A})$, specifically for $\delta = \mathbf{B} - \mathbf{A}$. It is, however, not unique any more as $|\lambda^{\bar{e}}_{\delta}| > 1$. In summary, starting from the Euclidean distance relation under scalings, which is complete and unique, but not invariant, we are able to derive a new relation that is invariant and complete, but not unique. Is there a way to also make it unique? Let's start with the modified $\Lambda^{\bar{e}}$ and select just one element $\dot{s}$ of the line, with $\dot{s} = |x|$. The new $\Lambda^{\dot{e}}$ then becomes

$$\Lambda^{\dot{e}} : \{x \to x + |x|\delta : \delta \in \mathbb{R}^n\}$$

Table 1 Characteristics of the four representations under perturbation

	Λ^e under Θ^T	Λ^e under Θ^S	$\Lambda^{\bar{e}}$ under Θ^S	$\Lambda^{\dot{e}}$ under Θ^S
Invariant	yes	no	yes	yes
Unique	yes	yes	no	yes
Complete	yes	yes	yes	no
$S-$Complete				$\mathbb{R}^n \setminus \{0\}$

Then

$$\begin{aligned} \theta^s(\lambda^{\dot{e}}_\delta(\mathbf{A})) &= \theta^s(\mathbf{A} + |\mathbf{A}|\delta) &= s(\mathbf{A} + |\mathbf{A}|\delta) \\ \lambda^{\dot{e}}_\delta(\theta^s(\mathbf{A})) &= \lambda^{\dot{e}}_\delta(s\mathbf{A}) &= s(\mathbf{A} + |\mathbf{A}|\delta). \end{aligned}$$

Hence, $\Lambda^{\dot{e}}$ is unique and invariant under Θ^S. It is important to realize, however, that it is not complete any more. To see this, consider that the representation for a particular pair $(\mathbf{A}, \mathbf{B})$ is the element $\lambda^{\dot{e}}_{\delta'} \in \Lambda^{\dot{e}}$ with

$$\delta' = \frac{(\mathbf{B} - \mathbf{A})}{|\mathbf{A}|}, \text{ where } |\mathbf{A}| \neq 0, \tag{2}$$

as $\mathbf{B}$ then is in the feasible set of $\mathbf{A}$:

$$\mathbf{B} \in \lambda^{\dot{e}}_{\delta'}(\mathbf{A}) = \mathbf{A} + |\mathbf{A}|(\mathbf{B} - \mathbf{A})/|\mathbf{A}| = \mathbf{B}.$$

Due to equation 2, any pair $(0, \mathbf{B})$ has no representation $\lambda^{\dot{e}}_\delta$ in $\Lambda^{\dot{e}}$ and thus it is not complete. It is, however, $S-$complete for $\mathbb{R}^n \setminus \{0\}$ (and possibly subsets). Table 1 lists the properties for the various combinations of Λ and perturbations Θ so far.

2.2 General Case

Are there always families of relations that are unique, complete, and invariant for a given family of perturbations? Certainly not if Θ^S has no inverse, as no Λ could be backward invariant and thus invariant. What about invertible perturbations? Let's look, for instance, at the set of all affinities with uniform scaling

$$\Theta^{st} : \{x \to sx + t : s \in \mathbb{R}, t \in \mathbb{R}^n\}$$

and an example pair for $n = 2$: $\mathbf{A} = (0, 1)$, $\mathbf{B} = (^1/_2, ^1/_2)$ as illustrated in figure 7. Without loss of generality we choose $\theta^{st} \in \Theta^{st}$ that transform $\mathbf{A}$ into $\mathbf{A}' = (0, 2)$. There are two such perturbations

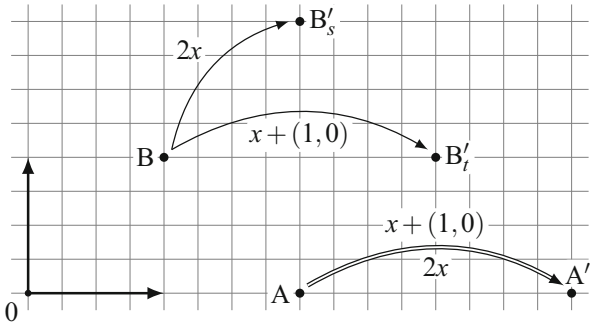


Fig. 7 Counterexample to the existence of invariant, complete, and unique Λ in the general case. There are two θ from the set of affinities with uniform scaling that transform $\mathbf{A}$ into the same $\mathbf{A}'$, but $\mathbf{B}$ into two different $\mathbf{B}'$.

$$x \to 2x$$
$$x \to x + (0, 1)$$

and they transform $\mathbf{B}$ into $\mathbf{B}' = (1, 1)$ and $\mathbf{B}' = ({}^1/_2, 1{}^1/_2)$ respectively. Now we assume that there is a representation λ for which $\mathbf{B}$ is in the feasible set of $\mathbf{A}$, i.e. $\mathbf{B} \in \lambda(\mathbf{A})$. As the true perturbation is unknown, invariance of Λ would require that $\{(1, 1), ({}^1/_2, 1{}^1/_2)\}$ is a subset of $\lambda(\mathbf{A}')$. This violates uniqueness, as $|\lambda(\mathbf{A}')| \geq 2$. One can certainly choose to make Λ unique by removing all but one element, e.g. $\lambda(\mathbf{A}') := \{(1, 1)\}$. However, then it is not invariant any more, as forward invariance for the tupel

$$(\mathbf{A}, \mathbf{B}, \lambda, \theta) = ((0, 1), ({}^1/_2, {}^1/_2), \lambda, x \to x + (0, 1))$$

is violated: the perturbed $\mathbf{B}' = ({}^1/_2, 1{}^1/_2)$ is not an element of $\lambda(\mathbf{A}') = \{(1, 1)\}$. This shows that there is no Λ in general that is invariant, unique, and complete without restricting the perturbations Θ, and in particular not for homeomorphisms as a superset of affinities.

This section introduced the concepts of invariance, uniqueness, and completeness for representations of pairwise point relations. The next section will consider the case where there is a signal, or more precisely, a vector-valued function $\mathbb{R}^n \to \mathbb{R}^m$ on the domain and explore what that means for the representation Λ.

3 Invariant Relations in Line Drawings

The following will focus on a subset of all possible signals, the set of all monochrome line drawings in $\mathbb{R}^2$. Figure 8 shows an example line drawing under a homeomorphism $\mathcal{H} \in \Theta^h$. There are two locations $\mathbf{A}$ and $\mathbf{B}$ marked with their transformed counterparts $\mathbf{A}' = \mathcal{H}(\mathbf{A})$ and $\mathbf{B}' = \mathcal{H}(\mathbf{B})$.

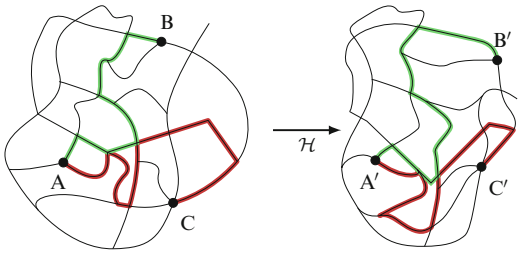


Fig. 8 Line drawing under Homeomorphism $\mathcal{H}$ with three locations **A**,**B**, and **C** marked. The green highlighted path relates **A** and **B** in terms of a topological description.

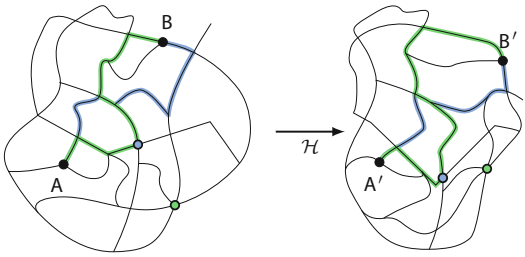


Fig. 9 Line drawing with two paths from **A** to **B** marked in green and blue. Both the green and blue path description individually may lead to other locations (marked by colored circles). However, **B** is the only location that can be reached by the blue and the green path description simultaneously.

One description λ from **A** to **B** may be transliterated as (indicated as green path in figure 8): "Start from **A** and traverse a line, there will be an intersection of order four. Follow the third branch counterclockwise and then follow the leftmost branch in the next two intersections. Take the second right and then again the second right. The next intersection is an element of the feasible set (which contains **B**)." The same holds for $\mathbf{A}'$ and $\mathbf{B}'$, i.e. the existence of the same path can be ascertained in the perturbed image. It does not matter how distorted the domain is, as long as there are no tears or rips that would interrupt the path. Consequently, this path description example is invariant for the pair $(\mathbf{A}, \mathbf{B})$ (figure 9). Is it also unique?

3.1 Uniqueness

Uniqueness would require that $\lambda(\mathbf{A}) \setminus \mathbf{B} = \emptyset$. However, when one starts with a different branch from **A**, one may end up in the location **C** instead, marked with the red path in figure 8. There are no further paths consistent with the description, thus the feasible set $\lambda(\mathbf{A}) = \{\mathbf{B}, \mathbf{C}\}$ (and via invariance $\lambda(\mathbf{A}') = \{\mathbf{B}', \mathbf{C}'\}$). In other words, this particular λ is not unique. One may certainly aim to make λ unique. As an example the above transliteration could be modified as (modification emphasized): "Start from **A** and traverse a line, [...] The next intersection is an

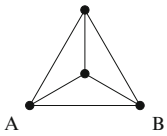


Fig. 10 Counterexample for existence of invariant, complete, and unique Λ under homeomorphisms.

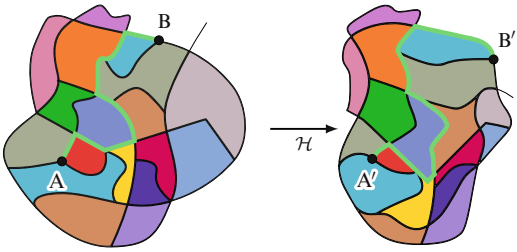


Fig. 11 From line drawings to piecewise constant images.

element of the feasible set *if the order of the intersection is three*." Only the green path in figure 8 is consistent with the modified λ, and thus only **B** is in its feasible set, making the modified λ unique.

However, in the general case uniqueness, completeness, and invariance cannot be achieved for this class of signals. Consider, for instance, the line drawing in figure 10. Due to symmetry, there is no unique way to describe a path from **A** to **B** that is invariant under arbitrary homeomorphisms. One can certainly always look at a more complex signal class and hope that uniqueness can be achieved there. One example may be the class of piecewise constant images, as shown in figure 11. The path description now would be: "Starting from **A** and following a path that has a gray patch on its left and a red patch on its right, follow paths with these consecutive left/right color attributes: (blue, red), (blue, yellow), (blue, brown), (blue, gray), (orange, gray), (orange, blue), (. . ., blue). The next intersection will be an element of the feasible set." Although this description is now unique in our example, uniqueness is not guaranteed for the general case either (as the set of line drawings is a subset of the set of piecewise constant images). However, once we start looking pragmatically at the set of natural images, the situation may not be as bleak. In the following section 4, we will derive a particular implementation of our representation based on topological connectedness that works well with natural images in practice.

3.2 Intersection of feasible sets

A generic scheme to deal with the potential lack of uniqueness is to use a multitude of descriptions for the same pair $(\mathbf{A}, \mathbf{B})$. Figure 9 shows in addition to the first description in green also the following description in blue: "Start from **A** and traverse a line, there will be an intersection of order four. Take the second path clockwise, then make a right, and three lefts. The next intersection is an element of

the feasible set." Both feasible sets contain two locations each, but their intersection only contains **B** (and **B**′). More formally, based on individual path descriptions $\lambda_1, \ldots, \lambda_n$, one can define the combined $\bar{\lambda}$ as

$$\bar{\lambda} = \bigcap_{i=1}^{n} \lambda_i$$

This scheme is motivated not so much from a principled perspective (as one could always formally treat $\bar{\lambda}$ as an "individual" description from the start, albeit a more complex one), but from considerations on computational implementations that will be elaborated later: an independent set of shorter path descriptions may be more computationally efficient than one long description.

3.3 Completeness

An interesting case is now the completeness of a hypothetical Λ. We choose to consider only Λ that are invariant under homeomorphisms. For completeness in this example we need to consider $\mathbb{R}^2$. The simplest example of f is the constant function, $f = c$. Any homeomorphism on the domain of f will not change that f is constant everywhere. That implies that we are not gaining anything by using f for the definition of Λ and we have no invariant, complete, and unique Λ as discussed in section 2.2. One can now appeal to practical considerations from a signal processing perspective and argue that not all parts of the signal are of equal importance. For instance, homogeneous regions of the signal may not carry much information (they carry some, such as "I am homogeneous"). Conversely, a signal region with high entropy may be considered interesting. In terms of our representation this implies that the modeling should focus on these interesting regions. In practice it may be sufficient to find a representation that is invariant, unique, but only $S-$complete, where the set S contains the relevant information for a particular problem.

Due to the lack of invariant, complete, and unique representations in the general case, one always has to find a trade-off between them, which may be specific to an application. The next section introduces a generic scheme in the continuous domain that allows to strike different such trade-offs and for which efficient discrete approximations are possible.

4 A Continuous Trace Model

This section introduces a practical representation Λ for common real-world signals. Without loss of generality we look at natural images as a running example. As alluded to earlier, we make a design choice at this point to factorize the

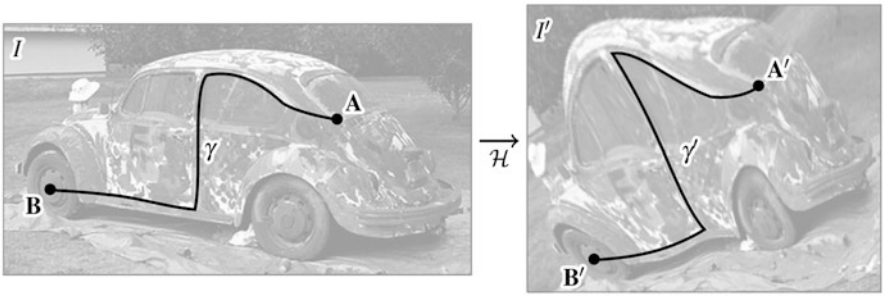


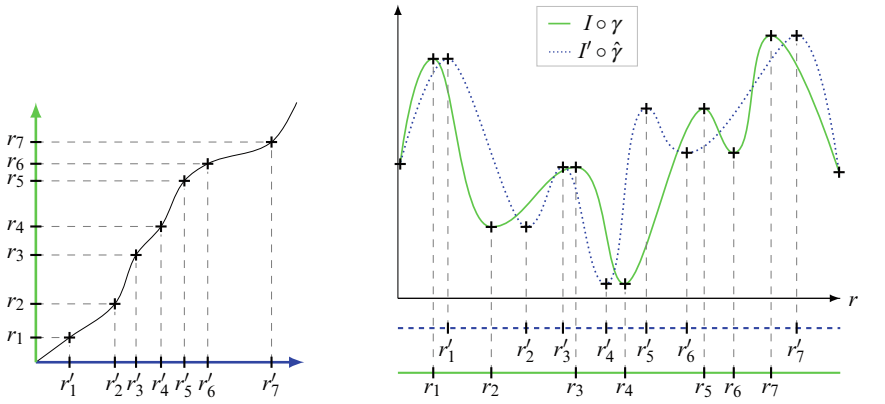
Fig. 12 An image I is spatially deformed by a homeomorphism $\mathcal{H}$ into the image I'. The points **A** and **B** are mapped into $\mathbf{A}'$ and $\mathbf{B}'$, and the spatial curve γ is mapped into the curve γ'.

representation into a set of one-dimensional problems due to later computational considerations. This is certainly not the only option and it is conceivable to pose the model differently. The following sections introduce the model formally and discuss its properties.

4.1 Definitions

Let I be a continuous image over the domain $\mathbb{R}^2$, i.e. $I : \mathbb{R}^2 \mapsto \mathbb{R}$ as shown in figure 12. The image is perturbed by a spatial deformation, in the most general case by a homeomorphism $\mathcal{H} : \mathbb{R}^2 \mapsto \mathbb{R}^2$. $\mathcal{H}$ is a continuous function with continuous inverse and maps the image I into its perturbed version $I' = I \circ \mathcal{H}^{-1}$, where the symbol "$\circ$" denotes function composition: $(f \circ g)(x) := f(g(x))$. Continuity implies that there are no tears, rips, or rifts in the mapping and that the local neighborhood structure is preserved between I and I'. Furthermore, the locations **A** and **B** in the image I are mapped into the locations $\mathbf{A}' = \mathcal{H}(\mathbf{A})$ and $\mathbf{B}' = \mathcal{H}(\mathbf{B})$ in the image I'. The goal is to find a way to relate **A** and **B** (and thus $\mathbf{A}'$ and $\mathbf{B}'$) by a description of the image information between them that is invariant to $\mathcal{H}$.

As pointed out in section 3, any such description cannot in general be unique and complete at the same time. The particular choice of the description determines the trade-off between uniqueness and completeness. The chosen approach in the next sections is based on the supposition that it is more useful in practice to have a complete description, i.e. to be able to describe all locations, even if it is not unique, or ambiguous. Spurious ambiguities may additionally be resolved in higher-level representations built on the presented framework, whereas it may be more difficult to recover from an inability to describe certain locations. Accordingly, the goal is to find a description for all possible $\mathbf{B}'$, i.e. a complete description which is then necessarily ambiguous for general signals. The challenge of minimizing ambiguity is then addressed by potentially using *all* available image information in the description by the intersection of feasible sets introduced in section 3.2.



(a) Monotonic function that reparametrizes r.

(b) The original trace $\tau = I \circ \gamma(r)$ and its transformed counterpart $\hat{\tau} = I' \circ \hat{\gamma}(r)$.

Fig. 13 An example profile trace under reparametrization of r.

4.2 Profile Trace

The continuous spatial curve $\gamma : [0, 1] \to \mathbb{R}^2$ as depicted in figure 12 connects the locations **A** and **B**, i.e. the endpoints of the curve coincide with the locations, respectively: $\gamma(0) = \mathbf{A}$, $\gamma(1) = \mathbf{B}$. Its equivalent in the transformed image under the homeography $\mathcal{H}$ is $\gamma' = \mathcal{H} \circ \gamma$. The spatially transformed curve γ' connects $\mathbf{A}'$ and $\mathbf{B}'$, i.e. $\gamma'(0) = \mathbf{A}'$, $\gamma'(1) = \mathbf{B}'$. The images I and I' as function of the curves γ and γ', respectively, have the profiles $\tau, \tau' : [0, 1] \mapsto \mathbb{R}$, where $\tau = I \circ \gamma$ and $\tau' = I' \circ \gamma'$ which shall be termed *profile trace* (or simply *trace* for short, as opposed to the curve γ). An example of a trace is shown in figure 13b and an example of a curve is shown in figure 12. The traces have the property that $\tau(r) = \tau'(r)$ at every point r because

$$\begin{aligned} \tau'(r) &= (I' \circ \gamma')(r) \\ &= (I' \circ \mathcal{H} \circ \gamma)(r) \\ &= (I \circ \gamma)(r) \\ &= \tau(r) \end{aligned} \tag{3}$$

In other words, the profile traces τ do not change under smooth deformations $\mathcal{H}$. This is not surprising however, as the perturbation $\mathcal{H}$ has been used explicitly in the construction of the perturbed trace τ'. In general, certainly, the perturbation $\mathcal{H}$ is not known a priori. Furthermore, equation 3 cannot immediately be used to construct some invariant property that is measurable in the image, as the curve γ' itself is not directly observable. A weaker, but ultimately more useful statement, is that *there*

exists some curve $\hat{\gamma}$ *between* $\mathbf{A}'$ *and* $\mathbf{B}'$ *with the same trace* $\hat{\tau} = I' \circ \hat{\gamma} = \tau$. This is strictly a weaker criterion as the curves $\hat{\gamma}$ and γ' are not necessarily one and the same. The critical realization is summarized in the following proposition:

Proposition 1 *The existence property of* $\hat{\gamma}$ *(or* γ'*) is not a function of* $\mathcal{H}$ *and thus invariant under* $\mathcal{H}$.

Now, the goal is to restrict the true location of $\mathbf{B}' = \mathcal{H}(\mathbf{B})$ given the image I', a profile trace τ, and the location $\mathbf{A}' = \mathcal{H}(\mathbf{A})$. The following holds with regard to the location of $\mathbf{B}'$:

Proposition 2 *A necessary condition for any* $\mathbf{C}'$ *being the true location* $\mathbf{C}' = \mathbf{B}'$ *is the existence of a curve* $\hat{\gamma}$ *such that* $\hat{\gamma}(0) = \mathbf{A}'$, $\hat{\gamma}(1) = \mathbf{C}'$ *and that the resulting trace* $\hat{\tau}$ *is equivalent to the trace* τ.

Another way of stating proposition 2 is to find an equivalent trace $\hat{\tau}$ up to some reparametrization of the underlying curve γ'. This is now the structure of our Λ and its elements have the form

$$\lambda_\tau(\mathbf{A}) = \{x : \; s.t. \, \exists \gamma : \gamma(0) = \mathbf{A}, \gamma(1) = x, I \circ \gamma = \tau\}. \tag{4}$$

While the considerations in this section are purely in terms of existence, i.e. not in terms of a particular computational approach, the structure of possible computational solutions can already be seen: Proposition 2 decomposes the problem of modeling the location of $\mathbf{B}$ into two components. The first component is an enumeration of paths between two locations and the second component is the establishing of equivalence between traces along these paths. This decomposition allows to consider the path identity and trace equivalence separately. An important consequence of this decomposition is that the spatial perturbations can largely be modeled independently from other signal perturbations. Furthermore, the trace can be designed specifically to address a given perturbation prior and specific trade-offs between accuracy and run-time performance. This is exemplified in section 6.

Figure 13b shows an illustrative example of a profile trace, i.e. the function $\tau = (I \circ \gamma)(r)$ between two locations $\mathbf{A} = \gamma(0)$ and $\mathbf{B} = \gamma(1)$. The abscissa of the graph is the curve parameter $r \in [0, 1]$ and the ordinate is the value of the image function I at the location $\gamma(r)$. The curve γ itself is not shown, figure 12 serves as an example. As shown above, the trace in the perturbed image I', $\tau' = (I' \circ \gamma')(r)$, is equivalent to τ and thus has the same graph. Let $\hat{\gamma}$ be the same curve as γ' up to a reparametrization of r, i.e. it is the same curve, but its speed varies. An example of a reparametrization is illustrated in figure 13a. $\hat{\gamma}$ also has the same domain and range as γ'. The domain of γ' is the interval $[0, 1]$ and its range is the set of locations in the image that it covers between $\mathbf{A}$ and $\mathbf{B}$. The corresponding traces $\tau' = \tau$ and $\hat{\tau}$ can then be understood as related by a dynamic time warp [22]. Figure 13b shows the original trace $\tau = I \circ \gamma$, as well as the corresponding warped trace $\hat{\tau} = I' \circ \hat{\gamma}$. It can be seen that the overall evolution of both graphs is the same with one curve trailing or leading the other. In order to make practical use of the trace model, proposition 2 requires that the equivalence between two

traces τ and $\hat{\tau}$ can be established. Naturally, this could be done by solving the dynamic time warp problem for exact point-wise equivalence. However, this may not be computationally favorable. An alternative is to relax the strict point-wise equivalence between traces, preferably with a relaxation that is invariant to dynamic time warps. The following section describes an example of such a relaxation based on rank order consistency.

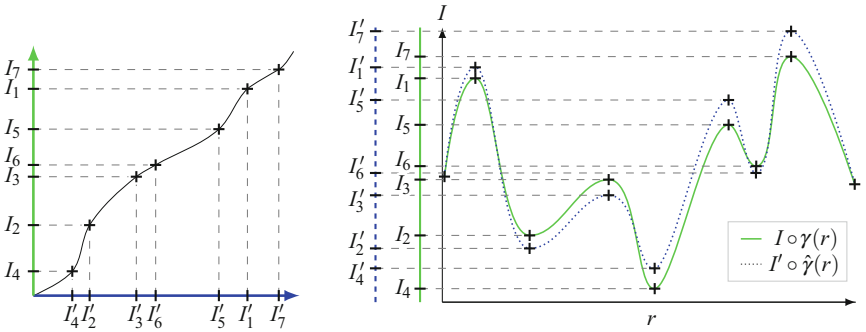
4.3 Invariance to Dynamic Warps

A countable set can be sorted according to a partial order relation. The rank of each element after sorting is its ordered rank, or *rank order*. Under monotonic transformation of the initial set, the rank order is retained, i.e. the rank order is invariant to the transformation [23]. The transformation of interest here is the dynamic time warp reparametrization of the curve $\hat{\gamma}$, whose monotonicity is illustrated in figure 13a. Figure 13b shows the original trace and its perturbed counterpart with the minima and maxima between **A** and **B** labeled sequentially from one to seven. The bottom lines depict the coordinates of the extremal points for both traces as $r_{1,\ldots,7}$ and $r'_{1,\ldots,7}$, respectively. The order of the extremal points with respect to the coordinate r does not change between the original and the perturbed trace. In other words, the sequence of extremal points is invariant to dynamic warps:

$$\{(I \circ \gamma)(r_1, \ldots, r_7)\} = \{(I' \circ \hat{\gamma})(r'_1, \ldots, r'_7)\}. \tag{5}$$

More importantly the coordinates r and r' do not need to be known explicitly, as the extrema can readily be estimated from the image profile, and the equivalence in equation 5 can be verified just from the extrema. Informally this relaxed equivalence states that two traces are equivalent, if their extremal points are the same and in the same order. Any signal property that has a similar topological nature such as zero crossings can be used in this way to define trace equivalence while being invariant to spatial perturbations. With respect to proposition 2 and strict point-wise equivalence, equation 5 is a necessary but not sufficient condition, i.e. it is an equivalence that is less strict and it may increase the ambiguity of the location of **B**. This is a trade-off between computational complexity and uniqueness.

In practice, signals undergo extraneous perturbations such as illumination change and image noise. The next sections extend the trace model to address these in more detail. Although it is not pursued in this work, it is in principle possible to statistically model the equivalence in equation 5 to incorporate the effects of noise, e.g. the likelihood of two extrema matching each other or the likelihood of two extrema changing rank orders.



(a) Non-linear monotonic transfer function between I and I'.

(b) The original trace $I \circ \gamma(r)$ and its transformed counterpart $I' \circ \hat{\gamma}(r)$.

Fig. 14 An example profile trace under non-linear monotonic illumination change.

4.4 Invariance to Monotonic Illumination Change

The previous section demonstrated how to achieve an equivalence function that is invariant to dynamic warps of the curve $\hat{\gamma}$. This section shows how to achieve invariance to non-linear monotonic illumination changes by a similar argument based on order consistency. Such changes in illumination arise due to a variety of reasons, such as a changing scene lighting or dynamic CCD camera effects [21] and they can be instantaneous or gradual [23]. Figure 14a shows a non-linear monotonic transfer function between I and I', and figure 14b shows its effect on the original trace $I \circ \gamma$ as the perturbed trace $I' \circ \hat{\gamma}$. For clarity of exposure in this example the curve parametrization r is assumed to be the same, i.e. there is no dynamic warp. As can be seen in the graphs, the values of the perturbed function change in a non-linear fashion. On the left side of the graph, the values of the extremal points for each trace are shown with the same ordering of the locations from left to right as in the previous example. The order of the extremal values of the image function I, when sorted from smallest to largest value for the original trace, is

$$(4, 2, 3, 6, 5, 1, 7).$$

By virtue of the monotonicity of the illumination change, the same order holds for the perturbed trace. Consequently the rank order of the extremal values can define trace equivalence. This equivalence will be invariant to monotonic illumination change. Specifically, two traces are considered equivalent if the order of their extremal points is the same.

4.5 Generalized Texture Trace

The profile trace is defined in terms of the profile of the image, i.e. the image I as a function of a spatial one-dimensional curve γ. This allows the derivation of invariant properties under the assumption that the image is transformed by a spatial process. Also, the previous section has shown how to include invariance to non-linear monotonic illumination change into the model. However, up to this point the model is based on the image function along γ which has two practically significant issues. Firstly, in certain applications the image function may have higher variability along the same curve between signal instances due to presently unmodeled perturbations. These may include sampling effects, intra-class variation, non-monotonic illumination change, and so on. Secondly, it does not use information in the immediate neighborhood of the profile curve which could be used to decrease the ambiguities of the model. This section proposes a formal extension of the profile trace to include neighborhoods along the spatial curve in order to decrease the ambiguity of the representation and to allow the incorporation of further invariances. Let

$$F(x_0) : \{(x, I(x)) : \|x - x_0\| \leq s\} \mapsto \mathbb{R}^n \tag{6}$$

be a function that assigns a vector to each location x_0 on the curve γ in the image based on the image information in a neighborhood s. This definition of F includes a wide variety of functions such as convolutions with finite support, edge detection, feature computations such as SIFT [15], SURF [4], etc. Informally, the idea is to use local characteristics or texture properties of the image instead of just the image profile in order to reduce ambiguities of the representation. This also allows another layer of abstraction above the image function. In the case that the function F itself is invariant to homeomorphisms, it can be incorporated into the trace definition while maintaining homeomorphic invariance of the trace model in a straightforward manner: Proposition 2 has to be modified such that the profile trace definition

$$\tau(r) = (I \circ \gamma)(r)$$

is augmented by the definition of the *texture trace*:

$$\tau(r) = (F \circ I \circ \gamma)(r) \tag{7}$$

A specific example of a homeomorphic invariant function F based on inflection points is

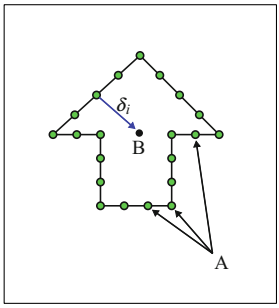
$$F : x_0 \mapsto \begin{cases} 0 & \text{if } \mathrm{x}_0 \text{ is minimum along } \gamma^{\perp}(x_0) \\ 1 & \text{if } \mathrm{x}_0 \text{ is maximum along } \gamma^{\perp}(x_0) \\ 2 & \text{if } \mathrm{x}_0 \text{ is saddle point along } \gamma^{\perp}(x_0) \\ 3 & \text{otherwise.} \end{cases} \tag{8}$$

where $\gamma^{\perp}(x_0)$ denotes the normal of the curve γ at x_0. This simple example considers infinitesimally small neighborhoods around each location on the curve along the normal of the tangent. Due to the differential definition, the inflection point properties are invariant to homeomorphisms. Additionally, due to the choice of order relations in equation 8 (essentially "greater than left and right," "smaller than left and right," etc.), this particular F is also invariant to monotonic changes of the illumination.

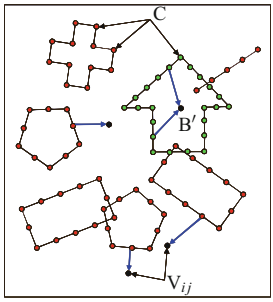
However, many practically interesting choices for local texture models such as SIFT are not homeomorphic invariant. They can still be employed, although the full invariance of the model is then relinquished, yielding a quasi-invariant representation. For a given choice of local texture model, the resulting extent of quasi-invariance is typically related to the neighborhood size, i.e. the smaller the neighborhood, the closer the representation comes to full invariance. The idea of the generalized texture trace to include further invariances as well as to decrease ambiguity will be employed and made concrete in the discrete texture trace of section 6.

5 Plugin for Metric Pairwise Relations

This section demonstrates how to plug in an invariant representation Λ into an existing algorithm in order to render it invariant to a selected set of transformations. Candidate algorithms need to use the Euclidean (or a related) metric in a pairwise fashion (this precludes, for instance, algorithms that inherently model tuples of points with size larger than two). Extensions of the presented model to beyond pairwise models are conceivable, but not explored here. In order to demonstrate the plug-in concept, we chose the well-known Generalized Hough Transform [3] for detecting shapes in images. Figure 15 illustrates the basic idea of the Generalized Hough Transform and algorithm 1 is a baseline version of the Hough algorithm. The objective is to detect a known object in the presence of distractor objects and perturbations of the domain. Initially, the reference object is *trained*, resulting in a *reference model*. The training consists of enumerating all points $\mathbf{A}$ on the boundary of the reference shape, choosing an arbitrary reference location $\mathbf{B}$ and storing their difference vectors δ_i as the reference model. The δ_i are also called *votes*. In the subsequent detection phase there is a set of shapes including the reference shape and it is not known which point belongs to which shape. The detection proceeds by applying all votes δ_i to all shape points $\mathbf{C}_j$ and storing the location $\mathbf{C}_j + \delta_i$ for which they vote.



(a) Reference shape for training.



(b) Target shape with distractor objects.

Fig. 15 Shape localization with the Hough transform. The foreground boundary points are marked in green, the distractor points are marked in red, and the reference location is marked with a cross in the reference shape. A selection of relations δ_i and their corresponding voting locations $\mathbf{V}_{ij}$ in the target space are marked with blue arrows and black dots, respectively. The indices are omitted for clarity of exposure.

Algorithm 1 Localization algorithm based on Euclidean distance

```
1:  procedure GetPointRelations(B, A_1, ..., A_N)        ▷ Get reference relations for B
2:      for i ← 1, ..., N do                             ▷ Iterate over all data locations
3:          δ_i ← B − A_i                                ▷ Store difference vector between B and A_i
4:      end for
5:  end procedure
6:
7:  procedure FindLocation(δ_1, ..., δ_N, C_1, ..., C_K)   ▷ Find location in target
8:      for j ← 1, ..., K do                             ▷ Iterate over locations
9:          for i ← 1, ..., N do                         ▷ Iterate over reference relations
10:             V_ij ← λ_δi(C_j) = C_j + δ_i             ▷ Gather vote for C_j and δ_i
11:         end for
12:     end for
13:     B ← AggregateVotes({V_ij})                       ▷ Assign target location by aggregating votes
14: end procedure
```

Under mild assumptions on the nature of the reference shape and the distractor shapes, there will be a concentration of votes in the vicinity of the true center $\mathbf{B}'$ of the reference shape. This concentration can be estimated by an aggregation of the votes, for instance by a kernel density estimate or by discretizing the voting domain. The basic voting scheme can be extended to include rotation, scale, and other perturbations of the domain by modifying the voting space [3] or extending to texture descriptors [11], but not easily to local perturbations such as homeomorphisms. For the present discussion however, the basic algorithm suffices to illustrate the analogy in the trace model.

Two steps of algorithm 1 need to be modified in order to plug in the new representation. The first is in the training phase in line 3, the second in the detection phase in line 10. For both steps, we interpret the Euclidean distance as a set relation

Algorithm 2 Localization algorithm based on trace model

1: **procedure** GETPOINTRELATIONS($\mathbf{B}, \mathbf{A}_1, \ldots, \mathbf{A}_N$) ▷ Get reference relations for $\mathbf{B}$
2: **for** $i \leftarrow 1, \ldots, N$ **do** ▷ Iterate over all data locations
3: $\mathcal{T}_i \leftarrow \{\tau(\mathbf{A}_i, \mathbf{B})\}$ ▷ Store all traces between $\mathbf{B}$ and $\mathbf{A}_i$
4: **end for**
5: **end procedure**
6:
7: **procedure** FINDLOCATION($\mathcal{T}_1, \ldots, \mathcal{T}_N, \mathbf{C}_1, \ldots, \mathbf{C}_K$) ▷ Find location in target
8: **for** $j \leftarrow 1, \ldots, K$ **do** ▷ Iterate over locations
9: **for** $i \leftarrow 1, \ldots, N$ **do** ▷ Iterate over reference relations
10: $\mathbf{V}_{ij} \leftarrow \lambda_{\mathcal{T}_i}(\mathbf{C}_j) = \bigcap \lambda_\tau(\mathbf{C}_j), \quad \tau \in \mathcal{T}_i$ ▷ Gather vote for $\mathbf{C}_j$ and $\mathcal{T}_i$ as in eq. 4
11: **end for**
12: **end for**
13: $\mathbf{B} \leftarrow$ AGGREGATEVOTES($\{\mathbf{V}_{ij}\}$) ▷ Assign target location by aggregating votes
14: **end procedure**

Table 2 Comparison of the initial Euclidean and the invariant algorithm as a result of plugging in the trace relation

	Euclidean algorithm	Trace algorithm
Modeled entity	Two-dimensional shape	Textured image patch
Observable	Points on the boundary	Points on the image patch
Model	Euclidean relation of points to center	Trace relation of points to center
Localization	All points vote for center location with all model relations	All points vote for center location with all model relations
Invariance	Translations	Homeomorphisms

as laid out earlier. In the Euclidean version, the first step in line 3 gathers the parameters $\delta_i \leftarrow \mathbf{B} - \mathbf{A}_i$ as the representation of $\mathbf{B}$ given $\mathbf{A}_i$ and the second step in line 10 determines the feasible sets $\lambda_{\delta_i}(\mathbf{C}_j) = \mathbf{C}_j + \delta_i$ of the estimated $\mathbf{B}$ given each $\mathbf{C}_j$ as the votes. Based on this interpretation we now plug in a different choice of pairwise relation, such as the profile trace as presented in section 4.2. Algorithm 2 lists the steps with the changes to algorithm 1 marked in blue. Instead of the feasible set $\lambda_{\delta_i}(\mathbf{C}_j)$ of the Euclidean relation, we use $\lambda_{\mathcal{T}_i}(\mathbf{C}_j)$ as the intersection of the feasible sets $\bigcap \lambda_\tau(\mathbf{C}_j)$ of all traces. Table 2 summarizes the resulting properties of the original Euclidean and the resulting invariant algorithm. This algorithm will be used in the next section to represent an image patch as a basic signal component.

6 Discrete Approximation

In this section we derive a discrete approximation of the continuous texture trace in eq. 7. In order to make the discrete solution computationally feasible, we have to make trade-offs when it comes to invariance. Two different trade-offs are chosen, one that results in full rotational invariance, one without full rotational invariance.

After discretizing the continuous trace τ into its discrete approximation t, we plug it into algorithm 2. The resulting representation is then studied empirically in the subsequent section 7 to validate its performance with real-world data.

Three obstacles need to be overcome to make the trace model practical:

Continuous Model All considerations so far have been under the assumption of a continuous signal domain. Optical digital images are discretely sampled and under the assumption of appropriate filtering, the original bandwidth-limited continuous signal can be extracted. It is thus in principle possible to formulate a computational approach to the trace model in the continuous domain (e.g., via spectral methods). However, the approach taken in this work is rather to discretize the trace model to achieve a practical computational approximation. The discretization is a coarse approximation due to computational constraints and the performance may depend on the discretization granularity. It is important to note that the proposed discretization is not the only possible choice.

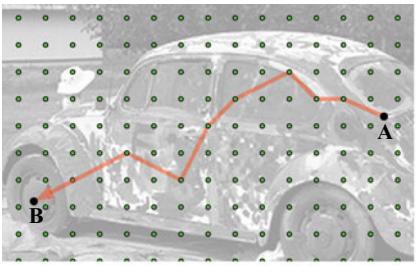
Lack of Completeness Not all point pairs can be described uniquely and invariantly, for instance in homogeneous signal regions. As pointed out earlier, homogeneous or otherwise ambiguous signal regions in practice often do not carry relevant or discriminative information. For this reason we choose to pragmatically accept the lack of completeness.

Lack of Uniqueness Path descriptions in real signals may be ambiguous. We address this by enumerating *all possible traces* between two locations given predefined bounds on the discrete representation length. This is in essence a complete topological characterization of the space between the two locations and it minimizes the ambiguity as much as possible within the bounds of the particular discretization parameters.

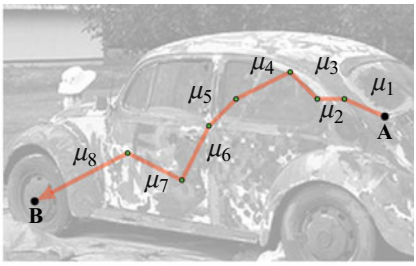
We will only sketch the steps in the discretization as this section primarily serves to demonstrate real-world implications of the overall scheme and the discretization is an incidental necessity. More detail on the discretization, detailed empirical evaluations on its effects on invariance and performance, and further applications beyond the following can be found in [7].

6.1 Discretization of Curve γ

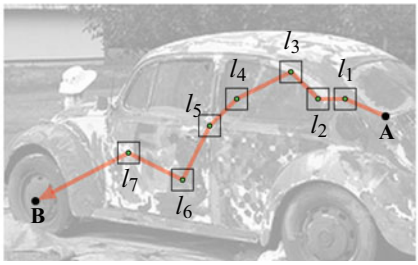
The objective of the discretization is a computational procedure that, given two locations $\mathbf{A}, \mathbf{B}$ in the image, allows the extraction of the traces between them as well as the determination of the feasible set of $\mathbf{B}'$ given $\mathbf{A}$ and the trace. An example curve for a trace is shown in figure 12. Figure 16a shows the first step, the discretization of the domain into a regular set of discrete locations (marked in green). The curve then is expressed in terms of only those locations, as shown in figures 16a–16d and the relative spatial relations between two points on the curve are described by a discrete set of local relations $\mu_i \in \Omega^\mu : \{(x, y) \rightarrow \{0, 1\}\}$. The relations μ_i are binary, i.e.



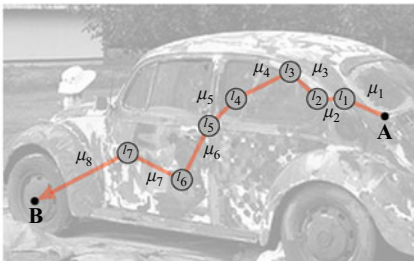
(a) Step one: Discretization of the curve γ.



(b) Step two: Assignment of discrete local relations μ.



(c) Step three: Assignment of discrete local texture labels l.



(d) The final model is a set of discrete texture labels that are locally connected via discrete spatialrelations.

Fig. 16 Steps in the discretization of the continuous trace model.

for any ordered pair of points (x, y) they either hold true or not. The last step of the discretization of the curve γ is to limit its discrete length to a finite number of steps n_d. The result is a discrete representation of γ: A finite sequence of discrete relations μ_i as illustrated in figure 16b.

6.2 *Quantization of Image Function*

Quantizing the image function in the discrete trace model implies that each location needs to be assigned a discrete label $l \in \Omega^L$ based on the image function. According to the generalization in section 4.5, any function may be used in a neighborhood around γ that is itself homeomorphic invariant or quasi-invariant. Functions with larger support may be more robust to image noise due to averaging effects while invariance to affine illumination changes may be achieved by using derivatives of the image function. There is a rich set of processes that assign quantized labels and also average over neighborhoods based on image derivatives, such as vector-quantized SIFT features [16] or other texture features [17]. Care has to be taken to account for the specific invariances of the texture features. In the following

we experimented with two different local features based on SIFT: one, where we compensate locally for orientation as estimated from the scale space, and one where we don't. These give rise to a rotation invariant and a rotation sensitive discrete representation accordingly.

Given a particular texture descriptor, a label l is assigned to a location y in the following manner. Firstly the texture feature is computed in a neighborhood around y and optionally compensated for the locally estimated orientation or other variants of the texture feature. Secondly, the resulting vector is quantized into $\left|\Omega^l\right|$ values by the use of a fixed code book. The code book can be generated, for instance, by a vector quantization scheme. In the following experiments, this is done via multiple repetitions of $k-$means [14] and choosing the instantiation with the least error in a large data set of images unrelated to the data in the results section.

6.3 Discretized Textured Trace

Putting everything together, a discrete texture trace then is defined as:

Definition 7 A discrete texture trace is a finite sequence of label-relationship pairs

$$t = ((l, \mu)_i : i = 1, \ldots, n_d) \in \Omega^t$$

of length n_d. Given a starting location a it induces the feasible set of locations b that are reachable from a via the trace t. A location b is reachable by t if there is a sequence of intermediate locations (y_k) such that

$$\mu_1(a, y_1) \times \mu_2(y_1, y_2) \times \ldots \times \mu_{n_d}(y_{n-1}, b) > 0$$

and the locations $(a, y_1, \ldots, y_{n_d-1})$ have labels $(l_1, l_2, \ldots, l_{n_d})$, respectively.

For a given input image, the locations y_k are sampled over the image domain with a fixed density. The discrete neighborhood structure and labeled landmarks y induce a graph $\mathcal{G} = (E, V)$ with the relations μ as edges E and the landmarks as labeled nodes V. The problem of determining the feasible sets then can be formulated as finding attributed paths in a graph. The set of *attributed adjacency matrices* $\left\{\mathbf{W}^{l\mu} : l \in \Omega^l, \mu \in \Omega^\mu\right\}$ of the graph $\mathcal{G}$ is defined as: $w_{ij}^{l\mu} > 0$ if the node i of label l has node j of arbitrary label connected to it by relation μ. Then, according to definition 7, the trace $t = (l, \mu)_i$ of length n_d relates the nodes a and b exactly if there is an intermediate sequence $Y = (y_1, \ldots, y_{n_d-1})$ such that

$$R(a, Y, b) = w_{(a,y_1)}^{(l\mu)_1} \left(\prod_{k=2}^{n_d-1} w_{(y_{k-1},y_k)}^{(l\mu)_k} \right) w_{(y_{n_d-1},b)}^{(l\mu)_{n_d}} > 0 \tag{9}$$

Equation 9 yields a computational approach for determining the feasible set of a trace t given a location a. Specifically, the feasible set $\lambda_t(a)$ is the set of all locations b that a relates to via the trace t and the above equation 9:

$$b \in \lambda_t(a) \Leftrightarrow \exists Y : R(a, Y, b) > 0. \tag{10}$$

The $\left|\Omega^l\right| \times |\Omega^\mu|$ matrices $\mathbf{W}^{l\mu}$ are large, but sparse and the existence of such a sequence can be established efficiently via sparse matrix multiplication. This concludes the derivation of a computable pairwise representation $\lambda_t \in \Lambda$, which will be used in the following experiments.

6.4 Patch Model

The remaining piece is how to model a part of an image or object given the trace representation. We now plug the λ_t from the previous section into algorithm 2. As discussed in section 5, we choose to represent an image region by one location on it and its trace relation to other points in the sense of the Hough transform. As the set of starting points $\mathbf{A}$ in algorithm 2 we simply choose all other locations in the image. The traces between two locations already incorporate the entire space between them, given the bounds on the representation length (in this case the maximum number of steps n_d taken). This amounts to a complete topological characterization of the points' neighborhood under a particular parametrization of the discretization parameters.

More formally, let b be the central point of the image patch or object, and y_k and a_i a set of locations sampled densely on the patch, as illustrated in figure 17. Then, b is represented by the subset $P \subseteq \Omega^t$ of all traces that have b in their feasible set for any location a on the patch with the locations y_k as intermediate nodes:

$$t \in P \Leftrightarrow \exists a : b \in \lambda_t(a) \tag{11}$$

The size of the subset P is the number of *reference traces* $n_{\text{ref}} = |P|$, which depends on the actual patch texture as well as the choice of discretization and sampling parameters. Given a center location b, the set P fully represents all information about the patch that can be expressed in the trace model with a given parametrization. The center location is contained in the feasible set of all traces P and via the construction of the (discrete) trace, this property is (quasi-) invariant to homeomorphisms. This implies that a perturbed version of the patch will also have its perturbed center location $b' = \mathcal{H}(b)$ in the feasible set of every trace in P given at least one starting location a'. Conversely, candidates for b' in a new image can be found by enumerating all feasible sets for all starting locations within the image (or a region of interest). Under the previous assumptions, the true b' then has to be contained in their intersection. Due to violations of the continuous assumptions and the particular discretization choices, only quasi-invariance is retained in practice. As a result there

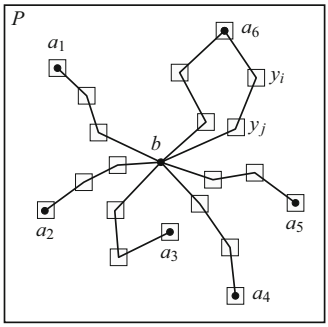


Fig. 17 An image patch is modeled as the set of all traces that end in its center location b, starting from any other location a_i on the patch. This example with $n_d = 3$ shows the intermediate locations y_i (two of them labeled), six selected start locations $a_{1...6}$ as well as seven traces. Each start node and intermediate location has a texture label assigned to it in a small neighborhood, illustrated by gray rectangles.

may be traces which do not have b' in their feasible set, implying that the correct location for b' is in the feasible set of less than n_{ref} of the reference traces. In order to address this, the following experimental section will use as candidates for b' locations that coincide with *as many feasible sets as possible*.

With this model for an image patch, we build a simple visual tracking algorithm based on template matching. It includes an optional incremental model updating mechanism that allows to model gradual changes of the tracked object through a video. The updating is performed by keeping a histogram of likely traces over time. Details of the tracking algorithm go beyond the scope of this chapter and are presented in depth in [7].

7 Results

We present results from two types of experiments:

1. Matching of image patches when the images undergo perturbations. This is a relevant application, for instance, in wide-baseline stereo or as a basic component in more complex algorithms.
2. Visual tracking of objects through videos.

For the first set of experiments we use the rotation invariant as well as the rotation sensitive version of the discretized texture trace. The rotation invariant texture trace was not included in the tracking results, as in-plane rotation is addressed by the incremental updating process and the rotation invariant texture trace has shown to perform slightly worse if there is no significant inter-frame rotation.

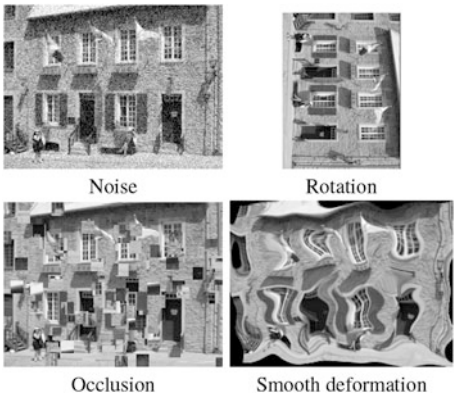


Fig. 18 Example synthetic perturbations. Occlusions are generated by randomly replacing image blocks by unrelated image blocks of various sizes, smooth deformations are generated by multi-scale Perlin noise of the domain with varying magnitudes.

7.1 Matching Image Patches

Figure 19a to 19d shows the likelihood that a patch can be detected under perturbations as the detection rate over a population of experiments. The perturbations were sampled over a range of different settings and synthetic data to improve statistical validity. Figure 18 shows example images. We compare to the Geodesic Intensity Histogram (GIH, [12]) as a representative for the state of the art in homeomorphic invariance in patch matching. The discretized texture trace representations outperform the GIH significantly under noise and occlusion. The reason for the discrepancy in the performance under noise may lie in the construction of the GIH. Both GIH and texture trace use the image function in a topological manner to define the patch representation. However, where the texture trace averages at each point over neighborhoods in the texture label quantization, which makes it more insensitive to noise, the GIH uses the image pixels directly to extract the geodesic contours. The latter may be very prone to image noise. In the case of occlusion, an explanation lies in the way two patches are compared. In the GIH, a patch is defined by the histogram of gray values at a set of geodesic distances from a common center, and two patches are compared via the χ^2 distance of their histograms. The χ^2 distance is not robust to occlusion of its dimensions, i.e. partial randomization of histogram entries. In contrast, the Hough algorithm has an independent voting-like structure, where occlusion is gracefully handled implicitly.

Under local perturbation, the texture trace outperforms the GIH for all but large perturbation magnitudes. At larger perturbations, it may suffer from the fixed neighborhood size for the texture quantization, where the labels cannot be assigned robustly any more as the perturbations within the neighborhood become too large. Except for the case of rotation, the rotation invariant texture trace performs slightly worse than the rotation sensitive trace. This is intuitively expected due to the trade off in the representation when additional invariance is added: one gains invariance towards one nuisance parameter, but potentially loses discriminative power for

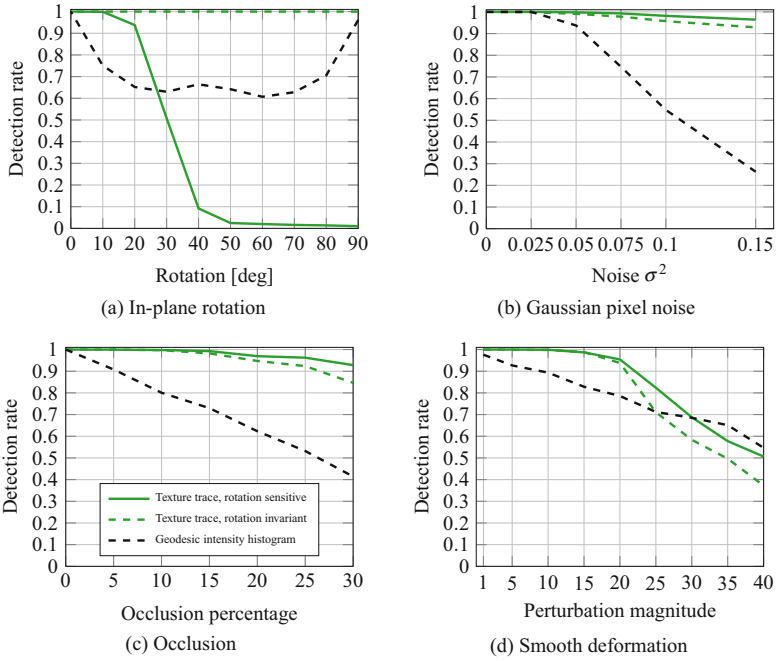


Fig. 19 Performance of rotation invariant and rotation sensitive discrete texture trace and the geodesic intensity histogram under various perturbations.

the others. The rotation invariant trace has near perfect performance in the case of rotation, demonstrating that it is indeed fully rotation invariant as designed (figure 19).

7.2 Tracking

Four video sequences from the literature were used for the visual tracking results with a wide array of perturbations, including motion blur, in- and out-of-plane rotation, occlusions, and illumination change. The protocol from [20] was used, adding to their comparison. The sequences "board," "box," "lemming," and "liquor" of [20] are evaluated by the PASCAL score [8] against the recent SPT [13] as well as PROST [20], MIL [2], FragTrack [1], ORF[19], and GRAD [10]. The GIH method was not included as it is unclear how to extend this method to incremental tracking. The PASCAL score measures the percentage of frames where the ground truth and detection overlap sufficiently to imply a correct detection. The results are shown in table 3, where *rsDTT* denotes the rotation sensitive discrete texture trace. The trace method has a consistently high score and is on par with the SPT with an

Table 3 PASCAL score for the four PROST sequences [20]. The best and second best method are highlighted in bold and underlined, respectively. The rotation sensitive texture trace is on par with the best compared method.

Method	Average	"board"	"box"	"lemming"	"liquor"
PROST	80.4	75.0	90.6	70.5	85.4
MIL	49.2	67.9	24.5	83.6	20.6
FragTrack	66.0	67.9	61.4	54.9	79.9
ORF	27.3	10.0	28.3	17.2	53.6
GRAD	88.9	94.3	91.8	78.0	91.4
SPT	<u>95.2</u>	<u>97.9</u>	**94.8**	<u>88.1</u>	**100**
rsDTT	**95.5**	**99.3**	<u>93.1</u>	**91.4**	<u>98.0</u>
rsDTT one-shot	86.6	96.4	77.3	81.3	91.4

overall PASCAL performance of 95.5%. It is important to realize that all of the high performing compared methods such as SPT use machine learning as an integral part of their representation.

7.3 One-Shot Tracking

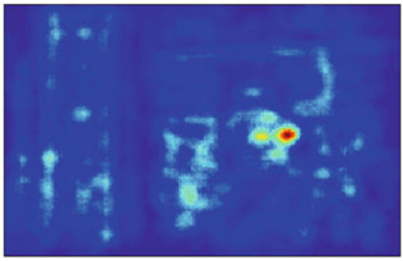
In order to get a better empirical understanding of the quasi-invariance properties of the texture trace representation, this section looks at the following question: how far can one get in tracking with *only using one frame for model building*, i.e. no continuous, incremental updating of the model? This *one-shot tracking* clearly stresses the invariance properties of any representation as only the first frame of a sequence is available during model building. The same tracking algorithm of the previous section is used, just without model updating after the first frame.

The resulting performance for the four sequences is shown in table 3 as *rsDTT one-shot*. When comparing the overall PASCAL performance of the one-shot method to the compared methods, one can see that it already outperforms four out of the six. In other words, with just using one initial frame and no elaborate machine learning apparatus, the texture trace-based tracker already takes third place out of seven, outperformed only by the GRAD and SPT methods. To illustrate the performance of the one-shot tracking, we applied it to the "dudek" sequence [9]. Figure 20b shows the initial image with the reference location marked in green, figure 20c the same image cropped to the given bounding box and several detections of the algorithm throughout the sequence. As an observation, the detected center point is always on the bridge of the nose between the eyes (as is the reference location in the first frame). Figure 20b shows one frame within the sequence and figure 20d the corresponding computed confidence map from the trace model. The overall PASCAL performance on this sequence with the one shot tracking is 99.5%, indicating that the detections are very precise.

(a) Annotated location in the first frame.

(b) Frame with detected maximum of confidence map.

(c) Close up of detections and reference image (top left)



(d) Confidence map (best viewed in color).

Fig. 20 Visualization of the detection result and the confidence map of the one-shot tracking for one frame of the "dudek" sequence [9]: (a) the reference location in the first frame, (b) one frame from the middle of the sequence, and (c) the rsDTT confidence map based only on the first frame.

8 Conclusion and Outlook

We described a generic scheme for constructing invariant and quasi-invariant signal representations based on topological connectedness and the preservation of neighborhood structure. The choice of defining it in terms of set relations allows the one-to-one transformation of certain algorithms based on metric relations into an invariant domain, effectively including the invariances without extra effort. Furthermore, we derived a particular instantiation, the trace model, and employed it in two applications. The underlying principle of the trace model is the relation of two signal locations by a description of the space between them based on one-dimensional paths, regardless of the signal's dimension. While this is only one possible derivation, it is motivated primarily by computational efficiency due to the factorization into a set of one-dimensional problems. The invariances of the specific instantiation of the model can be tuned to application-specific requirements. We demonstrate two versions of the trace model: one is rotation invariant, one is rotation sensitive. The computational backbone for both is matrix multiplication, for which there are efficient parallel implementations. Based on the discrete trace model we have shown results for two important problems in computer vision: patch-matching and visual tracking.

The main practical challenge of the trace implementation remains computational complexity. Empirical performance analysis not presented here shows that the representation greatly benefits from improving the discretization granularity, particularly the trace length. However, computation time increases significantly with the number of discretization steps. The model itself is highly parallelizable due to the independence of individual traces. On the other hand, many traces share redundant sub-paths which can be computed more efficiently as it is done here.

The trace model can readily be extended to higher-dimensional signals, such as videos or sequences of image volumes.Outside of image processing and computer vision, it may be used to substitute metrics in the sense of set relations in domains that can be attributed with a similar topological structure, where invariance to local deformations and robustness to occlusion of the domain is sought.

References

1. A. Adam, E. Rivlin, I. Shimshoni, Robust fragments-based tracking using the integral histogram, in *CVPR*, vol. 1 (2006), pp. 798–805
2. B. Babenko, M.-H. Yang, S. Belongie, Visual tracking with online multiple instance learning, in *CVPR* (2009), pp. 983–990
3. D.H. Ballard, Generalizing the hough transform to detect arbitrary shapes. Pattern Recognit. **13**(2), 111–122 (1981)
4. H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**, 346–359 (2008)
5. T.O. Binford, Visual perception by computer, in *Proceedings of the IEEE Conference on Systems and Control*, Miami (1971)
6. T. Binford, T. Levitt, Quasi-invariants: theory and exploitation, in *Proceedings of Defense Advanced Research Project Agency Image, Understanding Workshop* (1993), pp. 819–829
7. J. Ernst, The trace model for spatial invariance with applications in structured pattern recognition, image patch matching and incremental visual tracking. Dissertation. Shaker Verlag GmbH, Germany (2012)
8. M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL visual object classes challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
9. A.D. Jepson, D.J. Fleet, T.F. El-Maraghi, Robust online appearance models for visual tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(10), 1296–1311 (2003)
10. D.A. Klein, A.B. Cremers, Boosting scalable gradient features for adaptive real-time tracking, in *ICRA* (2011), pp. 4411–4416
11. B. Leibe, A. Leonardis, B. Schiele, Combined object categorization and segmentation with an implicit shape model, in *Workshop on Statistical Learning in Computer Vision (ECCV Workshop)* (2004)
12. H. Ling, D.W. Jacobs, Deformation invariant image matching. in *ICCV* (2005), pp. 1466–1473
13. B. Liu, J. Huang, L. Yang, C. Kulikowski, Robust tracking using local sparse appearance model and k-selection, in *CVPR* (2011), pp. 1313–1320
14. S. Lloyd, Least squares quantization in PCM. IEEE Trans. Inf. Theory **28**(2), 129–137 (1982)
15. D.G. Lowe, Object recognition from local scale-invariant features, in *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2, ICCV '99* (IEEE Computer Society, Washington, DC, 1999), p. 1150
16. D.G. Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)

17. K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1615–1630 (2005)
18. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Van Gool, A comparison of affine region detectors. Int. J. Comput. Vis. **65**(1–2), 43–72 (2005)
19. A. Saffari, C. Leistner, J. Santner, M. Godec, H. Bischof, On-line random forests, in *ICCV* (2009), pp. 1393–1400
20. J. Santner, C. Leistner, A. Saffari, T. Pock, H. Bischof, PROST parallel robust online simple tracking, in *CVPR* (2010)
21. Y. Tsin, V. Ramesh, T. Kanade, Statistical calibration of CCD imaging process, in *Proceedings. Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001*, vol. 1 (2001), pp. 480–487
22. T.K. Vintsyuk, Speech discrimination by dynamic programming. Cybernetics **4**, 52–57 (1968)
23. B. Xie, V. Ramesh, T. Boult, Sudden illumination change detection using order consistency. Image Vis. Comput. **22**(2), 117–125 (2004). Statistical Methods in Video Processing
24. M. Zerroug, R. Nevatia, Using invariance and quasi-invariance for the segmentation and recovery of curved objects, in *Proceedings of the Second Joint European - US Workshop on Applications of Invariance in Computer Vision* (Springer, London, 1994), pp. 317–340

Use of Quillen-Suslin Theorem for Laurent Polynomials in Wavelet Filter Bank Design

Youngmi Hur*

Abstract In this chapter we give an overview of a method recently developed for designing wavelet filter banks via the Quillen-Suslin Theorem for Laurent polynomials. In this method, the Quillen-Suslin Theorem is used to transform vectors with Laurent polynomial entries to other vectors with Laurent polynomial entries so that the matrix analysis tools that were not readily available for the vectors before the transformation can now be employed. As a result, a powerful and general method for designing non-redundant wavelet filter banks is obtained. In particular, the vanishing moments of the resulting wavelet filter banks can be controlled in a very simple way, which is especially advantageous compared to other existing methods for the multi-dimensional cases.

Keywords Laurent polynomials • Multi-dimensional wavelets • Non-redundant filter banks • Polyphase representation • Quillen-Suslin Theorem • Wavelet filter banks

1 Introduction

In this chapter we provide an overview of a recent method in [15] for designing non-redundant wavelet filter banks using the Quillen-Suslin Theorem for Laurent polynomials, which is a well-known result in Algebraic Geometry. The method works for any dimension but it would be the most useful for multi-dimensional cases, where the problem of designing wavelet filter banks can be quite challenging.

Wavelet representation [18], along with Fourier representation, has been one of the most commonly used data representations. Constructing 1-dimensional (1-D)

*This chapter is mainly based on the work with Hyungju Park and Fang Zheng presented in [15]. This research was partially supported by National Research Foundation of Korea (NRF) Grants 20151003262 and 20151009350.

Y. Hur (✉)
Department of Mathematics, Yonsei University, Seoul 03722, Korea
e-mail: yhur@yonsei.ac.kr

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4_12

wavelets is mostly well understood by now, but the situation is not the same for the multi-dimensional (multi-D) case. Taking the tensor product of 1-D functions is the most common approach, but the resulting separable wavelets have many unavoidable limitations. In order to overcome these limitations, various non-tensor-based approaches for constructing multi-D wavelets have been tried, but many of these methods show limitations in various aspects as well. For example, some work only for low spatial dimensions and cannot be easily extended to higher dimensions, whereas others assume that the lowpass filters or refinable functions satisfy additional conditions such as the interpolatory condition (see, for example, [10, 12–14] and the references therein). Therefore, the problem of constructing multi-D wavelets is still very challenging and calls for new ideas and insights.

Constructing wavelet filter banks is often reduced to solving an associated matrix problem with Laurent polynomial entries. Once the associated matrix problem is obtained, the wavelet filter bank design problem can be solved by using various techniques for the matrices with Laurent polynomial entries that have been developed in many different branches of mathematics. The method we look at in this chapter is based on a new way of applying the Quillen-Suslin Theorem for Laurent polynomials to the matrix problem, and it presents some advantages over the existing (both the tensor product and non-tensor-based) methods of multi-D wavelet construction: it works for any spatial dimension and for any dilation matrix, and it works without any additional assumptions, such as interpolatory condition, on the initial lowpass filters. Furthermore, it provides a simple algorithm for constructing wavelets with a prescribed number of vanishing moments.

2 Wavelet Filter Bank Design via Laurent Polynomial Matrices

Filters f are (real-valued) functions defined on the integer grids $\mathbb{Z}^n$. A filter bank (FB) consists of the analysis bank, which is a collection of, say p, filters used to analyze a given signal, and the synthesis bank, which is another (possibly different but with the same cardinality) collection of filters used to synthesize the analyzed coefficients or their modifications, depending on the application at hand, in order to get back to the original signal or its variant. We consider a special kind of FB, where one filter from each band is lowpass (i.e., $\sum_{k\in\mathbb{Z}^n} f(k) = \sqrt{q}$ where $q = |\det \Lambda|$ with dilation matrix Λ), and all the other filters are highpass (i.e., $\sum_{k\in\mathbb{Z}^n} f(k) = 0$), and we refer to such a FB as the *wavelet FB*. Only the FBs with finite impulse response filters and with the perfect reconstruction property will be considered, and in such a case we necessarily have $p \geq q$.

2.1 Polyphase Representation and Wavelet FB Design

The connection between the wavelet FB design problem and the Laurent polynomial matrix problem can be made via the polyphase decomposition [33]. Originally introduced for computationally efficient implementation of various filtering operations, the polyphase decomposition provides a way to transform filters and signals to vectors with Laurent polynomial entries, to which we refer as the *polyphase representation*. In particular, for an analysis filter h and a synthesis filter g, and for a dilation matrix Λ, the polyphase representation are given as the following Laurent polynomial vectors of length $q = |\det \Lambda| \geq 2$:

$$\mathtt{H}(z) := [H_{\nu_0}(z), \ldots, H_{\nu_{q-1}}(z)],$$
$$\mathtt{G}(z) := [G_{\nu_0}(z), \ldots, G_{\nu_{q-1}}(z)]^T,$$

respectively, where T is used for the transpose, $H_\nu(z)$ and $G_\nu(z)$ for the z-transform of the subfilters $h_\nu(k) := h(\Lambda k - \nu)$ and $g_\nu(k) := g(\Lambda k + \nu)$, respectively, and $\{\nu_0 := 0, \ldots, \nu_{q-1}\} =: \Gamma$ for a complete set of coset representatives of $\mathbb{Z}^n/\Lambda\mathbb{Z}^n$ containing 0.

In this setting, designing a FB is equivalent to finding a $p \times q$ analysis matrix $\mathtt{A}(z)$ and a $q \times p$ synthesis matrix $\mathtt{S}(z)$ with $\mathtt{S}(z)\mathtt{A}(z) = \mathtt{I}_q$. In this case, the FB is non-redundant if $p = q$, that is, if $\mathtt{A}(z)$ and $\mathtt{S}(z)$ are square. It is a wavelet FB if the first row of $\mathtt{A}(z)$ and the first column of $\mathtt{S}(z)$ are the polyphase representation of lowpass filters and all other rows of $\mathtt{A}(z)$ and all other columns of $\mathtt{S}(z)$ are the polyphase representation of highpass filters.

Understanding properties of a wavelet FB in terms of the polyphase representation is important. We recall that the filter f is lowpass (resp. highpass) if and only if $\sum_{\nu\in\Gamma} F_\nu(\mathtt{1}) = \sqrt{q}$ (resp. $\sum_{\nu\in\Gamma} F_\nu(\mathtt{1}) = 0$), where $\mathtt{1} \in \mathbb{R}^n$ is the vector of ones, and the lowpass filter f has positive accuracy if and only if $F_\nu(\mathtt{1}) = 1/\sqrt{q}$, for all $\nu \in \Gamma$ (cf. [10]). For a filter f, the number of zeros of $F(z)|_{z=e^{i\omega}}$ at $\omega \in \Gamma^*\backslash\{0\}$, where $F(z)$ is the z-transform of the filter f, is referred to as the *accuracy number* [28]. It is well known that the number of vanishing moments of each highpass filter in a non-redundant wavelet FB is at least the minimum of the accuracy numbers of the lowpass filters [5]. The number of vanishing moments is one of the important criteria in determining the approximation power of a wavelet system [19].

2.2 Quillen-Suslin Theorem and Wavelet FB Design

A row vector of length q with Laurent polynomial entries is called *unimodular* if it has a right inverse, which is a column vector of length q. A unimodular column vector is defined similarly.

Example 1 A row vector

$$\mathtt{H}(z) = [\frac{1}{2}, \frac{1}{4}z_1^{-1} + \frac{1}{4}, \frac{1}{4}z_2^{-1} + \frac{1}{4}, \frac{1}{4}z_1^{-1}z_2^{-1} + \frac{1}{4}] \tag{1}$$

is unimodular, because $[2, 0, 0, 0]^T$ is a right inverse of $\mathtt{H}(z)$. In fact, there are infinitely many right inverses of $\mathtt{H}(z)$, and one of them is the column vector

$$[-\frac{1}{8}z_1^{-1} - \frac{1}{8}z_2^{-1} - \frac{1}{8}z_1^{-1}z_2^{-1} + \frac{5}{4} - \frac{1}{8}z_1 - \frac{1}{8}z_2 - \frac{1}{8}z_1z_2, \frac{1}{4} + \frac{1}{4}z_1, \frac{1}{4} + \frac{1}{4}z_2, \frac{1}{4} + \frac{1}{4}z_1z_2]^T.$$

Clearly the former is simpler, but the latter may be preferred for a wavelet FB design because the lowpass filter associated with it has larger accuracy number: it is 2, whereas the one for the former is 0. □

More generally, a matrix with Laurent polynomial entries is called a *unimodular matrix* if its maximal minors generate 1. The Quillen-Suslin Theorem (also referred to as the unimodular completion), originally conjectured by J. P. Serre [26] and proved after about 20 years [24, 29], is a well-known result in Algebraic Geometry, and it asserts that any unimodular matrix over a polynomial ring can be completed to an invertible square matrix. This result, together with its generalization to Laurent polynomial ring [30] and their constructive and algorithmic proofs [1, 17, 23], has been used in various other disciplines including Signal Processing as well [3, 16]. The following special case of the unimodular completion over Laurent polynomial rings is used for the wavelet FB design method we look at in this chapter.

Theorem 1 (Quillen-Suslin Theorem for Laurent polynomials [30]) *Let* $\mathtt{D}(z)$ *be a unimodular column vector of length q with Laurent polynomial entries. Then there exists an invertible* $q \times q$ *matrix* $\mathtt{M}(z)$ *with Laurent polynomial entries such that* $\mathtt{M}(z)\mathtt{D}(z) = [1, 0, \ldots, 0]^T$.

Although the above result can be useful in designing non-redundant wavelet FBs (cf. [5]), there are still some important questions remained to be answered. For example, obtaining a pair of lowpass filters with a prescribed number of accuracy is a key step in such an approach, but this may not be straightforward to do so, especially in multi-D cases, as we illustrate below for the 2-dimensional case.

Example 2 When $n = 2$, the lowpass filter associated with the linear box spline has accuracy 2, and its polyphase representation is given as $\mathtt{H}(z)$ in (1) and thus, as we saw in Example 1, it has a right inverse $[2, 0, 0, 0]^T$. But the lowpass filter associated with $[2, 0, 0, 0]^T$ has 0 accuracy and, as a result, it cannot be a lowpass filter for a wavelet FB. Gröbner bases techniques ([6, 20, 22]) can be used to give the most general form of the right inverse for $\mathtt{H}(z)$:

$$\begin{bmatrix} 2 \\ 0 \\ 0 \\ 0 \end{bmatrix} - \frac{1}{2}u_1(z)\begin{bmatrix} z_1^{-1} + 1 \\ -2 \\ 0 \\ 0 \end{bmatrix} - \frac{1}{2}u_2(z)\begin{bmatrix} z_2^{-1} + 1 \\ 0 \\ -2 \\ 0 \end{bmatrix} - \frac{1}{2}u_3(z)\begin{bmatrix} z_1^{-1}z_2^{-1} + 1 \\ 0 \\ 0 \\ -2 \end{bmatrix}$$

(implemented via the Maple package *QuillenSuslin* by Anna Fabiańska), where $u_1(z), u_2(z), u_3(z)$ are any Laurent polynomials that are used as parameters. To find a right inverse of $\mathtt{H}(z)$ with positive accuracy, one can choose specific Laurent polynomials for parameters $u_1(z), u_2(z), u_3(z)$, which is usually done by fixing the total degree of Laurent polynomials and then increasing the total degree if needed [8, 21, 25]. However, this approach may not be the best strategy, especially if one looks for a right inverse for which the associated lowpass filter is supported in a non-rectangular region. □

3 New Quillen-Suslin based Method for Designing Wavelet FBs

In this section we discuss the main ingredients of the theory and algorithms in the new Quillen-Suslin Theorem based method for designing wavelet FBs presented in [15], and start our discussion by pointing out some motivation for the theory.

3.1 Motivation for the theory

For any lowpass filters h and g used for analysis and synthesis, respectively, their polyphase representation $\mathtt{H}(z)$ and $\mathtt{G}(z)$ satisfy the following simple matrix identity:

$$\begin{bmatrix} \mathtt{G}(z) & \mathtt{I}_q \end{bmatrix} \begin{bmatrix} \mathtt{H}(z) \\ \mathtt{I}_q - \mathtt{G}(z)\mathtt{H}(z) \end{bmatrix} = \mathtt{I}_q.$$

In fact, the above identity can be understood as a matrix-based interpretation of Laplacian pyramid (LP) algorithms [4], which is widely used in Signal Processing [9, 31, 32]. However, this matrix identity alone does not give a wavelet FB, because the filters associated with the column vectors of the matrix $\mathtt{I}_q$ in the synthesis matrix $\begin{bmatrix} \mathtt{G}(z) & \mathtt{I}_q \end{bmatrix}$ are not highpass, even if the lowpass filters h and g are chosen to have positive accuracy. If the lowpass filters have positive accuracy and they are biorthogonal, i.e. $\mathtt{H}(z)\mathtt{G}(z) = 1$, then another synthesis matrix $\begin{bmatrix} \mathtt{G}(z) & \mathtt{I}_q - \mathtt{G}(z)\mathtt{H}(z) \end{bmatrix}$ is available, and its use leads to the construction of wavelet FBs, as studied in [7, 11]. Actually, the most general LP synthesis matrix is known and it is

$$\mathtt{S}_{LP}(z) := \begin{bmatrix} \mathtt{G}(z) + \mathtt{F}(z)(1 - \mathtt{H}(z)\mathtt{G}(z)) & \mathtt{I}_q - \mathtt{F}(z)\mathtt{H}(z) \end{bmatrix},$$

where $\mathtt{F}(z)$ is any column vector of length q [2].

Another approach to design wavelet FBs based on LP algorithms is studied in [10] for the case including when the lowpass filter h satisfies the interpolatory condition. A lowpass filter is *interpolatory* if the first component of its polyphase

representation is constant, and in such a case the constant is necessarily $1/\sqrt{q}$, where $q = |\det \Lambda|$ for the dilation matrix Λ (cf. [10]). Suppose that h is interpolatory with positive accuracy. Since in this case, for any column vector $\mathtt{G}(z)$ of length q, the second row of the analysis matrix $\mathtt{A}_{LP}(z) := \begin{bmatrix} \mathtt{H}(z) \\ \mathtt{I}_q - \mathtt{G}(z)\mathtt{H}(z) \end{bmatrix}$ can be written in terms of the rest rows of the matrix, we have the following identity

$$\begin{bmatrix} 1 & 0 \\ \sqrt{q}(1-\mathtt{H}(z)\mathtt{G}(z)) & -\sqrt{q}\,\widetilde{\mathtt{H}}(z) \\ 0 & \mathtt{I}_{q-1} \end{bmatrix} \begin{bmatrix} 1\,0 & 0 \\ 0\,0 & \mathtt{I}_{q-1} \end{bmatrix} \mathtt{A}_{LP}(z) = \mathtt{A}_{LP}(z),$$

which in turn gives

$$\begin{aligned} \mathtt{I}_q &= \mathtt{S}_{LP}(z)\mathtt{A}_{LP}(z) \\ &= \left(\mathtt{S}_{LP}(z) \begin{bmatrix} 1 & 0 \\ \sqrt{q}(1-\mathtt{H}(z)\mathtt{G}(z)) & -\sqrt{q}\,\widetilde{\mathtt{H}}(z) \\ 0 & \mathtt{I}_{q-1} \end{bmatrix} \right) \left(\begin{bmatrix} 1\,0 & 0 \\ 0\,0 & \mathtt{I}_{q-1} \end{bmatrix} \mathtt{A}_{LP}(z) \right) \\ &= \begin{bmatrix} G_{v_0}(z) + \sqrt{q}\,(1-\mathtt{H}(z)\mathtt{G}(z)) & -\sqrt{q}\,\widetilde{\mathtt{H}}(z) \\ \widetilde{\mathtt{G}}(z) & \mathtt{I}_{q-1} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{q}} & \widetilde{\mathtt{H}}(z) \\ -\frac{1}{\sqrt{q}}\,\widetilde{\mathtt{G}}(z) & \mathtt{I}_{q-1} - \widetilde{\mathtt{G}}(z)\widetilde{\mathtt{H}}(z) \end{bmatrix} \\ &=: \mathtt{S}_{ECLP}(z)\mathtt{A}_{ECLP}(z) \end{aligned}$$

where $\widetilde{\mathtt{H}}(z)$ (resp. $\widetilde{\mathtt{G}}(z)$) is a subvector of $\mathtt{H}(z)$ (resp. $\mathtt{G}(z)$) obtained by removing the first entry. Therefore, as long as the lowpass filter g associated with $\mathtt{G}(z)$ has positive accuracy, we obtain a non-redundant wavelet FB whose analysis matrix is $\mathtt{A}_{ECLP}(z)$ and the synthesis matrix is $\mathtt{S}_{ECLP}(z)$ (cf. [10] for more details). In particular, the first column of $\mathtt{S}_{ECLP}(z)$, which is $\mathtt{G}(z) + [\sqrt{q}, 0, \ldots, 0]^T(1 - \mathtt{H}(z)\mathtt{G}(z))$, is the polyphase representation of the synthesis lowpass filter.

3.2 Main ingredients of the theory

In the approach outlined above, the fact that the vector $\mathtt{H}(z)$ for the interpolatory filter has a *unit*[1] in the Laurent polynomial ring as one of its entry is used essentially, and it is clear that this property does not hold true for the general lowpass filter.

Let $\mathtt{H}(z)$ be any polyphase representation for an analysis lowpass filter h with positive accuracy (that is not necessarily interpolatory). Suppose that we want to design a non-redundant wavelet FB for which its analysis lowpass filter is h. Then $\mathtt{H}(z)$ is necessarily unimodular, because, being square matrices, the analysis matrix

[1] An element in a ring is called a unit if its multiplicative inverse lies in the ring.

times the synthesis matrix equals to $\mathtt{I}_q$ as well, hence reading off $(1,1)$-entry of both sides in the identity guarantees the existence of a right inverse of $\mathtt{H}(z)$. Therefore we assume that the polyphase representation $\mathtt{H}(z)$ we start with is unimodular. The unimodularity of $\mathtt{H}(z)$ for the interpolatory h is trivial since $[\sqrt{q}, 0, \cdots, 0]^T$ is a right inverse of $\mathtt{H}(z)$.

From the unimodularity of $\mathtt{H}(z)$, we see that there exists a column vector $\mathtt{F}(z)$ of length q with Laurent polynomial entries such that $\mathtt{H}(z)\mathtt{F}(z) = 1$. Hence $\mathtt{F}(z)$ is unimodular as well. By Theorem 1, there exists an invertible $q \times q$ matrix $\mathtt{M}(z)$ such that $\mathtt{M}(z)\mathtt{F}(z) = [1, 0, \ldots, 0]^T$. Then $[\mathtt{M}(z)]^{-1}$ is a $q \times q$ matrix with Laurent polynomial entries, and $\mathtt{H}(z)[\mathtt{M}(z)]^{-1}$ is a left inverse of $\mathtt{M}(z)\mathtt{F}(z) = [1, 0, \ldots, 0]^T$, hence its first entry is 1, which is a unit. By letting the transformed row vector $\mathtt{H}^{\mathtt{M}}(z) := \mathtt{H}(z)[\mathtt{M}(z)]^{-1}$ play the role of $\mathtt{H}(z)$ in the interpolatory case as described in Section 3.1, for any column vector $\mathtt{G}(z)$ of length q, we get the following matrix identity:

$$\mathtt{I}_q = \begin{bmatrix} \mathtt{G}^{\mathtt{M}}(z) + \mathtt{F}^{\mathtt{M}}(z)(1 - \mathtt{H}^{\mathtt{M}}(z)\mathtt{G}^{\mathtt{M}}(z)) & \mathtt{I}_q - \mathtt{F}^{\mathtt{M}}(z)\mathtt{H}^{\mathtt{M}}(z) \end{bmatrix} \begin{bmatrix} \mathtt{H}^{\mathtt{M}}(z) \\ \mathtt{I}_q - \mathtt{G}^{\mathtt{M}}(z)\mathtt{H}^{\mathtt{M}}(z) \end{bmatrix}, \tag{2}$$

where $\mathtt{F}^{\mathtt{M}}(z) := \mathtt{M}(z)\mathtt{F}(z)$ and $\mathtt{G}^{\mathtt{M}}(z) := \mathtt{M}(z)\mathtt{G}(z)$. The transformed polyphase representation used here can be thought of a generalization of the valid polyphase representation studied in [27].

Following the previous discussions when $\mathtt{H}(z)$ is interpolatory, because the second row of the transformed analysis matrix (the second matrix in the right-hand side of (2)) can be written in terms of the rest rows of the matrix, by inserting

$$\begin{bmatrix} 1 & 0 \\ (1 - \mathtt{H}^{\mathtt{M}}(z)\mathtt{G}^{\mathtt{M}}(z)) & -\widetilde{\mathtt{H}^{\mathtt{M}}}(z) \\ 0 & \mathtt{I}_{q-1} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & \mathtt{I}_{q-1} \end{bmatrix}$$

between the two matrices in the right-hand side of (2), we obtain the matrix identity

$$\begin{aligned} \mathtt{I}_q &= \begin{bmatrix} G^{\mathtt{M}}_{\nu_0}(z) + (1 - \mathtt{H}^{\mathtt{M}}(z)\mathtt{G}^{\mathtt{M}}(z)) & -\widetilde{\mathtt{H}^{\mathtt{M}}}(z) \\ \widetilde{\mathtt{G}^{\mathtt{M}}}(z) & \mathtt{I}_{q-1} \end{bmatrix} \begin{bmatrix} 1 & \widetilde{\mathtt{H}^{\mathtt{M}}}(z) \\ -\widetilde{\mathtt{G}^{\mathtt{M}}}(z) & \mathtt{I}_{q-1} - \widetilde{\mathtt{G}^{\mathtt{M}}}(z)\widetilde{\mathtt{H}^{\mathtt{M}}}(z) \end{bmatrix} \\ &=: \mathtt{S}^{\mathtt{M}}_{ECLP}(z)\mathtt{A}^{\mathtt{M}}_{ECLP}(z), \end{aligned}$$

hence we get a non-redundant wavelet FB with the analysis matrix $\mathtt{A}^{\mathtt{M}}_{ECLP}(z)\mathtt{M}(z)$ and the synthesis matrix $[\mathtt{M}(z)]^{-1}\mathtt{S}^{\mathtt{M}}_{ECLP}(z)$, provided that the lowpass filter g associated with $\mathtt{G}(z)$ has positive accuracy. More precisely, in this wavelet FB, the polyphase representation for the synthesis lowpass filter is

$$[\mathtt{M}(z)]^{-1}\left(\mathtt{G}^{\mathtt{M}}(z) + [1, 0, \ldots, 0]^T(1 - \mathtt{H}^{\mathtt{M}}(z)\mathtt{G}^{\mathtt{M}}(z))\right) = \mathtt{G}(z) + \mathtt{F}(z)(1 - \mathtt{H}(z)\mathtt{G}(z)),$$

for the synthesis highpass filters are the 2nd through the last column vectors of $[\mathtt{I}_q - \mathtt{F}(z)\mathtt{H}(z)][\mathtt{M}(z)]^{-1}$, and for the analysis highpass filters are the 2nd through the last row vectors of $\mathtt{M}(z)[\mathtt{I}_q - \mathtt{G}(z)\mathtt{H}(z)]$.

Remark 1 *Although the case when* $\mathtt{M}(z)$ *satisfies* $\mathtt{M}(z)\mathtt{F}(z) = [1, 0, \ldots, 0]^T$ *is discussed above, all we need to run the above argument is for* $\mathtt{M}(z)\mathtt{F}(z)$ *to be a unimodular column vector with a unit in at least one of its components.*

Remark 2 *Unlike the classical approach in searching for a right inverse of* $\mathtt{H}(z)$ *for a non-redundant wavelet FB design (cf. Example 2), in the above approach, we do not need to look for a single right inverse of* $\mathtt{H}(z)$ *that has positive accuracy. Rather, one needs a pair of column vectors* $\mathtt{F}(z)$ *and* $\mathtt{G}(z)$ *such that* $\mathtt{F}(z)$ *is any right inverse of* $\mathtt{H}(z)$ *(with possibly no accuracy) and that* $\mathtt{G}(z)$ *has positive accuracy (but needs not be a right inverse of* $\mathtt{H}(z)$*), which is much easier to find.*

3.3 Main ingredients of the algorithms

The theory in the previous subsection provides an immediate algorithm for designing non-redundant wavelet FBs.

Algorithm 1 For a non-redundant wavelet FB from a lowpass filter.

Input: $\mathtt{H}(z)$: *unimodular polyphase representation of an analysis lowpass filter* h *with positive accuracy.*

Output: $\mathtt{D}(z)$: *polyphase representation of a synthesis lowpass filter,*
$\mathtt{J}_1(z), \ldots, \mathtt{J}_{q-1}(z)$: *polyphase representation of analysis highpass filters,*
$\mathtt{K}_1(z), \ldots, \mathtt{K}_{q-1}(z)$: *polyphase representation of synthesis highpass filters, such that, together with* $\mathtt{H}(z)$*, they form a non-redundant wavelet FB.*

Step 1: *Choose a lowpass filter* g *with positive accuracy, and let* $\mathtt{G}(z)$ *(as a column vector) be its polyphase representation.*

Step 2: *Choose a right inverse* $\mathtt{F}(z)$ *of* $\mathtt{H}(z)$.

Step 3: Set $\mathtt{D}(z) := \mathtt{G}(z) + \mathtt{F}(z)(1 - \mathtt{H}(z)\mathtt{G}(z))$.

Step 4: *Choose an invertible* $q \times q$ matrix $\mathtt{M}(z)$ *such that* $\mathtt{M}(z)\mathtt{F}(z) = [1, 0, \cdots, 0]^T$.

Step 5: *Set* $\mathtt{J}_1(z), \ldots, \mathtt{J}_{q-1}(z) :=$ *2nd through last rows of* $\mathtt{M}(z)[\mathtt{I}_q - \mathtt{G}(z)\mathtt{H}(z)]$.

Step 6: *Set* $\mathtt{K}_1(z), \ldots, \mathtt{K}_{q-1}(z) :=$ *2nd through last columns of* $[\mathtt{I}_q - \mathtt{F}(z)\mathtt{H}(z)][\mathtt{M}(z)]^{-1}$.

Given an analysis lowpass filter h, if one is interested in getting a synthesis lowpass filter d with positive accuracy, one can stop the algorithm after **Step 3** and use $\mathtt{D}(z)$ there as its polyphase representation. In fact, it can be shown that the accuracy of the lowpass filter d is at least $\min\{\alpha_h, \alpha_g, \alpha_f + \beta_h, \alpha_f + \beta_g\}$, where f is the lowpass filter having $\mathtt{F}(z)$ as its polyphase representation, and α_x and β_x are the accuracy number and the flatness number of a lowpass filter x, respectively (see [15] for details including the definition of the flatness number of a lowpass filter).

In general α_f can be zero, and β_h, β_g can be as small as 1 (they have to be positive because h and g are lowpass filters) even if h and g have large accuracy, hence as a result, the accuracy of d can be much smaller than α_h. This situation can be improved by choosing g with large accuracy, and iterating a part of **Algorithm 1** as shown in the next algorithm. Recalling the close relation between the number

of vanishing moments and the accuracy numbers of the lowpass filters for a non-redundant wavelet FB (cf. Section 2.1), the next algorithm provides a way to design wavelet FBs with large vanishing moments from a lowpass filter with large accuracy.

Algorithm 2 For a non-redundant wavelet FB with $\geq \alpha_h$ vanishing moments.

Input: $\mathtt{H}(z)$: *unimodular polyphase representation of an analysis lowpass filter h with accuracy α_h.*

Output: $\mathtt{D}(z)$: *polyphase representation of a synthesis lowpass filter,*
$\mathtt{J}_1(z), \ldots, \mathtt{J}_{q-1}(z)$: *polyphase representation of analysis highpass filters,*
$\mathtt{K}_1(z), \ldots, \mathtt{K}_{q-1}(z)$: *polyphase representation of synthesis highpass filters,*
such that, together with $\mathtt{H}(z)$*, they form a non-redundant wavelet FB with highpass filters having at least α_h vanishing moments.*

Step 1: *Set* $Ite := 1$.
Step 2: *Choose a lowpass filter g with at least α_h accuracy, and let* $\mathtt{G}(z)$ *(as a column vector) be its polyphase representation.*
Step 3: *Choose a right inverse* $\mathtt{F}(z)$ *of* $\mathtt{H}(z)$.
Step 4: *Set* $\mathtt{D}(z) := \mathtt{G}(z) + \mathtt{F}(z)(1 - \mathtt{H}(z)\mathtt{G}(z))$.
Step 5: *If* $\alpha_f + (Ite)\beta_h \geq \alpha_h$ *and* $\alpha_f + (Ite)\beta_g \geq \alpha_h$, *then go to* **Step 6**. *Otherwise, let* $Ite := Ite + 1$ *and* $\mathtt{F}(z) := \mathtt{D}(z)$, *and go to* **Step 4**.
Step 6: *Choose an invertible $q \times q$ matrix* $\mathtt{M}(z)$ *such that* $\mathtt{M}(z)\mathtt{F}(z) = [1, 0, \cdots, 0]^T$.
Step 7: *Set* $\mathtt{J}_1(z), \ldots, \mathtt{J}_{q-1}(z) :=$ *2nd through last rows of* $\mathtt{M}(z)[\mathtt{I}_q - \mathtt{G}(z)\mathtt{H}(z)]$.
Step 8: *Set* $\mathtt{K}_1(z), \ldots, \mathtt{K}_{q-1}(z) :=$ *2nd through last columns of* $[\mathtt{I}_q - \mathtt{F}(z)\mathtt{H}(z)][\mathtt{M}(z)]^{-1}$.

Because β_h and β_g are positive, each time the algorithm goes back to **Step 4** from **Step 5**, $\alpha_f + (Ite)\beta_h$ and $\alpha_f + (Ite)\beta_g$ strictly increase and they eventually satisfy the conditions $\alpha_f + (Ite)\beta_h \geq \alpha_h$ and $\alpha_f + (Ite)\beta_g \geq \alpha_h$, even if they did not initially. Therefore, by the time the algorithm reaches to **Step 6**, $\min\{\alpha_h, \alpha_g, \alpha_f + \beta_h, \alpha_f + \beta_g\} = \alpha_h$, and the accuracy number of the lowpass filter associated with $\mathtt{D}(z)$ is at least α_h.

In both algorithms, $\mathtt{G}(z)$, $\mathtt{F}(z)$, and $\mathtt{M}(z)$ need to be chosen. One can always choose $\mathtt{H}(z^{-1})^T$ as $\mathtt{G}(z)$. $\mathtt{F}(z)$ is nothing but the first column of $[\mathtt{M}(z)]^{-1}$, and $\mathtt{F}(z)$ and $\mathtt{M}(z)$ can be found by using Mathematical softwares such as Maple package *QuillenSuslin* mentioned earlier.

4 Conclusion

We presented some important ingredients of a recent method in [15] for designing non-redundant wavelet FBs, as well as some essential background material for the method including the Quillen-Suslin Theorem. The main advantage of this method compared to other existing wavelet FB design methods is the existence of a simple algorithm for designing a non-redundant wavelet FB with a prescribed number of vanishing moments.

References

1. M. Amidou, I. Yengui, An algorithm for unimodular completion over Laurent polynomial rings. Linear Algebra Appl. **429**(7), 1687–1698 (2008)
2. H. Bölcskei, F. Hlawatsch, H.G. Feichtinger, Frame-theoretic analysis of oversampled filter banks. IEEE Trans. Signal Process. **46**(12), 3256–3268 (1998)
3. B. Buchberger, Gröbner bases and systems theory. Multidim. Syst. Signal Process. **12**, 223–251 (2001)
4. P.J. Burt, E.H. Adelson, The Laplacian pyramid as a compact image code. IEEE Trans. Commun. **31**(4), 532–540 (1983)
5. D.-R. Chen, B. Han, S.D. Riemenschneider, Construction of multivariate biorthogonal wavelets with arbitrary vanishing moments. Adv. Comput. Math. **13**(2), 131–165 (2000)
6. D. Cox, J. Little, D. O'Shea, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra* (Springer, New York, 2006)
7. M.N. Do, M. Vetterli, Framing pyramids. IEEE Trans. Signal Process. **51**(9), 2329–2342 (2003)
8. B. Han, R.-Q. Jia, Optimal interpolatory subdivision schemes in multidimensional spaces. SIAM J. Numer. Anal. **36**, 105–124 (1998)
9. D.J. Heeger, J.R. Bergen, Pyramid-based texture analysis/synthesis, in *Proceedings of ACM SIGGRAPH* (1995), pp. 229–238
10. Y. Hur, Effortless critical representation of Laplacian pyramid. IEEE Trans. Signal Process. **58**, 5584–5596 (2010)
11. Y. Hur, A. Ron, CAPlets: wavelet representations without wavelets, preprint (2005). Available online: ftp://ftp.cs.wisc.edu/Approx/huron.ps
12. Y. Hur, A. Ron, L-CAMP: extremely local high-performance wavelet representations in high spatial dimension. IEEE Trans. Inf. Theory **54**, 2196–2209 (2008)
13. Y. Hur, A. Ron, High-performance very local Riesz wavelet bases of $L_2(\mathbf{R}^n)$. SIAM J. Math. Anal. **44**, 2237–2265 (2012)
14. Y. Hur, F. Zheng, Coset Sum: an alternative to the tensor product in wavelet construction. IEEE Trans. Inf. Theory **59**, 3554–3571 (2013)
15. Y. Hur, H. Park, F. Zheng, Multi-D wavelet filter bank design using Quillen-Suslin theorem for Laurent polynomials. IEEE Trans. Signal Process. **62**, 5348–5358 (2014)
16. Z. Lin, L. Xu, Q. Wu, Applications of Gröbner bases to signal and image processing: a survey. Linear Algebra Appl. **391**, 169–202 (2004)
17. A. Logar, B. Sturmfels, Algorithms for the Quillen-Suslin theorem. J. Algebra **145**, 231–239 (1992)
18. S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**(7), 674–693 (1989)
19. Y. Meyer, *Wavelets and Operators* (Cambridge University Press, Cambridge, 1992)
20. H. Park, A computational theory of Laurent polynomial rings and multidimensional FIR systems. Ph.D. dissertation, University of California, Berkeley (1995)
21. H. Park, Optimal design of synthesis filters in multidimensional perfect reconstruction FIR filter banks using Gröbner bases. IEEE Trans. Circuits Syst. **49**, 843–851 (2002)
22. H. Park, Symbolic computation and signal processing. J. Symb. Comput. **37**(2), 209–226 (2004)
23. H. Park, C. Woodburn, An algorithmic proof of Suslin's stability theorem for polynomial rings. J. Algebra **178**(1), 277–298 (1995)
24. D. Quillen, Projective modules over polynomial rings. Invent. Math. **36**(1), 167–171 (1976)
25. S.D. Riemenschneider, Z. Shen, Multidimensional interpolatory subdivision schemes. SIAM J. Numer. Anal. **34**, 2357–2381 (1997)
26. J.-P. Serre, Faisceaux algébriques cohérents. Ann. Math. **61**, 191–274 (1955)
27. A.K. Soman, P.P. Vaidyanathan, Generalized polyphase representation and application to coding gain enhancement. IEEE Trans. Circuits Syst. II Analog Digit. Signal Process. **41**(9), 627–630 (1994)

28. G. Strang, T. Nguyen, *Wavelets and Filter Banks* (Wellesley-Cambridge Press, Wellesley, 1997)
29. A.A. Suslin, Projective modules over a polynomial ring are free. Sov. Math. Dokl. **17**(4), 1160–1164 (1976)
30. R.G. Swan, Projective modules over Laurent polynomial rings. Trans. Am. Math. Soc **237**, 111–120 (1978)
31. S. Toelg, T. Poggio, Towards an example-based image compression architecture for video-conferencing, A.I. Memo No. 1494, MIT (1994)
32. M. Unser, An improved least squares Laplacian pyramid for image compression. Signal Process. **27**, 187–203 (1992)
33. P.P. Vaidyanathan, *Multirate Systems and Filter Banks* (Prentice-Hall, Englewood Cliffs, NJ, 1993)

A Fast Fourier Transform for Fractal Approximations

Calvin Hotchkiss and Eric S. Weber

Abstract We consider finite approximations of a fractal generated by an iterated function system of affine transformations on $\mathbb{R}^d$ as a discrete set of data points. Considering a signal supported on this finite approximation, we propose a Fast (Fractal) Fourier Transform by choosing appropriately a second iterated function system to generate a set of frequencies for a collection of exponential functions supported on this finite approximation. Since both the data points of the fractal approximation and the frequencies of the exponential functions are generated by iterated function systems, the matrix representing the Discrete Fourier Transform (DFT) satisfies certain recursion relations, which we describe in terms of Diţă's construction for large Hadamard matrices. These recursion relations allow for the DFT matrix calculation to be reduced in complexity to $O(N \log N)$, as in the case of the classical FFT.

Keywords Fractal • Fast Fourier Transform • Hadamard matrix

2000 *Mathematics Subject Classification*. Primary: 05B20, 65T50; Secondary: 28A80

1 Introduction

The Fast Fourier Transform (FFT) is celebrated as a significant mathematical achievement (see, for example, [1]). The FFT utilizes symmetries in the matrix representation of the Discrete Fourier Transform (DFT) [3]. For 2^N (equispaced) data points on [0,1), the matrix representation of the DFT is given by

$$\mathcal{F}_N = (e^{-2\pi i \frac{jk}{2^N}})_{jk}$$

C. Hotchkiss (✉) • E.S. Weber
Department of Mathematics, Iowa State University, 396 Carver Hall, Ames, IA 50011, USA
e-mail: hotchkis@iastate.edu; esweber@iastate.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4_13

where $0 \leq j, k < 2^N$. The FFT is obtained from the DFT by a permutation of the columns of $\mathcal{F}_N$:

$$\mathcal{F}_N = (e^{-2\pi i \frac{j\sigma(k)}{2^N}})_{jk}$$

for $0 \leq j, k < 2^N$, where

$$\sigma(k) = \begin{cases} 2k & 0 \leq k < 2^{N-1}, \\ 2k+1 & 2^{N-1} \leq k < 2^N. \end{cases}$$

The significance of the permutation is that the permuted matrix can be written in the following block form:

$$\mathcal{F}_N P = \begin{pmatrix} \mathcal{F}_{N-1} & D\mathcal{F}_{N-1} \\ \mathcal{F}_{N-1} & -D\mathcal{F}_{N-1} \end{pmatrix} \tag{1}$$

where D is a diagonal matrix. This block form reduces the computational complexity of the associated matrix multiplication; recursively, $\mathcal{F}_{N-1}$ can be permuted and written in block form as well. Repeated application of the column permutation reduces the computational complexity further, and results in overall complexity $O(N \cdot 2^N)$.

We take the view in the present paper that the DFT arises naturally in the context of iterated function systems, and the FFT arises as reordering of the iterated function system. Indeed, consider the following set of generators:

$$\tau_0(x) = \frac{x}{2}; \qquad \tau_1(x) = \frac{x+1}{2}.$$

The invariant set of this IFS is the interval $[0, 1]$, and the invariant measure is Lebesgue measure restricted to $[0, 1]$. Consider the approximation for the invariant set [10, 12] given by

$$\mathcal{S}_N := \{\tau_{j_{N-1}} \circ \tau_{j_{N-2}} \circ \cdots \circ \tau_{j_1} \circ \tau_{j_0}(0) : j_k \in \{0, 1\}\}.$$

This is an approximation in the sense that $[0, 1] = \overline{\cup_N \mathcal{S}_N}$, but the significance for our purposes is that $\mathcal{S}_N$ consists of 2^N equispaced-points:

$$\mathcal{S}_N = \{\frac{k}{2^N} : k \in \mathbb{Z},\ 0 \leq k < 2^N\}.$$

Define a second iterated function system generated by

$$\rho_0(x) = 2x; \qquad \rho_1(x) = 2x + 1.$$

Since these are not contractions, the IFS will not have a compact invariant set, but we consider the finite orbits of 0 under this IFS just as before. Define

$$\mathcal{T}_N := \{\rho_{j_N-1} \circ \rho_{j_N-2} \circ \cdots \circ \rho_{j_1} \circ \rho_{j_0}(0) : j_k \in \{0, 1\}\}.$$

Note that

$$\mathcal{T}_N = \{k : k \in \mathbb{Z},\ 0 \leq k < 2^N\}.$$

With the inherited ordering on $\mathcal{S}_N$ and $\mathcal{T}_N$ from $\mathbb{R}$, say $\mathcal{S}_N = \{s_0, s_1, \ldots, s_{2^N-1}\}$ and $\mathcal{T}_N = \{t_0, t_1, \ldots, t_{2^N-1}\}$, we obtain

$$\mathcal{F}_N = (e^{-2\pi i t_j s_k})_{jk}.$$

For $0 \leq k < 2^N$, we write $k = \sum_{n=0}^{N-1} j_n 2^n$ with $j_n \in \{0, 1\}$. Then

$$\tau_{j_N-1} \circ \tau_{j_N-2} \circ \cdots \circ \tau_{j_0}(0) = \frac{k}{2^N} = s_k. \tag{2}$$

However,

$$\rho_{j_0} \circ \rho_{j_1} \circ \cdots \circ \rho_{j_N-1}(0) = k = t_k. \tag{3}$$

We define a new ordering on $\mathcal{S}_N$ as follows:

$$\tilde{s}_k = \tau_{j_0} \circ \tau_{j_1} \circ \cdots \circ \tau_{j_N-1}(0) \tag{4}$$

where k is written in base 2. As we shall see in Theorem 9, this new ordering on $\mathcal{S}_N$ results in the following matrix equality:

$$(e^{-2\pi i t_j \tilde{s}_k})_{jk} = \mathcal{F}_N P \tag{5}$$

as in Equation (1).

We will call the compositions in Equations (3) and (4) the *obverse* ordering. The composition in Equation (2) will be called the *reverse* ordering. As suggested previously, and will be established in Theorem 9, if the elements of $\mathcal{S}_N$ and $\mathcal{T}_N$ are both ordered with the obverse compositions, then the permuted DFT matrix obtained is as in Equation (5). However, if both $\mathcal{S}_N$ and $\mathcal{T}_N$ are ordered using the reverse compositions, then the matrix becomes

$$(e^{-2\pi \tilde{i} t_j s_k})_{jk} = P\mathcal{F}_N = \begin{pmatrix} \mathcal{F}_{N-1} & \mathcal{F}_{N-1} \\ \mathcal{F}_{N-1}D & -\mathcal{F}_{N-1}D \end{pmatrix},$$

a block form that will allow the inverse $\mathcal{F}_N^{-1}$ to have a fast multiplication algorithm.

Consider the measure $\mu_N = \frac{1}{2^N}\sum_{s\in\mathcal{S}_N}\delta_s$; this sequence of measures converges weakly to Lebesgue measure restricted to $[0,1]$, the invariant measure for the IFS generated by $\{\tau_0,\tau_1\}$. Moreover, we consider the exponential functions $\{e^{2\pi it(\cdot)} : t \in \mathcal{T}_N\} \subset L^2(\mu_N)$; this set will be an orthonormal basis, and the DFT is the matrix representation of this basis (up to a scaling factor). Thus, the IFS generated by $\{\tau_0,\tau_1\}$ gives rise to a fractal, and the IFS generated by $\{\rho_0,\rho_1\}$ gives rise to the frequencies of an orthonormal set of exponentials.

A probability measure μ is *spectral* if there exists a set of frequencies $\Lambda \subset \mathbb{R}$ such that $\{e^{2\pi i\lambda(\cdot)} : \lambda \in \Lambda\} \subset L^2(\mu)$ is an orthonormal basis [5, 6]. If the measure is spectral, the set Λ is called a spectrum for μ. Jorgensen and Pederson [13] prove that the uniform measure supported on the middle-thirds Cantor set is not spectral. However, they prove that the invariant measure μ_4 for the iterated function system generated by

$$\tau_0(x) = \frac{x}{4}, \qquad \tau_1(x) = \frac{x+2}{4}$$

is spectral, and moreover, the spectrum is obtained via the iterated function system generated by

$$\rho_0(x) = 4x, \qquad \rho_1(x) = 4x+1.$$

In fact, the orbit of 0 under the iterated function system generated by $\{\rho_0,\rho_1\}$ is a spectrum for μ_4.

For a generic iterated function system $\{\psi_0,\ldots,\psi_{K-1}\}$ consisting of contractions on $\mathbb{R}^d$, we will consider an approximation $\mathcal{S}_N$ to the invariant set given by

$$\mathcal{S}_N := \{\psi_{j_{N-1}}\circ\psi_{j_{N-2}}\circ\cdots\circ\psi_{j_1}\circ\psi_{j_0}(0) : j_k \in \{0,1,\ldots,K-1\}\}.$$

This collection of points we will consider as the locations of data points. We then will choose a second iterated function system $\{\rho_0,\ldots,\rho_{K-1}\}$, and consider the finite orbit of 0:

$$\mathcal{T}_N := \{\rho_{j_{N-1}}\circ\rho_{j_{N-2}}\circ\cdots\circ\rho_{j_1}\circ\rho_{j_0}(0) : j_k \in \{0,1,\ldots,K-1\}\}.$$

These will be the frequencies for an exponential basis in $L^2(\mu_N)$, where $\mu_N = \frac{1}{K^N}\sum_{s\in\mathcal{S}_N}\delta_s$.

A necessary and sufficient condition to obtain an exponential basis for $L^2(\mu_N)$ from the frequencies in $\mathcal{T}_N$ is that the matrix

$$H_N = (e^{-2\pi i s_j t_k})_{j,k}$$

is invertible, where s_j and t_k range through $\mathcal{S}_N$ and $\mathcal{T}_N$ under any ordering, respectively. Preferably, the matrix H_N would be *Hadamard* [2, 7, 14], i.e. $H_N^* H_N = K^N I_{K^N}$ (since it automatically has entries of modulus 1), since this would correspond

to an orthogonal exponential basis. As we will show, if H_1 is invertible (Hadamard), then all H_N will be invertible (Hadamard, respectively).

Moreover, we will put an ordering (namely, the obverse ordering) on $\mathcal{S}_N$ and $\mathcal{T}_N$ so that under this ordering the matrix H_N has a block form in the manner of Diţă's construction for large Hadamard matrices. This block form will allow for the computational complexity of the matrix multiplication to be reduced. Then, $\mathcal{S}_N$ and $\mathcal{T}_N$ will be reordered (using the reverse ordering) so that the inverse of H_N will have a similar block form, again allowing for a fast algorithm for the matrix multiplication.

We note that for a generic IFS, the set $\mathcal{S}_N$ will consist of irregularly spaced points. We view the matrix H_N as being a Fourier transform for a signal (or set of data points) located at the points in $\mathcal{S}_N$, and thus H_N (and its block form as shown in Theorem 9) can be considered a non-equispaced FFT. We further note, however, that this is not a full irregularly spaced FFT, since all of the data point locations in $\mathcal{S}_N$ are rationally related. Please see [8, 9, 11] for the irregularly spaced FFT.

1.1 Diţă's Construction of Large Hadamard Matrices

Diţă's construction for large Hadamard matrices is as follows [4, 15]. If A is a $K \times K$ Hadamard matrix, B is an $M \times M$ Hadamard matrix, and $E_1, \ldots, E_{K-1}$ are $M \times M$ unitary diagonal matrices, then the $KM \times KM$ block matrix H is a Hadamard matrix:

$$H = \begin{pmatrix} a_{00}B & a_{01}E_1B & \ldots & a_{0(K-1)}E_{K-1}B \\ a_{10}B & a_{11}E_1B & \ldots & a_{1(K-1)}E_{K-1}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{(K-1)0}B & a_{(K-1)1}E_1B & \ldots & a_{(K-1)(K-1)}E_{K-1}B \end{pmatrix}. \tag{6}$$

Since we will also consider invertible matrices, not just Hadamard matrices, we show that for $A, B, E_1, \ldots, E_{K-1}$ invertible, H will also be invertible, and its inverse has a similar block form.

Proposition 1 *Suppose A and B are invertible, $E_1, \ldots, E_{K-1}$ are invertible and diagonal. Let $C = A^{-1}$. For the matrix H in Equation 6,*

$$H^{-1} = \begin{pmatrix} c_{00}B^{-1} & c_{01}B^{-1} & \ldots & c_{0(K-1)}B^{-1} \\ c_{10}B^{-1}E_1^{-1} & c_{11}B^{-1}E_1^{-1} & \ldots & c_{1(K-1)}B^{-1}E_1^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ c_{(K-1)0}B^{-1}E_{K-1}^{-1} & c_{1(K-1)}B^{-1}E_{K-1}^{-1} & \ldots & c_{(K-1)(K-1)}B^{-1}E_{K-1}^{-1} \end{pmatrix}. \tag{7}$$

Proof Let G be the block matrix in Equation (7), and let $E_0 = I_M$. Note that the product of H and G will have a block form. Multiplying the j-th row of H with the ℓ-th column of G, we obtain that the j, ℓ block of HG is:

$$\sum_{k=0}^{K-1}(a_{jk}E_kB)(c_{k\ell}B^{-1}E_k^{-1}) = \sum_{k=0}^{K-1} a_{jk}c_{k\ell}I_M.$$

Since $\sum_{k=0}^{K-1} a_{jk}c_{k\ell} = \delta_{j,\ell}$, we obtain $HG = I_{KM}$. □

If $A, B_0, \ldots, B_{K-1}, E_1, \ldots, E_{K-1}$ are all unitary, then the construction for H^{-1} gives H^*, so H is also unitary.

1.2 Complexity of Matrix Multiplication in Diţă's Construction

Let $\vec{v}$ be a vector of length KM. Consider $H\vec{v}$ where H is the block matrix as in Equation (6). We divide the vector $\vec{v}$ into K vectors of length M as follows:

$$\vec{v} = \begin{pmatrix} \vec{v}_0 \\ \vec{v}_1 \\ \vdots \\ \vec{v}_{K-1} \end{pmatrix}.$$

Then the matrix multiplication $H\vec{v}$ can be reduced in complexity, since

$$H\vec{v} = \begin{pmatrix} \sum_{j=0}^{K-1} a_{0j}E_jB\vec{v}_0 \\ \sum_{j=0}^{K-1} a_{1j}E_jB\vec{v}_1 \\ \vdots \\ \sum_{j=0}^{K-1} a_{(K-1)j}E_jB\vec{v}_{K-1} \end{pmatrix}.$$

Let $\mathcal{O}_M$ be the number of operations required to multiply the vector $\vec{w}$ of length M by the matrix B. The total number of operations required for each component of $H\vec{v}$ is $\mathcal{O}_M + M(K-1) + MK$ multiplications and $M(K-1)$ additions. The total number of operations for $H\vec{v}$ is then $K\mathcal{O}_M + 3MK^2 - 2MK$. We have just established the following proposition.

Proposition 2 *The product $H\vec{v}$ requires at most $K\mathcal{O}_M + 3MK^2 - 2MK$ operations.*

Since $\mathcal{O}_M = O(M^2)$, we obtain that the computational complexity of H is $O(M^2K + MK^2)$, whereas for a generic $KM \times KM$ matrix, the computational complexity is $O(K^2M^2)$. Thus, the block form of H reduces the computational complexity of the matrix multiplication.

2 A Fast Fourier Transform on $\mathcal{S}_N$

We consider an iterated function system generated by contractions $\{\psi_0, \psi_1, \ldots, \psi_{K-1}\}$ on $\mathbb{R}^d$ of the following form:

$$\psi_j(x) = A(x + \vec{b}_j)$$

where A is a $d \times d$ invertible matrix with $\|A\| < 1$. We require A^{-1} to have integer entries, the vectors $\vec{b}_j \in \mathbb{Z}^d$, and without loss of generality we suppose $\vec{b}_0 = \vec{0}$. We then choose a second iterated function system generated by $\{\rho_0, \rho_1, \ldots, \rho_{K-1}\}$ of the form

$$\rho_j(x) = Bx + \vec{c}_j$$

where $B = (A^T)^{-1}$, with $\vec{c}_j \in \mathbb{Z}^d$, and $\vec{c}_0 = \vec{0}$. We require the matrix

$$M_1 = (e^{-2\pi i \vec{c}_j \cdot A\vec{b}_k})_{j,k}$$

be invertible (or Hadamard). Note that depending on A and $\{\vec{b}_0, \vec{b}_1, \ldots, \vec{b}_{K-1}\}$, there may not be any choice $\{\vec{c}_0, \vec{c}_1, \ldots, \vec{c}_{K-1}\}$ so that M_1 is invertible. However, for many IFSs there is a choice:

Proposition 3 *If the set $\{\vec{b}_0, \vec{b}_1, \ldots, \vec{b}_{K-1}\}$ is such that for every pair $(j \neq k)$, $A\vec{b}_j - A\vec{b}_k \notin \mathbb{Z}^d$, then there exists $\{\vec{c}_0, \vec{c}_1, \ldots, \vec{c}_{K-1}\}$ such that the matrix M_1 is invertible.*

Proof The mappings $\phi_1 : \vec{x} \mapsto e^{2\pi i \vec{x} \cdot A\vec{b}_j}$ and $\phi_2 : \vec{x} \mapsto e^{2\pi i \vec{x} \cdot A\vec{b}_k}$ are characters on $G = \mathbb{Z}^d / B\mathbb{Z}^d$. Since $A\vec{b}_j - A\vec{b}_k \notin \mathbb{Z}^d$, the characters are distinct. Thus, by Schur orthogonality, $\sum_{x \in G} \phi_1(x)\overline{\phi_2(x)} = 0$. Therefore, the matrix $M = (e^{-2\pi i \vec{x}_k \cdot A\vec{b}_j})_{j,k}$, where $\{\vec{x}_k\}$ is any enumeration of G, has orthogonal columns. Thus, there is a choice of a square submatrix of M which is invertible. □

Even under the hypotheses of Proposition 3 there is not always a choice of $\vec{c}$'s so that M_1 is Hadamard; this is the case for the middle-third Cantor set, which is the attractor set for the IFS generated by $\psi_0(x) = \frac{x}{3}$, $\psi_1(x) = \frac{x+2}{3}$ (and is a reflection of the fact that μ_3 is not spectral).

Notation 1 We define our notation for compositions of the IFSs using two distinct orderings. Let $N \in \mathbb{N}$. For $j \in \{0, 1, \ldots, K^N - 1\}$, write $j = j_0 + j_1 K + \cdots + j_{N-1} K^{N-1}$ with $j_0, \ldots, j_{N-1} \in \{0, 1, \ldots K - 1\}$. We define

$$\Psi_{j,N} := \psi_{j_0} \circ \psi_{j_1} \circ \cdots \circ \psi_{j_{N-1}}$$
$$\mathcal{R}_{j,N} := \rho_{j_0} \circ \rho_{j_1} \circ \cdots \circ \rho_{j_{N-1}}.$$

These give rise to enumerations of $\mathcal{S}_N$ and $\mathcal{T}_N$ as follows:

$$\mathcal{S}_N = \{\Psi_{j,N}(0) : j = 0, 1, \ldots K^N - 1\}$$
$$\mathcal{T}_N = \{\mathcal{R}_{j,N}(0) : j = 0, 1, \ldots K^N - 1\}.$$

We call these the "obverse" orderings of $\mathcal{S}_N$ and $\mathcal{T}_N$.

Likewise, we define

$$\widetilde{\Psi}_{j,N} := \psi_{j_{N-1}} \circ \psi_{j_{N-2}} \circ \cdots \circ \psi_{j_0}$$
$$\widetilde{\mathcal{R}}_{j,N} := \rho_{j_{N-1}} \circ \rho_{j_{N-2}} \circ \cdots \circ \rho_{j_0}$$

which also enumerate $\mathcal{S}_N$ and $\mathcal{T}_N$. We call these the "reverse" orderings.

Remark 1 Note that for $N = 1$, $\Psi_{j,1} = \widetilde{\Psi}_{j,1}$ and $\mathcal{R}_{j,1} = \widetilde{\mathcal{R}}_{j,1}$.

We define the matrices M_N and $\widetilde{M}_N$ as follows:

$$[M_N]_{jk} = e^{-2\pi i \mathcal{R}_{j,N}(0)\cdot\Psi_{k,N}(0)}$$

and

$$[\widetilde{M}_N]_{jk} = e^{-2\pi i \widetilde{\mathcal{R}}_{j,N}(0)\cdot\widetilde{\Psi}_{k,N}(0)}.$$

Both of these are the matrix representations of the exponential functions with frequencies given by $\mathcal{T}_N$ on the data points given by $\mathcal{S}_N$. The matrix M_N corresponds to the obverse ordering on both $\mathcal{T}_N$ and $\mathcal{S}_N$, whereas the matrix $\widetilde{M}_N$ corresponds to the reverse ordering on both. Since these matrices arise from different orderings of the same sets, there exist permutation matrices P and Q such that

$$Q\widetilde{M}_N P = M_N. \tag{8}$$

Indeed, define for $j \in \{0, \ldots, K^N - 1\}$ a conjugate as follows: if $j = j_0 + j_1 K + \cdots + j_{N-1}K^{N-1}$, let $\tilde{j} = j_{N-1} + j_{N-2}K + \cdots + j_0 K^{N-1}$. Note then that $\tilde{\tilde{j}} = j$, and

$$\widetilde{\Psi}_{k,N} = \Psi_{\tilde{k},N} \qquad \widetilde{\mathcal{R}}_{k,N} = \mathcal{R}_{\tilde{k},N}. \tag{9}$$

Now, define a $K^N \times K^N$ permutation matrix P by $[P]_{mn} = 1$ if $n = \tilde{m}$, and 0 otherwise.

Lemma 4 *For P defined above,*

$$P\widetilde{M}_N P = M_N.$$

Proof We calculate

$$\begin{aligned}[P\widetilde{M}_N P]_{mn} &= \sum_k [P]_{mk} \sum_\ell [\widetilde{M}_N]_{k\ell}[P]_{\ell n} \\ &= [P]_{m\tilde{m}}[\widetilde{M}_N]_{\tilde{m}\tilde{n}}[P]_{\tilde{n}n} \\ &= e^{-2\pi i \widetilde{\mathcal{R}}_{\tilde{m},N}(0)\cdot\widetilde{\Psi}_{\tilde{n},N}(0)} \\ &= e^{-2\pi i \mathcal{R}_{m,N}(0)\cdot\Psi_{n,N}(0)} = [M_N]_{mn}\end{aligned}$$

by virtue of Equation (9). □

Proposition 5 *For scale* $N = 1$,

$$M_1 = \widetilde{M}_1 = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & \exp(2\pi i \vec{c}_1 \cdot A\vec{b}_1) & \dots & \exp(2\pi i \vec{c}_1 \cdot A\vec{b}_{K-1}) \\ \vdots & \vdots & \vdots\vdots & \vdots \\ 1 & \exp(2\pi i \vec{c}_{K-1} \cdot A\vec{b}_1) & \dots & \exp(2\pi i \vec{c}_{K-1} \cdot A\vec{b}_{K-1}) \end{pmatrix}.$$

Proof The proof follows from Remark 1. □

Lemma 6 *For* $N \in \mathbb{N}$, $0 \le j < K^N$, *and* $\vec{x}, \vec{y} \in \mathbb{R}^d$,

i) $\Psi_{j,N}\left(\vec{x} + \vec{y}\right) = \Psi_{j,N}(\vec{x}) + A^N\vec{y}$
ii) $\widetilde{\Psi}_{j,N}\left(\vec{x} + \vec{y}\right) = \widetilde{\Psi}_{j,N}(\vec{x}) + A^N\vec{y}$
iii) $\mathcal{R}_{j,N}\left(\vec{x} + \vec{y}\right) = R_{j,N}(\vec{x}) + B^N\vec{y}$
vi) $\widetilde{\mathcal{R}}_{j,N}\left(\vec{x} + \vec{y}\right) = \widetilde{R}_{j,N}(\vec{x}) + B^N\vec{y}$.

Proof We prove by induction on N. The base case is easily checked. Assume the equality in Item i) holds for $N-1$. For $j = j_0 + j_1K + \cdots + j_{N-1}K^{N-1}$, let $\ell = j - j_{N-1}K^{N-1}$. We have

$$\begin{aligned}\Psi_{j,N}\left(\vec{x} + \vec{y}\right) &= \Psi_{\ell,N-1}(\psi_{j_{N-1}}(\vec{x} + \vec{y})) \\ &= \Psi_{\ell,N-1}(\psi_{j_{N-1}}(\vec{x}) + A\vec{y}) \\ &= \Psi_{\ell,N-1}(\psi_{j_{N-1}}(\vec{x})) + A^{N-1}A\vec{y} \\ &= \Psi_{j,N}(\vec{x}) + A^N\vec{y}\end{aligned}$$

The proofs for the other three identities are analogous. □

Lemma 7 *For* $N \in \mathbb{N}$ *and* $0 \le j < K^N$,

i) $\Psi_{j,N}(0) = A^N\vec{z}$ *for some* $\vec{z} \in \mathbb{Z}^d$,
ii) $\widetilde{\Psi}_{j,N}(0) = A^N\vec{z}$ *for some* $\vec{z} \in \mathbb{Z}^d$,
iii) $\mathcal{R}_{j,N}(0) \in \mathbb{Z}^d$,
vi) $\widetilde{\mathcal{R}}_{j,N}(0) \in \mathbb{Z}^d$.

Proof We prove by induction on N. The base case is easily checked. Assume the equality in Item i) holds for $N-1$. For $j = j_0 + j_1 K + \cdots + j_{N-1}K^{N-1}$, let $q_j = j - j_{N-1}K^{N-1}$. We have

$$\begin{aligned}\Psi_{j,N}(0) &= \psi_{j_{N-1}}\left(\Psi_{q_j,N-1}(0)\right)\\ &= A\left(A^{N-1}\vec{z} + \vec{b}_j\right)\\ &= A^N(\vec{z} + A^{-(N-1)}\vec{b}_j)\end{aligned}$$

Since A^{-1} is an integer matrix, so is $A^{-(N-1)}$ and thus $\vec{z} + A^{-(N-1)}\vec{b}_j \in \mathbb{Z}^d$. Item ii) is analogous. For Item iii), note first that $\rho_j(\mathbb{Z}^d) \subset \mathbb{Z}^d$, so by induction, $\rho_{j_0} \circ \cdots \circ \rho_{j_{N-1}}(0) \in \mathbb{Z}^d$. Likewise for Item iv). □

Lemma 8 *Assume $N \geq 2$, let ℓ be an integer between 0 and $K-1$, and suppose $l \cdot K^{N-1} \leq j < (l+1)K^{N-1}$. Then,*

i) $\Psi_{j,N}(0) = \Psi_{j-l\cdot K^{N-1},N-1}(0) + A^N\vec{b}_l$,
ii) $\widetilde{\Psi}_{j,N}(0) = A\widetilde{\Psi}_{j-l\cdot K^{N-1},N-1}(0) + A\vec{b}_l$,
iii) $\mathcal{R}_{j,N}(0) = \mathcal{R}_{j-l\cdot K^{N-1},N-1}(0) + B^{N-1}\vec{c}_l$,
vi) $\widetilde{\mathcal{R}}_{j,N}(0) = B\widetilde{\mathcal{R}}_{j-l\cdot K^{N-1},N-1}(0) + \vec{c}_l$.

Proof For $l \cdot K^{N-1} \leq j < (l+1)K^{N-1}$, $j_{N-1} = l$, so we have

$$\begin{aligned}\Psi_{j,N}(0) &= \psi_{j_0} \circ \psi_{j_1} \circ \cdots \circ \psi_{j_{N-2}} \circ \psi_l(0)\\ &= \psi_{j_0} \circ \psi_{j_1} \circ \cdots \circ \psi_{j_{N-2}}\left(A(0 + \vec{b}_l)\right)\\ &= \Psi_{j-l\cdot K^{N-1},N-1}(0 + A\vec{b}_l).\end{aligned}$$

Applying Lemma 6 Item i) to $\Psi_{j-l\cdot K^{N-1},N-1}$:

$$\Psi_{j-l\cdot K^{N-1},N-1}(0 + A\vec{b}_l) = \Psi_{j-l\cdot K^{N-1},N-1}(0) + A^{N-1}A\vec{b}_l.$$

The proof of Item iii) is similar to Item i) with one crucial distinction, so we include the proof here. We have

$$\begin{aligned}\mathcal{R}_{j,N}(0) &= \rho_{j_0} \circ \rho_{j_1} \circ \cdots \circ \rho_{j_{N-2}} \circ \rho_l(0)\\ &= \rho_{j_0} \circ \rho_{j_1} \circ \cdots \circ \rho_{j_{N-2}}\left(B0 + \vec{c}_l\right)\\ &= \mathcal{R}_{j-l\cdot K^{N-1},N-1}(0 + \vec{c}_l).\end{aligned}$$

Applying Lemma 6 Item iii) to $R_{j-l\cdot K^{N-1},N-1}$:

$$\mathcal{R}_{j-l\cdot K^{N-1},N-1}(0 + \vec{c}_l) = \mathcal{R}_{j-l\cdot K^{N-1},N-1}(0) + B^{N-1}\vec{c}_l.$$

For Item ii), we have

$$\begin{aligned}\widetilde{\Psi}_{j,N}(0) &= \psi_\ell(\widetilde{\Psi}_{j-\ell\cdot K^{N-1},N-1}(0)) \\ &= A\widetilde{\Psi}_{j-\ell\cdot K^{N-1},N-1}(0) + A\vec{b}_\ell.\end{aligned}$$

The proof of Item iv) is analogous. □

Note that in Item i), the extra term involves A^N, whereas in Item iii) the extra term involves B^{N-1}. We are now in a position to prove our main theorem.

Theorem 9 *The matrix M_N representing the exponentials with frequencies given by $\mathcal{T}_N$ on the fractal approximation $\mathcal{S}_N$, when both are endowed with the obverse ordering, has the form*

$$M_N = \begin{pmatrix} m_{00}M_{N-1} & m_{01}D_{N,1}M_{N-1} & \dots & m_{0(K-1)}D_{N,K-1}M_{N-1} \\ m_{10}M_{N-1} & m_{11}D_{N,1}M_{N-1} & \dots & m_{1(K-1)}D_{N,K-1}M_{N-1} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ m_{(K-1)0}M_{N-1} & m_{(K-1)1}D_{N,1}M_{N-1} & \dots & m_{(K-1)(K-1)}D_{N,K-1}M_{N-1} \end{pmatrix}. \quad (10)$$

Here, $D_{N,m}$ are diagonal matrices with $[D_{N,m}]_{pp} = e^{-2\pi i\mathcal{R}_{p,N-1}(0)\cdot A^N\vec{b}_m}$, and $m_{jk} = [M_1]_{jk}$.

Proof Let us first subdivide M_N into blocks $B_{\ell m}$ of size $K^{N-1}\times K^{N-1}$, so that

$$M_N = \begin{pmatrix} B_{00} & \dots & B_{0(K-1)} \\ \vdots & \ddots & \vdots \\ B_{(K-1)0} & \dots & B_{(K-1)(K-1)} \end{pmatrix}.$$

Fix $0 \le j,k < K^N$ and suppose $\ell K^{N-1} \le j < (\ell+1)K^{N-1}$ and $mK^{N-1} \le k < (m+1)K^{N-1}$ with $0 \le \ell, m < K$. Let $q_j = j - \ell K^{N-1}$ and $q_k = k - mK^{N-1}$. Observe that

$$[M_N]_{jk} = [B_{\ell m}]_{q_j q_k}. \quad (11)$$

Using Lemma 8 Items ii) and iv), we calculate

$$\mathcal{R}_{j,N}(0)\cdot\Psi_{k,N}(0) = \left(\mathcal{R}_{q_j,N-1}(0) + B^{N-1}\vec{c}_\ell\right)\cdot\left(\Psi_{q_k,N-1}(0) + A^N\vec{b}_m\right).$$

By Lemma 7 Item i), for some $z\in\mathbb{Z}^d$,

$$B^{N-1}\vec{c}_\ell\cdot\Psi_{q_k,N-1}(0) = B^{N-1}\vec{c}_\ell\cdot A^{N-1}z = \vec{c}_\ell\cdot z\in\mathbb{Z}.$$

Note that

$$B^{N-1}\vec{c}_\ell\cdot A^N\vec{b}_m = \vec{c}_\ell\cdot A\vec{b}_m.$$

Therefore, combining the above, we obtain

$$\begin{aligned}[M_N]_{jk} &= e^{-2\pi i \mathcal{R}_{j,N}(0)\cdot\Psi_{k,N}(0)} \\ &= e^{-2\pi i \mathcal{R}_{q_j,N-1}(0)\cdot\Psi_{q_k,N-1}(0)} e^{-2\pi i \mathcal{R}_{q_j,N-1}(0)\cdot A^N\vec{b}_m} e^{-2\pi i \vec{c}_\ell\cdot A\vec{b}_m} \\ &= [M_{N-1}]_{q_j q_k} e^{-2\pi i \mathcal{R}_{q_j,N-1}(0)\cdot A^N\vec{b}_m} [M_1]_{\ell m}. \end{aligned} \tag{12}$$

Letting j vary between ℓK^{N-1} and $(\ell+1)K^{N-1}$ and k vary between mK^{N-1} and $(m+1)K^{N-1}$ corresponds to q_j and q_k varying between 0 and K^{N-1}. Therefore, we obtain from Equations (11) and (12) the matrix equation

$$B_{\ell m} = [M_1]_{\ell m} D_{N,m} M_{N-1}$$

where $[D_{N,m}]_{pp} = e^{-2\pi i \mathcal{R}_{p,N-1}(0)\cdot A^N\vec{b}_m}$ as claimed. □

Corollary 10 *The matrix M_N is invertible. If M_1 is Hadamard, then M_N is also Hadamard.*

Proof If M_1 is invertible, then by induction, M_N is invertible via Proposition 1. If M_1 is Hadamard, then again by induction, M_N is Hadamard by Diţă's construction. □

Theorem 11 *The matrix $\widetilde{M}_N$ representing the exponentials with frequencies given by $\mathcal{T}_N$ on the fractal approximation $\mathcal{S}_N$, when both are endowed with the reverse ordering, has the form*

$$\widetilde{M}_N = \begin{pmatrix} m_{00}\widetilde{M}_{N-1} & m_{01}\widetilde{M}_{N-1} & \cdots & m_{0(K-1)}\widetilde{M}_{N-1} \\ m_{10}\widetilde{M}_{N-1}\widetilde{D}_{N,1} & m_{11}\widetilde{M}_{N-1}\widetilde{D}_{N,1} & \cdots & m_{1(K-1)}\widetilde{M}_{N-1}\widetilde{D}_{N,1} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ m_{(K-1)0}\widetilde{M}_{N-1}\widetilde{D}_{N,K-1} & m_{(K-1)1}\widetilde{M}_{N-1}\widetilde{D}_{N,K-1} & \cdots & m_{(K-1)(K-1)}\widetilde{M}_{N-1}\widetilde{D}_{N,K-1} \end{pmatrix}. \tag{13}$$

Here, $\widetilde{D}_{N,q}$ is a diagonal matrix with $[\widetilde{D}_{N,\ell}]_{pp} = e^{-2\pi i c_\ell\cdot A(\widetilde{\Psi}_{p,N-1}(0))}$, and $m_{jk} = [M_1]_{jk}$.

Proof The proof proceeds similarly to the proof Theorem 9. Let us first subdivide $\widetilde{M}_N$ into $K^{N-1}\times K^{N-1}$ blocks $\widetilde{B}_{\ell m}$, so that

$$\widetilde{M}_N = \begin{pmatrix} \widetilde{B}_{00} & \cdots & \widetilde{B}_{0(K-1)} \\ \vdots & \ddots & \vdots \\ \widetilde{B}_{(K-1)0} & \cdots & \widetilde{B}_{(K-1)(K-1)} \end{pmatrix}.$$

Fix $0 \le j,k < K^N$ and suppose $\ell K^{N-1} \le j < (\ell+1)K^{N-1}$ and $mK^{N-1} \le k < (m+1)K^{N-1}$ with $0 \le \ell, m < K$. Let $q_j = j - \ell K^{N-1}$ and $q_k = k - mK^{N-1}$. Observe that

$$[\widetilde{M}_N]_{jk} = [\widetilde{B}_{\ell m}]_{q_j q_k}. \tag{14}$$

We calculate using Lemma 8 items ii) and iv):

$$\begin{aligned}\widetilde{\mathcal{R}}_{j,N}(0)\cdot\widetilde{\Psi}_{k,N}(0) &= (B\widetilde{\mathcal{R}}_{q_j,N-1}(0)+\vec{c}_\ell)\cdot(A\widetilde{\Psi}_{q_k,N-1}(0)+A\vec{b}_m)\\ &= \widetilde{\mathcal{R}}_{q_j,N-1}(0)\cdot\widetilde{\Psi}_{q_k,N-1}(0)+\vec{c}_\ell\cdot A\widetilde{\Psi}_{q_k,N-1}(0)\\ &\quad+\widetilde{\mathcal{R}}_{q_j,N-1}(0)\cdot\vec{b}_m+\vec{c}_\ell\cdot A\vec{b}_m.\end{aligned}$$

By Lemma 7 Item iv), $\widetilde{\mathcal{R}}_{q_j,N-1}(0)\cdot\vec{b}_m\in\mathbb{Z}$. Thus,

$$[\widetilde{B}_{\ell m}]_{q_jq_k}=[M_{N-1}]_{q_jq_k}e^{-2\pi i\vec{c}_\ell\cdot A\widetilde{\Psi}_{q_k,N-1}(0)}[M_1]_{\ell m}$$

and as in the proof of Theorem 9, we have

$$\widetilde{B}_{\ell m}=[M_1]_{\ell m}\widetilde{M}_{N-1}\widetilde{D}_{N,\ell}.$$

□

2.1 Computational Complexity of Theorems 9 and 11

As a consequence of Proposition 2, the matrix M_N can be multiplied by a vector of dimension K^N in at most $K\mathcal{P}_{N-1}+3K^{N+1}-2K^N$ operations, where $\mathcal{P}_{N-1}$ is the number of operations required by the matrix multiplication for M_{N-1}. Since M_{N-1} has the same block form as M_N, $\mathcal{P}_{N-1}$ can be determined by $\mathcal{P}_{N-2}$, etc. The proof of the following proposition is a standard induction argument, which we omit. Note that this says that the computational complexity for M_N is comparable to that for the FFT (recognizing the difference in the number of generators for the respective IFS's).

Proposition 12 *The number of operations to calculate the matrix multiplication $M_N\vec{v}$ is $\mathcal{P}_N=K^{N-1}\mathcal{P}_1+3(N-1)K^{N+1}-2(N-1)K^N$. Consequently, $\mathcal{P}_N=O(N\cdot K^N)$.*

The significance of Theorem 11 concerns the inverse of M_N. If P is the permutation matrix as in Lemma 4, then $M_N^{-1}=P\widetilde{M}_N^{-1}P$. By Proposition 1, $\widetilde{M}_N^{-1}$ has the form of Diţă's construction, and so the computational complexity of $\widetilde{M}_N^{-1}$ is the same as M_N. Thus, modulo multiplication by the permutation matrices P, the computational complexity of multiplication by M_N^{-1} is the same as that for M_N.

2.2 The Diagonal Matrices

The matrices M_N and $\widetilde{M}_N$ have the form of Diţă's construction as shown in Theorems 9 and 11. The block form of Diţă's construction involves diagonal matrices, which in Equations (10) and (13) are determined by the IFSs used to generate the matrices M_N and $\widetilde{M}_N$. As such, the diagonal matrices satisfy certain recurrence relations.

Theorem 13 *The diagonal matrices which appear in the block form of M_N (Equation (10)) satisfy the recurrence relation $D_{N,m} = D_{N-1,m} \otimes E_{N,m}$, where $E_{N,m}$ is the $K \times K$ diagonal matrix with $[E_{N,m}]_{uu} = e^{-2\pi i c_u \cdot A^N \vec{b}_m}$. That is:*

$$[D_{N,m}]_{pp} = [D_{N-1,m}]_{\widehat{p}\widehat{p}}\, e^{-2\pi i (c_{p_0} \cdot A^N \vec{b}_m)}$$

where $\widehat{p} = (p - p_0)/K$.

Likewise, the diagonal matrices which appear in the block form of $\widetilde{M}_N$ (Equation (13)) satisfy the recurrence relation $\widetilde{D}_{N,\ell} = \widetilde{D}_{N-1,\ell} \otimes \widetilde{E}_{N,\ell}$, where $\widetilde{E}_{N,\ell}$ is the $K \times K$ diagonal matrix with $[\widetilde{E}_{N,\ell}]_{uu} = e^{-2\pi i \vec{c}_\ell \cdot A^N \vec{b}_u}$. That is:

$$[\widetilde{D}_{N,\ell}]_{pp} = [D_{N-1,\ell}]_{\widehat{p}\widehat{p}}\, e^{-2\pi i \vec{c}_\ell \cdot A^N \vec{b}_{p_0}}.$$

Proof As demonstrated in Theorem 9, for $p = 0, 1, \ldots, K^{N-1}$, $[D_{N,m}]_{pp} = e^{-2\pi i \mathcal{R}_{p,N-1}(0) \cdot A^N \vec{b}_m}$. Note that $p_{N-1} = 0$, and $\rho_0(0) = 0$. We want to cancel one power of A in $A^N \vec{b}_m$, so we factor out a B from $\mathcal{R}_{p,N-1}(0)$:

$$\mathcal{R}_{p,N-1}(0) = \rho_{p_0} \circ \rho_{p_1} \circ \cdots \circ \rho_{p_{N-2}}(0) = B\left(\rho_{p_1} \circ \cdots \circ \rho_{p_{N-2}}(0)\right) + \vec{c}_{p_0}.$$

Since $\widehat{p} = p_1 + p_2 K + \cdots + p_{N-2} K^{N-3}$, $\mathcal{R}_{p,N-1}(0) = B R_{\widehat{p},N-2}(0) + \vec{c}_{p_0}$. Thus,

$$\begin{aligned}
[D_{N,m}]_{pp} &= e^{-2\pi i \mathcal{R}_{p,N-1}(0) \cdot A^N \vec{b}_m} \\
&= e^{-2\pi i (B\mathcal{R}_{\widehat{p},N-2}(0) \cdot A(A^{N-1}\vec{b}_m))} e^{-2\pi i (\vec{c}_{p_0} \cdot A^N \vec{b}_m)} \\
&= e^{-2\pi i (\mathcal{R}_{\widehat{p},N-2}(0) \cdot (A^{N-1}\vec{b}_m))} e^{-2\pi i (\vec{c}_{p_0} \cdot A^N \vec{b}_m)} \\
&= [D_{N-1,m}]_{\widehat{p}\widehat{p}}\, e^{-2\pi i (\vec{c}_{p_0} \cdot A^N \vec{b}_m)}.
\end{aligned}$$

Similarly, as demonstrated in Theorem 11, $[\widetilde{D}_{N,\ell}]_{pp} = e^{-2\pi i \vec{c}_\ell \cdot A(\widetilde{\Psi}_{p,N-1}(0))}$. We write:

$$\begin{aligned}
\widetilde{\Psi}_{p,N-1}(0) &= \psi_{p_{N-2}} \circ \psi_{p_{N-3}} \circ \cdots \circ \psi_{p_1} \circ \psi_{p_0}(0) \\
&= \psi_{p_{N-2}} \circ \psi_{p_{N-3}} \circ \cdots \circ \psi_{p_1}(0 + A\vec{b}_{p_0}) \\
&= \widetilde{\Psi}_{\widehat{p},N-2}(0 + A\vec{b}_{p_0}) \\
&= \widetilde{\Psi}_{\widehat{p},N-2}(0) + A^{N-1}\vec{b}_{p_0}.
\end{aligned}$$

where in the last equality we use Lemma 6 item ii). Therefore:

$$\begin{aligned}
[\widetilde{D}_{N,\ell}]_{pp} &= e^{-2\pi i c_\ell \cdot A(\widetilde{\Psi}_{p,N-1}(0))} \\
&= e^{-2\pi i \vec{c}_\ell \cdot A\left(\widetilde{\Psi}_{\widehat{p},N-2}(0)+A^{N-1}\vec{b}_{p0}\right)} \\
&= e^{-2\pi i \vec{c}_\ell \cdot A\left(\widetilde{\Psi}_{\widehat{p},N-2}(0)+A^{N}\vec{b}_{p0}\right)} \\
&= e^{-2\pi i \vec{c}_\ell \cdot A(\widetilde{\Psi}_{\widehat{p},N-2}(0))} e^{-2\pi i \vec{c}_\ell \cdot A^N \vec{b}_{p0}} \\
&= [\widetilde{D}_{N-1,\ell}]_{\widehat{p}\widehat{p}}\; e^{-2\pi i \vec{c}_\ell \cdot A^N \vec{b}_{p0}}.
\end{aligned}$$

□

References

1. L. Auslander, R. Tolimieri, Is computing with the finite Fourier transform pure or applied mathematics? Bull. Am. Math. Soc. (N.S.) **1**(6), 847–897 (1979). MR 546312 (81e:42020)
2. T. Banica, Quantum permutations, Hadamard matrices, and the search for matrix models. Banach Cent. Publ. **98**, 11–42 (2012)
3. J.W. Cooley, J.W. Tukey, An algorithm for the machine calculation of complex Fourier series. Math. Comput. **19**, 297–301 (1965). MR 0178586 (31 #2843)
4. P. Diţă, Some results on the parametrization of complex Hadamard matrices. J. Phys. A **37**(20), 5355–5374 (2004). MR 2065675 (2005b:15045)
5. D.E. Dutkay, P.E.T. Jorgensen, Fourier frequencies in affine iterated function systems. J. Funct. Anal. **247**(1), 110–137 (2007). MR MR2319756 (2008f:42007)
6. D.E. Dutkay, D. Han, Q. Sun, On the spectra of a Cantor measure. Adv. Math. **221**(1), 251–276 (2009). MR MR2509326
7. D.E. Dutkay, J. Haussermann, E. Weber, Spectral properties of small Hadamard matrices. Linear Algebra Appl. **506**, 363–381 (2016). ISSN:0024-3795, doi:10.1016/j.laa.2016.06.006, http://dx.doi.org/10.1016/j.laa.2016.06.006. MR CLASS 15B34 (05B20 11L05 65T50), MR NUMBER 3530685
8. A. Dutt, V. Rokhlin, Fast Fourier transforms for nonequispaced data. SIAM J. Sci. Comput. **14**(6), 1368–1393 (1993). MR 1241591 (95d:65114)
9. A. Dutt, V. Rokhlin, Fast Fourier transforms for nonequispaced data. II. Appl. Comput. Harmon. Anal. **2**(1), 85–100 (1995). MR 1313101 (95m:65222)
10. K. Falconer, *Fractal Geometry*. Mathematical Foundations and Applications (Wiley, Chichester, 1990). MR MR1102677 (92j:28008)
11. L. Greengard, J.-Y. Lee, Accelerating the nonuniform fast Fourier transform. SIAM Rev. **46**(3), 443–454 (2004). MR 2115056 (2006g:65222)
12. J.E. Hutchinson, Fractals and self-similarity. Indiana Univ. Math. J. **30**(5), 713–747 (1981). MR MR625600 (82h:49026)
13. P.E.T. Jorgensen, S. Pedersen, Dense analytic subspaces in fractal L^2-spaces. J. Anal. Math. **75**, 185–228 (1998). MR MR1655831 (2000a:46045)
14. J.H. McClellan, T.W. Parks, Eigenvalue and eigenvector decomposition of the discrete Fourier transform. IEEE Trans. Audio Electroacoust. **AU-20**(1), 66–74 (1972). MR 0399751 (53 #3593)
15. W. Tadej, K. Życzkowski, A concise guide to complex Hadamard matrices. Open Syst. Inf. Dyn. **13**(2), 133–177 (2006). MR 2244963 (2007f:15020)

Index

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4

Applied and Numerical Harmonic Analysis (77 volumes)

A. Saichev and W.A. Woyczyński: *Distributions in the Physical and Engineering Sciences* (ISBN 978-0-8176-3924-2)
C.E. D'Attellis and E.M. Fernandez-Berdaguer: *Wavelet Theory and Harmonic Analysis in Applied Sciences* (ISBN 978-0-8176-3953-2)
H.G. Feichtinger and T. Strohmer: *Gabor Analysis and Algorithms* (ISBN 978-0-8176-3959-4)
R. Tolimieri and M. An: *Time-Frequency Representations* (ISBN 978-0-8176-3918-1)
T.M. Peters and J.C. Williams: *The Fourier Transform in Biomedical Engineering* (ISBN 978-0-8176-3941-9)
G.T. Herman: *Geometry of Digital Spaces* (ISBN 978-0-8176-3897-9)
A. Teolis: *Computational Signal Processing with Wavelets* (ISBN 978-0-8176-3909-9)
J. Ramanathan: *Methods of Applied Fourier Analysis* (ISBN 978-0-8176-3963-1)
J.M. Cooper: *Introduction to Partial Differential Equations with MATLAB* (ISBN 978-0-8176-3967-9)
A. Procházka, N.G. Kingsbury, P.J. Payner, and J. Uhlir: *Signal Analysis and Prediction* (ISBN 978-0-8176-4042-2)
W. Bray and C. Stanojevic: *Analysis of Divergence* (ISBN 978-1-4612-7467-4)
G.T. Herman and A. Kuba: *Discrete Tomography* (ISBN 978-0-8176-4101-6)
K. Gröchenig: *Foundations of Time-Frequency Analysis* (ISBN 978-0-8176-4022-4)
L. Debnath: *Wavelet Transforms and Time-Frequency Signal Analysis* (ISBN 978-0-8176-4104-7)
J.J. Benedetto and P.J.S.G. Ferreira: *Modern Sampling Theory* (ISBN 978-0-8176-4023-1)
D.F. Walnut: *An Introduction to Wavelet Analysis* (ISBN 978-0-8176-3962-4)
A. Abbate, C. DeCusatis, and P.K. Das: *Wavelets and Subbands* (ISBN 978-0-8176-4136-8)

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 5*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-54711-4

O. Bratteli, P. Jorgensen, and B. Treadway: *Wavelets Through a Looking Glass* (ISBN 978-0-8176-4280-80)
H.G. Feichtinger and T. Strohmer: *Advances in Gabor Analysis* (ISBN 978-0-8176-4239-6)
O. Christensen: *An Introduction to Frames and Riesz Bases* (ISBN 978-0-8176-4295-2)
L. Debnath: *Wavelets and Signal Processing* (ISBN 978-0-8176-4235-8)
G. Bi and Y. Zeng: *Transforms and Fast Algorithms for Signal Analysis and Representations* (ISBN 978-0-8176-4279-2)
J.H. Davis: *Methods of Applied Mathematics with a MATLAB Overview* (ISBN 978-0-8176-4331-7)
J.J. Benedetto and A.I. Zayed: *Modern Sampling Theory* (ISBN 978-0-8176-4023-1)
E. Prestini: *The Evolution of Applied Harmonic Analysis* (ISBN 978-0-8176-4125-2)
L. Brandolini, L. Colzani, A. Iosevich, and G. Travaglini: *Fourier Analysis and Convexity* (ISBN 978-0-8176-3263-2)
W. Freeden and V. Michel: *Multiscale Potential Theory* (ISBN 978-0-8176-4105-4)
O. Christensen and K.L. Christensen: *Approximation Theory* (ISBN 978-0-8176-3600-5)
O. Calin and D.-C. Chang: *Geometric Mechanics on Riemannian Manifolds* (ISBN 978-0-8176-4354-6)
J.A. Hogan: *Time–Frequency and Time–Scale Methods* (ISBN 978-0-8176-4276-1)
C. Heil: *Harmonic Analysis and Applications* (ISBN 978-0-8176-3778-1)
K. Borre, D.M. Akos, N. Bertelsen, P. Rinder, and S.H. Jensen: *A Software-Defined GPS and Galileo Receiver* (ISBN 978-0-8176-4390-4)
T. Qian, M.I. Vai, and Y. Xu: *Wavelet Analysis and Applications* (ISBN 978-3-7643-7777-9)
G.T. Herman and A. Kuba: *Advances in Discrete Tomography and Its Applications* (ISBN 978-0-8176-3614-2)
M.C. Fu, R.A. Jarrow, J.-Y. Yen, and R.J. Elliott: *Advances in Mathematical Finance* (ISBN 978-0-8176-4544-1)
O. Christensen: *Frames and Bases* (ISBN 978-0-8176-4677-6)
P.E.T. Jorgensen, J.D. Merrill, and J.A. Packer: *Representations, Wavelets, and Frames* (ISBN 978-0-8176-4682-0)
M. An, A.K. Brodzik, and R. Tolimieri: *Ideal Sequence Design in Time-Frequency Space* (ISBN 978-0-8176-4737-7)
S.G. Krantz: *Explorations in Harmonic Analysis* (ISBN 978-0-8176-4668-4)
B. Luong: *Fourier Analysis on Finite Abelian Groups* (ISBN 978-0-8176-4915-9)
G.S. Chirikjian: *Stochastic Models, Information Theory, and Lie Groups, Volume 1* (ISBN 978-0-8176-4802-2)
C. Cabrelli and J.L. Torrea: *Recent Developments in Real and Harmonic Analysis* (ISBN 978-0-8176-4531-1)
M.V. Wickerhauser: *Mathematics for Multimedia* (ISBN 978-0-8176-4879-4)

B. Forster, P. Massopust, O. Christensen, K. Gröchenig, D. Labate, P. Vandergheynst, G. Weiss, and Y. Wiaux: *Four Short Courses on Harmonic Analysis* (ISBN 978-0-8176-4890-9)
O. Christensen: *Functions, Spaces, and Expansions* (ISBN 978-0-8176-4979-1)
J. Barral and S. Seuret: *Recent Developments in Fractals and Related Fields* (ISBN 978-0-8176-4887-9)
O. Calin, D.-C. Chang, and K. Furutani, and C. Iwasaki: *Heat Kernels for Elliptic and Sub-elliptic Operators* (ISBN 978-0-8176-4994-4)
C. Heil: *A Basis Theory Primer* (ISBN 978-0-8176-4686-8)
J.R. Klauder: *A Modern Approach to Functional Integration* (ISBN 978-0-8176-4790-2)
J. Cohen and A.I. Zayed: *Wavelets and Multiscale Analysis* (ISBN 978-0-8176-8094-7)
D. Joyner and J.-L. Kim: *Selected Unsolved Problems in Coding Theory* (ISBN 978-0-8176-8255-2)
G.S. Chirikjian: *Stochastic Models, Information Theory, and Lie Groups, Volume 2* (ISBN 978-0-8176-4943-2)
J.A. Hogan and J.D. Lakey: *Duration and Bandwidth Limiting* (ISBN 978-0-8176-8306-1)
G. Kutyniok and D. Labate: *Shearlets* (ISBN 978-0-8176-8315-3)
P.G. Casazza and P. Kutyniok: *Finite Frames* (ISBN 978-0-8176-8372-6)
V. Michel: *Lectures on Constructive Approximation* (ISBN 978-0-8176-8402-0)
D. Mitrea, I. Mitrea, M. Mitrea, and S. Monniaux: *Groupoid Metrization Theory* (ISBN 978-0-8176-8396-2)
T.D. Andrews, R. Balan, J.J. Benedetto, W. Czaja, and K.A. Okoudjou: *Excursions in Harmonic Analysis, Volume 1* (ISBN 978-0-8176-8375-7)
T.D. Andrews, R. Balan, J.J. Benedetto, W. Czaja, and K.A. Okoudjou: *Excursions in Harmonic Analysis, Volume 2* (ISBN 978-0-8176-8378-8)
D.V. Cruz-Uribe and A. Fiorenza: *Variable Lebesgue Spaces* (ISBN 978-3-0348-0547-6)
W. Freeden and M. Gutting: *Special Functions of Mathematical (Geo-)Physics* (ISBN 978-3-0348-0562-9)
A. Saichev and W.A. Woyczyński: *Distributions in the Physical and Engineering Sciences, Volume 2: Linear and Nonlinear Dynamics of Continuous Media* (ISBN 978-0-8176-3942-6)
S. Foucart and H. Rauhut: *A Mathematical Introduction to Compressive Sensing* (ISBN 978-0-8176-4947-0)
G. Herman and J. Frank: *Computational Methods for Three-Dimensional Microscopy Reconstruction* (ISBN 978-1-4614-9520-8)
A. Paprotny and M. Thess: *Realtime Data Mining: Self-Learning Techniques for Recommendation Engines* (ISBN 978-3-319-01320-6)
A. Zayed and G. Schmeisser: *New Perspectives on Approximation and Sampling Theory: Festschrift in Honor of Paul Butzer's 85th Birthday* (978-3-319-08800-6)
R. Balan, M. Begue, J. Benedetto, W. Czaja, and K.A Okoudjou: *Excursions in Harmonic Analysis, Volume 3* (ISBN 978-3-319-13229-7)

H. Boche, R. Calderbank, G. Kutyniok, J. Vybiral: *Compressed Sensing and its Applications* (ISBN 978-3-319-16041-2)
S. Dahlke, F. De Mari, P. Grohs, and D. Labate: *Harmonic and Applied Analysis: From Groups to Signals* (ISBN 978-3-319-18862-1)
G. Pfander: *Sampling Theory, a Renaissance* (ISBN 978-3-319-19748-7)
R. Balan, M. Begue, J. Benedetto, W. Czaja, and K.A Okoudjou: *Excursions in Harmonic Analysis, Volume 4* (ISBN 978-3-319-20187-0)
O. Christensen: *An Introduction to Frames and Riesz Bases, Second Edition* (ISBN 978-3-319-25611-5)
E. Prestini: *The Evolution of Applied Harmonic Analysis: Models of the Real World, Second Edition* (ISBN 978-1-4899-7987-2)
J.H. Davis: *Methods of Applied Mathematics with a Software Overview, Second Edition* (ISBN 978-3-319-43369-1)
M. Gilman, E. M. Smith, S. M. Tsynkov: *Transionospheric Synthetic Aperture Imaging* (ISBN 978-3-319-52125-1)
S. Chanillo, B. Franchi, G. Lu, C. Perez, E.T. Sawyer: *Harmonic Analysis, Partial Differential Equations and Applications* (ISBN 978-3-319-52741-3)
R. Balan, J. Benedetto, W. Czaja, M. Dellatorre, and K.A Okoudjou: *Excursions in Harmonic Analysis, Volume 5* (ISBN 978-3-319-54710-7)

For an up-to-date list of ANHA titles, please visit **http://www.springer.com/series/4968**