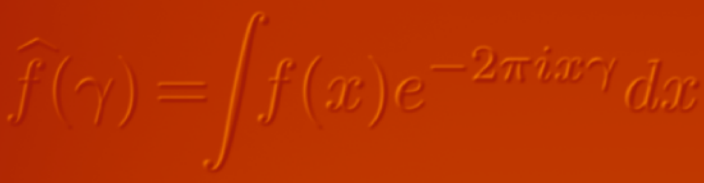
Applied and Numerical Harmonic Analysis



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

Excursions in Harmonic Analysis, Volume 4

The February Fourier Talks at the Norbert Wiener Center

Birkhäuser

Applied and Numerical Harmonic Analysis

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

Excursions in Harmonic Analysis, Volume 4

The February Fourier Talks at the Norbert Wiener Center

Birkhäuser

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

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

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

Matthew Begué
Department of Mathematics
Norbert Wiener Center
University of Maryland
College Park, MD, USA

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

ISSN 2296-5009 ISSN 2296-5017 (electronic)
Applied and Numerical Harmonic Analysis
ISBN 978-3-319-20187-0 ISBN 978-3-319-20188-7 (eBook)
DOI 10.1007/978-3-319-20188-7

Library of Congress Control Number: 2012951313

Mathematics Subject Classification (2010): 26-XX, 35-XX, 40-XX, 41-XX, 42-XX, 43-XX, 44-XX, 46-25 XX, 47-XX, 58-XX, 60-XX, 62-XX, 65-XX, 68-XX, 78-XX, 92-XX, 93-XX, 94-XX

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Dedicated to

Arne Beurling

sui generis among harmonic analysts

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
Biomedical signal processing
Digital signal processing
Fast algorithms
Gabor theory and applications
Image processing
Numerical partial differential equations
Prediction theory
Radar applications
Sampling theory
Spectral estimation
Speech processing
Time-frequency and time-scale analysis
Wavelet theory

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

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

A basic problem in Fourier analysis is to show how complicated phenomena, such as sound waves, can be described in terms of elementary harmonics. There are two aspects of this problem: first, to find, or even define properly, the harmonics or spectrum of a given phenomenon, e.g., the spectroscopy problem in optics; second, to determine which phenomena can be constructed from given classes of harmonics, as done, 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 these Volumes 3 and 4 have at least one author who spoke at the February Fourier Talks during the period 2002–2013 or at the workshop on Phaseless Reconstruction that immediately followed the 2013 February Fourier Talks. Volumes 1 and 2 were limited to the period 2006–2011.

The February Fourier Talks (FFT)

The *FFT* 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* have been organized by the NWC. The *FFT* have 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 two-day meeting.

The goals of the *FFT* are the following:

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

- To introduce young mathematicians and scientists to applied and pure harmonic analysis;
- To build bridges between pure harmonic analysis and applications thereof.

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

The *FFT* is formatted as a two-day single-track meeting consisting of 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 Lecturer	Keynote Address	Colloquium
• Ronald Coifman	• Peter Carr	• Rama Chellappa
• Ingrid Daubechies	• Barry Cipra	• Margaret Cheney
• Ronald DeVore	• James Coddington	• Charles Fefferman
• Richard Kadison	• Nathan Crone	• 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 were 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 Mathematical Engineering. Applied harmonic analysis and its theoretical underpinnings form the technological basis for this area. It can be confidently asserted that Mathematical Engineering will be to today's mathematics departments what Mathematical Physics was to those of a century ago. At that time, Mathematical Physics provided the impetus for tremendous advances within mathematics departments, with particular impact in fields such as differential equations, operator theory, and numerical analysis. Tools developed in these fields were essential in the advances of applied physics, e.g., the development of the solid state devices which now enable our information economy.

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

The golden age of Mathematical Engineering is upon us. The Norbert Wiener Center reflects the importance of integrating new mathematical technologies and algorithms in the context of current industrial and academic needs and problems.

The Norbert Wiener Center has three goals:

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

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

The Norbert Wiener Center 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

I Sampling Theory
II Remote Sensing
III Mathematics of Data Processing
IV Applications of Data Processing

Volume 2

V Measure Theory
VI Filtering
VII Operator Theory
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

College Park, MD, USA

Radu Balan
Matthew Begué
John J. Benedetto
Wojciech Czaja
Kasso A. Okoudjou

Acknowledgements

The editors of Volumes 3 and 4 gratefully acknowledge additional editorial assistance by Dr. Alfredo Nava-Tudela, as well as the support of Danielle Walker, Associate 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 SR2 Group, LLC. One of the successes of the February Fourier Talks has been the dynamic participation of graduate students 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 XIII
Theoretical Harmonic Analysis

The chapters in Part XIII illustrate the interplay between harmonic analysis and a number of areas in both pure and applied mathematics. In particular, the topics covered in this part include function space theory, frames and operators in Hilbert spaces, and nonlinear dispersive equations.

In the first chapter, ÁRPÁD BÉNYI, TADAHIRO OH , and OANA POCOVNICU introduce a randomization technique of functions defined on $\mathbb{R}^d$, that is naturally related to a Wiener-type decomposition of functions in the modulation spaces. Such technique allows them to consider the (almost sure) well-posedness of certain non-linear Schrödinger equations defined on $\mathbb{R}^d$. The chapter briefly surveys how randomization techniques have been used to establish well-posedness results for such equations when the space variable lies in a compact domain. Using this new (Wiener) randomization technique enables them to derive some improved (probabilistic) Strichartz estimates which in turn help established an almost sure well-posedness result for NLS.

In the second chapter, DAVID LARSON and SAM SCHOLZE give a detailed overview of a new operator theoretical approach to the erasure problem in frame theory. In a tutorial-style exposition, they describe a technique called bridging whose goal is to perfectly reconstruct signals/functions from a set of incomplete/lost frame or sampling coefficients. Using matrix analysis they consider the question both for the finite and infinite dimensional cases and provide conditions under which perfect reconstruction is possible. In addition, they give a number of algorithms (in the form of Matlab codes) showing how to implement their method in practice.

Chapter "Choosing Function Spaces in Harmonic Analysis" is an entertaining and highly informative survey of HANS FEICHTINGER on the question of choosing the right function space for a specific application. A parallel is drawn between this choice and the one that is made when buying a car! His point being that in the latter case one has a set of standards and options the desired automobile should possess. In this regard, the chapter argues that choosing a function space for a given application should be dealt with in a similar fashion. The chapter then proceeds with an inventory of different methods to construct function spaces and provides examples of such spaces in the form of the Wiener amalgam spaces and the modulation spaces. The chapter advocates for a new approach to Fourier analysis motivated by the use of certain ad hoc mathematical methods in signal processing.

In the final chapter of this part, MARCIN BOWNIK and JOHN JASPER survey the recent developments on the existence of (infinite dimensional) frames with prescribed norms and frame operator. In particular, they observe that such results can be viewed as Schur-Horn type theorems for the diagonals of positive self-adjoint operators with given spectral properties. Similar results in infinite dimensions were due to Kadison. The chapter presents some generalizations of Kadison's results dealing with characterization of certain frames with specified spectra.

Wiener randomization on unbounded domains and an application to almost sure well-posedness of NLS

Árpád Bényi*, Tadahiro Oh, and Oana Pocovnicu

Abstract We introduce a randomization of a function on $\mathbb{R}^d$ that is naturally associated to the Wiener decomposition and, intrinsically, to the modulation spaces. Such randomized functions enjoy better integrability, thus allowing us to improve the Strichartz estimates for the Schrödinger equation. As an example, we also show that the energy-critical cubic nonlinear Schrödinger equation on $\mathbb{R}^4$ is almost surely locally well posed with respect to randomized initial data below the energy space.

Key words: Nonlinear Schrödinger equation, Almost sure well-posedness, Modulation space, Wiener decomposition, Strichartz estimate, Fourier restriction norm method

* This work is partially supported by a grant from the Simons Foundation (No. 246024 to Árpád Bényi).

Á. Bényi (✉)
Department of Mathematics, Western Washington University, 516 High Street, Bellingham, WA 98225, USA
e-mail: arpad.benyi@wwu.edu

T. Oh
School of Mathematics, The University of Edinburgh, and The Maxwell Institute for the Mathematical Sciences, James Clerk Maxwell Building, The King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ, UK
e-mail: hiro.oh@ed.ac.uk

O. Pocovnicu
School of Mathematics, Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540, USA

Department of Mathematics, Princeton University, Fine Hall, Washington Rd., Princeton, NJ 08544-1000, USA

Department of Mathematics, Heriot-Watt University, Edinburgh, EH14 4AS, United Kingdom
e-mail: opocovnicu@math.princeton.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_1

Introduction

Background

The Cauchy problem of the nonlinear Schrödinger equation (NLS)

$$\begin{cases} i\partial_t u + \Delta u = \pm |u|^{p-1}u \\ u\big|_{t=0} = u_0 \in H^s(\mathbb{R}^d), \end{cases} \qquad (t,x) \in \mathbb{R}\times\mathbb{R}^d \tag{1}$$

has been studied extensively over recent years. One of the key ingredients in studying (1) is the dispersive effect of the associated linear flow. Such dispersion is often expressed in terms of the Strichartz estimates (see Lemma 1 below), which have played an important role in studying various problems on (1), in particular, local and global well-posedness issues.

It is well known that (1) is invariant under several symmetries. In the following, we concentrate on the dilation symmetry. The dilation symmetry states that if $u(t,x)$ is a solution to (1) on $\mathbb{R}^d$ with an initial condition u_0, then $u^\lambda(t,x) = \lambda^{-\frac{2}{p-1}} u(\lambda^{-2}t, \lambda^{-1}x)$ is also a solution to (1) with the λ-scaled initial condition $u_0^\lambda(x) = \lambda^{-\frac{2}{p-1}} u_0(\lambda^{-1}x)$. Associated to the dilation symmetry, there is a scaling-critical Sobolev index $s_c := \frac{d}{2} - \frac{2}{p-1}$ such that the homogeneous $\dot{H}^{s_c}$ norm is invariant under the dilation symmetry. For example, when $p = \frac{4}{d-2} + 1$, we have $s_c = 1$ and (1) is called energy critical. It is known that (1) is ill posed in the supercritical regime, that is, in H^s for $s < s_c$; see [1, 11, 16, 18].

In an effort to study the invariance of the Gibbs measure for the defocusing (Wick-ordered) cubic NLS on $\mathbb{T}^2$, Bourgain [6] considered random initial data of the form

$$u_0^\omega(x) = \sum_{n\in\mathbb{Z}^2} \frac{g_n(\omega)}{\sqrt{1+|n|^2}} e^{in\cdot x}, \tag{2}$$

where $\{g_n\}_{n\in\mathbb{Z}^2}$ is a sequence of independent complex-valued standard Gaussian random variables. Function (2) represents a typical element in the support of the Gibbs measure, more precisely, in the support of the Gaussian free-field on $\mathbb{T}^2$ associated to this Gibbs measure, and is critical with respect to the scaling. With a combination of deterministic PDE techniques and probabilistic arguments, Bourgain showed that the (Wick-ordered) cubic NLS on $\mathbb{T}^2$ is well posed almost surely with respect to random initial data (2). Burq-Tzvetkov [14] further explored the study of Cauchy problems with more general random initial data. They considered the cubic nonlinear wave equation (NLW) on a three-dimensional compact Riemannian manifold M without a boundary, where the scaling-critical Sobolev index s_c is

given by $s_c = \frac{1}{2}$. Given $u_0(x) = \sum_{n=1}^{\infty} c_n e_n(x) \in H^s(M)$, $s \geq \frac{1}{4}$, they proved almost sure local well-posedness with random initial data of the form[2]

$$u_0^{\omega}(x) = \sum_{n=1}^{\infty} g_n(\omega) c_n e_n(x), \tag{3}$$

where $\{g_n\}_{n=1}^{\infty}$ is a sequence of independent mean-zero random variables with a uniform bound on the fourth moments and $\{e_n\}_{n=1}^{\infty}$ is an orthonormal basis of $L^2(M)$ consisting of the eigenfunctions of the Laplace-Beltrami operator. It was also shown that u_0^{ω} in (3) has the same Sobolev regularity as the original function u_0 and is not smoother, almost surely. In particular, if $u_0 \in H^s(M) \setminus H^{\frac{1}{2}}(M)$, their result implies almost sure local well-posedness in the supercritical regime. There are several works on Cauchy problems of evolution equations with random data that followed these results, including some on almost sure global well-posedness: [7, 9, 10, 12, 13, 15, 19–22, 36–38, 43, 44, 49].

We point out that many of these works are on compact domains, where there is a countable basis of eigenfunctions of the Laplacian and thus there is a natural way to introduce a randomization. On $\mathbb{R}^d$, randomizations were introduced with respect to a countable basis of eigenfunctions of the Laplacian with a confining potential such as a harmonic oscillator $\Delta - |x|^2$; we note that functions in Sobolev spaces associated to the Laplacian with a confining potential have an extra decay in space. Our goal is to introduce a randomization for functions in the usual Sobolev spaces on $\mathbb{R}^d$ without such extra decay. For this purpose, we first review some basic notions and facts concerning the so-called *modulation spaces* of time-frequency analysis.

Modulation spaces

The modulation spaces were introduced by Feichtinger [23] in the early 1980s. In following collaborations with Gröchenig [24, 25, 27], they established the basic theory of these function spaces, in particular their invariance, continuity, embeddings, and convolution properties, see also [27]. The difference between the Besov spaces and the modulation spaces consists in the geometry of the frequency space employed: the dyadic annuli in the definition of the former spaces are replaced by unit cubes Q_n centered at $n \in \mathbb{Z}^d$ in the definition of the latter ones. Thus, the modulation spaces arise via a uniform partition of the frequency space $\mathbb{R}^d = \bigcup_{n \in \mathbb{Z}^d} Q_n$, which is commonly referred to as a *Wiener decomposition* [53]. In certain contexts, this decomposition allows for a finer analysis by effectively capturing the time-frequency concentration of a distribution.

For $x, \xi \in \mathbb{R}^d$, let $\mathscr{F}u(\xi) = \widehat{u}(\xi) = \int_{\mathbb{R}^d} u(x) e^{-2\pi i x \cdot \xi}\, dx$ denote the Fourier transform of a distribution u. Typically, the (weighted) modulation spaces $M_s^{p,q}(\mathbb{R}^d)$, $p, q > 0, s \in \mathbb{R}$, are defined by imposing the $L^p(dx)L^q(\langle \xi \rangle^s d\xi)$ integrability of

[2] For NLW, one needs to specify $(u, \partial_t u)|_{t=0}$ as an initial condition. For simplicity of presentation, we only displayed $u|_{t=0}$ in (3).

the short-time (or windowed) Fourier transform of a distribution $V_\phi u(x,\xi) := \mathscr{F}(u\overline{T_x\phi})(\xi)$. Here, $\langle\xi\rangle^s = (1+|\xi|^2)^{\frac{s}{2}}$, ϕ is some fixed non-zero Schwartz function, and T_x denotes the translation defined by $T_x(\phi)(y) = \phi(y-x)$. When $s=0$, one simply writes $M^{p,q}$. Modulation spaces satisfy some desirable properties: they are quasi-Banach spaces, two different windows ϕ_1, ϕ_2 yield equivalent norms, $M^{2,2}_s(\mathbb{R}^d) = H^s(\mathbb{R}^d)$, $(M^{p,q}_s(\mathbb{R}^d))' = M^{p',q'}_{-s}(\mathbb{R}^d)$, $M^{p_1,q_1}_{s_1}(\mathbb{R}^d) \subset M^{p_2,q_2}_{s_2}(\mathbb{R}^d)$ for $s_1 \geq s_2$, $p_1 \leq p_2$, and $q_1 \leq q_2$, and $\mathscr{S}(\mathbb{R}^d)$ is dense in $M^{p,q}_s(\mathbb{R}^d)$.

We prefer to use an equivalent norm on the modulation space $M^{p,q}_s$, which is induced by a corresponding Wiener decomposition of the frequency space. Given $\psi \in \mathscr{S}(\mathbb{R}^d)$ such that $\operatorname{supp}\psi \subset [-1,1]^d$ and $\sum_{n\in\mathbb{Z}^d}\psi(\xi-n) \equiv 1$, let

$$\|u\|_{M^{p,q}_s(\mathbb{R}^d)} = \big\| \langle n\rangle^s \|\psi(D-n)u\|_{L^p_x(\mathbb{R}^d)} \big\|_{\ell^q_n(\mathbb{Z}^d)}. \tag{4}$$

Note that $\psi(D-n)$ is just a Fourier multiplier with symbol χ_{Q_n} conveniently smoothed:

$$\psi(D-n)u(x) = \int_{\mathbb{R}^d} \psi(\xi-n)\widehat{u}(\xi)e^{2\pi i x\cdot\xi}\,d\xi.$$

It is worthwhile to compare definition (4) with the one for the Besov spaces which uses a dyadic partition of the frequency domain. Let $\varphi_0, \varphi \in \mathscr{S}(\mathbb{R}^d)$ such that $\operatorname{supp}\varphi_0 \subset \{|\xi| \leq 2\}$, $\operatorname{supp}\varphi \subset \{\frac{1}{2} \leq |\xi| \leq 2\}$, and $\varphi_0(\xi) + \sum_{j=1}^\infty \varphi(2^{-j}\xi) \equiv 1$. With $\varphi_j(\xi) = \varphi(2^{-j}\xi)$, we define the Besov spaces $B^{p,q}_s$ via the norm

$$\|u\|_{B^{p,q}_s(\mathbb{R}^d)} = \big\| 2^{js}\|\varphi_j(D)u\|_{L^p(\mathbb{R}^d)} \big\|_{\ell^q_j(\mathbb{Z}_{\geq 0})}. \tag{5}$$

There are several known embeddings between the Besov, Sobolev, and modulation spaces; see, for example, Okoudjou [39], Toft [50], Sugimoto-Tomita [47], and Kobayashi-Sugimoto [34].

Randomization adapted to the Wiener decomposition

Given a function ϕ on $\mathbb{R}^d$, we have

$$\phi = \sum_{n\in\mathbb{Z}^d} \psi(D-n)\phi,$$

where $\psi(D-n)$ is defined above. We introduce a randomization naturally associated to the Wiener decomposition, and hence to the modulation spaces, as follows. Let $\{g_n\}_{n\in\mathbb{Z}^d}$ be a sequence of independent mean zero complex-valued random variables on a probability space $(\Omega, \mathscr{F}, P)$, where the real and imaginary parts of g_n are independent and endowed with probability distributions $\mu_n^{(1)}$ and $\mu_n^{(2)}$. Then, we can define the *Wiener randomization of* ϕ by

$$\phi^\omega := \sum_{n\in\mathbb{Z}^d} g_n(\omega)\psi(D-n)\phi. \tag{6}$$

In the sequel, we make the following assumption: there exists $c > 0$ such that

$$\left| \int_{\mathbb{R}} e^{\gamma x} d\mu_n^{(j)}(x) \right| \leq e^{c\gamma^2} \tag{7}$$

for all $\gamma \in \mathbb{R}$, $n \in \mathbb{Z}^d$, $j = 1, 2$. Note that (7) is satisfied by standard complex-valued Gaussian random variables, standard Bernoulli random variables, and any random variables with compactly supported distributions.

It is easy to see that if $\phi \in H^s(\mathbb{R}^d)$, then the randomized function ϕ^ω is almost surely in $H^s(\mathbb{R}^d)$; see Lemma 3 below. One can also show that there is no smoothing upon randomization in terms of differentiability; see, for example, Lemma B.1 in [14]. Instead, the main point of this randomization is its improved integrability; if $\phi \in L^2(\mathbb{R}^d)$, then the randomized function ϕ^ω is almost surely in $L^p(\mathbb{R}^d)$ for any finite $p \geq 2$; see Lemma 4 below. Such results for random Fourier series are known as Paley-Zygmund's theorem [41]; see also Kahane's book [30] and Ayache-Tzvetkov [2].

Remark 1. One may fancy a randomization associated to Besov spaces of the form:

$$\phi^\omega := \sum_{j=0}^{\infty} g_n(\omega) \varphi(D)\phi.$$

In view of the Littlewood-Paley theory, such a randomization does not yield any improvement on differentiability or integrability and thus it is of no interest.

Main results

The Wiener randomization of an initial condition allows us to establish some improvements of the Strichartz estimates. In turn, these probabilistic Strichartz estimates yield an almost sure well-posedness result for NLS. First, we recall the usual Strichartz estimates on $\mathbb{R}^d$ for the reader's convenience. We say that a pair (q, r) is *Schrödinger admissible* if it satisfies

$$\frac{2}{q} + \frac{d}{r} = \frac{d}{2} \tag{8}$$

with $2 \leq q, r \leq \infty$, and $(q, r, d) \neq (2, \infty, 2)$. Let $S(t) = e^{it\Delta}$. Then, the following Strichartz estimates are known to hold.

Lemma 1 ([26, 31, 46, 54]). *Let (q, r) be Schrödinger admissible. Then, we have*

$$\|S(t)\phi\|_{L_t^q L_x^r(\mathbb{R}\times\mathbb{R}^d)} \lesssim \|\phi\|_{L_x^2(\mathbb{R}^d)}. \tag{9}$$

Next, we present improvements of the Strichartz estimates under the Wiener randomization. Proposition 1 will be then used for a local-in-time theory, while Propo-

sition 2 is useful for small data global theory. The proofs of Propositions 1 and 2 are presented in section "Probabilistic Strichartz estimates".

Proposition 1 (Improved local-in-time Strichartz estimate). *Given $\phi \in L^2(\mathbb{R}^d)$, let ϕ^ω be its randomization defined in (6), satisfying (7). Then, given $2 \le q, r < \infty$, there exist $C, c > 0$ such that*

$$P\Big(\|S(t)\phi^\omega\|_{L_t^q L_x^r([0,T]\times\mathbb{R}^d)} > \lambda\Big) \le C\exp\left(-c\frac{\lambda^2}{T^{\frac{2}{q}}\|\phi\|_{L^2}^2}\right) \tag{10}$$

for all $T > 0$ and $\lambda > 0$.

In particular, by setting $\lambda = T^\theta \|\phi\|_{L^2}$, we have

$$\|S(t)\phi^\omega\|_{L_t^q L_x^r([0,T]\times\mathbb{R}^d)} \lesssim T^\theta \|\phi\|_{L^2(\mathbb{R}^d)}$$

outside a set of probability at most $C\exp\big(-cT^{2\theta-\frac{2}{q}}\big)$. Note that as long as $\theta < \frac{1}{q}$, this probability can be made arbitrarily small by letting $T \to 0$. Moreover, for fixed $T > 0$, we have the following: given any small $\varepsilon > 0$, we have

$$\|S(t)\phi^\omega\|_{L_t^q L_x^r([0,T]\times\mathbb{R}^d)} \le C_T\left(\log\frac{1}{\varepsilon}\right)^{\frac{1}{2}}\|\phi\|_{L^2}$$

outside a set of probability $< \varepsilon$.

The next proposition states an improvement of the Strichartz estimates in the global-in-time setting.

Proposition 2 (Improved global-in-time Strichartz estimate). *Given $\phi \in L^2(\mathbb{R}^d)$, let ϕ^ω be its randomization defined in (6), satisfying (7). Given a Schrödinger admissible pair (q,r) with $q, r < \infty$, let $\widetilde{r} \ge r$. Then, there exist $C, c > 0$ such that*

$$P\Big(\|S(t)\phi^\omega\|_{L_t^q L_x^{\widetilde{r}}(\mathbb{R}\times\mathbb{R}^d)} > \lambda\Big) \le Ce^{-c\lambda^2\|\phi\|_{L^2}^{-2}} \tag{11}$$

for all $\lambda > 0$. In particular, given any small $\varepsilon > 0$, we have

$$\|S(t)\phi^\omega\|_{L_t^q L_x^{\widetilde{r}}(\mathbb{R}\times\mathbb{R}^d)} \lesssim \left(\log\frac{1}{\varepsilon}\right)^{\frac{1}{2}}\|\phi\|_{L^2}$$

outside a set of probability at most ε.

We conclude this introduction by discussing an example of almost sure local well-posedness of NLS with randomized initial data below a scaling critical regularity. In the following, we consider the energy-critical cubic NLS on $\mathbb{R}^4$:

$$i\partial_t u + \Delta u = \pm|u|^2 u, \qquad (t,x) \in \mathbb{R}\times\mathbb{R}^4. \tag{12}$$

Cazenave-Weissler [17] proved local well-posedness of (12) with initial data in the critical space $\dot{H}^1(\mathbb{R})$. See Ryckman-Vişan [45], Vişan [51], and Kenig-Merle [32] for global-in-time results. In the following, we state a local well-posedness result of (12) with random initial data below the critical space. More precisely, given $\phi \in H^s(\mathbb{R}^4) \setminus H^1(\mathbb{R}^4)$, $s \in (\frac{3}{5}, 1)$, and ϕ^ω its randomization defined in (6), we prove that (12) is almost surely locally well posed with random initial data ϕ^ω. Although ϕ and its randomization ϕ^ω lie in a supercritical regularity regime, the Wiener randomization essentially makes the problem *subcritical*. This is a common feature for many of the probabilistic well-posedness results.

Theorem 1. *Let $s \in (\frac{3}{5}, 1)$. Given $\phi \in H^s(\mathbb{R}^4)$, let ϕ^ω be its randomization defined in (6), satisfying (7). Then, the cubic NLS (12) on $\mathbb{R}^4$ is almost surely locally well posed with respect to the randomization ϕ^ω as initial data. More precisely, there exist $C, c, \gamma > 0$ and $\sigma = 1+$ such that for each $T \ll 1$, there exists a set $\Omega_T \subset \Omega$ with the following properties:*

(i) $P(\Omega \setminus \Omega_T) \leq C \exp\Big(- \frac{c}{T^\gamma \|\phi\|_{H^s}^2} \Big)$,

(ii) *For each $\omega \in \Omega_T$, there exists a (unique) solution u to (12) with $u|_{t=0} = \phi^\omega$ in the class*

$$S(t)\phi^\omega + C([-T,T] : H^\sigma(\mathbb{R}^4)) \subset C([-T,T] : H^s(\mathbb{R}^4)).$$

The details of the proof of Theorem 1 are presented in section "Almost sure local well-posedness". We discuss here a very brief outline of the argument. Denoting the linear and nonlinear parts of u by $z(t) = z^\omega(t) := S(t)\phi^\omega$ and $v(t) := u(t) - S(t)\phi^\omega$, respectively, we can reduce (12) to

$$\begin{cases} i\partial_t v + \Delta v = \pm |v+z|^2 (v+z) \\ v|_{t=0} = 0. \end{cases} \tag{13}$$

We then prove that the Cauchy problem (13) is almost surely locally well posed for v, viewing z as a random forcing term. This is done by using the standard subcritical $X^{s,b}$ spaces with $b > \frac{1}{2}$ defined by

$$\|u\|_{X^{s,b}(\mathbb{R}\times\mathbb{R}^4)} = \| \langle \xi \rangle^s \langle \tau + |\xi|^2 \rangle^b \widehat{u}(\tau, \xi) \|_{L^2_{\tau,\xi}(\mathbb{R}\times\mathbb{R}^4)}.$$

We point out that the uniqueness in Theorem 1 refers to uniqueness of the nonlinear part $v(t) = u(t) - S(t)\phi^\omega$ of a solution u.

We conclude this introduction with several remarks.

Remark 2. Theorem 1 holds for both defocusing and focusing cases (corresponding to the $+$ sign and the $-$ sign in (1), respectively) due to the local-in-time nature of the problem.

Remark 3. Theorem 1 can also be proven with the variants of the $X^{s,b}$ spaces adapted to the U^p and V^p spaces introduced by Koch, Tataru, and their collaborators [28, 29, 35]. These spaces are designed to handle problems in critical regularities. We decided to present the proof with the classical subcritical $X^{s,b}$ spaces,

$b > \frac{1}{2}$, to emphasize that the problem has become subcritical upon randomization. We should, however, point out that with the spaces introduced by Koch and Tataru, we can also prove probabilistic small data global well-posedness and scattering as a consequence of the probabilistic global-in-time Strichartz estimates (Proposition 2). See our paper [3] for an example of such results for the cubic NLS on $\mathbb{R}^d$, $d \geq 3$.

It is of interest to consider almost sure global existence for (12). While the mass of v in (13) has a global-in-time control, there is no energy conservation for v and thus we do not know how to proceed at this point. In [3], we establish almost sure global existence for (12), assuming an a priori control on the H^1 norm of the nonlinear part v of a solution. We also prove there, without any assumption, global existence with a large probability by considering a randomization, not on unit cubes but on dilated cubes this time.

In the context of the energy-critical defocusing cubic NLW on $\mathbb{R}^4$, one can obtain an a priori bound on the energy of the nonlinear part of a solution, see [15]. As a consequence, the third author [42] proved almost sure global well-posedness of the energy-critical defocusing cubic NLW on $\mathbb{R}^4$ below the scaling critical regularity.

Remark 4. In Theorem 1, we simply used $\sigma = 1+$ as the regularity of the nonlinear part v. It is possible to characterize the possible values of σ in terms of the regularity $s < 1$ of ϕ. However, for simplicity of presentation, we omitted such a discussion.

Remark 5. In probabilistic well-posedness results [5, 7, 19, 37] for NLS on $\mathbb{T}^d$, random initial data are assumed to be of the following specific form:

$$u_0^\omega(x) = \sum_{n \in \mathbb{Z}^d} \frac{g_n(\omega)}{(1+|n|^2)^{\frac{\alpha}{2}}} e^{in\cdot x}, \tag{14}$$

where $\{g_n\}_{n\in\mathbb{Z}^d}$ is a sequence of independent complex-valued standard Gaussian random variables. Expression (14) has a close connection to the study of invariant (Gibbs) measures and, hence, it is of importance. At the same time, due to the lack of a full range of Strichartz estimates on $\mathbb{T}^d$, one could not handle a general randomization of a given function as in (3). In Theorem 1, we consider NLS on $\mathbb{R}^4$ and thus we do not encounter this issue thanks to a full range of the Strichartz estimates. For NLW, finite speed of propagation allows us to use a full range of Strichartz estimates even on compact domains, at least locally in time; thus, in that context, one does not encounter such an issue.

Remark 6. In a recent preprint, Lührmann-Mendelson [36] considered the defocusing NLW on $\mathbb{R}^3$ with randomized initial data defined in (6) in a supercritical regularity and proved almost sure global well-posedness in the energy-subcritical case, following the method developed in [19]. For the energy-critical quintic NLW on $\mathbb{R}^3$, they obtained almost sure local well-posedness along with small data global existence and scattering.

Probabilistic Strichartz estimates

In this section, we state and prove some basic properties of the randomized function ϕ^ω defined in (6), including the improved Strichartz estimates (Propositions 1 and 2). First, recall the following probabilistic estimate. See [14] for the proof.

Lemma 2. *Assume* (7)*. Then, there exists $C > 0$ such that*

$$\Big\| \sum_{n\in\mathbb{Z}^d} g_n(\omega) c_n \Big\|_{L^p(\Omega)} \le C\sqrt{p}\|c_n\|_{\ell^2_n(\mathbb{Z}^d)}$$

for all $p \ge 2$ and $\{c_n\} \in \ell^2(\mathbb{Z}^d)$.

Given $\phi \in H^s$, it is easy to see that its randomization $\phi^\omega \in H^s$ almost surely, for example, if $\{g_n\}$ has a uniform finite variance. Under assumption (7), we have a more precise description on the size of ϕ^ω.

Lemma 3. *Given $\phi \in H^s(\mathbb{R}^d)$, let ϕ^ω be its randomization defined in* (6)*, satisfying* (7)*. Then, we have*

$$P\Big(\|\phi^\omega\|_{H^s(\mathbb{R}^d)} > \lambda \Big) \le Ce^{-c\lambda^2\|\phi\|_{H^s}^{-2}} \tag{15}$$

for all $\lambda > 0$.

Proof. By Minkowski's integral inequality and Lemma 2, we have

$$\begin{aligned}\Big(\mathbb{E}\|\phi^\omega\|^p_{H^s(\mathbb{R}^d)}\Big)^{\frac{1}{p}} &\le \Big\| \|\langle\nabla\rangle^s \phi^\omega\|_{L^p(\Omega)} \Big\|_{L^2_x(\mathbb{R}^d)} \lesssim \sqrt{p} \Big\| \|\psi(D-n)\langle\nabla\rangle^s\phi\|_{\ell^2_n} \Big\|_{L^2_x} \\ &\sim \sqrt{p}\|\phi\|_{H^s}\end{aligned}$$

for any $p \ge 2$. Thus, we have obtained

$$\mathbb{E}[\|\phi^\omega\|^p_{H^s}] \le C_0^p p^{\frac{p}{2}} \|\phi\|^p_{H^s}.$$

By Chebyshev's inequality, we have

$$P\Big(\|\phi^\omega\|_{H^s} > \lambda \Big) < \left(\frac{C_0 p^{\frac{1}{2}} \|\phi\|_{H^s}}{\lambda} \right)^p \tag{16}$$

for $p \ge 2$.

Let $p = \left(\frac{\lambda}{C_0 e \|\phi\|_{H^s}} \right)^2$. If $p \ge 2$, then by (16), we have

$$P\Big(\|\phi^\omega\|_{H^s} > \lambda \Big) < \left(\frac{C_0 p^{\frac{1}{2}} \|\phi\|_{H^s}}{\lambda} \right)^p = e^{-p} = e^{-c\lambda^2\|\phi\|_{H^s}^{-2}}.$$

Otherwise, i.e., if $p = \left(\frac{\lambda}{C_0 e \|\phi\|_{H^s}}\right)^2 \leq 2$, we can choose C such that $Ce^{-2} \geq 1$. Then, we have

$$P\Big(\|\phi^\omega\|_{H^s} > \lambda\Big) \leq 1 \leq Ce^{-2} \leq Ce^{-c\lambda^2\|\phi\|_{H^s}^{-2}},$$

thus giving the desired result.

The next lemma shows that if $\phi \in L^2(\mathbb{R}^d)$, then its randomization ϕ^ω is almost surely in $L^p(\mathbb{R}^d)$ for any $p \in [2,\infty)$.

Lemma 4. *Given $\phi \in L^2(\mathbb{R}^d)$, let ϕ^ω be its randomization defined in (6), satisfying (7). Then, given finite $p \geq 2$, there exist $C, c > 0$ such that*

$$P\Big(\|\phi^\omega\|_{L^p(\mathbb{R}^d)} > \lambda\Big) \leq Ce^{-c\lambda^2\|\phi\|_{L^2}^{-2}} \tag{17}$$

for all $\lambda > 0$. In particular, ϕ^ω is in $L^p(\mathbb{R}^d)$ almost surely.

Proof. By Lemma 2, we have

$$\begin{aligned}\Big(\mathbb{E}\|\phi^\omega\|_{L^p_x(\mathbb{R}^d)}^r\Big)^{\frac{1}{r}} &\leq \Big\|\|\phi^\omega\|_{L^r(\Omega)}\Big\|_{L^p_x(\mathbb{R}^d)} \lesssim \sqrt{r}\Big\|\|\psi(D-n)\phi\|_{\ell^2_n}\Big\|_{L^p_x} \\ &\leq \sqrt{r}\Big\|\|\psi(D-n)\phi\|_{L^p_x}\Big\|_{\ell^2_n} \leq \sqrt{r}\Big\|\|\psi(D-n)\phi\|_{L^2_x}\Big\|_{\ell^2_n} \\ &\sim \sqrt{r}\|\phi\|_{L^2_x}\end{aligned}$$

for any $r \geq p$. Then, (17) follows as in the proof of Lemma 3.

We conclude this section by presenting the proofs of the improved Strichartz estimates under randomization. Before continuing further, we briefly recall the definitions of the smooth projections from Littlewood-Paley theory. Let φ be a smooth real-valued bump function supported on $\{\xi \in \mathbb{R}^d : |\xi| \leq 2\}$ and $\varphi \equiv 1$ on $\{\xi : |\xi| \leq 1\}$. If $N > 1$ is a dyadic number, we define the smooth projection $\mathbf{P}_{\leq N}$ onto frequencies $\{|\xi| \leq N\}$ by

$$\widehat{\mathbf{P}_{\leq N} f}(\xi) := \varphi\big(\tfrac{\xi}{N}\big)\widehat{f}(\xi).$$

Similarly, we can define the smooth projection $\mathbf{P}_N$ onto frequencies $\{|\xi| \sim N\}$ by

$$\widehat{\mathbf{P}_N f}(\xi) := \Big(\varphi\big(\tfrac{\xi}{N}\big) - \varphi\big(\tfrac{2\xi}{N}\big)\Big)\widehat{f}(\xi).$$

We make the convention that $\mathbf{P}_{\leq 1} = \mathbf{P}_1$. Bernstein's inequality states that

$$\|\mathbf{P}_{\leq N} f\|_{L^q(\mathbb{R}^d)} \lesssim N^{\frac{d}{p}-\frac{d}{q}}\|\mathbf{P}_{\leq N} f\|_{L^p(\mathbb{R}^d)}, \qquad 1 \leq p \leq q \leq \infty. \tag{18}$$

The same inequality holds if we replace $\mathbf{P}_{\leq N}$ by $\mathbf{P}_N$. As an immediate corollary, we have

$$\|\psi(D-n)\phi\|_{L^q(\mathbb{R}^d)} \lesssim \|\psi(D-n)\phi\|_{L^p(\mathbb{R}^d)}, \qquad 1 \leq p \leq q \leq \infty, \tag{19}$$

for all $n \in \mathbb{Z}^d$. This follows from applying (18) to $\phi_n(x) := e^{2\pi i n\cdot x}\psi(D-n)\phi(x)$ and noting that $\operatorname{supp}\widehat{\phi_n} \subset [-1,1]^d$. The point of (19) is that once a function is (roughly) restricted to a cube, we do not need to lose any regularity to go from the L^q norm to the L^p norm, $q \geq p$.

Proof of Proposition 1. Let $q, r \geq 2$. We write L^q_T to denote $L^q_t([0,T])$. By Lemma 2 and (19), we have

$$\begin{aligned}\Big(\mathbb{E}\|S(t)\phi^\omega\|^p_{L^q_t L^r_x([0,T]\times\mathbb{R}^d)}\Big)^{\frac{1}{p}} &\leq \Big\|\|S(t)\phi^\omega\|_{L^p(\Omega)}\Big\|_{L^q_T L^r_x}\\ &\leq \sqrt{p}\Big\|\|\psi(D-n)S(t)\phi\|_{\ell^2_n}\Big\|_{L^q_T L^r_x} \leq \sqrt{p}\Big\|\|\psi(D-n)S(t)\phi\|_{L^r_x}\Big\|_{L^q_T \ell^2_n}\\ &\lesssim \sqrt{p}\Big\|\|\psi(D-n)S(t)\phi\|_{L^2_x}\Big\|_{L^q_T \ell^2_n} \lesssim T^{\frac{1}{q}}\sqrt{p}\|\phi\|_{L^2_x}\end{aligned}$$

for $p \geq \max(q,r)$. Then, (10) follows as in the proof of Lemma 3.

Proof of Proposition 2. Let (q,r) be Schrödinger admissible and $\widetilde{r} \geq r$. Then, proceeding as before, we have

$$\begin{aligned}\Big(\mathbb{E}\|S(t)\phi^\omega\|^p_{L^q_t L^{\widetilde{r}}_x(\mathbb{R}\times\mathbb{R}^d)}\Big)^{\frac{1}{p}} &\lesssim \sqrt{p}\Big\|\|\psi(D-n)S(t)\phi\|_{L^{\widetilde{r}}_x}\Big\|_{\ell^2_n L^q_t}\\ &\lesssim \sqrt{p}\Big\|\|\psi(D-n)S(t)\phi\|_{L^q_t L^r_x}\Big\|_{\ell^2_n}.\end{aligned}$$

By Lemma 1,

$$\lesssim \sqrt{p}\Big\|\|\psi(D-n)\phi\|_{L^2_x}\Big\|_{\ell^2_n} \sim \sqrt{p}\|\phi\|_{L^2_x}$$

for $p \geq \max(q,\widetilde{r})$. Finally, (11) follows as above.

Remark 7. The Cauchy problem (1) has also been studied for initial data in the modulation spaces $M^{p,1}_s$ for $1 \leq p \leq \infty$ and $s \geq 0$; see, for example, [4] and [52]. Thus, it is tempting to consider what happens if we randomize an initial condition in a modulation space $M^{p,q}_s$. In this case, however, there is no improvement in the Strichartz estimates in terms of integrability, i.e., p, hence, no improvement of well-posedness with respect to $M^{p,q}_s$ in terms of differentiability, i.e. in s. Indeed, by computing the moments of the modulation norm of the randomized function (6), one immediately sees that the modulation norm remains essentially unchanged due to the outside summation over n. In the proof of Propositions 1 and 2, the averaging effect of a linear combination of the random variables g_n played a crucial role. For the modulation spaces, we do not have such an averaging effect since the outside summation over n forces us to work on a piece restricted to each cube, i.e., each random variable at a time.

Almost sure local well-posedness

Given $\phi \in H^s(\mathbb{R}^d)$, let ϕ^ω be its randomization defined in (6). In the following, we consider the Cauchy problem (12) with random initial data $u|_{t=0} = \phi^\omega$. By letting $z(t) = z^\omega(t) := S(t)\phi^\omega$ and $v(t) := u(t) - S(t)\phi^\omega$, we can reduce (1) to

$$\begin{cases} i\partial_t v + \Delta v = \pm|v+z|^2(v+z) \\ v|_{t=0} = 0. \end{cases} \tag{20}$$

By expressing (20) in the Duhamel formulation, we have

$$v(t) = \mp i \int_0^t S(t-t')\mathcal{N}(v+z)(t')dt', \tag{21}$$

where $\mathcal{N}(u) = |u|^2 u = u\bar{u}u$. Let η be a smooth cutoff function supported on $[-2,2]$, $\eta \equiv 1$ on $[-1,1]$, and let $\eta_T(t) = \eta\left(\frac{t}{T}\right)$. Note that if v satisfies

$$v(t) = \mp i\eta(t) \int_0^t S(t-t')\eta_T(t')\mathcal{N}(\eta v + \eta_T z)(t')dt' \tag{22}$$

for some $T \ll 1$, then it also satisfies (21) on $[-T,T]$. Hence, we consider (22) in the following.

Given $z(t) = S(t)\phi^\omega$, define Γ by

$$\Gamma v(t) = \mp i\eta(t) \int_0^t S(t-t')\eta_T(t')\mathcal{N}(\eta v + \eta_T z)(t')dt'. \tag{23}$$

Then, the following nonlinear estimates yield Theorem 1.

Proposition 3. *Let $s \in \left(\frac{3}{5}, 1\right)$. Given $\phi \in H^s(\mathbb{R}^4)$, let ϕ^ω be its randomization defined in (6), satisfying (7). Then, there exists $\sigma = 1+$, $b = \frac{1}{2}+$, and $\theta = 0+$ such that for each small $T \ll 1$ and $R > 0$, we have*

$$\|\Gamma v\|_{X^{\sigma,b}} \le C_1 T^\theta (\|v\|^3_{X^{\sigma,b}} + R^3), \tag{24}$$

$$\|\Gamma v_1 - \Gamma v_2\|_{X^{\sigma,b}} \le C_2 T^\theta (\|v_1\|^2_{X^{\sigma,b}} + \|v_2\|^2_{X^{\sigma,b}} + R^2)\|v_1 - v_2\|_{X^{\sigma,b}} \tag{25}$$

outside a set of probability at most $C\exp\left(-c\frac{R^2}{\|\phi\|^2_{H^s}}\right)$.

We first present the proof of Theorem 1, assuming Proposition 3. Then, we prove Proposition 3 at the end of this section.

Proof of Theorem 1. Let B_1 denote the ball of radius 1 centered at the origin in $X^{\sigma,b}$. Then, given $T \ll 1$, we show that the map Γ is a contraction on B_1. Given $T \ll 1$, we choose $R = R(T) \sim T^{-\frac{\gamma}{2}}$ for some $\gamma \in (0, \frac{2\theta}{3})$ such that

$$C_1 T^\theta (1 + R^3) \le 1 \quad \text{and} \quad C_2 T^\theta (2 + R^2) \le \frac{1}{2}.$$

Then, for $v, v_1, v_2 \in B_1$, Proposition 3 yields

$$\|\Gamma v\|_{X^{\sigma,b}} \le 1,$$
$$\|\Gamma v_1 - \Gamma v_2\|_{X^{\sigma,b}} \le \frac{1}{2}\|v_1 - v_2\|_{X^{\sigma,b}}$$

outside an exceptional set of probability at most

$$C\exp\bigg(-c\frac{R^2}{\|\phi\|_{H^s}^2}\bigg) = C\exp\bigg(-\frac{c}{T^\gamma\|\phi\|_{H^s}^2}\bigg).$$

Therefore, by defining Ω_T to be the complement of this exceptional set, it follows that for $\omega \in \Omega_T$, there exists a unique $v^\omega \in B_1$ such that $\Gamma v^\omega = v^\omega$. This completes the proof of Theorem 1.

Hence, it remains to prove Proposition 3. Before proceeding further, we first present some lemmata on the basic $X^{s,b}$ estimates. See [5, 33, 48] for the basic properties of the $X^{s,b}$ spaces.

Lemma 5. (i) *Linear estimates: Let $T \in (0,1)$ and $b \in \big(\frac{1}{2}, \frac{3}{2}\big]$. Then, for $s \in \mathbb{R}$ and $\theta \in \big[0, \frac{3}{2} - b\big)$, we have*

$$\|\eta_T(t)S(t)\phi\|_{X^{s,b}(\mathbb{R}\times\mathbb{R}^4)} \lesssim T^{\frac{1}{2}-b}\|\phi\|_{H^s(\mathbb{R}^4)}, \tag{26}$$
$$\bigg\|\eta(t)\int_0^t S(t-t')\eta_T(t')F(t')dt'\bigg\|_{X^{s,b}(\mathbb{R}\times\mathbb{R}^4)} \lesssim T^\theta\|F\|_{X^{s,b-1+\theta}(\mathbb{R}\times\mathbb{R}^4)}.$$

(ii) *Strichartz estimates: Let (q,r) be Schrödinger admissible and $p \ge 3$. Then, for $b > \frac{1}{2}$ and $N_1 \le N_2$, we have*

$$\|u\|_{L_t^q L_x^r(\mathbb{R}\times\mathbb{R}^4)} \lesssim \|u\|_{X^{0,b}(\mathbb{R}\times\mathbb{R}^4)}, \tag{27}$$
$$\|u\|_{L_{t,x}^p(\mathbb{R}\times\mathbb{R}^4)} \lesssim \big\||\nabla|^{2-\frac{6}{p}}u\big\|_{X^{0,b}(\mathbb{R}\times\mathbb{R}^4)}, \tag{28}$$
$$\|\mathbf{P}_{N_1}u_1\mathbf{P}_{N_2}u_2\|_{L_{t,x}^2(\mathbb{R}\times\mathbb{R}^4)}$$
$$\lesssim N_1\bigg(\frac{N_1}{N_2}\bigg)^{\frac{1}{2}}\|\mathbf{P}_{N_1}u_1\|_{X^{0,b}(\mathbb{R}\times\mathbb{R}^4)}\|\mathbf{P}_{N_2}u_2\|_{X^{0,b}(\mathbb{R}\times\mathbb{R}^4)}. \tag{29}$$

Recall that (27) follows from the Strichartz estimate (9) and (28) follows from Sobolev inequality and (9), while (29) follows from a refinement of the Strichartz estimate by Bourgain [8] and Ozawa-Tsutsumi [40].

As a corollary to Lemma 5, we have the following estimates.

Lemma 6. *Given small $\varepsilon > 0$, let $\varepsilon_1 = 2\varepsilon+$. Then, for $N_1 \le N_2$, we have*

$$\|u\|_{L_{t,x}^{\frac{3}{1+\varepsilon_1}}(\mathbb{R}\times\mathbb{R}^4)} \lesssim \|u\|_{X^{0,\frac{1}{2}-2\varepsilon}(\mathbb{R}\times\mathbb{R}^4)}, \tag{30}$$

$$\begin{aligned}\|\mathbf{P}_{N_1}u_1\mathbf{P}_{N_2}u_2\|_{L_{t,x}^2(\mathbb{R}\times\mathbb{R}^4)}&\\ \lesssim C(N_1,N_2)\|\mathbf{P}_{N_1}u_1\|_{X^{0,\frac{1}{2}+}(\mathbb{R}\times\mathbb{R}^4)}&\|\mathbf{P}_{N_2}u_2\|_{X^{0,\frac{1}{2}-2\varepsilon}(\mathbb{R}\times\mathbb{R}^4)},\end{aligned} \tag{31}$$

where $C(N_1,N_2)$ is given by

$$C(N_1,N_2) = \begin{cases} N_1^{\frac{3}{2}+\varepsilon_1+}N_2^{-\frac{1}{2}+\varepsilon_1} & \text{if } N_1 \le N_2, \\ N_1^{-\frac{1}{2}+5\varepsilon_1+}N_2^{\frac{3}{2}-3\varepsilon_1} & \text{if } N_1 \ge N_2. \end{cases}$$

Proof. The first estimate (30) follows from interpolating (27) with $q = r = 3$ and $\|u\|_{L_{t,x}^2} = \|u\|_{X^{0,0}}$. The second estimate (31) follows from interpolating (29) and

$$\begin{aligned}\|\mathbf{P}_{N_1}u_1\mathbf{P}_{N_2}u_2\|_{L_{t,x}^2} &\le \|\mathbf{P}_{N_1}u_1\|_{L_{t,x}^\infty}\|\mathbf{P}_{N_2}u_2\|_{L_{t,x}^2} \\ &\lesssim \|\mathbf{P}_{N_1}u_1\|_{X^{2+,\frac{1}{2}+}}\|\mathbf{P}_{N_2}u_2\|_{X^{0,0}}.\end{aligned}$$

This completes the proof of Lemma 6.

We are now ready to prove Proposition 3.

Proof of Proposition 3. We only prove (24) since (25) follows in a similar manner. By Lemma 5 (i) and duality, we have

$$\begin{aligned}\|\Gamma v(t)\|_{X^{\sigma,b}} &\lesssim T^\theta \|\mathcal{N}(\eta v + \eta_T z)\|_{X^{\sigma,b-1+\theta}} \\ &= T^\theta \sup_{\|v_4\|_{X^{0,1-b-\theta}}\le 1} \left| \iint_{\mathbb{R}\times\mathbb{R}^4} \langle\nabla\rangle^\sigma \big[\mathcal{N}(\eta v + \eta_T z)\big] v_4 dxdt \right|.\end{aligned} \tag{32}$$

We estimate the right-hand side of (32) by performing case-by-case analysis of expressions of the form:

$$\left| \iint_{\mathbb{R}\times\mathbb{R}^4} \langle\nabla\rangle^\sigma (w_1w_2w_3) v_4 dxdt \right|, \tag{33}$$

where $\|v_4\|_{X^{0,1-b-\theta}} \le 1$ and $w_j = \eta v$ or $\eta_T z$, $j = 1,2,3$. Before proceeding further, let us simplify some of the notations. In the following, we drop the complex conjugate sign since it plays no role. Also, we often suppress the smooth cutoff function η (and η_T) from $w_j = \eta v$ (and $w_j = \eta_T z$) and simply denote them by v_j (and z_j, respectively). Lastly, in most of the cases, we dyadically decompose w_j, $j = 1,2,3$, and v_4 such that their spatial frequency supports are $\{|\xi_j| \sim N_j\}$ for some dyadic $N_j \ge 1$ but still denote them by w_j, $j = 1,2,3$, and v_4.

Let $b = \frac{1}{2} + \varepsilon$ and $\theta = \varepsilon$ for some small $\varepsilon > 0$ (to be chosen later) so that $1 - b - \theta = \frac{1}{2} - 2\varepsilon$. In the following, we set $\varepsilon_1 = 2\varepsilon+$.

Case (1): *vvv* case.

In this case, we do not need to perform dyadic decompositions and we divide the frequency spaces into $\{|\xi_1| \geq |\xi_2|, |\xi_3|\}$, $\{|\xi_2| \geq |\xi_1|, |\xi_3|\}$, and $\{|\xi_3| \geq |\xi_1|, |\xi_2|\}$. Without loss of generality, assume that $|\xi_1| \gtrsim |\xi_2|, |\xi_3|$. By $L^3 L^{\frac{6}{1-\varepsilon_1}} L^{\frac{6}{1-\varepsilon_1}} L^{\frac{3}{1+\varepsilon_1}}$-Hölder's inequality and Lemmata 5 and 6, we have

$$\begin{aligned}\bigg| \int_{\mathbb{R}\times\mathbb{R}^4} \langle\nabla\rangle^{\sigma} v_1 v_2 v_3 v_4 dxdt \bigg| &\leq \|\langle\nabla\rangle^{\sigma} v_1\|_{L^3_{t,x}} \|v_2\|_{L^{\frac{6}{1-\varepsilon_1}}_{t,x}} \|v_3\|_{L^{\frac{6}{1-\varepsilon_1}}_{t,x}} \|v_4\|_{L^{\frac{3}{1+\varepsilon_1}}_{t,x}} \\ &\lesssim \prod_{j=1}^{3} \|v_j\|_{X^{\sigma,\frac{1}{2}+}} \|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}} \lesssim \prod_{j=1}^{3} \|v_j\|_{X^{\sigma,b}}\end{aligned}$$

for $\sigma \geq 1 + \varepsilon_1 = 1 + 2\varepsilon+$.

Case (2): *zzz* case.

Without loss of generality, assume $N_3 \geq N_2 \geq N_1$.

• **Subcase (2.a):** $N_2 \sim N_3$.

By $L^{\frac{6}{1-2\varepsilon_1}} L^4 L^4 L^{\frac{3}{1+\varepsilon_1}}$-Hölder's inequality and Lemmata 5 and 6, we have

$$\begin{aligned}&\bigg| \int_{\mathbb{R}\times\mathbb{R}^4} z_1 z_2 \langle\nabla\rangle^{\sigma} z_3 v_4 dxdt \bigg| \\ &\qquad \lesssim \|z_1\|_{L^{\frac{6}{1-2\varepsilon_1}}_{t,x}} \big\|\langle\nabla\rangle^{\frac{\sigma}{2}} z_2\big\|_{L^4_{t,x}} \big\|\langle\nabla\rangle^{\frac{\sigma}{2}} z_3\big\|_{L^4_{t,x}} \|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}.\end{aligned}$$

Hence, by Proposition 1, the contribution to (33) in this case is at most $\lesssim R^3$ outside a set of probability

$$\leq C \exp\bigg(- c \frac{R^2}{T^{\frac{1-2\varepsilon_1}{3}} \|\phi\|^2_{L^2}} \bigg) + C \exp\bigg(- c \frac{R^2}{T^{\frac{1}{2}} \|\phi\|^2_{H^s}} \bigg) \tag{34}$$

provided that $s > \frac{\sigma}{2}$. Note that s needs to be strictly greater than $\frac{\sigma}{2}$ due to the summations over dyadic blocks. For the convenience of readers, we briefly show how this follows. In summing $\|\langle\nabla\rangle^{\frac{\sigma}{2}} \mathbf{P}_{N_3} z_3\|_{L^4_{t,x}}$ over dyadic blocks in N_3, we have

$$\begin{aligned}\sum_{\substack{N_3 \geq 1 \\ \text{dyadic}}} \|\langle\nabla\rangle^{\frac{\sigma}{2}} \mathbf{P}_{N_3} z_3\|_{L^4_{t,x}} &\leq \Big(\sum_{N_3} N_3^{0-} \Big)^{\frac{3}{4}} \|\langle\nabla\rangle^{\frac{\sigma}{2}+} \mathbf{P}_{N_3} z_3\|_{\ell^4_{N_3} L^4_{t,x}} \\ &= \Big(\sum_{N_3} N_3^{0-} \Big)^{\frac{3}{4}} \|\langle\nabla\rangle^{\frac{\sigma}{2}+} \mathbf{P}_{N_3} z_3\|_{L^4_{t,x} \ell^4_{N_3}} \\ &\leq \Big(\sum_{N_3} N_3^{0-} \Big)^{\frac{3}{4}} \|\langle\nabla\rangle^{\frac{\sigma}{2}+} \mathbf{P}_{N_3} z_3\|_{L^4_{t,x} \ell^2_{N_3}} \lesssim \|\langle\nabla\rangle^{\frac{\sigma}{2}+} z_3\|_{L^4_{t,x}},\end{aligned}$$

where the last inequality follows from the Littlewood-Paley theory. By Proposition 1 with $q = r = 4$, we obtain the second term in (34) as long as $s > \frac{\sigma}{2}$. Moreover, while

the terms with z_1 and z_2 also suffer a slight loss of derivative, we can hide the loss in N_1 and N_2 under the z_3 term since $N_3 \geq N_1, N_2$. Similar comments also apply in the sequel.

• **Subcase (2.b):** $N_3 \sim N_4 \gg N_1, N_2$.

○ Subsubcase (2.b.i): $N_1, N_2 \ll N_3^{\frac{1}{3}}$.

We include the detailed calculation only in this case, with similar comments applicable in the following. By Lemmata 5 (ii) and 6, with $b = \frac{1}{2}+$ and $\delta = 0+$, we have

$$\begin{aligned}\left| \int_{\mathbb{R}\times\mathbb{R}^4} z_1 z_2 \langle\nabla\rangle^{\sigma} z_3 v_4 dxdt \right| &\lesssim \|z_1 \langle\nabla\rangle^{\sigma} z_3\|_{L^2_{t,x}} \|z_2 v_4\|_{L^2_{t,x}} \\ &\lesssim N_1^{\frac{3}{2}} N_3^{-\frac{1}{2}+\sigma} N_2^{\frac{3}{2}+\varepsilon_1+\delta} N_4^{-\frac{1}{2}+\varepsilon_1} \prod_{j=1}^{3} \|z_j\|_{X^{0,b}} \|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}} \\ &\lesssim N_1^{\frac{3}{2}-s} N_2^{\frac{3}{2}+\varepsilon_1-s+\delta} N_3^{-\frac{1}{2}+\sigma-s} N_4^{-\frac{1}{2}+\varepsilon_1} \prod_{j=1}^{3} \|z_j\|_{X^{s,b}} \|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}\end{aligned}$$

By $N_1, N_2, \ll N_3^{\frac{1}{3}}$, $N_3 \sim N_4$, and Lemma 5 (i), we have

$$\ll T^{0-} N_3^{-\frac{5}{3}s+\sigma+\frac{4}{3}\varepsilon_1+\frac{1}{3}\delta} \prod_{j=1}^{3} \|\mathbf{P}_{N_j}\phi^{\omega}\|_{H^s} \|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}.$$

Here, we lost a small power of T in applying (26). Note that such a loss in T can be hidden under T^{θ} in (32) and does not cause a problem. Now, we want the power of the largest frequency N_3 to be strictly negative so that we can sum over dyadic blocks. This requires

$$\frac{5}{3}s > \sigma + \frac{4}{3}\varepsilon_1. \tag{35}$$

Provided this condition holds, using Lemma 3, we see that the contribution to (33) in this case is at most $\lesssim T^{0-}R^3$ outside a set of probability

$$\leq C \exp\left(-c \frac{R^2}{\|\phi\|^2_{H^s}} \right).$$

○ Subsubcase (2.b.ii): $N_2 \gtrsim N_3^{\frac{1}{3}} \gg N_1$.

By Lemmata 5 and 6, we have

$$\begin{aligned}\left| \int_{\mathbb{R}\times\mathbb{R}^4} z_1 z_2 \langle\nabla\rangle^{\sigma} z_3 v_4 dxdt \right| &\lesssim \|z_2\|_{L^4_{t,x}} \|\langle\nabla\rangle^{\sigma} z_3\|_{L^4_{t,x}} \|z_1 v_4\|_{L^2_{t,x}} \\ &\lesssim T^{0-} N_1^{\frac{3}{2}+\varepsilon_1-s+} N_2^{-s} N_3^{\sigma-s-\frac{1}{2}+\varepsilon_1} \|\mathbf{P}_{N_1}\phi^{\omega}\|_{H^s} \prod_{j=2}^{3} \|\langle\nabla\rangle^{s} z_j\|_{L^4_{t,x}} \|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}.\end{aligned}$$

Hence, by Lemma 3 and Proposition 1, the contribution to (33) in this case is at most $\lesssim T^{0-}R^3$ outside a set of probability

$$\leq C\exp\left(-c\frac{R^2}{\|\phi\|_{H^s}^2}\right)+C\exp\left(-c\frac{R^2}{T^{\frac{1}{2}}\|\phi\|_{H^s}^2}\right)$$

provided that (35) is satisfied.

◦ Subsubcase (2.b.iii): $N_1, N_2 \geq N_3^{\frac{1}{3}}$.

By $L^{\frac{9}{2-\varepsilon_1}}L^{\frac{9}{2-\varepsilon_1}}L^{\frac{9}{2-\varepsilon_1}}L^{\frac{3}{1+\varepsilon_1}}$-Hölder's inequality and Lemmata 5 and 6, we have

$$\left|\int_{\mathbb{R}\times\mathbb{R}^4} z_1 z_2 \langle\nabla\rangle^{\sigma} z_3 v_4 dxdt\right| \lesssim N_3^{\sigma-\frac{5}{3}s}\prod_{j=1}^{3}\big\|\langle\nabla\rangle^{s} z_j\big\|_{L_{t,x}^{\frac{9}{2-\varepsilon_1}}}\|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}.$$

Hence, by Proposition 1, the contribution to (33) in this case is at most $\lesssim R^3$ outside a set of probability

$$\leq C\exp\left(-c\frac{R^2}{T^{\frac{4-2\varepsilon_1}{9}}\|\phi\|_{H^s}^2}\right)$$

provided that

$$\frac{5}{3}s > \sigma. \tag{36}$$

Therefore, given $s > \frac{3}{5}$, we choose $\sigma = 1+$ and $\varepsilon = 0+$ for Case (2) such that (35) and (36) are satisfied.

Case (3): *vvz* case.

Without loss of generality, assume $N_1 \geq N_2$.

• **Subcase (3.a):** $N_1 \gtrsim N_3$.

By $L^3 L^{\frac{6}{1-\varepsilon_1}}L^{\frac{6}{1-\varepsilon_1}}L^{\frac{3}{1+\varepsilon_1}}$-Hölder's inequality and Lemmata 5 and 6, we have

$$\left|\int_{\mathbb{R}\times\mathbb{R}^4} \langle\nabla\rangle^{\sigma} v_1 v_2 z_3 v_4 dxdt\right| \lesssim \|v_1\|_{X^{\sigma,\frac{1}{2}+}}\|v_2\|_{X^{1+\varepsilon_1,\frac{1}{2}+}}\|z_3\|_{L_{t,x}^{\frac{6}{1-\varepsilon_1}}}\|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}.$$

Hence, by Proposition 1, the contribution to (33) in this case is at most $\lesssim R\prod_{j=1}^{2}\|v_j\|_{X^{\sigma,\frac{1}{2}+}}$ outside a set of probability

$$\leq C\exp\left(-c\frac{R^2}{T^{\frac{1-\varepsilon_1}{3}}\|\phi\|_{H^{0+}}^2}\right) \tag{37}$$

provided that $\sigma > 1+\varepsilon_1 = 1+2\varepsilon+$. Note that we have $\|\phi\|_{H^{0+}}$ instead of $\|\phi\|_{L^2}$ in (37) due to the summation over N_3.

• **Subcase (3.b):** $N_3 \sim N_4 \gg N_1 \geq N_2$.
By Lemmata 5 and 6, we have

$$\begin{aligned}\left|\int_{\mathbb{R}\times\mathbb{R}^4} v_1 v_2 \langle\nabla\rangle^{\sigma} z_3 v_4 dxdt\right| &\lesssim \|v_1\|_{L^4_{t,x}} \|\langle\nabla\rangle^{\sigma} z_3\|_{L^4_{t,x}} \|v_2 v_4\|_{L^2_{t,x}} \\ &\lesssim N_2^{\frac{3}{2}+\varepsilon_1-\sigma+} N_3^{\sigma-s} N_4^{-\frac{1}{2}+\varepsilon_1} \|v_1\|_{X^{\frac{1}{2},\frac{1}{2}+}} \|v_2\|_{X^{\sigma,\frac{1}{2}+}} \|\langle\nabla\rangle^{s} z_3\|_{L^4_{t,x}} \|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}} \\ &\lesssim N_1^{2-2\sigma+\varepsilon_1+} N_3^{\sigma-s-\frac{1}{2}+\varepsilon_1} \|v_1\|_{X^{\sigma,\frac{1}{2}+}} \|v_2\|_{X^{\sigma,\frac{1}{2}+}} \|\langle\nabla\rangle^{s} z_3\|_{L^4_{t,x}} \|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}.\end{aligned}$$

Hence, by Proposition 1, the contribution to (33) in this case is at most $\lesssim R \prod_{j=1}^{2} \|v_j\|_{X^{\sigma,\frac{1}{2}+}}$ outside a set of probability

$$\leq C \exp\left(-c \frac{R^2}{T^{\frac{1}{2}} \|\phi\|_{H^s}^2}\right)$$

provided that $2-2\sigma+\varepsilon_1 < 0$ and $s > \sigma - \frac{1}{2} + \varepsilon_1$. Given $s > \frac{1}{2}$, these conditions are satisfied by taking $\sigma = 1+$ and $\varepsilon = 0+$.

Case (4): *vzz* case.
Without loss of generality, assume $N_3 \geq N_2$.

• **Subcase (4.a):** $N_1 \gtrsim N_3$.
By $L^3 L^{\frac{6}{1-\varepsilon_1}} L^{\frac{6}{1-\varepsilon_1}} L^{\frac{3}{1+\varepsilon_1}}$-Hölder's inequality and Lemmata 5 and 6, we have

$$\left|\int_{\mathbb{R}\times\mathbb{R}^4} \langle\nabla\rangle^{\sigma} v_1 z_2 z_3 v_4 dxdt\right| \lesssim \|v_1\|_{X^{\sigma,\frac{1}{2}+}} \|z_2\|_{L^{\frac{6}{1-\varepsilon_1}}_{t,x}} \|z_3\|_{L^{\frac{6}{1-\varepsilon_1}}_{t,x}} \|v_4\|_{X^{0,\frac{1}{2}-\varepsilon}}.$$

Hence, by Proposition 1, the contribution to (33) in this case is at most $\lesssim R^2 \|v_1\|_{X^{\sigma,\frac{1}{2}+}}$ outside a set of probability

$$\leq C \exp\left(-c \frac{R^2}{T^{\frac{1-\varepsilon_1}{3}} \|\phi\|_{H^{0+}}^2}\right).$$

• **Subcase (4.b):** $N_3 \gg N_1$.
First, suppose that $N_2 \sim N_3$. Then, by Lemmata 5 and 6 (after separating the argument into two cases: $N_1 \leq N_4$ or $N_1 \geq N_4$), we have

$$\begin{aligned}\left|\int_{\mathbb{R}\times\mathbb{R}^4} v_1 z_2 \langle\nabla\rangle^{\sigma} z_3 v_4 dxdt\right| &\lesssim \|\langle\nabla\rangle^{\frac{\sigma}{2}} z_2\|_{L^4_{t,x}} \|\langle\nabla\rangle^{\frac{\sigma}{2}} z_3\|_{L^4_{t,x}} \|v_1 v_4\|_{L^2_{t,x}} \\ &\lesssim N_1^{1+2\varepsilon_1-\sigma+} N_3^{\sigma-2s} \|v_1\|_{X^{\sigma,\frac{1}{2}+}} \|\langle\nabla\rangle^{s} z_2\|_{L^4_{t,x}} \|\langle\nabla\rangle^{s} z_3\|_{L^4_{t,x}} \|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}.\end{aligned}$$

Hence, by Proposition 1, the contribution to (33) in this case is at most $\lesssim R^2 \|v_1\|_{X^{\sigma,\frac{1}{2}+}}$ outside a set of probability

$$\leq C\exp\bigg(-c\frac{R^2}{T^{\frac{1}{2}}\|\phi\|_{H^s}^2}\bigg)$$

provided that $\sigma > 1+2\varepsilon_1$ and $s > \frac{1}{2}\sigma$. Given $s > \frac{1}{2}$, these conditions are satisfied by taking $\sigma = 1+$ and $\varepsilon = 0+$.

Hence, it remains to consider the case $N_3 \sim N_4 \gg N_1, N_2$.

$\circ$ Subsubcase (4.b.i): $N_1, N_2 \ll N_3^{\frac{1}{3}}$.

By Lemmata 5 and 6, we have

$$\begin{aligned}
\bigg|\int_{\mathbb{R}\times\mathbb{R}^4} v_1 z_2 \langle\nabla\rangle^\sigma z_3 v_4 dxdt\bigg| &\lesssim \|v_1\langle\nabla\rangle^\sigma z_3\|_{L^2_{t,x}}\|z_2 v_4\|_{L^2_{t,x}}\\
&\lesssim T^{0-}N_1^{\frac{3}{2}-\sigma}N_2^{\frac{3}{2}+\varepsilon_1-s+}N_3^{\sigma-s-\frac{1}{2}}N_4^{-\frac{1}{2}+\varepsilon_1}\\
&\qquad\times \|v_1\|_{X^{\sigma,\frac{1}{2}+}}\prod_{j=2}^3\|\mathbf{P}_{N_j}\phi^\omega\|_{H^s}\|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}\\
&\lesssim T^{0-}N_3^{\frac{2}{3}\sigma-\frac{4}{3}s+\frac{4}{3}\varepsilon_1+}\|v_1\|_{X^{\sigma,\frac{1}{2}+}}\prod_{j=2}^3\|\mathbf{P}_{N_j}\phi^\omega\|_{H^s}\|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}.
\end{aligned}$$

Hence, by Lemma 3, the contribution to (33) in this case is at most $\lesssim T^{0-}R^2\|v_1\|_{X^{\sigma,\frac{1}{2}+}}$ outside a set of probability

$$\leq C\exp\bigg(-c\frac{R^2}{\|\phi\|_{H^s}^2}\bigg)$$

provided that

$$s > \frac{1}{2}\sigma + \varepsilon_1. \tag{38}$$

Given $s > \frac{1}{2}$, this condition is satisfied by taking $\sigma = 1+$ and $\varepsilon = 0+$.

$\circ$ Subsubcase (4.b.ii): $N_1 \ll N_3^{\frac{1}{3}} \lesssim N_2$.

By Lemmata 5 and 6, we have

$$\begin{aligned}
\bigg|\int_{\mathbb{R}\times\mathbb{R}^4} v_1 z_2 \langle\nabla\rangle^\sigma z_3 v_4 dxdt\bigg| &\lesssim \|z_2\|_{L^4_{t,x}}\|\langle\nabla\rangle^\sigma z_3\|_{L^4_{t,x}}\|v_1 v_4\|_{L^2_{t,x}}\\
&\lesssim N_1^{\frac{3}{2}+\varepsilon_1-\sigma+}N_2^{-s}N_3^{\sigma-s-\frac{1}{2}+\varepsilon_1}\|v_1\|_{X^{\sigma,\frac{1}{2}+}}\prod_{j=2}^3\|\langle\nabla\rangle^s z_j\|_{L^4_{t,x}}\|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}\\
&\lesssim N_3^{\frac{2}{3}\sigma-\frac{4}{3}s+\frac{4}{3}\varepsilon_1+}\|v_1\|_{X^{\sigma,\frac{1}{2}+}}\prod_{j=2}^3\|\langle\nabla\rangle^s z_j\|_{L^4_{t,x}}\|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}.
\end{aligned}$$

Hence, by Proposition 1, the contribution to (33) in this case is at most $\lesssim R^2\|v_1\|_{X^{\sigma,\frac{1}{2}+}}$ outside a set of probability

$$\leq C\exp\left(-c\frac{R^2}{T^{\frac{1}{2}}\|\phi\|_{H^s}^2}\right)$$

provided that (38) is satisfied.

∘ Subsubcase (4.b.iii): $N_2 \ll N_3^{\frac{1}{3}} \lesssim N_1$.
By Lemmata 5 and 6, we have

$$\begin{aligned}\left|\int_{\mathbb{R}\times\mathbb{R}^4} v_1 z_2 \langle\nabla\rangle^{\sigma} z_3 v_4 dxdt\right| &\lesssim \|v_1\|_{L^3_{t,x}} \|\langle\nabla\rangle^{\sigma} z_3\|_{L^6_{t,x}} \|z_2 v_4\|_{L^2_{t,x}} \\ &\lesssim T^{0-} N_1^{-\sigma} N_2^{\frac{3}{2}+\varepsilon_1-s+} N_3^{\sigma-s-\frac{1}{2}+\varepsilon_1} \\ &\qquad \times \|v_1\|_{X^{\sigma,\frac{1}{2}+}} \|\mathbf{P}_{N_2}\phi^{\omega}\|_{H^s} \|\langle\nabla\rangle^{s} z_3\|_{L^6_{t,x}} \|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}} \\ &\lesssim T^{0-} N_3^{\frac{2}{3}\sigma-\frac{4}{3}s+\frac{4}{3}\varepsilon_1+} \|v_1\|_{X^{\sigma,\frac{1}{2}+}} \|\mathbf{P}_{N_2}\phi^{\omega}\|_{H^s} \|\langle\nabla\rangle^{s} z_3\|_{L^6_{t,x}} \|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}.\end{aligned}$$

Hence, by Lemma 3 and Proposition 1, the contribution to (33) in this case is at most $\lesssim T^{0-}R^2\|v_1\|_{X^{\sigma,\frac{1}{2}+}}$ outside a set of probability

$$\leq C\exp\left(-c\frac{R^2}{\|\phi\|_{H^s}^2}\right) + C\exp\left(-c\frac{R^2}{T^{\frac{1}{3}}\|\phi\|_{H^s}^2}\right)$$

provided that (38) is satisfied.

∘ Subsubcase (4.b.iv): $N_1, N_2 \gtrsim N_3^{\frac{1}{3}}$.
By $L^3 L^{\frac{6}{1-\varepsilon_1}} L^{\frac{6}{1-\varepsilon_1}} L^{\frac{3}{1+\varepsilon_1}}$-Hölder's inequality and Lemmata 5 and 6, we have

$$\begin{aligned}\left|\int_{\mathbb{R}\times\mathbb{R}^4} v_1 z_2 \langle\nabla\rangle^{\sigma} z_3 v_4 dxdt\right| &\lesssim \|v_1\|_{L^3_{t,x}} \|z_2\|_{L^{\frac{6}{1-\varepsilon_1}}_{t,x}} \|\langle\nabla\rangle^{\sigma} z_3\|_{L^{\frac{6}{1-\varepsilon_1}}_{t,x}} \|v_4\|_{L^{\frac{3}{1+\varepsilon_1}}_{t,x}} \\ &\lesssim N_1^{-\sigma} N_2^{-s} N_3^{\sigma-s} \|v_1\|_{X^{\sigma,\frac{1}{2}+}} \prod_{j=2}^{3} \|\langle\nabla\rangle^{s} z_j\|_{L^{\frac{6}{1-\varepsilon_1}}_{t,x}} \|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}} \\ &\lesssim N_3^{\frac{2}{3}\sigma-\frac{4}{3}s} \|v_1\|_{X^{\sigma,\frac{1}{2}+}} \prod_{j=2}^{3} \|\langle\nabla\rangle^{s} z_j\|_{L^{\frac{6}{1-\varepsilon_1}}_{t,x}} \|v_4\|_{X^{0,\frac{1}{2}-2\varepsilon}}.\end{aligned}$$

Hence, by Proposition 1, the contribution to (33) in this case is at most $\lesssim R^2\|v_1\|_{X^{\sigma,\frac{1}{2}+}}$ outside a set of probability

$$\leq C\exp\left(-c\frac{R^2}{T^{\frac{1-\varepsilon_1}{3}}\|\phi\|_{H^s}^2}\right)$$

provided that $s > \frac{1}{2}\sigma$. Given $s > \frac{1}{2}$, this condition is satisfied by setting $\sigma = 1+$. This completes the proof of Proposition 3.

References

1. T. Alazard, R. Carles, Loss of regularity for supercritical nonlinear Schrödinger equations. Math. Ann. **343**(2), 397–420 (2009)
2. A. Ayache, N. Tzvetkov, L^p properties for Gaussian random series. Trans. Am. Math. Soc. **360**(8), 4425–4439 (2008)
3. Á. Bényi, T. Oh, O. Pocovnicu, On the probabilistic Cauchy theory of the cubic nonlinear Schrödinger equation on $\mathbb{R}^d$, $d \geq 3$. Trans. Am. Math. Soc. Ser. B **2**, 1–50 (2015)
4. Á. Bényi, K. Okoudjou, Local well-posedness of nonlinear dispersive equations on modulation spaces. Bull. Lond. Math. Soc. **41**(3), 549–558 (2009)
5. J. Bourgain, Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations. I. Schrödinger equations. Geom. Funct. Anal. **3**, 107–156 (1993)
6. J. Bourgain, Invariant measures for the 2D-defocusing nonlinear Schrödinger equation. Commun. Math. Phys. **176**(2), 421–445 (1996)
7. J. Bourgain, Invariant measures for the Gross-Piatevskii equation. J. Math. Pures Appl. (9) **76**(8), 649–702 (1997)
8. J. Bourgain, Refinements of Strichartz' inequality and applications to 2D-NLS with critical nonlinearity. Int. Math. Res. Not. **1998**(5), 253–283 (1998)
9. J. Bourgain, A. Bulut, Almost sure global well-posedness for the radial nonlinear Schrödinger equation on the unit ball II: the 3D case. J. Eur. Math. Soc. **16**(6), 1289–1325 (2014)
10. J. Bourgain, A. Bulut, Invariant Gibbs measure evolution for the radial nonlinear wave equation on the 3D ball. J. Funct. Anal. **266**(4), 2319–2340 (2014)
11. N. Burq, P. Gérard, N. Tzvetkov, Multilinear eigenfunction estimates and global existence for the three dimensional nonlinear Schrödinger equations. Ann. Sci. École Norm. Sup. (4) **38**(2), 255–301 (2005)
12. N. Burq, L. Thomann, N. Tzvetkov, Long time dynamics for the one dimensional nonlinear Schrödinger equation. Ann. Inst. Fourier (Grenoble) **63**(6), 2137–2198 (2013)
13. N. Burq, L. Thomann, N. Tzvetkov, Global infinite energy solutions for the cubic wave equation. Bull. Soc. Math. France. **143**(2) 301–313 (2015)
14. N. Burq, N. Tzvetkov, Random data Cauchy theory for supercritical wave equations. I. Local theory. Invent. Math. **173**(3), 449–475 (2008)
15. N. Burq, N. Tzvetkov, Probabilistic well-posedness for the cubic wave equation. J. Eur. Math. Soc. **16**(1), 1–30 (2014)
16. R. Carles, Geometric optics and instability for semi-classical Schrödinger equations. Arch. Ration. Mech. Anal. **183**(3), 525–553 (2007)
17. T. Cazenave, F. Weissler, Some remarks on the nonlinear Schrödinger equation in the critical case, in *Nonlinear Semigroups, Partial Differential Equations and Attractors (Washington, DC, 1987)*. Lecture Notes in Mathematics, vol. 1394 (Springer, Berlin, 1989), pp. 18–29
18. M. Christ, J. Colliander, T. Tao, Asymptotics, frequency modulation, and low-regularity ill-posedness of canonical defocusing equations. Am. J. Math. **125**(6), 1235–1293 (2003)
19. J. Colliander, T. Oh, Almost sure well-posedness of the cubic nonlinear Schrödinger equation below $L^2(\mathbb{T})$. Duke Math. J. **161**(3), 367–414 (2012)
20. Y. Deng, Two-dimensional nonlinear Schrödinger equation with random radial data. Anal. PDE **5**(5), 913–960 (2012)
21. A.-S. de Suzzoni, Invariant measure for the cubic wave equation on the unit ball of $\mathbb{R}^3$. Dyn. Partial Differ. Equ. **8**(2), 127–147 (2011)
22. A.-S. de Suzzoni, Consequences of the choice of a particular basis of $L^2(S^3)$ for the cubic wave equation on the sphere and the Euclidian space. Commun. Pure Appl. Anal. **13**(3), 991–1015 (2014)

23. H. Feichtinger, Modulation spaces of locally compact Abelian groups, Technical report, University of Vienna (1983), in *Proc. Internat. Conf. on Wavelets and Applications (Chennai, 2002)*, ed. by R. Radha, M. Krishna, S. Thangavelu (New Delhi Allied Publishers, New Delhi, 2003), pp. 1–56.
24. H. Feichtinger, K. Gröchenig, Banach spaces related to integrable group representations and their atomic decompositions, I. J. Funct. Anal. **86**, 307–340 (1989)
25. H. Feichtinger, K. Gröchenig, Banach spaces related to integrable group representations and their atomic decompositions, II. Monatsh. Math. **108**, 129–148 (1989)
26. J. Ginibre, G. Velo, Smoothing properties and retarded estimates for some dispersive evolution equations, Commun. Math. Phys. **144**(1), 163–188 (1992)
27. K. Gröchenig, *Foundations of Time-Frequency Analysis* (Birkhäuser, Boston, 2001), xvi+359 pp
28. M. Hadac, S. Herr, H. Koch, Well-posedness and scattering for the KP-II equation in a critical space. Ann. Inst. H. Poincaré Anal. Non Linéaire **26**(3), 917–941 (2009); Erratum to "Well-posedness and scattering for the KP-II equation in a critical space". Ann. Inst. H. Poincaré Anal. Non Linéaire **27**(3), 971–972 (2010)
29. S. Herr, D. Tataru, N. Tzvetkov, Global well-posedness of the energy critical Nonlinear Schrödinger equation with small initial data in $H^1(\mathbb{T}^3)$. Duke Math. J. **159**, 329–349 (2011)
30. J.P. Kahane, *Some Random Series of Functions*. Cambridge Studies in Advanced Mathematics, vol. 5, 2nd edn. (Cambridge University Press, Cambridge, 1985), xiv+305 pp
31. M. Keel, T. Tao, Endpoint Strichartz estimates. Am. J. Math. **120**(5), 955–980 (1998)
32. C. Kenig, F. Merle, Global well-posedness, scattering and blow-up for the energy-critical, focusing, non-linear Schrödinger equation in the radial case. Invent. Math. **166**(3), 645–675 (2006)
33. C. Kenig, G. Ponce, L. Vega, The Cauchy problem for the Korteweg-de Vries equation in Sobolev spaces of negative indices. Duke Math. J. **71**(1), 1–21 (1993)
34. M. Kobayashi, M. Sugimoto, The inclusion relation between Sobolev and modulation spaces. J. Funct. Anal. **260**(11), 3189–3208 (2011)
35. H. Koch, D. Tataru, A priori bounds for the 1D cubic NLS in negative Sobolev spaces. Int. Math. Res. Not. IMRN **2007**(16), Art. ID rnm053, 36 pp (2007)
36. J. Lührmann, D. Mendelson, Random data Cauchy theory for nonlinear wave equations of power-type on $\mathbb{R}^3$. Commun. Partial Differ. Equs. **39**(12), 2262–2283 (2014)
37. A. Nahmod, G. Staffilani, Almost sure well-posedness for the periodic 3D quintic NLS below the energy space. J. Eur. Math. Soc. (2012, to appear)
38. A. Nahmod, N. Pavlović, G. Staffilani, Almost sure existence of global weak solutions for supercritical Navier-Stokes equations. SIAM J. Math. Anal. **45**(6), 3431–3452 (2013)
39. K. Okoudjou, Embeddings of some classical Banach spaces into modulation spaces. Proc. Am. Math. Soc. **132**, 1639–1647 (2004)
40. T. Ozawa, Y. Tsutsumi, Space-time estimates for null gauge forms and nonlinear Schrödinger equations. Differ. Integr. Equ. **11**(2), 201–222 (1998)
41. R.E.A.C. Paley, A. Zygmund, On some series of functions (1), (2), (3), Proc. Camb. Philos. Soc. **26**, 337–357, 458–474 (1930); **28**, 190–205 (1932)
42. O. Pocovnicu, Almost sure global well-posedness for the energy-critical defocusing cubic nonlinear wave equation on $\mathbb{R}^d, d = 4$ and 5. J. Eur. Math. Soc. (2014, to appear)
43. A. Poiret, D. Robert, L. Thomann, Probabilistic global well-posedness for the supercritical nonlinear harmonic oscillator. Anal. PDE **7**(4), 997–1026 (2014)
44. G. Richards, Invariance of the Gibbs measure for the periodic quartic gKdV. Ann. Inst. H. Poincaré Anal. Non Linćaire (2012, to appear)
45. E. Ryckman, M. Vişan, Global well-posedness and scattering for the defocusing energy-critical nonlinear Schrödinger equation in $\mathbb{R}^{1+4}$. Am. J. Math. **129**(1), 1–60 (2007)
46. R.S. Strichartz, Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations. Duke Math. J. **44**(3), 705–714 (1977)
47. M. Sugimoto, N. Tomita, The dilation property of modulation spaces and their inclusion relation with Besov spaces. J. Funct. Anal. **248**(1), 79–106 (2007)

48. T. Tao, *Nonlinear Dispersive Equations. Local and Global Analysis*. CBMS Regional Conference Series in Mathematics, vol. 106. Published for the Conference Board of the Mathematical Sciences, Washington, DC (American Mathematical Society, Providence, 2006), xvi+373 pp
49. L. Thomann, Random data Cauchy problem for supercritical Schrödinger equations. Ann. Inst. H. Poincaré Anal. Non Linéaire **26**(6), 2385–2402 (2009)
50. J. Toft, Convolution and embeddings for weighted modulation spaces, in *Advances in Pseudo-Differential Operators*. Oper. Theory Adv. Appl., vol. 155 (Birkhäuser, Basel, 2004), pp. 165–186
51. M. Vişan, Global well-posedness and scattering for the defocusing cubic nonlinear Schrödinger equation in four dimensions. Int. Math. Res. Not. IMRN **2012**(5), 1037–1067 (2012)
52. B.X. Wang, L. Han, C. Huang, Global well-posedness and scatering for the derivative nonlinear Schrödinger equation with small rough data. Ann. Inst. H. Poincaré Anal. Non Linéaire **26**, 2253–2281 (2009)
53. N. Wiener, Tauberian theorems. Ann. Math. (2) **33**(1), 1–100 (1932)
54. K. Yajima, Existence of solutions for Schrödinger evolution equations. Commun. Math. Phys. **110**(3), 415–426 (1987)

Bridging erasures and the infrastructure of frames

David Larson and Sam Scholze

Abstract The purpose of this chapter is to give a detailed tutorial-style exposition of a method we have recently discovered for complete recovery from frame and sampling erasures by inverting a matrix of dimension the cardinality of the erasure set. This is usually simpler and more efficient than inverting the frame operator of the remaining coefficients because the erasure set is usually much smaller than the dimension of the space. The method is not hard and can be easily implemented on a computer, even for infinite frames and sampling schemes as long as one is dealing with only a finite set of erasures. The set of erasures that can be handled in this way can be very large. We call the method *nilpotent bridging*. We introduced this method in a recent research article, along with some new classification results and methods of measuring redundancy that are based on the matricial infrastructure of a dual frame pair: the set of all submatrices of the cross-Gramian of the pair. The first author gave a talk on this material in a conference at Vanderbilt University in May 2014 entitled "Building bridges (reconstruction from frame and sampling omissions)", and this chapter is based on the notes from that talk. In the process of developing the mathematics underlying the bridging technique, we discovered a second method for recovery from erasures in finitely many steps, and this chapter will also discuss this alternate method.

Key words: Finite frame, Omission, Erasure, Reconstruction, Bridging

D. Larson (✉) • S. Scholze
Department of Mathematics, Texas A&M University, College Station, TX, USA
e-mail: larson@math.tamu.edu; scholzes@math.tamu.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_2

Introduction

In a recent research article [25] we gave some new methods for perfect reconstruction from frame and sampling erasures in a small number of steps. This method is efficient in the sense that it only requires a matrix inversion of size $L \times L$, where L is the cardinality of the erasure set. The purpose of this chapter is to give a detailed tutorial-style exposition of this. By bridging an erasure set we mean replacing the erased Fourier coefficients of a function with respect to a frame by appropriate linear combinations of the non-erased coefficients. We prove that if a minimal redundancy condition is satisfied bridging can always be done to make the reduced error operator nilpotent of index 2 using a bridge set of indices no larger than the cardinality of the erasure set. This results in perfect reconstruction of the erased coefficients in one final matricial step. We also obtained a new direct formula for the inverse of an invertible partial reconstruction operator. This leads to a second method of perfect reconstruction from frame and sampling erasures in a small number of steps. This gives an alternative to the bridging method for many (but not all) cases. Our second method also only requires a matrix inversion of the same size as the cardinality of the erasure set. The methods we use employ matrix techniques only of the order of the cardinality of the erasure set and are applicable to rather large finite erasure sets for infinite frames and sampling schemes as well as for finite frame theory. Some new classification theorems for frames are obtained and some new methods of measuring redundancy are introduced based on our bridging theory.

When the first author served in the US Air Force a number of years ago, his unit worked on projects which essentially amounted to processing numerical data collected by instruments onboard aircraft for the purpose of performing mathematical and statistical computations relevant to photomapping large areas of rough terrain. In a given project, it was not uncommon that a number of points of data would be erased, omitted, or otherwise corrupted, and a *work-around-bad-points* procedure became routine and built into the data processing protocol. The data would be collected on a timeline at regular intervals, or at lattice points of a grid, so a common plan would be to simply replace a *bad* point with the numerical averaging of neighboring *good* points, according to an adopted protocol scheme that would ensure that the same data would be processed in exactly the same way by different units employing the same protocol. We will use the term *bridging* for this type of procedure. At that time, choosing an averaging protocol seemed to belong to the *art* of the subject rather than to the *science* because of the difficulty of proving mathematically that one choice of a protocol scheme was any better in some strong sense than another. The purpose of this chapter and the recent article [25] on which it is based is to show that a special bridging method we call *nilpotent bridging* can be used to effectively and efficiently achieve perfect reconstruction from erasures in cases where frame and sampling techniques are employed.

Frame and sampling techniques are often used to analyze and digitize signals and images when they are represented as vectors or functions in a Hilbert space.

There is a large literature on the pure and applied mathematics of this subject (cf. [3, 10, 12, 14, 15, 19]). A number of articles have been written on problems and methods for reconstruction from erasures (cf. [5, 8, 9, 17, 22]).

Let $\{f_j\}$ be a Parseval frame for a Hilbert space $\mathscr{H}$, or more generally let $\{f_j, g_j\}$ be a dual frame pair. (See definitions below.) Let f be a vector in $\mathscr{H}$, and let Λ be a finite subset of the index set. If f is analyzed with $\{g_j\}$ and if the frame coefficients $\langle f, g_j \rangle$ for Λ are erased, then by bridging the erasures we mean replacing the erased coefficients ($\langle f, g_j \rangle$ for $j \in \Lambda$) with appropriate linear combinations of the non-erased coefficients ($\langle f, g_j \rangle$ for $j \in \Lambda^c$). As mentioned above, we showed that bridging can always be done to make the resulting reduced error operator nilpotent of index 2 using a bridge set no larger than the cardinality of the erasure set. From this, an algorithm for perfect reconstruction from erasures follows in one final simple step. The resulting algorithms use only finite matrix methods of order the cardinality of the erasure set. Frames can be infinite, such as Gabor and wavelet frames. The only delimiter in a computational sense seems to be the size of the erasure set, which we take to be finite in this chapter. This method adapts equally well to sampling theory, such as Shannon-Whittaker sampling theory [4, 18, 29]. In the process of developing the mathematics underlying the bridging technique, we discovered a second, more direct method for recovery from erasures in finitely many steps, and we will also discuss this direct method. In [25] we also proved some results about dual frame pairs which are strongly robust with respect to bridging. These are the dual frame pairs we defined to have *full skew-spark*. We describe our results and include the proof of a new result for Parseval frames (Theorem 7) that was not included in [25]. We wanted to include it in [25] but the proof simply eluded us at the time of writing it. We have since worked out a proof and are pleased to include it in this chapter.

In government, industrial, and academic organizations, the word "infrastructure" basically means the set of ways in which separate components of an organization or a system relate to each other and reinforce each other. In architecture the term refers to the interrelationship between the different components of a building, the system of beams, weight-bearing walls, and the foundation, that keeps it standing. An organization which is simply a union of components that have no strong relationship with each other is not very "robust". A building with this kind of problem will collapse. If a component can be "down" for a time with the slack somehow picked up by the other components, this means the system has a degree of robustness.

For frames, loosely put, infrastructure should translate to *redundancy*. The concept of redundancy has been widely studied. In particular, the spark of a frame (cf. [1, 13]) is one measure of redundancy. If $F := \{f_j\}_{j\in\mathbb{J}}$, or more generally if $(F,G) := \{f_j, g_j\}_{j\in\mathbb{J}}$, is a dual frame pair (see definitions below), then for each pair of subsets Λ and Ω of the index set $\mathbb{J}$ we can consider the *cross-Gramian* matrix $(\langle f_k, g_j \rangle)_{k\in\Lambda, j\in\Omega}$ and examine its properties. These are just the submatrices of the full cross-Gramian matrix $Gr(F,G) := (\langle f_k, g_j \rangle)_{k,j\in\mathbb{J}}$, which is an idempotent (not-necessarily-self-adjoint projection) since (F,G) is a dual frame pair. We define the *(matricial) infrastructure* of a dual frame pair to be the set of all submatrices of $Gr(F,G)$. Each such submatrix gives a concrete matricial link between

subcollections $F_\Lambda := \{f_k\}_{k\in\Lambda}$ and $G_\Omega := \{g_j\}_{j\in\Omega}$. This is the coefficient matrix (the *bridge* matrix) for a special system of linear equations (6). When this system has a solution then Ω strongly reinforces Λ in the sense that if the frame coefficients of a signal over Λ are erased, the coefficients over Ω can be used to efficiently recover them. In the case where $|\Lambda| = |\Omega|$ so the bridge matrix is square, invertibility of the bridge matrix translates to strong infrastructure and leads to our concept of *skew-spark* as a measure of redundancy.

As with the article [25], we would like to thank Deguang Han for useful discussions on this work, and for piquing our interest in frame erasure problems in the recent interesting article [27]. We thank Stephen Rowe for useful Matlab and programming advice in the experimental phases of this work. Many of our mathematical results were obtained after numerous computer experiments, and we take the opportunity in this chapter to provide some details on these. Lastly, we thank the referees of [25] for providing several helpful suggestions for that paper that we have also incorporated in this chapter.

Preliminaries

A *frame* F for a Hilbert space $\mathscr{H}$ is a sequence of vectors $\{f_j\} \subset \mathscr{H}$ indexed by a finite or countable index set $\mathbb{J}$ for which there exist constants $0 < A \leq B < \infty$ such that, for every $f \in \mathscr{H}$,

$$A\|f\|^2 \leq \sum_{j\in\mathbb{J}} |\langle f, f_j\rangle|^2 \leq B\|f\|^2. \tag{1}$$

The constants A and B are known as the lower and upper *frame bounds*, respectively. The supremum over all lower frame bounds is called the *optimal lower frame bound*, and the infimum over all upper frame bounds is called the *optimal upper frame bound*. If the frame $\{f_j\}_{j\in\mathbb{J}}$ has optimal frame bounds A_0 and B_0 and $A_0 = B_0$, we call $\{f_j\}_{j\in\mathbb{J}}$ a *tight frame*. If $A_0 = B_0 = 1$, $\{f_j\}_{j\in\mathbb{J}}$ is called a *Parseval frame*. If we only require that a sequence $\{f_j\}$ satisfies the upper bound condition in (1), then $\{f_j\}$ is called a *Bessel sequence*. A frame which is a Schauder basis is called a Riesz basis. Orthonormal bases are special cases of Parseval frames. A Parseval frame $\{f_j\}$ for a Hilbert space $\mathscr{H}$ is an orthonormal basis if and only if each f_j is a unit vector.

The *analysis operator* Θ for a Bessel sequence $\{f_j\}$ is a bounded linear operator from $\mathscr{H}$ to $\ell^2(\mathbb{J})$ defined by

$$\Theta f = \sum_{j\in\mathbb{J}} \langle f, f_j\rangle e_j, \tag{2}$$

where $\{e_j\}$ is the standard orthonormal basis for $\ell^2(\mathbb{J})$. It is easily verified that

$$\Theta^* e_j = f_j, \ \ \forall j \in \mathbb{J}.$$

The Hilbert space adjoint Θ^* is called the *synthesis operator* for $\{f_j\}$. The positive operator $S := \Theta^*\Theta : \mathscr{H} \to \mathscr{H}$ is called the *frame operator* or sometimes the *Bessel operator* if the Bessel sequence is not a frame, and we have

$$Sf = \sum_{j\in\mathbb{J}} \langle f, f_j\rangle f_j, \quad \forall f \in \mathscr{H}. \tag{3}$$

In operator theory, rank-one operators are often represented as tensor products of vectors, and we will find it convenient to use this standard representation throughout this chapter. The notation $x \otimes y$ will denote the operator that maps a vector z to the vector $\langle z, y\rangle x$. So, $(x \otimes y)(z) = \langle z, y\rangle x$. Thus we can write (3) as

$$S = \sum_{j\in\mathbb{J}} f_j \otimes f_j.$$

Similarly, $\Theta = \sum_{j\in\mathbb{J}} e_j \otimes f_j$ and $\Theta^* = \sum_{j\in\mathbb{J}} f_j \otimes e_j$. The operator $\Theta\Theta^* : \ell^2(\mathbb{J}) \to \ell^2(\mathbb{J})$ is called the Gramian operator (or Gram Matrix) and is denoted $Gr(F)$. Then,

$$Gr(F) = \sum_{j,k\in\mathbb{J}} \left\langle f_k, f_j\right\rangle e_j \otimes e_k = \left(\left\langle f_k, f_j\right\rangle\right)_{j,k}.$$

An alternative to rank-one notation is *outer product notation*. Using this, we can write $S = \sum f_j f_j^*$, $\Theta = \sum e_j f_j^*$, $\Theta^* = \sum f_j e_j^*$, and $Gr(F) = \sum_{j,k} \left\langle f_k, f_j\right\rangle e_j e_k^*$. Throughout this chapter we will stick with the rank-one notation.

If $\{f_j\}_{j\in\mathbb{J}}$ forms a frame, from (3) we obtain the *reconstruction formula (or frame decomposition)*

$$f = \sum_{j\in\mathbb{J}} \langle f, S^{-1}f_j\rangle f_j = \sum_{j\in\mathbb{J}} \langle f, f_j\rangle S^{-1}f_j \ \ \forall f \in \mathscr{H},$$

where the convergence is in the norm of $\mathscr{H}$. The frame $\{S^{-1}f_j\}$ is called the *canonical or standard dual* of $\{f_j\}$.

In the case that $\{f_j\}$ is a Parseval frame for $\mathscr{H}$, we have $S = I$ and hence $f = \sum_{j\in\mathbb{J}} \langle f, f_j\rangle f_j, \ \ \forall f \in \mathscr{H}$. More generally, if a Bessel sequence $\{g_j\}$ satisfies a reconstruction formula

$$f = \sum_{j\in\mathbb{J}} \langle f, g_j\rangle f_j \ \ \forall f \in \mathscr{H},$$

then $\{g_j\}$ is called an *alternate dual* of $\{f_j\}$. (Hence $\{g_j\}$ is also necessarily a frame.) The canonical and alternate duals are often simply referred to as *duals*, and $(F, G) := \{f_j, g_j\}_{j\in\mathbb{J}}$ is called a *dual frame pair.* The second frame G in the ordered pair will be called the *analysis frame* and the first frame F will be called the *synthesis frame*.

It will be convenient to define a *frame pair* which is not necessarily a dual frame pair to be simply a pair of frames $F = \{f_j\}$ and $G = \{g_j\}$ indexed by the same set $\mathbb{J}$ for which the operator $\widetilde{S}f = \sum \langle f, g_j\rangle f_j$ is invertible. We will call the operator $\widetilde{S}$ the

cross-frame operator for F and G, and the operator $Gr(F,G) = \sum \langle f_k, g_j \rangle e_j \otimes e_k$ the *cross-Gramian*. If $\{f_1, \cdots, f_L\}$ and $\{g_1, \cdots, g_L\}$ are finite sets of vectors, we will write $Gr(\{f_1, \cdots, f_L\}, \{g_1, \cdots, g_L\})$ for the cross-Gram matrix,

$$Gr(\{f_1, \cdots, f_L\}, \{g_1, \cdots, g_L\}) = \left(\langle f_k, g_j \rangle\right)_{j,k} := \begin{pmatrix} \langle f_1, g_1 \rangle & \langle f_2, g_1 \rangle & \cdots & \langle f_L, g_1 \rangle \\ \langle f_1, g_2 \rangle & \langle f_2, g_2 \rangle & \cdots & \langle f_L, g_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle f_1, g_L \rangle & \langle f_2, g_L \rangle & \cdots & \langle f_L, g_L \rangle \end{pmatrix}. \tag{4}$$

We will use this notation mainly when $\{f_j\}$ and $\{g_j\}$ are frames or subsets of frames. It is useful to note that if $\{f_1, \cdots, f_L\}$ and $\{g_1, \cdots, g_L\}$ are two bases for the same Hilbert space $\mathscr{H}$, then $Gr(\{f_1, \cdots, f_L\}, \{g_1, \cdots, g_L\})$ is invertible. Indeed, if $\{e_j\}$ is an orthonormal basis for $\mathscr{H}$, and A and B are invertible matrices with $Ae_j = f_j$ and $Be_j = g_j$, then $Gr(\{f_1, \cdots, f_L\}, \{g_1, \cdots, g_L\})$ is just the matrix of B^*A with respect to $\{e_j\}$.

Throughout this chapter, we will include many blocks of Matlab code that implement our algorithms. We utilize a few of Matlab's built in shortcuts to help with the creation of vectors. The *colon notation* $a : k : b$ creates the vector $(a, a+k, a+2k, \cdots, b)$. The shortcuts $zeros(n,1)$ and $rand(n,1)$ create the $n \times 1$ column vectors consisting of zeros, and random numbers from the interval $(0,1)$, respectively. For $n \times 1$ or $1 \times n$ column vectors, v and w, the following table describes some of the built in Matlab functions we will use.

$norm(v)$	Computes the ℓ^2 norm of v.
$dot(v,w)$	Computes the dot product of v and w.
$min(v)$	Finds the smallest component of the vector v.
$max(v)$	Finds the largest component of the vector v.
$abs(v)$	Returns a vector containing the absolute values of the components of v.
$setdiff(v,w)$	Computes the set difference of v and w.

To create an $m \times n$ matrix consisting of either zeros or randomly selected numbers from the interval $(0,1)$ we use the shortcuts $zeros(m,n)$ and $rand(m,n)$, respectively. To create the $n \times n$ identity matrix, we use the shortcut $eye(n)$. The following table lists a few of the basic Matlab commands for $m \times n$ matrices A and B. We assume that the sizes of the matrices are valid for each of the following commands.

$norm(A)$	Computes the operator norm of A.
$A \backslash B$	Computes $A^{-1}B$.
$size(A)$	Returns the vector (m,n) for an $m \times n$ matrix A.
$rank(A)$	Computes the rank of the matrix A.
$eigs(A)$	Computes a vector that contains the largest eigenvalues in magnitude.

To access the submatrix of a matrix A that contains the rows indexed by the vector v and the columns indexed by the vector w, we use the code $A(v,w)$. If we would like either all of the rows or all of the columns, we use $A(:,w)$ or $A(v,:)$, respectively.

The function *sinc* we use is a user-defined function file that computes $\mathrm{sinc}(x)$ for a number x or applies the sinc function to every component of a vector. We use the convention $\mathrm{sinc}(x) = \frac{\sin x}{x}$ for $x \neq 0$ and $\mathrm{sinc}(x) = 1$ for $x = 0$.

Lastly, in Matlab *for* loops are indexed by vectors. For example to loop over the indices 1,5,7,8, and 25, we would use the code $for(j = [1,5,7,8,25])$. For more on Matlab, we refer the reader to [16].

Let $F = \{f_j\}_{j\in\mathbb{J}}$ be a frame. An *erasure set* for F is defined to be simply a finite subset of $\mathbb{J}$. We say that an erasure set Λ for a frame F satisfies the *minimal redundancy condition* if $\overline{\mathrm{span}}\{f_j : j \notin \Lambda\} = \mathscr{H}$. This is equivalent to the condition that the reduced sequence $\{f_j : j \in \Lambda^c\}$ is still a frame (cf. Theorem 5.4.7 in [14]).

Most of our work will concern dual frame pairs. If $(F,G) = \{f_j, g_j\}_{j\in\mathbb{J}}$ is a dual frame pair, then as we did above for single frames, we define an *erasure set* for (F,G) to be simply a finite subset of $\mathbb{J}$. We say that Λ satisfies the *minimal redundancy condition for the dual frame pair* (F,G) if $\overline{\mathrm{span}}\{g_j : j \notin \Lambda\} = \mathscr{H}$ (or equivalently if $\{g_j : j \in \Lambda^c\}$ is still a frame). We point out that the minimal redundancy condition for a dual frame pair (F,G) as we have defined it is a condition on only the analysis frame G. The redundancy properties of the synthesis frame F play a role here only in that it is required to be a dual frame to G. For the special case where G is the standard dual of F, F and G have the same linear redundancy properties. The Parseval frame case, where $F = G$, is a special case of this.

For a dual pair (F,G), if Λ satisfies the minimal redundancy condition, then since $\{g_j : j \in \Lambda^c\}$ is a frame for $\mathscr{H}$ it has *some* frame dual (in general many duals) that will yield the reconstruction of f from the coefficients over Λ^c, so there is enough information in $\{\langle f, g_j\rangle : j \in \Lambda^c\}$ to reconstruct f. On the other hand if Λ fails the minimal redundancy condition then some nonzero vector f will be orthogonal to g_j for all $j \in \Lambda^c$, and hence no reconstruction of f is possible using only the coefficients $\{\langle f, g_j\rangle : j \in \Lambda^c\}$. This justifies the use of the word "minimal" in the description of the minimal redundancy condition.

Let F be a *Parseval frame*. If Λ is an erasure set which satisfies the minimal redundancy condition, then the *partial reconstruction operator* $R_\Lambda := \sum_{j\in\Lambda^c} f_j \otimes f_j$ is the *frame operator* for the reduced frame $\{f_j\}_{j\in\Lambda^c}$, hence it is invertible. Let $f_R = R_\Lambda f$ be the partial reconstruction of the vector f. It is possible to reconstruct f from the "good" Fourier coefficients by $f = R_\Lambda^{-1} f_R$. However, given a *dual frame pair* (F,G) indexed by $\mathbb{J} = \{1,2,\cdots,N\}$ and an erasure set Λ satisfying the minimal redundancy condition, the partial reconstruction operator $R_\Lambda := \sum_{j\in\Lambda^c} f_j \otimes g_j$ need not be invertible. In fact invertibility of R_Λ can fail even if both F and G separately satisfy the minimal redundancy condition for Λ. The following simple example shows that this can happen and R_Λ can even be the zero operator.

Example 1. Let $\{f_j, g_j\}_{j=1}^N$ be a dual frame pair. Suppose

$$\begin{aligned} f_j &= f_{j+N} = f_{j+2N} \quad 1 \le j \le N \\ g_{j+N} &= -g_j \quad 1 \le j \le N \\ g_{j+2N} &= g_j \quad 1 \le j \le N. \end{aligned}$$

Then, it is easily verified that $\{f_j, g_j\}_{j=1}^{3N}$ is a dual frame pair, and $\Lambda = \{1, 2, \cdots, N\}$ satisfies the minimal redundancy condition with respect to both frames. However,

$$R_\Lambda = \sum_{j=N+1}^{3N} f_j \otimes g_j = \sum_{j=N+1}^{2N} f_j \otimes g_j + \sum_{j=2N+1}^{3N} f_j \otimes g_j = \sum_{j=1}^{N} f_j \otimes (-g_j) + \sum_{j=1}^{N} f_j \otimes g_j = 0.$$

□

Even when R_Λ is invertible, computing R_Λ^{-1} can be a computationally costly process. The error for the partial reconstruction is $f_E = f - f_R$, and the associated error operator for the partial reconstruction is $E_\Lambda = I - R_\Lambda = \sum_{j\in\Lambda} f_j \otimes g_j$. Then, $R_\Lambda^{-1} = (I - E_\Lambda)^{-1}$, and if the norm, or more generally the spectral radius of E_Λ, is strictly less than 1 then R_Λ^{-1} can be computed using the Neumann series expansion $R_\Lambda^{-1} = I + E + E^2 + \cdots = \sum_{j=0}^{\infty} E^j$.

For certain very special cases (F, G), with corresponding erasure set Λ, the error operator E_Λ will be nilpotent of index 2 (i.e., $E_\Lambda^2 = 0$) such as the example below. In this case, $R_\Lambda^{-1} = I + E_\Lambda$, and moreover, the error f_E of f can be obtained by applying the error operator to the partial reconstruction f_R instead of f. (That is, $f_E = E_\Lambda f = E_\Lambda(f_E + f_R) = E_\Lambda^2 f + E_\Lambda f_R = E_\Lambda f_R$.)

Example 2. Let $\{e_1, e_2\}$ be the standard orthonormal basis for $\mathbb{C}^2$. Let $F = \{e_1, -e_1, e_1, e_2\}$ and $G = \{e_2, e_2, e_1, e_2\}$. Let $\Lambda = \{1\}$. Then, $E_\Lambda = e_1 \otimes e_2$. So,

$$E_\Lambda^2 = (e_1 \otimes e_2)(e_1 \otimes e_2) = \langle e_1, e_2 \rangle (e_1 \otimes e_2) = 0.$$

Therefore, $R_\Lambda^{-1} = I + E_\Lambda$. □

Bridging

Let (F, G) be a dual frame pair for a Hilbert space $\mathscr{H}$. Let Λ be an erasure set, $\Omega \subset \Lambda^c$, and $f \in \mathscr{H}$. The main idea behind bridging is to replace each erased coefficient $\langle f, g_j \rangle$ for $j \in \Lambda$ with $\left\langle f, g'_j \right\rangle$ for $g'_j \in \operatorname{span}\{g_k : k \in \Omega\}$. (That is, $\left\langle f, g'_j \right\rangle$ is a weighted average of the $\langle f, g_k \rangle$ for $k \in \Omega$.) The point of this preconditioning is to make the inverse problem of recovering f more efficient than simply inverting R_Λ. There are many ways to select the g'_j, and we will summarize some of these strategies after we introduce some useful terminologies.

The partial reconstruction with bridging is

$$\widetilde{f} = f_R + f_B,$$

where $f_B = \sum_{j\in\Lambda} \left\langle f, g'_j \right\rangle f_j$. We call f_B the *bridging supplement* and $B_\Lambda := \sum_{j\in\Lambda} f_j \otimes g'_j$ the *bridging supplement operator*. The reduced error is $f_{\widetilde{E}} := f - \widetilde{f}$, and the associated *reduced error operator* is $\widetilde{E}_\Lambda = I - R_\Lambda - B_\Lambda$. We have

$$\widetilde{E}_\Lambda f = f_{\widetilde{E}} = \sum_{j\in\Lambda} \langle f, g_j - g'_j \rangle f_j.$$

There are two natural types of bridging methods:

- *metric bridging*. In this case we bridge by attempting to optimize the norm (or the Hilbert-Schmidt norm) of the reduced error operator. There has been some work relating to this in the literature (cf. [7]). Perfect reconstruction is not the goal of metric bridging.
- *spectral bridging*. In this case we bridge by attempting to optimize the spectral properties of the reduced error operator.

  (1) *Nilpotent bridging* is the special case of spectral bridging where we bridge the erased coefficients to make the reduced error operator nilpotent. In this case the Neumann series for the inverse of the reduced partial reconstruction operator is a finite sum, so perfect reconstruction can be obtained in finitely many steps. By k-nilpotent bridging we mean that we bridge to make the reduced error operator nilpotent of index k. Then, 2-nilpotent bridging is the best possible case from the point of view of simplicity of computations.
  (2) *Cardinal spectral reduction* is another case of spectral bridging. The goal of this method is to decrease the number of nonzero eigenvalues of the reduced error operator. When we choose a bridge set with smaller than optimal cardinality, cardinal spectral reduction occurs naturally.
  (3) *Radial spectral reduction* is the last case of spectral bridging. Here the goal is to reduce the spectral radius of the reduced error operator (cf. [27]).

In [25] we showed that 2-nilpotent bridging is always possible with $|\Omega| \leq |\Lambda|$ whenever Λ satisfies the minimal redundancy condition. Moreover, explicit formulas for the bridging and for the resultant perfect reconstruction were given. Using a smaller than optimal bridge set yields cardinal spectral reduction in a systematic way (see Theorem 4). At the outset of the authors' investigation of frame erasures, radial spectral reduction was the goal, and in this respect this project was inspired by the beautiful paper [27]. We ran into some difficulty here, but we got around the problem of reducing the spectral radius by figuring out that by using an optimal bridge set we can actually make the spectral radius zero. Currently the authors' only way of systematically reducing the spectral radius of the error operator is to actually make it zero. It is possible that a method could be devised to systematically shrink the spectral radius by choosing an appropriate much smaller bridge set. Our present results yield that such underbridging yields cardinal spectral reduction, but we do not know how to do it to actually shrink the spectral radius. This is another direction to pursue.

2-Nilpotent Bridging

Let (F,G) be a dual frame pair for a Hilbert space $\mathscr{H}$ with erasure set Λ and corresponding bridge set Ω. Recall that our reduced error operator is

$$\widetilde{E}_\Lambda = \sum_{j\in\Lambda} f_j \otimes (g_j - g_j')$$

for some choice of $g_j' \in \operatorname{span}\{g_j : j \in \Omega\}$. We wish to make the reduced error operator nilpotent of index 2. To do so, it is easily verified that

$$f_j \perp (g_k - g_k') \ \ \forall j,k \in \Lambda \tag{5}$$

forces $E_\Lambda^2 = 0$. So, writing

$$g_k' = \sum_{\ell\in\Omega} c_\ell^{(k)} g_\ell$$

we seek coefficients $c_\ell^{(k)}$ so that (5) is satisfied. We have

$$0 = \left\langle f_j, g_k - \sum_{\ell\in\Omega} c_\ell^{(k)} g_\ell \right\rangle = \langle f_j, g_k\rangle - \sum_{\ell\in\Omega} \overline{c_\ell^{(k)}} \langle f_j, g_\ell\rangle .$$

For each $k \in \Lambda$, we obtain a system of $|\Lambda|$ equations with $|\Omega|$ unknowns:

$$\langle f_j, g_k\rangle = \sum_{\ell\in\Omega} \overline{c_\ell^{(k)}} \langle f_j, g_\ell\rangle .$$

If we enumerate $\Lambda = \{\lambda_j\}_{j=1}^L$ and $\Omega = \{\omega_j\}_{j=1}^M$ we get the matrix equation

$$\begin{pmatrix} \langle f_{\lambda_1}, g_{\omega_1}\rangle & \langle f_{\lambda_1}, g_{\omega_2}\rangle & \cdots & \langle f_{\lambda_1}, g_{\omega_M}\rangle \\ \langle f_{\lambda_2}, g_{\omega_1}\rangle & \langle f_{\lambda_2}, g_{\omega_2}\rangle & \cdots & \langle f_{\lambda_2}, g_{\omega_M}\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle f_{\lambda_L}, g_{\omega_1}\rangle & \langle f_{\lambda_L}, g_{\omega_2}\rangle & \cdots & \langle f_{\lambda_L}, g_{\omega_M}\rangle \end{pmatrix} \begin{pmatrix} \overline{c_{\omega_1}^{(k)}} \\ \overline{c_{\omega_2}^{(k)}} \\ \vdots \\ \overline{c_{\omega_M}^{(k)}} \end{pmatrix} = \begin{pmatrix} \langle f_{\lambda_1}, g_k\rangle \\ \langle f_{\lambda_2}, g_k\rangle \\ \vdots \\ \langle f_{\lambda_L}, g_k\rangle \end{pmatrix} \tag{6}$$

for all $k \in \Lambda$. We call the matrix in (6) the *bridge matrix* and denote it $B(F,G,\Lambda,\Omega)$. Since the bridge matrix does not change depending on k, we can solve for all of the coefficients simultaneously with the equation

$$\left(\left\langle f_{\lambda_j}, g_{\omega_k}\right\rangle\right)_{j,k} \left(\overline{c_{\omega_j}^{\lambda_k}}\right)_{j,k} = \left(\left\langle f_{\lambda_j}, g_{\lambda_k}\right\rangle\right)_{j,k}. \tag{7}$$

We can rewrite this equation as

$$B(F,G,\Lambda,\Omega)C = B(F,G,\Lambda,\Lambda), \tag{8}$$

where C denotes our coefficient matrix (actually, C is the matrix of complex conjugates of the coefficients $c_{\omega_j}^{(\lambda_k)}$ in (7)).

Remark 1. (1) The transpose of the bridge matrix $B(F,G,\Lambda,\Omega)$ is a skew (i.e., diagonal-disjoint) minor of the cross-Gram matrix $Gr(F,G)$ of the frames F and G, and the transpose of $B(F,G,\Lambda,\Lambda)$ is a principal minor of $Gr(F,G)$.

(2) The form of the bridge matrix in equation (6) depends on the particular enumerations one takes of Λ and Ω. However, for two different enumerations one bridge matrix will transform into the other by interchanging appropriate rows and columns, and so the norm and the rank of the matrices will be the same. In particular, one will be invertible if and only if the other is.

Given a dual frame pair (F,G), and an erasure set Λ, a bridge set Ω is said to satisfy the *robust bridging condition* (or Ω is a *robust* bridge set) if equation (8) has a solution.

To get an idea of what this means, consider the case of one erasure. Let $\Lambda = \{k\}$, and choose a set $\Omega = \{\ell\}$. Then, $g_k' = c\, g_\ell$. For Nilpotent bridging, we require that $\langle f_k, g_k - g_k' \rangle = 0$. In solving for c, we get

$$0 = \langle f_k, g_k - g_k' \rangle = \langle f_k, g_k \rangle - \overline{c}\, \langle f_k, g_\ell \rangle .$$

So, if $\langle f_k, g_\ell \rangle \neq 0$, then Ω is a robust bridge set for Λ and

$$g_k' = \frac{\langle g_k, f_k \rangle}{\langle g_\ell, f_k \rangle} g_\ell. \tag{9}$$

In particular any singleton set $\{\ell\}$ is a robust bridge set for Λ provided $\langle f_k, g_\ell \rangle \neq 0$. So, in a suitably random frame, any singleton set disjoint from Λ will be a robust bridge set. □

Now, given $f \in \mathscr{H}$, recall that $f = f_{\widetilde{E}} + \widetilde{f}$. However, $\widetilde{E}_\Lambda (f - \widetilde{f}) = \widetilde{E}_\Lambda^2 f = 0$. Thus, $f_{\widetilde{E}} = \widetilde{E}_\Lambda \widetilde{f}$, and we can reconstruct f from the good Fourier coefficients by

$$f = \widetilde{f} + \widetilde{E}_\Lambda \widetilde{f}. \tag{10}$$

Furthermore, $f_B \in \operatorname{span}\{f_j : j \in \Lambda\}$, so by (5), $\widetilde{E}_\Lambda f_B = 0$. Therefore, to reconstruct f, we have

$$f = \widetilde{f} + \widetilde{E}_\Lambda f_R. \tag{11}$$

A simple algorithm for erasure recovery using 2-nilpotent bridging

By combining 2-nilpotent bridging and the final computational step using the reduced (2-nilpotent) error operator, we can write down a simple algorithm for accomplishing perfect reconstruction (or recovery) from erasures.

Let $\{f_j, g_j\}_{j \in \mathbb{J}}$ be a dual frame pair, Λ be an erasure set, and Ω be a corresponding robust bridge set. For $j \in \Lambda$ and $f \in \mathscr{H}$

$$\begin{aligned}\langle f, g_j\rangle &= \langle f, g'_j\rangle + \langle f, g_j - g'_j\rangle \\ &= \langle f, g'_j\rangle + \langle f - f_R, g_j - g'_j\rangle + \langle f_R, g_j - g'_j\rangle .\end{aligned}$$

Since $f - f_R \in \operatorname{span}\{f_j : j \in \Lambda\}$, equation (5) says that $f - f_R \perp g_j - g'_j$. So,

$$\begin{aligned}\langle f, g_j\rangle &= \langle f, g'_j\rangle + \langle f_R, g_j - g'_j\rangle \\ &= \langle f - f_R, g'_j\rangle + \langle f_R, g_j\rangle \\ &= \sum_{k\in\Omega} \overline{c_k^{(j)}}\, \langle f - f_R, g_k\rangle + \langle f_R, g_j\rangle .\end{aligned}$$

Therefore, we can recover the erased coefficients with the following equation:

$$(\langle f, g_j\rangle)_{j\in\Lambda} = C^T(\langle f - f_R, g_k\rangle)_{k\in\Omega} + (\langle f_R, g_j\rangle)_{j\in\Lambda} . \tag{12}$$

Remark 2. With the above algorithm, we have simplified the problem of inverting R_Λ (an $n \times n$ matrix) to inverting $B(F, G, \Lambda, \Omega)$ (an $L \times L$ matrix). This is particularly useful when the size of the erasure set is small when compared to the dimension of our underlying Hilbert space.

Remark 3. Equation (12) shows that an erased (or missing) Fourier coefficient of f over Λ can be precisely computed as the corresponding Fourier coefficient of the partial reconstruction f_R plus a bridging term which depends only on the Fourier coefficients of f_R and f over the bridge set, Ω. It follows that any noise or error in the computation of f_R can result in an error in the bridging term. This error can be large if the bridge system is ill conditioned, resulting in a potential error amplification. The good news is that only the errors in the Fourier coefficients of f_R over the indices in $\Lambda \cup \Omega$ will affect this amplification.

If for notational purposes we set $\alpha_j = \langle f, g_j\rangle$ and $\beta_j = \langle f_R, g_j\rangle$, then α_j for $j \in \Omega$ are known coefficients, and the β_j are all computable because f_R is computable. The above equation states that for $j \in \Lambda$ the erased coefficient $\alpha_j := \langle f, g_j\rangle$ can be computed as

$$\alpha_j = \beta_j + \sum_{k\in\Omega} \overline{c_k^{(j)}}(\alpha_k - \beta_k).$$

Implementation

In this subsection, we will implement the algorithm from the previous subsection. We will give several snippets of code followed by detailed explanations of what each block of code is used for. For the coding, we assume that the algorithm actually produces a solution. The section "Generic Duals and Infrastructure" discusses why in most cases this is a safe assumption.

```
n = 200;
N = 300;
L = [1:1:10];
W = [11:1:20];

F = rand(n,N);
S = F * F';
G = S \ F;
```

In this first bit of code, we are making a few basic variable declarations and randomly creating a dual frame pair, (F,G). The number n denotes the dimension of our Hilbert space (here we are using $\mathbb{R}^n$) and N denotes the length of our frame. The vectors L and W contain the erasure and the bridge indices, respectively. Here we have selected $\Lambda = \{1,2,\cdots,10\}$ and $\Omega = \{11,12,\cdots,20\}$. Any of these variables can be modified to the user's specifications prior to execution. The randomly computed $n \times N$ matrix F contains the frame vectors as its columns. Because of this, F actually denotes the synthesis matrix, Θ_F^* for our frame F. The matrix S is the frame operator for F, and G is the synthesis matrix for the standard dual to F.

```
f = rand(n,1);
f = f ./ norm(f,2);
FC = zeros(N,1);
for(k = 1:1:N)
  FC(k) = dot(f,G(:,k));
end
FC(L) = zeros(size(L'));
f_R = zeros(n,1);
for(k = 1:1:N)
  f_R = f_R + FC(k) * F(:,k);
end
```

In this piece of code, we create a random, unit norm vector $f \in \mathbb{R}^n$ which will be our test vector for the reconstruction. The vector FC denotes the Fourier coefficients for f with respect to the frame G. In the seventh line, we erase the Fourier coefficients that are indexed by L (the erasure set). In the last three lines, we compute f_R, the partial reconstruction of f.

```
FRCL = zeros(size(L))';
for(j = 1:1:max(size(L)))
  FRCL(j,1) = dot(f_R,G(:,L(j)));
end

FRCW = zeros(size(W))';
for(k = 1:1:max(size(W)))
  FRCW(k,1) = dot(f_R,G(:,W(k)));
end

C = (F(:,L)'*G(:,W))\(F(:,L)'*G(:,L));

FC(L) = C' * (FC(W) - FRCW) + FRCL;
```

The main purpose of this code is to recover the missing Fourier coefficients. This code is a direct implementation of the algorithm in the previous subsection. The vectors *FRCL* and *FRCW* (standing for Fourier coefficients of f_R with respect to Λ and Fourier coefficients of f_R with respect to Ω, respectively) denote $\left(\langle f_R, g_j\rangle\right)_{j\in\Lambda}$ and $\left(\langle f_R, g_j\rangle\right)_{j\in\Omega}$, respectively. In the second to last line we see the bridge equation written in Matlab code. It is easy to see that the matrix product $\Theta_{F,\Lambda}\Theta^*_{G,\Omega}$ is the bridge matrix, where $\Theta_{F,\Lambda}$ denotes the minor of Θ_F created by only using the rows indexed by Λ, and $\Theta^*_{G,\Omega}$ denotes the minor of Θ^*_G formed by only using the columns indexed by Ω. Thus, we see that $F(:,L)' * G(:,W)$ computes the bridge matrix, $B(F,G,\Lambda,\Omega)$. Similarly, $F(:,L)' * G(:,L)$ computes $B(F,G,\Lambda,\Lambda)$. Thus, the second to last line is the computation of our coefficient matrix, C. The last line is equation (12) which recovers the lost coefficient data.

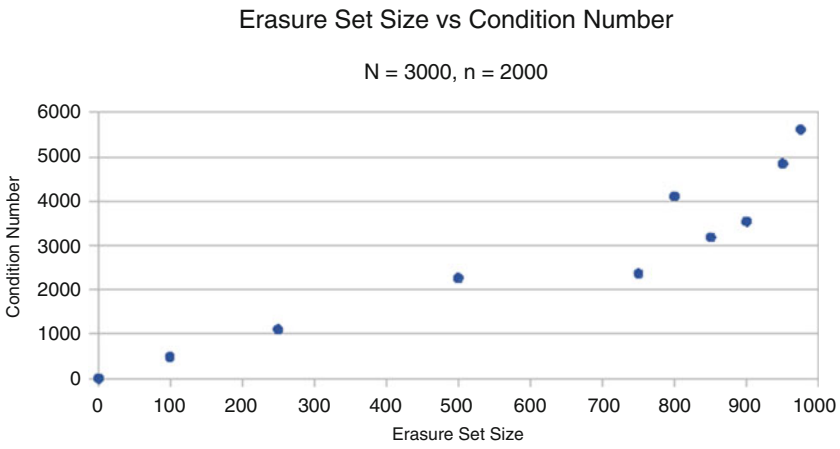
```
g = f_R;
for(j = L)
  g = g + FC(j) * F(:,j);
end

norm(f - g,2)
```

This last bit of code adds back our lost Fourier coefficient data, and we get the reconstructed vector g. To test the accuracy of our reconstruction, the last line computes the ℓ^2 norm difference of the original vector we randomly selected against our final reconstructed vector g.

Numerical Considerations

The main concern about the bridging method is the stability of the bridge matrix inversion. To examine the stability, we designed an experiment where we use frames of length $N = 3000$ in $n = 2000$ dimensions. For our experiment, we used erasure set sizes of 1, 100, 250, 500, 750, 800, 850, 900, 950, 975, and 1000. For each of these sizes, we performed 20 trials. In each trial, we computed a new random frame F and used the standard dual. We recorded the median of condition number of $B(F,G,\Lambda,\Omega)$ from the 20 trials, and due to the high variability of the data we recorded the standard deviation from the 20 trials. The following is a graph of these data.



Note that $|\Lambda| = 1000$ is omitted to avoid distortion. The median condition number at this point was 6.14×10^4. As we can see, the condition numbers "blow up" as $|\Lambda|$ approaches $N - n$. It is worth noting that even in the case of $|\Lambda| = 1000$, the largest condition number observed was on the order of 10^7 and the worst ℓ^2 error norm was on the order of 10^{-7}. Unfortunately, for this method there is a high degree of variability, with the standard deviation for each set of 20 trials usually exceeding the median condition number for those 20 trials.

Theoretical Considerations

The following result provides a necessary and sufficient condition for the existence of a robust bridge set for a given erasure set.

Theorem 1. *Let (F,G) be a dual frame pair, and let Λ be an erasure set. Then, there is a robust bridge set Ω for Λ if and only if Λ satisfies the minimal redundancy condition for G. In this case we can take $|\Omega| = \dim(\mathscr{F})$, where $\mathscr{F} = span\{f_j : j \in \Lambda\}$.*

For a rigorous proof of the above theorem we refer to Theorem 3.7 in [25]. We note that the structure of the proof gives some intuition into the role of the bridge matrix in the theory. Essentially, it is a change of basis matrix modulo the orthogonal complement of span$\{f_j : j \in \Lambda\}$. With some work it can be shown that the rank of the matrix $(\langle f_j, g_k \rangle)_{1\le j\le L, 1\le k\le M}$ is

$$\dim\left(\text{span}\{g_k\}_{1\le k\le M}/\left(\text{span}\{f_j\}_{1\le j\le L}\right)^{\perp}\right) = \dim\left(\text{span}\{f_j\}_{1\le j\le L}/\left(\text{span}\{g_k\}_{1\le k\le M}\right)^{\perp}\right).$$

The following is a useful criterion for sufficiency of robustness of a bridge set. We can compute $\dim(\mathscr{F})$ as the rank of the Gramian of $\mathscr{F}$. So in Matlab one can simply check whether the rank of $B(F,G,\Lambda,\Omega)$ equals the rank of $Gr(F)$. If his happens, then Ω is a robust bridge set. If it fails, it can still happen that Ω is a robust bridge set. We refer the reader to Theorem 3.8 in [25] for a proof.

Theorem 2. *Let (F,G) be a dual frame pair and Λ be an erasure set. If $\Omega \subset \Lambda^c$ is a bridge set for which*

$$\text{rank}(B(F,G,\Lambda,\Omega)) = \dim(\mathscr{F}) \tag{13}$$

where $\mathscr{F} = span\{f_j : j \in \Lambda\}$, then Ω is a robust bridge set. In particular if $|\Lambda| = |\Omega|$ and $B(F,G,\Lambda,\Omega)$ is invertible, then Ω is a robust bridge set.

The above theorem says that the rank condition (13) on the bridge matrix is sufficient for robustness of Ω. In the general case it is not necessary, as shown by Example 2. In that case, the unreduced error operator is already nilpotent of index 2, so any bridge set is robust for it. From experiments, it appears that the minimal rank possible of the bridge matrix for a robust bridge set and the minimal size of Ω is linked to the number of nonzero eigenvalues of the unreduced error operator. (See Theorem 4 for a result relating to this.) However, for Parseval frames, the converse of Theorem 2 holds. This is Corollary 3.10 in [25].

Theorem 3. *Let F be a Parseval frame. If Λ is an erasure set for F and $\Omega \subset \Lambda^c$, then Ω is a robust bridge set for Λ if and only if rank$(B(F,G,\Lambda,\Omega)) = \dim(\mathscr{F})$, where $\mathscr{F} = span\{f_j : j \in \Lambda\}$. In particular, if $\{f_j : j \in \Lambda\}$ is linearly independent and $|\Omega| = |\Lambda|$, then Ω is a robust bridge set for Λ if and only if $B(F,G,\Lambda,\Omega)$ is invertible.*

We give two examples that illustrate the relationship between the minimal redundancy condition and the invertibility of R_Λ. For the examples, we consider the dual frame pair

$$F = \left\{(1,1)^T, (-1,1)^T, (-1,-1)^T, (1,-1)^T\right\}$$

and

$$G = \left\{(1,0)^T, \left(\frac{1}{2},\frac{1}{2}\right)^T, \left(\frac{1}{2},-\frac{1}{2}\right)^T, (1,0)^T\right\}.$$

Our first example is an example where the 2-nilpotent bridging algorithm works, but R_Λ is not invertible.

Example 3. Let $\Lambda = \{1\}$,

$$R_\Lambda = \sum_{j=2}^{4} f_j \otimes g_j = I - f_1 \otimes g_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ -1 & 1 \end{pmatrix}$$

is not invertible. Therefore, methods that require the inversion of R_Λ will not work. Furthermore,

$$E_\Lambda = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

is idempotent, so Neumann series approximations also fail. However, since $\langle f_1, g_2 \rangle \neq 0$ and $\langle f_1, g_4 \rangle \neq 0$, equation (9) shows that nilpotent bridging works with $\Omega = \{2\}$ or $\Omega = \{4\}$. Note that $\Omega = \{3\}$ will not work for Nilpotent bridging since $\langle f_1, g_3 \rangle = 0$. □

While for robustness Λ needs to satisfy the minimal redundancy condition with respect to G, the second example shows that Λ need not satisfy the minimal redundancy condition with respect to F.

Example 4. Let $\Lambda = \{2,4\}$ and $\Omega = \{1,3\}$. Then, Λ does not satisfy the minimal redundancy condition for F. But, we have

$$f_2, f_4 \perp g_2 - 0g_1 - 0g_3 \text{ and}$$
$$f_2, f_4 \perp g_4 - g_1 - 0g_3.$$

Letting $f = (4,2)^T$, we get

$$f_R = R_\Lambda f = (f_1 \otimes g_1)(f) + (f_3 \otimes g_3)(f) = (3,3)^T$$

and

$$f_B = B_\Lambda f = (f_2 \otimes 0)(f) + (f_4 \otimes g_1)(f) = (4,-4)^T.$$

So,

$$\widetilde{f} = f_R + f_B = (7,-1)^T.$$

We have

$$f_{\widetilde{E}} = \widetilde{E}_\Lambda f_R = (f_2 \otimes (g_2 - 0g_1 - 0g_3))(f_R) + (f_4 \otimes (g_4 - g_1 - 0g_3))(f_R) = (-3,3)^T.$$

Therefore we recover our original vector as

$$\widetilde{f} + f_{\widetilde{E}} = (4,2)^T.$$

□

Consider a dual frame pair (F,G) with erasure set Λ and bridge set Ω. Computer experiments indicated that if $|\Omega| < |\Lambda|$, then $|\sigma(\widetilde{E}_\Lambda) \setminus \{0\}| = |\Lambda| - |\Omega|$. So, if one chooses a bridge set that is too small, $\widetilde{E}_\Lambda$ will have nonzero eigenvalues, but may have fewer nonzero eigenvalues than E_Λ (the error operator without bridging).

This is what we call cardinal spectral reduction. The following gives a mathematical proof of this fact. It is Theorem 3.13 in [25], but we include the proof here because the (simple) proof tells the story of why underbridging yields reduction in the size of the spectrum. We first observed this phenomenon in running some computer experiments with large frames, and it seemed quite mysterious. So, we worked out the proof.

Theorem 4. *Let (F,G) be a dual frame pair. Assume Λ satisfies the minimal redundancy condition with respect to G, and $|\Lambda| = L$. Then, there is a bridge set Ω of any size $M \leq L$ so that $|\sigma(\widetilde{E}_\Lambda) \setminus \{0\}| \leq L - M$.*

Proof. By Theorem 1, we can find a robust bridge set $\Omega' \subset \Lambda^c$ satisfying $|\Omega'| \leq L$. That is, for each $k \in \Lambda$ we can find

$$g'_k = \sum_{j \in \Omega'} c_j^{(k)} g_j$$

so that $g'_k \perp \text{span}\{f_j : j \in \Lambda\}$. Assume that $\Omega' = \{\omega_1, \cdots, \omega_{|\Omega'|}\}$. Let $\Omega = \{\omega_1, \cdots, \omega_M\}$ and

$$g''_k = \sum_{j \in \Omega} c_j^{(k)} g_j.$$

Then,

$$\widetilde{E}_\Lambda = \sum_{k \in \Lambda} f_k \otimes (g_k - g''_k) = \sum_{k \in \Lambda} f_k \otimes (g_k - g'_k) + \sum_{k \in \Lambda} f_k \otimes (g'_k - g''_k).$$

Let $N = \widetilde{E}_\Lambda = \sum_{k \in \Lambda} f_k \otimes (g_k - g'_k)$, and $A = \sum_{k \in \Lambda} f_k \otimes (g'_k - g''_k)$. Then, it is easily verified that N is nilpotent of index 2, and $NA = 0$. Since $\text{range}(A^*) \subset \{g'_k - g''_k : k \in \Lambda\} \subset \{g_{\omega_k} : k = M+1, \cdots, |\Omega'|\}$, the rank of A is at most $L - M$.

Let $\lambda \in \sigma(N+A) \setminus \{0\}$. Both N and A are finite rank operators, so λ must be an eigenvalue of $N+A$. Thus, there exists $x \in \mathscr{H}$ so that

$$(N+A)x = \lambda x.$$

Multiplying by N on the left on both sides yields

$$0 = \lambda N x.$$

Since $\lambda \neq 0$, we have $Nx = 0$. Thus, $Ax = \lambda x$ and $\lambda \in \sigma(A)$. Since A can have at most $L - M$ distinct eigenvalues, it follows that $\widetilde{E}_\Lambda$ has at most $L - M$ nonzero eigenvalues. □

Applications to Sampling Theory

There are well-known deep established connections between frame theory and modern sampling theory. We cite for instance the excellent references [4, 18, 29].

We note that a good account of sampling theory for our purposes is contained in Chapter 9 of [21]. Let X be a metric space and let μ be a Borel measure on X. Let $\mathscr{H}$ be a closed subspace of $L^2(X,\mu)$ consisting of continuous functions. Let $T = \{t_j\}_{j\in\mathbb{J}} \subset X$ and define the sampling transform Θ mapping $\mathscr{H}$ into the complex sequences by $\Theta(f) = (f(t_j))_{j\in\mathbb{J}}$. If $\Theta : \mathscr{H} \to \ell^2(\mathbb{J})$ is bounded, then the point evaluation functionals $\gamma_j : \mathscr{H} \to \mathbb{C}$ defined by $\gamma_j(f) = f(t_j)$ are bounded, and hence by the Riesz Representation Theorem, $\gamma_j(f) = \langle f, g_j\rangle$ for some $g_j \in \mathscr{H}$. If the sampling transform is also bounded below, then $\{g_j\}_{j\in\mathbb{J}}$ forms a frame for $\mathscr{H}$, and thus we can find some dual $F := \{f_j\}_{j\in\mathbb{J}}$. We then have the identity

$$f = \sum_{j\in\mathbb{J}} \langle f, g_j\rangle f_j = \sum_{j\in\mathbb{J}} f(t_j) f_j \quad \forall f \in \mathscr{H}. \tag{14}$$

We will refer to (X,F,T) as a sampling scheme for $\mathscr{H}$. The most well-known sampling scheme comes from the Shannon-Whittaker Sampling Theorem. For this scheme, $\mathscr{H} = PW[-\pi,\pi]$, $T = p\mathbb{Z}$ ($p \in (0,1]$), and $f_j = \operatorname{sinc}(\pi(t-jp))$. Then, $g_j = p\operatorname{sinc}(\pi(t-jp))$, where $\operatorname{sinc}(x) = \frac{\sin x}{x}$.

Let Λ be an erasure set for a sampling scheme (X,F,T), with corresponding bridge set Ω. We can think of the erased coefficients as either $\langle f, g_j\rangle$ or $f(t_j)$ for $j \in \Lambda$. For this case, the bridge matrix is

$$B(F,G,\Lambda,\Omega) = \left(\langle f_j, g_k\rangle\right)_{j\in\Lambda,k\in\Omega} = \left(f_j(t_k)\right)_{j\in\Lambda,k\in\Omega}. \tag{15}$$

Similarly, $B(F,G,\Lambda,\Lambda) = (f_j(t_k))_{j,k\in\Lambda}$. Note that these matrices only involve the sampled values of the $\{f_j\}$ over the points $\{t_k\}$ and do not explicitly involve the $\{g_k\}$. Let us simply write $B(\Lambda,\Omega)$ and $B(\Lambda,\Lambda)$ for these two matrices. Then, the algorithm in subsection "A simple algorithm for erasure recovery using 2-nilpotent bridging" becomes the following Theorem:

Theorem 5. *Let (X,F,T) be a sampling scheme with erasure set Λ satisfying the minimal redundancy condition, and Ω be a robust bridge set for Λ. Suppose $C = \left(\overline{c_j^{(k)}}\right)_{j\in\Omega,k\in\Lambda}$ solves the bridging equation*

$$B(\Lambda,\Omega)C = B(\Lambda,\Lambda),$$

where $B(\Lambda,\Omega) = \left(f_j(t_k)\right)_{j\in\Lambda,k\in\Omega}$ and $B(\Lambda,\Lambda) = \left(f_j(t_k)\right)_{j,k\in\Lambda}$. Then,

$$(f(t_j))_{j\in\Lambda} = C^T\left((f(t_j))_{j\in\Omega} - (f_R(t_j))_{j\in\Omega}\right) + (f_R(t_j))_{j\in\Lambda}.$$

Remark 4. As in Remark 3, the error in the sampled values of the partial reconstruction f_R over the bridge set can result in a potential error amplification in our reconstruction. Unfortunately, to perfectly compute the sampled values of f_R over the bridge set we require an infinite sum, so a truncation of f_R is necessary. Thus, error amplification is unavoidable. A further analysis of this error term is given in the

next subsection, and the following subsection discusses methods to improve condition numbers for sampling schemes, which will lead to less amplification of error in our reconstruction.

N-Term Approximation Error

In practice we cannot deal with infinite sums. So, we must chop off our infinite sums, and use N-term approximations. To significantly simplify the computations, we assume that (X,F,T) is a sampling scheme and that F is the standard dual to G, where $\langle f, g_j\rangle = f(t_j)$ for all $j \in \mathbb{J}$. Let S denote the frame operator for G. Throughout this subsection, we assume that Λ is an omission set satisfying the minimal redundancy property, that Ω is a robust bridge set for Λ, and that the matrix $C = \left(\overline{c_k^{(n)}}\right)_{k\in\Omega, n\in\Lambda}$ solves the bridge equation. We assume $\Lambda, \Omega \subset \mathbb{J}_N \subset \mathbb{J}$, where $\mathbb{J}_N$ has cardinality N. We would like to know what the ensuing error in our bridging algorithm is when we only know $f(t_j)$ for $j \in \mathbb{J}_N \setminus \Lambda$.

Recall that we have the reconstruction formula $f = f_R + f_B + \widetilde{E}_\Lambda f_R$. Thus, it is natural to expect that $f \approx \widetilde{f} := f_R^{(N)} + f_B + \widetilde{E}_\Lambda f_R^{(N)}$, where we have

$$f_R^{(N)} := \sum_{j\in\mathbb{J}_N\setminus\Lambda} f(t_j) f_j \text{ and}$$

$$\widetilde{E}_\Lambda f_R^{(N)} := \sum_{n\in\Lambda} \left(f_R^{(N)}(t_j) - \sum_{k\in\Omega} c_k^{(n)} f_R^{(N)}(t_k) \right) f_n.$$

Note that since $\Lambda, \Omega \subset \mathbb{J}_N$, f_B is the same as its N-term approximation.

A useful identity for the error bound is the following equality:

$$\|f_j\|_{L^2(\mu)}^2 = \langle f_j, f_j\rangle = \langle f_j, S^{-1} g_j\rangle = \langle S^{-1} f_j, g_j\rangle = (S^{-1} f_j)(t_j). \tag{16}$$

The following proposition provides an upper bound for the error in our N-term approximation.

Proposition 1. *With the above terminology, we have*

$$\|f - \widetilde{f}\|_{L^2(\mu)} \le \sum_{j\in\mathbb{J}_N^c} \left| f(t_j) \sqrt{(S^{-1} f_j)(t_j)} \right| + \sum_{j\in\Lambda} \sum_{m\in\mathbb{J}_N^c} \left| f(t_m) f_m(t_j) - \sum_{k\in\Omega} c_k^{(j)} f(t_m) f_m(t_k) \right| \sqrt{(S^{-1} f_j)(t_j)}.$$

Proof. We have

$$\begin{aligned}\|f-\widetilde{f}\|_{L^2(\mu)} &= \|f_R+f_B+\widetilde{E}_\Lambda f_R-f_R^{(N)}-f_B-\widetilde{E}_\Lambda f_R^{(N)}\|_{L^2(\mu)}\\ &\le \|f_R-f_R^{(N)}\|_{L^2(\mu)}+\|\widetilde{E}_\Lambda f_R-\widetilde{E}_\Lambda f_R^{(N)}\|_{L^2(\mu)}.\end{aligned}$$

We first find a bound for the first piece:

$$\begin{aligned}|f_R-f_R^{(N)}|^2 &= \left|\sum_{j\in\mathbb{J}_N^c} f(t_j)f_j\right|^2\\ &\le \left(\sum_{j\in\mathbb{J}_N^c}\left|f(t_j)f_j\right|\right)^2\\ &= \sum_{j,m\in\mathbb{J}_N^c}|f(t_j)f(t_m)||f_jf_m|.\end{aligned}$$

Thus,

$$\begin{aligned}\|f_R-f_R^{(N)}\|_{L^2(\mu)}^2 &\le \sum_{j,m\in\mathbb{J}_N^c}|f(t_j)f(t_m)|\|f_jf_m\|_{L^1(\mu)}\\ &\le \sum_{j,m\in\mathbb{J}_N^c}|f(t_j)f(t_m)|\|f_j\|_{L^2(\mu)}\|f_m\|_{L^2(\mu)}\\ &= \sum_{j,m\in\mathbb{J}_N^c}|f(t_j)f(t_m)|\sqrt{(S^{-1}f_j)(t_j)}\sqrt{(S^{-1}f_m)(t_m)}\\ &= \left(\sum_{j\in\mathbb{J}_N^c}\left|f(t_j)\sqrt{(S^{-1}f_j)(t_j)}\right|\right)^2.\end{aligned}$$

where the third line holds by equation (16). Hence,

$$\|f_R-f_R^{(N)}\|_{L^2(\mu)} \le \sum_{j\in\mathbb{J}_N^c}\left|f(t_j)\sqrt{(S^{-1}f_j)(t_j)}\right|. \tag{17}$$

For the second piece, we have

$$\begin{aligned}|\widetilde{E}_\Lambda f_R-\widetilde{E}_\Lambda f_R^{(N)}|^2 &= \left|\sum_{j\in\Lambda}\left\{(f_R(t_j)-f_R^{(N)}(t_j))-\sum_{k\in\Omega}c_k^{(j)}((f_R(t_k)-f_R^{(N)}(t_k)))\right\}f_j\right|^2\\ &\le \left(\sum_{j\in\Lambda}\left|\sum_{m\in\mathbb{J}_N^c}f(t_m)f_m(t_j)-\sum_{k\in\Omega}c_k^{(j)}\sum_{m\in\mathbb{J}_N^c}f(t_m)f_m(t_k)\right||f_j|\right)^2\\ &\le \left(\sum_{j\in\Lambda}\sum_{m\in\mathbb{J}_N^c}\left|f(t_m)f_m(t_j)-\sum_{k\in\Omega}c_k^{(j)}f(t_m)f_m(t_k)\right||f_j|\right)^2.\end{aligned}$$

To simplify the calculation, let

$$a_j = \sum_{m\in\mathbb{J}_N^c} \left| f(t_m)f_m(t_j) - \sum_{k\in\Omega} c_k^{(j)} f(t_m)f_m(t_k) \right|. \tag{18}$$

Then, we have

$$|\widetilde{E}_\Lambda f_R - \widetilde{E}_\Lambda f_R^{(N)}|^2 \le \sum_{j,\ell\in\Lambda} a_j a_\ell |f_j f_\ell|. \tag{19}$$

So,

$$\begin{aligned}
\|\widetilde{E}_\Lambda f_R - \widetilde{E}_\Lambda f_R^{(N)}\|_{L^2(\mu)}^2 &\le \sum_{j,\ell\in\Lambda} a_j a_\ell \|f_j f_\ell\|_{L^1(\mu)} \\
&\le \sum_{j,\ell\in\Lambda} a_j a_\ell \|f_j\|_{L^2(\mu)} \|f_\ell\|_{L^2(\mu)} \\
&= \left(\sum_{j\in\Lambda} a_j \|f_j\|_{L^2(\mu)} \right)^2 \\
&= \left(\sum_{j\in\Lambda} \sum_{m\in\mathbb{J}_N^c} \left| f(t_m)f_m(t_j) - \sum_{k\in\Omega} c_k^{(j)} f(t_m)f_m(t_k) \right| \sqrt{(S^{-1}f_j)(t_j)} \right)^2
\end{aligned}$$

where the last equality holds by (16). Taking square roots gives

$$\|\widetilde{E}_\Lambda f_R - \widetilde{E}_\Lambda f_R^{(N)}\|_{L^2(\mu)} \le \sum_{j\in\Lambda} \sum_{m\in\mathbb{J}_N^c} \left| f(t_m)f_m(t_j) - \sum_{k\in\Omega} c_k^{(j)} f(t_m)f_m(t_k) \right| \sqrt{(S^{-1}f_j)(t_j)}. \tag{20}$$

Therefore, adding (17) and (20) yields the error bound

$$\begin{aligned}
\|f - \widetilde{f}\|_{L^2(\mu)} \le & \sum_{j\in\mathbb{J}_N^c} \left| f(t_j)\sqrt{(S^{-1}f_j)(t_j)} \right| \\
& + \sum_{j\in\Lambda} \sum_{m\in\mathbb{J}_N^c} \left| f(t_m)f_m(t_j) - \sum_{k\in\Omega} c_k^{(j)} f(t_m)f_m(t_k) \right| \sqrt{(S^{-1}f_j)(t_j)}.
\end{aligned}$$

□

Numerical Considerations

Not only do we have to look out for truncation error but we also need to be careful in our choice of bridge set. Consider the Shannon-Whittaker sampling scheme with $p = \frac{1}{2}$. If we assume Λ and Ω only contain even integers, then

$$B(\Lambda,\Omega) = (f_j(t_k))_{j\in\Lambda,k\in\Omega} = \left(\operatorname{sinc}\left(\frac{\pi}{2}(k-j) \right) \right)_{j\in\Lambda,k\in\Omega}.$$

However, $k-j$ is even, and so $B(\Lambda,\Omega)$ is the zero matrix. Not only this, but in general, if we randomly select a bridge set for a given erasure set, $B(\Lambda,\Omega)$ will be extremely poorly conditioned. For example, when $|\Lambda|=5$ and using 20001 Fourier coefficients as our truncation we have run experiments where the condition numbers have been on the order of 10^{14}. This problem can be fixed by choosing a bridge set which is "close" to our original erasure set. A problem arises when the bridge and erasure sets are sufficiently separated because the components of the bridge matrix are of the form $\operatorname{sinc}\left(\frac{\pi}{2}(k-j)\right)$, and thus the components are very small.

Suppose (X,F,T) is a sampling scheme and let d denote the metric on X. We say that $\Omega=\{\omega_j : j=1,\cdots,L\}$ is a *close* bridge set to $\Lambda=\{\lambda_j : j=1,2,\cdots,L\}$ if it minimizes

$$d(\Lambda,\Omega) := \sum_{j=1}^{L} d(\lambda_j,\omega_j),$$

where we take the minimum over all robust bridge sets, Ω. For example, for the Shannon-Whittaker sampling scheme with $p=\frac{1}{2}$, $\Omega=\{0,1,2,\cdots,M\}$ is a close bridge set for $\Lambda=\{\frac{2j+1}{2} : j=0,1,\cdots,M\}$, provided it is a robust bridge set. Now, if Ω is a close bridge set for Λ, the components along the diagonal are more dominant, and the off diagonal entries are smaller in absolute value. That is, the bridge matrix is *diagonally dominant*. Therefore, it is reasonable to expect that close bridge sets are much more well conditioned. Indeed, experimentally even with $|\Lambda|=1000$, for a truncation using 20001 Fourier coefficients, with 10 runs of our Matlab code, and Λ chosen randomly, the median condition number of $B(\Lambda,\Omega)$ was 11.89.

The following algorithm searches for a bridge set Ω that is "near" Λ. We say near because it does not always attain a close bridge, but is still "good enough for practical purposes".

```
Pos = setdiff(1:2*N+1,La);
B = zeros(L,L);
for(j = 1:1:L)
  [mini,position] = min(abs(Pos-La(j)));
  Om(j) = Pos(position);
  Pos = setdiff(Pos,Om(j));
  B(:,j) = sinc(pi*p*(Om(j)-La));
  while(rank(B(1:j,:)) < j)
    [mini,position] = min(abs(Pos-La(j)));
    Om(j) = Pos(position);
    Pos = setdiff(Pos,Om(j));
    B(:,j) = sinc(pi*p*(Om(j)-La));
  end
end
```

In the above code, $2N+1$ denotes the number of sampled values we are using for our truncation, $L = |\Lambda|$, *La* denotes Λ, the vector *Om* denotes Ω, B is our bridge matrix, and *Pos* is a vector that tells us what values are still eligible for the bridge set. In our indexing scheme, the index 1 stands for $-pN$ and the index $2N+1$ stands for pN, where N is the number of positive Fourier coefficients we are using. Thus, we are using $2N+1$ total Fourier coefficients. In the notation of the previous subsection, we are using $\mathbb{J}_{2N+1} = \{-pN, -(p-1)N, \cdots, (p-1)N, pN\}$ as our truncation.

The first line of code throws out the elements of *La* from the list of possible candidates of the bridge set, *Pos*. Inside of the *for* loop, we are building the matrix B column by column and ensuring that B has rank j at the end of each iteration. The first line in the for loop finds the closest element of *Pos* to the *jth* element of the erasure set and stores that index value in the variable *position*. The next line assigns the value to the j^{th} entry in Ω. The following line deletes this index from the possibilities for the next element of the bridge. The next line computes the j^{th} column of the bridge matrix, and if the first j columns of the matrix have rank less than j, the while loop keeps repeating the above procedure until it finds an index so that B has rank j.

Unfortunately, if too many consecutively sampled values are erased, even close bridges will be poorly conditioned, and a good reconstruction will be impossible. When considering sampling at the half integers, if we take $\Lambda = \{\frac{j}{2} : j = 1, 2, \cdots, 15\}$ and considering $f(t) = \text{sinc}(\pi t)$, our algorithm gives a reconstruction whose Fourier coefficients are off by as much as 0.324. The condition number of the bridge matrix in this case is 7.093×10^6.

Generic Duals and Infrastructure

We consider *spark*, *m-linear*, and related properties for finite frames with an eye toward bridging applications. Let $\mathscr{H}$ be an n-dimensional Hilbert space. Denote the set of N-tuples of vectors in $\mathscr{H}$ by $\mathscr{H}^N$. For a frame $F \in \mathscr{H}^N$, we define $\mathscr{D}(F) = \{G \in \mathscr{H}^N : (F,G) \text{ is a dual frame pair}\}$ and call it the *dual set* of F. It is easily shown that $\mathscr{D}(F)$ is a closed convex subset of $\mathscr{H}^N$ in any of the equivalent linear space norms. In this section it will be convenient to adopt the norm $\|F\| := \max_{1 \le j \le N} \|f_j\|$ for $F = \{f_j\}_{j=1}^N \in \mathscr{H}^N$, where $\|f_j\|$ is the usual Hilbert space norm on $\mathscr{H}$. Since $\mathscr{D}(F)$ is closed, it is a complete metric space with the norm topology inherited from $\mathscr{H}^N$.

In the frame literature, a class of frames is sometimes called *generic* if it is open and dense in the set of all frames (cf. [1, 2, 26]). We will say that a class of duals to a given frame F is generic if it is open and dense in the relative topology on $\mathscr{D}(F)$ inherited as a subspace of $\mathscr{H}^N$.

The next lemma shows that in the presence of the minimal redundancy condition, one can explicitly construct "*designer duals*" that satisfy certain conditions with respect to Λ.

Lemma 1. *Let Λ be an erasure set for a frame F with the minimal redundancy condition and $\{g_j\}_{j \in \Lambda}$ be assigned arbitrarily. Then,*

1. $\{g_j\}_{j\in\Lambda}$ *can be extended to* $\{g_j\}_{j=1}^N$ *so that* $\sum_{j=1}^N f_j \otimes g_j = 0$.
2. $\{g_j\}_{j\in\Lambda}$ *can be extended to* $\{g_j\}_{j=1}^N \in \mathscr{D}(F)$.

Proof. For (1), let $A = \sum_{j\in\Lambda} f_j \otimes g_j$. Let $\{h_j\}_{j\in\Lambda^c}$ be a dual to the reduced frame $\{f_j\}_{j\in\Lambda^c}$. Then,

$$A = \left(\sum_{j\in\Lambda^c} f_j \otimes h_j\right) A = \sum_{j\in\Lambda^c} f_j \otimes (A^* h_j).$$

For each $j \in \Lambda^c$, let $g_j = -A^* h_j$. Then,

$$\begin{aligned}\sum_{j=1}^N f_j \otimes g_j &= \sum_{j\in\Lambda^c} f_j \otimes g_j + \sum_{j\in\Lambda} f_j \otimes g_j = -\sum_{j\in\Lambda^c} f_j \otimes A^* h_j + A \\ &= A - \left(\sum_{j\in\Lambda^c} f_j \otimes h_j\right) A = A - IA = 0.\end{aligned}$$

To prove (2), let $\{g'_j\}_{j=1}^N \in \mathscr{D}(F)$. Let $h_j = g_j - g'_j$ for $j \in \Lambda$. By the part (1), we can extend $\{h_j\}_{j\in\Lambda}$ to $\{h_j\}_{j=1}^N$ so that $\sum_{j=1}^N f_j \otimes h_j = 0$. For all j, let $\widetilde{g_j} = g'_j + h_j$. Then,

$$\sum_{j=1}^N f_j \otimes \widetilde{g_j} = \sum_{j=1}^N f_j \otimes g'_j + \sum_{j=1}^N f_j \otimes h_j = I + 0 = I.$$

Thus, $\{\widetilde{g_j}\}_{j=1}^N \in \mathscr{D}(F)$. Furthermore, for $j \in \Lambda$,

$$\widetilde{g_j} = g'_j + h_j = g'_j + g_j - g'_j = g_j.$$

Thus, $\{\widetilde{g_j}\}_{j=1}^N$ is the desired extension of $\{g_j\}_{j\in\Lambda}$. □

In the frame literature (cf. [1, 13]) a frame F in $\mathscr{H}$ is said to have *spark k* if every collection of $k-1$ vectors in F is linearly independent and there is a collection of k vectors in F which is linearly dependent, and it is said to be *full spark* if it has spark $n+1$, where n is the dimension of $\mathscr{H}$. A frame is said to be *m-independent* (cf. [20, 27]) if every collection of m vectors in F is linearly independent. So, spark k means m-independent for all $m < k$ and **not** k-independent, and full spark means n-independent. It is known that the set of full spark frames is an open dense set in $\mathscr{H}^N$ (cf. [1, 26]). That is, full spark frames are generic. Note that m-independence is hereditary in the sense that it implies j-independence for all $j \le m$.

Lemma 2. *Let (F,G) be a dual frame pair for an n-dimensional Hilbert space with length N. Let Λ be a set, and Ω be a bridge set satisfying $|\Lambda| = |\Omega|$. A necessary (but not sufficient) condition for $B(F,G,\Lambda,\Omega)$ to be an invertible matrix is*

$$|\Lambda| \le \min\left\{n, N-n, \frac{N}{2}\right\} \tag{21}$$

(Note that in case $N \ge 2n$ this condition reduces to $|\Lambda| \le n$.)

Proof. If $|\Lambda| > n$, then the rows of the bridge matrix $B(F,G,\Lambda,\Omega)$ will be linearly dependent (since $\mathscr{H}$ is an n-dimensional space). Thus, $B(F,G,\Lambda,\Omega)$ will fail to be invertible.

Assume that $B(F,G,\Lambda,\Omega)$ is invertible, and $|\Lambda| > N-n$. Then, since the bridge equation $B(F,G,\Lambda,\Omega)C = B(F,G,\Lambda,\Lambda)$ has a solution ($C = B(F,G,\Lambda,\Omega)^{-1}$ $B(F,G,\Lambda,\Lambda)$), Theorem 1 asserts that Λ satisfies the minimal redundancy condition with respect to G. Therefore, $|\Lambda^c| \geq n$. So, $N = |\Lambda| + |\Lambda^c| > N-n+n > N$. This is a contradiction, and therefore if $B(F,G,\Lambda,\Omega)$ is invertible, then $|\Lambda| \leq N-n$.

If $|\Lambda| > \frac{N}{2}$, then $|\Lambda| + |\Omega| > N$. This is a contradiction since Λ and Ω are disjoint subsets of $\{1,\cdots,N\}$. □

Corollary 1. *Assume that $F \in \mathscr{H}^N$ is full spark. Let Λ be an erasure set satisfying $|\Lambda| \leq \min\{n, N-n, \frac{N}{2}\}$ and Ω be a bridge set satisfying $|\Lambda| = |\Omega|$ and $\Lambda \cap \Omega = \emptyset$. Then, there exists a dual frame G to F so that $B(F,G,\Lambda,\Omega)$ is invertible.*

Proof. Define a bijection $\varphi : \Omega \to \Lambda$. Let $\{g_j\}_{j\in\Omega} = \{f_{\varphi(j)}\}_{j\in\Omega}$. By Lemma 1, we can extend $\{g_j\}_{j\in\Omega}$ to a dual frame G for F. Then, $B(F,G,\Lambda,\Omega)$ is identical to the Gram matrix of a permutation of the finite sequence $\{f_j : j \in \Lambda\}$, which is invertible since $\{f_j : j \in \Lambda\}$ is linearly independent. □

We say that a dual frame pair (F,G) has *skew-spark k* if for every erasure set Λ with $|\Lambda| < k$ and any bridge set $\Omega \subset \Lambda^c$ satisfying $|\Lambda| = |\Omega|$, the bridge matrix $B(F,G,\Lambda,\Omega)$ is invertible. If (F,G) has skew-spark $\min\{\frac{N}{2}, n, N-n\}+1$, then (F,G) is said to have *full skew-spark*.

Proposition 2. *If the dual frame pair (F,G) for $\mathscr{H}$ has skew-spark k, then F and G each have spark at least k.*

Proof. Let Λ be an erasure set of cardinality $k-1$. Let Ω be any subset of Λ^c of cardinality $k-1$. By hypothesis the matrix $B(F,G,\Lambda,\Omega)$ is invertible, so its rows and columns are linearly independent. This implies that $\{f_j : j \in \Lambda\}$ is linearly independent. Since Λ was arbitrary, this shows that F has spark k. The proof for G is analogous. □

This also shows that if $n \leq \min\{\frac{N}{2}, N-n\}$, then full skew-spark implies full spark.

Let $\mathscr{G} = \{G \in \mathscr{D}(F) : (F,G) \text{ has full skew-spark}\}$.

Theorem 6. *Assume that F has full spark. Then,*

$$\mathscr{G} := \{G \in \mathscr{D}(F) : (F,G) \textit{ has full skew-spark}\}$$

is generic in $\mathscr{D}(F)$.

Proof. Let $\Gamma = \{\Lambda \subset \{1,\cdots,N\} : |\Lambda| \leq \min\{\frac{N}{2}, N-n, n\}\}$. For a given $\Lambda \in \Gamma$, let $\Phi_\Lambda = \{\Omega \subset \{1,\cdots,N\} : |\Omega| = |\Lambda|, \Omega \cap \Lambda = \emptyset\}$. Then, $\mathscr{G} = \bigcap_{\Lambda\in\Gamma} \bigcap_{\Omega\in\Phi_\Lambda} \mathscr{G}_{\Lambda,\Omega}$, where $\mathscr{G}_{\Lambda,\Omega} = \{G \in \mathscr{D}(F) : \det(B(F,G,\Lambda,\Omega)) \neq 0\}$. Since we are intersecting over

all possible erasure sets and all corresponding bridge sets, the above intersection is finite. So, by the Baire Category Theorem, if we show that each $\mathscr{G}_{\Lambda,\Omega}$ is open and dense, then $\mathscr{G}$ will also be open and dense.

Fix an erasure set Λ and a corresponding bridge set Ω. It is easily verified that the maps $G \overset{\alpha}{\mapsto} B(F,G,\Lambda,\Omega)$ and $B(F,G,\Lambda,\Omega) \mapsto \det(B(F,G,\Lambda,\Omega))$ are continuous. So, $\mathscr{G}_{\Lambda,\Omega} = (\det\circ\alpha)^{-1}(\mathbb{C}\setminus\{0\})$ is an open set.

To show density of $\mathscr{G}_{\Lambda,\Omega}$, let $\varepsilon > 0$, and assume that $G_0 \in \mathscr{D}(F)\setminus\mathscr{G}_{\Lambda,\Omega}$. Since F has full spark, Λ satisfies the minimal redundancy condition with respect to F. Thus, by Corollary 1, there is a $G_1 \in \mathscr{D}(F)$ so that $\det(B(F,G_1,\Lambda,\Omega)) \neq 0$. Let $G_t = (1-t)G_0 + tG_1$. By proposition 2, $G_t \in \mathscr{D}(F)$. Furthermore, $\det(B(F,G_t,\Lambda,\Omega))$ is a polynomial in t satisfying $\det(B(F,G_t,\Lambda,\Omega))(0) = 0$ and $\det(B(F,G_t,\Lambda,\Omega))(1) \neq 0$. Thus, $\det(B(F,G_t,\Lambda,\Omega))$ has only finitely many zeros. So, we can find $0 < t_0 < \frac{\varepsilon}{\|G_1-G_0\|}$ so that $G_{t_0} \in \mathscr{G}_{\Lambda,\Omega}$. Furthermore,

$$\|G_{t_0} - G_0\| = \|(1-t_0)G_0 + t_0G_1 - G_0\| = \|t_0(G_1-G_0)\| \leq t_0\|G_1-G_0\| < \varepsilon.$$

Hence, $\mathscr{G}_{\Lambda,\Omega}$ is dense in $\mathscr{D}(F)$.

Therefore, by the Baire Category Theorem, $\mathscr{G}$ is generic in $\mathscr{D}(F)$. □

In short, what we have proven in this section so far is that for most frames $F \in \mathscr{H}^N$, and most duals G to F, the pair (F,G) has full skew-spark. Arguably, the most important class of dual frame pairs are those of the form (F,F) where F is a Parseval frame. The above theorem gives little information about this important class. Our initial computer experiments largely involved Parseval frames, and these experiments provided ample evidence to us that randomly computed Parseval frames probably have full skew-spark. This provided quite a bit of our motivation for proving the above generic results for dual frame pairs. However, at the time of writing [25] we did not have a rigorous proof for the class of Parseval frames. We have since worked out a proof, which we include here:

Full Skew-Spark Parseval Frames

Let

$$PF_N(\mathscr{H}) = \{F \in \mathscr{H}^N : F \text{ is a Parseval frame}\}.$$

It is easy to verify that $PF_N(\mathscr{H})$ is a closed subset of $\mathscr{H}^N$, and hence it is a complete metric space with the norm topology induced from $\mathscr{H}^N$.

Theorem 7. *The set* $\mathscr{P} = \{F \in PF_N(\mathscr{H}) : F \text{ has full skew-spark}\}$ *is generic in* $PF_N(\mathscr{H})$.

To prove this theorem, we require two lemmas.

Lemma 3. *There exists a Parseval frame* F *such that* $B(F,F,\Lambda,\Omega)$ *is* $\frac{1}{2}I_L$ *(the* $L\times L$ *identity matrix), where* $L = |\Lambda| = |\Omega| \leq min\left\{n, N-n, \frac{N}{2}\right\}$ *and* $\Omega\cap\Lambda = \emptyset$.

Proof. Enumerate $\Lambda = \{\lambda_j\}_{j=1}^L$ and $\Omega = \{\omega_j\}_{j=1}^L$. Let $\{e_j\}_{j=1}^L$ be an orthonormal set in $\mathscr{H}$. For each $1 \leq j \leq L$, set $f_{\lambda_j} = f_{\omega_j} = \frac{1}{\sqrt{2}} e_j$. Let $\mathscr{F} = \text{span}\{f_j : j \in \Lambda \cup \Omega\}$. Then, $\dim(\mathscr{F}^\perp) = n - L \leq N - 2L$ since $n \leq N - L$. So, since $|\{1, 2, \cdots, L\} \setminus (\Lambda \cup \Omega)| = N - 2L$, one can find a Parseval frame $\{f_j : j \in (\Lambda \cup \Omega)^c\}$ for $\mathscr{F}^\perp$. Then, $\mathscr{F} = \{f_j\}_{j=1}^N$ is a frame for $\mathscr{H}$ for which $B(F, F, \Lambda, \Omega) = \frac{1}{2} I_L$. □

Lemma 4. *Let F be a Parseval frame for an n-dimensional Hilbert space $\mathscr{H}$. Let Λ be an erasure set with $|\Lambda| \leq \min\left\{n, N-n, \frac{N}{2}\right\}$ and Ω be a bridge set for Λ with $\Lambda \cap \Omega = \emptyset$ and $|\Lambda| = |\Omega|$. Then, given $\varepsilon > 0$, there exists a Parseval frame $\widetilde{F}$ with $\|F - \widetilde{F}\| < \varepsilon$ so that $B(\widetilde{F}, \widetilde{F}, \Lambda, \Omega)$ is invertible.*

Proof. Enumerate $\Lambda = \{\lambda_j\}_{j=1}^L$ and $\Omega = \{\omega_j\}_{j=1}^L$. Assume without loss of generality that $B(F, F, \Lambda, \Omega)$ is singular. By Lemma 3, there is a Parseval frame $F_1 = \{f_j^{(1)}\}_{j=1}^N$ such that $B(F_1, F_1, \Lambda, \Omega)$ is invertible. Let $F = F_0 = \{f_j^{(0)}\}_{j=1}^N$. For $0 < t < 1$, define $F_t = \{f_j^{(t)}\}_{j=1}^N$, where $f_j^{(t)} = (1-t) f_j^{(0)} + t f_j^{(1)}$. Then, for each $t \in [0, 1]$, F_t is a set of vectors in $\mathscr{H}$, but they need not span $\mathscr{H}$, and hence are not necessarily a frame for $\mathscr{H}$. Let S_t be the frame operator (or Bessel operator in the case that F_t is not a frame) for F_t. Then,

$$S_t = \sum_{j=1}^N f_j^{(t)} \otimes f_j^{(t)}.$$

If $\{e_j\}_{j=1}^n$ is the standard orthonormal basis for $\mathscr{H}$, the matrix coordinate functions $m_{j,k}(t) = \left\langle S_t e_k, e_j \right\rangle$ of the matrix $M(t)$ of S_t with respect to $\{e_j\}_{j=1}^n$ are quadratic functions of t. Thus, $\det M(t)$ is a polynomial function of t. Hence, the *formal inverse* matrix valued function, which we denote by $Q(t)$, that is given by the *adjoint* formula (or Cramer's rule) for the inverse of an invertible matrix has the form $Q(t) = \left(q_{jk}(t)\right)_{j,k=1}^n$, where the coordinate functions $q_{jk}(t)$ are *rational* functions of t.

At points where S_t is invertible, let P_t be the Parseval frame $P_t = S_t^{-\frac{1}{2}} F_t$, where $S_t^{-\frac{1}{2}} = (S_t^{-1})^{\frac{1}{2}}$ is the positive square root of S_t^{-1}. Then, $P_0 = F_0$. We have

$$B(P_t, P_t, \Lambda, \Omega) = \left(\left\langle S_t^{-\frac{1}{2}} f_{\lambda_j}^{(t)}, S_t^{-\frac{1}{2}} f_{\omega_k}^{(t)} \right\rangle \right)_{j,k=1}^L = \left(\left\langle S_t^{-1} f_{\lambda_j}^{(t)}, f_{\omega_k}^{(t)} \right\rangle \right)_{j,k=1}^L.$$

If $[f_{\lambda_j}^{(t)}]_E$ denotes the coordinate vector of $f_{\lambda_j}^{(t)}$ with respect to $E = \{e_j\}_{j=1}^n$, and we define $[f_{\omega_k}^{(t)}]_E$ similarly, then for points $t \in [0, 1]$ for which S_t is invertible,

$$\left\langle S_t^{-1} f_{\lambda_j}^{(t)}, f_{\omega_k}^{(t)} \right\rangle = \left\langle Q(t) [f_{\lambda_j}^{(t)}]_E, [f_{\omega_k}^{(t)}]_E \right\rangle.$$

Since $Q(t)$ is an $n \times n$ matrix valued function with rational coordinate functions, and $[f_{\lambda_j}^{(t)}]_E$ and $[f_{\omega_k}^{(t)}]_E$ are matrix valued vectors with polynomial coordinate functions,

this yields a formal rational matrix valued function on $[0,1]$ with rational coordinate functions

$$b_{j,k}(t) := \left\langle Q(t)[f_{\lambda_j}^{(t)}]_E, [f_{\omega_k}^{(t)}]_E \right\rangle.$$

Let $b(t) := (b_{j,k}(t))_{j,k=1}^{L}$. Let $\delta(t) := \det(b(t))$ denote the formal determinant. At points t where S_t is invertible we have $\delta(t) = \det(B(P_t, P_t, \Lambda, \Omega))$. By hypothesis, F_0 and F_1 are Parseval frames, so $S_0 = S_1 = I_n$ (the $n \times n$ identity matrix). Thus $P_0 = F_0 = F$ and $P_1 = F_1$. By hypothesis $B(F, F, \Lambda, \Omega)$ is singular, so $\delta(0) = 0$. By construction, $B(F_1, F_1, \Lambda, \Omega)$ is invertible, and so $\delta(1) \neq 0$. A nonconstant rational function can have at most finitely many points where it is undefined, and at most finitely many zeros. So, there is an $\alpha > 0$ so that $\delta(t)$ is defined and nonzero. Since S_0 is invertible and the map $t \mapsto S_t$ is continuous at $t = 0$, the map $t \mapsto S_t^{-\frac{1}{2}}$ is continuous at $t = 0$. Thus, the map $t \mapsto P_t$ is continuous at $t = 0$. So, there exists $\alpha_1 \in (0, \alpha)$ so that $\|F - P_t\| < \varepsilon$ wherever $t \in [0, \alpha_1]$. Choose $\widetilde{t} \in (0, \alpha_1)$ and let $\widetilde{F} = P_{\widetilde{t}}$. Then, $\|F - \widetilde{F}\| < \varepsilon$ and $\delta(\widetilde{t}) \neq 0$, so $B(\widetilde{F}, \widetilde{F}, \Lambda, \Omega)$ is invertible as required. □

Proof (Proof of Theorem 7). Let $\Gamma = \left\{\Lambda \subset \{1, 2, \cdots, N\} : |\Lambda| \leq \min\left\{n, N-n, \frac{N}{2}\right\}\right\}$, and given $\Lambda \in \Gamma$, define $\Phi_\Lambda = \{\Omega : \Omega \text{ is a bridge set for } \Lambda\}$. Then,

$$\mathscr{P} = \cap_{\Omega \in \Phi_\Lambda} \cap_{\Lambda \in \Gamma} \mathscr{P}_{(\Lambda, \Omega)}.$$

By Lemma 4, each $\mathscr{P}_{(\Lambda,\Omega)}$ is dense in $PF_N(\mathscr{H})$.

Define $\delta_{(\Lambda,\Omega)} : \mathscr{P}_{(\Lambda,\Omega)} \to \mathbb{C}$ by

$$\delta_{(\Lambda,\Omega)}(F) = \det(B(F, F, \Lambda, \Omega)).$$

Then, since $\delta_{(\Lambda,\Omega)}$ is continuous,

$$\mathscr{P}_{(\Lambda,\Omega)} = \delta_{(\Lambda,\Omega)}^{-1}(\mathbb{C} \setminus \{0\}).$$

Thus, each $\mathscr{P}_{(\Lambda,\Omega)}$ is open.

Therefore, by the Baire Category Theorem, $\mathscr{P}$ is open and dense in $PF_N(\mathscr{H})$. □

Remark 5.

Applying the Baire Category Theorem in the above proofs is a bit of overkill: all one needs is the elementary fact that in any metric space a finite intersection of open dense sets is open and dense. However, most of the concepts in this section extend to infinite frames in an infinite dimensional space, as long as the erasure sets are finite, and we think that it is likely that this type of result will remain true. In this case an application of the Baire Category Theorem indeed would seem necessary.

Remark 6.

We found it convenient to present and prove the topological results of this section for the metric topology. A similar argument can be used to obtain these for the Zariski topology.

Computing an Inverse for R_Λ

In this section, we obtain a basis-free closed-form formula for the inverse of the partial reconstruction operator R_Λ for a finite erasure set. By basis free we mean that the computations do not depend on any preassigned basis for the space, and by closed-form we mean that it is of the same general form as R_Λ is given in and does not require an iterative process such as the Neumann series formula. This gives a second method of perfect reconstruction from frame and sampling erasures in finitely many steps that applies when R_Λ^{-1} exists. Furthermore, this method only requires a matrix inversion of size $|\Lambda| \times |\Lambda|$, and Neumann series techniques can be applied to this small matrix to speed up the matrix inversion.

The motivation for this formula comes from an observation that a rank-1 perturbation of the identity operator, $A = I - x \otimes y$, is invertible if and only if $\langle x, y\rangle \neq 1$, and in this case the inverse is

$$A^{-1} = I + \frac{1}{1 - \langle x, y\rangle} x \otimes y. \tag{22}$$

This formula does not seem to have been used in the frame literature for 1-erasures. In attempting to generalize this to any finite number of erasures we discovered the general formula in Theorem 8 below.

Let (F, G) be a dual frame pair indexed by $\mathbb{J}$ and Λ be an erasure set. Recall that

$$R_\Lambda = \sum_{j \in \mathbb{J} \setminus \Lambda} f_j \otimes g_j = I - \sum_{j \in \Lambda} f_j \otimes g_j.$$

Motivated by this, we derive a simple method for computing inverses of operators of the form

$$R = I - \sum_{j=1}^{L} f_j \otimes g_j.$$

Remark 7. If $R = I - \sum_{j=1}^{L} f_j \otimes g_j$ is invertible, an elementary operator theory proof (cf. Proposition 6.1 in [25]) shows that R^{-1} has the form $I + \sum_{j,k=1}^{L} c_{jk} f_j \otimes g_k$ for some $c_{jk} \in \mathbb{C}$.

Although the elementary tensors $f_j \otimes g_k$ in the representation of R^{-1} in Proposition 6.1 are generally not linearly independent and hence the coefficients $\{c_{jk}\}_{j,k=1}^{L}$ are not unique, we can derive a simple matricial formula that gives a valid choice of the c_{jk}.

Theorem 8. *Let $R = I - \sum_{j=1}^{L} f_j \otimes g_j$, where $\{f_j\}_{j=1}^{L}, \{g_j\}_{j=1}^{L}$ are finite sequences and $\{f_j\}_{j=1}^{L}$ is linearly independent. Assume I denotes the $L \times L$ identity matrix and*

$$M = Gr(\{f_1, \ldots, f_L\}, \{g_1, \ldots, g_L\}) := \begin{pmatrix} \langle f_1, g_1\rangle & \langle f_2, g_1\rangle & \cdots & \langle f_L, g_1\rangle \\ \langle f_1, g_2\rangle & \langle f_2, g_2\rangle & \cdots & \langle f_L, g_2\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle f_1, g_L\rangle & \langle f_2, g_L\rangle & \cdots & \langle f_L, g_L\rangle \end{pmatrix}. \tag{23}$$

Then, R is invertible if and only if $(I-M)$ is invertible. Moreover, if $(I-M)^{-1}$ exists, setting

$$C := (c_{jk})_{j,k=1}^{L} = (I-M)^{-1}, \tag{24}$$

we have

$$R^{-1} = I + \sum_{j,k=1}^{L} c_{jk} f_j \otimes g_k. \tag{25}$$

Proof. Assume R^{-1} exists. By remark 7 we can write

$$R^{-1} = I + \sum_{j=1}^{L}\sum_{k=1}^{L} c_{jk} f_j \otimes g_k$$

for some $c_{jk} \in \mathbb{C}$. Compute

$$\begin{aligned}
I &= R^{-1}R \\
&= \left(I + \sum_{j=1}^{L}\sum_{k=1}^{L} c_{jk} f_j \otimes g_k\right)\left(I - \sum_{j=1}^{L} f_j \otimes g_j\right) \\
&= I + \sum_{j=1}^{L}\sum_{k=1}^{L} c_{jk} f_j \otimes g_k - \sum_{j=1}^{L} f_j \otimes g_j - \sum_{j=1}^{L}\sum_{k=1}^{L}\sum_{\ell=1}^{L} c_{jk}(f_j \otimes g_k)(f_\ell \otimes g_\ell) \\
&= I + \sum_{j=1}^{L}\sum_{k=1}^{L} c_{jk} f_j \otimes g_k - \sum_{j=1}^{L} f_j \otimes g_j - \sum_{j=1}^{L}\sum_{\ell=1}^{L}\sum_{k=1}^{L} c_{jk} \langle f_\ell, g_k\rangle (f_j \otimes g_\ell) \\
&= I + \sum_{j=1}^{L}\sum_{k=1}^{L} c_{jk} f_j \otimes g_k - \sum_{j=1}^{L} f_j \otimes g_j - \sum_{\ell=1}^{L}\sum_{j=1}^{L}\sum_{k=1}^{L} c_{j\ell} \langle f_k, g_\ell\rangle (f_j \otimes g_k).
\end{aligned}$$

In the last sum, we switched indices k and ℓ. Thus,

$$\sum_{j=1}^{L} f_j \otimes g_j = \sum_{j=1}^{L}\sum_{k=1}^{L} c_{jk} f_j \otimes g_k - \sum_{\ell=1}^{L}\sum_{j=1}^{L}\sum_{k=1}^{L} c_{j\ell} \langle f_k, g_\ell\rangle (f_j \otimes g_k).$$

By simply setting the coefficients of the $f_j \otimes g_k$ equal to $\delta_{j,k}$, we obtain the following system of equations:

$$c_{jk} - \sum_{\ell=1}^{L} c_{j\ell} \langle f_k, g_\ell\rangle = \delta_{jk}. \tag{26}$$

For a fixed value of j, we have the system

$$(\delta_{jk})_{k=1,\cdots,L}^{T} = \begin{pmatrix} 1-\langle f_1,g_1\rangle & -\langle f_1,g_2\rangle & \cdots & -\langle f_1,g_L\rangle \\ -\langle f_2,g_1\rangle & 1-\langle f_2,g_2\rangle & \cdots & -\langle f_2,g_L\rangle \\ \vdots & \vdots & \ddots & \vdots \\ -\langle f_L,g_1\rangle & -\langle f_L,g_2\rangle & \cdots & 1-\langle f_L,g_L\rangle \end{pmatrix} (c_{jk})_{k=1,\cdots,L}^{T}.$$

Let $C = \left(c_{jk}\right)_{j,k}$. Combining the equations for all j gives

$$I = (I - M^T)C^T,$$

where $M = Gr(\{f_1, \ldots, f_L\}, \{g_1, \ldots, g_L\})$. So, $C(I - M) = I$.

We will show that under the hypothesis that $\{f_1, \cdots f_L\}$ is linearly independent the matrix $I - M$ is invertible, so this system has a unique solution. This will yield a valid choice of the c_{jk}. If $I - M$ were singular then 1 would be an eigenvalue of M. So, there would exist a nonzero vector $x = (x_k)_{k=1}^L \in \mathbb{C}^n$ so that $Mx = x$. Computing gives $\sum_{j=1}^L \langle f_j, g_k \rangle x_j = x_k$ for each k. Let $z = \sum_{j=1}^L x_j f_j$. Since x is nonzero not all of the x_j are zero. By hypothesis $\{f_1, \cdots f_L\}$ is linearly independent, so z cannot be the zero vector. Compute

$$Rz = z - \sum_{k=1}^{L} \langle z, g_k \rangle f_k = z - \sum_{k=1}^{L} \sum_{j=1}^{L} x_j \langle f_j, g_k \rangle f_k = z - \sum_{k=1}^{L} x_k f_k = z - z = 0.$$

So, z is in the kernel of R contradicting our hypothesis that R is invertible. Thus $I - M$ is a nonsingular matrix, and the system has the unique solution $C = (I - M)^{-1}$ as claimed.

Conversely, if $(I - M)^{-1}$ exists, the above computations give an explicit formula for R^{-1}. □

Remark 8. If $\{f_j : j \in \Lambda\}$ is not linearly independent, to apply Theorem 8, one must first use linearity of the elementary tensors $f \otimes g$ in the first component and conjugate linearity in the second component to precondition R to the form $I - \sum_{j=1}^L f'_j \otimes g'_j$ with the first component set $\{f'_j : j \in \Lambda\}$ linearly independent. In many cases this will be simple and even automatic, but in other cases this may be computationally expensive. The main point is that if R is invertible, the computation above, perhaps with preconditioning, always yields a formula for the inverse. Furthermore, it may be useful to note that since we are solving a matrix equation, it follows that the coefficients c_{jk} are given by rational functions of the $\langle f_j, g_k \rangle$. In this sense the formula is indeed basis free.

There are two methods to implement Theorem 8. Notice that we can either directly invert $I - M$ or we can utilize a Neumann series. The first method we call the *direct inversion method*, and the second we call the *Neumann iterative method*. The implementation and numerical stability of each of these methods is discussed below.

Implementation of the Direct Inversion Method

With the exception of the bridge set, the first two code snippets in the bridging section are also used in the direct inversion method. The next piece is given here.

```
M = (F(:,L)' * G(:,L))';
C = (eye(max(size(L))) - M) \ eye(max(size(L)));
```

In this block of code, we compute the matrix M from equation (23) (in a similar way to the creation of the bridging matrix in the implementation section for the bridging method) and use it to calculate $(I-M)^{-1}$ to determine the coefficients given by equation (24).

```
g = f_R;
for(j = 1:1:max(size(L)))
   for(k = 1:1:max(size(L)))
      g = g + C(j,k) * dot(f_R,G(:,k)) * F(:,j);
   end
end

norm(f-g,2)
```

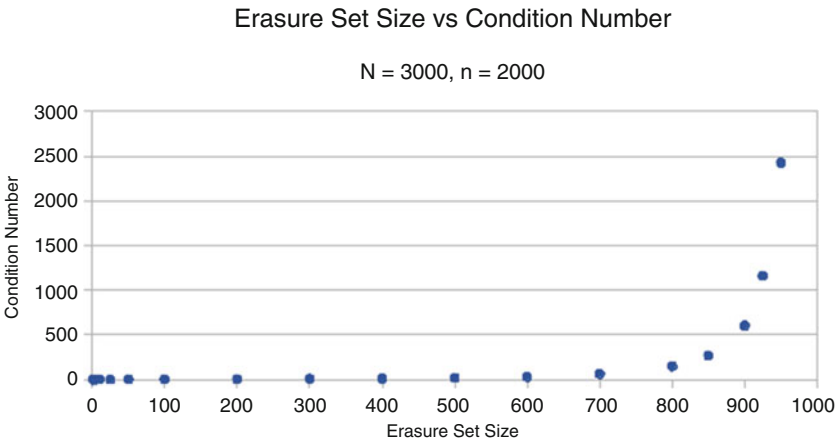
Once we know the coefficients as in Theorem 8, the nested for loop above computes

$$f_R + \sum_{j,k\in\Lambda} c_{j,k} \langle f_R, g_k \rangle f_j.$$

Lastly, to check the accuracy of the method, the last line calculates the ℓ^2 norm of the difference of the reconstructed vector and the original vector. Unfortunately, the nested *for* loop slows down the computations. If properly parallelized, this code could run much more efficiently.

Numerical Considerations

For this method, we are mostly concerned with the stability of the inversion of $I-M$. To understand the stability, we designed an experiment with frames of length $N = 3000$ in $n = 2000$ dimensions. For the experiment, we use erasure set sizes of 1, 5, 10, 25, 50, multiples of 100 from 100 to 1000, 850, 925, 950, and 975. For each size, 10 trials were run and of the 10 trials, the median condition number was recorded. For each separate trial, we used a new random frame and its standard dual. The following graph displays the data collected.



Note that the point $|\Lambda| = N - n = 1000$ was omitted in the graph to avoid distortion. The median condition number at this point was 1.23×10^7. This shows that as $|\Lambda|$ gets close to $N - n$, the condition numbers start to "blow up". However, we can still get a reconstruction for $|\Lambda| = N - n$ as with bridging. In this case, the ℓ^2 norm of the error in the reconstruction of a unit norm vector is still on the order of 10^{-8}.

Implementation of the Neumann Iterative Method

For this method, instead of directly inverting in equation (24), we use the approximation

$$(I - M)^{-1} \approx \sum_{j=1}^{J} M^j.$$

Unfortunately, there is no clear-cut way to get a relation between the number of iterations used, J, and an upper bound on the error in our reconstruction. However, it is clear that the smaller the error in the approximation of $(I - M)^{-1}$, the better our approximation will be. Notice that in the real case, and using the standard dual frame, $M^* = M$. So, $\|M\| = r(M)$. Let $M_J = \sum_{j=0}^{J} M^j$. If we know $r(M) = \|M\|$, we can compute the corresponding J so that $\|(I - M)^{-1} - M_J\| < \varepsilon$. We have

$$\|(I - M)^{-1} - M_J\| \leq \sum_{j=J+1}^{\infty} \|M\|^j = \frac{\|M\|^{J+1}}{1 - \|M\|}. \tag{27}$$

Hence, to achieve an ε tolerance a sufficient condition is

$$\frac{\|M\|^{J+1}}{1 - \|M\|} < \varepsilon. \tag{28}$$

Therefore, the number of Neumann iterations required to get an ε tolerance is given by

$$J > \log_{\|M\|}(\varepsilon(1-\|M\|)) - 1. \tag{29}$$

Thus, in place of the second block of code in the direct inversion method we use the following Matlab code.

```
tolerance = 10^(-10);

M = (F(:,L)' * G(:,L))';
Mnorm = max(abs(eigs(M)));

J = log(tolerance*(1-Mnorm))/log(Mnorm)-1;
M_J = eye(max(size(L)));
for(j = 1:1:(J+1))
  M_J = eye(max(size(L))) + M*M_J;
end

C = M_J;
```

Here, tolerance is the amount of tolerance we allow in $\|(I-M)^{-1} - M_J\|$. The variable *Mnorm* is the spectral radius and norm of M. J is given by equation (29). Lastly, the for loop computes $M_J \approx C$. As with the Direct inversion, we still have the nested *for* loop slowing down our computations.

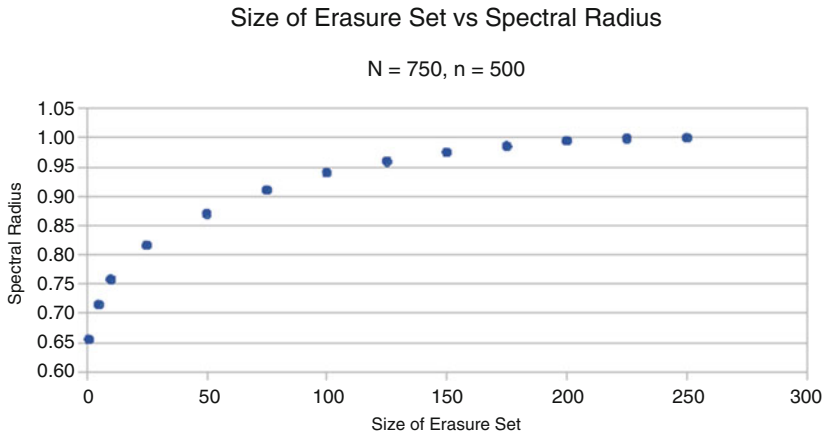
Numerical Considerations

Questions we are interested in for this method are the following:

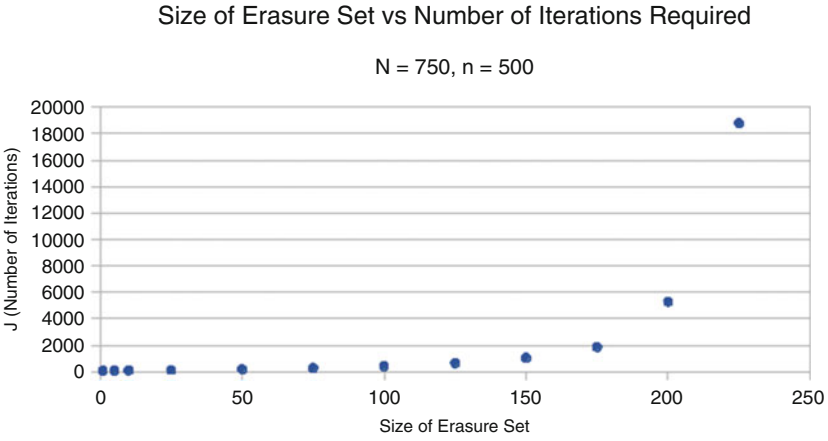
1. How frequently is $\|M\| < 1$ or $r(M) < 1$?
2. How many iterations are needed to get an accurate reconstruction?

To answer our questions, we designed an experiment with frames of length $N = 750$ on $n = 500$ dimensions. We used erasure set sizes of 1, 5, 10, and multiples of 25 from 25 to 250. For each size, 10 trials were run and the median spectral radius was computed. For each trial, we used a new random frame and its standard dual. Based on the median spectral radius, we then computed the number of iterations required so that $\|M - M_J\| < 10^{-10}$.

To answer question 1, the following graph suggests that as long as $|\Lambda| < N - n = 250$, $\|E_\Lambda\| = r(E_\Lambda) < 1$.



For these spectral radii, we can compute the corresponding J from equation (29) so that $\|M - M_J\| < 10^{-10}$. The following plot is of these computed values of J so that $\|M - M_J\| < 10^{-10}$. It shows that as $|\Lambda|$ approaches $N - n$, J evidently increases without bound.



Concluding Remarks

The main results in this chapter are presented for reconstruction from finite erasure subsets of frames, so much of our theory is finite dimensional. However, the reconstruction results can be applied to finite subsets of infinite frames, including the well-known classes of Gabor (Weyl-Heisenberg) frames, Laurent frames, infinite group frames, and wavelet frames, as well as abstract sampling theory. There may be applications to the pure and applied aspects of these classes, including classification results. In fact, our initial computer experiments suggest to us that many of these natural classes of infinite frames may be full skew-spark in the sense that they

have skew-spark k for all finite k. But mathematical proofs of general theorems on this have eluded us so far. In addition, there may be applications to the three closely related topics that deal with frames in *blocks*: operator-valued frames, fusion frames, and G-frames (cf. [12, 24, 28]). Finally, we should mention that we expect that there will be applications to the more abstract theories: frames for Banach spaces and related topics of Banach frames, atomic decompositions, and framings (cf. [11]), the theory of frames for Hilbert C* modules, and in the purely algebraic direction: frames for other fields such as p-adic frames and binary frames (cf. [6, 23]).

References

1. B. Alexeev, J. Cahill, D. Mixon, Full spark frames. J. Fourier Anal. Appl. **18**(6), 1167–1194 (2012)
2. R. Balan, P.G. Casazza, D. Edidin, Equivalence of reconstruction from the absolute value of the frame coefficients to a sparse representation problem. IEEE Signal Process. Lett. **14**(5), 341–343 (2007)
3. R. Balan, B.G. Bodmann, P.G. Casazza, D. Edidin, Frames for linear reconstruction without phase, in *The 42nd Annual Conference on Information Sciences and Systems* (2008), pp. 721–726
4. J. Benedetto, P.J.S.G. Ferriera (eds.), *Modern Sampling Theory* (Birkhäuser, Boston, 2001)
5. B.G. Bodmann, V.I. Paulsen, Frames, graphs and erasures. Linear Algebra Appl. **404**, 118–146 (2005)
6. B.G. Bodmann, M. Le, L. Reza, M. Tobin, M. Tomforde, Frame theory for binary vector spaces. Involve **2**(5), 589–602 (2009)
7. P. Boufounos, A.V. Oppenheim, V.K. Goyal, Causal compensation for erasures in frame representations. IEEE Trans. Signal Process. **56**(3), 1071–1082 (2008)
8. P.G. Casazza, J. Kovačević, Equal-norm tight frames with erasures. Adv. Comput. Math. **18**, 387–430 (2003)
9. P.G. Casazza, G. Kutyniok, Robustness of fusion frames under erasures of subspaces and of local frame vectors. Contemp. Math. **464**, 149–160 (2008)
10. P.G. Casazza, G. Kutyniok, *Finite Frames: Theory and Application*. Applied and Numerical Harmonic Analysis (Birkhäuser, Basel, 2014)
11. P.G. Casazza, D. Han, D.R. Larson, Frames for Banach spaces, in *The Functional and Harmonic Analysis of Wavelets and Frames* (San Antonio, TX, 1999). Contemporary Mathematics, vol. 247 (American Mathematical Society, 1999), pp. 149–182
12. P.G. Casazza, G. Kutyniok, S. Li, Fusion frames and distributed processing. Appl. Comput. Harmon. Anal. **25**, 114–132 (2008)
13. P.G. Casazza, R.G. Lynch, J.C. Tremain, L.M. Woodland, Integer frames. Houst. J. Math. (to appear)
14. O. Christensen, *An Introduction to Frames and Riesz Bases* (Birkhäuser, Basel, 2003)
15. R.J. Duffin, A.C. Schaeffer, A class of nonharmonic Fourier series. Trans. Am. Math. Soc. **72**, 341–366 (1952)
16. A. Gilat, *Matlab: An Introduction with Applications* (Wiley, New York, 2004)
17. V.K. Goyal, J. Kovačević, J.A. Kelner, Quantized frame expansions with erasures. Appl. Comput. Harmon. Anal. **10**, 203–233 (2001)
18. K. Gröchenig, *Foundations of Time Frequency Analysis*. Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, 2001)

19. D. Han, D.R. Larson, Frames, bases and group representations. Mem. Am. Math. Soc. **697** (2000)
20. D. Han, W. Sun, Reconstruction of signals from frame coefficients with erasures at unknown locations. IEEE Trans. Inf. Theory (to appear)
21. D. Han, K. Kornelson, D. Larson, E. Weber. *Frames for Undergraduates*. American Mathematical Society Student Mathematical Library, vol. 40 (American Mathematical Society, Providence, RI, 2007)
22. R. Holmes, V.I. Paulsen, Optimal frames for erasures. Linear Algebra Appl. **377**, 31–51 (2004)
23. R. Hotovy, D.R. Larson, S. Scholze, Binary frames. Houst. J. Math. (to appear)
24. V. Kaftal, D.R. Larson, S. Zhang, Operator valued frames. Trans. Am. Math. Soc., **361**, 6349–6385 (2009)
25. D. Larson, S. Scholze, Signal reconstruction from frame and sampling erasures. J. Fourier Anal. Appl. (to appear)
26. Y.M. Lu, M.N. Do, A theory for sampling signals from a union of subspaces. IEEE Trans. Signal Process. **56**(6), 2334–2345 (2008)
27. S. Pehlivan, D. Han, R. Mohapatra, Linearly connected sequences and spectrally optimal dual frames for erasures. J. Funct. Anal. (to appear)
28. W. Sun, G-frames and g-Riesz bases. J. Math. Anal. Appl. **232**, 437–452 (2006)
29. A.I. Zayed, *Advances in Shannon's Sampling Theory* (CRC Press, Boca Raton, 1993)

Choosing Function Spaces in Harmonic Analysis

Hans G. Feichtinger

Abstract Without doubt function spaces play a crucial role in Harmonic Analysis. Moreover many function spaces arose from questions in Fourier analysis. Here we would like to draw the attention to the question: "Which function spaces are useful for which problem?" Looking into the books on Fourier analysis one may come to the conclusion that it has almost become a dogma that one has to study *Lebesgue integration* and $\boldsymbol{L}^p$-spaces properly in order to have a chance to understand the Fourier transform. For the study of PDEs one has to resort to *Sobolev spaces*, or the Schwartz theory of *tempered distribution*, where suddenly Lebesgue spaces play a minor role. Finally, numerical applications make use of the FFT (Fast Fourier Transform), which has a vast range of applications in signal processing, but in the corresponding engineering books neither Lebesgue nor Schwartz theory plays a significant role. "Strange objects" like a Dirac distribution or Dirac combs (used to prove sampling theorems) are often used in a mysterious way, divergent integrals are giving magically useful result s. Cautious authors provide some hint to the fact that "mathematicians know how to give those objects a correct meaning"[1] More recently other function systems, such as wavelets and Gabor expansions, have come into the picture, as well as the theory of spline-type spaces and irregular sampling have gained importance. In this context the classical function spaces such as $\boldsymbol{L}^p$-spaces or even Sobolev and Besov spaces are not really helpful and do not allow to derive good results. Instead, *Wiener amalgam spaces* and *modulation spaces* are playing a major role there. It is the purpose of this chapter to initiate a discussion about the "information content" of function spaces and their "usefulness". In fact,

[1] But they are giving the engineering students at the same time the feeling that it is too complicated, that only the pedantic mathematicians have to take care of such theoretical foundations, while for them it is not worthwhile the effort, or too time consuming to even try to understand this theory properly.

H.G. Feichtinger (✉)
Faculty of Mathematics, University of Vienna, Vienna, Austria
e-mail: hans.feichtinger@univie.ac.at

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_3

even the discussion of the meaning of such words may be a stimulating challenge for the community and worth the effort. When I illustrate this circle of problems in the context of time-frequency analysis, but also with respect to potential usefulness for the teaching of the subject to engineers, I do not mean to specifically promote my favorite spaces, but rather show - in a context very familiar to me - how I want to understand the question. Of course such a description is subjective, while, on the other hand, it provides a kind of experience report, indicating that I personally found those spaces useful for many of the things I have been doing in the last decades. It also favors obviously less popular spaces over the well-known and frequently used ones.

Context and Background Information

This chapter is part of a *series of survey papers by the authors* (with varying coauthors) which tries to lay the ground for an *alternative approach to Fourier analysis* overall. It is meant to replace the *top-down* approach (starting from Lebesgue spaces over LCA groups and then going to distribution theory) by a *bottom-up* approach, starting from finite Abelian groups and linear algebra, going straight to a theory of generalized functions. The reservoir of objects which can be treated in this way should be comprehensive enough to cover most cases (e.g., for engineering applications or for the discussion of questions of abstract harmonic analysis), but still technically much less involved than the usual theory of (Schwartz-Bruhat) distributions [22].

The articles published in this (informal) series are the following ones: First [77], showing an access to Gabor Analysis via linear algebra, followed by the article [78] describing the theory of Gabor expansions over finite Abelian groups. Here one finds the *algebraic* backbone of Gabor analysis (without the trouble of functional analytic issues arising in the Euclidean context).

The last article published so far is [60], where the idea of *Conceptual Harmonic Analysis* (CHA) is introduced. Recall that *Abstract Harmonic Analysis* (as promoted, e.g., in [37, 81, 99, 100, 117]) views Fourier analysis over different LCA groups in parallel and provides consistent terminology, such as *characters, the dual group, orthogonal subgroups, etc.* In contrast, CHA takes in addition to this unifying terminology the connections between the different settings into account. How can we approximate (functions or distribution) on $\mathbb{R}^d$ resp. their continuous Fourier transforms using functions over finite Abelian groups, such as $\mathbb{Z}_N^d$? Can we use code for Gabor analysis over finite Abelian groups in order to approximately compute approximate dual Gabor windows (see [27, 79])? Typical first answers on such questions are given, for example, in [108]. Since these results require the operations of regular sampling and periodization these operations have to be properly described using Dirac combs within the context of generalized functions.

The papers [69] and [74] which can be considered "classical" by now provide the technical background in the setting of *modulation spaces*, specifically the Segal algebra $\boldsymbol{S}_0(\mathbb{R}^d)$ and its dual space, presented in a more condensed form in [31], featuring specifically *Banach Gelfand Triples*. A summary of applications and useful facts concerning this specific topic will also be given in [71].

Function Spaces: the Current Situation

If we look the publications in the field of function spaces, one can easily come to the impression that it has become an industry (like many other fields of science), with comprehensive output, concerning a large variety of (old and new) function spaces. Individual papers typically compare certain spaces, look at the mapping properties of concrete (families of) operators on certain spaces, improve the range of parameters for such operators, reformulate known results in the setting of new spaces, and so on. It is hard to keep up with these developments and make effective use of these systematic studies and find the information content (if it is there) hidden in the various propositions and theorems of the published manuscripts.

It seems that for many authors more and more the main concern in the preparation of publishable results appears to be this one: "How can I find a question, that has not been treated in the literature, which is within my reach and hence will lead to a cited publication?"

This is a mind-set quite different from the "old" one, where one would first formulate a problem, even just out of curiosity, then formulate questions that arise naturally and which one would like to see answered, and then use the appropriate function spaces (if appropriate) to describe the solution, or provide new constructions which may allow to answer the pertinent question in a better or more natural way. In order to make the point let me compare the situation with the situation with *car industry*!

Comparison with car industry

I claim that the history of *function spaces* is - in some sense - comparable with the history of cars[2]. Let me try to explain in which sense one can establish parallel developments in these two apparently different areas of human activity!

First one is glad to work with concrete objects, even before the terminology is established. This means that people are able to build cars which are really *automobiles* in the sense of being able to move autonomously thanks to a motor, driven by gasoline (for example). I want to compare this with the following situation in Fourier analysis: after a long development (starting with J.P. Fourier's courageous statement concerning Fourier series) H. Lebesgue has given the community the correct spaces $\left(\boldsymbol{L}^p(\mathbb{R}^d), \|\cdot\|_p\right)$, for $1 \leq p \leq \infty$, for a proper treatment of Fourier integrals.

Then follows the phase where it was clear what kind of objects one is looking for, and different inventors developed objects of the kind under consideration which have different strengths and weaknesses. We have a *market* of cars and different

[2] Of course one can replace cars by "computers" or other objects of daily life, such as "washing machines" or "HiFi radio devices". Equally well one could choose other scientific topics, different from *function spaces* and still have a very similar situation, the key problem of the information age: Find the relevant information value (= added value for the customer) in the collection of data provided by the producers.

companies praise their products, try to build good and fashionable cars, tell the users about their great innovations and the advantages which their own brand is providing to the customers. On the other hand, it becomes more and more clear in this period what the things are which can be used in the competitive process: Maximal speed, horse-powers, acceleration parameters, gasoline consumption, trunk-volume, reliability, look, and so on. Such criteria are then found in the catalogues and shown in advertisement and are supposed to help the customer to choose in a market of abundance.

We are probably now in a comparable stage concerning function spaces. Individual results show that one space is strictly larger than another one, therefore extending an operator from the small to the large space (or in the opposite direction, demonstrating that the range is in fact in the smaller of two spaces, and not in the larger) is really logically speaking an improvement. Yes, a car with more horse powers may really be better when it comes to overtake a heavy track on a cross-country road. But would we necessarily pay a higher price for this, would we accept the increased gasoline consumption? The reader may anticipate that at this moment we are coming to arguments rarely asked in mathematical analysis, but rather familiar from *numerical analysis*. It is this area, where one asks not only for the efficiency of an algorithm (i.e., gain per iteration), but naturally compares it with the *computational costs*, the *memory requirements*. The suitability of a given algorithm will finally depend on the problem type, the computational environment and other, potentially more subjective parameters, like availability of code, familiarity with the method, and other more psychological aspects.

To summarize, concerning function spaces we are at the age of informal wisdom in the community, which spaces could be good for which kind of applications. Producers are offering their more and more sophisticated function spaces, demonstrating their "power" or "strength" through concrete examples, often the traditional ones. In a way it is a fairground, and journals are the market place to show their achievements to the public.

Up to this point the development of car industry and function space development appears to be comparable to me, and concerning function spaces it is the status quo. But the story **concerning cars meanwhile went much further!** With time it became clear, what kind of criteria a "good car" has to satisfy (low gasoline consumption, many horse-powers, spacious trunk, good price or whatever else), and specialized journals started to offer reports, comparing the different new car models. They follow more or less standardized tests and combine them with "personal impressions". In many cases the test criteria are well described or at least they are known to car industry (and to the *educated costumer*, consulting such reports).

In contrast, despite the great variety of function spaces on the "market", and the almost industrial production of new spaces we lack a similar guidance, and even evaluation criteria. We believe that it is time *to start this discussion*, even if it may be controversial at times. In any case it will provide valuable insight into what "users expect". The final result of a probably longer discussion process will be the identification of the "unique best function space", but - absolutely comparable with a good consumer report - guidelines which allow to identify those function spaces which are likely to be useful on a concrete context.

Having worked myself long enough as a pure mathematician (also with some function spaces on the records, [47–49, 53, 63]) I have nothing against the invention of new function spaces, but - if possible - one should see a *chance* for possible usefulness. It does not have to be motivated by real-world applications, but "l'art pour l'art" may be only interesting if it exhibits unexpected facts, incredible counterexamples, or warnings, that one should not try to prove *impossible results*.

Gabor Analysis and Wavelet Bases

An import "impossibility result" that comes to mind is the well-known *Balian-Low* theorem, which tells us that it is impossible to obtain a *Gaborian Riesz basis*, i.e. a Riesz basis for $\boldsymbol{L}^2(\mathbb{R}^d)$ which is of the form $(\pi(\lambda)g)_{\lambda\in\Lambda}$, where $\Lambda \lhd \mathbb{R}^d \times \widehat{\mathbb{R}}^d$ and $g \in \mathscr{S}(\mathbb{R}^d)$ (or even $g \in \boldsymbol{S}_0(\mathbb{R}^d)$). There exists meanwhile a substantial body of literature concerning this specific result (see, e.g., [11, 13, 14, 94, 126]). On the one hand, it stops of course the search for (orthonormal or Riesz) bases of $\boldsymbol{L}^2(\mathbb{R}^d)$ of Gaborian type, while, on the other hand, it forces the community to look out for good Gabor systems (because they are considered valuable due to the interpretation of the coefficients arising from Gabor expansions) *which have one of the two properties which a basis has* (think of linear algebra):

We teach in our linear algebra courses that a basis is the perfect case where a set is both linear independent *and* a generating system. Hence *every vector* in a finite dimensional Hilbert space will have a *unique* representation as a *finite linear combination* of the basis elements. Clearly one comes to a similar notion in the context of Hilbert spaces, which is exactly the concept of *Riesz bases*. One way of thinking of Riesz basis is to view them as *distorted orthonormal bases*. In fact, a Riesz basis in a Hilbert space $\mathscr{H}$ is a system $(g_i)_{i\in I}$ which can be obtained by the same bounded and invertible linear mapping T from an orthonormal basis $(h_i)_{i\in I}$. Any Riesz basis allows for a unique set of coefficients $(c_i)_{i\in I}$ in $\boldsymbol{\ell}^2(I)$ with unconditional convergence. The mapping $f \mapsto (c_i)_{i\in I}$ is a linear mapping, realized as a family of linear (!) functionals, or in other words: there is a uniquely determined *biorthogonal system* $(\widetilde{h}_i)_{i\in I}$ such that $c_i = \langle f, \widetilde{h}_i\rangle, i \in I$.

If it is not possible to have a Riesz basis for the Hilbert space $\left(\boldsymbol{L}^2(\mathbb{R}^d), \|\cdot\|_2\right)$, then one has to give up some of the requirements. On the one hand, one may look out for potentially linearly dependent[3] *generating systems*. If one adds the aspect of *numerical stability*, one is coming to the - by now well established - concept of *frames* in Hilbert spaces. One thus has to ask whether it is possible to have *good Gabor frames* (and later on: when is a Gabor frame "better" than another one, and similar questions). The attempt to answer such questions has led to the meanwhile comprehensive theory of Gabor frames and Gabor multipliers.

[3] in a suitable sense, to be discussed separately, because it is not the classical notion of linear dependence, but rather a more functional analytic version of it: some - in fact typically any - element can be written as a well convergent *series* of the other elements of the system!

Alternatively one may give up the request to represent *all elements* of the given Hilbert space, and try to be rather content with the fact that a Gaborian *Riesz basic sequence* is a Riesz basis for a closed (but proper!) subspace of $\left(\boldsymbol{L}^2(\mathbb{R}^d), \|\cdot\|_2\right)$, let us call it $\mathscr{H}_0$. Such a situation is convenient for mobile communication. Since (modulated) Gauss functions are good (joint) *approximate eigen-vectors* for all *slowly varying* channels (as they arise in the description of mobile communication channels) such systems are in fact quite useful for applications in mobile communication. Assume that the information to be sent is coded using suitable linear combinations of such a Riesz basic sequence. Then one can hope that after application of such a linear system (typically an *underspread* operator) the series is still only slightly perturbed, such that consequently the use of a biorthogonal Riesz basic sequence (which happens again to be of Gabor type) can be used to recover the coefficients (containing the information to be transmitted).

Quite in contrast to this situation it has been possible in wavelet theory (reportedly to his own surprise, by the pioneering work of Yves Meyer) to find *orthonormal bases* of *constant shape*, which are nowadays known as *orthonormal wavelet bases* (see [114, 122, 123]). There are various reasons, why the family of wavelet orthonormal bases has gained so quickly high importance both for applications and theory. First of all one can interpret the size of a cofficient as an indicator for the amount of energy within a signal at a given (dyadic) scale and a specified location. But good wavelet bases (i.e. those having good time decay and satisfying a few moment conditions) are also well suited to characterize within the realm of tempered distributions those distributions which belong to the classical function spaces, such as Sobolev or Besov spaces (typically viewed as generalized Lipschitz spaces with respect to $\mathbf{L}p$-norms), but also Bessel potential spaces, or more generally the Triebel-Lizorkin spaces. Since these spaces have been well established tools at the time when wavelets appeared they found immediate application in many branches of analysis.

These function spaces are used nowadays in analysis, in particular within the theory of PDEs, because they appear as the "natural" method to describe the smoothness, and at first sight it is natural to associate differentiation with the loss of smoothness. Taking a second order derivative of a $C^{(k)}$-function on $\mathbb{R}^d$ of course results in a $C^{(k-2)}$-function. But when it comes to the discussion of pseudo-differential operators sooner or later the use of the classical function spaces turns out to have its limitations, and therefore other/new function spaces are needed. And not always is the link to classical differentiability the most important aspect (although this claim is certainly a potentially controversial one).

However, the recent development in time-frequency analysis, including the discussion of pseudo-differential operators and Fourier integral operators indicates that also in this context modulation spaces might be the more natural spaces [30, 32–34, 90, 92, 93, 102, 136].

Consumer Reports Needed

As in the world of consumer goods, where customers are eager to compare the quality of services or the quality of products, certain *standards* have to be established. *Hotels* have their stars, *Vacation on the Farm* has its flowers, and restaurants have their Guide Michelin stars. It is part of the deal that it is generally agreed on the level of service which is guaranteed by a certain number of "stars", so that customers have a good idea of what they can expect. Despite the large variety of offers found at different places one can be assured that a high ranking means "highly recommendable", but one will not be surprised if it is also very expensive.

Since we want to discuss the *quality* or *level of information* provided by the membership of a function or distribution in a certain function space we will first require the development of widely accepted criteria for corresponding statements. The final list is probably the outcome of the longer discussion within the community, most likely via internet, at conferences or in the common room of a mathematical faculty anywhere in the world.

Note that we are at a very *early phase of this development* and the community has not given much thought to the formulation of such criteria[4], but it is clear that a list (possibly far too short at the moment) of properties are likely to play a role.

One has to say that "as of now" the academic system has not yet established a system of setting standards in the sense of *quality criteria* that a good "consumer report" should satisfy. Since production of new results and originality are these day the basis for a successful academic career the meaningful collection of information is not properly recognized as a valuable contribution to the advancement of (mathematical) sciences. Nevertheless I am convinced that it will become more and more important in the coming years.

While cooperative research is getting more and more important, and the share of single author papers is decreasing (see T. Tao's talk, cf. below) we still put "originality" ahead of "usefulness" when we have to judge papers as reviewers. This is - in the long term - a potentially critical aspect of academic life, which certainly has not taken into account the current mass-production of "scientific results".

It is certainly not yet easy nowadays to publish a paper or even a report just comparing the results obtained by a group of authors, and providing a comparative description of their relevance. *Although it might be quite informative for young researchers in the field to understand the state of the art before starting their own research*, the effort going into the establishment of knowledge of this kind is usually an *individual effort*, and the results are most of the time not well documented except for a short list of references referring to earlier contributions.

We suggest to view this article only as an incentive to the community to work towards such a slit. If we want to have an idea how it could look like, we only have to look at modern consumer reports or feedback systems in the internet, which help users to identify the product optimally matched to the individual need.

[4] This challenge is one of the main reasons for the formulation of this chapter!

Although clearly any envisaged descriptive system which may convey implicit recommendations about the usefulness and information content provided by the various function spaces will contain some partial ordering, it is a much more complex framework. Ideally one would like to provide pre-formulated settings for certain groups of users, such as "engineering students" or "researchers interested in pseudo-differential operators", in the same way as hotel-guides allow to optimize for the needs of "families with small children" or "young female individual travelers". In this sense Hotel booking systems are nowadays much better suited to describe what this author has in mind. One can look up at which price one can get which service at a given time and location, and choose a location which is well suited for a family but maybe completely un-interesting for a young couple, even if it is well priced.

Remember that it is easy to sort cars by horse-powers, the total weight, or just the price, but this is not at all a good sorting criterion for choosing a car. We also went beyond the time where a son used to stay with the brand which his father liked, but in science it appears that the tradition of staying with the function spaces suggested by the advisors dictates to a surprisingly large extent which function spaces the next generation is using. It might be an interesting topic for a psychological study to find out whether it is the primarily the familiarity with the technical details of the subject, or the understanding that the questions related to a particular area are of specific relevance. In any case, however, the questions suggested above (concerning potential outside relevance, usefulness of the achieved results, etc.) should be allowed, should be discussed and inspire a critical and constructive discussion process.

As we tried to explain the local rankings will be only part of an overall and more complex *expert system*, which is based on the particular observations. This *bigger picture* will need many of the results established in the last decades, but should go far beyond it.

In addition to the aspect of clarification of concepts and the gain of better insight into the "information content of function spaces" in a given context the consideration indicated above will have an additional effect. The *market* will react on *needs of the costumers*. Providers delivering better quality at a lower price (this could be easy to use function spaces with a wide range of applications) would find their *products more and more widely used*, whereas fancy offers would have to be really *interesting or really different from existing ones* in order to be attractive to the users. No bank would support the building of a new *standard hotel* in a *bad neighborhood*, because it might not be accepted by the guests and may end up in financial disaster.

To summarize this comparison with the hotel business: We can expect that a good hotel in an attractive place will be widely accepted, we also can hope that a very nice hotel in a not-yet attractive neighborhood might be a good idea, if it can offer clean rooms at a fair rate. But a standard hotel in a bad place may fail to find customers. But many - especially younger researchers - are in the danger of trying to build a new hotel, according to standard arguments, and hope to publish results just based on the fact that such results are (i) *new* (in the sense of not found in the literature so far) and (ii) that they are *technically feasible*, without any motivation, with very little discussion in which sense the results are a true extension (*in the conceptual or methodological sense*) of existing results, not only in a formal, logical sense.

Towards a Ranking of Spaces

Thinking in general terms of function spaces I found that the following list of properties might play a role for the ranking:

Property	Explanation: how it applies to function spaces
Size of space	size of space (big or small);
Universality	it can be used in *many* situations;
Easy to Use	it is *easy* to use;
Standard-Compare	it compares *well* with standard spaces;
Auxiliary	it is a good intermediate/auxiliary space;
Needed	it is exactly the right space for *some applications;*
Family	it belongs to an important family of spaces;
Variety	it allows for a *large variety* of equivalent descriptions;

Let us just illustrate some of the above-mentioned points, including Pros and Cons: For many of these properties there is no general claim which of these properties is important or which version of the property is "more useful". This clearly depends on the context. For an operator it is interesting to be able to show that its domain is *big*. If one has verified a few properties of a function, then the possibility of concluding that it must belong to the smaller of two spaces is clearly a stronger statement.

If a linear operator maps two spaces into themselves, one cannot say, which of these statements is the stronger one. It may be of interest to demonstrate that the Fourier transform is even well defined on the space $\mathscr{S}'(\mathbb{R}^d)$ of tempered distributions, while, on the other hand, (i.e., that the extended Fourier transform has a *large domain*), but if $\sigma \in \boldsymbol{S}_0'(\mathbb{R}^d) \subsetneq \mathscr{S}'(\mathbb{R}^d)$ it is a stronger claim which says that its Fourier transform $\hat{\sigma}$ belongs to $\boldsymbol{S}_0'(\mathbb{R}^d)$ as well.

Knowledge about the membership in one of two function spaces $\boldsymbol{B}^1$ and $\boldsymbol{B}^2$ may be considered to be more informative if one of these spaces is smaller, say it appears as more valuable to know that $f \in \boldsymbol{B}^2$ if $\boldsymbol{B}^2 \subseteq \boldsymbol{B}^1$. But if this extra information is not used afterwards, and the verification of the claim "$f \in \boldsymbol{B}^1$" is much easier to establish, it may still be preferable to use $\boldsymbol{B}^1$, e.g. in order to have an elegant and simple proof.

Even if there would be a perfect space for each problem arising in analysis, it would not be very smart (nor feasible) to really use too many of those spaces, because it makes things very complicated, and only experts can handle this complex situation. Instead, a small selection, typical for various application areas and with different functionality is what we look for.

Right now there are certainly some well-established spaces (such as Lebesgue or Sobolev spaces), so together with any new space it is advisable to list not only its properties, but also to indicate how it compares to those standard spaces. To investigate the boundedness of linear operators on $\boldsymbol{L}^p$-spaces appears as a natural "first question", even if only rarely the usefulness of such theorems is discussed in a critical way.

It may be also interesting to know, which role a particular function space has for a given area of analysis. Having no other choice than using a function space for the description of a given important operator (a good example is certainly the real Hardy space with respect to the Hardy-Littlewood maximal function) is certainly a very strong reason to have some general knowledge about such a space, even if it is not used regularly.

A space can have a large range of applications and can therefore be extremely useful if it has a *variety of equivalent characterizations*. Each of them may be useful in a different context, but if one knows that they are all equivalent one can use the same space for many applications.

Construction Principles

While the topics mentioned above indicate why certain spaces could be in use (e.g., because they are universally applicable and easy to use) the listing above does not take into account *how those function spaces are constituted*. Obviously the history of *construction principles* would be another interesting topic, which we do not want to fully expand here. Nevertheless we believe that there are some fundamental principles which come up regularly, while others are only relevant for very specific situations. Most of the time new constructions are just a simple recombination of known construction principles, which may lead to almost useless spaces and possibly also to an elegant new approach to known or stronger results. It is also clear that a *good understanding of construction principles and their mutual compatibility* may help to find orientation within the extended family of function spaces.

Again we cannot try more than just giving a selection of the most important methods:

1. *Rearrangement invariant function spaces* (can be defined over general measure spaces) and allow to create Banach spaces of functions described by the distribution of their values (how large is the set of points for which a given function $|f|$ takes values larger than some positive value $\alpha > 0$. The classical $\boldsymbol{L}^p$-spaces are the prototypes of this family, but also Lorentz- and Orlicz spaces (see [124, 131, 132]) belong to this class; such spaces are also *solid*, i.e. satisfy:
$$f \in \boldsymbol{B}, |g(x)| \leq |f(x)| \; a.e. \quad \Rightarrow \quad g \in \boldsymbol{B} \quad \text{and} \quad \|g\|_{\boldsymbol{B}} \leq \|f\|_{\boldsymbol{B}}.$$
2. *Weighted spaces* $\boldsymbol{B}_w := \{f \,|\, fw \in \boldsymbol{B}\}$ or obtained from solid spaces, using a strictly positive weight function w on the same domain; among those weight functions sub-multiplicative or moderate weights are the most important ones (see [44, 91]);
3. *Mixed norm spaces* (the prototypical construction being described in the paper by Benedek-Panzone, [12]); clearly here the order in which the various p-norms are used is important for the spaces (and can be seen as responsible for the difference between Besov and Triebel-Lizorkin spaces);

4. Another variant of $\boldsymbol{L}^p$-theory is the theory of $\boldsymbol{L}^p$-spaces with *variable exponent*, see [39, 40]. One has seen a number of classical results being adapted to this case in the last decade, but it may be difficult to appreciate the information coming the membership on one of these spaces. An additional complication might be their high sensitivity (unlike $\boldsymbol{L}^p$-spaces, which are stable under any measure preserving transformations) to small perturbations, even simple translations.
5. *Domains* of (unbounded) operators, such as the classical *Sobolev spaces*: given an unbounded operator one has most of the time a "natural domain" for such an operator (e.g., differentiation in $\boldsymbol{L}^2(\mathbb{R}^d)$);
6. Smoothness described by some *modulus of continuity*: this is the original way of introducing the *Besov spaces* as *generalized Lipschitz spaces* (with respect to $\boldsymbol{L}^p$-norms);
7. Real Hardy spaces $\left(\boldsymbol{H}^1(\mathbb{R}^d), \|\cdot\|_{\mathscr{H}}\right)$ (see [28]) are the prototypical examples of Banach spaces characterized by *atomic decompositions* of their elements. Many "minimal spaces" share this property (see, e.g., [6, 50, 121]) such spaces are often useful in order to establish boundedness properties of operators. Once it is possible to verify that a linear operator T on such a space maps *atoms* into *molecules* (which in turn are typically decomposed naturally into a well convergent series of atoms) it is usually not difficult to verify that T acts boundedly on the given "atomic space".
8. Taking dual spaces often creates interesting new spaces, but they are function spaces not always; typically, one obtains *Banach spaces of generalized functions* if test functions are dense in the given Banach space [21, 138];
9. *Real and complex interpolation methods* play an important role in generating *families of function spaces* (sometimes just *scales of spaces*) from a given pair. The complex method appears to be better suited for the interpolation of Banach algebras, while the real method (most relevant are the K-method and the J-method, which are, however, equivalent); one can say that Lorentz spaces $\boldsymbol{L}(p,q)$ are obtained from the scale of $\boldsymbol{L}^p$-spaces by real interpolation methods (see [15, 16]);
10. *Approximation spaces* are characterized by the approximation error of their elements with respect to some given family of subspaces (or a sequence of compact, typically regularizing operators), see S.M. Nikolskij's book [125], or the paper by A. Pietsch [130].
11. *Decomposition spaces*[5] (with a *local* and a *global component*), among them *Wiener amalgam spaces* (see [23, 48, 82]), allow to describe a space by its global behavior of a certain local property. While Wiener amalgam spaces rely on *uniform* decompositions [48, 97] of the underlying group, the Fourier characterization of Besov spaces uses *dyadic partitions* of unity (see the work of Frazier-Jawerth, [83, 84]). The general theory of decomposition spaces allows for a much more general covering situation, using the so-called BAPUs (*bounded, admissible partitions of unity*, see [20, 62]). The important property

[5] Not to be confused with spaces with atomic decompositions: in the current setting decompositions are domain decompositions which are used to cut the given function into "pieces".

of such partitions is the fact that every element overlaps only with a finite maximal number of "neighbors". Corresponding weights are then called moderate if they are more or less constant over such clusters (of neighbors), a rather general approach is described in [62]. The so-called α-modulation spaces [87] are just a concrete class within this general context; the so-called Herz-spaces are special decomposition spaces with local $\boldsymbol{L}^p$-components over dyadic intervals [98, 105].

12. *Banach module constructions* [21] allow to create "essential parts" and "relative completions" of Banach modules. In harmonic analysis *homogeneous Banach spaces* (as defined by Y. Katznelson in [109], see also [134]) are the prototypical examples of this kind, with *Segal algebra* in the sense of Reiter as a subclass (those which are inside of $\left(\boldsymbol{L}^1(\mathscr{G}), \|\cdot\|_1\right)$). By combining two module actions (e.g., convolution and pointwise multiplication) one can create from a given space a variety of up to six spaces, all with the same norm! Details are given in [21].
13. *Tensor product constructions* have a long tradition in harmonic analysis, especially in connection with the Eymard's Fourier algebras [41] $\boldsymbol{A}(\mathscr{G})$ and (for $p \neq 2$) the Figa-Talamanca-Herz Banach algebras (pointwise) $\boldsymbol{A}_p(\mathscr{G})$ defined over general locally compact groups (see [38, 80, 98, 140]). They are intimately related to the study of convolution operators on $\boldsymbol{L}^p$-spaces;
14. *Isomorphic images* of given Banach spaces inherit of course most of their abstract properties. Sometimes the isomorphism in use moves a problem into a more transparent setting. For example, the transform may turn an abstract Hilbert space into a RKH (*reproducing kernel Hilbert space*, [10]). The best way to understand (fractional) Sobolev spaces is to identify them (via the *Fourier transform*, which is an isometric isomorphism on $\left(\boldsymbol{L}^2(\mathbb{R}^d), \|\cdot\|_2\right)$, thanks to Plancherel's theorem) with a corresponding weighted $\boldsymbol{L}^2$-space over the dual group (the Fourier domain, $\widehat{\mathbb{R}^d}$), see [135]; in a similar way, the space $\mathscr{F}\boldsymbol{L}^p$ is just the image of $\left(\boldsymbol{L}^p(\mathbb{R}^d), \|\cdot\|_p\right)$ under the (distributional) Fourier transform, with the norm inherited from the $\boldsymbol{L}^p$ (over the "time-domain");
15. A very similar idea can be carried out with, e.g., *spaces of multipliers*, or sometimes the so-called Fourier multipliers. In the classical setting (see, for example, [113]) $\boldsymbol{L}^p$-Fourier multipliers are just pointwise multipliers of $\mathscr{F}\boldsymbol{L}^p(\mathbb{R}^d)$. But of course one can view them also as those tempered distributions which define *convolution operators* on $\left(\boldsymbol{L}^p(\mathbb{R}^d), \|\cdot\|_p\right)$ (with the operator norm as norm);
16. *Traces* of given function spaces to lower dimensional subspaces are also often an interesting source for new spaces; in many cases (e.g., Besov or modulation spaces), they belong to the same family of spaces (and this is then an interesting property of the family), but it is not always true.
17. *Coorbit spaces* have been developed in a series of papers [63, 64, 88]. The elements of such spaces are characterized by a typical behavior (described through some solid, translation invariant Banach space of function on the group $\mathscr{G}$) of the representation coefficients of its elements for a suitable group representation π of $\mathscr{G}$ on some Hilbert space. These representation coefficients are also called *voice transforms* or (generalized) continuous *wavelet transform* and are

taken with respect to suitably chosen "mother wavelets". Up to this (typically isometric) transfer coorbit spaces are thus just isomorphic copies of closed (and left-invariant) function spaces on the group $\mathscr{G}$.

At the beginning coorbit theory was more about finding a unified approach to the continuous wavelet transform and the STFT, using group theoretical methods. It also provided ways to establish by analogy properties of *Moebius invariant Banach spaces of analytic functions* (see [7–9]). However, more recently, other groups and other function spaces (also allowing *atomic decompositions*, frames, etc.) have obtained some attention, among them function spaces related to the Blaschke group (see [127, 128]) and in particular the theory of *shearlets* [35, 36, 112].

18. *Intersections* (of finitely or even infinite, e.g. parameterized families of spaces, see [17–19]), or correspondingly the *linear sum* of two (or more) function spaces (see [116]) may also generate new spaces; often even the intersection of two members of a family of spaces is not in the same family of spaces (e.g., the intersection of two $\boldsymbol{L}^p$-spaces, or two Wiener amalgam spaces with different local resp. global components: one space may have a smaller local component, the other one a smaller global component). A systematic study of such *lattices* of spaces is undertaken in [5].

    Another situation where the intersection of $\boldsymbol{L}^p$ with its multiplier space gives an interesting new space (called the *tempered elements*) is given in [42];
19. Given a Banach spaces (such as a weighted $\boldsymbol{L}^p$-space $\boldsymbol{L}^p_w(\mathbb{R}^d)$) one may ask whether there is a *smallest space* containing the given space, but which has better translation properties, e.g. (to take the most simple case) is isometrically invariant under translation (see, e.g., [106] [43, 47]);
20. In a similar way one can look for the *dilation invariant hull* of a given space, if it is not isometrically dilation invariant. The *exotic space* $\mathscr{B}_0$ described in [75] is an example obtained in this way (using $\boldsymbol{L}^2$-normalized dilation, starting from the space $\boldsymbol{S}_0(\mathbb{R}^d)$, which itself is both an atomic and an amalgam space); the observation that certain dyadic decomposition spaces can also be described as the dilation invariant hull of all functions in the unit ball of some $\boldsymbol{L}^q$-space with fixed support was the key to the proof of Wiener's Third Tauberian theorem in [51].
21. Function (resp. distribution) spaces correspond sometimes to certain Banach spaces of operators. For example, it is known via the *kernel theorem* that functions in $\boldsymbol{L}^2(\mathbb{R}^{2d})$ are exactly the integral kernels of Hilbert-Schmidt operators on $\left(\boldsymbol{L}^2(\mathbb{R}^d), \|\cdot\|_2\right)$; but one can also look at the space of kernels of operators from $\boldsymbol{L}^p(\mathbb{R}^d)$ to $\boldsymbol{L}^q(\mathbb{R}^d)$, or the space of all Weyl (or Kohn-Nirenberg) symbols of a class of operators (see [69]); more often function spaces are used in order to describe the mapping properties, e.g. in the context of *quantization*, mapping from function spaces to operator ideals;
22. Capacity based spaces are describing their elements by the *capacities of their level sets* (as opposed to the Lebesgue spaces using the measure of these sets); one may consider these spaces a somewhat less observed family of spaces ([2, 26, 110]).

23. *Quotients* and *closed* resp. *complemented subspaces* of a given family of spaces are sometimes also considered.

Out of this (probably still incomplete catalogue) of possible methods of gaining new spaces from given ones let us choose a few examples in order to explain our understanding of "usefulness" of function spaces, respectively, in order to give a concrete interpretation of the abstract principles mentioned in the section about *rankings*. We are going to talk about Wiener amalgam spaces, because they appear to be useful for *quite a wide range of problems in analysis*, but also because they are not yet so widely known or used, despite the fact that they are *easy to understand* and *easy to use*.

Wiener Amalgam Spaces

Wiener amalgam spaces (introduced as *Wiener-type* spaces in [48], and renamed following a suggestion by John Benedetto, see [97] for a gentle introduction) have been useful in a large number of places and have in some way become the backbone of many papers by the author concerning coorbit theory [63, 64], or sampling theory [54, 55, 65, 66] or spline-type space [3, 4, 56].

Wiener Amalgam Spaces require a *local component*, which can be any double module space $(\boldsymbol{B}, \|\cdot\|_{\boldsymbol{B}})$ (cf. [61]), and a description of the global behavior of the local properties of f described by means of the $\boldsymbol{B}$-norm of localized pieces of f. We restrict our attention here (for simplicity) to global $\boldsymbol{\ell}^p$-spaces.

One can define $\boldsymbol{W}(\boldsymbol{B}, \boldsymbol{\ell}^p)$ by means of a BUPU, a *uniform partition of unity*, ideally a collection of functions (ψ_i) which have uniformly small compact support (like cubic B-splines), are bounded in the multiplier algebra $\boldsymbol{A}$ of $(\boldsymbol{B}, \|\cdot\|_{\boldsymbol{B}})$, and have controlled overlap of their supports. Members of this space are functions belonging locally to $\boldsymbol{B}$, and have the following (norm) expression finite:

$$\|f|\boldsymbol{W}(\boldsymbol{B}, \boldsymbol{\ell}^p)\| := \left(\sum_{i \in I} \|f \cdot \psi_i\|_{\boldsymbol{B}}^p \right)^{1/p} < \infty.$$

Especially the choices $\boldsymbol{B} = \boldsymbol{C}_0(\mathbb{R}^d), \boldsymbol{L}^2(\mathbb{R}^d)$ and $\mathscr{F}\boldsymbol{L}^1(\mathbb{R}^d)$ are useful. We have $\boldsymbol{S}_0(\mathbb{R}^d) = \boldsymbol{W}(\mathscr{F}\boldsymbol{L}^1, \boldsymbol{\ell}^1)$ and $\boldsymbol{S}_0'(\mathbb{R}^d) = \boldsymbol{W}(\mathscr{F}\boldsymbol{L}^\infty, \boldsymbol{\ell}^\infty)(\mathbb{R}^d)$.

Previously the corresponding spaces (going in fact back to work of N. Wiener, e.g. in the context of his book [142]) had been used just with local components of the form $\boldsymbol{L}^r$, for some $r \in [1, \infty]$, see, e.g., [23, 82, 101].

When $\boldsymbol{L}^p$-spaces over $\mathbb{R}^d$ are compared the following question arises:
Is it more informative to know that $f \in \boldsymbol{L}^r$ or to know $f \in \boldsymbol{L}^s$, for $s \neq r$, say $r < s$?

The answer is of course: this depends on side information that one may have. If f is compactly supported, $f \in \boldsymbol{L}^s$ is the stronger claim (any bounded function is integrable, square integrable function as well, etc.). On the other hand, if f is band-limited, $f \in \boldsymbol{L}^r$ is the stronger statement, because only the *global* properties

matter. Certainly such questions can be better answered by means of Wiener amalgam spaces, which allow to separate local and global properties (resp. obstacles for inclusion results).

Standard reference: [43, 48] for Wiener's algebra resp. Banach conv. algebras.

Wiener amalgam spaces arose as a technique which was supposed to imitate the construction of Besov spaces in the context of LCA groups (such as $\mathscr{G} = \mathbb{R}^d$), but without the use of dilation. They were designed having already the subsequent definition of modulation spaces in mind (as inverse images of Wiener amalgams). From the various descriptions of Besov spaces available at that time the characterization using dyadic partitions of unity (related to Paley-Littlewood theory) it became soon clear that one needs smooth partitions of unity, which have to be uniformly bounded in the *Fourier algebra* $\left(\mathscr{F}\boldsymbol{L}^1(\mathbb{R}^d), \|\cdot\|_{\mathscr{F}\boldsymbol{L}^1}\right)$, and so eventually it turned out that so-called BUPUs $\Psi = (\psi_i)_{i\in I}$ (*Uniform Bounded Partitions of Unity*) are the right way to go. Here uniformity refers to the size of the support of the building blocks ψ_i, while boundedness refers to boundedness in the Fourier algebra $\left(\mathscr{F}\boldsymbol{L}^1(\mathbb{R}^d), \|\cdot\|_{\mathscr{F}\boldsymbol{L}^1}\right)$.

The model for these spaces have been the (ordinary) *amalgam spaces* $\boldsymbol{\ell}^q(\boldsymbol{L}^p)$ (generalizing Wiener's construction) where a kind of trivial (regular) decomposition of $\mathbb{R}^d$ could be used. In principle this two-parameter family of spaces uses *local* $\boldsymbol{L}^p$-norms with global restrictions in the form of an outer $\boldsymbol{\ell}^q$-summability. In this way one has locally the natural inclusions known from the compact case (e.g., the torus group, with $\boldsymbol{L}^{p_1}(\mathbb{U}) \subseteq \boldsymbol{L}^{p_2}(\mathbb{U})$ if and only if $p_1 \geq p_2$, while the opposite inclusion (same as for the $\boldsymbol{\ell}^q$-spaces over $\mathbb{Z}$) is valid in the *global components*.

Product-convolution operators [23] have motivated the corresponding (and highly useful) convolution theorem given in [48], showing that local and global convolution result can be combined to convolution results for Wiener Amalgam spaces. This in turn is quite useful in order to derive good properties of spaces of smooth functions. There are also natural results concerning interpolation of amalgam spaces [46].

On the one hand, one can obtain Plancherel-Polya type theorems for band-limited functions in $\boldsymbol{L}^p$-spaces, on the other hand results of the type of a Sobolev embedding.

In the first case we argue that any band-limited function f with $\operatorname{spec}(f) = \operatorname{supp}(\hat{f}) \subseteq \boldsymbol{\Omega}$ allows to have some function $\boldsymbol{W}(\boldsymbol{C}_0, \ell^1)(\mathbb{R}^d)$ with $\hat{h}(\omega) = 1$ for $\omega \in \boldsymbol{\Omega}$, hence

$$h = h * f \in \boldsymbol{L}^p * \boldsymbol{W}(\boldsymbol{C}_0, \ell^1) \subset \boldsymbol{W}(\boldsymbol{L}^1, \boldsymbol{\ell}^p) * \boldsymbol{W}(\boldsymbol{C}_0, \ell^1) \subset \boldsymbol{W}(\boldsymbol{C}_0, \boldsymbol{\ell}^p).$$

On the other hand, we have for the case that $1/w \in \boldsymbol{L}^2(\mathbb{R}^d)$ that

$$\boldsymbol{L}^2_w(\mathbb{R}^d) = \boldsymbol{W}(\boldsymbol{L}^2, \ell^2_w) \subseteq \boldsymbol{W}(\boldsymbol{L}^2, \boldsymbol{\ell}^1),$$

and hence one has for the corresponding Sobolev space $\boldsymbol{H}_w(\mathbb{R}^d) := \mathscr{F}^{-1}(\boldsymbol{L}^2_w(\mathbb{R}^d)) \subset \boldsymbol{W}(\mathscr{F}\boldsymbol{L}^1, \boldsymbol{\ell}^2)$, due to the Wiener amalgam version of the Hausdorff-Young theorem [53]. It follows therefrom that we have the continuous embedding (under the usual assumption of Sobolev's embedding) $\boldsymbol{H}_w(\mathbb{R}^d) \hookrightarrow \boldsymbol{W}(\boldsymbol{C}_0, \ell^2)(\mathbb{R}^d)$, and consequently one can estimate for any lattice $\Lambda \lhd \mathbb{R}^d$ (up to some constant $C > 0$ depending only on w and the lattice)

$$\left(\sum_{\lambda \in \Lambda} |f(\lambda)|^2 \right)^{1/2} \leq C \|f\|_{\boldsymbol{H}_w} \quad \forall f \in \boldsymbol{H}_w.$$

Similar estimates hold for "irregular" discrete sets, as long as they are *well-spread*, i.e. as long as they are finite unions of δ-separated sets. And of course such results imply that analogous estimates hold true for band-limited functions (see also [73]). Since such a claim cannot be deduced from the standard inclusion (Sobolev embedding theorem) stating $\boldsymbol{H}_w \subset \boldsymbol{L}^2 \cap \boldsymbol{C}_0$ this is a more informative statement.

Wiener amalgam spaces over locally compact groups (allowing an integrable representation π) have been the technical cornerstone of *coorbit theory* (see [63, 64, 88]) but also the decisive tool for the analysis of *iterative reconstruction methods for irregular sampling problems* for band-limited or spline-type functions [4, 54, 66, 67, 72, 76].

Outlining a New Approach to Fourier Analysis

If one looks into the (many) classical books about Fourier analysis which have been published in the last centenary, one finds only relatively little differences. This is perhaps not surprising, since *Fourier analysis* is by now a very mature field, with well-established basic principles, with the "right" function spaces being well known mostly in the form of *Lebesgues spaces*). There are also various generalizations which make Fourier analysis as such also a very important part of the modern theory of PDE (recalling, e.g., the concept of *microlocal analysis* as introduced by L. Hörmander) or for the theory of pseudo-differential operators. It has become common practice to ask about the boundedness of linear operators between (resp. on) $\boldsymbol{L}^p$-spaces, and only the study of Calderón- Zygmund and related operators (like the Hilbert transform) brought the insight that at least for the limiting cases (which are $\boldsymbol{L}^1(\mathbb{R}^d)$ and $\boldsymbol{L}^\infty(\mathbb{R}^d)$) one should seek appropriate replacements, namely the *real Hardy space* $\boldsymbol{H}^1(\mathbb{R}^d)$ its dual, the $\boldsymbol{BMO}(\mathbb{R}^d)$-space. For details see the seminal paper [28].

The Traditional Approach to Fourier Analysis

Most of the current literature on Fourier analysis takes for granted that the *natural domain for the Fourier transform* are simply the Lebesgue spaces $\left(\boldsymbol{L}^1(\mathscr{G}), \|\cdot\|_1\right)$ of Lebesgue integrable functions on a group $\mathscr{G}$. This despite the fact that one has to introduce first the Haar measure on $\mathscr{G}$ and then a measure space, whereas bounded measures could be easily introduced as the dual space of $\boldsymbol{C}_0(\mathscr{G})$.

This (by now widely accepted!) viewpoint is based on the observation that the Fourier transform - as originally defined - is coming up as an integral transform, and the historical development has shown that obviously the Lebesgue integral (on

general LCA groups the Haar measure) is the right tool. It does not only provide a proper domain to the Fourier transform, but it also ensures that $\left(\boldsymbol{L}^1(\mathscr{G}), \|\cdot\|_1\right)$ is a Banach space (in contrast to the space of Riemann integrable functions). Furthermore, the *Lemma of Riemann-Lebesgue* allows to describe the FT as non-expansive mapping from $\left(\boldsymbol{L}^1(\mathscr{G}), \|\cdot\|_1\right)$ into $\boldsymbol{C}_0(\widehat{\mathscr{G}})$, endowed with the sup-norm $\|\hat{f}\|_\infty$, since $\|\hat{f}\|_\infty \leq \|f\|_1$ for any $f \in \boldsymbol{L}^1(\mathscr{G})$. It is even a very convenient setting to prove the so-called *convolution theorem*, because it only requires to make proper use of Fubini's Theorem in order to find out that convolution, given by

$$f * g(x) := \int_G g(x-y) f(y) dy \quad \forall f, g \in \boldsymbol{L}^1(\mathscr{G}), \tag{1}$$

is well defined (almost everywhere), satisfies the *submultiplicativity property*

$$\|f * g\|_1 \leq \|f\|_1 \|g\|_1 \quad f, g \in \boldsymbol{L}^1(\mathscr{G}).$$

On the other hand, Fubini's Theorem is also the key ingredient for the proof of the so-called *convolution theorem*, stating that

$$\mathscr{F}(f * g) = \mathscr{F}(f) \cdot \mathscr{F}(g) \quad \forall f, g \in \boldsymbol{L}^1(\mathscr{G}). \tag{2}$$

For many applications this is a good choice, but there come also several restrictions with this approach. First of all it makes the Fourier transform, although it can be extended to a unitary operator on $\left(\boldsymbol{L}^2(\mathscr{G}), \|\cdot\|_2\right)$ (Plancherel's theorem) not symmetric, because the domain of the inverse Fourier transform, which should be of course $\mathscr{F}\boldsymbol{L}^1(\widehat{\mathscr{G}}) := \{\hat{f} \,|\, f \in \boldsymbol{L}^1(\mathscr{G})\}$, is *not in general contained* in $\boldsymbol{L}^1(\widehat{\mathscr{G}})$, and hence the inverse Fourier transform (although it has almost the same integral kernel as the forward FT) has problems to be realized as a true Lebesgue integral, despite the good properties of this integral. In fact, even for $\boldsymbol{L}^2(\mathbb{R}^d)$ the realization of the Fourier-Plancherel Transform requires an approximation argument and cannot be carried out as a "literal integral transform".

There is another important area where convolution and Fourier transforms play a role: the theory of *translation invariant linear systems T* (TILS). Here one expects (looking at the discrete situation) that every such operator is a convolution operator with some impulse response (or mathematically speaking convolution kernel) or equivalently, that such systems can be described as Fourier multipliers, or a so-called *transfer function* $h(\omega)$, which can be any bounded function, in fact an arbitrary element from $\boldsymbol{L}^\infty(\widehat{\mathscr{G}})$. Thus T is either of the form $T(f) = \sigma * f$ (for "some" impulse response function σ (thought as $\sigma = T(\delta_0)$ whenever it makes sense) or equivalently (in which sense) $\mathscr{F}(T(f)) = h \cdot \hat{f}$.

Even for simple cases one can see the limitation of the viewpoint restricting the attention to Lebesgue integrals: Given $\chi(x) = e^{ix^2}$ (a so-called chirp signal) on $\mathbb{R}$ one can show that the operator, well defined on functions with compact support, extends in a unique way to bounded (in fact unitary) convolution operator on $\boldsymbol{L}^2(\mathbb{R})$, with transfer $h = \mathscr{F}(\chi) = \chi$ (see [25]). But for a function such as the well-known SINC-function (inverse Fourier transform of a box-function) the convolution $\chi * \mathrm{sinc}$

cannot be written as a convolution integral, and also the claim the chirp function is invariant under the Fourier transform cannot be shown if one is restricted to the Lebesgue viewpoint.

Although detailed investigations of the classical Fourier transform are a valuable contribution to mathematical analysis, one has to see that the modeling of real-world problems must not make unnatural restrictions which are purely based on the method that mathematicians would like to use. So we come to the conclusion that a more general perspective on the Fourier transform has to be taken, in order to give a good and clear meaning of the ingredients needed for a proper description.

What is Fourier Analysis all about?

Even if it may be considered a strange question in this context let us nevertheless ask the question: What are the main applications of Fourier analysis? Is it just an exciting chapter of mathematical analysis, or does it have applications in the real world?

Of course the answers to this question will depend very much on the individuals providing the answer, and thus be to some extent subjective answers. *But* there are some common reasons, why it is good to have the Fourier transform, and why it is playing a natural role in many places.

First of all one should recognize that any linear system (linear mapping between function spaces on a group) which commutes with translations can be thought of as a convolution operator, at least in a sufficiently general context. In the engineering literature, or properly speaking in the context of *discrete* groups one can define the *impulse response* of the system T as the output which is generated from the unit vector at the neutral element (the Dirac measure at zero, if you want so). By making use of the translation invariance of the system (often justified by the time-invariance of physical laws) one can then show that the action of T on a general input signal can be described as a convolution of the input signal with that impulse response function. Especially when it comes to the question of inversion of such an operator (or other symbolic operators, such as the square root) it is important to know that the Fourier transform is *diagonalizing* all those operators and consequently inversion is easy via pointwise inversion on the Fourier transform side. In fact, the composition of systems corresponds to pointwise multiplication of the corresponding transfer functions, which can be defined as the Fourier transforms of the individual impulse response objects (functions, measures, or distributions).

Classical Fourier Analysis and L^p-spaces

In this section we are taking a short "tour d'horizon" through the early history of Fourier analysis and discuss some aspects of the theory of function spaces connected with Fourier analysis. Classical Fourier analysis starts with Fourier series. Since the Fourier coefficients of a $\mathbb{Z}$-periodic function f are given by an integral of the form

$$\hat{f}[k] := \int_0^1 f(t)\overline{\chi_k(t)}\,dt = \int_0^1 f(t)e^{-2\pi ikt}\,dt \tag{3}$$

where $\chi_k(t) = e^{2\pi ikt}$ is the *pure frequency*, it is reasonable to consider the Lebesgue space $\boldsymbol{L}^1([0,1])$ or in fact better $\left(\boldsymbol{L}^1(\mathbb{U}), \|\cdot\|_1\right)$ as the natural domain for the Fourier transform, with the norm $\|f\|_1 := \int_0^1 |f(t)|dt$, and verify that

$$\|\hat{f}\|_\infty := \sup_{k\in\mathbb{Z}} |f(k)| \le \|f\|_1. \tag{4}$$

Unfortunately it is not completely clear how to come back from the Fourier coefficients to the original function $f \in \boldsymbol{L}^1(\mathbb{U})$. The fact that one has for decent functions (e.g., twice continuously differentiable ones) the possibility of *synthesizing* the function from its *pure frequencies* $\chi_k, k \in \mathbb{Z}$ (with amplitudes obtained through the *Fourier analysis* described by formula (4) above), namely

$$f(t) = \sum_{k\in\mathbb{Z}} \hat{f}[k]\,\chi_k(t), \tag{5}$$

with absolute (hence unconditional) and uniform convergence, suggested to focus on the most general circumstances where this sum is convergent (at which points t, depending on the local behavior of f near t), either directly or via the use of various kinds of sophisticated *summability procedures*.

These summability methods, associated with famous names in the history of Fourier analysis (such as Fejer, De La Vallee Poissin, Weierstrass and others) provide more stable methods of reconstruction than just the partial sums (which are fine in the $\boldsymbol{L}^2$-sense, but not for other norms!). Typically they work like this: multiply the Fourier coefficients with a family of sequences which have properties of converging to the constant $(1)_{k\in\mathbb{Z}}$ sequence, while decaying to zero in a more decent way than just the sequence which represents the indicator function of $[-n,n]\cap\mathbb{Z}$.

During this (still) early phase of Fourier analysis the discovery that one has a fairly simple situation if one views everything for the space $\boldsymbol{L}^2(\mathbb{U})$ of (locally) square integrable, periodic functions. Nowadays we would say that this is clear, because $\left(\boldsymbol{L}^2(\mathbb{U}), \|\cdot\|_2\right)$ is a Hilbert space with respect to the scalar product

$$\langle f,g\rangle_{\boldsymbol{L}^2(\mathbb{U})} := \int_0^1 f(t)\overline{g(t)}dt. \tag{6}$$

Within this *separable Hilbert space* the family $(\chi_k)_{k\in\mathbb{Z}}$ forms a *complete orthonormal basis* and therefore it is clear that one has (recalling corresponding statements

which are well known nowadays in the context of linear algebra) the following unconditional expansion:

$$f = \sum_{k\in\mathbb{Z}} \langle f, \chi_k \rangle \, \chi_k = \sum_{k\in\mathbb{Z}} \hat{f}[k] \, \chi_k \tag{7}$$

where the sum is *unconditionally*[6] convergent with respect to the norm of the Hilbert space $\left(\boldsymbol{L}^2(\mathbb{U}), \|\cdot\|_2\right)$. Due to *Euler's formula*

$$e^{ix} = \cos(x) + i\sin(x), \tag{8}$$

resp. the (equivalent) pair of equations

$$\cos(x) = (e^{ix} + e^{-ix})/2; \quad \sin(x) = (e^{ix} - e^{-ix})/(2i) \tag{9}$$

it is clear that one can rewrite Fourier series also in a more classical form, using as building blocks the system of functions $(\cos(2\pi kt))_{k\in\mathbb{N}_0}$ together with the system $(\sin(2\pi kt))_{k\in\mathbb{N}}$, with similar properties.

As it turned out the question of pointwise (obviously only to be expected in the almost everywhere sense, i.e. up to some set of measures zero, depending on f) was much harder to tackle, and was solved only half a century later by Lennart Carleson [24] in the context of $\boldsymbol{L}^2$.

Our first summary of classical Fourier analysis already indicates that the spaces $\boldsymbol{L}^1(\mathbb{U})$ and $\boldsymbol{L}^2(\mathbb{U})$ play an important role, and hence it is not surprising that the space "in between those spaces" also came quickly into the focus of people working in analysis, the so-called $\boldsymbol{L}^p$-spaces.

Obviously they are very important, among others because they have been among the first Banach spaces to be treated as such, including the characterization of the dual spaces ($\boldsymbol{L}^q$ is dual to $\boldsymbol{L}^p$ for $1 \le p < \infty$, with $1/q + 1/p = 1$), or the discussion of *reflexive* Banach spaces. In the Fourier context one has the so-called *Hausdorff-Young* inequality, which states that for $p \in [1,2]$ one can control the ℓ^q-norm (again for $1/q + 1/p = 1$) of the sequence of Fourier coefficients:

$$\|\hat{f}\|_{\ell^q(\mathbb{Z})} \le \|f\|_{\boldsymbol{L}^p(\mathbb{U})} := \left(\int_0^1 |f(t)|^p\right)^{1/p}. \tag{10}$$

Unfortunately $p = 2$ is the only case where the mapping from $\boldsymbol{L}^p(\mathbb{U})$ to $\boldsymbol{\ell}^q(\mathbb{Z})$ is surjective, while for $p > 2$ not more than square summability of $\hat{f}$ can be claimed.

Fourier Analysis over Euclidean Spaces and LCA groups

The first decades of the 20th century saw the development of Fourier analysis over the real line (non-periodic functions) and over $\mathbb{R}^d$, closely connected with the

[6] i.e., independent of the enumeration of $\mathbb{Z}$, etc.

beginning of functional analysis as we know it these days. It was this time when the Lebesgue spaces gained their importance as a well-defined family (closed under complex interpolation) of Banach spaces, which is closed under duality. The multiplier question (characterization of linear operators on $\boldsymbol{L}^p$-spaces commuting with translations) resp. of the identification of $\boldsymbol{L}^p$-Fourier multipliers were challenging problems, with ground-breaking results by Marcinkiewicz.

In the middle of the last century after Rudolf Lipschitz and Antoni Zygmund a more detailed analysis of smoothness was undertaken, nowadays associated with the names of Sergei Lwowitsch Sobolev, Oleg Besov, Sergei Michailowitsch Nikolskii, Elias Stein, Jaak Peetre [129], Hans Triebel [137, 139], and many others. Smoothness was then understood as a fine form of differentiability. Among others the expression of "generalized Lipschitz spaces" was used for spaces not known as Besov spaces. They have been characterized in many different ways, mostly using moduli of continuity, or higher order differences. Only later (and in the attempt to better understand the interpolation theoretic properties of this family) the intensive use of *dyadic Fourier decompositions* (Paley-Littlewood) theory came into place (see the books of Triebel, or the precursors of wavelet theory by Frazier-Jawerth, see [83–85]).

Once it was clear that there are lot of analogies between the classical theory of Fourier series for periodic functions (of one or several variables), or equivalently for functions defined on the torus group and the corresponding theory of functions over $\mathbb{R}^d$ it was natural to ask for the most general context in which Fourier analysis could take place.

There are essentially two ways into this direction. The book by A. Weil [141] starts from A. Haar's invariant integral for locally compact groups $\mathscr{G}$ [96], and uses the existence of sufficiently many characters (pure frequencies) for the Abelian case in order to do Fourier analysis over LCA groups $\mathscr{G}$.

Another approach comes from Gelfand's theory of commutative Banach algebras with involution, which - when applied to $\boldsymbol{L}^1(\mathscr{G})$, for some LCA group $\mathscr{G}$ - provides another abstract approach to Fourier analysis over Abelian groups.

Theory of Generalized Functions

There is no doubt that a proper treatment of the Fourier transform will require some form of distribution theory. It is fair to speak of a *theory of generalized functions*, because many manipulations which can be carried out for ordinary functions using pointwise operators (such as affine transformation of the argument, rotation, and others) have a natural extension to the setting of distributions. Even the meaning of the *support* of a distribution is well defined.

When it comes to Fourier analysis the theory of tempered distributions, introduced by L. Schwartz in ([133]) appears to be the natural one. It is also the basis for dealing with PDEs in an appropriate context, and was the foundation for the work of Lars Hörmander in this field (see, e.g., [103]).

Unfortunately it requires a bit of theory concerning topological vector spaces which is normally not studied by engineers, and therefore only the basics of this theory of generalized functions is discussed in the corresponding courses.

The attempts by Lighthill ([115]) of a simplified approach (pursued later by Jones, [107]), or the approach by Howell ([104]) have not really changed the situation.

The Banach Gelfand Triple $(\boldsymbol{S}_0, \boldsymbol{L}^2, \boldsymbol{S}_0')$ may provide a good compromise in this context as well. At least it allows to define the Fourier transform, the sampling or periodization procedures in a mathematical clean way (in the context of $\boldsymbol{S}_0'(\mathbb{R}^d)$) without reference to the theory of topological vector spaces.

Fourier Analysis and L^p-spaces

Overall the previous subsection indicates that, on the one hand, the theory of $\boldsymbol{L}^p$-spaces plays an important role for the development of functional analysis in general, and in particular for Fourier analysis, but overall the family of $\boldsymbol{L}^p$-spaces as such is not really well suited for a description of properties of the Fourier transforms in terms of properties of the function (and vice versa), through the membership in one of the spaces $\left(\boldsymbol{L}^p(\mathbb{R}^d), \|\cdot\|_p\right)$, for $1 \leq p \leq \infty$, for example.

This is in fact not surprising, because it is easy to note that the decay of the Fourier transform is (both qualitatively and quantitatively) related to the smoothness of a function. It is a simple consequence of the mean-value theorem that for a function which does not deviate much (locally) from its mean-value (due to smoothness) the integral against a highly oscillating exponential function will be close to zero. On the other hand, one can expect rightfully that a function which synthesized mostly from low frequencies (to use engineering terminology at this place) will behave like a linear combination of pure frequencies up to some maximal frequency, hence will show a lot of smoothness.

This situation also does not change too much if one allows for weighted $\boldsymbol{L}^p$-spaces (on both sides). Only in rare cases one can characterize the membership of a function in one of the smoothness spaces by the membership of its Fourier transform in such a weighted $\boldsymbol{L}^p$-space. The only really important subfamily is the family of Sobolev spaces ($p = 2$), where one has an exact match between a space (typically denoted by $(\mathscr{H}_s(\mathbb{R}^d), \|\cdot\|_{\mathscr{H}_s})$) and its Fourier image, the space $\boldsymbol{L}^2_w(\mathbb{R}^d)$, with $w(\omega) = w_s(\omega) = (1+|\omega|^2)^{s/2}$. These spaces are very useful, e.g. in the context of PDE. Moreover one can show that for $s > d/2$ they are continuously embedded into $\left(\boldsymbol{C}_0(\mathbb{R}^d), \|\cdot\|_\infty\right)$, hence point-evaluations are continuous linear functionals, i.e. the spaces $\mathscr{H}_s(\mathbb{R}^d)$ are in fact a so-called RKH (reproducing kernel Hilbert space). Furthermore, under the same condition they are Banach algebras under pointwise multiplication.

Time-Frequency Analysis and Function Spaces

Time-Frequency analysis (see [89] for a comprehensive mathematical introduction to the field) has become a flourishing branch of mathematical analysis in the last three decades. Sometimes it is described as the branch of mathematics which is based on the two basic operations arising in the context of Fourier analysis, namely the *time-* and the *frequency shifts* (which are shift operators on the FT side, resp. modulation operators, doing a multiplication with a complex exponential function, a so-called *pure frequency*). There are also obvious connections to the *Schrödinger representation* of the (reduced) *Heisenberg group*.

One of the basic objects of a subfield of analysis called "time-frequency analysis" is the **Short Time Fourier Transform** (STFT) or *sliding window Fourier transform*, which requires to choose some so-called window function g (well localized near 0), which we assume to be a bounded, continuous function with good decay, perhaps with compact support and $\|g\|_2 = 1$, in order to define the STFT $V_g(f)$, the *localized Fourier transform* or *sliding window FT* of f by the following scalar product in the $\boldsymbol{L}^2$-sense:

$$V_g(f)(t,\omega) := \langle f, M_\omega T_t g \rangle \tag{11}$$

According to *Moyal's equality* one has then

$$\|V_g(f)\|_2 = \|g\|_2 \|f\|_2 \quad \text{for} \quad f,g \in \boldsymbol{L}^2(G). \tag{12}$$

which in turn implies (under the assumption $\|g\|_2 = 1$) the weak reconstruction formula

$$f = \int_{\mathscr{G}\times\widehat{\mathscr{G}}} V_g(f)(\lambda)\pi(\lambda)g \, d\lambda. \tag{13}$$

with $\pi(\lambda)(g) = \pi(t,\omega)(g) = M_\omega T_t g$.

Although these results may give the impression that again the Hilbert space $\left(\boldsymbol{L}^2(\mathbb{R}^d), \|\cdot\|_2\right)$ is the right setting to formulate questions in time-frequency analysis, it turns out that this is more or less the only Lebesgue space which is relevant for TF-analysis. Also the first impression that (in contract to wavelet theory) no admissibility condition is required, because (13) is valid for arbitrary pairs $f,g \in \boldsymbol{L}^2(\mathscr{G})$

One of the fundamental papers in this field is due to D. Gabor [86] who claimed that "every complex-valued function" on $\mathbb{R}$ can be written as a superposition (in fact a double series) of the Gauss-function, shifted *"in time and frequency"* along the integer lattice $\Lambda = \mathbb{Z}\times\mathbb{Z}$. Nowadays so-called *atomic decompositions* of this form, i.e. of the form

$$f = \sum_{\lambda\in\Lambda} c_\lambda \pi(\lambda)g \tag{14}$$

are called *regular Gabor expansions* of the function or distribution f whenever Λ is a lattice in $\mathbb{R}^d \times \widehat{\mathbb{R}}^d$, i.e. a discrete and co-compact subgroup of *phase-space*, resp. a set of the form $\mathring{A} * \mathbb{Z}^{2d}$, for some non-singular matrix $\mathring{A}$, and the building block g (not necessarily the Gauss-function, but typically a function well concentrated near zero and of some smoothness) is termed *Gabor atom* (or *Gabor-window*) (see e.g. [52]).

The study of the convergence of such series, the determination of suitable coefficients and the comparison of properties of the function and the (global, i.e. decay and summability) properties of the coefficient sequence $(c_\lambda)_{\lambda\in\Lambda}$ requires to work with suitable sequence spaces, in order to define the appropriate function spaces (again it is more proper to talk of the Banach space of distributions).

Depending on the viewpoint and the circumstances one makes use of small or comprehensive families of such function spaces, which are nowadays summarized under the name of *modulation spaces*. While the (now classical) modulation spaces $\boldsymbol{M}^s_{p,q}(\mathbb{R}^d)$ (see [49, 52]) are modeled as a family of spaces with maximal similarity to the well established family of Besov spaces $\boldsymbol{B}^s_{p,q}(\mathbb{R}^d)$, with $s\in\mathbb{R}, 1\le p,q\le\infty$, the general class of modulation spaces allows much for freedom in the choice of weights, including radial symmetric weights (leading to the so-called Shubin classes) or weights which allow to model the idea of variable band-width ([1]), but also to choose weighted variants with moderate weight functions growing exponentially.

The Linear Algebra Background

When we deal with questions of Fourier analysis or Time-Frequency analysis over finite Abelian groups [78] all the function spaces are finite dimensional, and thus questions about linear operators can be handled in the setting of linear algebra, using matrices. This allows to compute eigenvalues of hermitian matrices, apply the FFT to signals of finite length (resp. the $FFT2$ for digital images) and one does not have to deal with convergence issues. However it is still of interest to take the analytic viewpoint as suggested by the functional analytic questions which have to be considered in the continuous case, because it is not irrelevant to know whether a matrix (e.g., the matrix describing the frame operator of some Gabor system) is badly conditioned or not. Clearly it is important to the user to have well-conditioned systems and efficient algorithms, which are often derived using algebraic properties of the setting. The theory of Banach Gelfand triples $(\boldsymbol{S}_0,\boldsymbol{L}^2,\boldsymbol{S}_0')(\mathscr{G})$ [31, 59, 71] allows to transfer this situation into the realm of *continuous variables*, in particular to the setting of continuous variables in the following way:

- Whatever result is valid in the context of finite Abelian groups can be expected to provide valid claims in for *nice test functions*, i.e. functions from $\boldsymbol{S}_0(\mathscr{G})$;
- most of the time corresponding statements can be expanded (using abstract approximation principles) to the Hilbert space setting (i.e., to $\boldsymbol{L}^2(\mathscr{G})$;
- the further extension to $\boldsymbol{S}_0'(\mathscr{G})$ is entailed using the w^*-density of $\boldsymbol{S}_0(\mathscr{G})$ or $\boldsymbol{L}^2(\mathscr{G})$ in $\boldsymbol{S}_0'(\mathscr{G})$.

Just to give a simple example: In the context of the Banach Gelfand triple $(\boldsymbol{S}_0,\boldsymbol{L}^2,\boldsymbol{S}_0')(\mathbb{R}^d)$ the *Fourier transform* is considered as a *unitary Banach Gelfand triple automorphism*, which means simply that

1. The Fourier transform is well defined (together with its inverse Fourier transform, taken as integrals in the traditional Riemann sense!) as an isometric mapping on $\left(\boldsymbol{S}_0(\mathbb{R}^d), \|\cdot\|_{\boldsymbol{S}_0}\right)$;
2. It is shown to be isometric also with respect to the standard $\boldsymbol{L}^2$-norm, hence by a simple density argument it can be extended to in fact a unitary automorphism of the Hilbert space $\left(\boldsymbol{L}^2(\mathbb{R}^d), \|\cdot\|_2\right)$; this step only requires to interpret $\left(\boldsymbol{L}^2(\mathbb{R}^d), \|\cdot\|_2\right)$ as the completion of $(\boldsymbol{S}_0(\mathbb{R}^d), \|\cdot\|_2)$;
3. Finally, by duality the Fourier transform of $\sigma \in \boldsymbol{S}_0'(\mathbb{R}^d)$ is given by

$$\hat{\sigma}(f) = \sigma(\hat{f}),\ f \in \boldsymbol{S}_0(\mathbb{R}^d),$$

describing also the unique $w^* - w^*$ continuous extension of the Fourier-Plancherel transform to a linear automorphism of $(\boldsymbol{S}_0'(\mathbb{R}^d), \|\cdot\|_{\boldsymbol{S}_0'})$.

Among all the Banach Gelfand triples one can characterize then $\mathscr{F}$ (the Fourier transform) by its characteristic property (known from the FFT):

$$\mathscr{F}(\chi_s) = \delta_s, \quad s \in \mathbb{R}^d,$$

i.e. the Fourier transform (and also its inverse, up to change of signs) identifies *pure frequencies* with *Dirac measures*. Such a statement would be expressed by a theoretical physicist as the claim that it performs a *change of basis* from the continuous family of pure frequencies to the basis of Diracs (and vice versa).

When it comes to realization of such an abstract mapping it is important to have a good way of expressing the connection between this continuous (generalized or classical) Fourier transform and corresponding linear mappings which can be performed on the computer. Clearly the FFT algorithm by Cooley-Tukey ([29]) comes to mind, which is an efficient way of realizing the DFT (discrete Fourier transform, resp. the abstract Fourier transform over the cyclic group of order n, especially for n being rich of divisors, such as $n = 2^k$, for some k, such as $n = 1024$ or 512).

Although one may expect that the application of the FFT algorithm applied to regular samples taken from a nice function has something to do with the samples of the Fourier transform (understood as integral transform), this is not true in the strict sense. In fact, only the combination of *sampling* and *periodization* allows to make the link to the FFT. Details concerning this questions are found in a paper by N. Kaiblinger ([108]), based on [68]. It shows that - suitable positioned - among others: The piecewise linear interpolators to the FFT-transformed values of a regular sampled function $f \in \boldsymbol{S}_0(\mathbb{R})$ converge to $\hat{f}$ in the space $\left(\boldsymbol{S}_0(\mathbb{R}), \|\cdot\|_{\boldsymbol{S}_0}\right)$. No similar result (without even stronger assumptions) appears to be known in the literature for the standard spaces (most likely because it is not possible to prove such results, or only under some much stronger assumptions). Note that $\boldsymbol{S}_0(\mathbb{R})$-convergence implies the convergence in $\left(\boldsymbol{L}^p(\mathbb{R}^d), \|\cdot\|_p\right)$, for any $p \in [1, \infty]$. In [108] it is also shown that Gabor analysis computations for $\mathbb{R}^d$ can be approximately be performed using corresponding finite Gabor families (hence discrete versions of the STFT).

Modulation Spaces

The first extensive report on modulation spaces was formulated in 1983 and published in 2003 [57]. The link to Gabor expansions and the resulting atomic decompositions (comparable to the Frazier-Jawerth Theory for Besov-Triebel-Lizorkin spaces as given in [83]) was given in [52] (making a result public which was presented at a conference in 1986 in Edmonton). Only slowly it appeared that modulation spaces can be characterized not only by very specific atomic decompositions of Gaborian type, but also through a continuous transform, namely the STFT. In this way the connection to representation theory of the Heisenberg group was established.

Later on the comparison between the family of function spaces (mostly Besov-Triebel-Lizorkin spaces) characterized via expansions and the corresponding characterization of modulation spaces using the STFT led to the creation of *coorbit theory* (see [63, 64, 88]), which is still expanding and exploited more and more, in order to create a more systematic approach to the large variety of function spaces.

In this context the by now classical modulation spaces $\left(\boldsymbol{M}^s_{p,q}(\mathbb{R}^d), \|\cdot\|_{\boldsymbol{M}^s_{p,q}}\right)$ and their various generalizations, including Shubin classes $\boldsymbol{Q}(\mathbb{R}^d)$ or functions of variable bandwidth [1] are obtained as special coorbit spaces from the Schrödinger representation of the Heisenberg group, or more practically speaking they are characterized by the membership of their STFT (with respect to some Schwartz window g) in some solid and translation invariant function space $(\boldsymbol{Y}, \|\cdot\|_{\boldsymbol{Y}})$ over the time-frequency plane $\mathbb{R}^d \times \widehat{\mathbb{R}}^d$.

As suggested in [58] we will call all the spaces arising in this way as *modulation spaces*. The most important are the space $\left(\boldsymbol{M}^p(\mathbb{R}^d), \|\cdot\|_{\boldsymbol{M}^p}\right)$, with $p \in [1,\infty]$. These spaces are all Fourier invariant and can be characterized by the fact that their elements allow for a Gabor expansion (with respect to any good Gabor system) with $\boldsymbol{\ell}^p$-coefficients. Specifically the space $\boldsymbol{M}^1(\mathbb{R}^d) = \boldsymbol{S}_0(\mathbb{R}^d), \boldsymbol{M}^2(\mathbb{R}^d) = \boldsymbol{L}^2(\mathbb{R}^d)$ and $\boldsymbol{M}^\infty(\mathbb{R}^d) = \boldsymbol{S}_0'(\mathbb{R}^d)$ are relevant. $\boldsymbol{M}^1(\mathbb{R}^d) = \boldsymbol{S}_0(\mathbb{R}^d)$ (cf. [47, 89]) is the smallest space in the family of all Banach spaces with are isometrically TF-invariant, i.e. which satisfy $\|\pi(\lambda)g\|_{\boldsymbol{B}} = \|g\|_{\boldsymbol{B}}$. It also appears to be one of the most useful spaces, not only for TF-analysis, but also for classical Fourier analysis (in the form of classical Fourier analysis).

Banach Gelfand Triples

Although modulation spaces as described in the previous section form a comprehensive family of function spaces, or more correctly, a family of Banach spaces of (ultra-) distributions it is sometimes beneficial to restrict the attention to a so-called Banach Gelfand triple, typically the Banach Gelfand triple $(\boldsymbol{S}_0, \boldsymbol{L}^2, \boldsymbol{S}_0')(\mathbb{R}^d)$, consisting of the Segal algebra $\boldsymbol{S}_0(\mathbb{R}^d)$ (which coincides with the modulation space $\boldsymbol{M}^1(\mathbb{R}^d)$, and is used with this symbol in [89] extensively), the Hilbert space $\boldsymbol{L}^2(\mathbb{R}^d)$

and the dual space $\boldsymbol{S}_0'(\mathbb{R}^d)$. We also like to think of this triple as a *rigged Hilbert space*, i.e. an enrichment of the Hilbert space structure.

An indication of the usefulness of this particular Banach Gelfand triple and explanations, why it is a convenient setting for many basic facts, including a so-called kernel theorem or the description of the (extended) Fourier transform can be found in [31] and [71]. While Plancherel's theorem is just describing the preservation of energy, making the Fourier transform a unitary automorphism of $\boldsymbol{L}^2(\mathbb{R}^d)$, the extended setting provides the information that it maps a pure frequency ($t \mapsto exp(2\pi ist)$ exactly to the corresponding Dirac measure $\delta_s \in \boldsymbol{S}_0'(\mathbb{R}^d)$, a statement which cannot be made using only $\boldsymbol{L}^2(\mathbb{R}^d)$, because both of these objects are not belonging to this Hilbert space. However, when it comes to realize the Fourier transform (or its inverse) as an integral transform or to explicitly describe the transition from the kernel $K \in \boldsymbol{S}_0(\mathbb{R}^{2d})$ to the corresponding Kohn-Nirenberg symbol it is very convenient to use functions from the space of test functions, because due to the good properties of the functor $\mathscr{G} \mapsto \left(\boldsymbol{S}_0(G), \|\cdot\|_{\boldsymbol{S}_0}\right)$ it is no problem to carry out all the necessary operations (automorphisms, partial Fourier transforms) without problems, all the integrals converge in the best possible way (namely as absolutely convergent Riemannian sums) and no discussion of sets of measure zero nor the application of summability methods is required.

Let us just mention that we compare in our teaching the situation with the treatment of the fields $\mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$. When we do computations using these different field we normally do *not distinguish* anymore between the different objects, although a priori they are constituted in a completely different way. Rational numbers are described as pairs of integers, while real numbers are typically viewed as infinite decimal expressions. But since there are natural embeddings of the smaller structure into the larger all the relevant structures of a field can be used at any level, in fact one will typically work at the most convenient level. Certainly the completeness makes $\mathbb{R}$ the favorite setting (comparable to the Hilbert space with the triple), while certain operations, such as forming the multiplicative inverse: $x \to 1/x$ are done more accurately and swiftly within $\mathbb{Q}$ (and are extended in a unique, but in practice *cumbersome* way to $\mathbb{R}$). It is a fantastic achievement of analysis that we have the field of real numbers, and that we know that $\gamma = 1/\pi^2$ is a well-defined number, even if we never think about this in a constructive way, meaning how to compute the "infinite decimal expression" representing this number $\gamma \in \mathbb{R}$! It should suffice to mention Moivre's formulas as a convenient tool to derive the addition theorem for the trigonometric functions via Euler's formula from the exponential law, also valid over the field $\mathbb{C}$. In the same way the "outer level" of the Banach Gelfand triple gives us the right perspective and the most elegant way of proving (and understanding or interpreting) things, which are often obvious for the case of finite Abelian groups, resp. for discrete and periodic signals (where the FFT is a unitary matrix mapping discrete pure frequencies to unit vectors and vice versa).

Outlook and Further Aspects

The information age is not only influencing the way how science in general is performed, it also enables and encourages alternative forms of knowledge accumulation, respectively, information processing.

For this reason we want to mention at the end of this article some of them, because we think that some aspects of this change have to do with the considerations formulated in the early part of this article.

Once it is agreed within a community of researchers that a concrete question is of great relevance, that a certain area should be explored more systematically, or that one needs function spaces with certain prescribed properties, it is also very likely that especially young researchers will be motivated to go into such a direction. Such choices may be more fruitful than just following the advice of their advisors (who may like to see their own, old problems solved), but one has also to be aware that at the same time the movement towards "fashion" or "bubbles" arise. But the following might help to discriminate between potentially more relevant and less interesting questions, which could be given as a recommendation: *Would I be interested in the answer to the question I formulate and answer in my manuscript BEYOND the possibility of having a publication?* Why would the answer provided (as complicated as it may be) be interesting to anyone?

Meta-Studies and Survey Activities

There is an interesting development in the *medical sciences*, where it has been realized quite a while ago that despite the large number research projects it is very hard for a normal member of the medical service to keep up with the amount of information available concerning medications and therapies. One consequence is increasing *specialization*, but this does not necessarily lead to a better treatment, because the specialist in one field may not see the connections between different symptoms or may not be aware of the effects arising from the combination of different therapies.

In order to fight this problem the Cochrane Institute was created, see www.cochrane.org, which is supported by a group of top scientists. They have established a *system of information condensation* suitable for the practitioner, who can find well-prepared meta-studies concerning concrete questions, with well-structured information about the studies which have been used to prepare a given report, *and the evaluation criteria used* to establish the comparison.

According to their own web-page *Cochrane* is a global independent network of health practitioners, researchers, patient advocates, and others, responding to the challenge of making the vast amounts of evidence generated through research useful for informing decisions about health. We are a not-for-profit organization with collaborators from over 120 countries working together to produce credible, accessible health information that is free from commercial sponsorship and other conflicts of interest. They write further:

Our vision is a world of improved health where decisions about health and health care are informed by high-quality, relevant and up-to-date synthesized research evidence.

Our mission is to promote evidence-informed health decision-making by producing high-quality, relevant, accessible systematic reviews and other synthesized research evidence.

Our work is internationally recognized as the benchmark for high quality information about the effectiveness of health care.

They provide an answer to the following key question: **What is a systematic review?** For this question they provide the following answer: *A systematic review summarizes the results of available carefully designed healthcare studies (controlled trials) and provides a high level of evidence on the effectiveness of healthcare interventions. Judgments may be made about the evidence and inform recommendations for healthcare.*

These reviews are complicated and depend largely on what clinical trials are available, how they were carried out (the quality of the trials) and the health outcomes that were measured. Review authors pool numerical data about effects of the treatment through a process called meta-analyses. Then authors assess the evidence for any benefits or harms from those treatments. In this way, systematic reviews are able to summarise the existing clinical research on a topic.

When we think of *pioneering work in the mathematical sciences* our first ideas are typically related to the level of *originality* and difficulty of a problem that has been solved. The quality of the presentation, the level of information provided to the community by the article is often considered secondary. Fortunately some publishers promote the publication of good summaries and compilations of material, even if it is not completely new.

Writing a good survey is in many cases equally challenging (and sometimes even more useful for others) than just adding a few more details to an existing body of knowledge, hoping that "some day somebody may make use of it". In reality such small pieces of information get rather lost and results needed in a concrete situation will be easily reproved whenever they occur, and nobody will search resp. find those results in the huge pile of general publications. In signal processing one would call this *publication noise*, i.e. "noise" which makes it in fact more difficult to find the relevant information, the "true signal" in the avalanche of technical details.

We are not saying that all such results are useless. According to the terminology established by Thomas Kuhn ([111]) they constitute *standard research* which may prepare the ground for real innovation, the so-called change of paradigms. In fact, the partial results making up the main body of published research nowadays should go into more systematic summaries of Cochrane type.

Unfortunately such summaries are not (!yet) highly valuated within the mathematical community, nor can we find many activities within our field that would be comparable with the Cochrane Institute. Of course I do not ignore the valuable contributions of learned societies such as SIAM, AMS, EMS, and so on, but finding ways to improve the situation, to open appropriate platforms, to have a discussion on such issues, should certainly be intensified.

Collaborative Research Efforts

As a last topic I would like to mention the style of cooperation which may also carry some of the spirit I try to support. If we have common goals, if a free exchange of ideas is the best way to get good results which are also useful for science, technology, or for the quality of life, then we have to work together. This is a change of paradigm in the workstyle within the mathematical community, where achievements by great individuals mark the history.

In his talk at the occasion of the 2015 Breakthrough Prize in Mathematics Symposium Terence Tao was reporting on the experiences of the POLYMATH project which he has been involved (together with Timothy Gowers) in the last year.

It provides another model for cooperative research which might also be useful in the context of function spaces. See for details http://youtube/elWIDVI6b18

https://www.youtube.com/watch?v=elWIDVI6b18

I would like to leave this to discussions in the community and thank the reader who has read the article to this point for her/his patience. Feedback to the author is welcome (hans.feichtinger@univie.ac.at).

Acknowledgements

The author would like to thank CIRM, *Centre international de rencontres mathematiques* at Luminy (Marseille) for the hospitality. This article was written at CIRM during the MORLET CHAIR semester of the author (August 2014 to January 2015).

Personal thanks concerning inspiration go to my many coauthors over the years, who contributed to a sharpening of my perspectives in many different ways.

Aside from the obvious intensive cooperation with Charly Gröchenig, with whom I developed the *theory of coorbit spaces* and put up - together with Thomas Strohmer - the foundations of *iterative reconstruction of band-limited functions from irregular samples* I want to specifically mention Werner Kozek and Franz Luef. Werner joined the NuHAG team around 1994 and helped us to establish many connections between abstract and applied harmonic analysis. Many of the questions arising in time-frequency analysis and in particular Gabor analysis required to rethink the function space concepts and look out for function spaces which are more appropriate than the usual $\boldsymbol{L}^p$-spaces or even the Schwartz space of tempered distributions. Fortunately the Segal algebra $\boldsymbol{S}_0(\mathscr{G})$ and its dual were already available at that time (see [45, 47]). The articles [69] and [74] are the precursors of the more elegant set of results concerning Banach Gelfand triples. Franz in turn established interesting connections to *non-commutative geometry* (see [118–120]), and many discussions concerning the foundations of our field and the usefulness of the Banach Gelfand Triple in this field have certainly contributed to the solidification of the viewpoints presented here (see [31, 70, 77, 78]).

Finally the author would like to acknowledge inspiration taken from a talk of Joachim Buhmann at the final meeting of the DFG priority program 1324 (see [95] for a related article). At this occasion he was presenting very interesting ideas concerning the *information content* of algorithms, which obviously influenced some of the sections of this chapter. The concrete set of ideas which we found most important concern the suggestion to switch from an *inventor's viewpoint* to a user's perspective. The comparison with medical science was equally enlightening: a patient does not care whether a certain pharmacy is the best for some disease, she or he is just interested in receiving the best possible treatment for her/his situation. The simple observation that this is still a perspective rarely taken in mathematical discussions was one of the main motivations for the perhaps sometimes far-fetched comparisons used in this chapter.

References

1. R. Aceska, H.G. Feichtinger, Reproducing kernels and variable bandwidth. J. Funct. Spaces Appl., Art. ID 469341 (2012)
2. D.R. Adams, Sets and functions of finite L^p-capacity. Indiana Univ. Math. J. **27**(4), 611–627 (1978)
3. A. Aldroubi, H.G. Feichtinger, Non-uniform sampling: exact reconstruction from non-uniformly distributed weighted-averages, in *Wavelet Analysis: Twenty Years Developments Proceedings of the International Conference of Computational Harmonic Analysis, Hong Kong, China, June 4–8, 2001, volume 1 of Ser. Anal.*, ed. by D.X. Zhou (World Science Publication, Singapore, 2002), pp. 1–8
4. A. Aldroubi, K. Gröchenig, Nonuniform sampling and reconstruction in shift-invariant spaces. SIAM Rev. **43**(4), 585–620 (2001)
5. J.-P. Antoine, C. Trapani, *Partial Inner Product Spaces - Theory and Applications*. Lecture Notes in Mathematics, vol. 1986 (Springer, Berlin, 2009)
6. J. Arazy, S.D. Fisher, Some aspects of the minimal, Möbius-invariant space of analytic functions on the unit disc, in *Interpolation Spaces and Allied Topics in Analysis*, ed. by M. Cwikel, J. Peetre. Proceedings of the Conference held in Lund, Sweden, August 29–September 1, 1983. Lecture Notes in Mathematics, vol. 1070 (Springer, Berlin, 1984), pp. 24–44
7. J. Arazy, S.D. Fisher, The uniqueness of the Dirichlet space among Möbius-invariant Hilbert spaces. Ill. J. Math. **29**(3), 449–462 (1985)
8. J. Arazy, S. Fisher, J. Peetre, Möbius invariant function spaces. J. Reine Angew. Math. **363**, 110–145 (1985)
9. J. Arazy, S. Fisher, J. Peetre, Möbius invariant spaces of analytic functions, in *Complex Analysis I. Proc Spec Year, College Park/Md 1985-86*, ed. by C.A. Berenstein. Lecture Notes in Mathematics, vol. 1275 (Springer, Berlin, 1987), pp. 10–22
10. N. Aronszajn, Theory of reproducing kernels. Trans. Am. Math. Soc. **68**, 337–404 (1950)
11. G. Ascensi, H.G. Feichtinger, N. Kaiblinger, Dilation of the Weyl symbol and Balian-low theorem. Trans. Am. Math. Soc. **366**(7), 3865–3880 (2014)
12. A. Benedek, R. Panzone, The space L^p, with mixed norm. Duke Math. J. **28**(3), 301–324 (1961)
13. J.J. Benedetto, C. Heil, D.F. Walnut, Differentiation and the Balian-low theorem. J. Fourier Anal. Appl. **1**(4), 355–402 (1995)
14. J.J. Benedetto, W. Czaja, A.M. Powell, J. Sterbenz, An endpoint $(1,\infty)$ Balian-low theorem. Math. Res. Lett. **13**(3), 467–474 (2006)
15. C. Bennett, R.C. Sharpley, *Interpolation of Operators* (Academic, London, 1988)

16. J. Bergh, J. Löfström, *Interpolation Spaces. An Introduction*. Grundlehren der Mathematischen Wissenschaften, vol. 223 (Springer, Berlin, 1976)
17. J.-P. Bertrandias, C. Dupuis, Transformation de Fourier sur les espaces $l^p(L^{p'})$. Ann. Inst. Fourier (Grenoble) **29**(1), 189–206 (1979)
18. J.-P. Bertrandias, C. Datry, C. Dupuis, Unions et intersections d'espaces L^p invariantes par translation ou convolution. Ann. Inst. Fourier (Grenoble) **28**(2), 53–84 (1978)
19. A. Beurling, Construction and analysis of some convolution algebras. Ann. Inst. Fourier (Grenoble) **14**(2), 1–32 (1964)
20. L. Borup, M. Nielsen, Frame decomposition of decomposition spaces. J. Fourier Anal. Appl. **13**(1), 39–70 (2007)
21. W. Braun, H.G. Feichtinger, Banach spaces of distributions having two module structures. J. Funct. Anal. **51**, 174–212 (1983)
22. F. Bruhat, Distributions sur un groupe localement compact et applications à l etude des représentations des groupes p-adiques. Bull. Soc. Math. France **89**, 43–75 (1961)
23. R.C. Busby, H.A. Smith, Product-convolution operators and mixed-norm spaces. Trans. Am. Math. Soc. **263**, 309–341 (1981)
24. L. Carleson, On convergence and growth of partial sums of Fourier series. Acta Math. **116**, 135–157 (1966)
25. P. Cartier, Über einige integralformeln in der theorie der quadratischen Formen. Math. Z. **84**, 93–100 (1964)
26. J. Cerda, J. Martin, P. Silvestre, Capacitary function spaces. Collect. Math. **62**(1), 95–118 (2011)
27. O. Christensen, R.S. Laugesen, Approximately dual frame pairs in Hilbert spaces and applications to Gabor frames. Sampl. Theory Signal Image Process. **9**(1–3), 77–89 (2010)
28. R.R. Coifman, G. Weiss, Extensions of Hardy spaces and their use in analysis. Bull. Am. Math. Soc. **83**(4), 569–645 (1977)
29. J. Cooley, J. Tukey, An algorithm for the machine calculation of complex Fourier series. Math. Comput. **19**, 297–301 (1965)
30. E. Cordero, F. Nicola, Pseudodifferential operators on l^p, Wiener amalgam and modulation spaces. Int. Math. Res. Notices **2010**(10), 1860–1893 (2010)
31. E. Cordero, H.G. Feichtinger, F. Luef, Banach Gelfand triples for Gabor analysis, in *Pseudo-Differential Operators*, ed. by L. Rodino, M.W. Wong. Lecture Notes in Mathematics, vol. 1949 (Springer, Berlin, 2008), pp. 1–33
32. E. Cordero, F. Nicola, L. Rodino, On the global boundedness of Fourier integral operators. Ann. Global Anal. Geom. **38**(4), 373–398 (2010)
33. E. Cordero, K. Gröchenig, F. Nicola, Approximation of Fourier integral operators by Gabor multipliers. J. Fourier Anal. Appl. **18**(4), 661–684 (2012)
34. E. Cordero, J. Toft, P. Wahlberg, Sharp results for the Weyl product on modulation spaces. J. Funct. Anal. **267**(8), 3016–3057 (2014)
35. S. Dahlke, G. Kutyniok, G. Steidl, G. Teschke, Shearlet coorbit spaces and associated Banach frames. Appl. Comput. Harmon. Anal. **27**(2), 195–214 (2009)
36. S. Dahlke, G. Steidl, G. Teschke, The continuous shearlet transform in arbitrary space dimensions. J. Fourier Anal. Appl., **16**(3), 340–364 (2010)
37. A. Deitmar, *A First Course in Harmonic Analysis*. Universitext (Springer, New York, NY, 2002)
38. A. Derighetti, Closed subgroups as Ditkin sets. J. Funct. Anal. **266**(3), 1702–1715 (2014)
39. L. Diening, Riesz potential and Sobolev embeddings on generalized Lebesgue and Sobolev spaces $L^{p(\cdot)}$ and $W^{k,p(\cdot)}$. Math. Nachr. **268**, 31–43 (2004)
40. L. Diening, P. Hästö, S. Roudenko, Function spaces of variable smoothness and integrability. J. Funct. Anal. **256**(6), 1731–1768 (2009)
41. P. Eymard, L'algèbre de Fourier d'un groupe localement compact. Bull. Soc. Math. France **92**, 181–236 (1964)
42. H.G. Feichtinger, Multipliers of Banach spaces of functions on groups. Math. Z. **152**, 47–58 (1976)

43. H.G. Feichtinger, A characterization of Wiener's algebra on locally compact groups. Arch. Math. (Basel) **29**, 136–140 (1977)
44. H.G. Feichtinger, Gewichtsfunktionen auf lokalkompakten Gruppen. Sitzungsber. Österr. Akad.Wiss. **188**, 451–471 (1979)
45. H.G. Feichtinger, Un espace de Banach de distributions tempérées sur les groupes localement compacts abéliens. C. R. Acad. Sci. Paris S'er. A–B **290**(17), 791–794 (1980)
46. H.G. Feichtinger, Banach spaces of distributions of Wiener's type and interpolation, in *Proceedings of Conference on Functional Analysis and Approximation, Oberwolfach August 1980*, ed. by P. Butzer, S. Nagy, E. Görlich. International Series of Numerical Mathematics, vol. 69 (Birkhäuser, Boston, Basel, 1981), pp. 153–165
47. H.G. Feichtinger, On a new Segal algebra. Monatsh. Math. **92**, 269–289 (1981)
48. H.G. Feichtinger, Banach convolution algebras of Wiener type, in *Proceedings of Conference on Functions, Series, Operators, Budapest 1980*, ed. by B.S. Nagy, J. Szabados. Colloquia Mathematica Societatis Janos Bolyai, vol. 35 (North-Holland, Amsterdam, 1983), pp. 509–524
49. H.G. Feichtinger, Modulation spaces on locally compact Abelian groups. Technical Report, Jan 1983
50. H.G. Feichtinger Minimal Banach spaces and atomic representations. Publ. Math. Debrecen **34**(3–4), 231–240 (1987)
51. H.G. Feichtinger, An elementary approach to Wiener's third Tauberian theorem for the Euclidean *n*-space, in *Symposia Math., Volume XXIX of* Analisa Armonica, Cortona, pp. 267–301, 1988
52. H.G. Feichtinger, Atomic characterizations of modulation spaces through Gabor-type representations, in *Proceedings of Conference on Constructive Function Theory*. Rocky Mountain Journal of Mathematics, vol. 19, pp. 113–126 (1989)
53. H.G. Feichtinger, Generalized amalgams, with applications to Fourier transform. Can. J. Math. **42**(3), 395–409 (1990)
54. H.G. Feichtinger, New results on regular and irregular sampling based on Wiener amalgams, in *Proceedings of Conference on Function Spaces, Edwardsville/IL (USA) 1990*, ed. by K. Jarosz. Lecture Notes in Pure Applied Mathematics, vol. 136 (Marcel Dekker, New York, 1992), pp.107–121
55. H.G. Feichtinger, Wiener amalgams over Euclidean spaces and some of their applications, in *Proceedings of Conference on Function spaces*, 1990. Lect. Notes Pure Appl. Math **136**, (Marcel Dekker, Edwardsville, 1992), pp. 123–137. Zbl 0833.46030
56. H.G. Feichtinger, Spline-type spaces in Gabor analysis, in *Wavelet Analysis: Twenty Years Developments Proceedings of the International Conference of Computational Harmonic Analysis, Hong Kong, China, 4–8 June 2001, Ser. Anal.*, vol. 1, ed. by D.X. Zhou (World Scientific Publishing, River Edge, NJ, 2002), pp. 100–122
57. H.G. Feichtinger, Modulation spaces of locally compact Abelian groups, in *Proceedings of International Conference on Wavelets and Applications, Chennai, January 2002*, ed. by R. Radha, M. Krishna, S. Thangavelu (Allied Publishers, New Delhi, 2003), pp. 1–56
58. H.G. Feichtinger, Modulation spaces: looking back and ahead. Sampl. Theory Signal Image Process. **5**(2), 109–140 (2006)
59. H.G. Feichtinger, Banach Gelfand triples for applications in physics and engineering, in *AIP Conference Proceedings*, vol. 1146 (American Institute of Physics, New York, 2009), pp. 189–228
60. H.G. Feichtinger, *Elements of Postmodern Harmonic Analysis* (Springer, Berlin, 2014), p. 27
61. H.G. Feichtinger, W. Braun, Banach spaces of distributions with double module structure and twisted convolution, in *Proceedings of Alfred Haar Memorial Conference*. Colloquia Mathematica Societatis Janos Bolyai (North Holland Publ. Comp., Amsterdam, Zbl 0515.46045 1985)
62. H.G. Feichtinger, P. Gröbner, Banach spaces of distributions defined by decomposition methods. I. Math. Nachr. **123**, 97–120 (1985)

63. H.G. Feichtinger, K. Gröchenig, Banach spaces related to integrable group representations and their atomic decompositions, I. J. Funct. Anal. **86**(2), 307–340 (1989)
64. H.G. Feichtinger, K. Gröchenig, Banach spaces related to integrable group representations and their atomic decompositions, II. Monatsh. Math. **108**(2–3), 129–148 (1989)
65. H.G. Feichtinger, K. Gröchenig, Irregular sampling theorems and series expansions of band-limited functions. J. Math. Anal. Appl. **167**(2), 530–556 (1992)
66. H.G. Feichtinger, K. Gröchenig, Iterative reconstruction of multivariate band-limited functions from irregular sampling values. SIAM J. Math. Anal. **23**(1), 244–261 (1992)
67. H.G. Feichtinger, K. Gröchenig, Error analysis in regular and irregular sampling theory. Appl. Anal. **50**(3–4), 167–189 (1993)
68. H.G. Feichtinger, N. Kaiblinger, Quasi-interpolation in the Fourier algebra. J. Approx. Theory **144**(1), 103–118 (2007)
69. H.G. Feichtinger, W. Kozek, Quantization of TF lattice-invariant operators on elementary LCA groups, in *Gabor Analysis and Algorithms*, ed. by H.G. Feichtinger, T. Strohmer. Applied and Numerical Harmonic Analysis. (Birkhäuser, Boston, MA, 1998), pp. 233–266
70. H.G. Feichtinger, F. Luef, Wiener amalgam spaces for the fundamental identity of Gabor analysis. Collect. Math. **57**(Extra Volume (2006)), 233–253 (2006)
71. H.G. Feichtinger, F. Luef, Banach Gelfand triples for analysis. Notices Am. Math. Soc. (2016, in preparation)
72. H.G. Feichtinger, D. Onchis, Constructive realization of dual systems for generators of multi-window spline-type spaces. J. Comput. Appl. Math. **234**(12), 3467–3479 (2010)
73. H.G. Feichtinger, T. Werther, Robustness of regular sampling in Sobolev algebras, in *Sampling, Wavelets and Tomography*, ed. by J. Benedetto (Birkhäuser, Boston, 2004), pp. 83–113
74. H.G. Feichtinger, G. Zimmermann, A Banach space of test functions for Gabor analysis, in *Gabor Analysis and Algorithms: Theory and Applications*, ed. by H.G. Feichtinger, T. Strohmer. Applied and Numerical Harmonic Analysis. (Birkhäuser, Boston, MA, 1998), pp. 123–170
75. H.G. Feichtinger, G. Zimmermann, An exotic minimal Banach space of functions. Math. Nachr. **239–240**, 42–61 (2002)
76. H.G. Feichtinger, K. Gröchenig, T. Strohmer, Efficient numerical methods in non-uniform sampling theory. Numer. Math. **69**(4), 423–440 (1995)
77. H.G. Feichtinger, F. Luef, T. Werther, A guided tour from linear algebra to the foundations of Gabor analysis, in *Gabor and Wavelet Frames*. Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore, vol. 10 (World Scientific Publishing, Hackensack, 2007), pp. 1–49
78. H.G. Feichtinger, W. Kozek, F. Luef, Gabor Analysis over finite Abelian groups. Appl. Comput. Harmon. Anal. **26**(2), 230–248 (2009)
79. H.G. Feichtinger, A. Grybos, D. Onchis, Approximate dual Gabor atoms via the adjoint lattice method. Adv. Comput. Math. **40**(3), 651–665 (2014)
80. A. Figa Talamanca, Translation invariant operators in L^p. Duke Math. J. **32**, 495–501 (1965)
81. G.B. Folland, *A Course in Abstract Harmonic Analysis* (CRC Press, Boca Raton, 1994)
82. J.J.F. Fournier, J. Stewart, Amalgams of L^p and ℓ^q. Bull. Amer. Math. Soc. (N.S.) **13**, 1–21 (1985)
83. M. Frazier, B. Jawerth, Decomposition of Besov spaces. Indiana Univ. Math. J. **34**, 777–799 (1985)
84. M. Frazier, B. Jawerth, A discrete transform and decompositions of distribution spaces. J. Funct. Anal. **93**(1), 34–170 (1990)
85. M.W. Frazier, B.D. Jawerth, G. Weiss, *Littlewood-Paley Theory and the Study of Function Spaces* (American Mathematical Society, Providence, RI, 1991)
86. D. Gabor, Theory of communication. J. IEE **93**(26), 429–457 (1946)
87. P. Gröbner, Banachräume glatter Funktionen und Zerlegungsmethoden. Ph.D. thesis, University of Vienna, 1992
88. K. Gröchenig, Describing functions: atomic decompositions versus frames. Monatsh. Math. **112**(3), 1–41 (1991)

89. K. Gröchenig, *Foundations of Time-Frequency Analysis*. Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, MA, 2001)
90. K. Gröchenig, Time-Frequency analysis of Sjörstrand's class. Rev. Mat. Iberoam. **22**(2), 703–724 (2006)
91. K. Gröchenig, Weight functions in time-frequency analysis, in *Pseudodifferential Operators: Partial Differential Equations and Time-Frequency Analysis*, ed. by L. Rodino, et al. Fields Institute Communications, vol. 52 (American Mathematical Society, Providence, RI, 2007), pp. 343–366
92. K. Gröchenig, C. Heil, Modulation spaces and pseudodifferential operators. Integr. Equat. Oper. Theory **34**(4), 439–457 (1999)
93. K. Gröchenig, T. Strohmer, Pseudodifferential operators on locally compact abelian groups and Sjöstrand's symbol class. J. Reine Angew. Math. **613**, 121–146 (2007)
94. K. Gröchenig, D. Han, C. Heil, G. Kutyniok, The Balian-Low theorem for symplectic lattices in higher dimensions. Appl. Comput. Harmon. Anal. **13**(2), 169–176 (2002)
95. A. Gronskiy, J. Buhmann, How informative are Minimum Spanning Tree algorithms? in *IEEE International Symposium on Information Theory (ISIT), 2014*, pp. 2277–2281, IEEE, June 2014
96. A. Haar, Der Massbegriff in der Theorie der kontinuierlichen Gruppen. Ann. Math. **34**(1), 147–169 (1933)
97. C. Heil, An introduction to weighted Wiener amalgams, in *Wavelets and Their Applications (Chennai, January 2002)*, ed. by M. Krishna, R. Radha, S. Thangavelu (Allied Publishers, New Delhi, 2003), pp. 183–216
98. C. Herz, Lipschitz spaces and Bernstein's theorem on absolutely convergent Fourier transforms. J. Math. Mech. **18**, 283–323 (1968)
99. E. Hewitt, K.A. Ross, *Abstract Harmonic Analysis. Vol. II: Structure and Analysis for Compact Groups. Analysis on Locally Compact Abelian Groups* (Springer, Berlin/Heidelberg/New York, 1970)
100. E. Hewitt, K.A. Ross, *Abstract Harmonic Analysis. Vol. 1: Structure of Topological Groups; Integration Theory; Group Representations*, 2nd edn. (Springer, Berlin/Heidelberg/New York, 1979)
101. F. Holland, Harmonic analysis on amalgams of L^p and ℓ^q. J. Lond. Math. Soc. **10**, 295–305 (1975)
102. A. Holst, J. Toft, P. Wahlberg, *Weyl Product Algebras and Classical Modulation Spaces* (Polish Academy of Sciences, Institute of Mathematics, Banach Center Publications, Warszawa, 2010)
103. L. Hörmander, Pseudo-differential operators. Commun. Pure Appl. Anal. **18**, 501–517 (1965)
104. K.B. Howell, *Principles of Fourier Analysis* (Chapman and Hall/CRC, Boca Raton, FL, 2001)
105. R. Johnson, Temperatures, Riesz potentials, and the Lipschitz spaces of Herz. Proc. Lond. Math. Soc. III. Ser. **27**, 290–316 (1973)
106. R. Johnson, Maximal subspaces of Besov spaces invariant under multiplication by characters. Trans. Am. Math. Soc. **249**, 387–407 (1979)
107. D.S. Jones, *The Theory of Generalised Functions*. Reprint of the 1982, 2nd Hardback edn. (Cambridge University Press, Cambridge, 2009)
108. N. Kaiblinger, Approximation of the Fourier transform and the dual Gabor window. J. Fourier Anal. Appl. **11**(1), 25–42 (2005)
109. Y. Katznelson, *An Introduction to Harmonic Analysis*, 3rd Corr. edn. (Cambridge University Press, Cambridge, 2004)
110. T. Kilpeläinen, Weighted Sobolev spaces and capacity. Ann. Acad. Sci. Fenn. Ser. A I Math. **19**(1), 95–113 (1994)
111. T. Kuhn, *The Structure of Scientific Revolutions* (University of Chicago Press, Chicago, 1996)
112. G. Kutyniok, D. Labate, *Shearlets Multiscale Analysis for Multivariate Data*. Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, MA, 2012)

113. R. Larsen, *An Introduction to the Theory of Multipliers* (Springer, New York, 1971)
114. P.G. Lemarié, Y. Meyer, Ondelettes et bases hilbertiennes (Wavelets and Hilbert bases). Rev. Mat. Iberoam. **2**, 1–18 (1986)
115. M.J. Lighthill, *Introduction to Fourier Analysis and Generalised Functions* (Students' Edition) (Cambridge University Press, Cambridge, 1962)
116. T.S. Liu, A.C.M. van Rooij, Sums and intersections of normed linear spaces. Math. Nachr. **42**(1–3), 29–42 (1969)
117. L. Loomis, *An Introduction to Abstract Harmonic Analysis* (Van Nostrand and Co, Toronto/New York/London, 1953)
118. F. Luef, Gabor analysis, noncommutative tori and Feichtinger's algebra, in *Gabor and Wavelet Frames*. Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore, vol. 10. World Scientific Publishing, Hackensack, 2007), pp. 77–106
119. F. Luef, Projective modules over non-commutative tori are multi-window Gabor frames for modulation spaces. J. Funct. Anal. **257**(6), 1921–1946 (2009)
120. F. Luef, Projections in noncommutative tori and Gabor frames. Proc. Am. Math. Soc. **139**(2), 571–582 (2011)
121. Y. Meyer, Minimalité de certains espaces fonctionnels et applications à la théorie des opérateurs (Minimality of certain functional spaces and applications to operator theory). *Séminaire Équations aux Dérivées Partielles*, pp. 1–12 (1985)
122. Y. Meyer, De la recherche petrolière à la géometrie des espaces de Banach en passant par les paraproduite (From petroleum research to Banach space geometry by way of paraproducts), in *Séminaire sur les Équations aux Dérivées Partielles, 1985–1986, Exp. No. I* (Ecole Polytech, Palaiseau, 1986), p. 11
123. Y. Meyer, Constructions de bases orthonormées d'ondelettes (construction of orthonormal bases of wavelets). Rev. Mat. Iberoam. **4**(1), 31–39 (1988)
124. J. Musielak, *Orlicz Spaces and Modular Spaces* (Springer, Berlin, 1983)
125. S.M. Nikol'skij, *Approximation of Functions of Several Variables and Imbedding Theorems*. (Translated from the Russian by J. M. Danskin.) Grundlehren der mathematischen Wissenschaften, Band 205. (Springer, Berlin/Heidelberg/New York, 1975), VIII + 420 p.
126. S. Nitzan, J.-F. Olsen, A quantitative Balian-Low theorem. J. Fourier Anal. Appl. **19**(5), 1078–1092 (2013)
127. M. Pap, Properties of the voice transform of the Blaschke group and connections with atomic decomposition results in the weighted Bergman spaces. J. Math. Anal. Appl. **389**(1), 340–350 (2012)
128. M. Pap, F. Schipp, The voice transform on the Blaschke group I. Pure Math. Appl. (PUMA) **17**(3–4), 387–395 (2006)
129. J. Peetre, *New Thoughts on Besov Spaces*. Duke University Mathematics Series, vol. 1 (Mathematics Department, Duke University, Durham, 1976)
130. A. Pietsch, Approximation spaces. J. Approx. Theory **32**(2), 115–134 (1981)
131. M. Rao, Z. Ren, *Theory of Orlicz Spaces*. Pure and Applied Mathematics, vol. 146 (Marcel Dekker, New York, 1991)
132. Y.B. Rutitskij, M. Krasnoselskij, *Convex Functions and Orlicz Spaces*. (P. Noordhoff Ltd. IX, Groningen, The Netherlands, 1961), 249 p.
133. L. Schwartz, *Theorie des Distributions* (1959)
134. H.S. Shapiro, *Topics in Approximation Theory*. Lecture Notes in Mathematics, vol. 187 (Springer, Berlin, 1971)
135. E.M. Stein, *Singular Integrals and Differentiability Properties of Functions* (Princeton University Press, Princeton, NJ, 1970)
136. J. Toft, Continuity and Schatten properties for Toeplitz operators on modulation spaces, in *Modern Trends in Pseudo-Differential Operators*, ed. by J. Toft, M.W. Wong, H. Zhu. Operator Theory: Advances and Applications, vol. 172 (Birkhäuser, Basel, 2007), pp. 313–328
137. H. Triebel, *Spaces of Besov-Hardy-Sobolev Type*. (B. G. Teubner, Leipzig, 1978)

138. H. Triebel, *Theory of Function Spaces*. Monographs in Mathematics, vol. 78. (Birkhäuser, Basel, 1983)
139. H. Triebel, *Theory of Function Spaces II*. Monographs in Mathematics, vol. 84 (Birkhäuser, Basel, 1992)
140. N. Varopoulos, Tensor algebras and harmonic analysis. Acta Math. **119**, 51–112 (1967)
141. A. Weil, *L'integration dans les Groupes Topologiques et ses Applications*. (Hermann and Cie, Paris, 1940)
142. N. Wiener, *The Fourier Integral and Certain of Its Applications* (Cambridge University Press, Cambridge, 1933)

Existence of frames with prescribed norms and frame operator

Marcin Bownik and John Jasper

Abstract In this chapter we survey several recent results on the existence of frames with prescribed norms and frame operator. These results are equivalent to Schur-Horn type theorems which describe possible diagonals of positive self-adjoint operators with specified spectral properties. The first infinite dimensional result of this type is due to Kadison who characterized diagonals of orthogonal projections. Kadison's theorem automatically gives a characterization of all possible sequences of norms of Parseval frames. We present some generalizations of Kadison's result such as (a) the lower and upper frame bounds are specified, (b) the frame operator has two point spectrum, and (c) the frame operator has a finite spectrum.

Key words: Frame, Frame operator, Diagonals of self-adjoint operators, The Schur-Horn theorem, The Pythagorean theorem, The Carpenter theorem, Spectral theory

Frames and the Schur-Horn Theorem

The concept of frames in Hilbert spaces was originally introduced in the context of nonharmonic Fourier series by Duffin and Schaeffer [18] in the 1950's. The advent of wavelet theory brought a renewed interest in frame theory as is attested by now classical books of Daubechies [16], Meyer [31], and Mallat [30]. For an introduction to frame theory we refer to the book by Christensen [15].

M. Bownik (✉)
Department of Mathematics, University of Oregon, Eugene, OR 97403–1222, USA
e-mail: mbownik@uoregon.edu

J. Jasper
Department of Mathematics, University of Missouri, Columbia, MO 65211–4100, USA
e-mail: jasperj@missouri.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_4

Definition 1. A sequence $\{f_i\}_{i\in I}$ in a Hilbert space $\mathscr{H}$ is called a *frame* if there exists $0 < A \leq B < \infty$ such that

$$A||f||^2 \leq \sum_{i\in I} |\langle f, f_i\rangle|^2 \leq B||f||^2 \qquad \text{for all } f \in \mathscr{H}. \tag{1}$$

The numbers A and B are called the *frame bounds*. The supremum over all As and infimum over all Bs which satisfy (1) are called the *optimal frame bounds*. If $A = B$, then $\{f_i\}$ is said to be a *tight frame*. In addition, if $A = B = 1$, then $\{f_i\}$ is called a *Parseval frame*. The frame operator is defined by

$$Sf = \sum_{i\in I} \langle f, f_i\rangle f_i.$$

It is well known that S is a self-adjoint operator satisfying $A\mathbf{I} \leq S \leq B\mathbf{I}$.

The construction of frames with desired properties is a vast subject that is central to frame theory. Among the recently studied classes of frames with desired features are Grassmanian frames, equiangular frames, equal norm tight frames, finite frames for sigma-delta quantization, fusion frames, and frames for signal reconstruction without the phase. In particular, the construction of frames with prescribed norms and frame operator has been studied by many authors.

Problem 1. Characterize all possible sequences of norms $\{||f_i||\}_{i\in I}$ of frames $\{f_i\}_{i\in I}$ with prescribed frame operator S.

In the finite dimensional case Casazza and Leon [12, 13] gave explicit and algorithmic construction of tight frames with prescribed norms. Moreover, Casazza, Fickus, Kovačević, Leon, and Tremain [14] characterized norms of finite tight frames in terms of their "fundamental frame inequality" using frame potential methods of Benedetto and Fickus [5]. An alternative approach using projection decomposition was undertaken by Dykema, Freeman, Kornelson, Larson, Ordower, and Weber [19], which yields some necessary and some sufficient conditions for infinite dimensional Hilbert spaces [29]. A significantly refined eigenstep method for constructing finite frames with prescribed spectrum and diagonal was recently introduced by Cahill, Fickus, Mixon, Poteet, and Strawn [11, 20]. These results are described in Section "Finite dimensional frames".

Significant progress in the area became possible thanks to the Schur-Horn theorem as noted by Antezana, Massey, Ruiz, and Stojanoff [1] and Tropp, Dhillon, Heath, and Strohmer [34]. In particular, the authors of [1] established the following connection between Schur-Horn-type theorems and the existence of frames with prescribed norms and frame operator, see [1, Proposition 4.5] and [6, Proposition 2.3].

Theorem 1 (Antezana-Massey-Ruiz-Stojanoff). *Let S be a positive self-adjoint operator on a Hilbert space $\mathscr{H}$. Let $\{d_i\}_{i\in I}$ be a bounded sequence of positive numbers. Then the following are equivalent:*

1. *there exists a frame $\{f_i\}_{i\in I}$ in $\mathscr{H}$ with the frame operator S such that $d_i = ||f_i||^2$ for all $i \in I$,*

2. *there exists a larger Hilbert space $\mathscr{K} \supset \mathscr{H}$ and a self-adjoint operator E acting on $\ell^2(I)$, which is unitarily equivalent with $S \oplus \mathbf{0}$, where $\mathbf{0}$ is the zero operator acting on $\mathscr{K} \ominus \mathscr{H}$, such that its diagonal $\langle Ee_i, e_i \rangle = d_i$ for all $i \in I$.*

Thus, Problem 1 of characterizing sequences $\{\|f_i\|^2\}_{i\in I}$ for all frames $\{f_i\}_{i\in I}$ with frame operator S is subsumed by the Schur-Horn problem:

Problem 2. Characterize diagonals $\{\langle Ee_i, e_i \rangle\}_{i\in I}$ of a self-adjoint operator E, where $\{e_i\}_{i\in I}$ is any orthonormal basis of $\mathscr{H}$.

In Section "Finite dimensional frames" we discuss the finite dimensional Schur-Horn theorem and its connection to finite dimensional frames. In Section "Infinite dimensional frames" we present the infinite dimensional results. A beautifully simple and complete characterization of Parseval frame norms was given by Kadison [25, 26], which easily extends to tight frames by scaling. The authors [6] have extended this result to the non-tight setting by characterizing frame norms with prescribed optimal frame bounds. The second author [24] has characterized diagonals of self-adjoint operators with three points in the spectrum. This yields a characterization of frame norms whose frame operator has two point spectrum. Finally, the authors [7, 10] have recently extended this result to operators with finite spectrum.

Finite dimensional frames

The classical Schur-Horn theorem [22, 33] characterizes diagonals of self-adjoint (Hermitian) matrices with given eigenvalues. It can be stated as follows, where $\mathscr{H}_N$ is N-dimensional Hilbert space over $\mathbb{R}$ or $\mathbb{C}$, i.e., $\mathscr{H}_N = \mathbb{R}^N$ or $\mathbb{C}^N$.

Theorem 2 (Schur-Horn). *Let $\{\lambda_i\}_{i=1}^N$ and $\{d_i\}_{i=1}^N$ be real sequences in nonincreasing order. There exists a self-adjoint operator $E : \mathscr{H}_N \to \mathscr{H}_N$ with eigenvalues $\{\lambda_i\}$ and diagonal $\{d_i\}$ if and only if*

$$\sum_{i=1}^{N} \lambda_i = \sum_{i=1}^{N} d_i \quad \text{and} \quad \sum_{i=1}^{n} d_i \le \sum_{i=1}^{n} \lambda_i \quad \text{for all } 1 \le n \le N. \tag{2}$$

The necessity of (2) is due to Schur [33] and the sufficiency of (2) is due to Horn [22]. It should be noted that (2) can be stated by the equivalent convexity condition

$$(d_1, \ldots, d_N) \in \operatorname{conv}\{(\lambda_{\sigma(1)}, \ldots, \lambda_{\sigma(N)}) : \sigma \in S_N\}, \tag{3}$$

where S_N is a permutation group on N elements.

Using Theorem 1 we obtain a complete solution to Problem 1 for finite frames.

Theorem 3. *Let S be a positive self-adjoint invertible $M \times M$ matrix with eigenvalues $\{\lambda_i\}_{i=1}^M$ in nonincreasing order. Let $\{d_i\}_{i=1}^N$ be a nonnegative nonincreasing sequence. There exists a frame $\{f_i\}_{i=1}^N$ for $\mathscr{H}_M$ with frame operator S and $\|f_i\|^2 = d_i$ for each $i = 1, \dots, N$ if and only if*

$$\sum_{i=1}^{N} d_i = \sum_{i=1}^{M} \lambda_i \quad \text{and} \quad \sum_{i=1}^{k} d_i \le \sum_{i=1}^{k} \lambda_i \quad \text{for all } 1 \le k \le M.$$

Though this completely answers the question of existence of a frame with prescribed norms and frame operator, it does not give a construction of the desired frame. Indeed, the early proofs of the Schur-Horn theorem were existential, and there have been several recent papers [12, 13, 17, 20] on algorithms for the construction of the matrices in Theorem 2. Therefore, Theorem 3 is not the final word on constructing frames with a given frame operator and set of lengths.

Example 1. Let $\{e_1, e_2\}$ be an orthonormal basis for $\mathbb{C}^2$. Consider the frames $\{f_i\}_{i=1}^4$ and $\{g_i\}_{i=1}^4$ given by

$$f_1 = e_1,\ f_2 = e_2,\ f_3 = e_1,\ f_4 = e_2$$

and

$$g_1 = e_1,\ g_2 = e_2,\ g_3 = 2^{-1/2}(e_1 + e_2),\ g_4 = 2^{-1/2}(e_1 - e_2).$$

A simple calculation shows that each is a frame with frame operator $2\mathbf{I}$ and the norms of the frame vectors are all 1. However, these frames are fundamentally different. Indeed, $\{g_i\}$ is not unitarily equivalent (or even isomorphic) to any reordering of $\{f_i\}$.

Example 1 shows that for a given positive invertible operator S and a sequence of lengths $\{d_i\}$ there may be many different frames with frame operator S and lengths $\{d_i\}$. To understand the set of all frames with a given frame operator and set of lengths authors of [11] consider the problem of constructing every such frame.

For a given vector f in a Hilbert space $\mathscr{H}$, let ff^* denote the rank one operator given by

$$ff^*(g) = \langle g, f\rangle f \qquad \text{for all } g \in \mathscr{H}.$$

We will use the following standard result from linear algebra [23, Propositions 4.3.4 and 4.3.10].

Proposition 1. *Let S be an $M \times M$ self-adjoint matrix with eigenvalue list $\lambda_1 \ge \dots \ge \lambda_M$. If $f \in \mathbb{C}^M$ with $\gamma = \|f\|^2$ and $\mu_1 \ge \dots \ge \mu_M$ is the eigenvalue list of $S + ff^*$, then*

$$\mu_1 \ge \lambda_1 \ge \mu_2 \ge \lambda_2 \ge \dots \ge \mu_M \ge \lambda_M \tag{4}$$

and

$$\sum_{i=1}^{M} \mu_i = \gamma + \sum_{i=1}^{M} \lambda_i. \tag{5}$$

Conversely, for any sequence $\{\mu_i\}_{i=1}^M$ *and* γ *satisfying (4) and (5), there exists a vector* $f \in \mathbb{C}^M$ *with* $\|f\|^2 = \gamma$ *such that* $S + ff^*$ *has eigenvalue list* $\{\mu_i\}_{i=1}^M$.

Two sequences $\{\lambda_i\}_{i=1}^M$ and $\{\mu_i\}_{i=1}^M$ satisfying (4) are said to *interlace*, in symbols $\{\lambda_i\}_{i=1}^M \sqsubseteq \{\mu_i\}_{i=1}^M$. For a frame $\{f_i\}_{i=1}^N$ define the *jth partial frame operator*

$$S_j = \sum_{i=1}^{j} f_i f_i^*.$$

By applying Proposition 1 to the partial frame operators, starting with the zero operator, we obtain the following result [11, Theorem 2].

Theorem 4 (Cahill-Fickus-Mixon-Poteet-Strawn). *Let S be a positive self-adjoint operator with eigenvalue list* $\lambda_1 \geq \ldots \geq \lambda_M > 0$. *Let* $\{d_i\}_{i=1}^N$ *be a nonnegative sequence.*

(i) If there is a sequence of nonincreasing nonnegative sequences $\{\{\lambda_{i,j}\}_{i=1}^M\}_{j=0}^N$ *such that*

$$\lambda_{i,0} = 0 \text{ and } \lambda_{i,N} = \lambda_i \qquad \text{for all } i = 1,\ldots,M, \tag{6}$$

$$\{\lambda_{i,j}\}_{i=1}^M \sqsubseteq \{\lambda_{i,j+1}\}_{i=1}^M, \tag{7}$$

$$\sum_{i=1}^{M} \lambda_{i,j} = \sum_{i=1}^{j} d_i \qquad \text{for each } j = 0,\ldots,N-1, \tag{8}$$

then there exists a frame $\{f_i\}_{i=1}^N$ *with* $\|f_i\|^2 = d_i$ *such that* $\{\lambda_{i,j}\}_{i=1}^M$ *is the eigenvalue sequence of the jth partial frame operator.*

(ii) Conversely, if there exists a frame $\{f_i\}_{i=1}^N$ *with frame operator S,* $\|f_i\|^2 = d_i$ *for* $i = 1,\ldots,N$, *and* $\{\lambda_{i,j}\}_{i=1}^M$ *is the eigenvalue sequence of the jth partial frame operator, then* $\{\{\lambda_{i,j}\}_{i=1}^M\}_{j=0}^N$ *satisfies (6), (7), and (8).*

A doubly indexed sequence $\{\{\lambda_{i,j}\}_{i=1}^M\}_{j=0}^N$ satisfying (6), (7), and (8) is called a sequence of *eigensteps*. In Example 1 we saw that a given frame operator and set of lengths does not determine a frame. In light of Theorem 4 one might think that for a given sequence of eigensteps $\{\{\lambda_{i,j}\}_{i=1}^M\}_{j=0}^N$ there is a unique (up to unitarily equivalence) frame such that $\{\lambda_{i,j}\}_{i=1}^M$ is the eigenvalue sequence of the jth partial frame operator. The following simple example shows that this is not true.

Example 2. Let $d_i = 2$ for $i = 1,2,3$. Let $\lambda_1 = 3$, $\lambda_2 = 2$, and $\lambda_3 = 1$. A sequence of eigensteps is given by the following table

j	0	1	2	3
$\lambda_{3,j}$	0	0	0	1
$\lambda_{2,j}$	0	0	2	2
$\lambda_{1,j}$	0	2	2	3

We may choose $f_1 \in \mathbb{C}^3$ to be any vector with $\|f_1\|^2 = 2$. Next, we may choose f_2 to be any vector with $\|f_2\|^2 = 2$ in $\text{span}\{f_1\}^\perp$. Finally, let $f_3 = x + y$ where

$x \in \text{span}\{f_1, f_2\}$ with $\|x\|^2 = \frac{1}{2}$, and $y \in \text{span}\{f_1, f_2\}^\perp$ with $\|y\|^2 = \frac{3}{2}$. A simple calculation shows that $\{\lambda_{i,j}\}_{i=1}^3$ is the eigenvalue sequence of the jth partial frame operator of $\{f_i\}_{i=1}^3$.

In [11, Theorem 2], Cahill, Fickus, Mixon, Poteet, and Strawn give an algorithm to prove Theorem 4(ii), that is, an algorithm to produce a frame from a given sequence of eigensteps. At step j the vector f_j is chosen and added to the frame so that the partial frame operator has the desired spectrum. As we see in Example 2, at each step there will be a set of choices for the next frame vector. Crucially, their algorithm identifies *all* choices for the jth frame vector, and thus the algorithm constructs all finite frames. In [20, Theorem 5], Fickus, Mixon, Poteet, and Strawn give an explicit algorithm for producing a sequence of eigensteps which they dubbed *Top Kill* since it iteratively "kills" as much as possible from the top portion of staircases starting with the final eigenvalue sequence $\{\lambda_{i,N}\}_{i=1}^M$ and proceeding backward to the zero sequence $\{\lambda_{i,0}\}_{i=1}^M$. Finally, in [20, Theorem 2] they gave an explicit description of all possible sequences of eigensteps as in Theorem 4(i). For details we refer to [11] and [20].

Infinite dimensional frames

The infinite dimensional case of Problem 1 was considered by Kornelson and Larson in [29]. Their result gives a sufficient condition for the sequence of frame norms in terms of the essential norm of the frame operator

$$\|S\|_{ess} = \inf\{\|S+K\| : K \text{ a compact operator}\}.$$

We observe that in light of Kadison's Theorem 6, [29, Proposition 7 and Corollaries 8 and 9] are incorrect as stated, see Example 3. However, this does not affect the correctness of [29, Theorem 6] which can be stated as follows.

Theorem 5 (Kornelson-Larson). *Let S be a positive invertible operator on a Hilbert space $\mathscr{H}$. Suppose that $\{d_i\}_{i\in I}$ is a sequence in $[0,\infty)$ such that*

$$\sum_{i\in I} d_i = \infty \qquad \text{and} \qquad \sup_{i\in I} d_i < \|S\|_{ess}.$$

Then, there exists a frame $\{f_i\}_{i\in I} \subset \mathscr{H}$ with $d_i = \|f_i\|^2$ for all $i \in I$ such that its frame operator is S.

Antezana, Massey, Ruiz, and Stojanoff have refined Theorem 5 by giving a necessary condition [1, Theorem 5.1] and some sufficient conditions [1, Theorem 5.4] for the sequence of frame norms with a given frame operator. Thus, these results give a partial answer to Problem 1.

Kadison's Pythagorean Theorem

The first complete characterization of frame norms was achieved by Kadison for Parseval frames, i.e., when the frame operator $S = \mathbf{I}$, which easily extends to tight frames by scaling. In his influential work [25, 26] Kadison discovered a characterization of diagonals of orthogonal projections acting on $\mathscr{H}$.

Theorem 6 (Kadison). *Let $\{d_i\}_{i\in I}$ be a sequence in $[0,1]$ and $\alpha \in (0,1)$. Define*

$$C(\alpha) = \sum_{d_i<\alpha} d_i, \qquad D(\alpha) = \sum_{d_i\geq\alpha} (1-d_i).$$

Then the following are equivalent:

1. *there exists an orthogonal projection of $\ell^2(I)$ onto a subspace $\mathscr{H}$ with diagonal $\{d_i\}_{i\in I}$,*
2. *there exists a Parseval frame $\{f_i\}_{i\in I}$ on a Hilbert space $\mathscr{H}$ such that $||f_i||^2 = d_i$ for all $i \in I$,*
3. *we have $C(\alpha) = \infty$ or $D(\alpha) = \infty$, or*

$$C(\alpha) < \infty \quad \text{and} \quad D(\alpha) < \infty \quad \text{and} \quad C(\alpha) - D(\alpha) \in \mathbb{Z}. \tag{9}$$

One can easily show that if (9) is satisfied for some $\alpha \in (0,1)$, then (9) holds for all $\alpha \in (0,1)$. Hence, Kadison's Theorem is often formulated for a specific choice of $\alpha = 1/2$. Alternative proofs of Theorem 6 were provided by the authors [8] and Argerami [2].

Example 3. For a given $\eta \in [0,1]$ define a sequence $\{d_i\}_{i\in\mathbb{Z}}$ by

$$\left(\ldots, \frac{1}{2^n}, \ldots, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, \eta, \frac{1}{2}, \frac{3}{4}, \frac{7}{8}, \ldots, \frac{2^n-1}{2^n}, \ldots\right). \tag{10}$$

Since this sequence is symmetric around $1/2$ with the exception of the entry η, we have $C(1/2) - D(1/2) \equiv \eta \mod 1$. Hence, by Theorem 6, $\{d_i\}_{i\in\mathbb{Z}}$ is a diagonal of projection if and only if $\eta = 0$ or $\eta = 1$.

Characterization of frame norms from frame bounds

The non-tight extension of Theorem 6 was considered by the authors [6] who characterized all possible sequences of norms of a frame with prescribed optimal bounds $A < B$. The special tight case $A = B$ follows immediately from Kadison's Pythagorean Theorem 6 by scaling. Theorem 7 can be also thought as an infinite dimensional Schur-Horn theorem for a class of self-adjoint operators with prescribed lower and upper bounds and with zero in the spectrum.

Theorem 7. *Let $0 < A < B < \infty$ and $\{d_i\}$ be a nonsummable sequence in $[0,B]$. Define the numbers*

$$C(A) = \sum_{d_i < A} d_i \quad \text{and} \quad D(A) = \sum_{d_i \geq A} (B - d_i). \tag{11}$$

Then the following are equivalent:

1. *there exists a positive operator E on $\ell^2(I)$ with the spectrum satisfying $\{A,B\} \subseteq \sigma(E) \subseteq \{0\} \cup [A,B]$ and the diagonal $\{d_i\}_{i\in I}$,*
2. *there exists a frame $\{f_i\}_{i\in I}$ on some Hilbert space $\mathscr{H}$ with optimal frame bounds A and B, and $d_i = ||f_i||^2$ for all $i \in I$,*
3. *one of the following holds:*

a. $C(A) = \infty$,
b. $D(A) = \infty$,
c. $C(A), D(A) < \infty$ *and there exists $n \in \mathbb{N}_0$ such that*

$$nA \leq C(A) \leq A + B(n-1) + D(A). \tag{12}$$

Note that the assumption of $\{d_i\}$ being nonsummable in Theorem 7 is not a true limitation. Indeed, the summable case requires more restrictive conditions which are omitted here. One should also emphasize that the non-tight case is not a mere generalization of the tight case $A = B$, since it is qualitatively different from the tight case. Indeed, by setting $A = B$ in Theorem 7 we do not get the correct necessary and sufficient condition (9) previously discovered by Kadison.

Example 4. For a given nonsummable sequence $\{d_i\}_{i\in\mathbb{N}}$ in $[0,1]$ we define

$$\mathscr{A} = \{A \in (0,1] : \exists \text{ frame } \{f_i\}_{i\in\mathbb{N}} \text{ with } ||f_i||^2 = d_i \text{ that has optimal frame bounds } A \text{ and } 1\}.$$

Without loss of generality we can assume that $\sup d_i = 1$. Indeed, if $\sup d_i < 1$, then by Theorem 7 we have always $\mathscr{A} = (0,1]$, and this case is not interesting.

In [6, Theorem 7.1] the authors have shown the set $\mathscr{A} \cup \{0,1\}$ is always closed. In general, determining the set $\mathscr{A}$ is not an easy task since it boils down to checking condition (12) for all possible values of $0 < A < 1$ by computing countably many infinite series (11). If $\{d_i\}_{i\in\mathbb{N}}$ happens to be a geometric series this task actually reduces to checking a finite number of conditions, see [6]. Indeed, for a given $\beta \in (0,1)$ define a sequence $\{d_i\}_{i\in\mathbb{N}}$ by

$$d_i = 1 - \beta^i \qquad \text{for } i = 1,2,\ldots$$

Using Theorem 7 the authors have shown [6] that

$$\mathscr{A} = [(1-2\beta)/(1-\beta), 1-\beta] \qquad \text{for } 0 < \beta < 1/2.$$

Theorem 7 has an analogue for operators without the zero in the spectrum. This result is much easier to show and it leads to a characterization of norms of Riesz bases with prescribed bounds.

Theorem 8. *Let $0 < A \leq B < \infty$ and $\{d_i\}_{i\in I}$ be a sequence in $[A,B]$. Then the following are equivalent:*

1. *there exists a positive operator E on $\ell^2(I)$ with the spectrum satisfying $\{A,B\} \subseteq \sigma(E) \subseteq [A,B]$ and the diagonal $\{d_i\}_{i\in I}$,*
2. *there exists a Riesz basis $\{f_i\}_{i\in I}$ with optimal bounds A and B and $d_i = ||f_i||^2$ for all $i \in I$,*
3. *we have*
$$C(A), D(A) \geq B - A. \tag{13}$$

Characterization of frame norms with finite spectrum frame operator

Another extension of Kadison's result [25, 26] was obtained by the second author [24] who characterized the set of diagonals of operators with three points in the spectrum.

Theorem 9. *Let $0 < A < B < \infty$ and $\{d_i\}_{i\in I}$ be a sequence in $[0,B]$ with $\sum d_i = \sum(B - d_i) = \infty$. Define $C(A)$ and $D(A)$ as in (11). Then the following are equivalent:*

1. *there exists a self-adjoint operator E with diagonal $\{d_i\}_{i\in I}$ and spectrum $\sigma(E) = \{0,A,B\}$,*
2. *there exists a frame $\{f_i\}_{i\in I}$ on some Hilbert space $\mathscr{H}$ with $d_i = ||f_i||^2$ for all $i \in I$ such that its frame operator S has spectrum $\sigma(S) = \{A,B\}$,*
3. *one of the following holds:*

   a. $C(A) = \infty$,
   b. $D(A) = \infty$,
   c. $C(A), D(A) < \infty$ and there exists $N \in \mathbb{N}$ and $k \in \mathbb{Z}$ such that

$$C(A) - D(A) = NA + kB \quad \text{and} \quad C(A) \geq (N+k)A.$$

By combining Theorems 6 and 9 we can show the existence of frames with at most two point spectrum for every nonsummable sequence of frame norms.

Theorem 10. *Let $\{d_i\}_{i\in I}$ be a sequence in $[0,B]$ such that $\sum d_i = \sum(B - d_i) = \infty$, where $B > 0$. Then, there exists a frame $\{f_i\}_{i\in I}$ with $||f_i||^2 = d_i$ for all $i \in I$ such that its frame operator S satisfies either:*

1. *$S = B\mathbf{I}$, i.e., $\{f_i\}_{i\in I}$ is a tight frame with frame bounds $A = B$, or*
2. *S has spectrum $\sigma(S) = \{A,B\}$ for some $0 < A < B$.*

Proof. If $C(B/2) = \infty$ or $D(B/2) = \infty$, then by Kadison's Theorem 6, there exists a tight frame $\{f_i\}_{i\in I}$ with frame bounds $A = B$. The same conclusion holds when $C(B/2) < \infty$ and $D(B/2) < \infty$, and $C(B/2) - D(B/2) \in B\mathbb{Z}$. Hence, without loss of generality we can assume that $C(B/2) < \infty$, $D(B/2) < \infty$, and

$$C(B/2) - D(B/2) \equiv \eta \mod B \qquad \text{for some } 0 < \eta < B.$$

Using Theorem 9 the authors have shown in [9, Theorem 3.4] that there exists a self-adjoint operator E on $\ell^2(I)$ with diagonal $\{d_i\}_{i\in I}$ and spectrum $\sigma(S) = \{0, A, B\}$, where $A = \eta$. Applying Theorem 1 yields conclusion 2 of Theorem 10.

Finally, the authors [7] showed the following characterization of diagonals of self-adjoint operators with finite spectrum. A variant of Theorem 11 with prescribed multiplicities, which has a more complicated statement and proof, was shown by the authors in [10]. In light of Theorem 1, the results in [7, 10] answer Problem 1 in the case when frame operator S has finite spectrum.

Theorem 11. *Let $\{A_j\}_{j=0}^{n+1}$ be an increasing sequence of real numbers such that $A_0 = 0$ and $A_{n+1} = B$, $n \in \mathbb{N}$. Let $\{d_i\}_{i\in I}$ be a sequence in $[0, B]$ with $\sum d_i = \sum(B - d_i) = \infty$. For each $\alpha \in (0, B)$, define*

$$C(\alpha) = \sum_{d_i < \alpha} d_i \quad \text{and} \quad D(\alpha) = \sum_{d_i \geq \alpha} (B - d_i). \tag{14}$$

Then the following are equivalent:

1. *there exists a self-adjoint operator E with diagonal $\{d_i\}_{i\in I}$ and $\sigma(E) = \{A_0, A_1, \ldots, A_{n+1}\}$,*
2. *there exists a frame $\{f_i\}_{i\in I}$ on some Hilbert space $\mathscr{H}$ with $d_i = ||f_i||^2$ for all $i \in I$ such that its frame operator S has spectrum $\sigma(S) = \{A_1, \ldots, A_{n+1}\}$,*
3. *one of the following holds:*

   a. *$C(B/2) = \infty$,*
   b. *$D(B/2) = \infty$,*
   c. *$C(B/2) < \infty$ and $D(B/2) < \infty$ (and thus $C(\alpha), D(\alpha) < \infty$ for all $\alpha \in (0, B)$), and there exist $N_1, \ldots, N_n \in \mathbb{N}$ and $k \in \mathbb{Z}$ such that*

$$C(B/2) - D(B/2) = \sum_{j=1}^{n} A_j N_j + kB, \tag{15}$$

and for all $r = 1, \ldots, n$,

$$(B - A_r)C(A_r) + A_r D(A_r) \geq (B - A_r) \sum_{j=1}^{r} A_j N_j + A_r \sum_{j=r+1}^{n} (B - A_j) N_j. \tag{16}$$

We remark that the assumption $\sum d_i = \sum(B - d_i) = \infty$ is not a true limitation of Theorems 9 and 11. Indeed, the summable case $\sum d_i < \infty$, or its symmetric variant $\sum(B - d_i) < \infty$, leads to a finite rank Schur-Horn theorem. This case requires

a different set of conditions which are closely related to the classical Schur-Horn majorization. Finally, the assumption $A_0 = 0$ is made only for simplicity; the general case follows immediately by a translation argument.

We shall illustrate Theorem 11 by considering the problem of describing possible frame operators with prescribed frame norms. This can be thought as the converse to Problem 1. A rigorous formulation of this problem and a solution was shown by the authors in [10, Theorem 8.2]. Here we shall concentrate only on the spectral variant of this problem studied in [9].

Definition 2. Suppose that $\{d_i\}_{i\in I}$ is a sequence in $[0,1]$. For a given $n \in \mathbb{N}$ and a sequence $\{d_i\}_{i\in I}$ in $[0,1]$ we consider the set

$$\mathscr{A}_n(\{d_i\}) = \big\{(A_1,\dots,A_n) \in (0,1)^n : \forall_{j\neq k}\, A_j \neq A_k \;\exists \text{ frame } \{f_i\}_{i\in I} \text{ s. t.}$$
$$\forall_{i\in I}\, d_i = ||f_i||^2 \text{ and spectrum of frame operator } \sigma(S) = \{A_1,\dots,A_n,1\}\big\}.$$

Subsequently we shall assume that $\sum d_i = \sum(1-d_i) = \infty$, so that we can apply Theorem 11. The authors have shown in [9, Theorem 3.8] that the sets $\mathscr{A}_n(\{d_i\})$ are nonempty for each $n \geq 2$ provided that infinitely many d_is satisfy $d_i \in (0,1)$. However, the set $\mathscr{A}_1(\{d_i\})$ might be empty as illustrated by Example 5. Moreover, the second author [24, Theorem 7.1] has shown that

$$\mathscr{A}_1(\{d_i\}) = \begin{cases} (0,1) & \text{if } C(1/2) = \infty \text{ or } D(1/2) = \infty, \\ \text{a finite set} & \text{if } C(1/2), D(1/2) < \infty. \end{cases}$$

Example 5. Let $\beta \in (0,1/2)$ and define the sequence $\{d_i\}_{i\in\mathbb{Z}\setminus\{0\}}$ by

$$d_i = \begin{cases} 1-\beta^i & i > 0 \\ \beta^{-i} & i < 0. \end{cases}$$

Using Theorem 9 the second author has shown in [24] that

$$\mathscr{A}_1(\{d_i\}) = \begin{cases} \{\frac{1}{3},\frac{1}{2},\frac{2}{3}\} & \frac{-1+\sqrt{13}}{6} \leq \beta < 1/2, \\ \{\frac{1}{2}\} & 1/3 \leq \beta < \frac{-1+\sqrt{13}}{6}, \\ \varnothing & 0 < \beta < 1/3. \end{cases}$$

The following result [9, Corollary 3.13] describes the spectral sets $\mathscr{A}_2(\{d_i\})$.

Theorem 12. *Let $\{d_i\}_{i\in I}$ be a sequence in $[0,1]$. If $C(1/2) = \infty$ or $D(1/2) = \infty$, then*

$$\mathscr{A}_2(\{d_i\}) = (0,1)^2 \setminus \Delta, \qquad \textit{where } \Delta = \{(x,x) : x \in (0,1)\}.$$

Otherwise, if $C(1/2), D(1/2) < \infty$, then the set $\mathscr{A}_2(\{d_i\})$ is nonempty and it consists of a countable union of line segments. Moreover, one end point of each of these line segments must lie in the boundary of the unit square.

Example 6. For a given $\eta \in [0,1]$ consider the sequence $\{d_i\}_{i\in\mathbb{Z}\setminus\{0\}}$ from Example 3. Using the characterization from Theorem 11 and numerical calculations performed with *Mathematica*, Figures 1, 2, 3, and 4 depict the set $\mathscr{A}_2(\{d_i\})$ for different values of the parameter η.

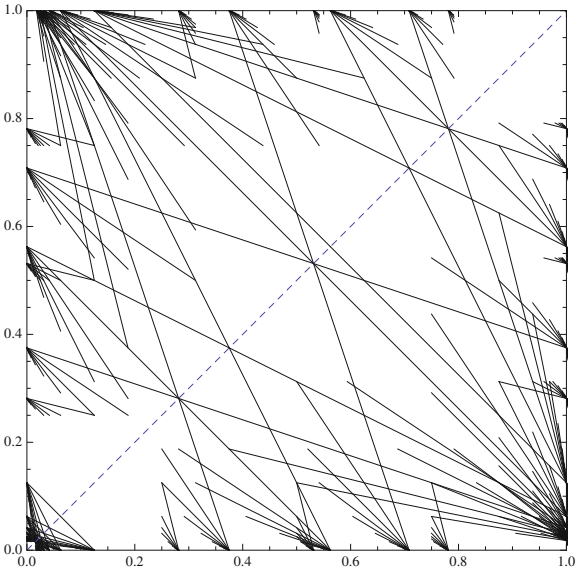


Fig. 1 The set $\mathscr{A}_2(\{d_i\})$ for the sequence (10) with $\eta = \frac{1}{8}$.

Other Schur-Horn-type theorems

We shall finish this chapter by mentioning other important developments in extending the Schur-Horn Theorem 2 to infinite dimensional Hilbert spaces. Neumann [32] gave an infinite dimensional version of the Schur-Horn theorem phrased in terms of ℓ^∞ closure of the convexity condition (3). Neumann's result can be considered an initial, albeit somewhat crude, solution to Problem 2. The first fully satisfactory progress was achieved by Kadison [25, 26] who solved Problem 2 for orthogonal projections. The work by Gohberg and Markus [21] and Arveson and Kadison [4] extended the Schur-Horn Theorem 2 to positive trace class operators. This has been further extended to compact positive operators by Kaftal and Weiss [28]. These results are stated in terms of majorization inequalities as in (2). Other notable progress includes the work of Arveson [3] on diagonals of normal operators with finite spectrum. For a detailed survey of recent progress on infinite Schur-Horn majorization theorems and their connections to operator ideals we refer to the paper of Kaftal and Weiss [27].

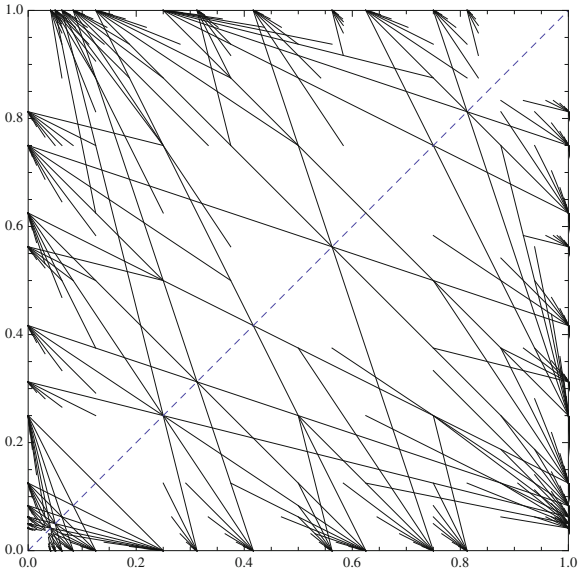


Fig. 2 The set $\mathscr{A}_2(\{d_i\})$ for the sequence (10) with $\eta = \frac{1}{4}$.

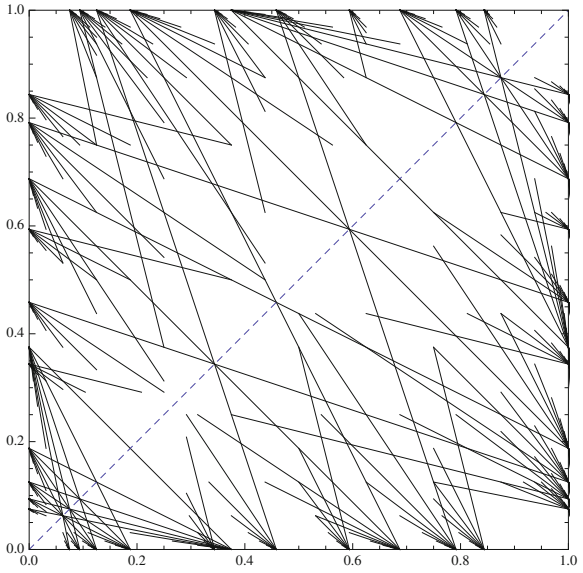


Fig. 3 The set $\mathscr{A}_2(\{d_i\})$ for the sequence (10) with $\eta = \frac{3}{8}$.

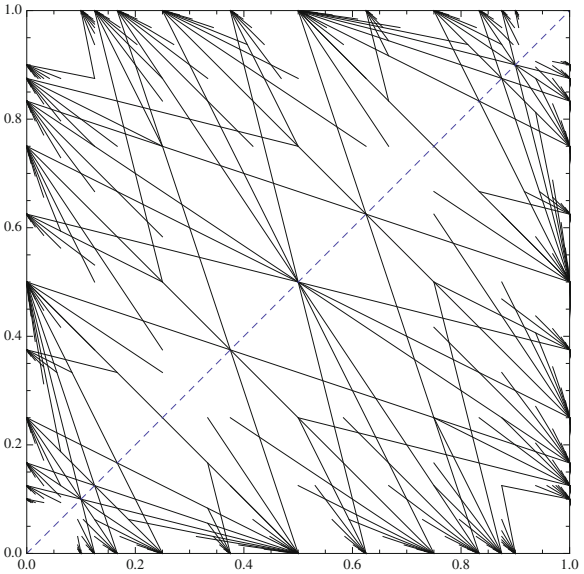


Fig. 4 The set $\mathscr{A}_2(\{d_i\})$ for the sequence (10) with $\eta = \frac{1}{2}$.

References

1. J. Antezana, P. Massey, M. Ruiz, D. Stojanoff, The Schur-Horn theorem for operators and frames with prescribed norms and frame operator. Illinois J. Math. **51**, 537–560 (2007)
2. M. Argerami, Majorisation and the carpenter's theorem. Integr. Equ. Oper. Theory **82**(1), 33–49 (2015)
3. W. Arveson, Diagonals of normal operators with finite spectrum. Proc. Natl. Acad. Sci. USA **104**, 1152–1158 (2007)
4. W. Arveson, R. Kadison, Diagonals of self-adjoint operators, in *Operator Theory, Operator Algebras, and Applications*, ed. by M. Amélia Bastos, A. Lebre, S. Samko, I.M. Spitkovsky. Contemporary Mathematics, vol. 414 (American Mathematical Society, Providence, 2006), pp. 247–263
5. J. Benedetto, M. Fickus, Finite normalized tight frames. Adv. Comput. Math. **18**, 357–385 (2003)
6. M. Bownik, J. Jasper, Characterization of sequences of frame norms. J. Reine Angew. Math. **654**, 219–244 (2011)
7. M. Bownik, J. Jasper, Diagonals of self-adjoint operators with finite spectrum (2014, preprint)
8. M. Bownik, J. Jasper, Constructive proof of the Carpenter's theorem. Can. Math. Bull. **57**, 463–476 (2014)
9. M. Bownik, J. Jasper, Spectra of frame operators with prescribed frame norms, in *Harmonic Analysis and Partial Differential Equations*, ed. by P. Cifuentes et al., Contemporary Mathematics, vol. 612 (American Mathematical Society, Providence, 2014), pp. 65–79
10. M. Bownik, J. Jasper, The Schur-Horn theorem for operators with finite spectrum. Trans. Am. Math. Soc. **367**(7), 5099–5140 (2015)
11. J. Cahill, M. Fickus, D.G. Mixon, M.J. Poteet, N. Strawn, Constructing finite frames of a given spectrum and set of lengths. Appl. Comput. Harmon. Anal. **35**, 52–73 (2013)
12. P. Casazza, M. Leon, *Existence and construction of finite tight frames*. J. Concr. Appl. Math. **4**, 277–289 (2006)

13. P. Casazza, M. Leon, Existence and construction of finite frames with a given frame operator. Int. J. Pure Appl. Math. **63**, 149–157 (2010)
14. P. Casazza, M. Fickus, J. Kovačević, M. Leon, J. Tremain, A physical interpretation of tight frames, in *Harmonic Analysis and Applications*, ed. by C. Heil. Applied and Numerical Harmonic Analysis (Birkhäuser Boston, Boston, 2006), pp. 51–76
15. O. Christensen, *An Introduction to Frames and Riesz Bases*. Applied and Numerical Harmonic Analysis (Birkhäuser Boston Inc., Boston, MA, 2003)
16. I. Daubechies, *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61 (SIAM, Philadelphia, PA, 1992)
17. I.S. Dhillon, R.W. Heath Jr., M. Sustik, J.A. Tropp, Generalized finite algorithms for constructing Hermitian matrices with prescribed diagonal and spectrum. SIAM J. Matrix Anal. Appl. **27**, 61–71 (2005)
18. R.J. Duffin, A.C. Schaeffer, *A class of nonharmonic Fourier series*. Trans. Am. Math. Soc. **72**, 341–366 (1952)
19. K. Dykema, D. Freeman, K. Kornelson, D. Larson, M. Ordower, E. Weber, Ellipsoidal tight frames and projection decompositions of operators. Illinois J. Math. **48**, 477–489 (2004)
20. M. Fickus, D.G. Mixon, M.J. Poteet, N. Strawn, Constructing all self-adjoint matrices with prescribed spectrum and diagonal. Adv. Comput. Math. **39**, 585–609 (2013)
21. I.C. Gohberg, A.S. Markus, Some relations between eigenvalues and matrix elements of linear operators. Mat. Sb. (N.S.) **64**, 481–496 (1964)
22. A. Horn, Doubly stochastic matrices and the diagonal of a rotation matrix. Am. J. Math. **76**, 620–630 (1954)
23. R. Horn, C. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, 1985)
24. J. Jasper, The Schur-Horn theorem for operators with three point spectrum. J. Funct. Anal. **265**, 1494–1521 (2013)
25. R. Kadison, The Pythagorean theorem. I. The finite case. Proc. Natl. Acad. Sci. USA **99**, 4178–4184 (2002)
26. R. Kadison, The Pythagorean theorem. II. The infinite discrete case. Proc. Natl. Acad. Sci. USA **99**, 5217–5222 (2002)
27. V. Kaftal, G. Weiss, A survey on the interplay between arithmetic mean ideals, traces, lattices of operator ideals, and an infinite Schur-Horn majorization theorem, in *Hot Topics in Operator Theory*, ed. by R.G. Douglas, J. Esterle, D. Gaspar, D. Timotin, F.-H. Vasilescu. Theta Series in Advanced Mathematics, vol. 9 (Theta, Bucharest, 2008), pp. 101–135
28. V. Kaftal, G. Weiss, An infinite dimensional Schur-Horn theorem and majorization theory. J. Funct. Anal. **259**, 3115–3162 (2010)
29. K. Kornelson, D. Larson, Rank-one decomposition of operators and construction of frames, in *Wavelets, Frames and Operator Theory*, ed. by C. Heil, P.E.T Jorgensen, D.R. Larson. Contemporary Mathematics, vol. 345 (American Mathematical Society, Providence, RI, 2004), pp. 203–214
30. S. Mallat, *A Wavelet Tour of Signal Processing* (Academic, San Diego, CA, 1998)
31. Y. Meyer, *Wavelets and Operators* (Cambridge University Press, Cambridge, 1992)
32. A. Neumann, An infinite-dimensional version of the Schur-Horn convexity theorem. J. Funct. Anal. **161**, 418–451 (1999)
33. I. Schur, Über eine Klasse von Mittelbildungen mit Anwendungen auf die Determinantentheorie. Sitzungsber. Berl. Math. Ges. **22**, 9–20 (1923)
34. J. Tropp, I. Dhillon, R. Heath, T. Strohmer, Designing structured tight frames via an alternating projection method. IEEE Trans. Inf. Theory **51**, 188–209 (2005)

Part XIV
Sparsity

The notion of *sparsity* plays a fundamental role in modern harmonic analysis. In common English its meaning is often described as small in numbers or amount, or thinly dispersed or scattered over a large area. The mathematical understanding of sparsity, although similar in principle, has a much more positive outlook: it means that the analyzed signal can be represented in such a way that only a few coefficients may play a significant role in the description of the data.

Examples of areas of mathematics which depend on sparse representations abound. Besides data compression (naturally), they include sampling theory, compressed sensing, and various topics in numerical linear algebra. Although the importance of sparsity became truly evident with the emergence of Big Data or Data Science, the theory is deeply rooted in classical mathematics. For example, the simplex algorithm of George Dantzig (who received his bachelor's degrees in mathematics and physics from the University of Maryland) has been an inspiration for many optimization problems that rely on sparsity of the underlying data.

The phase retrieval problem is another example of a scientific field, where sparsity made an enormous impact. Phase retrieval deals with the loss of information about the phase, which may occur in physical measurements. The problem arose in x-ray crystallography, where it appears in the context of determining the structure of a crystal from the diffraction data. It was solved by Hebert Hauptman (another alumnus of the University of Maryland, with a Ph.D. in mathematics under the direction of Professor Richard Good) and Jerome Karle, who were awarded the Nobel Prize in Chemistry for their discovery. The problem also appears naturally in such diverse fields as electron microscopy, quantum state tomography, and diffraction imaging. DUSTIN MIXSON discusses some of these most recent applications of the phase retrieval problem, together with a detailed survey of the most recent mathematical results in phase retrieval. In this chapter the reader is presented with some of the state-of-the-art results on the role of injectivity in phase transition, rich in mathematical details and bibliographic references.

PHILIP SCHNITER and SUNDEEP RANGAN continue with the themes of sparsity and phase retrieval, detailing for us a novel, probabilistic approach to the phase retrieval problem. Their approach is based on the generalized approximate message passing algorithm, which is an example of Bayesian inference strategy and which helps reveal probabilistic structures in analyzed problems. The performance of the algorithm is illustrated with many numerical results, which show its excellent phase transition behavior, robustness, and low computational complexity.

Another important application of sparsity is proposed in the next chapter, where IVAN W. SELESNICK discusses its role in signal denoising. Denoising (also known as noise reduction) is another example of a classical problem, which we can now view in a new light thanks to results from sparse sampling and compressive sensing. SELESNICK introduces sparsity-assisted signal smoothing - a new algorithm for denoising of 1-dimensional signals. Thanks to formulating the problem of noise reduction as a sparse-regularized linear inverse problem, he is able to take advantage of existing fast solvers, producing as a result a computationally efficient algorithm. The chapter provides a high level of mathematical detail, together with well-explained applications.

The last chapter in this part focuses on the relationship between sparsity assumptions and sampling strategies. RACHEL WARD introduces here the concept of importance sampling. This intuitive idea of sampling signals with respect to a density which takes into account the significance of certain regions for the given signal is presented from the perspective of its applications to three important problems in modern signal processing: stochastic optimization, compressive sensing, and low-rank matrix approximation.

Phase Transitions in Phase Retrieval

Dustin G. Mixon

Abstract Consider a scenario in which an unknown signal is transformed by a known linear operator, and then the pointwise absolute value of the unknown output function is reported. This scenario appears in several applications, and the goal is to recover the unknown signal – this is called phase retrieval. Phase retrieval has been a popular subject of research in the last few years, both in determining whether complete information is available with a given linear operator and in finding efficient and stable phase retrieval algorithms in the cases where complete information is available. Interestingly, there are a few ways to measure information completeness, and each way appears to be governed by a phase transition of sorts. This chapter will survey the state of the art with some of these phase transitions, and identify a few open problems for further research.

Key words: Phase retrieval, Phase transition, Informationally complete, Full spark, Almost injectivity, Unit norm tight frames

Introduction

Various applications feature an inverse problem called *phase retrieval*, in which one is given the pointwise absolute value of a known linear transformation of the desired signal. Note that such information will never completely determine the unknown signal since, for example, negating the input will lead to the same output. Indeed, the best one can do is recover $\{\omega x : |\omega| = 1\}$ if x is the unknown signal, but this global phase factor of ambiguity tends not to be an issue in application (for example, in some applications, the true signal is actually nonnegative everywhere,

D.G. Mixon
Department of Mathematics and Statistics, Air Force Institute of Technology,
Wright-Patterson AFB, OH 45433, USA
e-mail: dustin.mixon@afit.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_5

thereby removing all ambiguity). What follows is a brief overview of some of the applications of phase retrieval:

- *Coherent diffractive imaging.* This is a technique to image a nanoscale object by striking it with a highly coherent beam of X-rays to produce a diffraction pattern. The diffraction pattern is the Fourier transform of the object, but only the intensity of the pattern can be physically measured (by counting photons in different regions) [11, 33, 39, 41].
- *Optics.* This application enjoys various instances of phase retrieval: (1) When imaging a star by a lens, one receives the pointwise absolute value of the Fourier transform of the desired pupil distribution [53]. (2) For a high-resolution image, one can apply interferometric techniques to approximate the spatial coherence function (which is the Fourier transform of the desired object intensity), though the phase of this function is difficult to estimate accurately, so it is discarded [21]. (3) Soon after NASA launched its Hubble Space Telescope, they discovered that the primary mirror in the telescope suffered from a large spherical aberration; the extent of this aberration was established by determining the pupil function from the intensity of its Fourier transform (the point spread function) [29].
- *Quantum state tomography.* When measuring a pure state (i.e., a unit vector) x with a positive operator–valued measure of rank-1 elements $\{\varphi_n\varphi_n^*\}_{n=1}^N$ (i.e., the outer products of Parseval frame elements), the random outcome Z of the measurement has a distribution given by $\Pr(Z = n) = |\langle x, \varphi_n\rangle|^2$. As such, repeated measurements will produce an empirical estimate of the distribution, which is the pointwise absolute value (squared) of a linear transformation of the desired signal x [30, 31, 35].
- *Speech processing.* One common method of speech signal denoising is to take the short-time Fourier transform (STFT) and perform a smoothing operation on the magnitudes of the coefficients [4]. Instead of inverting the STFT with the noisy phases, one can recover the denoised version by phase retrieval [49].

Though there are many applications of phase retrieval, the task is often impossible; in particular, discarding the phases of the Fourier transform is not at all injective. This fact has led many researchers to invoke a priori knowledge of the desired signal, since injectivity might be gotten by restricting to a smaller signal class. For example, for the optics applications, the pupil distribution is only supported within the aperture of the optical system, and so this compact-support constraint combined with the intensity measurements might uniquely determine the desired signal. Introducing such information has led to various ad hoc phase retrieval algorithms, and while they have found some success (e.g., in correcting the Hubble Space Telescope), such algorithms often fail to work unexpectedly. Overall, this route has yet to produce algorithms with practical performance guarantees.

Thankfully, an alternative route was introduced in 2006 by Balan, Casazza, and Edidin [4]: Seek injectivity, not by restricting to a smaller signal class, but rather by designing a larger ensemble of measurement vectors. Unbeknownst to Balan et al. at the time, this idea had already been in the air in the quantum community (for

quantum state tomography [30, 31]), but posing the idea to the signal processing community led to a flurry of research in search of practical phase retrieval guarantees [2, 3, 5, 7, 13–16, 23, 26, 51, 52]. One popular method called *PhaseLift* recasts phase retrieval as a semidefinite program [13–16], another called *PhaseCut* reformulates it in terms of MaxCut [51, 52], and yet another uses the polarization identity along with angular synchronization to quickly solve certain instances [2, 7]. In this same line of research, a new methodology for coherent diffractive imaging emerged [15]: Instead of taking a single exposure and attempting phase retrieval with possibly incomplete information, take multiple exposures of the same object with different masks or diffraction gratings. Not only can such a process produce complete information, there are also provably efficient (and apparently stable) phase retrieval algorithms for this setting [7, 14]. Considering phase retrieval has a wide range of applications, it would be interesting to find other areas to apply this philosophy of taking more measurements to obtain injectivity.

Another effect of the paper [4] by Balan, Casazza, and Edidin was the community's desire for a deeper understanding of injectivity for phase retrieval. In particular, what are the conditions for injectivity, and how many measurements are required? Understanding this will help in determining how many (and what type of) exposures are necessary for coherent diffractive imaging; also, such phase retrieval results can be directly interpreted as fundamental limits of quantum state tomography. The purpose of this chapter is to survey several results along these lines, and to identity some remaining open problems. Section "Injectivity" focuses on injectivity in both the real and complex cases, and then the third section considers a relaxed version of injectivity called *almost injectivity*. First, we discuss the notation we use throughout this chapter as well as some preliminaries:

Notation and Preliminaries

Given a collection of vectors $\Phi = \{\varphi_n\}_{n=1}^N$ in $V = \mathbb{R}^M$ or $\mathbb{C}^M$, we will identify such a collection with the $M \times N$ matrix whose columns form the collection. Consider the intensity measurement process defined by

$$(\mathscr{A}(x))(n) := |\langle x, \varphi_n \rangle|^2.$$

Note that $\mathscr{A}(x) = \mathscr{A}(y)$ whenever $y = cx$ for some scalar c of unit modulus. As such, the mapping $\mathscr{A}: V \to \mathbb{R}^N$ is necessarily not injective. To resolve this (technical) issue, throughout this chapter, we consider sets of the form V/S, where V is a vector space and S is a multiplicative subgroup of the field of scalars. By this notation, we mean to identify vectors $x, y \in V$ for which there exists a scalar $c \in S$ such that $y = cx$; we write $y \equiv x \bmod S$ to convey this identification. Most (but not all) of the time, V/S is either $\mathbb{R}^M/\{\pm 1\}$ or $\mathbb{C}^M/\mathbb{T}$ (here, $\mathbb{T}$ is the complex unit circle), and we view the intensity measurement process as a mapping $\mathscr{A}: V/S \to \mathbb{R}^N$; it is in this way that we will consider the measurement process to be injective.

As the title suggests, the focus of this chapter is phase transitions in phase retrieval. As an example of a phase transition, consider what it takes for a collection of members of a vector space V to span V. Certainly, no collection of size less than the dimension of V has a chance of spanning, and we expect most collections of size at least the dimension to span. As such, we might say that the notion of spanning V exhibits a phase transition at the dimension of V. We make this definition explicit in the following:

Definition 1. Let $\mathrm{A}[\Phi;\mathbb{F}^{M\times N}]$ be a statement about a matrix $\Phi\in\mathbb{F}^{M\times N}$, and consider a function $f\colon\mathbb{N}\to\mathbb{N}$. We say $\mathrm{A}[\Phi;\mathbb{F}^{M\times N}]$ exhibits a *phase transition* at $N=f(M)$ if for each $M\geq 2$,

(a) $\mathrm{A}[\Phi;\mathbb{F}^{M\times N}]$ does not hold whenever $N<f(M)$, and
(b) for each $N\geq f(M)$, there exists an open, dense subset $S\subseteq\mathbb{F}^{M\times N}$ such that $\mathrm{A}[\Phi;\mathbb{F}^{M\times N}]$ holds for every $\Phi\in S$.

Based on experience, both parts (a) and (b) of a phase transition are established by first studying necessary and sufficient conditions for the property of interest $\mathrm{A}[\Phi;\mathbb{F}^{M\times N}]$. For part (b) in particular, algebraic geometry consistently plays a key role. Viewing $\Phi\in\mathbb{F}^{M\times N}$ as a point in real Euclidean space, consider the collection of Φ's for which $\mathrm{A}[\Phi;\mathbb{F}^{M\times N}]$ does not hold. If there exists a real, nontrivial algebraic variety that contains these points (which is often the case), then every point Φ in the complement of the variety (which is open and dense in $\mathbb{F}^{M\times N}$) satisfies $\mathrm{A}[\Phi;\mathbb{F}^{M\times N}]$. As such, for part (b), it suffices to identify the appropriate variety.

Throughout this chapter, we will continually follow a certain procedure for studying phase transitions. Given a property $\mathrm{A}[\Phi;\mathbb{F}^{M\times N}]$, we start by studying various necessary and sufficient conditions for that property. We then attempt to prove a phase transition $N=f(M)$ using the conditions available. Later, we consider explicit constructions of $M\times f(M)$ matrices which satisfy $\mathrm{A}[\Phi;\mathbb{F}^{M\times f(M)}]$; these minimal constructions are certainly mathematically interesting, and they are also optimal measurement designs for applications like quantum state tomography.

Injectivity

In this section, we study the phase transition for injectivity. As we will see, this phase transition is much better understood in the real case than in the complex case, and the distinction is rather interesting. It is highly recommended that the reader enjoys the real case before venturing into the complex case.

Injectivity in the Real Case

We start by defining the important concepts of this subsection:

Definition 2.

(a) $\mathrm{Inj}[\Phi;\mathbb{R}^{M\times N}]$ denotes the statement that $\mathscr{A}\colon \mathbb{R}^M/\{\pm 1\} \to \mathbb{R}^N$ defined by $(\mathscr{A}(x))(n) := |\langle x,\varphi_n\rangle|^2$ is injective, where $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{R}^M$.
(b) $\mathrm{CP}[\Phi;\mathbb{F}^{M\times N}]$ denotes the statement that $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{F}^M$ satisfies the *complement property*: for every $S \subseteq \{1,\dots,N\}$, either $\{\varphi_n\}_{n\in S}$ or $\{\varphi_n\}_{n\in S^c}$ spans $\mathbb{F}^M$.

Interestingly, the complement property characterizes injectivity in the real case. Much of the work in this chapter was inspired by the proof of the following result:

Theorem 1 (Theorem 2.8 in [4]). $\mathrm{Inj}[\Phi;\mathbb{R}^{M\times N}]$ *if and only if* $\mathrm{CP}[\Phi;\mathbb{R}^{M\times N}]$. *In words, $\mathscr{A}$ is injective if and only if Φ satisfies the complement property.*

Proof. We will prove both directions by obtaining the contrapositives.

($\Rightarrow$) Assume Φ does not satisfy the complement property. Then there exists $S \subseteq \{1,\dots,N\}$ such that neither $\{\varphi_n\}_{n\in S}$ nor $\{\varphi_n\}_{n\in S^c}$ spans $\mathbb{R}^M$. This implies that there are nonzero vectors $u,v \in \mathbb{R}^M$ such that $\langle u,\varphi_n\rangle = 0$ for all $n \in S$ and $\langle v,\varphi_n\rangle = 0$ for all $n \in S^c$. For each n, we then have

$$\begin{aligned}|\langle u \pm v,\varphi_n\rangle|^2 &= |\langle u,\varphi_n\rangle|^2 \pm 2\,\mathrm{Re}\langle u,\varphi_n\rangle\overline{\langle v,\varphi_n\rangle} + |\langle v,\varphi_n\rangle|^2\\ &= |\langle u,\varphi_n\rangle|^2 + |\langle v,\varphi_n\rangle|^2.\end{aligned}$$

Since $|\langle u+v,\varphi_n\rangle|^2 = |\langle u-v,\varphi_n\rangle|^2$ for every n, we have $\mathscr{A}(u+v) = \mathscr{A}(u-v)$. Moreover, u and v are nonzero by assumption, and so $u+v \neq \pm(u-v)$.

($\Leftarrow$) Assume that $\mathscr{A}$ is not injective. Then there exist vectors $x,y \in \mathbb{R}^M$ such that $x \neq \pm y$ and $\mathscr{A}(x) = \mathscr{A}(y)$. Taking $S := \{n : \langle x,\varphi_n\rangle = -\langle y,\varphi_n\rangle\}$, we have $\langle x+y,\varphi_n\rangle = 0$ for every $n \in S$. Otherwise when $n \in S^c$, we have $\langle x,\varphi_n\rangle = \langle y,\varphi_n\rangle$ and so $\langle x-y,\varphi_n\rangle = 0$. Furthermore, both $x+y$ and $x-y$ are nontrivial since $x \neq \pm y$, and so neither $\{\varphi_n\}_{n\in S}$ nor $\{\varphi_n\}_{n\in S^c}$ spans $\mathbb{R}^M$.

Having identified an equivalent condition (the complement property) to injectivity in the real case, we now use this condition to identify the phase transition. First, we note that a spanning set for $\mathbb{R}^M$ must have size at least M. As such, we know $\mathrm{CP}[\Phi;\mathbb{R}^{M\times N}]$ does not hold whenever $N < 2M-1$, since taking S to be the first $M-1$ members will leave S^c with $\leq M-1$ members. This suggests a phase transition of $N = 2M-1$, but it remains to prove part (b). To get this, we first introduce the notion of full spark: An $M\times N$ matrix with $M \leq N$ is said to be *full spark* if every $M\times M$ submatrix is invertible. Note that any full spark Φ with $N \geq 2M-1$ necessarily satisfies the complement property, since the larger of S and S^c necessarily has at least M elements, which necessarily span. As such, it suffices to show that full spark matrices form an open and dense subset. To this end, we note that the product of the determinants of all $M\times M$ submatrices forms a polynomial of the matrix entries whose zero set contains every Φ such that $\mathrm{CP}[\Phi;\mathbb{R}^{M\times N}]$ does not hold. Moreover, this polynomial is nonzero since the Vandermonde matrix

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & N \\ \vdots & \vdots & & \vdots \\ 1^{M-1} & 2^{M-1} & \cdots & N^{M-1} \end{bmatrix}$$

is full spark. The complement of this polynomial's zero set is therefore open and dense, as desired. This implies our first phase transition result:

Theorem 2 ([4]). $\mathrm{Inj}[\Phi;\mathbb{R}^{M\times N}]$ *exhibits a phase transition at* $N = 2M-1$.

Now that we have identified the phase transition, we consider the minimal constructions, i.e., the $M \times (2M-1)$ real matrices which satisfy the complement property. Here, we note that for every S of size M, S^c has size $M-1$, meaning S must index a spanning set. As such, the matrices in this extreme case are precisely the $M \times (2M-1)$ full spark matrices. For more information about full spark matrices, see [1].

Injectivity in the Complex Case

In the previous subsection, we quickly identified a characterization of injectivity in the real case that enabled the phase transition of interest to be completely studied. The complex case appears to be a bit more involved. For example, the actual phase transition is the subject of an open conjecture, though there has been a lot of progress on this conjecture recently. Since the complex case is so much more involved, this subsection is broken into different labeled parts, concerning necessary and sufficient conditions, the phase transition, and minimal constructions.

Conditions for Injectivity in the Complex Case

We begin by defining our symbol for injectivity in the complex case:

Definition 3. $\mathrm{Inj}[\Phi;\mathbb{C}^{M\times N}]$ denotes the statement that $\mathscr{A}\colon \mathbb{C}^M/\mathbb{T} \to \mathbb{R}^N$ defined by $(\mathscr{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$ is injective, where $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$.

What follows is a characterization of injectivity in the complex case:

Theorem 3 (Theorem 4 in [6]). *Consider* $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$, *and viewing* $\{\varphi_n\varphi_n^* u\}_{n=1}^N$ *as vectors in* $\mathbb{R}^{2M}$, *denote* $S(u) := \mathrm{span}_{\mathbb{R}}\{\varphi_n\varphi_n^* u\}_{n=1}^N$. *Then the following are equivalent:*

(a) $\mathrm{Inj}[\Phi;\mathbb{C}^{M\times N}]$.
(b) $\dim S(u) \geq 2M-1$ *for every* $u \in \mathbb{C}^M \setminus \{0\}$.
(c) $S(u) = \mathrm{span}_{\mathbb{R}}\{\mathrm{i}u\}^{\perp}$ *for every* $u \in \mathbb{C}^M \setminus \{0\}$.

Before proving this theorem, note that unlike the characterization in the real case, it is not clear whether this characterization can be tested in finite time; instead of being a statement about all (finitely many) partitions of $\{1,\dots,N\}$, this is a statement about all $u \in \mathbb{C}^M \setminus \{0\}$. However, we can view this characterization as an analog to the real case in some sense: In the real case, the complement property is equivalent to having $\operatorname{span}\{\varphi_n\varphi_n^* u\}_{n=1}^N = \mathbb{R}^M$ for all $u \in \mathbb{R}^M \setminus \{0\}$. As the following proof makes precise, the fact that $\{\varphi_n\varphi_n^* u\}_{n=1}^N$ fails to span all of $\mathbb{R}^{2M}$ is rooted in the fact that more information is lost with phase in the complex case. There is actually a nice differential geometric interpretation of this result, and we will discuss it after the proof:

Proof (Proof of Theorem 3). (a) $\Rightarrow$ (c): Suppose $\mathscr{A}$ is injective. We need to show that $\{\varphi_n\varphi_n^* u\}_{n=1}^N$ spans the set of vectors orthogonal to $\mathrm{i}u$. Here, orthogonality is with respect to the real inner product, which can be expressed as $\langle a,b\rangle_{\mathbb{R}} = \operatorname{Re}\langle a,b\rangle$. Note that

$$|\langle u \pm v, \varphi_n\rangle|^2 = |\langle u,\varphi_n\rangle|^2 \pm 2\operatorname{Re}\langle u,\varphi_n\rangle\langle\varphi_n,v\rangle + |\langle v,\varphi_n\rangle|^2,$$

and so subtraction gives

$$|\langle u+v,\varphi_n\rangle|^2 - |\langle u-v,\varphi_n\rangle|^2 = 4\operatorname{Re}\langle u,\varphi_n\rangle\langle\varphi_n,v\rangle = 4\langle\varphi_n\varphi_n^* u, v\rangle_{\mathbb{R}}. \tag{1}$$

In particular, if the right-hand side of (1) is zero, then injectivity implies that there exists some ω of unit modulus such that $u+v = \omega(u-v)$. Since $u \neq 0$, we know $\omega \neq -1$, and so rearranging gives

$$v = -\frac{1-\omega}{1+\omega}u = -\frac{(1-\omega)(1+\overline{\omega})}{|1+\omega|^2}u = -\frac{2\operatorname{Im}\omega}{|1+\omega|^2}\,\mathrm{i}u.$$

This means $S(u)^\perp \subseteq \operatorname{span}_{\mathbb{R}}\{\mathrm{i}u\}$. To prove $\operatorname{span}_{\mathbb{R}}\{\mathrm{i}u\} \subseteq S(u)^\perp$, take $v = \alpha\mathrm{i}u$ for some $\alpha \in \mathbb{R}$ and define $\omega := \frac{1+\alpha\mathrm{i}}{1-\alpha\mathrm{i}}$, which necessarily has unit modulus. Then

$$u+v = u+\alpha\mathrm{i}u = (1+\alpha\mathrm{i})u = \frac{1+\alpha\mathrm{i}}{1-\alpha\mathrm{i}}(u-\alpha\mathrm{i}u) = \omega(u-v).$$

Thus, the left-hand side of (1) is zero, meaning $v \in S(u)^\perp$.

(b) $\Leftrightarrow$ (c): First, (b) immediately follows from (c). For the other direction, note that $\mathrm{i}u$ is necessarily orthogonal to every $\varphi_n\varphi_n^* u$:

$$\langle\varphi_n\varphi_n^* u, \mathrm{i}u\rangle_{\mathbb{R}} = \operatorname{Re}\langle\varphi_n\varphi_n^* u, \mathrm{i}u\rangle = \operatorname{Re}\langle u,\varphi_n\rangle\langle\varphi_n,\mathrm{i}u\rangle = -\operatorname{Re}\mathrm{i}|\langle u,\varphi_n\rangle|^2 = 0.$$

Thus, $\operatorname{span}_{\mathbb{R}}\{\mathrm{i}u\} \subseteq S(u)^\perp$, and by (b), $\dim S(u)^\perp \leq 1$, both of which gives (c).

(c) $\Rightarrow$ (a): This portion of the proof is inspired by Mukherjee's analysis in [45]. Suppose $\mathscr{A}(x) = \mathscr{A}(y)$. If $x = y$, we are done. Otherwise, $x-y \neq 0$, and so we may apply (c) to $u = x-y$. First, note that

$$\langle\varphi_n\varphi_n^*(x-y), x+y\rangle_{\mathbb{R}} = \operatorname{Re}\langle\varphi_n\varphi_n^*(x-y), x+y\rangle = \operatorname{Re}(x+y)^*\varphi_n\varphi_n^*(x-y),$$

and so expanding gives

$$\begin{aligned}\langle \varphi_n \varphi_n^*(x-y), x+y\rangle_{\mathbb{R}} &= \mathrm{Re}\left(|\varphi_n^* x|^2 - x^* \varphi_n \varphi_n^* y + y^* \varphi_n \varphi_n^* x - |\varphi_n^* y|^2\right) \\ &= \mathrm{Re}\left(-x^* \varphi_n \varphi_n^* y + \overline{x^* \varphi_n \varphi_n^* y}\right) = 0.\end{aligned}$$

Since $x+y \in S(x-y)^\perp = \mathrm{span}_{\mathbb{R}}\{\mathrm{i}(x-y)\}$, there exists $\alpha \in \mathbb{R}$ such that $x+y = \alpha\mathrm{i}(x-y)$, and so rearranging gives $y = \frac{1-\alpha\mathrm{i}}{1+\alpha\mathrm{i}}x$, meaning $y \equiv x \bmod \mathbb{T}$.

To better understand the above result, we will first develop a deeper understanding of the set $\mathbb{C}^M/\mathbb{T}$. If we remove zero, this set happens to be something called a *smooth manifold*, which means we can cover the set with overlapping patches, each with smooth coordinates, and with smooth coordinate transformations between the overlapping portions. To see this, consider the patches defined by

$$U_m := \{Z \in (\mathbb{C}^M \setminus \{0\})/\mathbb{T} : Z_m \neq 0\}, \qquad m = 1, \ldots, M.$$

We define the following coordinates over the patch U_m:

$$(z_1, \ldots, z_M) = \frac{|Z_m|}{Z_m}(Z_1, \ldots, Z_M),$$

where $(Z_1, \ldots, Z_M) \in \mathbb{C}^M \setminus \{0\}$ denotes any representative of the corresponding point in $(\mathbb{C}^M \setminus \{0\})/\mathbb{T}$. As such, each patch has its own homeomorphism to a set of coordinates, which is an open subset of $\mathbb{R}^{2M-1}$ (we lost a degree of freedom since the mth complex coordinate has no imaginary part). If we denote the mth homeomorphism by $f_m : U_m \to \mathbb{R}^{2M-1}$, then it is not difficult to show that each f_m, and as the transition maps $\tau_{m,m'} : f_m(U_m \cap U_{m'}) \to f_{m'}(U_m \cap U_{m'})$ defined by $\tau_{m,m'} := f_{m'} \circ f_m^{-1}$ are all smooth.

So we now understand that $(\mathbb{C}^M \setminus \{0\})/\mathbb{T}$ is a smooth manifold with $2M-1$ real dimensions. If we consider the function $\mathscr{A}$ over $(\mathbb{C}^M \setminus \{0\})/\mathbb{T}$, we can take its derivative at a given point u in terms of some chosen local coordinates. This amounts to taking the Jacobian at $u \in U_m$, whose rows are $\{2\varphi_n \varphi_n^* u\}_{n=1}^N$ as vectors in $\mathbb{R}^{2M}$, but with the column corresponding to the mth imaginary component removed. With this in mind, and using ideas from the proof of Lemma 22 in [6], Theorem 3 can be reinterpreted as follows:

Theorem 4. *$\mathscr{A}$ is injective if and only if the derivative of $\mathscr{A}$ is injective at every point in $(\mathbb{C}^M \setminus \{0\})/\mathbb{T}$.*

This says quite a bit about the intensity measurement mapping $\mathscr{A}$. Indeed, one can imagine a smooth mapping of a circle to a figure-eight curve, in which the derivative is injective at every point of the circle, but the mapping is certainly not injective. One could identify injectivity of the derivative as a local form of injectivity, and so it is rather surprising to have this be equivalent to the traditional (global) form of injectivity. Unfortunately, it is not clear what one can glean from such a feature of $\mathscr{A}$. Indeed, the above theorems leave a lot to be desired; compared to

the complement property in the real case, it is still unclear what it takes for a complex ensemble to yield injective intensity measurements. While in pursuit of a more clear understanding, the following bizarre characterization was stumbled upon: A complex ensemble yields injective intensity measurements precisely when it yields injective *phase-only* measurements (in some sense). This is made more precise in the following theorem statement:

Theorem 5 (Theorem 5 in [6]). *Consider* $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$ *and the mapping* $\mathscr{A}\colon \mathbb{C}^M/\mathbb{T} \to \mathbb{R}^N$ *defined by* $(\mathscr{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$. *Then* $\mathscr{A}$ *is injective if and only if the following statement holds: If for every* $n = 1, \ldots, N$, *either* $\arg(\langle x, \varphi_n\rangle^2) = \arg(\langle y, \varphi_n\rangle^2)$ *or one of the sides is not well defined, then* $x = 0$, $y = 0$, *or* $y \equiv x \bmod \mathbb{R} \setminus \{0\}$.

Proof. By Theorem 3, $\mathscr{A}$ is injective if and only if

$$\forall x \in \mathbb{C}^M \setminus \{0\}, \qquad \operatorname{span}_{\mathbb{R}}\{\varphi_n \varphi_n^* x\}_{n=1}^N = \operatorname{span}_{\mathbb{R}}\{\mathrm{i}x\}^{\perp}. \tag{2}$$

Taking orthogonal complements of both sides, note that regardless of $x \in \mathbb{C}^M \setminus \{0\}$, we know $\operatorname{span}_{\mathbb{R}}\{\mathrm{i}x\}$ is necessarily a subset of $(\operatorname{span}_{\mathbb{R}}\{\varphi_n \varphi_n^* x\}_{n=1}^N)^{\perp}$, and so (2) is equivalent to

$$\begin{aligned} \forall x \in \mathbb{C}^M \setminus \{0\}, \qquad & \operatorname{Re}\langle \varphi_n \varphi_n^* x, \mathrm{i}y\rangle = 0 \quad \forall n = 1, \ldots, N \\ & \implies \quad y = 0 \text{ or } y \equiv x \bmod \mathbb{R} \setminus \{0\}. \end{aligned}$$

Thus, we need to determine when $\operatorname{Im}\langle x, \varphi_n\rangle\overline{\langle y, \varphi_n\rangle} = \operatorname{Re}\langle \varphi_n \varphi_n^* x, \mathrm{i}y\rangle = 0$. We claim that this is true if and only if $\arg(\langle x, \varphi_n\rangle^2) = \arg(\langle y, \varphi_n\rangle^2)$ or one of the sides is not well defined. To see this, we substitute $a := \langle x, \varphi_n\rangle$ and $b := \langle y, \varphi_n\rangle$. Then to complete the proof, it suffices to show that $\operatorname{Im} a\bar{b} = 0$ if and only if $\arg(a^2) = \arg(b^2)$, $a = 0$, or $b = 0$.

($\Leftarrow$) If either a or b is zero, the result is immediate. Otherwise, if $2\arg(a) = \arg(a^2) = \arg(b^2) = 2\arg(b)$, then 2π divides $2(\arg(a) - \arg(b))$, and so $\arg(a\bar{b}) = \arg(a) - \arg(b)$ is a multiple of π. This implies that $a\bar{b} \in \mathbb{R}$, and so $\operatorname{Im} a\bar{b} = 0$.

($\Rightarrow$) Suppose $\operatorname{Im} a\bar{b} = 0$. Taking the polar decompositions $a = re^{\mathrm{i}\theta}$ and $b = se^{\mathrm{i}\phi}$, we equivalently have that $rs \sin(\theta - \phi) = 0$. Certainly, this can occur whenever r or s is zero, i.e., $a = 0$ or $b = 0$. Otherwise, a difference formula then gives $\sin\theta\cos\phi = \cos\theta\sin\phi$. From this, we know that if θ is an integer multiple of $\pi/2$, then ϕ is as well, and vice versa, in which case $\arg(a^2) = 2\arg(a) = \pi = 2\arg(b) = \arg(b^2)$. Else, we can divide both sides by $\cos\theta\cos\phi$ to obtain $\tan\theta = \tan\phi$, from which it is evident that $\theta \equiv \phi \bmod \pi$, and so $\arg(a^2) = 2\arg(a) = 2\arg(b) = \arg(b^2)$.

The notion of injective phase-only measurements appears similar to the notion of parallel rigidity in certain location estimation problems (for example, see [46] and the references therein). It would be interesting to further investigate this relationship, but at the very least, it is rather striking that injectivity in one setting is equivalent to injectivity in the other. In [6], this equivalence is used to prove that the complement property is necessary for injectivity in the complex case. Contrary to what is claimed

in [4], the first part of the proof of Theorem 1 does not suffice: It demonstrates that $u+v \neq \pm(u-v)$, but fails to establish that $u+v \not\equiv u-v \bmod \mathbb{T}$; for instance, it could very well be the case that $u+v = \mathrm{i}(u-v)$, and so injectivity would not be violated in the complex case. Overall, the complement property is necessary but not sufficient for injectivity. To see that it is not sufficient, consider measurement vectors $(1,0)$, $(0,1)$ and $(1,1)$. These certainly satisfy the complement property, but $\mathscr{A}((1,\mathrm{i})) = (1,1,2) = \mathscr{A}((1,-\mathrm{i}))$, despite the fact that $(1,\mathrm{i}) \not\equiv (1,-\mathrm{i}) \bmod \mathbb{T}$; in general, real measurement vectors fail to yield injective intensity measurements in the complex setting since they do not distinguish complex conjugates.

The theorem that follows provides one last characterization of injectivity in the complex case, and it will play a key role in our understanding of the phase transition. Before stating the result, define the real M^2-dimensional space $\mathbb{H}^{M\times M}$ of self-adjoint $M\times M$ matrices; note that this is not a vector space over complex scalars since the diagonal of a self-adjoint matrix must be real. Given an ensemble of measurement vectors $\{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$, define the *super analysis operator* $\mathbf{A}\colon \mathbb{H}^{M\times M} \to \mathbb{R}^N$ by $(\mathbf{A}H)(n) = \langle H, \varphi_n\varphi_n^*\rangle_{\mathrm{HS}}$; here, $\langle\cdot,\cdot\rangle_{\mathrm{HS}}$ denotes the Hilbert-Schmidt inner product, which induces the Frobenius matrix norm. Note that $\mathbf{A}$ is a linear operator, and yet

$$\begin{aligned}(\mathbf{A}xx^*)(n) &= \langle xx^*, \varphi_n\varphi_n^*\rangle_{\mathrm{HS}} = \mathrm{Tr}[\varphi_n\varphi_n^*xx^*]\\ &= \mathrm{Tr}[\varphi_n^*xx^*\varphi_n] = \varphi_n^*xx^*\varphi_n = |\langle x,\varphi_n\rangle|^2 = (\mathscr{A}(x))(n).\end{aligned}$$

In words, the class of vectors identified with x modulo $\mathbb{T}$ can be "lifted" to xx^*, thereby linearizing the intensity measurement process at the price of squaring the dimension of the vector space of interest; this identification has been exploited by some of the most noteworthy strides in modern phase retrieval [5, 16]. As the following lemma shows, this identification can also be used to characterize injectivity:

Theorem 6 (Lemma 9 in [6], cf. Corollary 1 in [35]). *$\mathscr{A}$ is not injective if and only if there exists a matrix of rank* 1 *or* 2 *in the null space of* $\mathbf{A}$.

Proof. ($\Rightarrow$) If $\mathscr{A}$ is not injective, then there exist $x,y \in \mathbb{C}^M/\mathbb{T}$ with $x \not\equiv y \bmod \mathbb{T}$ such that $\mathscr{A}(x) = \mathscr{A}(y)$. That is, $\mathbf{A}xx^* = \mathbf{A}yy^*$, and so $xx^* - yy^*$ is in the null space of $\mathbf{A}$.

($\Leftarrow$) First, suppose there is a rank-1 matrix H in the null space of $\mathbf{A}$. Then there exists $x \in \mathbb{C}^M$ such that $H = xx^*$ and $(\mathscr{A}(x))(n) = (\mathbf{A}xx^*)(n) = 0 = (\mathscr{A}(0))(n)$. But $x \not\equiv 0 \bmod \mathbb{T}$, and so $\mathscr{A}$ is not injective. Now suppose there is a rank-2 matrix H in the null space of $\mathbf{A}$. Then by the spectral theorem, there are orthonormal $u_1,u_2 \in \mathbb{C}^M$ and nonzero $\lambda_1 \geq \lambda_2$ such that $H = \lambda_1u_1u_1^* + \lambda_2u_2u_2^*$. Since H is in the null space of $\mathbf{A}$, the following holds for every n:

$$\begin{aligned}0 &= \langle H, \varphi_n\varphi_n^*\rangle_{\mathrm{HS}}\\ &= \langle \lambda_1u_1u_1^* + \lambda_2u_2u_2^*, \varphi_n\varphi_n^*\rangle_{\mathrm{HS}} = \lambda_1|\langle u_1,\varphi_n\rangle|^2 + \lambda_2|\langle u_2,\varphi_n\rangle|^2.\end{aligned} \tag{3}$$

Taking $x := |\lambda_1|^{1/2}u_1$ and $y := |\lambda_2|^{1/2}u_2$, note that $y \not\equiv x \bmod \mathbb{T}$ since they are nonzero and orthogonal. We claim that $\mathscr{A}(x) = \mathscr{A}(y)$, which would complete the proof. If λ_1 and λ_2 have the same sign, then by (3), $|\langle x,\varphi_n\rangle|^2 + |\langle y,\varphi_n\rangle|^2 = 0$

for every n, meaning $|\langle x,\varphi_n\rangle|^2 = 0 = |\langle y,\varphi_n\rangle|^2$. Otherwise, $\lambda_1 > 0 > \lambda_2$, and so $xx^* - yy^* = \lambda_1 u_1 u_1^* + \lambda_2 u_2 u_2^* = A$ is in the null space of $\mathbf{A}$, meaning $\mathscr{A}(x) = \mathbf{A}xx^* = \mathbf{A}yy^* = \mathscr{A}(y)$.

Of the three characterizations of injectivity in the complex case that we provided, this is by far the easiest to grasp, and perhaps due to its simplicity, much of our current understanding of the phase transition is based on this one. Still, comparing to our understanding in the real case helps to identify areas for improvement. For example, it is unclear how to test whether matrices of rank 1 or 2 lie in the null space of an arbitrary super analysis operator. Indeed, we have yet to find a "good" sufficient condition for injectivity in the complex case like the complement property or full spark provide in the real case.

The Phase Transition for Injectivity in the Complex Case

At this point, we wish to study the phase transition (presuming it exists) for $\mathrm{Inj}[\boldsymbol{\Phi};\mathbb{C}^{M\times N}]$. To this end, we introduce the following subproblem (i.e., part (a) of the phase transition):

Problem 1. For any dimension M, what is the smallest number $N^*(M)$ of injective intensity measurements?

Interestingly, this problem has some history in the quantum mechanics literature. For example, [50] presents *Wright's conjecture* that three observables suffice to uniquely determine any pure state. In phase retrieval parlance, the conjecture states that there exist unitary matrices U_1, U_2, and U_3 such that $\boldsymbol{\Phi} = [U_1\ U_2\ U_3]$ yields injective intensity measurements. Note that Wright's conjecture actually implies that $N^*(M) \le 3M-2$; indeed, U_1 determines the norm (squared) of the signal, rendering the last column of both U_2 and U_3 unnecessary. Finkelstein [30] later proved that $N^*(M) \ge 3M-2$; combined with Wright's conjecture, this led many to believe that $N^*(M) = 3M-2$ (for example, see [15]). However, both this and Wright's conjecture were recently disproved in [35], in which Heinosaari, Mazzarella, and Wolf invoked embedding theorems from differential geometry to prove that

$$N^*(M) \ge \begin{cases} 4M - 2\alpha(M-1) - 3 \text{ for all } M, \\ 4M - 2\alpha(M-1) - 2 \text{ if } M \text{ is odd, } \alpha(M-1) \equiv 2 \bmod 4, \\ 4M - 2\alpha(M-1) - 1 \text{ if } M \text{ is odd, } \alpha(M-1) \equiv 3 \bmod 4, \end{cases} \tag{4}$$

where $\alpha(M-1) \le \log_2(M)$ is the number of 1's in the binary representation of $M-1$; apparently, this result had previously appeared in [43, 44] as well. By comparison, Balan, Casazza, and Edidin [4] proved that $N^*(M) \le 4M-2$, and so we at least have the asymptotic expression $N^*(M) = (4+o(1))M$.

At this point, we should clarify some intuition for $N^*(M)$ by explaining the nature of these best known lower and upper bounds. First, the lower bound (4) follows from an older result that complex projective space $\mathbb{C}\mathbf{P}^n$ does not smoothly embed into $\mathbb{R}^{4n-2\alpha(n)}$ (and other slight refinements which depend on n); this is

due to Mayer [38], but we highly recommend James's survey on the topic [36]. To prove (4) from this, suppose $\mathscr{A}\colon \mathbb{C}^M/\mathbb{T} \to \mathbb{R}^N$ were injective. Then $\mathscr{E}$ defined by $\mathscr{E}(x) := \mathscr{A}(x)/\|x\|^2$ embeds $\mathbb{C}\mathbf{P}^{M-1}$ into $\mathbb{R}^N$, and as Heinosaari et al. show, the embedding is necessarily smooth; considering $\mathscr{A}(x)$ is made up of rather simple polynomials, the fact that $\mathscr{E}$ is smooth should not come as a surprise. As such, the nonembedding result produces the best known lower bound. To evaluate this bound, first note that Milgram [40] constructs an embedding of $\mathbb{C}\mathbf{P}^n$ into $\mathbb{R}^{4n-\alpha(n)+1}$, establishing the importance of the $\alpha(n)$ term, but the constructed embedding does not correspond to an intensity measurement process. In order to relate these embedding results to our problem, consider the real case: It is known that for odd $n \geq 7$, real projective space $\mathbb{R}\mathbf{P}^n$ smoothly embeds into $\mathbb{R}^{2n-\alpha(n)+1}$ [48], which means the analogous lower bound for the real case would necessarily be smaller than $2(M-1)-\alpha(M-1)+1 = 2M-\alpha(M-1)-1 < 2M-1$. This indicates that the $\alpha(M-1)$ term in (4) might be an artifact of the proof technique, rather than of $N^*(M)$.

We now consider our previous analysis of injectivity to help guide intuition about the possible phase transition. Theorem 6 indicates that we want the null space of $\mathbf{A}$ to avoid nonzero matrices of rank ≤ 2. Intuitively, this is easier when the "dimension" of this set of matrices is small. To get some idea of this dimension, let's count real degrees of freedom. By the spectral theorem, almost every matrix in $\mathbb{H}^{M\times M}$ of rank ≤ 2 can be uniquely expressed as $\lambda_1 u_1 u_1^* + \lambda_2 u_2 u_2^*$ with $\lambda_1 \leq \lambda_2$. Here, (λ_1,λ_2) has two degrees of freedom. Next, u_1 can be any vector in $\mathbb{C}^M$, except its norm must be 1. Also, since u_1 is only unique up to global phase, we take its first entry to be nonnegative without loss of generality. Given the norm and phase constraints, u_1 has a total of $2M-2$ real degrees of freedom. Finally, u_2 has the same norm and phase constraints, but it must also be orthogonal to u_1, that is, $\mathrm{Re}\langle u_2,u_1\rangle = \mathrm{Im}\langle u_2,u_1\rangle = 0$. As such, u_2 has $2M-4$ real degrees of freedom. All together, we can expect the set of matrices in question to have $2+(2M-2)+(2M-4) = 4M-4$ real dimensions.

If the set S of matrices of rank ≤ 2 formed a subspace of $\mathbb{H}^{M\times M}$ (it doesn't), then we could expect the null space of $\mathbf{A}$ to intersect that subspace nontrivially whenever $\dim\mathrm{null}(\mathbf{A}) + (4M-4) > \dim(\mathbb{H}^{M\times M}) = M^2$. By the rank-nullity theorem, this would indicate that injectivity requires

$$N \geq \mathrm{rank}(\mathbf{A}) = M^2 - \dim\mathrm{null}(\mathbf{A}) \geq 4M-4. \tag{5}$$

Of course, this logic is not technically valid since S is not a subspace. It is, however, a special kind of set: a real projective variety. To see this, let's first show that it is a real algebraic variety, specifically, the set of members of $\mathbb{H}^{M\times M}$ for which all 3×3 minors are zero. Of course, every member of S has this minor property. Next, we show that members of S are the only matrices with this property: If the rank of a given matrix is ≥ 3, then it has an $M\times 3$ submatrix of linearly independent columns, and since the rank of its transpose is also ≥ 3, this $M\times 3$ submatrix must have 3 linearly independent rows, thereby implicating a full-rank 3×3 submatrix. This variety is said to be projective because it is closed under scalar multiplication. If S were a projective variety over an algebraically closed field (it's not), then the projective dimension

theorem (Theorem 7.2 of [34]) says that S intersects $\operatorname{null}(\mathbf{A})$ nontrivially whenever the dimensions are large enough: $\dim \operatorname{null}(\mathbf{A}) + \dim S > \dim \mathbb{H}^{M\times M}$, thereby implying that injectivity requires (5). Unfortunately, this theorem is not valid when the field is $\mathbb{R}$; for example, the cone defined by $x^2+y^2-z^2=0$ in $\mathbb{R}^3$ is a projective variety of dimension 2, but its intersection with the 2-dimensional xy-plane is trivial, despite the fact that $2+2>3$.

In the absence of a proof, we pose the natural conjecture:

Conjecture 1 (The $4M-4$ conjecture). $\operatorname{Inj}[\Phi;\mathbb{C}^{M\times N}]$ exhibits a phase transition at $N=4M-4$.

As incremental progress toward solving the $4M-4$ conjecture, we offer the following result:

Theorem 7 (Theorem 10 in [6]). *The $4M-4$ conjecture is true when $M=2$.*

In this case, injectivity is equivalent to $\mathbf{A}$ having a trivial null space by Theorem 6, meaning $\mathbf{A}$ must have rank $M^2=4=4M-4$ for injectivity, implying part (a). For $N=4M-4$, $\mathbf{A}$ has a square matrix representation, and so injectivity is equivalent to having $\det\mathbf{A}\neq 0$. As such, part (b) is proved by considering the real algebraic variety $V:=\{\mathbf{A}:\operatorname{Re}\det\mathbf{A}=\operatorname{Im}\det\mathbf{A}=0\}$ and showing that V^{c} is nonempty.

We can also prove the $M=3$ case, but we first introduce Algorithm 1, namely the *HMW test* for injectivity; this is named after Heinosaari, Mazarella, and Wolf, who implicitly introduce this algorithm in their paper [35].

Algorithm 1 The HMW test for injectivity when $M=3$

Input: Measurement vectors $\{\varphi_n\}_{n=1}^N\subseteq\mathbb{C}^3$
Output: Whether $\mathscr{A}$ is injective

```
Define A: H^{3×3} → R^N such that AH = {⟨H, φ_n φ_n^*⟩_HS}_{n=1}^N
if dim null(A) = 0 then
    "INJECTIVE"                                   {if A is injective, then 𝒜 is injective}
else
    Pick H ∈ null(A), H ≠ 0
    if dim null(A) = 1 and det(H) ≠ 0 then
        "INJECTIVE"                               {if A only maps nonsingular matrices to zero, then 𝒜 is injective}
    else
        "NOT INJECTIVE"                           {in the remaining case, A maps differences of rank-1 matrices to zero}
    end if
end if
```

Theorem 8 (Theorem 11 in [6], cf. Proposition 6 in [35]). *When $M=3$, the HMW test correctly determines whether $\mathscr{A}$ is injective.*

Proof. First, if $\mathbf{A}$ is injective, then $\mathscr{A}(x)=\mathbf{A}xx^*=\mathbf{A}yy^*=\mathscr{A}(y)$ if and only if $xx^*=yy^*$, i.e., $y\equiv x \bmod \mathbb{T}$. Next, suppose $\mathbf{A}$ has a 1-dimensional null space. Then Lemma 6 gives that $\mathscr{A}$ is injective if and only if the null space of $\mathbf{A}$ is spanned by a matrix of full rank. Finally, if the dimension of the null space is 2 or more,

then there exist linearly independent (nonzero) matrices A and B in this null space. If $\det(A) = 0$, then it must have rank 1 or 2, and so Lemma 6 gives that $\mathscr{A}$ is not injective. Otherwise, consider the map

$$f\colon t \mapsto \det(A\cos t + B\sin t) \qquad \forall t \in [0,\pi].$$

Since $f(0) = \det(A)$ and $f(\pi) = \det(-A) = (-1)^3\det(A) = -\det(A)$, the intermediate value theorem gives that there exists $t_0 \in [0,\pi]$ such that $f(t_0) = 0$, i.e., the matrix $A\cos t_0 + B\sin t_0$ is singular. Moreover, this matrix is nonzero since A and B are linearly independent, and so its rank is either 1 or 2. Lemma 6 then gives that $\mathscr{A}$ is not injective.

As an example, we may run the HMW test on the columns of the following matrix:

$$\Phi = \begin{bmatrix} 2 & 1 & 1 & 0 & 0 & 0 & 1 & \mathrm{i} \\ -1 & 0 & 0 & 1 & 1 & -1 & -2 & 2 \\ 0 & 1 & -1 & 1 & -1 & 2\mathrm{i} & \mathrm{i} & -1 \end{bmatrix}. \tag{6}$$

In this case, the null space of $\mathbf{A}$ is 1-dimensional and spanned by a nonsingular matrix. As such, $\mathscr{A}$ is injective. We are now ready to approach the $4M-4$ conjecture in the $M = 3$ case:

Theorem 9 (Theorem 12 in [6]). *The $4M-4$ conjecture is true when $M = 3$.*

This is proved using the HMW test. In this case, $4M - 4 = 8 = M^2 - 1$, meaning when $N = 4M-4$, the null space of $\mathbf{A}$ should typically be 1-dimensional. As such, the null space can be described algebraically in terms of a generalized cross product, and this is leveraged along with the HMW test to construct a real algebraic variety containing all $\mathbf{A}$'s which are not injective; the construction in (6) is then used to prove that the complement of this variety is nonempty, thereby proving part (b).

Recently, the American Institute of Mathematics hosted a workshop called "Frame Theory Intersects Geometry", where experts from the two communities discussed various problems, including the $4M-4$ conjecture. One outcome of this workshop was a paper by Conca, Edidin, Hering, and Vinzant [20], which makes a major stride toward solving the $4M-4$ conjecture:

Theorem 10 (Theorem 1.1 and Proposition 5.4 in [20]).

(a) Part (a) of the $4M-4$ conjecture is true whenever $M = 2^k + 1$.
(b) Part (b) of the $4M-4$ conjecture is true.

Proof (sketch). The proof of (a) uses certain integrality conditions, similar to the proofs of embedding results for complex projective space. Part (b) is proved using the following basic ideas: Consider the set of all $M \times N$ complex matrices Φ. This set has real dimension $2MN$. The goal is to show that the set of "bad" Φ's (those which fail to yield injectivity) has strictly smaller dimension. To do this, note from Theorem 6 that $\Phi = \{\varphi_n\}_{n=1}^N$ is bad precisely when there is an $M \times M$ matrix Q of rank ≤ 2 and Frobenius norm 1 such that $\varphi_n^* Q \varphi_n = 0$ for every $n = 1,\ldots,N$. As such,

we lift to the set of pairs (Φ, Q) which satisfy this relation. Counting the dimension of this lifted set, we note that the set of Q's has $4M-5$ real dimensions, and for each Q and each n, there is a $(2M-1)$-dimensional set of φ_n's such that $\varphi_n^* Q \varphi_n = 0$. Thus, the total dimension of bad pairs (Φ, Q) is $4M-5+(2M-1)N$. Recall that the bad set we care about is the set of Φ's for which there exists a Q such that (Φ, Q) is bad, and so we get our set by projection. Also, projections never increase the dimension of a set, and so the dimension of our set of bad Φ's is $\leq 4M-5+(2M-1)N$. As such, to ensure that this dimension is less than the ambient dimension $2MN$, it suffices to have $4M-5+(2M-1)N < 2MN$, or equivalently, $N \geq 4M-4$. Thus, generic $M \times N$ Φ's with $N \geq 4M-4$ are not bad.

Note that this result contains the previous cases where $M = 2, 3$, and the first remaining open case is $M = 4$. On my blog, I offer a US$100 prize for a proof of the $4M-4$ conjecture, and a can of Coca-Cola for a disproof [42].

Minimal Constructions with Injectivity in the Complex Case

In the absence of a "good" characterization of injectivity in the complex case, it is interesting to see explicit minimal constructions, as these shed some insight into the underlying structure of such ensembles. To this end, there are presently two general constructions, which we describe here.

The construction of Bodmann and Hammen [10] considers the case where the measurement vectors form a certain harmonic frame. Specifically, take the $(2M-1) \times (2M-1)$ discrete Fourier transform matrix and collect the first M rows; then, the resulting columns are the M-dimensional measurement vectors. Note that if the original signal is known to be real, then $2M-1$ measurements are injective whenever the measurement vectors are full spark, as this particular harmonic frame is (since all $M \times M$ submatrices are Vandermonde with distinct bases). Analogously, Bernhard and Hammen exploit the Fejer–Riesz spectral factorization theorem to say that these measurements completely determine signals from another interesting class. To be clear, identify a vector $(c_m)_{m=0}^{M-1}$ with the polynomial $\sum_{m=0}^{M-1} c_m z^m$; then, Bernhard and Hammen uniquely determine the vector if the roots of the corresponding polynomial all lie outside the open unit disk in the complex plane. In general, they actually say the polynomial is one of 2^M possibilities; here, the only ambiguity is whether a given root is at z_m or $1/\overline{z_m}$, that is, we can flip any root from outside to inside the disk. Note that this is precisely how much ambiguity we have in the real case after taking only M measurements, and in that case, we know it suffices to take only $M-1$ additional measurements. Next, in addition to taking these $2M-1$ measurements from before (viewed as equally spaced points on the complex unit circle), they also take $2M-1$ measurements corresponding to equally spaced points on another unit circle in the complex plane, this one being the image of the real line under a specially chosen (think "sufficiently irrational") Cayley map. However, this makes a total of $4M-2$ measurements, whereas the goal is to find $4M-4$ injective measurements – to fix this, they actually pick the second circle in such a way that

it intersects the first at two points, and that these intersection points correspond to measurements from both circles, so we might as well throw two of them away.

The second known construction is due to Fickus, Nelson, Wang, and the author [27]. Here, we apply two main ideas: (1) a signal's intensity measurements with the Fourier transform is the Fourier transform of the signal's autocorrelation; and (2) if a real, even function is sufficiently zero-padded, then it can be recovered (up to a global sign factor) from its autocorrelation. (Verifying (2) in small dimensions is a fun exercise.) In [27], we show how to generalize (2) to completely determine zero-padded complex functions, and then we identify these autocorrelations as the inverse Fourier transforms of intensity measurements with $4M-2$ truncated and modulated discrete cosine functions. At this point, we identify certain redundancies in our intensity measurements – two of them are completely determined by the others, so we can remove them.

Considering these minimal constructions, it is striking that they both come from a construction of size $4M-2$. This begs the following question: Does every injective ensemble of size $4M-2$ contain an injective ensemble of size $4M-4$?

Almost Injectivity

In both the real and complex cases, there appears to be a phase transition which governs how many intensity measurements are necessary and generically sufficient for injectivity. Interestingly, one can save a factor of 2 in this number of measurements by slightly weakening the desired notion of injectivity [4, 31]. To be explicit, we start with the following definition:

Definition 4. The intensity measurement mapping $\mathscr{A}$ is said to be *almost injective* if $\mathscr{A}^{-1}(\mathscr{A}(x)) = \{\omega x : |\omega| = 1\}$ for every x in an open, dense subset of $\mathbb{F}^M$. $\mathrm{AlmInj}[\Phi, \mathbb{F}^{M\times N}]$ denotes the statement that the intensity measurement mapping $\mathscr{A}$ associated with Φ is almost injective.

This section studies the phase transition for almost injectivity. Much like injectivity, we have a much better understanding of the real case than the complex case, and we consider these separately.

Almost Injectivity in the Real Case

In this section, we start by characterizing ensembles of measurement vectors which yield almost injective intensity measurements, and similar to the characterization of injectivity, the basic idea behind the analysis is to consider sums and differences of signals with identical intensity measurements. Our characterization starts with the following lemma:

Theorem 11 (Lemma 9 in [27]). *Consider* $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{R}^M$ *and the intensity measurement mapping* $\mathscr{A}\colon \mathbb{R}^M/\{\pm 1\} \to \mathbb{R}^N$ *defined by* $(\mathscr{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$. *Then* $\mathscr{A}$ *is almost injective if and only if almost every* $x \in \mathbb{R}^M$ *is not in the Minkowski sum* $\operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$ *for all* $S \subseteq \{1,\dots,N\}$. *More precisely,* $\mathscr{A}^{-1}(\mathscr{A}(x)) = \{\pm x\}$ *if and only if* $x \notin \operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$ *for any* $S \subseteq \{1,\dots,N\}$.

Proof. By the definition of the mapping $\mathscr{A}$, for $x, y \in \mathbb{R}^M$ we have $\mathscr{A}(x) = \mathscr{A}(y)$ if and only if $|\langle x, \varphi_n\rangle| = |\langle y, \varphi_n\rangle|$ for all $n \in \{1,\dots,N\}$. This occurs precisely when there is a subset $S \subseteq \{1,\dots,N\}$ such that $\langle x, \varphi_n\rangle = -\langle y, \varphi_n\rangle$ for every $n \in S$ and $\langle x, \varphi_n\rangle = \langle y, \varphi_n\rangle$ for every $n \in S^c$. Thus, $\mathscr{A}^{-1}(\mathscr{A}(x)) = \{\pm x\}$ if and only if for every $y \neq \pm x$ and for every $S \subseteq \{1,\dots,N\}$, either there exists an $n \in S$ such that $\langle x+y, \varphi_n\rangle \neq 0$ or an $n \in S^c$ such that $\langle x-y, \varphi_n\rangle \neq 0$. We claim that this occurs if and only if x is not in the Minkowski sum $\operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$ for all $S \subseteq \{1,\dots,N\}$, which would complete the proof. We verify the claim by seeking the contrapositive in each direction.

($\Rightarrow$) Suppose $x \in \operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$. Then there exists $u \in \operatorname{span}(\Phi_S)^\perp \setminus \{0\}$ and $v \in \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$ such that $x = u + v$. Taking $y := u - v$, we see that $x + y = 2u \in \operatorname{span}(\Phi_S)^\perp \setminus \{0\}$ and $x - y = 2v \in \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$, which means that there is no $n \in S$ such that $\langle x+y, \varphi_n\rangle \neq 0$ nor $n \in S^c$ such that $\langle x-y, \varphi_n\rangle \neq 0$. Furthermore, u and v are nonzero, and so $y \neq \pm x$.

($\Leftarrow$) Suppose $y \neq \pm x$ and for every $S \subseteq \{1,\dots,N\}$ there is no $n \in S$ such that $\langle x+y, \varphi_n\rangle \neq 0$ nor $n \in S^c$ such that $\langle x-y, \varphi_n\rangle \neq 0$. Then $x + y \in \operatorname{span}(\Phi_S)^\perp \setminus \{0\}$ and $x - y \in \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$. Since $x = \frac{1}{2}(x+y) + \frac{1}{2}(x-y)$, we have that $x \in \operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$.

The above characterization can be simplified to form the following partial characterization of almost injectivity:

Theorem 12 (Theorem 10 in [27]). *Consider* $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{R}^M$ *and the intensity measurement mapping* $\mathscr{A}\colon \mathbb{R}^M/\{\pm 1\} \to \mathbb{R}^N$ *defined by* $(\mathscr{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$. *Suppose* Φ *spans* $\mathbb{R}^M$ *and each* φ_n *is nonzero. Then* $\mathscr{A}$ *is almost injective if and only if the Minkowski sum* $\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp$ *is a proper subspace of* $\mathbb{R}^M$ *for each nonempty proper subset* $S \subseteq \{1,\dots,N\}$.

Note that the above result is not terribly surprising considering Theorem 11, as the new condition involves a simpler Minkowski sum in exchange for additional (reasonable and testable) assumptions on Φ. The proof of this theorem amounts to measuring the difference between the two Minkowski sums:

Proof (Proof of Theorem 12). First note that the spanning assumption on Φ implies

$$\operatorname{span}(\Phi_S)^\perp \cap \operatorname{span}(\Phi_{S^c})^\perp = \big(\operatorname{span}(\Phi_S) + \operatorname{span}(\Phi_{S^c})\big)^\perp = \operatorname{span}(\Phi)^\perp = \{0\},$$

and so one can prove the following identity:

$$\operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$$

$$= \left(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\right) \setminus \left(\operatorname{span}(\Phi_S)^\perp \cup \operatorname{span}(\Phi_{S^c})^\perp\right). \tag{7}$$

From Theorem 11 we know that $\mathscr{A}$ is almost injective if and only if almost every $x \in \mathbb{R}^M$ is not in the Minkowski sum $\operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$ for any $S \subseteq \{1,\dots,N\}$. In other words, the Lebesgue measure (which we denote by $\mathrm{Leb}[\cdot]$) of this Minkowski sum is zero for each $S \subseteq \{1,\dots,N\}$. By (7), this equivalently means that the Lebesgue measure of $\left(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\right) \setminus \left(\operatorname{span}(\Phi_S)^\perp \cup \operatorname{span}(\Phi_{S^c})^\perp\right)$ is zero for each $S \subseteq \{1,\dots,N\}$. Since Φ spans $\mathbb{R}^M$, this set is empty (and therefore has Lebesgue measure zero) when $S = \emptyset$ or $S = \{1,\dots,N\}$. Also, since each φ_n is nonzero, we know that $\operatorname{span}(\Phi_S)^\perp$ and $\operatorname{span}(\Phi_{S^c})^\perp$ are proper subspaces of $\mathbb{R}^M$ whenever S is a nonempty proper subset of $\{1,\dots,N\}$, and so in these cases both subspaces must have Lebesgue measure zero. As such, we have that for every nonempty proper subset $S \subseteq \{1,\dots,N\}$,

$$\begin{aligned}
&\mathrm{Leb}\left[\left(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\right) \setminus \left(\operatorname{span}(\Phi_S)^\perp \cup \operatorname{span}(\Phi_{S^c})^\perp\right)\right] \\
&\quad \geq \mathrm{Leb}\left[\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\right] - \mathrm{Leb}\left[\operatorname{span}(\Phi_S)^\perp\right] - \mathrm{Leb}\left[\operatorname{span}(\Phi_{S^c})^\perp\right] \\
&\quad = \mathrm{Leb}\left[\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\right] \\
&\quad \geq \mathrm{Leb}\left[\left(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\right) \setminus \left(\operatorname{span}(\Phi_S)^\perp \cup \operatorname{span}(\Phi_{S^c})^\perp\right)\right].
\end{aligned}$$

In summary, $\left(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\right) \setminus \left(\operatorname{span}(\Phi_S)^\perp \cup \operatorname{span}(\Phi_{S^c})^\perp\right)$ having Lebesgue measure zero for each $S \subseteq \{1,\dots,N\}$ is equivalent to $\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp$ having Lebesgue measure zero for each nonempty proper subset $S \subseteq \{1,\dots,N\}$, which in turn is equivalent to the Minkowski sum $\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp$ being a proper subspace of $\mathbb{R}^M$ for each nonempty proper subset $S \subseteq \{1,\dots,N\}$, as desired.

At this point, consider the following stronger restatement of Theorem 12: "Suppose each φ_n is nonzero. Then $\mathscr{A}$ is almost injective if and only if Φ spans $\mathbb{R}^M$ and the Minkowski sum $\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp$ is a proper subspace of $\mathbb{R}^M$ for each nonempty proper subset $S \subseteq \{1,\dots,N\}$". Note that we can move the spanning assumption into the condition because if Φ does not span, then we can decompose almost every $x \in \mathbb{R}^M$ as $x = u + v$ such that $u \in \operatorname{span}(\Phi)$ and $v \in \operatorname{span}(\Phi)^\perp$ with $v \neq 0$, and defining $y := u - v$ then gives $\mathscr{A}(y) = \mathscr{A}(x)$ despite the fact that $y \neq \pm x$. As for the assumption that the φ_n's are nonzero, we note that having $\varphi_n = 0$ amounts to having the nth entry of $\mathscr{A}(x)$ be zero for all x. As such, Φ yields almost injectivity precisely when the nonzero members of Φ together yield almost injectivity. With this identification, the stronger restatement of Theorem 12 above can be viewed as a complete characterization of almost injectivity. Next, we will replace the Minkowski sum condition with a rather elegant condition involving the ranks of Φ_S and Φ_{S^c}:

Theorem 13 (Theorem 11 in [27]). *Consider $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{R}^M$ and the intensity measurement mapping $\mathscr{A}: \mathbb{R}^M/\{\pm 1\} \to \mathbb{R}^N$ defined by $(\mathscr{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$.*

Suppose each φ_n is nonzero. Then $\mathscr{A}$ is almost injective if and only if Φ spans $\mathbb{R}^M$ and $\operatorname{rank}\Phi_S + \operatorname{rank}\Phi_{S^c} > M$ *for each nonempty proper subset* $S \subseteq \{1,\dots,N\}$.

Proof. Considering the discussion after the proof of Theorem 12, it suffices to assume that Φ spans $\mathbb{R}^M$. Furthermore, considering Theorem 12, it suffices to characterize when $\dim\left(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\right) < M$. By the inclusion-exclusion principle for subspaces, we have

$$\begin{aligned}&\dim\left(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\right)\\ &= \dim\left(\operatorname{span}(\Phi_S)^\perp\right) + \dim\left(\operatorname{span}(\Phi_{S^c})^\perp\right) - \dim\left(\operatorname{span}(\Phi_S)^\perp \cap \operatorname{span}(\Phi_{S^c})^\perp\right).\end{aligned}$$

Since Φ is assumed to span $\mathbb{R}^M$, we also have that $\operatorname{span}(\Phi_S)^\perp \cap \operatorname{span}(\Phi_{S^c})^\perp = \{0\}$, and so

$$\begin{aligned}&\dim\left(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\right)\\ &\quad= \left(M - \dim\left(\operatorname{span}(\Phi_S)\right)\right) + \left(M - \dim\left(\operatorname{span}(\Phi_{S^c})\right)\right) - 0\\ &\quad= 2M - \operatorname{rank}\Phi_S - \operatorname{rank}\Phi_{S^c}.\end{aligned}$$

As such, $\dim\left(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\right) < M$ precisely when $\operatorname{rank}\Phi_S + \operatorname{rank}\Phi_{S^c} > M$.

At this point, we point out some interesting consequences of Theorem 13. First of all, Φ cannot be almost injective if $N < M+1$ since $\operatorname{rank}\Phi_S + \operatorname{rank}\Phi_{S^c} \le |S| + |S^c| = N$. Also, in the case where $N = M+1$, we note that Φ is almost injective precisely when Φ is full spark, that is, every size-M subcollection is a spanning set (note this implies that all of the φ_n's are nonzero). In fact, every full spark Φ with $N \ge M+1$ yields almost injective intensity measurements, which in turn implies that a generic Φ yields almost injectivity when $N \ge M+1$ [4]. This is in direct analogy with injectivity in the real case; here, injectivity requires $N \ge 2M-1$, injectivity with $N = 2M-1$ is equivalent to being full spark, and being full spark suffices for injectivity whenever $N \ge 2M-1$ [4]. Another thing to check is that the condition for injectivity implies the condition for almost injectivity (it does). Overall, we have the following phase transition result:

Theorem 14 ([4]). $\operatorname{AlmInj}[\Phi, \mathbb{R}^{M\times N}]$ *exhibits a phase transition at* $N = M+1$.

Having established that full spark ensembles of size $N \ge M+1$ yield almost injective intensity measurements, we note that checking whether a matrix is full spark is NP-hard in general [37]. Granted, there are a few explicit constructions of full spark ensembles which can be used [1, 47], but it would be nice to have a condition which is not computationally difficult to test in general. We provide one such condition in the following theorem, but first, we briefly review the requisite frame theory.

A *frame* is an ensemble $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{R}^M$ together with *frame bounds* $0 < A \le B < \infty$ with the property that for every $x \in \mathbb{R}^M$,

$$A\|x\|^2 \leq \sum_{n=1}^{N} |\langle x, \varphi_n \rangle|^2 \leq B\|x\|^2.$$

When $A = B$, the frame is said to be *tight*, and such frames come with a painless reconstruction formula:

$$x = \frac{1}{A} \sum_{n=1}^{N} \langle x, \varphi_n \rangle \varphi_n.$$

To be clear, the theory of frames originated in the context of infinite-dimensional Hilbert spaces [22, 24], and frames have since been studied in finite-dimensional settings, primarily because this is the setting in which they are applied computationally. Of particular interest are so-called *unit norm tight frames (UNTFs)*, which are tight frames whose frame elements have unit norm: $\|\varphi_n\| = 1$ for every $n = 1, \ldots, N$. Such frames are useful in applications; for example, if one encodes a signal x using frame coefficients $\langle x, \varphi_n \rangle$ and transmits these coefficients across a channel, then UNTFs are optimally robust to noise [32] and one erasure [17]. Intuitively, this optimality comes from the fact that frame elements of a UNTF are particularly well-distributed in the unit sphere [8]. Another pleasant feature of UNTFs is that it is straightforward to test whether a given frame is a UNTF: Letting $\Phi = [\varphi_1 \cdots \varphi_N]$ denote an $M \times N$ matrix whose columns are the frame elements, then Φ is a UNTF precisely when each of the following occurs simultaneously:

(a) the rows have equal norm
(b) the rows are orthogonal
(c) the columns have unit norm

(This is a direct consequence of the tight frame's reconstruction formula and the fact that a UNTF has unit-norm frame elements; furthermore, since the columns have unit norm, it is not difficult to see that the rows will necessarily have norm $\sqrt{N/M}$.) In addition to being able to test that an ensemble is a UNTF, various UNTFs can be constructed using *spectral tetris* [18] (though such frames necessarily have $N \geq 2M$), and *every* UNTF can be constructed using the recent theory of *eigensteps* [12, 28]. Now that UNTFs have been properly introduced, we relate them to almost injectivity for phase retrieval:

Theorem 15 (Theorem 12 in [27]). *If M and N are relatively prime, then every unit norm tight frame $\Phi = \{\varphi_n\}_{n=1}^{N} \subseteq \mathbb{R}^M$ yields almost injective intensity measurements.*

Proof. Pick a nonempty proper subset $S \subseteq \{1, \ldots, N\}$. By Theorem 13, it suffices to show that $\operatorname{rank} \Phi_S + \operatorname{rank} \Phi_{S^c} > M$, or equivalently, $\operatorname{rank} \Phi_S \Phi_S^* + \operatorname{rank} \Phi_{S^c} \Phi_{S^c}^* > M$. Note that since Φ is a unit norm tight frame, we also have

$$\Phi_S \Phi_S^* + \Phi_{S^c} \Phi_{S^c}^* = \Phi \Phi^* = \tfrac{N}{M} I,$$

and so $\Phi_S \Phi_S^*$ and $\Phi_{S^c} \Phi_{S^c}^*$ are simultaneously diagonalizable, i.e., there exists a unitary matrix U and diagonal matrices D_1 and D_2 such that

$$UD_1U^* + UD_2U^* = \Phi_S\Phi_S^* + \Phi_{S^c}\Phi_{S^c}^* = \tfrac{N}{M}I.$$

Conjugating by U^*, this then implies that $D_1 + D_2 = \frac{N}{M}I$. Let $L_1 \subseteq \{1,\dots,M\}$ denote the diagonal locations of the nonzero entries in D_1, and $L_2 \subseteq \{1,\dots,M\}$ similarly for D_2. To complete the proof, we need to show that $|L_1| + |L_2| > M$ (since $|L_1| + |L_2| = \operatorname{rank}\Phi_S\Phi_S^* + \operatorname{rank}\Phi_{S^c}\Phi_{S^c}^*$). Note that $L_1 \cup L_2 \neq \{1,\dots,M\}$ would imply that $D_1 + D_2$ has at least one zero in its diagonal, contradicting the fact that $D_1 + D_2$ is a nonzero multiple of the identity; as such, $L_1 \cup L_2 = \{1,\dots,M\}$ and $|L_1| + |L_2| \geq M$. We claim that this inequality is strict due to the assumption that M and N are relatively prime. To see this, it suffices to show that $L_1 \cap L_2$ is nonempty. Suppose to the contrary that L_1 and L_2 are disjoint. Then since $D_1 + D_2 = \frac{N}{M}I$, every nonzero entry in D_1 must be N/M. Since S is a nonempty proper subset of $\{1,\dots,N\}$, this means that there exists $K \in (0,M)$ such that D_1 has K entries which are N/M and $M - K$ which are 0. Thus,

$$|S| = \operatorname{Tr}[\Phi_S^*\Phi_S] = \operatorname{Tr}[\Phi_S\Phi_S^*] = \operatorname{Tr}[UD_1U^*] = \operatorname{Tr}[D_1] = K(N/M),$$

implying that $N/M = |S|/K$ with $K \neq M$ and $|S| \neq N$. Since this contradicts the assumption that N/M is in lowest form, we have the desired result.

In general, whether a UNTF Φ yields almost injective intensity measurements is determined by whether it is *orthogonally partitionable*: Φ is orthogonally partitionable if there exists a partition $S \sqcup S^c = \{1,\dots,N\}$ such that $\operatorname{span}(\Phi_S)$ is orthogonal to $\operatorname{span}(\Phi_{S^c})$. Specifically, a UNTF yields almost injective intensity measurements precisely when it is not orthogonally partitionable. Historically, this property of UNTFs has been pivotal to the understanding of singularities in the algebraic variety of UNTFs [25], and it has also played a key role in solutions to the Paulsen problem [9, 19]. However, it is not clear in general how to efficiently test for this property; this is why Theorem 15 is so powerful.

Almost Injectivity in the Complex Case

The complex case is not understood nearly as well as the real case, but the phase transition is arguably better understood than the one for injectivity in the complex case. However, almost injectivity hasn't received as much attention, so there are no known characterizations in the complex case, let alone "useful" ones. To begin our discussion of the phase transition, we consider the following lemma (the proof is enjoyable):

Theorem 16. *Suppose $\mathscr{A}: \mathbb{R}^P \to \mathbb{R}^N$ has a continuous Jacobian J over some open set $U \subseteq \mathbb{R}^P$. If $\operatorname{rank}(J(x)) < P$ for every $x \in U$, then $\mathscr{A}$ is not injective when restricted to U.*

Proof. Let z be a point in U which maximizes $\operatorname{rank}(J(x))$, and let K denote the rank of $J(z)$. Then there are K linearly independent columns of $J(z)$ forming the

submatrix $J_{\mathscr{K}}(z)$. Furthermore, these columns remain linearly independent in any sufficiently small neighborhood B of z in U. (This can be established using the continuous mapping $x \mapsto \det[(J_{\mathscr{K}}(x))^* J_{\mathscr{K}}(x)]$.) As such, we can define the continuous mapping

$$x \mapsto P(x) := I - J_{\mathscr{K}}(x)[(J_{\mathscr{K}}(x))^* J_{\mathscr{K}}(x)]^{-1}(J_{\mathscr{K}}(x))^*$$

for all $x \in B$. By construction, $P(x)$ is the orthogonal projection onto the null space of $J(x)$. Pick some nonzero member v of the null space of $J(z)$, and consider the continuous vector field $x \mapsto P(x)v$ over B. By the Peano existence theorem, there exists $\varepsilon > 0$ and $\gamma\colon [0,\varepsilon] \to B$ such that $\gamma(0) = z$ and $\gamma'(t) = P(\gamma(t))v$ for every $t \in [0,\varepsilon]$. Since $\gamma'(t) = P(\gamma(t))v$ is in the null space of $J(\gamma(t))$, we then have

$$0 = J(\gamma(t))\gamma'(t) = \frac{d}{dt}\Big(\mathscr{A}(\gamma(t))\Big)$$

for every $t \in [0,\varepsilon]$, meaning $\mathscr{A}(x)$ is constant over all $x \in \gamma([0,\varepsilon])$. Furthermore, $\gamma([0,\varepsilon])$ contains more than a single point, since otherwise γ is constant, contradicting $\gamma'(0) = v \neq 0$. As such, $\mathscr{A}$ is not injective over any sufficiently small neighborhood of z, let alone U.

Take an almost injective intensity measurement mapping $\mathscr{A}$ and restrict it to an open set S of x's for which $\mathscr{A}^{-1}(\mathscr{A}(x)) = \{\omega x : |\omega| = 1\}$. Note that $\mathscr{A}$ is injective over S by assumption. Considering $(\mathbb{C}^M \setminus \{0\})/\mathbb{T}$ is a smooth manifold of real dimension $P = 2M-1$, we can intersect S with a patch to get an open set U, and consider the Jacobian of $\mathscr{A}$ in the patch's local coordinates. By the contrapositive of Theorem 16, we have that $N \geq \operatorname{rank}(J(x)) \geq P = 2M-1$ for some $x \in U$. This may lead one to believe that $\operatorname{AlmInj}[\Phi, \mathbb{C}^{M\times N}]$ exhibits a phase transition at $N = 2M-1$, but there is evidence to suggest that is this off by 1 due to algebraic properties of intensity measurements:

Conjecture 2. $\operatorname{AlmInj}[\Phi, \mathbb{C}^{M\times N}]$ exhibits a phase transition at $N = 2M$.

To be clear, part (b) of this conjecture was proved by Balan, Casazza, and Edidin [4], whereas a sketch of the proof of part (a) is provided in [31]. However, the latter sketch leaves much to be desired – while the argument is believable in principle, it is unclear whether their use of real algebraic geometry is sufficiently rigorous. For explicit minimal constructions in this case (assuming the conjecture is true), see [30, 31].

Acknowledgments

The author was supported by NSF Grant No. DMS-1321779. The views expressed in this chapter are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

References

1. B. Alexeev, J. Cahill, D.G. Mixon, Full spark frames. J. Four. Anal. Appl. **18**, 1167–1194 (2012)
2. B. Alexeev, A.S. Bandeira, M. Fickus, D.G. Mixon, Phase retrieval with polarization. SIAM J. Imaging Sci. **7**, 35–66 (2014)
3. R. Balan, Reconstruction of signals from magnitudes of redundant representations. Available online: 1207.1134 (2012)
4. R. Balan, P. Casazza, D. Edidin, On signal reconstruction without phase. Appl. Comput. Harmon. Anal. **20**, 345–356 (2006)
5. R. Balan, B.G. Bodmann, P.G. Casazza, D. Edidin, Painless reconstruction from magnitudes of frame coefficients. J. Four. Anal. Appl. **15**, 488–501 (2009)
6. A.S. Bandeira, J. Cahill, D.G. Mixon, A.A. Nelson, Saving phase: injectivity and stability for phase retrieval. Appl. Comput. Harmon. Anal. **37**, 106–125 (2014)
7. A.S. Bandeira, Y. Chen, D.G. Mixon, Phase retrieval from power spectra of masked signals (2013). Available online: arXiv:1303.4458
8. J.J. Benedetto, M. Fickus, Finite normalized tight frames. Adv. Comput. Math. **18**, 357–385 (2003)
9. B.G. Bodmann, P.G. Casazza, The road to equal-norm Parseval frames. J. Funct. Anal. **258**, 397–420 (2010)
10. B.G. Bodmann, N. Hammen, Stable phase retrieval with low-redundancy frames. Adv. Comput. Math. **14**(2), 317–331 (2013). Available online: arXiv:1302.5487
11. O. Bunk, A. Diaz, F. Pfeiffer, C. David, B. Schmitt, D.K. Satapathy, J.F. van der Veen, Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels. Acta Cryst. **A63**, 306–314 (2007)
12. J. Cahill, M. Fickus, D.G. Mixon, M.J. Poteet, N. Strawn, Constructing finite frames of a given spectrum and set of lengths. Appl. Comput. Harmon. Anal. **35**, 52–73 (2013)
13. E.J. Candès, X. Li, Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. Found. Comput. Math. **14**, 1017–1026 (2014)
14. E.J. Candès, X. Li, M. Soltanolkotabi, Phase retrieval from coded diffraction patterns (2013). Available online: arXiv:1310.3240
15. E.J. Candès, Y.C. Eldar, T. Strohmer, V. Voroninski, Phase retrieval via matrix completion. SIAM J. Imaging Sci. **6**, 199–225 (2013)
16. E.J. Candès, T. Strohmer, V. Voroninski, PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. Commun. Pure Appl. Math. **66**, 1241–1274 (2013)
17. P.G. Casazza, J. Kovačević, Equal-norm tight frames with erasures. Adv. Comput. Math. **18**, 387–430 (2003)
18. P.G. Casazza, M. Fickus, D.G. Mixon, Y. Wang, Z. Zhou, Constructing tight fusion frames. Appl. Comput. Harmon. Anal. **30**, 175–187 (2011)
19. P.G. Casazza, M. Fickus, D.G. Mixon, Auto-tuning unit norm frames. Appl. Comput. Harmon. Anal. **32**, 1–15 (2012)
20. A. Conca, D. Edidin, M. Hering, C. Vinzant, An algebraic characterization of injectivity in phase retrieval. Appl. Comput. Harmon. Anal. **38**(2), 346–356 (2013). Available online: arXiv:1312.0158
21. J.C. Dainty, J.R. Fienup, Phase retrieval and image reconstruction for astronomy, in *Image Recovery: Theory and Application*, ed. by H. Stark (Academic Press, New York, 1987)
22. I. Daubechies, A. Grossmann, Y. Meyer, Painless nonorthogonal expansions. J. Math. Phys. **27**, 1271–1283 (1986)
23. L. Demanet, P. Hand, Stable optimizationless recovery from phaseless linear measurements. J. Fourier Anal. Appl. **20**, 199–221 (2014)
24. R.J. Duffin, A.C. Schaeffer, A class of nonharmonic Fourier series. Trans. Am. Math. Soc. **72**, 341–366 (1952)

25. K. Dykema, N. Strawn, Manifold structure of spaces of spherical tight frames. Int. J. Pure Appl. Math. **28**, 217–256 (2006)
26. Y.C. Eldar, S. Mendelson, Phase retrieval: stability and recovery guarantees. Appl. Comput. Harmon. Anal. **36**(3), 473–494 (2012). Available online: arXiv:1211.0872
27. M. Fickus, D.G. Mixon, A.A. Nelson, Y. Wang, Phase retrieval from very few measurements. Linear Algebra Appl. **449**, 475–499 (2013). Available online: arXiv:1307.7176
28. M. Fickus, D.G. Mixon, M.J. Poteet, N. Strawn, Constructing all self-adjoint matrices with prescribed spectrum and diagonal. Adv. Comput. Math. **39**, 585–609 (2013)
29. J.R. Fienup, J.C. Marron, T.J. Schulz, J.H. Seldin, Hubble Space Telescope characterized by using phase-retrieval algorithms. Appl. Opt. **32**, 1747–1767 (1993)
30. J. Finkelstein, Pure-state informationally complete and "really" complete measurements. Phys. Rev. A **70**, 052107 (2004)
31. S.T. Flammia, A. Silberfarb, C.M. Caves, Minimal informationally complete measurements for pure states. Found. Phys. **35**, 1985–2006 (2005)
32. V.K. Goyal, M. Vetterli, N.T. Thao, Quantized overcomplete expansions in $\mathbb{R}^N$: Analysis, synthesis, and algorithms. IEEE Trans. Inf. Theory **44**, 1–31 (1998)
33. R.W. Harrison, Phase problem in crystallography. J. Opt. Soc. Am. A **10**, 1046–1055 (1993)
34. R. Hartshorne, *Algebraic Geometry*. Graduate Texts in Mathematics (Springer, New York, 1977)
35. T. Heinosaari, L. Mazzarella, M.M. Wolf, Quantum tomography under prior information. Commun. Math. Phys. **318**, 355–374 (2013)
36. I.M. James, Euclidean models of projective spaces. Bull. Lond. Math. Soc. **3**, 257–276 (1971)
37. L. Khachiyan, On the complexity of approximating extremal determinants in matrices. J. Complex. **11**, 138–153 (1995)
38. K.H. Mayer, Elliptische Differentialoperatoren und Ganzzahligkeitssätze für charakteristische Zahlen. Topology **4**, 295–313 (1965)
39. J. Miao, T. Ishikawa, Q. Shen, T. Earnest, Extending X-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. Annu. Rev. Phys. Chem. **59**, 387–410 (2008)
40. R.J. Milgram, Immersing projective spaces. Ann. Math. **85**, 473–482 (1967)
41. R.P. Millane, Phase retrieval in crystallography and optics. J. Opt. Soc. Am. A **7**, 394–411 (1990)
42. D.G. Mixon, Saving phase: Injectivity and stability for phase retrieval, Short, Fat Matrices (blog) (2013). Available online: http://dustingmixon.wordpress.com/2013/03/19/saving-phase-injectivity-and-stability-for-phase-retrieval/
43. B.Z. Moroz, Reflections on quantum logic. Int. J. Theor. Phys. **23**, 497–498 (1984)
44. B.Z. Moroz, A.M. Perelomov, On a problem posed by Pauli, Theor. Math. Phys. **101**, 1200–1204 (1994)
45. A. Mukherjee, Embedding complex projective spaces in Euclidean space. Bull. Lond. Math. Soc. **13**, 323–324 (1981)
46. O. Ozyesil, A. Singer, R. Basri, Camera motion estimation by convex programming (2013). Available online: arXiv:1312.5047
47. M. Püschel, J. Kovačević, Real, tight frames with maximal robustness to erasures, in *Proceedings of Data Compression Conference* (2005), pp. 63–72
48. B. Steer, On the embedding of projective spaces in euclidean space. Proc. Lond. Math. Soc. **21**, 489–501 (1970)
49. D.L. Sun, J.O. Smith III, Estimating a signal from a magnitude spectrogram via convex optimization (2012). Available online: arXiv:1209.2076
50. A. Vogt, Position and momentum distributions do not determine the quantum mechanical state, in *Mathematical Foundations of Quantum Theory*, ed. by A.R. Marlow (Academic Press, New York, 1978)

51. V. Voroninski, A comparison between the PhaseLift and PhaseCut algorithms (2012). Available online: http://math.berkeley.edu/~vladv/PhaseCutProofs.pdf
52. I. Waldspurger, A. d'Aspremont, S. Mallat, Phase recovery, MaxCut and complex semidefinite programming. Math. Progr. **149**(1–2), 47–81 (2012). Available online: arXiv:1206.0102
53. A. Walther, The question of phase retrieval in optics. Opt. Acta **10**, 41–49 (1963)

Sparsity-Assisted Signal Smoothing

Ivan W. Selesnick

Abstract This chapter describes a method for one-dimensional signal denoising that simultaneously utilizes both sparse optimization principles and conventional linear time-invariant (LTI) filtering. The method, called 'sparsity-assisted signal smoothing' (SASS), is based on modeling a signal as the sum of a low-pass component and a piecewise smooth component. The problem is formulated as a sparse-regularized linear inverse problem. We provide simple direct methods to set the regularization and non-convexity parameters, the latter if a non-convex penalty is utilized. We derive an iterative optimization algorithm that harnesses the computational efficiency of fast solvers for banded systems. The SASS approach performs a type of wavelet denoising, but does so through sparse optimization rather than through wavelet transforms. The approach is relatively free of the pseudo-Gibbs phenomenon that tends to arise in wavelet denoising.

Key words: Filtering, Total variation denoising, Wavelet denoising, Convex optimization, Sparse optimization

Introduction

This chapter develops a method for noise reduction, called 'sparsity-assisted signal smoothing' (SASS). The proposed SASS approach models the unknown signal of interest, $x(t)$, as the sum

$$x(t) = f(t) + g(t), \tag{1}$$

*This research was support by the NSF under grant CCF-1018020.

I.W. Selesnick (✉)
Department of Electrical and Computer Engineering, School of Engineering, New York University
e-mail: selesi@nyu.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_6

where $f(t)$ is a low-pass signal and $g(t)$ has a sparse order-K derivative. We consider the problem of estimating $x(t)$ from noisy data $y(t) = x(t) + w(t)$, where $w(t)$ is white Gaussian noise. The SASS denoising approach combines the principles of sparse optimization with conventional linear time-invariant (LTI) filtering.[1]

The SASS method involves low-pass filtering and the minimization of a non-differentiable objective function that promotes sparsity of the order-K derivative of $g(t)$. We formulate the SASS approach as a sparse-regularized linear inverse problem, which, after a change of variables, is shown to be a sparse deconvolution problem. Both convex and non-convex regularizations are considered. In both cases, we provide a simple, direct method to set the regularization parameter. In the non-convex case, we also provide a method to set the parameter that controls the degree of non-convexity.

A computationally efficient iterative optimization algorithm is developed for the SASS approach. The SASS approach is intentionally constructed using banded matrices exclusively, so fast solvers for banded systems can be used for its implementation. The optimization algorithm calls for no parameters (step sizes, etc.). In addition, we describe a method for dealing with the zero-locking issue, which can arise in the non-convex case. The method detects and corrects zero locking, when it occurs.

The SASS approach builds upon and extends the practice and capabilities of conventional low-pass filtering [44]. The first step of the method involves the specification of a low-pass filter, the cutoff frequency of which can be set as usual, according to knowledge of the frequency spectrum of the signals in question. In one limiting case ($\lambda \to \infty$), SASS amounts to low-pass filtering. In another limiting case ($\lambda \to 0$), SASS performs no filtering and the output of the method is the noisy input data. For practical values of λ, the SASS approach can be understood as an enhancement of the low-pass filter, as will be illustrated.

Related work

Several recent works have utilized an approach wherein a signal is explicitly expressed as the sum of a low-pass signal and a piecewise constant (i.e., sparse-derivative) signal [30, 52, 53]. In each approach, an inverse problem is formulated where total variation regularization [49] is used to estimate the piecewise constant signal component. These methods differ in the way the low-pass signal component is modeled and estimated. The low-pass component is modeled as locally polynomial in [52], while Tikhonov regularization is used in [30] and conventional low-pass filtering in [53].

The signal model used in these works (low-pass plus piecewise constant) is well suited for applications where additive step discontinuities are observed in the presence of a relatively slow varying signal. For example, the methods described in [52]

[1] Software is available at http://eeweb.poly.edu/iselesni/sass/.

and [53] are demonstrated, respectively, on data produced by a nanoparticle biosensor [16] and a near-infrared spectroscopic (NIRS) imaging device [1]. However, this model is limited because many natural (e.g., physiological) signals do not exhibit additive step discontinuities. Instead, they are more accurately modeled as having discontinuities in a higher order derivative.

The problem addressed in this chapter is an extension of one of the problems addressed in [53]. Specifically, SASS extends the 'LPF/TVD' problem [53] from $K = 1$ to $K > 1$, where K is the order of the sparse derivative. The LPF/TVD algorithm in [53] is not effective for the case $K > 1$ because it estimates the sparse-derivative component by the integration of a sparse signal. Using this approach for $K > 1$ leads to K-order integration which is very unstable; the obtained sparse order-K derivative component will inevitably be unbounded. The SASS approach in this chapter circumvents this problem by estimating the sum $x = f + g$ in (1), without estimating f and g individually.

The SASS approach can also be viewed as an extension of one-dimensional total variation (TV) denoising [8, 9, 49]. TV denoising is based on the assumption that the derivative of $x(t)$ is sparse, i.e., that $x(t)$ is approximately piecewise constant. Total variation denoising is notable due to its ability to preserve discontinuities and the absence of pseudo-Gibbs phenomenon. However, TV denoising is afflicted by staircase artifacts and performs poorly for more general signals. Hence, numerous generalizations for TV have been proposed to make TV denoising more widely effective, e.g., higher-order TV and directional TV. [3, 32, 34, 39]. The proposed SASS approach also uses higher-order TV, but in contrast to these methods, SASS incorporates a low-pass filter (LPF). The incorporation of the low-pass filter enhances both the prospective sparsity of the order-K derivative and the flexibility of high-order TV regularization. In effect, the LPF lifts some of the burden off the high-order total variation regularization.

Wavelet-based signal denoising is also suitably applied to signals of the form considered in this work. Several wavelet-domain algorithms have been developed specifically to account for the presence of singularities (i.e., discontinuities in the signal or its derivatives). In order to suppress spurious oscillations (the pseudo-Gibbs phenomenon) around singularities, which arise due to the modification of wavelet coefficients, these algorithms generally impose some model or constraints in the wavelet domain. Examples of such approaches include wavelet hidden Markov tree (HMT) [15], singularity detection [31, 33], wavelet footprints [20, 56], total variation regularization [21, 22], and singularity approximation [4, 5]. The proposed SASS approach is similar to these techniques in that it accounts for singularities in the signal; however, it does so through sparse optimization instead of wavelet transforms.

Preliminaries

Notation

We represent finite-length discrete-time signals as vectors and denote them in lower case, e.g., we represent the N-point signal $\mathbf{x} \in \mathbb{R}^N$ as

$$\mathbf{x} = [x(0), \ldots, x(N-1)]^{\mathsf{T}}, \tag{2}$$

where $[\cdot]^{\mathsf{T}}$ denotes the transpose. Matrices are denoted in upper case, e.g., $\mathbf{H} \in \mathbb{R}^{L \times N}$.

The notations $\|\mathbf{v}\|_1$ and $\|\mathbf{v}\|_2$ denote the ℓ_1 and ℓ_2 norms of the vector $\mathbf{v}$, respectively, i.e.,

$$\|\mathbf{v}\|_1 = \sum_n |v(n)|, \qquad \|\mathbf{v}\|_2^2 = \sum_n |v(n)|^2. \tag{3}$$

We denote the order-K difference matrix by $\mathbf{D}$. For example, the second-order difference matrix ($K = 2$) is given by

$$\mathbf{D} = \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & & \ddots & & \\ & & & 1 & -2 & 1 \end{bmatrix}, \tag{4}$$

where $\mathbf{D}$ is of size $(N-2) \times N$. The second-order difference of an N-point signal $\mathbf{x}$ is then given by $\mathbf{Dx}$.

In general, $\mathbf{D}$ is a Toeplitz matrix of size $(N-K) \times N$. For $K = 1$, the first row of $\mathbf{D}$ is $[-1, 1]$. For $K = 3$, the first row is $[-1, 3, -3, 1]$. In general, the first row of $\mathbf{D}$ consists of the coefficients of $(1-z)^K$. The order-K difference, $\mathbf{D} : \mathbb{R}^N \to \mathbb{R}^{N-K}$, is precisely defined by $\mathbf{y} = \mathbf{Dx}$, where

$$y(n) = \sum_{k=0}^{K} (-1)^k \binom{K}{k} x(n+K-k),$$

for $0 \leqslant n \leqslant N-K-1$. Note that $\mathbf{D}$ annihilates polynomial of degree $K-1$.

Filters

LTI filters are usually implemented via recursive difference equations [44]; however, in this work, we use banded Toeplitz matrices. We do so because this facilitates incorporating LTI filtering into the sparse optimization framework. It also provides a simple mechanism to perform zero-phase noncausal recursive filtering of finite-length signals. For example, the difference equation

$$a_1 y(n+1) + a_0 y(n) + a_1 y(n-1) = b_1 x(n+1) + b_0 x(n) + b_1 x(n-1) \tag{5}$$

can be written and implemented as

$$\mathbf{y} = \mathbf{A}^{-1}\mathbf{B}\mathbf{x}, \tag{6}$$

where $\mathbf{A}$ is a square banded Toeplitz matrix of the form

$$\mathbf{A} = \begin{bmatrix} a_0 & a_1 & & & \\ a_1 & a_0 & a_1 & & \\ & \ddots & & & \\ & & a_1 & a_0 & a_1 \\ & & & a_1 & a_0 \end{bmatrix}, \tag{7}$$

and $\mathbf{B}$ is similarly a banded Toeplitz matrix (not necessarily square), e.g.,

$$\mathbf{B} = \begin{bmatrix} b_1 & b_0 & b_1 & & \\ & b_1 & b_0 & b_1 & \\ & & \ddots & & \\ & & & b_1 & b_0 & b_1 \end{bmatrix}. \tag{8}$$

We define the filter matrix, $\mathbf{H}$, as

$$\mathbf{H} = \mathbf{A}^{-1}\mathbf{B}, \tag{9}$$

which can be implemented using fast solvers for banded systems [46, Sect 2.4]. Note that the order-K difference matrix, $\mathbf{D}$, represents the filter with transfer function $D(z) = (1 - z^{-1})^K$.

In this work, we use the high-pass filter

$$H(z) = \frac{B(z)}{A(z)} = \frac{(-z + 2 - z^{-1})^d}{(-z + 2 - z^{-1})^d + \alpha (z + 2 + z^{-1})^d}, \tag{10}$$

where $\alpha > 0$ is used to set the cutoff frequency. This is a zero-phase Butterworth filter of order $2d$. The filter has a zero at $z = 1$ of order $2d$. We implement this filter as $\mathbf{H} = \mathbf{A}^{-1}\mathbf{B}$.

Note that, if $K \leqslant 2d$, then the numerator, $\mathbf{B}$, of this high-pass filter satisfies

$$\mathbf{B} = \mathbf{B}_1\mathbf{D}, \tag{11}$$

where $\mathbf{B}_1$ is banded and $\mathbf{D}$ is the order-K difference matrix. In terms of transfer functions, (11) means that $B(z) = B_1(z)D(z)$, and hence $B_1(z)$ has a zero of multiplicity $2d - K$ at $z = 1$.

In Section. "Change of Variables", the filter $\mathbf{A}^{-1}\mathbf{B}_1$ will arise, which we denote as $\mathbf{H}_1$.The transfer function of this filter, $H_1(z)$, is the same as $H(z)$ in (10), but with K fewer zeros at $z = 1$. Further details about the filters are given in [53].

The implementation of a filter as $\mathbf{y} = \mathbf{A}^{-1}\mathbf{B}\mathbf{x}$ does tend to produce transients at the beginning and end of the signal. In practice, we alleviate these transients by polynomial smoothing of the first and last few signal values prior to filtering, as discussed in Section "Preprocessing to avoid start and end transients".

Problem Formulation

In the sparsity-assisted signal smoothing (SASS) approach, it is assumed that the N-point discrete-time data, y, is of the form

$$\mathbf{y} = \mathbf{f} + \mathbf{g} + \mathbf{w}, \quad \mathbf{y}, \mathbf{f}, \mathbf{g}, \mathbf{w} \in \mathbb{R}^N, \tag{12}$$

where f is a low-pass signal, g is a signal with (approximately) sparse order-K derivative, and w is white Gaussian noise. The assumption that the order-K derivative of g is sparse implies that g is (approximately) piecewise polynomial, where each polynomial segment is of degree $K-1$. However, the approach described here does not explicitly parameterize the signal in terms of its polynomial coefficients, nor does it explicitly segment the signal into polynomial segments (cf. splines).

We further assume that if the signal component g were absent in (12), then a low-pass filter can be used to satisfactorily estimate f from the data, y. That is, denoting the low-pass filter as LPF, we assume that $f \approx \mathsf{LPF}\{f + w\}$. Such a low-pass filter should have a zero-phase frequency response (otherwise phase distortion precludes an accurate approximation).

Given the signal y in (12), we aim to estimate the noise-free signal, i.e., $x = f + g$. We note that the approach taken here does not ultimately estimate f and g individually. There is an inherent nonuniqueness regarding f and g: they are each defined only up to an additive piecewise polynomial signal (with polynomial segments of degree $K-1$). This is because a piecewise polynomial signals is both low-pass and has a sparse order-K derivative. By estimating the sum $f+g$, instead of f and g individually, we avoid this ambiguity and improve the conditioning of the estimation problem.

Our approach to estimate x ($x = f + g$) from y is based on the low-pass filter, LPF, which we assume is known. Note that, if an estimate $\hat{g}$ were available, then we may estimate f by low-pass filtering, i.e., $\hat{f} = \mathsf{LPF}\{y - \hat{g}\}$. Then, an estimate $\hat{x}$ is given by

$$\begin{aligned} \hat{x} &= \hat{f} + \hat{g} && (13)\\ &= \mathsf{LPF}\{y - \hat{g}\} + \hat{g} && (14)\\ &= \mathsf{LPF}\{y\} + \mathsf{HPF}\{\hat{g}\}, && (15) \end{aligned}$$

where HPF is the high-pass filter defined by $\mathsf{HPF} = \mathsf{I} - \mathsf{LPF}$. Here, I is the identity operator.

It remains to estimate g. We assumed above that $f \approx \mathsf{LPF}\{f\}$, therefore we have $\mathsf{HPF}\{f\} \approx 0$. Consequently, applying the high-pass filter to (12), we obtain

$$\mathsf{HPF}\{y - \hat{g}\} \approx w. \tag{16}$$

Equation (16) implies that an accurate estimate of g is one that, when subtracted from the data y, yields a signal similar to noise, subsequent to high-pass filtering. Formulating the estimation of g as a linear inverse problem, (16) suggests the data fidelity term should have the form $\|\mathbf{H}(\mathbf{y} - \mathbf{g})\|_2^2$ where $\mathbf{H}$ is the high-pass filter matrix.

Cost function

Based on the foregoing discussion, we formulate the estimation of g as a penalized least squares problem. The penalty term is chosen to promote the behavior of g that we have assumed, i.e., the order-K derivative of g is sparse. Specifically, we formulate SASS as the problem

$$\mathbf{g}^* = \arg\min_{\mathbf{g}} \frac{1}{2}\|\mathbf{H}(\mathbf{y}-\mathbf{g})\|_2^2 + \lambda \sum_n \phi([\mathbf{D}\mathbf{g}]_n), \tag{17}$$

where $\lambda > 0$, $\mathbf{g} \in \mathbb{R}^N$, and $\mathbf{D}$ is the order-K difference matrix. In (17), the notation $[\mathbf{v}]_n$ denotes the nth component of vector $\mathbf{v}$.

We take $\mathbf{H}$ to be a high-pass filter of the form $\mathbf{H} = \mathbf{A}^{-1}\mathbf{B}$ where $\mathbf{A}$ and $\mathbf{B}$ are banded and where, furthermore, $\mathbf{B} = \mathbf{B}_1\mathbf{D}$ where $\mathbf{D}$ is the order-K difference matrix and $\mathbf{B}_1$ is banded (cf. (9) and (11)). Using the high-pass filter in (10), these conditions are satisfied when $K \leq 2d$.

The penalty function $\phi : \mathbb{R} \to \mathbb{R}$ is chosen so as to promote sparsity. Examples of sparsity-promoting functions include

$$\phi(u) = |u| \quad \text{and} \quad \phi(u) = \frac{1}{a}\log(1+a|u|). \tag{18}$$

Numerous other penalty functions have also been utilized for sparse optimization [10, 42]. If $\phi(u)$ is convex, then the optimization problem (17) is convex.

Formulation (17) generalizes the LPF/TVD problem in [53] to the higher-order case ($K > 1$) and to more general (non-convex) penalty functions.

Change of variables

We use a change of variables that simplifies problem (17) in two ways. First, note that the value of the cost function in (17) is unaffected if a polynomial of degree $K-1$ is added to g. This is because $\mathbf{D}$ annihilates such polynomials, as does $\mathbf{H}$ (because $\mathbf{D}$ is a right factor of $\mathbf{H}$). Hence, the minimizer of (17) is unique only up to an additive polynomial. Note in addition that the penalty term in (17) is non-separable; i.e., the elements of g are coupled. This coupling complicates the optimization problem and optimality conditions.

Both issues are eliminated by the change of variables

$$\mathbf{u} = \mathbf{D}\mathbf{g}, \quad \mathbf{u} \in \mathbb{R}^{N-K},\ \mathbf{g} \in \mathbb{R}^N, \tag{19}$$

where $\mathbf{D}$ is the order-K difference matrix of size $(N-K) \times N$. Note that

$$\mathbf{H}\mathbf{g} = \mathbf{A}^{-1}\mathbf{B}\mathbf{g} = \mathbf{A}^{-1}\mathbf{B}_1\mathbf{D}\mathbf{g} = \mathbf{A}^{-1}\mathbf{B}_1\mathbf{u}. \tag{20}$$

Hence, the optimization problem (17) becomes

$$\mathbf{u}^* = \arg\min_{\mathbf{u}} \left\{ F(\mathbf{u}) = \frac{1}{2}\|\mathbf{H}\mathbf{y} - \mathbf{A}^{-1}\mathbf{B}_1\mathbf{u}\|_2^2 + \lambda \sum_n \phi(u(n)) \right\}. \tag{21}$$

Note that the cost function, $F : \mathbb{R}^{N-K} \to \mathbb{R}$, is non-differentiable.

The change of variables (19) effectively eliminates $\mathbf{D}$ from both the penalty term and $\mathbf{H}$. As a result, the solution $\mathbf{u}^*$ is unique and the elements of $\mathbf{u}$ in the penalty term of (21) are decoupled.

Note that, given $\mathbf{u}$, the signal $\mathbf{g}$ cannot be uniquely determined by (19). Hence, we are unable to accurately determine $\mathbf{g}$. However, as we now show, this poses no issue in the estimation of the sum $\mathbf{x} = \mathbf{f} + \mathbf{g}$. From (15), we estimate $\mathbf{x}$ as

$$\hat{\mathbf{x}} = \mathbf{L}\mathbf{y} + \mathbf{H}\mathbf{g}, \tag{22}$$

where $\mathbf{L}$ is the low-pass filter matrix, $\mathbf{L} = \widetilde{\mathbf{I}} - \mathbf{H}$. Note that because $\mathbf{H}$ is not quite square, $\widetilde{\mathbf{I}}$ cannot be exactly the identity matrix. Instead, $\widetilde{\mathbf{I}}$ is obtained from the identity matrix by removing the first and last few rows [53]. The matrix $\widetilde{\mathbf{I}}$ truncates a signal to be compatible with $\mathbf{H}$. Using (20), we have

$$\hat{\mathbf{x}} = \mathbf{L}\mathbf{y} + \mathbf{A}^{-1}\mathbf{B}_1\mathbf{u}, \tag{23}$$

where $\mathbf{u}$ is the sparse signal obtained by minimizing F. Hence, we estimate $\mathbf{x}$ (i.e., $\mathbf{f} + \mathbf{g}$) without estimating $\mathbf{f}$ and $\mathbf{g}$ individually.

The change of variables (19) was also used in [53] for the special case $K = 1$. However, in that work, the component g was explicitly estimated by integration. For $K > 1$, that procedure is very unstable as it leads to K-order integration. The SASS method solves that problem by avoiding the explicit estimation of g; instead, the sum $\hat{x} = \hat{f} + \hat{g}$ is estimated (using (23)).

Optimality condition

When ϕ is convex, then $\mathbf{u}^* \in \mathbb{R}^{N-K}$ minimizes F in (21) if and only if

$$\mathbf{0} \in \partial F(\mathbf{u}^*), \tag{24}$$

where ∂F is the subgradient of F [26]. The subgradient of F is given by

$$\partial F(\mathbf{u}) = \mathbf{H_1}^{\mathsf{T}}(\mathbf{A}^{-1}\mathbf{B}_1\mathbf{u} - \mathbf{H}\mathbf{y}) + \lambda\, \partial\phi(\mathbf{u}), \tag{25}$$

where $\mathbf{H}_1 = \mathbf{A}^{-1}\mathbf{B}_1$. So, the optimality condition (24) can be written as

$$\frac{1}{\lambda}\mathbf{B}_1^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}(\mathbf{B}\mathbf{y} - \mathbf{B}_1\mathbf{u}^*) \in \partial\phi(\mathbf{u}^*). \tag{26}$$

For ℓ_1 norm regularization, i.e., $\phi(u) = |u|$, we have $\partial\phi(u) = \text{sign}(u)$ where sign is the set-valued function,

$$\text{sign}(u) = \begin{cases} \{1\}, & u > 0 \\ [-1, 1], & u = 0 \\ \{-1\}, & u < 0. \end{cases} \tag{27}$$

In this case, the optimality condition (26) can be written as

$$\frac{1}{\lambda}\left[\mathbf{B}_1^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}(\mathbf{B}\mathbf{y} - \mathbf{B}_1\mathbf{u}^*)\right]_n \begin{cases} \in [-1,\ 1], & u^*(n) = 0 \\ = 1, & u^*(n) > 0 \\ = -1, & u^*(n) < 0, \end{cases} \tag{28}$$

for all $0 \leqslant n \leqslant N - K - 1$. This condition can be used to easily validate the optimality of a solution produced by a numerical algorithm. It can also be used to gauge the convergence of an algorithm minimizing F.

This optimality condition can be illustrated graphically as a scatter plot. The values in (28), when plotted versus $u(n)$, must lie on the graph of the sign function; for example, see Fig. 4(a). We used such a scatter plot also in [53] to verify optimality in the convex case (i.e., ℓ_1 norm). Here, we use a scatter plot of this form also in the non-convex case, to verify that the solution is (locally) optimal (see Fig. 4(b)), and more importantly in the non-convex case, to identify and correct falsely locked zeros arising in the optimization process (see Fig. 8), as described in Sec. .

Setting λ

The solution $\mathbf{u}^*$ depends significantly on the regularization parameter, λ. Several methods can be used to set λ. Following [27], we describe a simple method for setting λ. The derivation is based on the optimality condition (28). We used this approach in [53] for the convex case; here we use it also for the non-convex case, as described in Sec. , which is justified if the non-convex regularizer is suitably constrained [51].

Suppose, in some realization of $\mathbf{y}$ in (12), that $\mathbf{g}$ is identically zero. In this case, we hope that $\mathbf{u}^*$ is also identically zero. Equivalently, using (28),

$$\frac{1}{\lambda}\left[\mathbf{B}_1^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{B}\mathbf{y}\right]_n \in [-1,\ 1], \quad \forall n, \tag{29}$$

with $\mathbf{y} = \mathbf{f} + \mathbf{w}$, where $\mathbf{f}$ and $\mathbf{w}$ are the low-pass signal and the additive noise signal, respectively ($\mathbf{g}$ being zero). Note that the matrix in (29) incorporates the high-pass filter $\mathbf{H}$ as a factor; hence, $\mathbf{B}_1^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{B}\mathbf{f} \approx 0$ because the high-pass filter annihilates the low-pass signal, $\mathbf{f}$. Therefore, replacing $\mathbf{y}$ in (29) by $\mathbf{w}$, (29) still holds approximately. Hence, (29) suggests that λ be set as

$$\lambda \approx \max_n \left|\left[\mathbf{B}_1^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{B}\mathbf{w}\right]_n\right|. \tag{30}$$

Equation (30) can be interpreted in the sense of the 'three-sigma' rule, i.e., most observations of a random variable fall within three standard deviations of its mean. Accordingly, we approximate (30) as

$$\lambda \approx 3\,\mathrm{std}\{\mathbf{B}_1^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{B}\mathbf{w}\}. \tag{31}$$

We define $\mathbf{r}$ as the random signal $\mathbf{r} = \mathbf{B}_1^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{B}\mathbf{w}$. If the noise $\mathbf{w}$ is zero-mean white Gaussian with standard deviation σ, then, disregarding start and end transients, the signal $\mathbf{r}$ is stationary and its standard deviation, σ_r, is given by $\sigma_r = \|\mathbf{p}\|_2\,\sigma$ where $\mathbf{p}$ represents the impulse response of the LTI system $\mathbf{P} = \mathbf{B}_1^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{B} = \mathbf{H}_1^{\mathsf{T}}\mathbf{H}$. Hence, in the case of additive white Gaussian noise, (31) can be written as

$$\lambda \approx 3\,\|\mathbf{p}\|_2\,\sigma. \tag{32}$$

This approach sets λ proportional to the noise standard deviation, σ.

Optimization Algorithm

In this section, we describe an iterative algorithm for sparsity-assisted signal smoothing (SASS). The algorithm minimizes F in (21). Numerous algorithmic frameworks for sparse optimization can be used, e.g., FOCUSS [47], iterative reweighted least squares (IRLS) [19], generalizations of IRLS [48, 59], ISTA [18, 23], and proximal splitting methods [13, 14], among others.

The SASS algorithm we propose takes advantage of the fact that the filter matrices, $\mathbf{A}$ and $\mathbf{B}$, and the order-K difference matrix, $\mathbf{D}$, are all banded. As a result, the algorithm is computationally efficient because it can be implemented using fast solvers for banded systems. The algorithm does not require the user to specify additional parameters.

Majorization-minimization

We use the majorization-minimization (MM) optimization framework [25] to develop an iterative algorithm to minimize (21). The MM procedure minimizes a function F by defining an iterative algorithm via

$$\mathbf{u}^{(i+1)} = \arg\min_{\mathbf{u}} G(\mathbf{u}, \mathbf{u}^{(i)}), \tag{33}$$

where i is the iteration index and G is some suitably chosen majorizer of F. Specifically, G should satisfy $G(\mathbf{u},\mathbf{v}) \geqslant F(\mathbf{u})\,\forall \mathbf{u}$, and $G(\mathbf{v},\mathbf{v}) = F(\mathbf{v})$. The MM process is most effective when the chosen majorizer, G, is easily minimized. With initialization $\mathbf{u}^{(0)}$, the update (33) produces a sequence $\mathbf{u}^{(i)}$ converging to a minimizer of F under mild assumptions (or a local minimizer, if F is not convex). Here we use the

MM procedure for both convex and non-convex cases. In either case, the MM procedure ensures the cost function decreases each iteration. For more details, see [25] and references therein.

To construct an easily minimized majorizer of F in (21), we define a quadratic majorizer of the penalty term, $\lambda \sum_n \phi(u(n))$. Assuming the penalty function, $\phi(u)$, is concave on $\mathbb{R}_+$, symmetric, and differentiable for $u \neq 0$, such as the those in (18), a quadratic majorizer can be readily found. Specifically, a majorizer of $\phi(u)$ is given by

$$g(u,v) = \frac{\phi'(v)}{2v}\, u^2 + \phi(v) - \frac{v}{2}\phi'(v), \tag{34}$$

as derived in [50] and illustrated in Fig. 1. That is, for $v \neq 0$,

$$g(u,v) \geqslant \phi(u), \quad \text{for all } u \in \mathbb{R} \tag{35}$$

$$g(v,v) = \phi(v). \tag{36}$$

Note that $g(u,v)$ is quadratic in u.

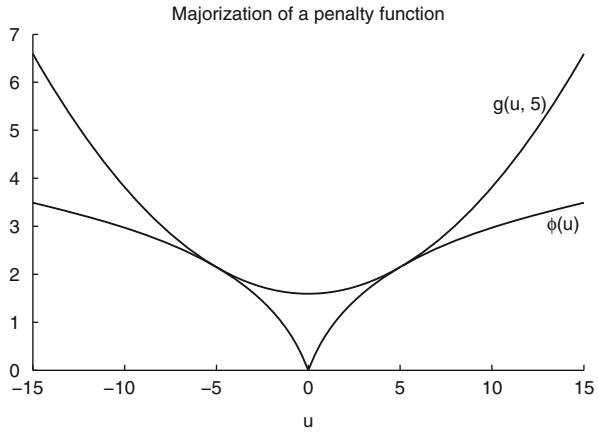


Fig. 1 Majorization of penalty function. The function $g(u,v)$ majorizes the function $\phi(u)$ and is equal to $\phi(u)$ at $u = v$. In the figure, $v = 5$, at which point the two functions are tangent.

The majorizer g can be used to obtain a majorizer for F in (21). If $\mathbf{u}$ and $\mathbf{v}$ are vectors, then

$$\lambda \sum_n g(u(n), v(n)) \geqslant \lambda \sum_n \phi(u(n)), \tag{37}$$

with equality if $\mathbf{u} = \mathbf{v}$. That is, the left-hand side of (37) is a majorizer of $\lambda \sum_n \phi(u(n))$. Moreover, the left-hand side of (37) can be written compactly as

$$\lambda \sum_n g(u(n), v(n)) = \frac{1}{2}\mathbf{u}^{\mathsf{T}} [\mathbf{\Lambda}(\mathbf{v})]^{-1} \mathbf{u} + c, \tag{38}$$

where $\mathbf{\Lambda}(\mathbf{v})$ is a diagonal matrix defined by

$$[\mathbf{\Lambda}(\mathbf{v})]_{n,n} = \frac{1}{\lambda} \frac{v(n)}{\phi'(v(n))}, \tag{39}$$

and c is a constant that does not depend on $\mathbf{u}$. Therefore, using (37), a majorizer of F in (21) is given by

$$G(\mathbf{u},\mathbf{v}) = \frac{1}{2}\|\mathbf{H}\mathbf{y} - \mathbf{A}^{-1}\mathbf{B}_1\mathbf{u}\|_2^2 + \frac{1}{2}\mathbf{u}^{\mathsf{T}}[\mathbf{\Lambda}(\mathbf{v})]^{-1}\mathbf{u} + c.$$

$G(\mathbf{u},\mathbf{v})$ is quadratic in $\mathbf{u}$. Hence, minimizing $G(\mathbf{u},\mathbf{v})$ with respect to $\mathbf{u}$ can be achieved by setting the gradient to zero and solving the resulting linear system. Hence, the MM update equation, (33), leads to

$$\mathbf{u}^{(i+1)} = \arg\min_{\mathbf{u}} G(\mathbf{u},\mathbf{u}^{(i)}) \tag{40}$$

$$= \left(\mathbf{B}_1^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{B}_1 + [\mathbf{\Lambda}(\mathbf{u}^{(i)})]^{-1}\right)^{-1}\mathbf{B}_1^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{B}\mathbf{y} \tag{41}$$

where $\mathbf{\Lambda}(\mathbf{u}^{(i)})$ depends on $\mathbf{u}^{(i)}$ per (39).

There are two numerical complications with (41). First, it calls for the solution to a large $(N \times N)$ dense system of equations, which is computationally costly. Second, the elements of the diagonal matrix $\mathbf{\Lambda}$ go to zero as $\mathbf{u}^{(i)}$ converges to a sparse vector. Therefore, many 'divide-by-zero' errors arise in practice when implementing (41) directly.

To avoid both problems, the matrix inverse lemma (MIL) can be used as described in [24]. Using the MIL, the inverse of the system matrix can be rewritten as

$$\begin{aligned}&\left(\mathbf{B}_1^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{B}_1 + [\mathbf{\Lambda}^{(i)}]^{-1}\right)^{-1} \\ &\quad = \mathbf{\Lambda}^{(i)} - \mathbf{\Lambda}^{(i)}\mathbf{B}_1^{\mathsf{T}}\left(\mathbf{A}\mathbf{A}^{\mathsf{T}} + \mathbf{B}_1\mathbf{\Lambda}^{(i)}\mathbf{B}_1^{\mathsf{T}}\right)^{-1}\mathbf{B}_1\mathbf{\Lambda}^{(i)},\end{aligned} \tag{42}$$

where we use the abbreviation $\mathbf{\Lambda}^{(i)} := [\mathbf{\Lambda}(\mathbf{u}^{(i)})]$, i.e.,

$$[\mathbf{\Lambda}^{(i)}]_{n,n} = \frac{1}{\lambda}\frac{u^{(i)}(n)}{\phi'(u^{(i)}(n))}. \tag{43}$$

Therefore, (41) can be implemented as

$$\mathbf{b} = \mathbf{B}_1^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{B}\mathbf{y} \tag{44}$$

$$\mathbf{Q}^{(i)} = \mathbf{A}\mathbf{A}^{\mathsf{T}} + \mathbf{B}_1\mathbf{\Lambda}^{(i)}\mathbf{B}_1^{\mathsf{T}} \tag{45}$$

$$\mathbf{u}^{(i+1)} = \mathbf{\Lambda}^{(i)}[\mathbf{b} - \mathbf{B}_1^{\mathsf{T}}[\mathbf{Q}^{(i)}]^{-1}\mathbf{B}_1\mathbf{\Lambda}^{(i)}\mathbf{b}]. \tag{46}$$

Note that the matrices, $\mathbf{A}\mathbf{A}^{\mathsf{T}}$ in (44) and $\mathbf{Q}^{(i)}$ in (45), are banded. Therefore, the inverses in (44) and (46) can be implemented very efficiently using fast solvers for

Table 1 Sparsity-assisted signal smoothing (SASS) algorithm

Input: $\mathbf{y} \in \mathbb{R}^N, K, \lambda, \phi, \mathbf{A}, \mathbf{B}$

1. $\mathbf{b} = \mathbf{B}_1^{\mathsf{T}} (\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1} \mathbf{B}\mathbf{y}$
2. $\mathbf{u} = \mathbf{D}\mathbf{y}$ (initialization)

repeat

3. $\boldsymbol{\Lambda}_{n,n} = \dfrac{1}{\lambda} \dfrac{u(n)}{\phi'(u(n))}$ ($\boldsymbol{\Lambda}$ is diagonal)
4. $\mathbf{Q} = \mathbf{A}\mathbf{A}^{\mathsf{T}} + \mathbf{B}_1 \boldsymbol{\Lambda} \mathbf{B}_1^{\mathsf{T}}$ ($\mathbf{Q}$ is banded)
5. $\mathbf{u} = \boldsymbol{\Lambda}[\mathbf{b} - \mathbf{B}_1^{\mathsf{T}} \mathbf{Q}^{-1} \mathbf{B}_1 \boldsymbol{\Lambda} \mathbf{b}]$ (MM update)

until convergence

6. $\mathbf{x} = \widetilde{\mathbf{I}}\mathbf{y} - \mathbf{A}^{-1}\mathbf{B}\mathbf{y} + \mathbf{A}^{-1}\mathbf{B}_1\mathbf{u}$.

output: $\mathbf{x}$, $\mathbf{u}$

banded systems [46, Sect. 2.4]. The complexity of these algorithms are linear in N. In addition, all other matrices appearing in (44)–(46) are diagonal or banded; hence, the matrix multiplications are also efficiently implemented.

The SASS algorithm is summarized in Table 1. Note that the algorithm does not require the user to specify any parameters, such as a step-size parameter. The only parameters are those that define the objective function, (21), i.e., K, λ, ϕ, $\mathbf{A}$, $\mathbf{B}$.

Non-convex penalty functions

Non-convex penalty functions can promote sparsity more strongly than convex penalty functions; hence, they can yield superior results in some sparse optimization problems. Generally, non-convex penalty functions lead to non-convex optimization problems (see [51] for exceptions). Consequently, algorithms have been developed specifically for the non-convex case, based on iterative reweighted ℓ_1 and/or least squares [7, 37, 41, 55, 57, 58], splitting [11, 29], and other optimization methods [28, 42]. Several methods target ℓ_0 pseudo-norm minimization, for example, by single best replacement (SBR) [54] or iterative thresholding [2, 36, 45]. Most of these methods could be applied to the minimization of (21).

We apply the SASS algorithm in both the convex and non-convex cases, the case depending on ϕ. Note that ϕ influences the algorithm only in line 3 of Table 1, i.e., equation (43). So, if we define $\psi : \mathbb{R} \to \mathbb{R}$ as

$$\psi(u) := \frac{u}{\phi'(u)}, \tag{47}$$

then ψ encapsulates the role of the penalty function in the algorithm.

Table 2 Sparse penalties and corresponding weight functions

Penalty, $\phi(u)$	Weight, $\psi(u)$
$\lvert u\rvert$ (i.e., ℓ_1 norm)	$\lvert u\rvert$
$\frac{1}{a}\log(1+a\lvert u\rvert)$	$\lvert u\rvert(1+a\lvert u\rvert)$
$\frac{2}{a\sqrt{3}}\left(\tan^{-1}\left(\frac{1+2a\lvert u\rvert}{\sqrt{3}}\right)-\frac{\pi}{6}\right)$	$\lvert u\rvert(1+a\lvert u\rvert+a^2\lvert u\rvert^2)$

Three penalty functions and the corresponding weight functions are summarized in Table 2. The second and third penalties, involving the logarithmic (log) and arctangent (atan) functions, are both non-convex (strictly concave on $\mathbb{R}_+$) and parameterized by the parameter $a > 0$. Increasing a increases the non-convexity of ϕ. The atan penalty, derived in [51] as a natural extension of the log penalty, promotes sparsity more strongly than the log penalty. Compared to the commonly used ℓ_p psuedo-norm with $p < 1$, the log and atan functions have the advantage that for suitably constrained a, the penalized least squares problem (e.g., (21)) can be convex even when the penalty function is not [51]. For the ℓ_p pseudo-norm, this possibility is precluded due to the fact that the derivative of $|u|^p$ goes to infinity at zero for $p < 1$.

Unfortunately, the use of a general non-convex penalty function complicates the setting of the regularization parameter, λ. For the non-convex case, the simple guideline (32) is not valid in general because that equation is derived based on ℓ_1 norm regularization, i.e., convex regularization. However, if F is convex (even if ϕ is not convex), then the guideline (32) is still valid [51]. Specifically, if the log or atan penalty is used and if a is suitably constrained, then (32) is still useful for setting λ.

When using the logarithmic (log) and arctangent (atan) penalty functions, how should the parameter a be specified? Note that a controls the extent to which the functions are non-convex. As $a \to 0$, the log and atan penalties approach the ℓ_1 norm. For $a > 0$, the log and atan penalties are strictly concave on $\mathbb{R}_+$. An upper bound on a guaranteeing F is convex, can be obtained by semidefinite programming (SDP) [51]. Here, we describe a heuristic to set a, so as to avoid the computational complexity of SDP.

To derive a heuristic to set the non-convexity parameter, a, in the log and atan functions, we make a simplifying assumption. Namely, we assume the sparse vector, $\mathbf{u}^*$, minimizing F, contains only a single nonzero entry. While not satisfied in practice, with this assumption we obtain a value for a above which F is definitely non-convex. Using corollary 1 of [51], this assumption leads to an upper bound on a of $\|\mathbf{h}_1\|_2^2/\lambda$ where $\mathbf{h}_1$ represents the impulse response of the system $\mathbf{H}_1 := \mathbf{A}^{-1}\mathbf{B}_1$. (In computing the impulse response, start and end transients due to signal boundaries should be omitted.) Because the assumption is idealized, this upper bound is

expected to be too high (i.e., not guaranteeing convexity of F); hence a smaller, more conservative value of a is appropriate. In the examples below, we use half this value; i.e., we set

$$a = 0.5\,\|\mathbf{h}_1\|_2^2/\lambda\,. \tag{48}$$

Zero locking

The SASS algorithm in Table 1 is susceptible to 'zero-locking' behavior. That is, if a component $u^{(i)}(m)$ is zero on some iteration i for some m, then it will be zero for all subsequent iterations, i.e., $u^{(k)}(m) = 0$ for all $k > i$. The zero is "locked in". This occurs because if $u^{(i)}(m) = 0$, then $[\mathbf{\Lambda}^{(i)}]_{m,m} = 0$, and consequently $u^{(i+1)}(m) = 0$. For this reason, the algorithm should be initialized with nonzeros.

This zero-locking behavior (or 'singularity issue') is a well-known phenomenon in reweighted methods for sparse optimization [25, 27, 43]. But it does not necessarily impede the convergence of algorithms in which it occurs [25, 27, 43]. We have found that in the convex case (i.e., ℓ_1 norm regularization), the algorithm converges reliably to an optimal solution in practice. We validate optimality using (28).

In the non-convex case, convergence to only a local optimal solution can be expected. We have found experimentally that in the non-convex case, the zero-locking issue sometimes causes the algorithm to converge to a solution that is not locally optimal. This can be recognized using condition (26). Graphically, some points in the scatter plot will lie off the graph of $\phi'(u)$, as illustrated in Example 2 (Fig. 8(a)) below.

In this way, we identify those values, $u(n)$, if any, that are incorrectly locked to zero. Those values can then be perturbed, and the SASS algorithm can be run a second time. In our implementation, we perturb these values using least squares. Specifically, we hold the other components of $\mathbf{u}$ fixed, and solve $\mathbf{Hy} \approx \mathbf{A}^{-1}\mathbf{Bu}$ in the least squares sense over the components of $\mathbf{u}$ that are identified as incorrectly locked to zero. (In our experiments, there are few, if any, incorrectly locked zeros; hence, the least squares problem is small in size, and computationally negligible.) After running the SASS algorithm a second time, we obtain a new sparse vector, $\mathbf{u}$, then we generate a new scatter plot and check for incorrectly locked zeros. In our experiments, we have found that, if λ is set according to (32), then only a second or third run of SASS is sufficient to correct all falsely locked zeros, when they occur. In Example 2 below, we illustrate the use of this process to overcome the zero-locking issue.

Preprocessing to avoid start and end transients

The implementation of a recursive digital filter as $\mathbf{H} = \mathbf{A}^{-1}\mathbf{B}$, where $\mathbf{A}$ and $\mathbf{B}$ are banded matrices, can induce undesirable transients at the start and end of the signal.

In the examples, we alleviate this issue using a simple preprocessing step. Namely, we perform low-order least squares polynomial approximation on the first and last segments of the noisy data, before using the SASS algorithm. In particular, we replace the first and last 15 points of the noisy signal by polynomials of degree r. In Examples 1 and 2, we use polynomials of degree 1 and 2, respectively. In Example 1, we use a second-order high-pass filter, $\mathbf{H}$, (with $d = 1$), which perfectly annihilates the degree 1 polynomial. Likewise, in Example 2, we use a fourth-order high-pass filter ($d = 2$) which annihilates the degree 2 polynomial. Hence, transients due to the signal boundaries are avoided. This is effective, provided the signal has no singularities too close to either of its end points.

Example 1

This example illustrates sparsity-assisted signal smoothing (SASS) to estimate the piecewise smooth signal, of the form $f + g$, shown in Fig. 2(a). The signal, f, consists of several low-frequency sinusoids; g is piecewise linear.

The noisy signal, $y = f + g + w$, illustrated in Fig. 2(b), has additive white Gaussian noise (AWG) with standard deviation $\sigma = 0.5$. We set the low-pass filter, $\mathbf{L} = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{A}^{-1}\mathbf{B}$, to be a second-order zero-phase Butterworth filter ($d = 1$) with a cutoff frequency of $f_c = 0.02$ cycles/sample (10) [53]. The cutoff frequency is set so that the pass-band encompasses the sinusoidal components in f. The output of the low-pass filter, shown in Fig. 2(c), is relatively free of noise. However, parts of the signal are heavily distorted, specifically those parts where the derivative is discontinuous, due to the piecewise linear component, g.

To use the SASS algorithm, the parameters K and λ must be set. In this example, we set $K = 2$; that is, we model g as having a sparse order-2 derivative. Implicitly, we model g as approximately piecewise linear. To set the regularization parameter, λ, we use (32). For the logarithmic and arctangent penalties, we also need to specify the additional parameter, a, which controls the degree of non-convexity of the penalty function. We set a using (48) as described in Section "Non-convex penalty functions". In this example, we have run SASS for 100 iterations. The run time was 36 milliseconds, measured on a 2013 MacBook Pro (2.5 GHz Intel Core i5) running Matlab R2011a.

The output of the SASS algorithm using the ℓ_1 norm penalty is shown in Fig. 2(d). Note that the signal is similar to the low-pass filtered signal, but it exhibits sharp features not possible with low-pass filtering. The RMSE (root-mean-square error) is also substantially better than the RMSE of the low-pass filter. The output of the SASS algorithm using the logarithmic penalty function is shown in Fig. 2(e). The result is similar to that obtained with the ℓ_1 norm, but the sharp features are of somewhat greater amplitude; it also has a lower RMSE. Note that the SASS output signals, in Fig. 2(d, e), are relatively free of Gibbs' phenomenon. There is negligible ringing around the singularities of the signal. This is due, in part, to the SASS

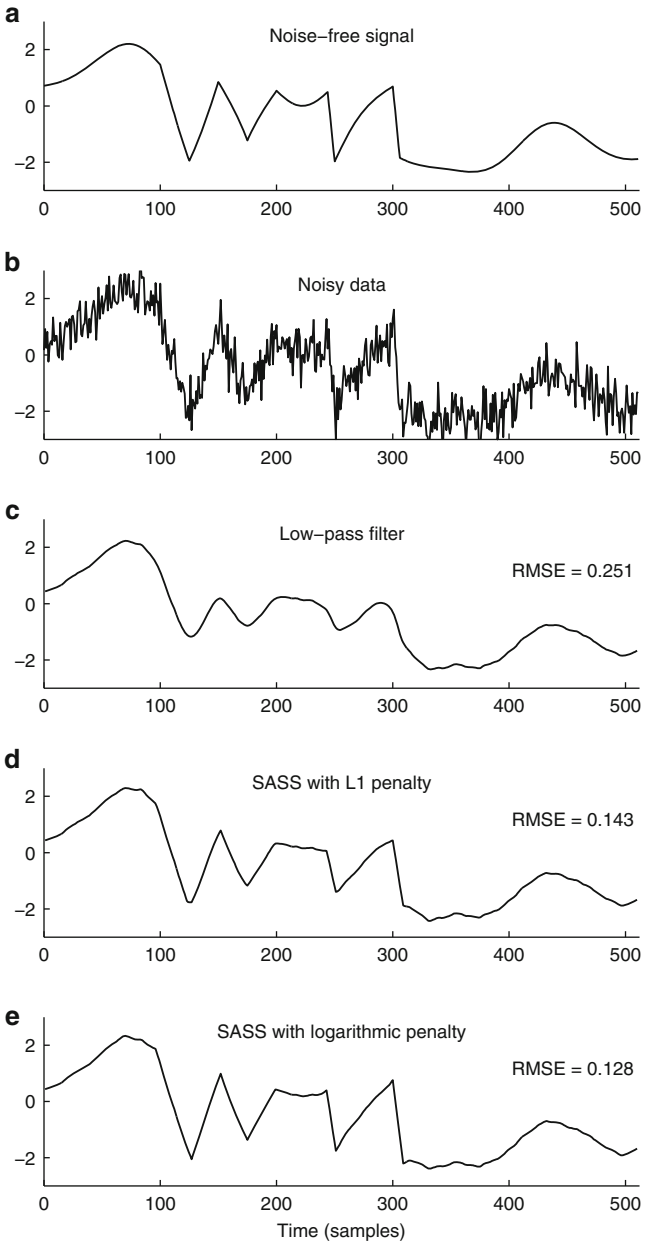


Fig. 2 Example 1. Sparsity-assisted signal smoothing (SASS). (a) Noise-free signal. (b) Noisy data, **y**. (c) Output of low-pass filter, **Ly**. (d, e) Output of SASS algorithm with ℓ_1 norm and logarithmic penalties, respectively.

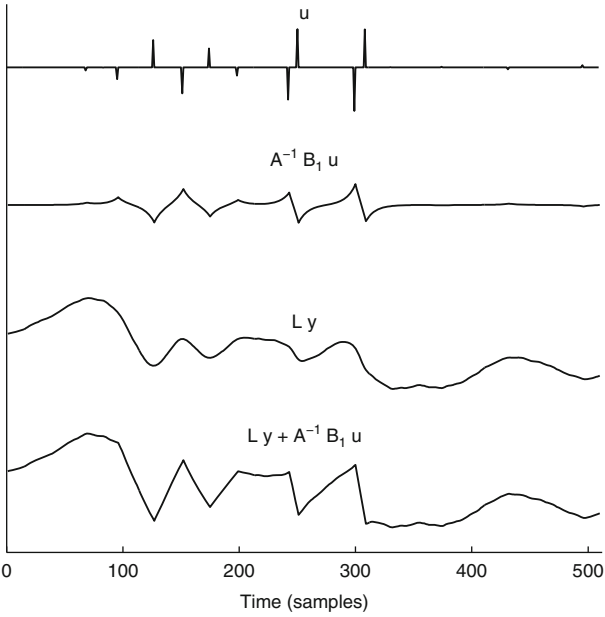


Fig. 3 Example 1. Components of the SASS output shown in Fig. 2(e). The sparse signal, $\mathbf{u}$, is obtained by sparse optimization. The SASS output is given by (23).

approach being based on total variation (TV) denoising, which is free of Gibbs' phenomenon.

We illustrate the SASS solution further in Fig. 3. The sparse vector, $\mathbf{u}$, minimizing F in (21), with the logarithmic penalty, is shown in the figure. Once $\mathbf{u}$ is obtained, we compute $\mathbf{A}^{-1}\mathbf{B}_1\mathbf{u}$ in (23). As shown in the figure, this signal exhibits points where the derivative is discontinuous. The final SASS output is then obtained by adding the low-pass filtered signal, $\mathbf{Ly}$, and $\mathbf{A}^{-1}\mathbf{B}_1\mathbf{u}$, according to (23). Note that the signal, $\mathbf{A}^{-1}\mathbf{B}_1\mathbf{u}$, can be considered an additive correction, or enhancement, obtained via sparse optimization, of the conventional low-pass filter output.

The optimality of the SASS solutions is validated in Fig. 4. Specifically, the points in the scatter plot lie on the graph of $\partial\phi$, which, for the ℓ_1 norm, is the (set-valued) sign function, as illustrated in Fig. 4(a). The preponderance of points on the line, $u = 0$, corresponds to the fact that $\mathbf{u}$ is sparse.

For the logarithmic penalty, which is not convex, the scatter plot can be used to validate locally optimality, only. The points in the scatter plot should lie on the graph of ϕ' for $u \neq 0$, and in the interval $[-1, 1]$ for $u = 0$. For the log penalty, we have

$$\phi'(u) = \frac{1}{a|u| + 1} \operatorname{sign}(u), \quad u \neq 0. \tag{49}$$

As Fig. 4(b) shows, the scatter plot conforms to this condition.

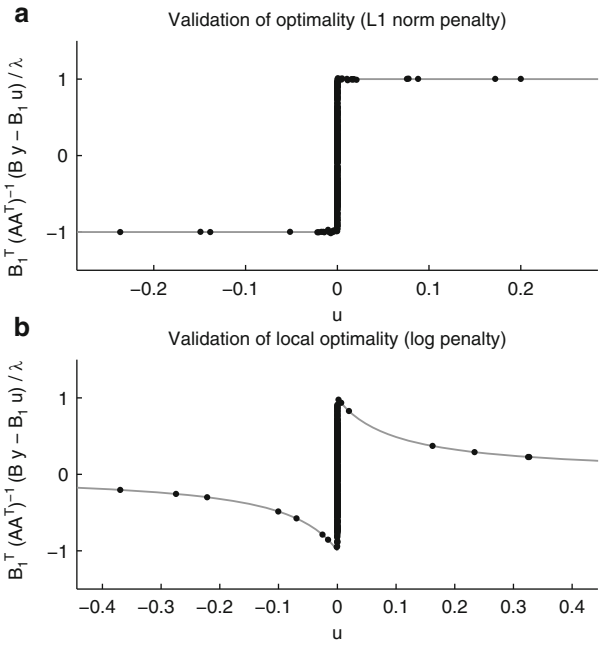


Fig. 4 Example 1: Scatter plots to validate optimality. (a) Optimality of the ℓ_1 norm solution. (b) Local optimality of the logarithmic penalty solution.

For the purpose of comparison, Fig. 5 shows the result of several other techniques suitable for piecewise smooth signal denoising. Fig. 5(a) shows the result of LASIP, which is based on local polynomial approximation over adaptively determined windows [35]. Fig. 5(b) shows the result of discrete wavelet transform hard-thresholding. Fig. 5(c) shows the result of wavelet denoising using a hidden Markov tree (HMT) model [15], which performs substantially better than simple thresholding. Fig. 5(d) shows the result of translation-invariant denoising using an undecimated wavelet transform (UDWT) [12, 38], which also performs well, among wavelet-based methods. For each wavelet method, we used a 5-level transform and the orthonormal Daubechies wavelet with three-vanishing moments [17].

Note that the HMT and UDWT results are qualitatively quite different from each other, although they are based on the same wavelet transform and achieve approximately the same RMSE, which shows the influence of the utilized wavelet-domain model (implicit or explicit). A main issue in obtaining good denoising results using wavelet transforms is the suppression of the pseudo-Gibbs phenomenon which tends to arise in wavelet denoising. This requires using the wavelet transform in conjunction with additional models, penalty terms, constraints, etc. Since the SASS approach does not utilize wavelet transforms, and is based on a simple, explicit model (cf. (1)), it is relatively unaffected by the pseudo-Gibbs phenomenon.

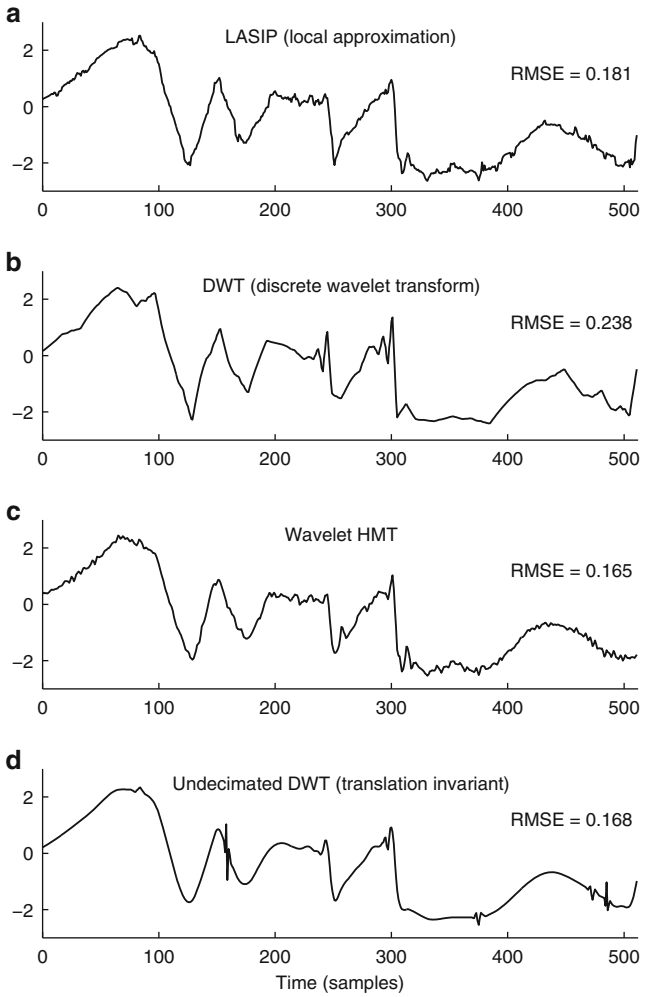


Fig. 5 Example 1. (a) LASIP (local approximation) [35]. (b) DWT (discrete wavelet transform). (c) Wavelet-HMT (hidden Markov tree) [15]. (d) Undecimated wavelet [12]. Compare with Fig. 2.

Example 2 (ECG Denoising)

This example illustrates sparsity-assisted signal smoothing (SASS) on the problem of denoising electrocardiogram (ECG) signals. We use the ECG waveform simulator, ECGSYN [40], to generate synthetic ECG signals, with a sampling rate of 256 samples/second. An example of two seconds of simulated ECG is shown in Fig. 6(a).

The noisy signal, shown in Fig. 6(b), has AWGN with $\sigma = 0.1$. In this example, we set the filter to be a fourth-order zero-phase Butterworth filter ($d = 2$) with a

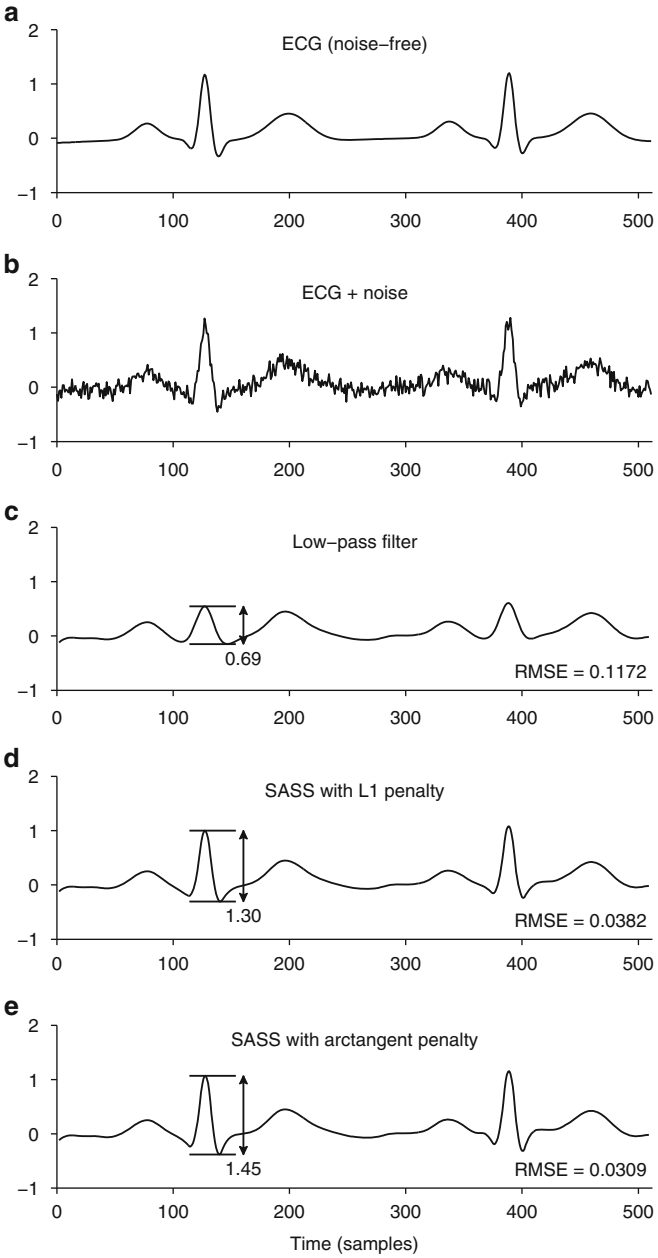


Fig. 6 Example 2. Sparsity-assisted signal smoothing (SASS). (a) Noise-free signal. (b) Noisy data, $\mathbf{y}$. (c) Output of low-pass filter, $\mathbf{L}\mathbf{y}$. (d, e) Output of SASS algorithm with ℓ_1 norm and arctangent penalties, respectively.

cutoff frequency of $f_c = 0.03$ cycles/sample (10). The output of the low-pass filter, shown in Fig. 6(c), is relatively free of noise; however, the peaks of the two QRS complexes are substantially attenuated. The peak-to-peak (P-P) amplitude of the first QRS complex is indicated in the figure. The peaks can be better preserved by using a low-pass filter with a higher cutoff frequency; however, the filter will then let more noise through.

For the SASS algorithm, we use $K = 3$ in this example, which corresponds to modeling the component g as having a sparse order-3 derivative (i.e., approximately piecewise polynomial with polynomial segments of order 2). We set λ using (32) and a using (48). To obtain the atan solution, we initialize the SASS algorithm with the ℓ_1 norm solution.

Figure 6(d, e) shows the output of the SASS algorithm using both the ℓ_1 norm and the arctangent penalty functions. As can be seen, SASS preserves the QRS waveform much more accurately than the low-pass filter. The ℓ_1 solution has a P-P amplitude of 1.30, almost twice that of the LPF. The atan solution has an even higher P-P amplitude of 1.45. The atan penalty function induces less bias and promotes sparsity more strongly, than the ℓ_1 norm penalty. The solution obtained with the logarithmic penalty (not shown) is similar to the atan solution (the P-P amplitude of the log solution is 1.43).

We illustrate, in Fig. 7, the components of the SASS solution obtained using the arctangent penalty. As shown, $\mathbf{u}$ is sparse, and the signal, $\mathbf{A}^{-1}\mathbf{B}_1\mathbf{u}$, is composed of a few zero-mean oscillatory waveforms. The final SASS output is obtained by adding the low-pass filtered signal, $\mathbf{Ly}$, and $\mathbf{A}^{-1}\mathbf{B}_1\mathbf{u}$, according to (23).

The optimality scatter plot and zero-correction procedure are illustrated in Fig. 8 for the SASS solution obtained using the arctangent penalty. For the arctangent

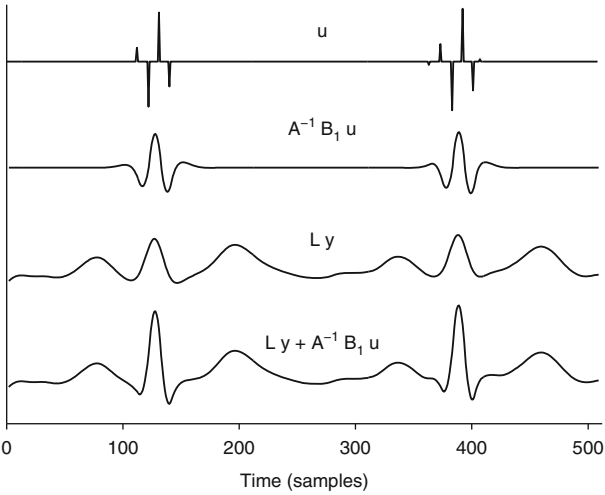


Fig. 7 Example 2. Components of the arctangent solution shown in Fig. 6(e).

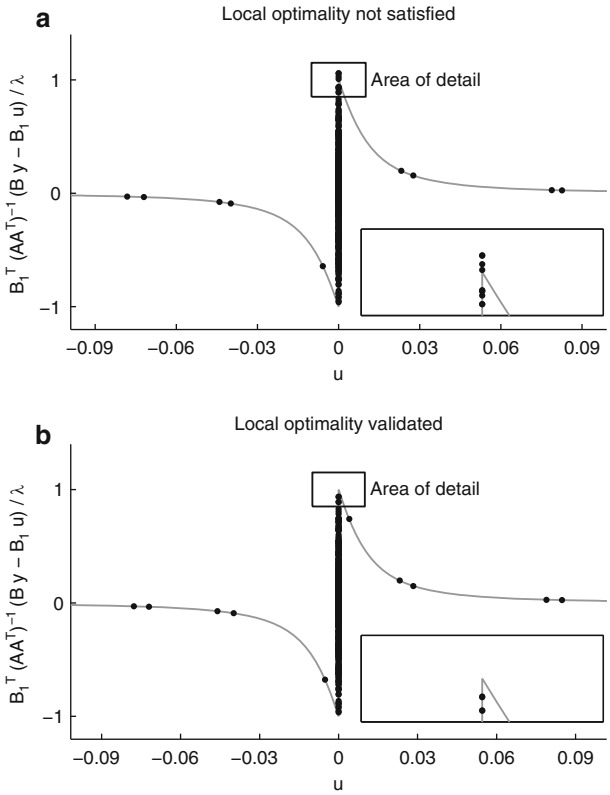


Fig. 8 Example 2: Correction of zero-locking phenomenon. (a) Some points in the scatter plot do not lie on the graph of $\phi'(u)$. The solution is not optimal. (b) After the zero-locking correction procedure, all points in the scatter plot lie on the graph of $\phi'(u)$. The solution is locally optimal.

penalty, which is not convex, the scatter plot can be used to validate locally optimality, only. The points in the scatter plot should lie on the graph of ϕ' for $u \neq 0$, and in $[-1, 1]$ for $u = 0$. For the atan penalty, we have

$$\phi'(u) = \frac{1}{a^2u^2 + a|u| + 1} \operatorname{sign}(u), \quad u \neq 0. \tag{50}$$

The output of the SASS algorithm with the arctangent penalty yields the scatter plot shown in Fig. 8(a). Note, Fig. 8(a) shows that four components of **u** are positive (dots on the graph of $\phi'(u)$, with $u > 0$). Upon close inspection, as shown in the area of detail, it can be seen that three points lie off the graph, on the line, $u = 0$. These values are incorrectly locked to zero, due to the zero-locking phenomena discussed above. Hence, the SASS algorithm has converged to a solution that is not a local optimum. Having identified, in this way, a set of points falsely locked to zero, we perturb these points away from zero and run the SASS algorithm a second time. The perturbation is performed using least squares. The result of the second run is shown

in Fig. 8(b). As can be seen, the points in the scatter plot are entirely on the graph of ϕ', i.e., no values are found to be incorrectly locked to zero. Note, Fig. 8(b) shows that five components of $\mathbf{u}$ are positive. That is, one of the components of $\mathbf{u}$ which was incorrectly locked to zero in the first solution is positive in the second solution. The atan solution shown in Fig. 6(e) is the optimal solution obtained as the result of the second run. We comment that the two atan solutions (local optimum and not) are visually quite similar.

Wavelet Functions

The SASS denoising algorithm can be viewed in terms of wavelet denoising. Both methods are based on the same basic signal model; namely, the representation of a signal as the sum of a low-pass (coarse) component and a piecewise smooth (detail) component [6, 17], i.e., f and g, respectively, in (1).

The similarity can be further illustrated by exhibiting the wavelet-like functions of the SASS algorithm. Note from (23) that the low-pass (coarse) component is given by $\mathbf{Ly}$, i.e., low-pass filtering of the data, $\mathbf{y}$. The piecewise smooth (detail) component is given by $\mathbf{A}^{-1}\mathbf{B}_1\mathbf{u}$, where $\mathbf{u}$ is obtained by sparse optimization. That is, the detail component is a weighted superposition of translations of the impulse response of the filter $\mathbf{H}_1 = \mathbf{A}^{-1}\mathbf{B}_1$. Hence, the impulse response of $\mathbf{H}_1$, denoted $\mathbf{h}_1$, can be considered a "wavelet". The detail component, i.e., $\mathbf{H}_1\mathbf{u}$, is a linear combination of translations of the wavelet. In a wavelet transform, the detail component is a linear combination of both translations and *dilations* of the wavelet. Hence, the SASS approach lacks the multiscale properties of wavelet denoising.

It is informative to examine the wavelet, as it determines the characteristics of the detail component. Figure 9 shows the wavelet for several values of the parameters, d and K. In the SASS approach, the parameter, K, is the order of the difference matrix, $\mathbf{D}$. The parameter, d, determines the order of the high-pass filter $\mathbf{H}$, i.e., $\mathbf{H}$ is of order $2d$, see (10). Both d and K should be small positive integers, with $1 \leqslant K \leqslant 2d$.

As Fig. 9 illustrates, K determines the regularity of the wavelet. For $K = 1$, the wavelet is discontinuous. For $K = 2$, the wavelet is continuous but its derivative is not. For $K = 3$, both the wavelet and its derivative are continuous. The number of vanishing wavelet moments can also be expressed in terms of K and d. Note that the transfer function, $H_1(z)$, has a zero of multiplicity $2d - K$ zeros at $z = 1$; therefore, convolution with the impulse response, $\mathbf{h_1}$, annihilates polynomials of degree $2d - K - 1$. Hence, the wavelet can be understood to have $2d - K$ vanishing moments. Note that when $K = 2d$, the wavelet has no vanishing moments; but, as illustrated in Example 1, this does not preclude its effectiveness as it would for a wavelet transform, due to the role of sparse optimization in SASS. The cutoff frequency, f_c, of the filter, $\mathbf{H}$, influences the scale (width) of the wavelet. Varying f_c has the effect of dilating/contracting the wavelet.

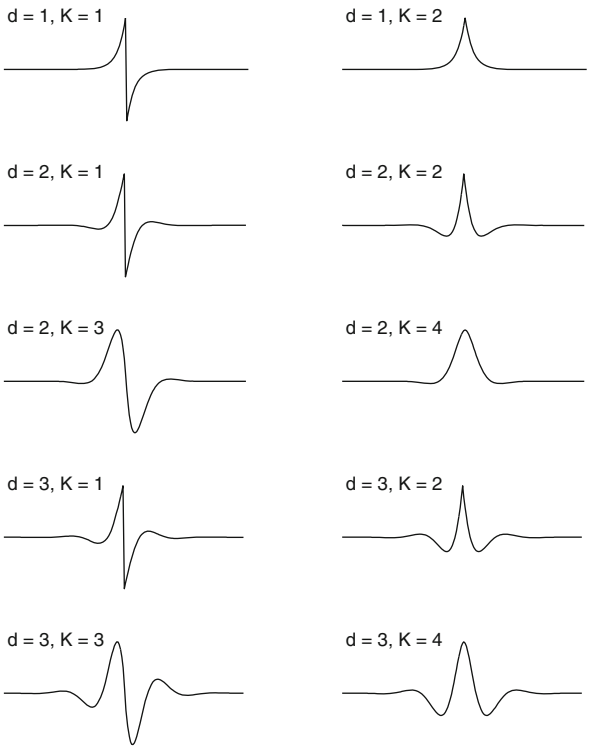


Fig. 9 Impulse response of $\mathbf{H}_1 = \mathbf{A}^{-1}\mathbf{B}_1$ for several values of (d, K), i.e., "wavelets".

Suitable values for d and K, for a particular class of signals, can be chosen based on the characteristics of the wavelet. For example, if it is known that the signals of interest contain additive step discontinuities, then it is reasonable to set $K = 1$, as in [53]. For many signals (e.g., ECG signals), better denoising results are obtained with $K > 1$. (Similarly, wavelets with more than one vanishing moment often produce better results than the Haar wavelet.)

Conclusion

We have described a method for signal denoising that utilizes both conventional filtering principles and sparse optimization principles. The method, 'sparsity-assisted signal smoothing' (SASS), is applicable for denoising signals that are piecewise smooth in a general sense, i.e., for which the order-K derivative can be modeled as (approximately) sparse. The SASS approach is based on the formulation of a sparse-regularized linear inverse problem. We provide simple direct approaches to specify the regularization parameter, λ, and the non-convexity parameter, a, the latter if a

non-convex penalty is utilized. The reweighted least squares algorithm we present, derived using the majorization-minimization principle, is devised so as to maintain the banded property of the involved matrices. Hence, the algorithm is computationally efficient due to the use of fast solvers for banded systems. The optimization algorithm calls for no additional parameters (step sizes, etc.).

The underlying signal model, i.e., a low-pass (coarse) component plus a piecewise smooth (detail) component, also underlies wavelet signal representations. The effectiveness of wavelet-based signal processing is largely due to the sparsity of the wavelet representation of piecewise smooth signals. The proposed SASS approach exploits sparse representations directly via optimization, rather than indirectly through a wavelet transform. Although SASS is not multi-scale, it is relatively free of pseudo-Gibbs phenomenon (oscillations around singularities) that often arises in wavelet processing.

Note that the SASS approach will likely be suboptimal for signals having singularities of two (or more) distinct orders (e.g., signals with both additive step discontinuities and 'ramp' discontinuities). The denoising of signals, having singularities of multiple orders, calls for a generalization of SASS.

References

1. R. Al abdi, H.L. Graber, Y. Xu, R.L. Barbour, Optomechanical imaging system for breast cancer detection. J. Opt. Soc. Am. A **28**(12), 2473–2493 (2011)
2. K. Bredies, D.A. Lorenz, Regularization with non-convex separable constraints. Inverse Prob. **25**(8), 085011 (2009)
3. K. Bredies, K. Kunisch, T. Pock, Total generalized variation. SIAM J. Imag. Sci. **3**(3), 492–526 (2010)
4. V. Bruni, D. Vitulano, Wavelet-based signal de-noising via simple singularities approximation. Signal Process. **86**(4), 859–876 (2006)
5. V. Bruni, B. Piccoli, D. Vitulano, A fast computation method for time scale signal denoising. Signal Image Video Process. **3**(1), 63–83 (2008)
6. C.S. Burrus, R.A. Gopinath, H. Guo, *Introduction to Wavelets and Wavelet Transforms* (Prentice Hall, Upper Saddle River, 1997)
7. E.J. Candès, M.B. Wakin, S. Boyd, Enhancing sparsity by reweighted l1 minimization. J. Fourier Anal. Appl. **14**(5), 877–905 (2008)
8. A. Chambolle, P.-L. Lions, Image recovery via total variation minimization and related problems. Numer. Math. **76**, 167–188 (1997)
9. T.F. Chan, S. Osher, J. Shen, The digital TV filter and nonlinear denoising. IEEE Trans. Image Process. **10**(2), 231–241 (2001)
10. P. Charbonnier, L. Blanc-Feraud, G. Aubert, M. Barlaud, Deterministic edge-preserving regularization in computed imaging. IEEE Trans. Image Process. **6**(2), 298–311 (1997)
11. R. Chartrand, Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data, in *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 262–265, July 2009
12. R.R. Coifman, D.L. Donoho, Translation-invariant de-noising, in *Wavelet and Statistics*, ed. by A. Antoniadis, G. Oppenheim (Springer, Berlin, 1995), pp. 125–150
13. P.L. Combettes, J.-C. Pesquet, Proximal thresholding algorithm for minimization over orthonormal bases. SIAM J. Optim. **18**(4), 1351–1376 (2008)

14. P.L. Combettes, J.-C. Pesquet, Proximal splitting methods in signal processing, in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, ed. by H.H. Bauschke et al. (Springer, Berlin/New York, 2011)
15. M.S. Crouse, R.D. Nowak, R.G. Baraniuk, Wavelet-based signal processing using hidden Markov models. IEEE Trans. Signal Process. **46**(4), 886–902 (1998)
16. V.R. Dantham, S. Holler, V. Kolchenko, Z. Wan, S. Arnold, Taking whispering gallery-mode single virus detection and sizing to the limit. Appl. Phys. Lett. **101**(4), 043704 (2012)
17. I. Daubechies, *Ten Lectures on Wavelets* (SIAM, Philadelphia, 1992)
18. I. Daubechies, M. Defriese, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math **LVII**, 1413–1457 (2004)
19. I. Daubechies, R. DeVore, M. Fornasier, C. Gunturk, Iteratively reweighted least squares minimization for sparse recovery. Commun. Pure Appl. Math. **63**(1), 1–38 (2010)
20. P.L. Dragotti, M. Vetterli, Wavelet footprints: theory, algorithms, and applications. IEEE Trans. Signal Process. **51**(5), 1306–1323 (2003)
21. S. Durand, J. Froment, Reconstruction of wavelet coefficients using total variation minimization. SIAM J. Sci. Comput. **24**(5), 1754–1767 (2003)
22. S. Durand, M. Nikolova, Denoising of frame coefficients using ℓ^1 data-fidelity term and edge-preserving regularization. Multiscale Model. Simul. **6**(2), 547–576 (2007)
23. M. Figueiredo, R. Nowak, An EM algorithm for wavelet-based image restoration. IEEE Trans. Image Process. **12**(8), 906–916 (2003)
24. M. Figueiredo, J. Bioucas-Dias, J.P. Oliveira, R.D. Nowak, On total-variation denoising: a new majorization-minimization algorithm and an experimental comparison with wavelet denoising, in *Proceedings of IEEE International Conference on Image Processing*, 2006
25. M. Figueiredo, J. Bioucas-Dias, R. Nowak, Majorization-minimization algorithms for wavelet-based image restoration. IEEE Trans. Image Process. **16**(12), 2980–2991 (2007)
26. J.-J. Fuchs, On sparse representations in arbitrary redundant bases. IEEE Trans. Inf. Theory **50**(6), 1341–1344 (2004)
27. J.-J. Fuchs, Convergence of a sparse representations algorithm applicable to real or complex data. IEEE. J. Sel. Top. Signal Process. **1**(4), 598–605 (2007)
28. G. Gasso, A. Rakotomamonjy, S. Canu, Recovering sparse signals with a certain family of nonconvex penalties and DC programming. IEEE Trans. Signal Process. **57**(12), 4686–4698 (2009)
29. A. Gholami, S.M. Hosseini, A general framework for sparsity-based denoising and inversion. IEEE Trans. Signal Process. **59**(11), 5202–5211 (2011)
30. A. Gholami, S.M. Hosseini, A balanced combination of Tikhonov and total variation regularizations for reconstruction of piecewise-smooth signals. Signal Process. **93**(7), 1945–1960 (2013)
31. T.-C. Hsung, D.P. Lun, W.-C. Siu, Denoising by singularity detection. IEEE Trans. Signal Process. **47**(11), 3139–3144 (1999)
32. Y. Hu, M. Jacob, Higher degree total variation (HDTV) regularization for image recovery. IEEE Trans. Image Process. **21**(5), 2559–2571 (2012)
33. B. Jalil, O. Beya, E. Fauvet, O. Laligant, Subsignal-based denoising from piecewise linear or constant signal. Opt. Eng. **50**(11), 117004 (2011)
34. F.I. Karahanoglu, I. Bayram, D. Van De Ville, A signal processing approach to generalized 1-d total variation. IEEE Trans. Signal Process. **59**(11), 5265–5274 (2011)
35. V. Katkovnik, K. Egiazarian, J. Astola, *Local Approximation Techniques in Signal and Image Processing* (SPIE Press, Bellingham, 2006)
36. N. Kingsbury, T. Reeves, Redundant representation with complex wavelets: how to achieve sparsity, in *Proceedings of IEEE International Conference on Image Processing*, 2003
37. I. Kozlov, A. Petukhov, Sparse solutions of underdetermined linear systems, in *Handbook of Geomathematics*, ed. by W. Freeden et al. (Springer, New York, 2010)
38. M. Lang, H. Guo, J.E. Odegard, C.S. Burrus, R.O. Wells Jr., Noise reduction using an undecimated discrete wavelet transform. IEEE Signal Process. Lett. **3**(1), 10–12 (1996)

39. S.-H. Lee, M.G. Kang, Total variation-based image noise reduction with generalized fidelity function. IEEE Signal Process. Lett. **14**(11), 832–835 (2007)
40. P.E. McSharry, G.D. Clifford, L.Tarassenko, L.A. Smith, A dynamical model for generating synthetic electrocardiogram signals. Trans. Biomed. Eng. **50**(3), 289–294 (2003)
41. H. Mohimani, M. Babaie-Zadeh, C. Jutten, A fast approach for overcomplete sparse decomposition based on smoothed l0 norm. IEEE Trans. Signal Process. **57**(1), 289–301 (2009)
42. M. Nikolova, M.K. Ng, C.-P. Tam, Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction. IEEE Trans. Image Process. **19**(12), 3073–3088 (2010)
43. J. Oliveira, J. Bioucas-Dias, M.A.T. Figueiredo, Adaptive total variation image deblurring: a majorization-minimization approach. Signal Process. **89**(9), 1683–1693 (2009)
44. T.W. Parks, C.S. Burrus, *Digital Filter Design* (Wiley, New York, 1987)
45. J. Portilla, L. Mancera, L0-based sparse approximation: two alternative methods and some applications, in *Proceedings of SPIE*, vol. 6701 (Wavelets XII), 2007
46. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd edn. (Cambridge University Press, Cambridge, 1992)
47. B.D. Rao, K. Engan, S.F. Cotter, J. Palmer, K. Kreutz-Delgado, Subset selection in noise based on diversity measure minimization. IEEE Trans. Signal Process. **51**(3), 760–770 (2003)
48. P. Rodriguez, B. Wohlberg, Efficient minimization method for a generalized total variation functional. IEEE Trans. Image Process. **18**(2), 322–332 (2009)
49. L. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms. Phys. D **60**, 259–268 (1992)
50. I. Selesnick, Penalty and shrinkage functions for sparse signal processing. Connexions (2012). http://www.cnx.org/content/m45134/
51. I.W. Selesnick, I. Bayram, Sparse signal estimation by maximally sparse convex optimization. IEEE Trans. Signal Process. **62**(5), 1078–1092 (2014)
52. I.W. Selesnick, H.L. Graber, D.S. Pfeil, R.L. Barbour, Simultaneous low-pass filtering and total variation denoising. IEEE Trans. Signal Process. **62**(5), 1109–1124 (2014)
53. I.W. Selesnick, H.L. Graber, D.S. Pfeil, R.L. Barbour, Simultaneous low-pass filtering and total variation denoising. IEEE Trans. Signal Process. (2014, in press). Preprint at http://www.eeweb.poly.edu/iselesni/lpftvd/
54. C. Soussen, J. Idier, D. Brie, J. Duan, From Bernoulli-Gaussian deconvolution to sparse signal restoration. IEEE Trans. Signal Process. **59**(10), 4572–4584 (2011)
55. X. Tan, W. Roberts, J. Li, P. Stoica, Sparse learning via iterative minimization with application to MIMO radar imaging. IEEE Trans. Signal Process. **59**(3), 1088–1101 (2011)
56. D. Van De Ville, B. Forster-Heinlein, M. Unser, T. Blu, Analytical footprints: compact representation of elementary singularities in wavelet bases. IEEE Trans. Signal Process. **58**(12), 6105–6118 (2010)
57. Y. Wang, W. Yin, Sparse signal reconstruction via iterative support detection. SIAM J. Imag. Sci. **3**(3), 462–491 (2010)
58. D. Wipf, S. Nagarajan, Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. IEEE. J. Sel. Top. Signal Process. **4**(2), 317–329 (2010)
59. B. Wohlberg, P. Rodriguez, An iteratively reweighted norm algorithm for minimization of total variation functionals. IEEE Signal Process. Lett. **14**(12), 948–951 (2007)

A Message-Passing Approach to Phase Retrieval of Sparse Signals

Philip Schniter and Sundeep Rangan

Abstract In phase retrieval, the goal is to recover a signal $\boldsymbol{x} \in \mathbb{C}^N$ from the magnitudes of linear measurements $\boldsymbol{Ax} \in \mathbb{C}^M$. While recent theory has established that $M \approx 4N$ intensity measurements are necessary and sufficient to recover generic $\boldsymbol{x}$, there is great interest in reducing the number of measurements through the exploitation of sparse $\boldsymbol{x}$, which is known as compressive phase retrieval. In this work, we detail a novel, probabilistic approach to compressive phase retrieval based on the generalized approximate message passing (GAMP) algorithm. We then present a numerical study of the proposed PR-GAMP algorithm, demonstrating its excellent phase-transition behavior, robustness to noise, and runtime. For example, to successfully recover K-sparse signals, approximately $M \geq 2K\log_2(N/K)$ intensity measurements suffice when $K \ll N$ and $\boldsymbol{A}$ has i.i.d Gaussian entries. When recovering a 6k-sparse 65k-pixel grayscale image from 32k randomly masked and blurred Fourier intensity measurements, PR-GAMP achieved 99% success rate with a median runtime of only 12.6 seconds. Compared to the recently proposed CPRL, sparse-Fienup, and GESPAR algorithms, experiments show that PR-GAMP has a superior phase transition and orders-of-magnitude faster runtimes as the problem dimensions increase.

Key words: Phase retrieval, Compressed sensing, Sparsity, Belief propagation, Message passing

P. Schniter (✉)
Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43202, USA
e-mail: schniter@ece.osu.edu

S. Rangan
Department of Electrical and Computer Engineering, Polytechnic Institute of New York University, Brooklyn, NY 11201, USA
e-mail: srangan@poly.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_7

1 Introduction

1.1 Phase retrieval

In phase retrieval, the goal is to recover a signal $\boldsymbol{x} \in \mathbb{C}^N$ from the *magnitudes* $y_m = |u_m|$ of possibly noisy linear measurements $\boldsymbol{u} = [u_1, \ldots, u_M]^{\mathsf{T}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w} \in \mathbb{C}^M$. This problem is motivated by the fact that it is often easier to build detectors (e.g., photographic plates or CCDs) that measure intensity rather than phase [1, 2]. Imaging applications of phase retrieval include X-ray diffraction imaging [3], X-ray crystallography [4, 5], array imaging [6], optics [7], speckle imaging in astronomy [8], and microscopy [9]. Nonimaging applications include acoustics [10], interferometry [11], and quantum mechanics [12].

To reconstruct $\boldsymbol{x} \in \mathbb{C}^N$ (up to a global phase uncertainty), it has been recently established that $M \geq 4N - o(N)$ intensity measurements are necessary [13] and $M \geq 4N - 4$ are sufficient [14] through appropriate design of the linear transform $\boldsymbol{A}$. Meanwhile, to reconstruct $\boldsymbol{x} \in \mathbb{R}^N$ (up to a global sign uncertainty), it has been shown that $M \geq 2N - 1$ measurements are both necessary and sufficient [10]. However, there exist applications where far fewer measurements are available, such as sub-wavelength imaging [15, 16], Bragg sampling from periodic crystalline structures [17], and waveguide-based photonic devices [18]. To facilitate these *compressive* phase retrieval tasks, it has been proposed to exploit *sparsity*[1] in $\boldsymbol{x}$. In fact, very recent theory confirms the potential of this approach: to reconstruct K-sparse N-length $\boldsymbol{x}$ using a generic (e.g., i.i.d Gaussian) $\boldsymbol{A}$, only $M \geq 4K - 2$ intensity measurements suffice in the complex case and $M \geq 2K$ suffice in the real case (where $M \geq 2K$ is also necessary) when $K < N$ [19]. While these bounds are extremely encouraging, achieving them with a practical algorithm remains elusive.

To our knowledge, the first algorithm for compressive phase retrieval was proposed by Moravec, Romberg, and Baraniuk in [20] and worked by incorporating an ℓ_1 norm *constraint* into a traditional Fienup-style [1] iterative algorithm. However, this approach requires that the ℓ_1 norm of the true signal is known, which is rarely the case in practice. Recently, a more practical sparse-Fienup algorithm was proposed by Mukherjee and Seelamantula [21], which requires knowledge of only the signal sparsity K but is applicable only to measurement matrices $\boldsymbol{A}$ for which $\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A} = \boldsymbol{I}$. Although this algorithm guarantees that the residual error $\|\boldsymbol{y} - |\boldsymbol{A}\widehat{\boldsymbol{x}}(t)|\|_2^2$ is nonincreasing over the iterations t, it succumbs to local minima and, as we show in Section 4.4, is competitive only in the highly sparse regime.

To circumvent the local minima problem, Ohlsson, Yang, Dong, and Sastry proposed the *convex relaxation* known as Compressive Phase Retrieval via Lifting (CPRL) [22], which adds ℓ_1 regularization to the well-known PhaseLift algorithm [6, 23]. Both CPRL and PhaseLift "lift" the unknown vector $\boldsymbol{x} \in \mathbb{C}^N$ into the space

[1] $\boldsymbol{x}$ may represent the sparse transform coefficients of a non-sparse signal-of-interest $\boldsymbol{s} = \boldsymbol{\Psi}\boldsymbol{x}$ in a sparsifying basis (or frame) $\boldsymbol{\Psi}$, in which case the intensity measurements would be $\boldsymbol{y} = |\boldsymbol{\Phi}\boldsymbol{s} + \boldsymbol{w}|$ and $\boldsymbol{A} \triangleq \boldsymbol{\Phi}\boldsymbol{\Psi}$.

of $N \times N$ rank-one matrices and solve a semidefinite program in the lifted space, requiring $O(N^3)$ complexity, which is impractical for practical image sizes N. Subsequent theoretical analysis [19] revealed that while $M \gtrsim O(K^2 \log N)$ intensity measurements suffice for CPRL when $\boldsymbol{x} \in \mathbb{R}^N$, $M \gtrsim O(K^2/\log^2 N)$ measurements are *necessary*, which is disappointing because this greatly exceeds the $2K$ measurements that suffice for the optimal solver [19]. More recently, a cleverly initialized alternating minimization (AltMin) approach was proposed by Natrapalli, Jain, and Sanghavi in [24] that gives CPRL-like guarantees/performance with only $O(NK^3)$ complexity, but this is still too complex for practical sparsities K (which tend to grow linearly with image size N).

Recently, Shechtman, Beck, and Eldar proposed the GrEedy Sparse PhAse Retrieval (GESPAR) algorithm [25], which applies fast 2-opt local search [26] to a sparsity constrained nonlinear optimization formulation of the phase-retrieval problem. Numerical experiments (see Section 4.4) show that GESPAR handles higher sparsities K than the sparse-Fienup technique from [21], but at the cost of significantly increased runtime. In fact, due to the combinatorial nature of GESPAR's support optimization, its complexity scales rapidly in K, making it impractical for many problems of interest.

In this work, we describe a novel[2] approach to compressive retrieval that is based on loopy belief propagation and, in particular, the *generalized approximate message passing* (GAMP) algorithm from [27]. In addition to describing and deriving our phase retrieval GAMP (PR-GAMP) algorithm, we present a detailed numerical study of its performance. For i.i.d Gaussian, Fourier, and masked Fourier matrices $\boldsymbol{A}$, we demonstrate that PR-GAMP performs far better than existing compressive phase retrieval algorithms in terms of both success rate and runtime for large values K and N. Interestingly, we find that PR-GAMP requires approximately $4\times$ the number of measurements as phase-oracle GAMP (i.e., GAMP given the magnitude-and-phase measurements $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$), which generalizes what is known about phase retrieval of *non*-sparse signals in $\mathbb{C}^N$, where the ratio of magnitude-only to magnitude-and-phase measurements necessary and sufficient for perfect recovery is also 4 for large N [13, 14]. We also find that PR-GAMP is robust to additive noise, giving mean-squared error that is only 3 dB worse than phase-oracle GAMP over a wide SNR range.

Notation: For matrices, we use boldface capital letters like $\boldsymbol{A}$, and we use $\boldsymbol{A}^{\mathsf{T}}$, $\boldsymbol{A}^{\mathsf{H}}$, and $\|\boldsymbol{A}\|_F$ to denote the transpose, Hermitian transpose, and Frobenius norm, respectively. For vectors, we use boldface small letters like $\boldsymbol{x}$, and we use $\|\boldsymbol{x}\|_p = (\sum_n |x_n|^p)^{1/p}$ to denote the ℓ_p norm, with $x_n = [\boldsymbol{x}]_n$ representing the n*th* element of $\boldsymbol{x}$. For random variable X, we write the pdf as $p_X(x)$, the expectation as $\mathrm{E}\{X\}$, and the variance as $\mathrm{var}\{X\}$. In some cases where it does not cause confusion, we drop the subscript on $p_X(x)$ and write the pdf simply as $p(x)$. For a circular-Gaussian random variable X with mean m and variance v, we write the pdf as $p_X(x) = \mathcal{N}(x; m, v) \triangleq \frac{1}{\pi v}\exp(-|x-m|^2/v)$. For the point mass at $x = 0$, we use the Dirac delta distribution

[2] We previously described PR-GAMP in the conference paper [28] and the workshop presentation [29].

$\delta(x)$. Finally, we use $\mathbb{R}$ for the real field, $\mathbb{C}$ for the complex field, $\mathrm{Re}\{x\}$ and $\mathrm{Im}\{x\}$ for the real and imaginary parts of x, and x^* for the complex conjugate of x.

2 Background on GAMP

The approximate message passing (AMP) algorithm was recently proposed by Donoho, Maleki, and Montanari [30, 31] for the task of estimating a signal vector $\boldsymbol{x} \in \mathbb{R}^N$ from linearly transformed and additive-Gaussian-noise corrupted measurements[3]

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w} \in \mathbb{C}^M. \tag{1}$$

The Generalized-AMP (GAMP) algorithm proposed by Rangan [27] then extends the methodology of AMP to the generalized linear measurement model

$$\boldsymbol{y} = q(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}) \in \mathbb{C}^M, \tag{2}$$

where $q(\cdot)$ is a component-wise nonlinearity. This nonlinearity affords the application of AMP to phase retrieval.

Both AMP and GAMP can be derived from the perspective of *belief propagation* [32], a Bayesian inference strategy that is based on a factorization of the signal posterior pdf $p(\boldsymbol{x}|\boldsymbol{y})$ into a product of simpler pdfs that, together, reveal the probabilistic structure in the problem. Concretely, if we model the signal coefficients in $\boldsymbol{x}$ and noise samples in $\boldsymbol{w}$ from (1)-(2) as statistically independent, so that $p(\boldsymbol{x}) = \prod_{n=1}^N p_{X_n}(x_n)$ and $p(\boldsymbol{y}|\boldsymbol{z}) = \prod_{m=1}^M p_{Y|Z}(y_m|z_m)$ for $\boldsymbol{z} \triangleq \boldsymbol{A}\boldsymbol{x}$, then we can factor the posterior pdf as

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) \tag{3}$$

$$= \prod_{m=1}^M p_{Y|Z}\big(y_m\big|[\boldsymbol{A}\boldsymbol{x}]_m\big) \prod_{n=1}^N p_{X_n}(x_n), \tag{4}$$

yielding the factor graph in Fig. 1.

In belief propagation [32], beliefs about the unknown variables are passed among the nodes of the factor graph until all agree on a common set of beliefs. The set of beliefs passed into a given variable node are then used to determine the posterior pdf of that variable, or an approximation thereof. The sum-product algorithm [33] is perhaps the most well-known incarnation of belief propagation, wherein the messages take the form of pdfs and exact posteriors are guaranteed whenever the graph does not have loops. For graphs with loops, exact inference is known to be NP hard,

[3] Here and elsewhere, we use $\boldsymbol{y}$ when referring to the M measurements that are available for signal reconstruction. In the canonical (noisy) compressive sensing problem, the measurements take the form $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$, but in the (noisy) compressive phase retrieval problem, the measurements instead take the form $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}|$.

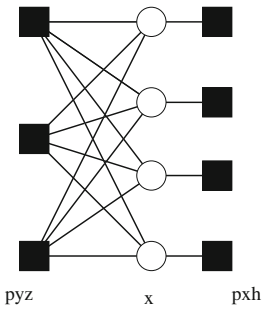


Fig. 1 GAMP factor graph, with white circles denoting random variables and black squares denoting pdf factors, for the case $M = 3$ and $N = 4$.

and so loopy belief propagation (LBP) is not guaranteed to produce correct posteriors. Still, LBP has shown state-of-the-art performance on many problems in, e.g., decoding, computer vision, and compressive sensing [34].

The conventional wisdom surrounding LBP says that accurate inference is possible only when the circumference of the loops is relatively large. With (1)-(2), this would require that $\boldsymbol{A}$ is a sparse matrix, which precludes most interesting cases of compressive inference, including compressive phase retrieval. Hence, the recent realization by Donoho, Maleki, Montanari, and Bayati that LBP-based compressive sensing is not only feasible [30] for *dense* matrices $\boldsymbol{A}$, but provably accurate [35, 36], was a breakthrough. In particular, they established that, in the large system limit (i.e., as $M, N \to \infty$ with M/N fixed) and under i.i.d sub-Gaussian $\boldsymbol{A}$, the iterations of AMP are governed by a state evolution whose fixed points describe the algorithm's performance. To derive the AMP algorithm, Donoho et al. [30] proposed an ingenious set of message-passing approximations that become exact in the limit of large sub-Gaussian $\boldsymbol{A}$.

Remarkably, the "approximate message passing" (AMP) principles in [30]—including the state evolution—can be extended from the linear model (1) to the generalized linear model in (2), as established in [27]. The GAMP algorithm from [27] is summarized in Table 1, where $\mathcal{N}(z; \widehat{z}, \nu^z)$ is used to denote the circular-Gaussian pdf in variable z with mean $\widehat{z}$ and variance ν^z. In the sequel, we detail how GAMP and some extensions of GAMP, allow us to tackle the phase retrieval problem.

3 Phase Retrieval GAMP

To apply the GAMP algorithm outlined in Table 1 to compressive phase retrieval, we specify a measurement likelihood function $p_{Y|Z}(y_m|\cdot)$ that models the lack of phase information in the observations y_m and a signal prior pdf $p_{X_n}(\cdot)$ that facilitates measurement compression, e.g., a sparsity-inducing pdf. In addition, we propose several modifications to the GAMP algorithm that aim to improve its robustness,

and we propose an expectation-maximization method to learn the noise variance that parameterizes $p_{Y|Z}(y_m|\cdot)$.

Table 1 The GAMP Algorithm from [27] with $T_{\max}$ iterations.

input $\boldsymbol{A}, \{p_{X_n}(\cdot), \widehat{x}_n(1), \nu^x_n(1)\}_{n=1}^N, \{p_{Y|Z}(y_m|\cdot), \widehat{s}_m(0)\}_{m=1}^M$

define

$$p_{Z|Y,P}(z|y,\widehat{p};\nu^p) = \frac{p_{Y|Z}(y|z)\mathcal{N}(z;\widehat{p},\nu^p)}{\int_{z'} p_{Y|Z}(y|z')\mathcal{N}(z';\widehat{p},\nu^p)} \quad \text{(D1)}$$

$$g_{\mathsf{out},m}(\widehat{p},\nu^p) = \frac{1}{\nu^p}\left(\mathrm{E}_{Z|Y,P}\{Z|y_m,\widehat{p};\nu^p\} - \widehat{p}\right) \quad \text{(D2)}$$

$$g'_{\mathsf{out},m}(\widehat{p},\nu^p) = \frac{1}{\nu^p}\left(\frac{\mathrm{var}_{Z|Y,P}\{Z|y_m,\widehat{p};\nu^p\}}{\nu^p} - 1\right) \quad \text{(D3)}$$

$$p_{X_n|R_n}(x|\widehat{r};\nu^r) = \frac{p_{X_n}(x)\mathcal{N}(x;\widehat{r},\nu^r)}{\int_{x'} p_{X_n}(x')\mathcal{N}(x';\widehat{r},\nu^r)} \quad \text{(D4)}$$

$$g_{\mathsf{in},n}(\widehat{r},\nu^r) = \mathrm{E}_{X_n|R_n}\{X_n|\widehat{r};\nu^r\} \quad \text{(D5)}$$

$$g'_{\mathsf{in},n}(\widehat{r},\nu^r) = \mathrm{var}_{X_n|R_n}\{X_n|\widehat{r};\nu^r\} \quad \text{(D6)}$$

for $t = 1,2,3,\ldots,T_{\max}$

$$\forall m: \nu^p_m(t) = \textstyle\sum_{n=1}^N |a_{mn}|^2 \nu^x_n(t) \quad \text{(R1)}$$

$$\forall m: \widehat{p}_m(t) = \textstyle\sum_{n=1}^N a_{mn}\widehat{x}_n(t) - \nu^p_m(t)\widehat{s}_m(t-1) \quad \text{(R2)}$$

$$\forall m: \widehat{s}_m(t) = g_{\mathsf{out},m}(\widehat{p}_m(t), \nu^p_m(t)) \quad \text{(R3)}$$

$$\forall m: \nu^s_m(t) = -g'_{\mathsf{out},m}(\widehat{p}_m(t), \nu^p_m(t)) \quad \text{(R4)}$$

$$\forall n: \nu^r_n(t) = \left(\textstyle\sum_{m=1}^M |a_{mn}|^2 \nu^s_m(t)\right)^{-1} \quad \text{(R5)}$$

$$\forall n: \widehat{r}_n(t) = \widehat{x}_n(t) + \nu^r_n(t) \textstyle\sum_{m=1}^M a^*_{mn}\widehat{s}_m(t) \quad \text{(R6)}$$

$$\forall n: \nu^x_n(t+1) = \nu^r_n(t) g'_{\mathsf{in},n}(\widehat{r}_n(t), \nu^r_n(t)) \quad \text{(R7)}$$

$$\forall n: \widehat{x}_n(t+1) = g_{\mathsf{in},n}(\widehat{r}_n(t), \nu^r_n(t)) \quad \text{(R8)}$$

end

output $\{\widehat{x}_n(T_{\max}+1), \nu^x_n(T_{\max}+1)\}_{n=1}^N, \{\widehat{s}_m(T_{\max})\}_{m=1}^M$

3.1 Likelihood function

Before deriving the likelihood function $p_{Y|Z}(y_m|\cdot)$, we introduce some notation. First, we will denote the noiseless transform outputs by

$$z_m \triangleq \boldsymbol{a}_m^{\mathsf{H}}\boldsymbol{x} = |z_m|e^{j\phi_m} \text{ with } \phi_m \in [0, 2\pi), \quad (5)$$

where $\boldsymbol{a}_m^{\mathsf{H}}$ is the mth row of $\boldsymbol{A}$ and $j \triangleq \sqrt{-1}$. Next, we will assume the presence of additive noise w_m and denote the noisy transform outputs by

$$u_m \triangleq z_m + w_m = |u_m|e^{j\theta_m} \text{ with } \theta_m \in [0, 2\pi). \quad (6)$$

Our (noisy) intensity measurements are then

$$y_m = |u_m| \text{ for } m = 1,\ldots,M, \quad (7)$$

Henceforth, we assume additive white circular-Gaussian noise (AWGN) $w_m \sim \mathcal{N}(0, \nu^w)$. Thus, if we condition on z_m, then u_m is circular Gaussian with mean z_m and variance ν^w, and y_m is Rician with pdf [37]

$$p_{Y|Z}(y_m|z_m;\nu^w) = \frac{2y_m}{\nu^w}\exp\left(-\frac{y_m^2+|z_m|^2}{\nu^w}\right) I_0\left(\frac{2y_m|z_m|}{\nu^w}\right) 1_{y_m\geq 0}, \tag{8}$$

where $I_0(\cdot)$ is the $0th$-order modified Bessel function of the first kind.

The functions $g_{\mathsf{out},m}(\cdot,\cdot)$ and $g'_{\mathsf{out},m}(\cdot,\cdot)$ defined in steps (D1)–(D3) of Table 1 can be computed using the expressions

$$\mathrm{E}_{Z|Y,P}\{Z|y_m,\widehat{p}_m;\nu_m^p\} = \frac{\int_{\mathbb{C}} z\, p_{Y|Z}(y_m|z;\nu^w)\mathcal{N}(z;\widehat{p}_m,\nu_m^p)dz}{\int_{\mathbb{C}} p_{Y|Z}(y_m|z';\nu^w)\mathcal{N}(z';\widehat{p}_m,\nu_m^p)dz'} \tag{9}$$

$$= \left(\frac{y_m}{1+\nu^w/\nu_m^p}R_0(\rho_m) + \frac{|\widehat{p}_m|}{\nu_m^p/\nu^w+1}\right)\frac{\widehat{p}_m}{|\widehat{p}_m|} \tag{10}$$

and

$$\begin{aligned}&\mathrm{var}_{Z|Y,P}\{Z|y_m,\widehat{p}_m;\nu_m^p\}\\ &= \frac{\int_{\mathbb{C}}|z|^2\, p_{Y|Z}(y_m|z;\nu^w)\mathcal{N}(z;\widehat{p}_m,\nu_m^p)dz}{\int_{\mathbb{C}} p_{Y|Z}(y_m|z';\nu^w)\mathcal{N}(z';\widehat{p}_m,\nu_m^p)dz'} - |\mathrm{E}_{Z|Y,P}\{Z|y_m,\widehat{p}_m;\nu_m^p\}|^2 \qquad (11)\\ &= \frac{y_m^2}{(1+\nu^w/\nu_m^p)^2} + \frac{|\widehat{p}_m|^2}{(\nu_m^p/\nu^w+1)^2} + \frac{1+\rho_m R_0(\rho_m)}{1/\nu^w+1/\nu_m^p}\\ &\quad - |\mathrm{E}_{Z|Y,P}\{Z|y_m,\widehat{p}_m;\nu_m^p\}|^2, \qquad (12)\end{aligned}$$

where

$$R_0(\rho_m) \triangleq \frac{I_1(\rho_m)}{I_0(\rho_m)} \text{ and } \rho_m \triangleq \frac{2y_m|\widehat{p}_m|}{\nu^w+\nu_m^p}, \tag{13}$$

as shown in Appendix A.

3.2 EM update of the noise variance

Above, the noise variance ν^w was treated as a known parameter. In practice, however, ν^w may be unknown, in which case it is not clear what value to use in (10) and (12). To address this problem, we now describe how ν^w can be learned using an expectation-maximization (EM) [38] procedure. The methodology is similar to that proposed in [39] for the case of a Gaussian $p_{Y|Z}(y_m|\cdot)$, but the details differ due to the form of $p_{Y|Z}(y_m|\cdot)$ in (8).

Choosing $\boldsymbol{x}$ as the hidden data, the ith iteration EM update of the ν^w estimate is [38]

$$\widehat{\nu^w}[i+1] = \arg\max_{\nu^w\geq 0} \mathrm{E}\left\{\ln p(\boldsymbol{y},\boldsymbol{x};\nu^w)\middle|\boldsymbol{y};\widehat{\nu^w}[i]\right\}, \tag{14}$$

where square brackets are used to distinguish EM iterations from GAMP iterations (recall Table 1). After a somewhat lengthy derivation, Appendix B shows that the EM update can be approximated as

$$\widehat{v^w}[i+1] \approx \frac{2}{M} \sum_{m=1}^{M} \left(y_m - |\boldsymbol{a}_m^{\mathsf{H}} \widehat{\boldsymbol{x}}[i]| \right)^2, \tag{15}$$

where $\widehat{\boldsymbol{x}}[i]$ denotes the posterior mean of $\boldsymbol{x}$ under the hypothesis $v^w = \widehat{v^w}[i]$. In practice, we use GAMP's estimate of the posterior mean (i.e., the GAMP output $\widehat{\boldsymbol{x}}(t)$ from Table 1 after the final GAMP iteration $t = T_{\max}$) in place of the true one, because computation of the latter is NP hard in general [40].

3.3 Signal prior distribution

GAMP offers great flexibility with respect to the choice of prior distribution on the signal vector $\boldsymbol{x}$. In this work, we focus on separable priors, which have the form $p(\boldsymbol{x}) = \prod_{n=1}^{N} p_{X_n}(x_n)$ with arbitrary $p_{X_n}(\cdot)$ (recalling (4)), but we note that various forms of non-separable priors can be supported using the "turbo GAMP" formulation proposed in [41] or the "analysis GAMP" formulation proposed in [42].

For separable priors, $p_{X_n}(\cdot)$ should be chosen to reflect whatever form of probabilistic structure is known about coefficient x_n. For example, if $\boldsymbol{x} \in \mathbb{C}^N$ is known to be K-sparse, but nothing is know about the support, then it is typical to choose the Bernoulli-Gaussian (BG) model

$$p_{X_n}(x_n) = (1-\lambda)\delta(x_n) + \lambda \mathcal{N}(x_n; 0, \varphi), \tag{16}$$

with sparsity rate $\lambda = \frac{K}{N}$ and nonzero coefficient variance φ that, if unknown, can be estimated from the observations via [39, eqn. (71)]

$$\varphi = \frac{\|\boldsymbol{y}\|_2^2 - M v^w}{\lambda \|\boldsymbol{A}\|_F^2}, \tag{17}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. For this BG prior, expressions for the thresholding functions $g_{\mathsf{in},n}(\cdot,\cdot)$ and $g'_{\mathsf{in},n}(\cdot,\cdot)$ defined in steps (D5)–(D6) of Table 1 were given in [41]. When the sparsity rate λ in (16) is unknown, it can be learned using the EM-BG procedure described in [39]. In most cases, improved performance is obtained when a Gaussian mixture (GM) pdf is used in place of the Gaussian pdf in (16) [39].

Various extensions of the above are possible. For example, when all coefficients x_n are known to be real valued or positive, the circular-Gaussian pdf in (16) should be replaced by a real-Gaussian or truncated-Gaussian pdf, respectively, or even a truncated-GM [43]. Furthermore, when certain coefficient subsets are known to be more or less sparse than others, a nonuniform sparsity [44] rate λ_n should be used in (16).

Table 2 GAMP steps with variance normalization $\alpha(t)$ and damping parameter $\beta \in (0,1]$.

for $t = 1,2,3,\ldots,T_{\max}$

$$\forall m: \nu_m^p(t) = \beta \textstyle\sum_{n=1}^N |a_{mn}|^2 \nu_n^x(t) + (1-\beta)\nu_m^p(t-1) \quad \text{(S1)}$$

$$\alpha(t) = \tfrac{1}{M} \textstyle\sum_{m=1}^M \nu_m^p(t) \quad \text{(S2)}$$

$$\forall m: \widehat{p}_m(t) = \textstyle\sum_{n=1}^N a_{mn}\widehat{x}_n(t) - \frac{\nu_m^p(t)}{\alpha(t)}\widehat{\underline{s}}_m(t-1) \quad \text{(S3)}$$

$$\forall m: \widehat{\underline{s}}_m(t) = \beta\alpha(t) g_{\mathsf{out},m}(\widehat{p}_m(t), \nu_m^p(t)) + (1-\beta)\widehat{\underline{s}}_m(t-1) \quad \text{(S4)}$$

$$\forall m: \underline{\nu}_m^s(t) = -\beta\alpha(t) g'_{\mathsf{out},m}(\widehat{p}_m(t), \nu_m^p(t)) + (1-\beta)\underline{\nu}_m^s(t-1) \quad \text{(S5)}$$

$$\forall n: \underline{\nu}_n^r(t) = \big(\textstyle\sum_{m=1}^M |a_{mn}|^2 \underline{\nu}_m^s(t)\big)^{-1} \quad \text{(S6)}$$

$$\forall n: \bar{x}_n(t) = \beta\widehat{x}_n(t) + (1-\beta)\bar{x}_n(t-1) \quad \text{(S7)}$$

$$\forall n: \widehat{r}_n(t) = \bar{x}_n(t) + \underline{\nu}_n^r(t) \textstyle\sum_{m=1}^M a_{mn}^* \widehat{\underline{s}}_m(t) \quad \text{(S8)}$$

$$\forall n: \nu_n^x(t+1) = \alpha(t)\underline{\nu}_n^r(t) g'_{\mathsf{in},n}\big(\widehat{r}_n(t), \alpha(t)\underline{\nu}_n^r(t)\big) \quad \text{(S9)}$$

$$\forall n: \widehat{x}_n(t+1) = g_{\mathsf{in},n}\big(\widehat{r}_n(t), \alpha(t)\underline{\nu}_n^r(t)\big) \quad \text{(S10)}$$

end

3.4 GAMP normalization and damping

To increase the numerical robustness of GAMP, it helps to normalize certain internal GAMP variables. To do this, we define $\alpha(t) \triangleq \frac{1}{M}\sum_{m=1}^M \nu_m^p(t)$ (which tends to grow very small with t at high SNR), normalize both $\widehat{s}_m(t)$ and $\nu_m^s(t)$ (which tend to grow very large) by $1/\alpha(t)$, and normalize $\nu_n^r(t)$ (which tends to grow very small) by $\alpha(t)$. The resulting GAMP iterations are shown in Table 2, with normalized quantities denoted by underbars. We note that, under infinite precision, these normalizations would cancel each other out and have no effect.

To reduce the chance of GAMP misconvergence, we find that it helps to "damp" the iterations. Damping helps to slow the algorithm using a stepsize $\beta \in (0,1]$ that is incorporated into GAMP as shown in Table 2. Based on our experiments, the value $\beta = 0.25$ seems to work well for phase retrieval. One consequence of the damping modification is the existence of additional state variables like $\bar{x}_n(t)$. To avoid the need to initialize these variables, we use $\beta = 1$ during the first iteration. We note that the damping modifications described here are the ones included in the public domain GAMPmatlab implementation,[4] which differ slightly from the ones described in [45].

3.5 Avoiding bad local minima

As is well known, the compressive phase retrieval problem is plagued by bad local minima. We now propose methods to initialize and restart PR-GAMP that aims to avoid these local minima. Based on our experience (see Section 4), these methods are much more important for Fourier $\boldsymbol{A}$ than randomized (e.g., i.i.d Gaussian or masked Fourier) $\boldsymbol{A}$.

[4] http://sourceforge.net/projects/gampmatlab/.

3.5.1 GAMP initialization

The GAMP algorithm in Table 1 requires an initialization of the signal coefficient estimates $\{\widehat{x}_n(1)\}_{n=1}^N$, their variances $\{v_n^x(1)\}_{n=1}^N$, and the state variables $\{\widehat{s}_m(0)\}_{m=1}^M$ (which can be interpreted as Lagrange multipliers [45]). As recommended in [27], the standard procedure uses the fixed choices $\widehat{x}_n(1) = \mathrm{E}\{X_n\}$, $v_n^x(1) = \mathrm{var}\{X_n\}$, $\widehat{s}_m(0) = 0$. For phase retrieval, we instead suggest to set each $\widehat{x}_n(1)$ using an independent draw of the random variable X_n and to set $v_n^x(1) = \frac{1}{N}\sum_{k=1}^N |\widehat{x}_k(1) - \mathrm{E}\{X_k\}|^2 \ \forall n$. This initialization, however, only applies to the first EM iteration; for subsequent EM iterations, GAMP should be warm started using the outputs of the previous EM iteration.

3.5.2 EM initialization

For the EM algorithm described in Section 3.2, we must choose the initial noise-variance estimate $\widehat{v^w}[0]$. Even when accurate knowledge of v^w is available, we find that setting $\widehat{v^w}[0]$ at a large value helps to avoid bad local minima. In particular, our empirical experience leads us to suggest setting $\widehat{v^w}[0]$ in correspondence with an initial SNR estimate of 10, i.e., $\widehat{v^w}[0] = \frac{\|\boldsymbol{y}\|_2^2}{M(\mathsf{SNR}_{\mathsf{init}}+1)}$ with $\mathsf{SNR}_{\mathsf{init}} = 10$.

3.5.3 Multiple restarts

To further facilitate the avoidance of bad local minima, we propose to run multiple attempts of EM-GAMP, each using a different random GAMP initialization (constructed as above). The attempt leading to the lowest normalized residual ($\mathsf{NR} \triangleq \|\boldsymbol{y} - |\boldsymbol{A}\widehat{\boldsymbol{x}}|\|_2^2 / \|\boldsymbol{y}\|_2^2$) is then selected as the algorithm output. The efficacy of multiple attempts is numerically investigated in Section 4.

Furthermore, to avoid unnecessary restarts, we allow the algorithm to be stopped as soon as the NR drops below a user-defined stopping tolerance of $\mathsf{NR}_{\mathsf{stop}}$. When the true SNR is known, we suggest setting $\mathsf{NR}_{\mathsf{stop}}\mathsf{dB} = -(\mathsf{SNR}_{\mathsf{true}}\mathsf{dB} + 2)$.

3.5.4 Algorithm summary

The PR-GAMP algorithm is summarized in Table 3, where $A_{\max}$ controls the number of attempts, $\mathsf{SNR}_{\mathsf{init}}$ controls the initial SNR, and $\mathsf{NR}_{\mathsf{stop}}$ controls the stopping tolerance.

Table 3 The proposed PR-GAMP algorithm with $A_{\max}$ attempts, SNR initialization $\mathsf{SNR}_{\mathsf{init}}$, and stopping residual $\mathsf{NR}_{\mathsf{stop}}$.

input $\boldsymbol{y}, \boldsymbol{A}, \{p_{X_n}(\cdot)\}_{n=1}^N, \mathsf{SNR}_{\mathsf{init}}, \mathsf{NR}_{\mathsf{stop}}$

$\widehat{v^w}[0] = \dfrac{\|\boldsymbol{y}\|_2^2}{M(\mathsf{SNR}_{\mathsf{init}}+1)}$

$\forall m : \widehat{s}_m[0] = 0$

$\mathsf{NR}_{\mathsf{best}} = \infty$

for $a = 1, 2, 3, \ldots, A_{\max}$,

 draw random $\widehat{\boldsymbol{x}}[0]$

 $\forall n : v_n^x[0] = \|\widehat{\boldsymbol{x}}[0]\|_2^2 / N$

 for $i = 1, 2, 3, \ldots, I_{\max}$

 $\big(\widehat{\boldsymbol{x}}[i], \widehat{\boldsymbol{v}}^x[i], \widehat{\boldsymbol{s}}[i]\big) = \mathsf{GAMP}\big(\boldsymbol{A}, \{p_{X_n}(\cdot)\}_{n=1}^N, \{p_{Y|Z}(y_m|\cdot; \widehat{v^w}[i-1])\}_{m=1}^M, \widehat{\boldsymbol{x}}[i-1], \widehat{\boldsymbol{v}}^x[i-1], \widehat{\boldsymbol{s}}[i-1]\big)$

 $\widehat{v^w}[i] = \frac{2}{M}\|\boldsymbol{y} - |\boldsymbol{A}\widehat{\boldsymbol{x}}[i]|\|_2^2$

 end

 $\mathsf{NR} = \|\boldsymbol{y} - |\boldsymbol{A}\widehat{\boldsymbol{x}}[i]|\|_2^2 / \|\boldsymbol{y}\|_2^2$

 if $\mathsf{NR} < \mathsf{NR}_{\mathsf{best}}$

 $\widehat{\boldsymbol{x}}_{\mathsf{best}} = \widehat{\boldsymbol{x}}[I_{\max}]$

 $\mathsf{NR}_{\mathsf{best}} = \mathsf{NR}$

 end

 if $\mathsf{NR} < \mathsf{NR}_{\mathsf{stop}}$

 stop

 end

end

output $\widehat{\boldsymbol{x}}_{\mathsf{best}}$

4 Numerical Results

In this section, we numerically investigate the performance of PR-GAMP[5] under various scenarios and in comparison to several existing algorithms: Compressive Phase Retrieval via Lifting (CPRL) [22], GrEedy Sparse PhAse Retrieval (GESPAR) from [25], and the sparse-Fienup technique from [21], As a benchmark, we also compare to "phase oracle" (PO) GAMP, i.e., GAMP operating on the magnitude-and-phase measurements $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ rather than on the intensity measurements $\boldsymbol{y} = |\boldsymbol{u}|$.

Unless otherwise noted, we generated random realizations of the true signal vector $\boldsymbol{x}$ as K-sparse length N with support chosen uniformly at random and with nonzero coefficients drawn i.i.d zero-mean circular-Gaussian. Then, for a given matrix $\boldsymbol{A}$, we generated M noisy intensity measurements $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}|$, where $\boldsymbol{w}$ was i.i.d circular-Gaussian with variance selected to achieve a target signal-to-noise ratio of $\mathsf{SNR} \triangleq \|\boldsymbol{A}\boldsymbol{x}\|_2^2 / \mathrm{E}\{\|\boldsymbol{w}\|_2^2\}$. Finally, each algorithm computed an estimate $\widehat{\boldsymbol{x}}$ from $(\boldsymbol{y}, \boldsymbol{A})$ in an attempt to best match $\boldsymbol{x}$ up to a tolerated ambiguity. For $\boldsymbol{A}$ with i.i.d

[5] PR-GAMP is part of the GAMPmatlab package at http://sourceforge.net/projects/gampmatlab/.

random entries, we tolerate only a phase rotation on $\widehat{\boldsymbol{x}}$, while for Fourier $\boldsymbol{A}$ and real-valued $\boldsymbol{x}$, we tolerate a flip, circular shift, and phase rotation on $\widehat{\boldsymbol{x}}$. Performance was then assessed using normalized mean-squared error on the disambiguated estimate:

$$\mathsf{NMSE}(\widehat{\boldsymbol{x}}) \triangleq \min_{\boldsymbol{\Theta}} \frac{\|\boldsymbol{x} - \mathsf{disambig}(\widehat{\boldsymbol{x}}, \boldsymbol{\Theta})\|_2^2}{\|\boldsymbol{x}\|_2^2}, \tag{18}$$

where $\boldsymbol{\Theta}$ are the ambiguity parameters. When computing empirical phase-transition curves, we defined a "successful" recovery as one that produced $\mathsf{NMSE} < 10^{-6}$.

4.1 Empirical phase transitions: i.i.d Gaussian A

First we investigated the phase-transition performance of PR-GAMP with i.i.d circular-Gaussian sensing matrices $\boldsymbol{A}$. Figure 2 plots the empirical success rate (averaged over 100 independent problem realizations) as a function of signal sparsity K and measurement length M for a fixed signal length of $N = 512$. Here we used $\mathsf{SNR} = 100$ dB, which makes the observations essentially "noiseless", and we allowed PR-GAMP up to 10 attempts from random initializations (i.e., $A_{\max} = 10$ in Table 3). The figure shows a "phase-transition" behavior that separates the (K, M) plane into two regions: perfect recovery in the top-left and failure in the bottom-right. Moreover, the figure shows that, for $K \ll N$, approximately $M \geq 2K\log_2(N/K)$ measurements suffice for PR-GAMP.

To see how well (versus how often) PR-GAMP recovers the signal, we plot the median NMSE over the same problem realizations in Fig. 3. There we see that the signal estimates are extremely accurate throughout the region above the phase transition.

To investigate the effect of number of attempts $A_{\max}$, we extracted the 50%-success contour (i.e., the phase-transition curve) from Fig. 2 and plotted it in Fig. 4, along with the corresponding contours obtained under different choices of $A_{\max}$. Figure 4 shows that in the case of i.i.d $\boldsymbol{A}$, there is relatively little to gain from multiple restarts from random realizations. With Fourier $\boldsymbol{A}$, however, we will see in the sequel that multiple restarts are indeed important.

Figure 4 also plots the phase-transition curve of phase-oracle (PO)-GAMP calculated from the same problem realizations. Comparing the PO-GAMP phase transition to that of PR-GAMP, we conclude that PR-GAMP requires approximately $4\times$ the number of measurements as PO-GAMP, regardless of sparsity rate K. Remarkably, this "$4\times$" rule generalizes what is known about the recovery of *non*-sparse signals in $\mathbb{C}^N$, where the ratio of (necessary and sufficient) magnitude-only to magnitude-and-phase measurements is also 4 (as $N \to \infty$) [13, 14].

Overall, Figures 2–4 demonstrate that PR-GAMP is indeed capable of *compressive* phase retrieval, i.e., successful $\mathbb{C}^N$-signal recovery from $M \ll 4N$ intensity measurements, when the signal is sufficiently sparse. Moreover, to our knowledge,

these phase transitions are far better than those of any other algorithm reported in the literature.

4.2 Robustness to noise

We now demonstrate the robustness of PR-GAMP to nontrivial levels of additive white circular-Gaussian noise $\boldsymbol{w}$ in the M intensity measurements $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}|$. As before, we use an $N = 512$-length K-sparse signal with an i.i.d Gaussian $\boldsymbol{A}$, but now we focus on the case $(K, M) = (4, 256)$, which is on the good side of the phase transition in Fig. 2. Figure 5 shows median NMSE performance over 200 independent

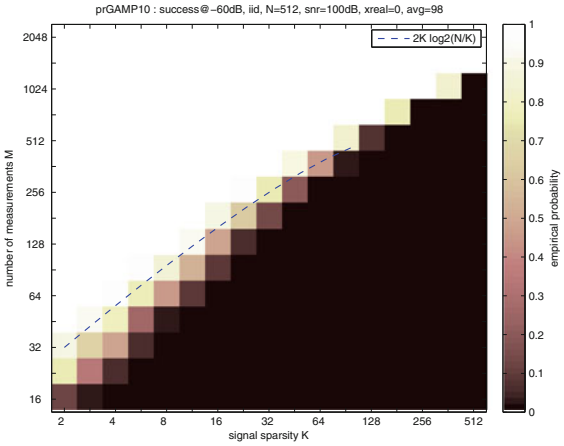


Fig. 2 Empirical probability of successful PR-GAMP recovery of an $N = 512$-length signal versus signal sparsity K and number of intensity measurements M, using i.i.d Gaussian $\boldsymbol{A}$ at SNR $=$ 100 dB. Here, PR-GAMP was allowed up to 10 attempts from different random initializations.

problem realizations as a function of $\mathsf{SNR} \triangleq \|\boldsymbol{A}\boldsymbol{x}\|_2^2 / \|\boldsymbol{w}\|_2^2$. At larger values of SNR (i.e., SNR ≥ 30 dB), there we see that throughout the tested SNR range, PR-GAMP performs only about 3 dB worse than PO-GAMP. The existence of a 3-dB gap can be explained by the fact that PO-GAMP is able to average the noise over twice as many real-valued measurements as PR-GAMP (i.e., $\{\mathrm{Re}\{u_m\}, \mathrm{Im}\{u_m\}\}_{m=1}^M$ versus $\{|u_m|\}_{m=1}^M$).

4.3 Comparison to CPRL

In this section, we present compare PR-GAMP to the state-of-the-art convex-relaxation approach to compressive phase retrieval, CPRL [22]. To implement CPRL, we used the authors' CVX-based matlab code[6] under default algorithmic

[6] http://users.isy.liu.se/rt/ohlsson/code/CPRL.zip.

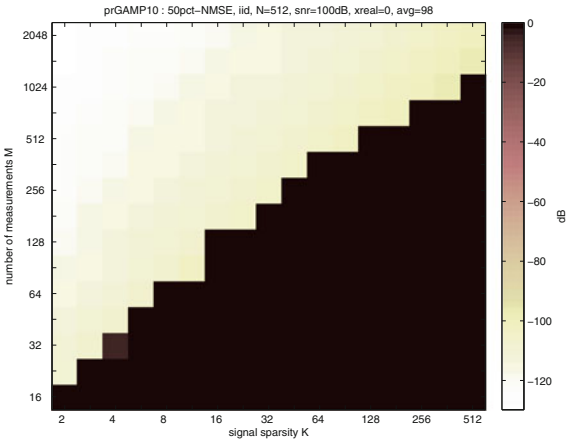


Fig. 3 Median NMSE for PR-GAMP recovery of an $N = 512$-length signal versus signal sparsity K and number of intensity measurements M, using i.i.d Gaussian $\boldsymbol{A}$ at SNR = 100 dB. Here, PR-GAMP was allowed up to 10 attempts from different random initializations.

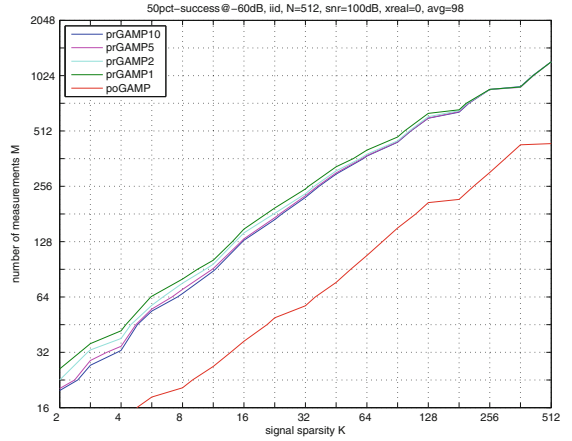


Fig. 4 50%-success contours for PR-GAMP and phase-oracle GAMP recovery of an $N = 512$-length signal versus signal sparsity K and number of intensity measurements M, using i.i.d Gaussian $\boldsymbol{A}$ at SNR = 100 dB. PR-GAMP-$A_{\max}$ denotes PR-GAMP under a maximum of $A_{\max}$ attempts.

settings. We also tried the authors' ADMM implementation, but found that it gave significantly worse performance. As before, we examine the recovery of a K-sparse signal in $\mathbb{C}^N$ from M intensity measurements $\boldsymbol{y} = |\boldsymbol{Ax}+\boldsymbol{w}|$, but now we use $\boldsymbol{A} = \boldsymbol{\Phi F}$ with i.i.d circular-Gaussian $\boldsymbol{\Phi}$ and discrete Fourier transform (DFT) $\boldsymbol{F}$, to be consistent with the setup assumed in [22].

Table 4 shows empirical success[7] rate and runtime (on a standard personal computer) for a problem with sparsity $K = 1$, signal lengths $N \in \{32, 48, 64\}$, and compressive measurement lengths $M \in \{20, 30, 40\}$. The table shows that, over 100

[7] Since CPRL rarely gave NMSE $< 10^{-6}$, we reduced the definition of "success" to NMSE $< 10^{-4}$ for this subsection only.

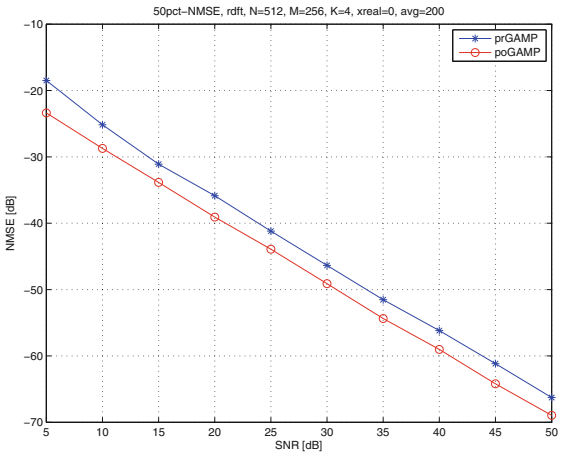


Fig. 5 Median NMSE for PR-GAMP and phase-oracle GAMP recovery of an $N = 512$-length $K = 4$-sparse signal versus SNR, from $M = 256$ measurements and i.i.d Gaussian $\boldsymbol{A}$.

problem realizations, both algorithms were 100% successful in recovering the signal at all tested combinations of (M,N). But the table also shows that CPRL's runtime increases rapidly with the signal dimensions, whereas that of PR-GAMP remains orders of magnitude smaller and independent of (M,N) over the tested range.[8]

Table 5 repeats the experiment carried out in Table 4, but at the sparsity $K = 2$. For this more difficult problem, the table shows that CPRL is much less successful at recovering the signal than PR-GAMP. Meanwhile, the runtimes reported in Table 5 again show CPRL complexity scaling rapidly with the problem dimension, whereas GAMP complexity stays orders-of-magnitude smaller and constant over the tested problem dimensions. In fact, the comparisons conducted in this section were restricted to very small problem dimensions precisely due to the poor complexity scaling of CPRL.

Table 4 Empirical success rate and median runtime over 100 problem realizations for several combinations of signal length N, measurement length M, and signal sparsity $K = 1$.

	$(M,N) = (20,32)$	$(M,N) = (30,48)$	$(M,N) = (40,64)$
CPRL	1.00 (3.4 sec)	1.00 (37 sec)	1.00 (434 sec)
PR-GAMP	1.00 (0.22 sec)	1.00 (0.20 sec)	1.00 (0.18 sec)

[8] Although the complexity of GAMP is known to scale as $O(MN)$ for this type of $\boldsymbol{A}$, the values of M and N in Table 4 are too small for this scaling law to manifest.

Table 5 Empirical success rate and median runtime over 100 problem realizations for several combinations of signal length N, measurement length M, and signal sparsity $K = 2$.

	$(M,N) = (20,32)$	$(M,N) = (30,48)$	$(M,N) = (40,64)$
CPRL	0.55 (4.1 sec)	0.65 (42 sec)	0.66 (496 sec)
PR-GAMP	0.99 (0.28 sec)	0.99 (0.24 sec)	1.00 (0.22 sec)

4.4 Comparison to sparse-Fienup and GESPAR: Fourier A

In this section, we compare PR-GAMP to the sparse-Fienup [21] and GESPAR[9] [25] algorithms, which requires[10] us to restrict our attention to Fourier-based $\boldsymbol{A}$ and real-valued sparse vectors $\boldsymbol{x}$. For the experiments below, we generated realizations of $\boldsymbol{x}$ the same was as above, but with the nonzero elements drawn from a real-Gaussian distribution. Also, we used $ITER = 6400$ in GESPAR as recommended by the authors in [25], and we allowed sparse-Fienup 1000 attempts from random initializations.

We first consider 2D Fourier $\boldsymbol{A}$, which is especially important for imaging applications. In particular, we repeat an experiment from [25], where the measurement and signal lengths were fixed at $M = N$ and the signal sparsity K was varied. For $N = 1024$, Fig. 6 shows the empirical success rate (over 200 realizations) for PR-GAMP, GESPAR, and sparse-Fienup. Meanwhile, Fig. 7 shows the corresponding median runtime for each algorithm, where all algorithms leveraged fast Fourier transform (FFT) implementations of $\boldsymbol{A}$. From Fig. 6, we can see that PR-GAMP yields a significantly better phase transition than GESPAR and sparse-Fienup. Meanwhile, from Fig. 7 we see that for the challenging case of $K \geq 40$, PR-GAMP-10 has uniformly better runtime *and* success rate than GESPAR and sparse-Fienup.

Next we consider 1D Fourier $\boldsymbol{A}$. Again, we repeat an experiment from [25], where the measurement and signal lengths were fixed at $M = 2N$ and the signal sparsity K was varied. For $N = 1024$, Fig. 8 shows the empirical success rate (over 200 realizations) for PR-GAMP, GESPAR, and sparse-Fienup, and Fig. 7 shows the corresponding median runtimes. From Fig. 8, we can see that PR-GAMP yields a significantly better phase transition than GESPAR and sparse-Fienup. Meanwhile, from Fig. 9 we see that for the challenging case of $K \geq 40$, PR-GAMP-20 has uniformly better runtime *and* success rate than GESPAR and sparse-Fienup.

Comparing the results in this section to those in Section 4.1, we conclude that compressive phase retrieval is much more difficult with Fourier matrices $\boldsymbol{A}$ than with i.i.d matrices $\boldsymbol{A}$. This phenomenon has been noticed by other authors as well,

[9] For GESPAR, we used the November 2013 version of the Matlab code provided by the authors at https://sites.google.com/site/yoavshechtman/resources/software.

[10] The sparse-Fienup from [21] requires $\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}$ to be a (scaled) identity matrix. Although GESPAR can in principle handle generic $\boldsymbol{A}$, the implementation provided by the authors is based on 1D and 2D Fourier $\boldsymbol{A}$ and is not easily modified.

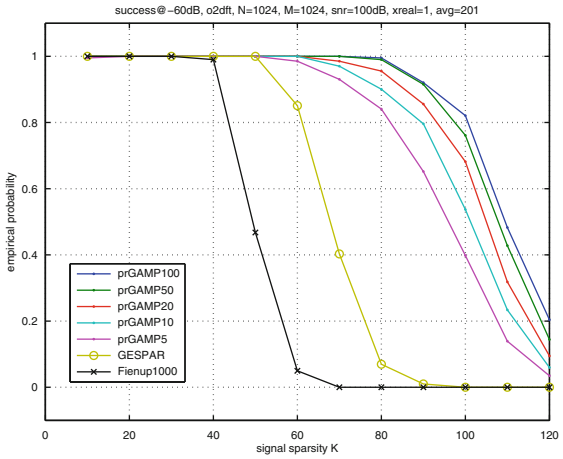


Fig. 6 Empirical success rate versus sparsity K in the recovery of an $N = 1024$-length real-valued signal from $M = 1024$ 2D-Fourier intensities at $\mathsf{SNR} = 100$dB. PR-GAMP-A denotes PR-GAMP under a maximum of A attempts.

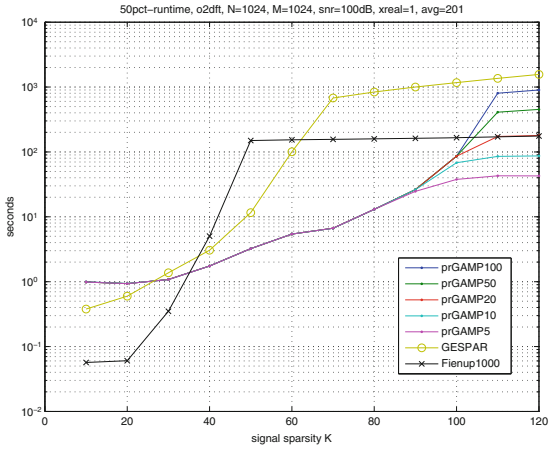


Fig. 7 Median runtime versus sparsity K in the recovery of an $N = 1024$-length real-valued signal from $M = 1024$ 2D-Fourier intensities at $\mathsf{SNR} = 100$dB. PR-GAMP-A denotes PR-GAMP under a maximum of A attempts.

which has led to proposals for randomized Fourier-based phase retrieval (e.g., using binary masks [46]). Also, we notice that the use of multiple restarts in PR-GAMP is much more important with Fourier $\boldsymbol{A}$ than it is with i.i.d $\boldsymbol{A}$.

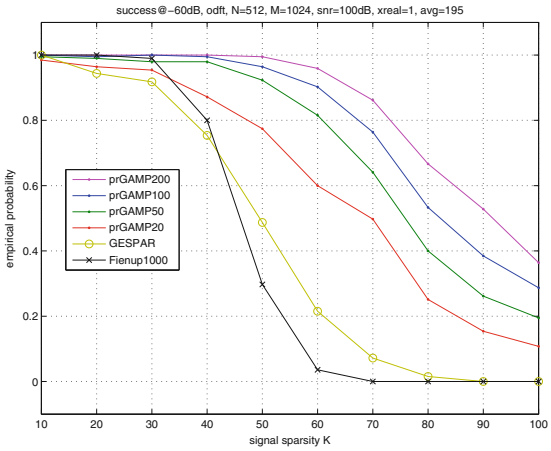


Fig. 8 Empirical success rate versus sparsity K in the recovery of an $N = 512$-length real-valued signal from $M = 1024$ 1D-Fourier intensities at $\mathsf{SNR} = 100$dB. PR-GAMP-A denotes PR-GAMP under a maximum of A attempts.

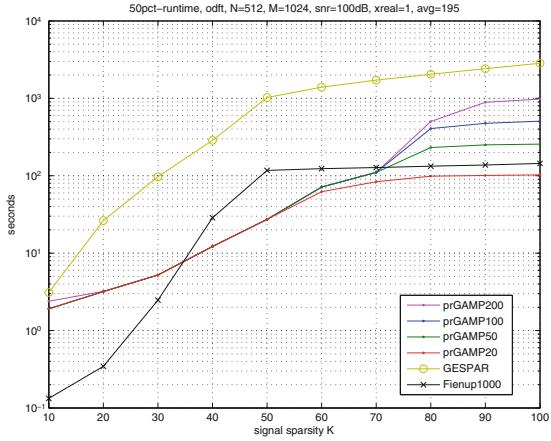


Fig. 9 Median runtime versus sparsity K in the recovery of an $N = 512$-length real-valued signal from $M = 1024$ 1D-Fourier intensities at $\mathsf{SNR} = 100$dB. PR-GAMP-A denotes PR-GAMP under a maximum of A attempts.

4.5 Practical image recovery with masked Fourier A

Finally, we demonstrate practical image recovery from compressed intensity measurements. For this experiment, the signal $\boldsymbol{x}$ was the $N = 65536$-pixel grayscale image shown on the left of Fig. 10, which has a sparsity of $K = 6678$. Since this image is real and nonnegative, we ran PR-GAMP with a nonnegative-real-BG prior [43], as opposed to the BG prior (16) used in previous experiments.

For the first set of experiments, we used a "masked" Fourier transformation $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ of the form

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{J}_1\boldsymbol{F}\boldsymbol{D}_1 \\ \boldsymbol{J}_2\boldsymbol{F}\boldsymbol{D}_2 \\ \boldsymbol{J}_3\boldsymbol{F}\boldsymbol{D}_3 \\ \boldsymbol{J}_4\boldsymbol{F}\boldsymbol{D}_4 \end{bmatrix}, \tag{19}$$

where $\boldsymbol{F}$ was a 2D DFT matrix of size $N \times N$, $\boldsymbol{D}_i$ were diagonal "masking" matrices of size $N \times N$ with diagonal entries drawn uniformly at random from $\{0,1\}$, and $\boldsymbol{J}_i$ were "selection" matrices of size $\frac{M}{4} \times N$ constructed from rows of the identity matrix drawn uniformly at random. The matrices $\boldsymbol{D}_i$ and $\boldsymbol{J}_i$ help to "randomize" the DFT, and they circumvent unicity issues such as shift and flip ambiguities. For phase retrieval, the use of image masks was discussed in [46]. Note that because $\boldsymbol{D}_i$ and $\boldsymbol{J}_i$ are sparse and $\boldsymbol{F}$ has a fast FFT-based implementation, the overall matrix $\boldsymbol{A}$ has a fast implementation.

To eliminate the need for the expensive matrix multiplications with the element-wise-squared versions of $\boldsymbol{A}$ and $\boldsymbol{A}^{\mathsf{H}}$, as specified in steps (S1) and (S6) of Table 2, GAMP was run in "uniform variance" mode, meaning that $\{\nu^p_m(t)\}_{m=1}^M$ were approximated by $\nu^p(t) \triangleq \frac{1}{M}\sum_{m'=1}^M \nu^p_{m'}(t)$; similar was done with $\{\underline{\nu}^s_m(t)\}_{m=1}^M$, $\{\nu^r_n(t)\}_{n=1}^N$, and $\{\nu^x_n(t)\}_{n=1}^N$. The result is that lines (S1)–(S2) in Table 2 become $\nu^p(t) = \beta\|\boldsymbol{A}\|_F^2\nu^x(t)/M + (1-\beta)\nu^p(t-1) = \alpha(t)$ and line (S6) becomes $\underline{\nu}^r(t) = \left(\|\boldsymbol{A}\|_F^2\underline{\nu}^s(t)/N\right)^{-1}$.

As before, the observations took the form $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}|$, but now the noise variance was adjusted to yield a nontrivial $\mathsf{SNR} = 30$ dB. To demonstrate *compressive* phase retrieval, only $M = N = 65536$ intensity measurements were used. Running PR-GAMP on 100 problem realizations (each with different random $\boldsymbol{A}$ and $\boldsymbol{w}$, and allowing at most 10 restarts per realization), a *99% success rate* was observed, where for this noisy problem "success" was defined as $\mathsf{NMSE} < \mathsf{SNR}^{-1} = -30$ dB. Furthermore, PR-GAMP's median runtime over these realizations was only *8.4 seconds*. The right subplot in Fig. 10 shows a typical PR-GAMP recovery.

For the second set of experiments, we "blurred" the masked Fourier outputs to further randomize $\boldsymbol{A}$, which allowed us to achieve similar recovery performance using *half* the intensity measurements, i.e., $M = N/2 = 32768$. In particular, we used a linear transformation $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ of the form

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{B}_1\boldsymbol{F}\boldsymbol{D}_1 \\ \boldsymbol{B}_2\boldsymbol{F}\boldsymbol{D}_2 \end{bmatrix}, \tag{20}$$

where $\boldsymbol{F}$ and $\boldsymbol{D}_i$ were as before[11] and $\boldsymbol{B}_i$ were banded[12] matrices of size $\frac{M}{2} \times N$ with 10 nonzero i.i.d circular-Gaussian entries per column. The use of blurring to

[11] Here, since we used only two masks, we ensured invertibility by constructing the diagonal of $\boldsymbol{D}_1$ using exactly $N/2$ unit-valued entries positioned uniformly at random and constructing the diagonal of $\boldsymbol{D}_2$ as its complement, so that $\boldsymbol{D}_1 + \boldsymbol{D}_2 = \boldsymbol{I}$.

[12] Since each $\boldsymbol{B}_i$ was a wide matrix, its nonzero band was wrapped from bottom to top when necessary.

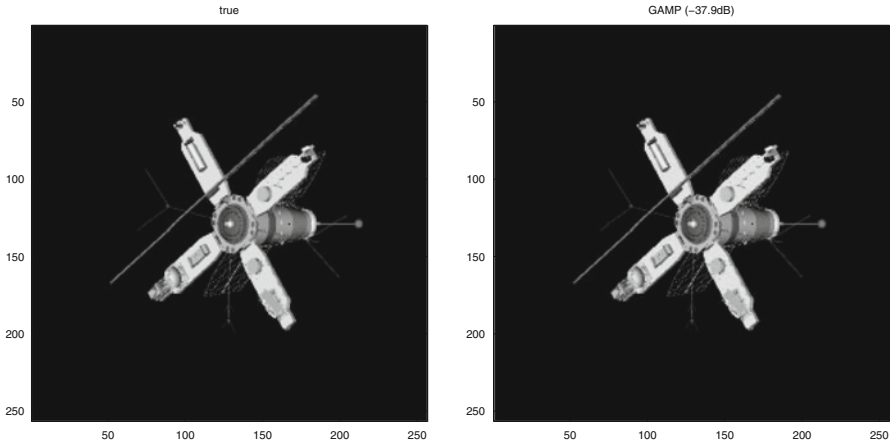


Fig. 10 Original image (left) and a typical PR-GAMP recovery (right) from $M = N$ masked Fourier intensity measurements at $\mathsf{SNR} = 30$ dB, which took 2.2 seconds.

enhance phase retrieval was discussed in [47]. As with (19), the $\boldsymbol{A}$ in (20) has a fast implementation. Running PR-GAMP as before on 100 problem realizations at $\mathsf{SNR} = 30$ dB, a *99% success rate* was observed with a median runtime of only *12.6 seconds*.

To our knowledge, no existing algorithms are able to perform compressive phase retrieval on images of this size and sparsity with such high speed and accuracy. To put our results in perspective, we recall the image recovery experiment in [25], which shows an example of GESPAR taking 80 seconds to recover a $K = 15$-sparse image whose support was effectively constrained to $N = 225$ pixels from $M = 38025$ 2D Fourier intensity measurements. In contrast, Fig. 10 shows PR-GAMP taking 2.2 seconds to recover a $K = 6678$-sparse image whose support was constrained to $N = 65536$ pixels from $M = 65536$ masked 2D Fourier intensity measurements.

5 Conclusions

In this chapter, we proposed a novel approach to compressive phase retrieval based on the generalized approximate message passing (GAMP) algorithm. Numerical results showed that the proposed PR-GAMP algorithm has excellent phase-transition behavior, noise robustness, and runtime. In particular, for successful recovery of synthetic K-sparse signals PR-GAMP requires approximately 4 times the number of measurements as phase-oracle GAMP and achieves NMSE that is only 3 dB worse than phase-oracle GAMP. For recovery of a real-valued 65532-pixel image from 32768 pre-masked and post-blurred Fourier intensities, PR-GAMP was successful 99% of the time with a median runtime of only 12.6 seconds. Comparison

to the recently proposed CPRL, sparse-Fienup, and GESPAR algorithms revealed PR-GAMP's superior phase transitions and orders of magnitude faster runtimes at large K.

Appendix A: Output Thresholding Rules

In this appendix, we derive expressions (10) and (12) that are used to compute the functions $g_{\text{out},m}$ and $g'_{\text{out},m}$ defined in lines (D2) and (D3) of Table 1.

To facilitate the derivations in this appendix,[13] we first rewrite $p_{Y|Z}(y|z)$ in a form different from (8). In particular, recalling that—under our AWGN assumption—the noisy transform outputs $u = z + w$ are conditionally distributed as $p(u|z) = \mathcal{N}(u; z, \nu^w)$, we first transform $u = ye^{j\theta}$ from rectangular to polar coordinates to obtain

$$p(y, \theta|z) = 1_{y\geq 0} 1_{\theta\in[0,2\pi)} \mathcal{N}(ye^{j\theta}; z, \nu^w)\, y, \tag{21}$$

where y is the Jacobian of the transformation, and then integrate out the unobserved phase θ to obtain

$$p_{Y|Z}(y|z) = 1_{y\geq 0}\, y \int_0^{2\pi} \mathcal{N}(ye^{j\theta}; z, \nu^w)\, d\theta. \tag{22}$$

We begin by deriving the integration constant

$$\begin{aligned} C(y, \nu^w, \widehat{p}, \nu^p) &\triangleq \int_{\mathbb{C}} p_{Y|Z}(y|z)\, \mathcal{N}(z; \widehat{p}, \nu^p) dz \\ &= y\, 1_{y\geq 0} \int_0^{2\pi} \int_{\mathbb{C}} \mathcal{N}(ye^{j\theta}; z, \nu^w) \mathcal{N}(z; \widehat{p}, \nu^p) dz d\theta \qquad (23) \\ &= y\, 1_{y\geq 0} \int_0^{2\pi} \mathcal{N}(ye^{j\theta}; \widehat{p}, \nu^w + \nu^p) d\theta, \qquad (24) \end{aligned}$$

where we used the Gaussian-pdf multiplication rule[14] in (24). Noting the similarity between (24) and (22), the equivalence between (22) and (8) implies that

$$C(y, \nu^w, \widehat{p}, \nu^p) = \frac{2y}{\nu^w + \nu^p} \exp\left(-\frac{y^2 + |\widehat{p}|^2}{\nu^w + \nu^p} \right) I_0\left(\frac{2y|\widehat{p}|}{\nu^w + \nu^p} \right) 1_{y\geq 0}. \tag{25}$$

In the sequel, we make the practical assumption that $y > 0$, allowing us to drop the indicator "$1_{y\geq 0}$" and invert C.

Next, we derive the conditional mean

$$\mathrm{E}_{Z|Y,P}\{Z|y, \widehat{p}; \nu^p\} = C(y, \nu^w, \widehat{p}, \nu^p)^{-1} \int_{\mathbb{C}} z\, p_{Y|Z}(y|z; \nu^w) \mathcal{N}(z; \widehat{p}, \nu^p) dz. \tag{26}$$

[13] For notational brevity, the subscript "m" is omitted throughout this appendix.

[14] $\mathcal{N}(z; a, A)\mathcal{N}(z; b, B) = \mathcal{N}\left(z; \frac{\frac{a}{A}+\frac{b}{B}}{\frac{1}{A}+\frac{1}{B}}, \frac{1}{\frac{1}{A}+\frac{1}{B}}\right) \mathcal{N}(a; b, A+B)$.

Plugging (22) into (26) and applying the Gaussian-pdf multiplication rule,

$$\begin{aligned}
&\mathrm{E}_{Z|Y,P}\{Z|y,\widehat{p};\nu^p\} \\
&= C^{-1} y \int_0^{2\pi} \int_{\mathbb{C}} z\,\mathcal{N}(z;ye^{j\theta},\nu^w)\mathcal{N}(z;\widehat{p},\nu^p)dzd\theta && (27)\\
&= C^{-1} y \int_0^{2\pi} \int_{\mathbb{C}} z\,\mathcal{N}\left(z;\frac{ye^{j\theta}/\nu^w+\widehat{p}/\nu^p}{1/\nu^w+1/\nu^p},\frac{1}{1/\nu^w+1/\nu^p}\right) \\
&\quad\times\mathcal{N}(ye^{j\theta};\widehat{p},\nu^w+\nu^p)dzd\theta && (28)\\
&= C^{-1} y \int_0^{2\pi} \frac{ye^{j\theta}/\nu^w+\widehat{p}/\nu^p}{1/\nu^w+1/\nu^p}\mathcal{N}(ye^{j\theta};\widehat{p},\nu^w+\nu^p)d\theta && (29)\\
&= \frac{y/\nu^w}{1/\nu^w+1/\nu^p} C^{-1} y \int_0^{2\pi} e^{j\theta}\mathcal{N}(ye^{j\theta};\widehat{p},\nu^w+\nu^p)d\theta \\
&\quad+\frac{\widehat{p}/\nu^p}{1/\nu^w+1/\nu^p} C^{-1} y \int_0^{2\pi} \mathcal{N}(ye^{j\theta};\widehat{p},\nu^w+\nu^p)d\theta && (30)\\
&= \frac{y}{\nu^w/\nu^p+1} C^{-1} y \int_0^{2\pi} e^{j\theta}\mathcal{N}(ye^{j\theta};\widehat{p},\nu^w+\nu^p)d\theta+\frac{\widehat{p}}{\nu^p/\nu^w+1}. && (31)
\end{aligned}$$

Expanding the $\mathcal{N}$ term, the integral in (31) becomes

$$\begin{aligned}
&\int_0^{2\pi} e^{j\theta}\mathcal{N}(ye^{j\theta};\widehat{p},\nu^w+\nu^p)d\theta \\
&= \frac{1}{\pi(\nu^w+\nu^p)}\exp\left(-\frac{y^2+|\widehat{p}|^2}{\nu^w+\nu^p}\right)\int_0^{2\pi} e^{j\theta}\exp\left(\frac{2y|\widehat{p}|}{\nu^w+\nu^p}\cos(\theta-\psi)\right)d\theta && (32)\\
&= \frac{1}{\pi(\nu^w+\nu^p)}\exp\left(-\frac{y^2+|\widehat{p}|^2}{\nu^w+\nu^p}\right)e^{j\psi}\int_0^{2\pi} e^{j\theta'}\exp\left(\frac{2y|\widehat{p}|}{\nu^w+\nu^p}\cos(\theta')\right)d\theta' && (33)\\
&= \frac{2e^{j\psi}}{\nu^w+\nu^p}\exp\left(-\frac{y^2+|\widehat{p}|^2}{\nu^w+\nu^p}\right)I_1\left(\frac{2y|\widehat{p}|}{\nu^w+\nu^p}\right), && (34)
\end{aligned}$$

where ψ denotes the phase of $\widehat{p}$, and where the integral in (33) was resolved using the expression in [48, 9.6.19]. Plugging (34) into (31) gives

$$\mathrm{E}_{Z|Y,P}\{Z|y,\widehat{p};\nu^p\} = \frac{\widehat{p}}{\nu^p/\nu^w+1}+\frac{ye^{j\psi}}{\nu^w/\nu^p+1}\frac{I_1\left(\frac{2y|\widehat{p}|}{\nu^w+\nu^p}\right)}{I_0\left(\frac{2y|\widehat{p}|}{\nu^w+\nu^p}\right)}, \tag{35}$$

which agrees with (10).

Finally, we derive the conditional covariance

$$\begin{aligned}
\mathrm{var}_{Z|Y,P}\{Z|y,\widehat{p};\nu^p\} = C(y,\nu^w,\widehat{p},\nu^p)^{-1}\int_{\mathbb{C}}|z|^2\,p_{Y|Z}(y|z;\nu^w)\mathcal{N}(z;\widehat{p},\nu^p)dz \\
-|\mathrm{E}_{Z|Y,P}\{Z|y,\widehat{p};\nu^p\}|^2. && (36)
\end{aligned}$$

Focusing on the first term in (36), if we plug in (22) and apply the Gaussian-pdf multiplication rule, we get

$$
\begin{aligned}
&C(y,\nu^w,\widehat{p},\nu^p)^{-1}\int_{\mathbb{C}}|z|^2\,p_{Y|Z}(y|z;\nu^w)\mathcal{N}(z;\widehat{p},\nu^p)dz \\
&=C^{-1}y\int_0^{2\pi}\int_{\mathbb{C}}|z|^2\,\mathcal{N}\left(z;\frac{ye^{j\theta}/\nu^w+\widehat{p}/\nu^p}{1/\nu^w+1/\nu^p},\frac{1}{1/\nu^w+1/\nu^p}\right)dz \\
&\quad\times\mathcal{N}(ye^{j\theta};\widehat{p},\nu^w+\nu^p)d\theta \qquad (37)\\
&=C^{-1}y\int_0^{2\pi}\left(\left|\frac{ye^{j\theta}/\nu^w+\widehat{p}/\nu^p}{1/\nu^w+1/\nu^p}\right|^2+\frac{1}{1/\nu^w+1/\nu^p}\right)\mathcal{N}(ye^{j\theta};\widehat{p},\nu^w+\nu^p)d\theta \qquad (38)\\
&=C^{-1}y\int_0^{2\pi}\frac{|y|^2/(\nu^w)^2+|\widehat{p}|^2/(\nu^p)^2+2y|\widehat{p}|/(\nu^w\nu^p)\mathrm{Re}\{e^{j(\theta-\psi)}\}}{(1/\nu^w+1/\nu^p)^2} \\
&\quad\times\mathcal{N}(ye^{j\theta};\widehat{p},\nu^w+\nu^p)d\theta+\frac{1}{1/\nu^w+1/\nu^p} \qquad (39)\\
&=\frac{|y|^2/(\nu^w)^2+|\widehat{p}|^2/(\nu^p)^2}{(1/\nu^w+1/\nu^p)^2}+\frac{1}{1/\nu^w+1/\nu^p} \\
&\quad+\frac{2y|\widehat{p}|/(\nu^w\nu^p)}{(1/\nu^w+1/\nu^p)^2}C^{-1}y\mathrm{Re}\left\{e^{-j\psi}\int_0^{2\pi}e^{j\theta}\,\mathcal{N}(ye^{j\theta};\widehat{p},\nu^w+\nu^p)d\theta\right\} \qquad (40)\\
&=\frac{|y|^2/(\nu^w)^2+|\widehat{p}|^2/(\nu^p)^2}{(1/\nu^w+1/\nu^p)^2}+\frac{1}{1/\nu^w+1/\nu^p} \\
&\quad+\frac{2y|\widehat{p}|/(\nu^w\nu^p)}{(1/\nu^w+1/\nu^p)^2}C^{-1}y\frac{2}{\nu^w+\nu^p}\exp\left(-\frac{y^2+|\widehat{p}|^2}{\nu^w+\nu^p}\right)I_1\left(\frac{2y|\widehat{p}|}{\nu^w+\nu^p}\right) \qquad (41)\\
&=\frac{|y|^2/(\nu^w)^2+|\widehat{p}|^2/(\nu^p)^2}{(1/\nu^w+1/\nu^p)^2}+\frac{1}{1/\nu^w+1/\nu^p}+\frac{2y|\widehat{p}|/(\nu^w\nu^p)}{(1/\nu^w+1/\nu^p)^2}\frac{I_1\left(\frac{2y|\widehat{p}|}{\nu^w+\nu^p}\right)}{I_0\left(\frac{2y|\widehat{p}|}{\nu^w+\nu^p}\right)}, \qquad (42)
\end{aligned}
$$

where (41) used (34) and (42) used (25). By plugging (42) back into (36), we obtain the expression given in (12).

Appendix B: EM Update for Noise Variance

Noting that

$$
\begin{aligned}
\ln p(\boldsymbol{y},\boldsymbol{x};\nu^w)&=\ln p(\boldsymbol{y}|\boldsymbol{x};\nu^w)+\ln p(\boldsymbol{x};\nu^w) \qquad (43)\\
&=\sum_{m=1}^{M}\ln p_{Y|Z}(y_m|\boldsymbol{a}_m^{\mathsf{H}}\boldsymbol{x};\nu^w)+\text{const} \qquad (44)\\
&=\sum_{m=1}^{M}\ln\left(y_m\int_0^{2\pi}\mathcal{N}(y_me^{j\theta_m};\boldsymbol{a}_m^{\mathsf{H}}\boldsymbol{x},\nu^w)d\theta_m\right)+\text{const}, \qquad (45)
\end{aligned}
$$

where (45) used the expression for $p_{Y|Z}$ from (22), we have

$$\begin{aligned}&\mathrm{E}\big\{\ln p(\boldsymbol{y},\boldsymbol{x};v^w)\big|\boldsymbol{y};\widehat{v^w}[i]\big\}\\&=\int_{\mathbb{C}^N} p(\boldsymbol{x}|\boldsymbol{y};\widehat{v^w}[i])\sum_{m=1}^{M}\ln\Big(\int_0^{2\pi}\mathcal{N}(y_m e^{j\theta_m};\boldsymbol{a}_m^{\mathsf{H}}\boldsymbol{x},v^w)d\theta_m\Big)d\boldsymbol{x}.\end{aligned}\tag{46}$$

To circumvent the high-dimensional integral in (46), we use the same large system limit approximation used in the derivation of GAMP [27]: for sufficiently dense $\boldsymbol{A}$, as $N\to\infty$, the central limit theorem (CLT) suggests that $\boldsymbol{a}_m^{\mathsf{H}}\boldsymbol{x}=z_m$ will become Gaussian. In particular, when $\boldsymbol{x}\sim p(\boldsymbol{x}|\boldsymbol{y};\widehat{v^w}[i])$, the CLT suggests that $\boldsymbol{a}_m^{\mathsf{H}}\boldsymbol{x}\sim\mathcal{N}(\widehat{z}_m,v_m^z)$, where

$$\widehat{z}_m[i]\triangleq\sum_{n=1}^{N}a_{mn}\widehat{x}_n[i],\tag{47}$$

$$v_m^z[i]\triangleq\sum_{n=1}^{N}|a_{mn}|^2v_n^x[i],\tag{48}$$

such that $\widehat{x}_n[i]$ and $v_n^x[i]$ are the mean and variance of the marginal posterior pdf $p(x_n|\boldsymbol{y};\widehat{v^w}[i])$. In this case,

$$\begin{aligned}&\mathrm{E}\big\{\ln p(\boldsymbol{y},\boldsymbol{x};v^w)\big|\boldsymbol{y};\widehat{v^w}[i]\big\}\\&=\sum_{m=1}^{M}\int_{\mathbb{C}}\mathcal{N}(z_m;\widehat{z}_m[i],v_m^z[i])\ln\int_0^{2\pi}\mathcal{N}(y_m e^{j\theta_m};z_m,v^w)d\theta_m dz_m.\end{aligned}\tag{49}$$

From (14), we see that any solution $\widehat{v^w}[i+1]>0$ is necessarily a value of v^w that zeros the derivative of the expected log-pdf. Thus, using the expected log-pdf expression from (49),

$$0=\sum_{m=1}^{M}\int_{\mathbb{C}}\mathcal{N}(z_m;\widehat{z}_m[i],v_m^z[i])\frac{\int_0^{2\pi}\frac{\partial}{\partial v^w}\mathcal{N}(y_m e^{j\theta_m};z_m,\widehat{v^w}[i+1])d\theta_m}{\int_0^{2\pi}\mathcal{N}(y_m e^{j\theta'_m};z_m,\widehat{v^w}[i+1])d\theta'_m}dz_m.\tag{50}$$

Plugging the derivative expression (see [39])

$$\begin{aligned}&\frac{\partial}{\partial v^w}\mathcal{N}(y_m e^{j\theta_m};z_m,\widehat{v^w}[i+1])\\&=\frac{\mathcal{N}(y_m e^{j\theta_m};z_m,\widehat{v^w}[i+1])}{\widehat{v^w}[i+1]^2}\big(|y_m e^{j\theta_m}-z_m|^2-\widehat{v^w}[i+1]\big),\end{aligned}\tag{51}$$

into (50) and multiplying both sides by $\widehat{v^w}[i+1]^2$, we find

$$\widehat{v^w}[i+1]=\frac{1}{M}\sum_{m=1}^{M}\int_{\mathbb{C}}\mathcal{N}(z_m;\widehat{z}_m[i],v_m^z[i])$$

$$\times \frac{\int_0^{2\pi} |y_m e^{j\theta_m} - z_m|^2 \mathcal{N}(y_m e^{j\theta_m}; z_m, \widehat{\nu^w}[i+1]) d\theta_m}{\int_0^{2\pi} \mathcal{N}(y_m e^{j\theta'_m}; z_m, \widehat{\nu^w}[i+1]) d\theta'_m} dz_m \tag{52}$$

$$= \frac{1}{M} \sum_{m=1}^{M} \int_{\mathbb{C}} \mathcal{N}(z_m; \widehat{z}_m[i], \nu_m^z[i]) \times \int_0^{2\pi} |y_m e^{j\theta_m} - z_m|^2 p(\theta_m; z_m, \widehat{\nu^w}[i+1]) d\theta_m dz_m \tag{53}$$

with the newly defined pdf

$$p(\theta_m; z_m, \widehat{\nu^w}[i+1]) \triangleq \frac{\mathcal{N}(y_m e^{j\theta_m}; z_m, \widehat{\nu^w}[i+1])}{\int_0^{2\pi} \mathcal{N}(y_m e^{j\theta'_m}; z_m, \widehat{\nu^w}[i+1]) d\theta'_m} \tag{54}$$

$$\propto \exp\Big(- \frac{|z_m - y_m e^{j\theta_m}|^2}{\widehat{\nu^w}[i+1]} \Big) \tag{55}$$

$$\propto \exp\big(\kappa_m \cos(\theta_m - \phi_m)\big) \text{ for } \kappa_m \triangleq \frac{2|z_m| y_m}{\widehat{\nu^w}[i+1]}, \tag{56}$$

where ϕ_m is the phase of z_m (recall (5)). The proportionality (56) identifies this pdf as a von Mises distribution [49], which can be stated in normalized form as

$$p(\theta_m; z_m, \widehat{\nu^w}[i+1]) = \frac{\exp(\kappa_m \cos(\theta_m - \phi_m))}{2\pi I_0(\kappa_m)}. \tag{57}$$

Expanding the quadratic in (53) and plugging in (57), we get

$$\widehat{\nu^w}[i+1] = \frac{1}{M} \sum_{m=1}^{M} \int_{\mathbb{C}} \mathcal{N}(z_m; \widehat{z}_m[i], \nu_m^z[i]) \Big(y_m^2 + |z_m|^2 - 2y_m|z_m| \int_0^{2\pi} \cos(\theta_m - \phi_m) \frac{\exp(\kappa_m \cos(\theta_m - \phi_m))}{2\pi I_0(\kappa_m)} d\theta_m \Big) dz_m \tag{58}$$

$$= \frac{1}{M} \sum_{m=1}^{M} \int_{\mathbb{C}} \mathcal{N}(z_m; \widehat{z}_m[i], \nu_m^z[i]) \times \left(y_m^2 + |z_m|^2 - 2y_m|z_m| R_0\left(\frac{2|z_m| y_m}{\widehat{\nu^w}[i+1]} \right) \right) dz_m, \tag{59}$$

where $R_0(\cdot)$ is the modified Bessel function ratio defined in (13) and (59) follows from [48, 9.6.19]. To proceed further, we make use of the expansion $R_0(\kappa) = 1 - \frac{1}{2\kappa} - \frac{1}{8\kappa^2} - \frac{1}{8\kappa^3} + o(\kappa^{-3})$ from [50, Lemma 5] to justify the high-SNR approximation

$$R_0(\kappa) \approx 1 - \frac{1}{2\kappa}, \tag{60}$$

which, when applied to (59), yields

$$\widehat{v^w}[i+1] \approx \frac{2}{M} \sum_{m=1}^{M} \int_{\mathbb{C}} \big(y_m - |z_m|\big)^2 \mathcal{N}(z_m; \widehat{z}_m[i], v_m^z[i]) dz_m. \tag{61}$$

Although (61) can be reduced to an expression that involves the mean of a Rician distribution, our empirical experience suggests that it suffices to assume $v_m^z[i] \approx 0$ in (61), after which we obtain the much simpler expression given in (15).

References

1. J.R. Fienup, Phase retrieval algorithms: a comparison. Appl. Opt. **21**, 2758–2769 (1982)
2. R.P. Millane, Recent advances in phase retrieval. Int. Soc. Opt. Eng. **6316** (2006)
3. O. Bunk, A. Diaz, F. Pfeiffer, C. David, B. Schmitt, D.K. Satapathy, J.F. Veen, Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels. Acta Crystallogr. Sect. A Found. Crystallogr. **63**(4), 306–314 (2007)
4. R.W. Harrison, Phase problem in crystallography. J. Opt. Soc. Am. A **10**(5), 1046–1055 (1993)
5. R.P. Millane, Phase retrieval in crystallography and optics. J. Opt. Soc. Am. A **7**, 394–411 (1990)
6. A. Chai, M. Moscoso, G. Papanicolaou, Array imaging using intensity-only measurements. Inverse Prob. **27**(1), 1–16 (2011)
7. A. Walther, The question of phase retrieval in optics. Opt. Acta **10**(1), 41–49 (1963)
8. J.C. Dainty, J.R. Fienup, Phase retrieval and image construction for astronomy, in *Image Recovery: Theory and Application*, ed. by H. Stark, ch. 7 (Academic Press, New York, 1987), pp. 231–275
9. J. Miao, T. Ishikawa, Q. Shen, T. Earnest, Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. Annu. Rev. Phys. Chem. **59**, 387–410 (2008)
10. R. Balan, P.G. Casazza, D. Edidin, On signal reconstruction without phase. Appl. Comput. Harmon. Anal. **20**, 345–356 (2006)
11. L. Demanet, V. Jugnon, Convex recovery from interferometric measurements, *arXiv:1307.6864* (2013)
12. J.V. Corbett, The pauli problem, state reconstruction and quantum-real numbers. Rep. Math. Phys. **57**(1) (2006)
13. T. Heinosaari, L. Mazzarella, M.M. Wolf, Quantum tomography under prior information, *arXiv:1109.5478* (2011)
14. B.G. Bodmann, N. Hammen, Stable phase retrieval with low-redundancy frames, *arXiv:1302.5487* (2013)
15. S. Gazit, A. Szameit, Y.C. Eldar, M. Segev, Super-resolution and reconstruction of sparse sub-wavelength images. Opt. Express, **17**(26), 23920–23946 (2009)
16. A. Szameit, Y. Shechtman, E. Osherovich, E. Bullkich, P. Sidorenko, H. Dana, S. Steiner, E.B. Kley, S. Gazit, T. Cohen-Hyams, S. Shoham, M. Zibulevsky, I. Yavneh, Y.C. Eldar, O. Cohen, M. Segev, Sparsity-based single-shot subwavelength coherent diffractive imaging. Nat. Mater. **11**, 455–459 (2012)
17. S. Marchesini, Ab initio compressive phase retrieval, *arXiv:0809.2006* (2008)
18. Y. Shechtman, E. Small, Y. Lahini, M. Verbin, Y. Eldar, Y. Silberberg, M. Segev, Sparsity-based super-resolution and phase-retrieval in waveguide arrays. Opt. Express **21**(20), 24015–24024 (2013)
19. X. Li, V. Voroninski, Sparse signal recovery from quadratic measurements via convex programming, *arXiv:1209.4785* (2012)

20. M.L. Moravec, J.K. Romberg, R. Baraniuk, Compressive phase retrieval, in *SPIE Conf. Series*, vol. 6701, San Diego (2007)
21. S. Mukherjee, C.S. Seelamantula, An iterative algorithm for phase retrieval with sparsity constraints: application to frequency domain optical coherence tomography, in *Proc. IEEE Int. Conf. Acoust. Speech & Signal Process.*, Kyoto (2012), pp. 553–556
22. H. Ohlsson, A.Y. Yang, R. Dong, S.S. Sastry, CPRL – an extension of compressive sensing to the phase retrieval problem, in *Proc. Neural Inform. Process. Syst. Conf.* (2012) (Full version at *arXiv:1111.6323*)
23. E.J. Candès, T. Strohmer, V. Voroninski, Phase lift: exact and stable signal recovery from magnitude measurements via convex programming. Commun. Pure Appl. Math. **66**, 1241–1274 (2011)
24. P. Netrapalli, P. Jain, S. Sanghavi, Phase retrieval using alternating minimization, in *Proc. Neural Inform. Process. Syst. Conf.* (2013) (See also *arXiv:1306.0160*)
25. Y. Shechtman, A. Beck, Y.C. Eldar, GESPAR: efficient phase retrieval of sparse signals. IEEE Trans. Signal Process. **62**, 928–938 (2014)
26. C. Papadimitriou, K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity* (Dover, New York, 1998)
27. S. Rangan, Generalized approximate message passing for estimation with random linear mixing, in *Proc. IEEE Int. Symp. Inform. Thy.*, Saint Petersburg (2011), pp. 2168–2172. (Full version at *arXiv:1010.5141*)
28. P. Schniter, S. Rangan, Compressive phase retrieval via generalized approximate message passing, in *Proc. Allerton Conf. Commun. Control Comput.*, Monticello (2012)
29. P. Schniter, Compressive phase retrieval via generalized approximate message passing, in *February Fourier Talks (FFT) Workshop on Phaseless Reconstruction*, College Park (2013)
30. D.L. Donoho, A. Maleki, A. Montanari, Message passing algorithms for compressed sensing. Proc. Natl. Acad. Sci. **106**, 18914–18919 (2009)
31. D.L. Donoho, A. Maleki, A. Montanari, Message passing algorithms for compressed sensing: I. Motivation and construction, in *Proc. Inform. Theory Workshop*, Cairo (2010), pp. 1–5
32. J. Pearl, *Probabilistic Reasoning in Intelligent Systems* (Morgan Kaufman, San Mateo, 1988)
33. F.R. Kschischang, B.J. Frey, H.-A. Loeliger, Factor graphs and the sum-product algorithm. IEEE Trans. Inform. Theory **47**, 498–519 (2001)
34. B.J. Frey, D.J.C. MacKay, A revolution: belief propagation in graphs with cycles, in *Proc. Neural Inform. Process. Syst. Conf.*, Denver (1997), pp. 479–485
35. M. Bayati, A. Montanari, The dynamics of message passing on dense graphs, with applications to compressed sensing. IEEE Trans. Inf. Theory **57**, 764–785 (2011)
36. M. Bayati, M. Lelarge, A. Montanari, Universality in polytope phase transitions and iterative algorithms, in *Proc. IEEE Int. Symp. Inf. Theory*, Boston, (2012), pp. 1–5 (see also *arXiv:1207.7321*)
37. S.O. Rice, Mathematical analysis of random noise. Bell Syst. Tech. J. **24**(1), 46–156 (1945)
38. A. Dempster, N.M. Laird, D.B. Rubin, Maximum-likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. **39**, 1–17 (1977)
39. J.P. Vila, P. Schniter, Expectation-maximization Gaussian-mixture approximate message passing. IEEE Trans. Signal Process. **61**, 4658–4672 (2013)
40. G.F. Cooper, The computational complexity of probabilistic inference using Bayesian belief networks. Artif. Intell. **42**, 393–405 (1990)
41. P. Schniter, Turbo reconstruction of structured sparse signals, in *Proc. Conf. Inform. Science & Syst.*, Princeton (2010), pp. 1–6
42. M. Borgerding and P. Schniter, Generalized approximate message passing for the cosparse analysis model, *arXiv:1312.3968* (2013)
43. J. Vila, P. Schniter, An empirical-Bayes approach to recovering linearly constrained non-negative sparse signals, in *Proc. IEEE Workshop Comp. Adv. Multi-Sensor Adaptive Process.*, Saint Martin (2013), pp. 5–8 (Full version at *arXiv:1310.2806*)
44. S. Som, L.C. Potter, P. Schniter, On approximate message passing for reconstruction of non-uniformly sparse signals, in *Proc. National Aerospace and Electronics Conf.*, Dayton (2010)

45. S. Rangan, P. Schniter, A. Fletcher, On the convergence of generalized approximate message passing with arbitrary matrices, in *Proc. IEEE Int. Symp. Inform. Thy.*, Honolulu (2014) (Full version at *arXiv:1402.3210*)
46. E.J. Candès, X. Li, M. Soltanolkotabi, Phase retrieval from coded diffraction patterns, *arXiv:1310.3240* (2013)
47. G. Zheng, R. Horstmeyer, C. Yang, Wide-field, high-resolution Fourier ptychographic microscopy. Nat. Photon. **7**, 739–745 (2013)
48. M. Abramowitz, I.A. Stegun, eds., *Handbook of Mathematical Functions* (Dover, New York, 1964)
49. K.V. Mardia, P.E. Jupp, *Directional Statistics* (Wiley, New York, 2009)
50. C. Robert, Modified Bessel functions and their applications in probability and statistics. Stat. Probab. Lett. **9**, 155–161 (1990)

Importance sampling in signal processing applications

Rachel Ward

Abstract *Importance sampling* is a technique originating in Monte Carlo simulation whereby one samples from a different, *weighted* distribution, in order to reduce variance of the resulting estimator. More recently, variations of importance sampling have emerged as a means for reducing computational and sample complexity in different problems of modern signal processing. Here we review importance sampling as it is manifested in three such problems: stochastic optimization, compressive sensing, and low-rank matrix approximation. In keeping with a general trend in convex optimization towards the analysis of phase transitions for exact recovery, importance sampling in compressive sensing and low-rank matrix recovery can be used to effectively push the phase transition for exact recovery towards fewer measurements.

Key words: Complexity, Compressive sensing, Importance sampling, Matrix completion, Measurements, Stochastic gradient, Weighted sampling

Introduction

Importance sampling in simulation

The usual setup for importance sampling is in Monte Carlo simulation: one wants to compute an integral of the form $\int_{\mathscr{D}} f(x)p(x)dx$, where $p(x)$ is a probability density: $\int_{\mathscr{D}} p(x)dx = 1$. An easy and computationally efficient way to approximate such an integral is to consider the integral as an expectation, $\mu = \mathbb{E}(f(x)) = \int_{\mathscr{D}} f(x)p(x)dx$, and approximate the expectation as a sample average,

R. Ward (✉)
Department of Mathematics, University of Texas at Austin, Austin, TX, USA
e-mail: rward@math.utexas.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_8

$$\int_{\mathscr{D}} f(x)dx \approx \frac{1}{m}\sum_{i=1}^{m} f(x_i), \quad x_i \sim p,$$

where the random variables x_i are independent and ideally distributed. Validity of this approximation is ensured by the law of large numbers, but the number of samples m needed for a given approximation accuracy grows with the variance of the random variable $f(x)$. In particular, if $f(x)$ is nearly zero on its domain $\mathscr{D}$ except in a region $A \subset \mathscr{D}$ for which $\mathbb{P}(x \in A)$ is small, then such standard Monte Carlo sampling may fail to have even one point inside the region A. It is clear intuitively that in this situation, we would benefit from getting some samples from the interesting or important region A. What *importance sampling* means is to sample from a different density $q(x)$ which overweights this region, rescaling the resulting quantity in order that the estimate remain unbiased.

More precisely, if x has probability density $p(x)$, then

$$\begin{aligned}\mu = \mathbb{E}[f(x)] &= \int_{\mathscr{D}} f(x)p(x)dx \\ &= \int_{\mathscr{D}} f(x)\frac{p(x)}{q(x)}q(x)dx = \mathbb{E}_q[f(x)w(x)],\end{aligned} \tag{1}$$

where $w(\cdot) \equiv \frac{p(\cdot)}{q(\cdot)}$ is the *weighting* function. By (1), the estimator

$$\widehat{\mu} = \frac{1}{m}\sum_{i=1}^{m} f(x_i)w(x_i), \qquad x_i \sim q, \tag{2}$$

is also an unbiased estimator for μ. The *importance sampling problem* then focuses on finding a biasing density $q(x)$ which overweights the important region close to an "optimal" way, at least such the *variance* of the importance sampling estimator is smaller than the variance of the general Monte Carlo estimate, so that fewer samples m are required to achieve a prescribed estimation error. In general, the density q^* with minimal variance σ^2_{q*} is proportional to $|f(x)|p(x)$, which is unknown a priori; still, there are many techniques for estimating or approximating this optimal distribution, see [31, Chapter 9].

Importance sampling beyond simulation

In recent times, probabilistic and stochastic algorithms have seen an explosion of growth as we move towards *bigger* data problems in *higher* dimensions. Indeed, we are often in the situation where at least one of the following is true:

1. Taking measurements is expensive, and we would like to reduce the number of measurements needed to reach a prescribed approximation accuracy

2. Optimizing over the given data is expensive, and we would like to reduce the number of computations needed to get within a prescribed tolerance of the optimal solution.

Importance sampling has proved to be helpful in both regimes. Whereas in simulation, importance sampling has traditionally been used for approximating *linear* estimates such as expectations/integrals, recent applications in signal processing and machine learning have considered importance sampling in approximating or even exactly recovering nonlinear estimates as well.

We consider here three case studies where the principle of importance sampling has been applied; this is by no means a complete list of all applications of importance sampling to machine learning and signal processing problems.

1. **Stochastic optimization:** Towards minimizing $F : \mathbb{R}^n \to \mathbb{R}$ of the form $F(x) = \sum_{i=1}^m f_i(x)$ via stochastic gradient descent, one iterates $x_{k+1} = x_k - \gamma w(i_k) \nabla f_{i_k}(x_k)$ with i_k randomly chosen from $\{1,2,\ldots,m\}$ so that

$$\mathbb{E}_{i_k}[x_{k+1}] = x_k - \gamma \sum_{i=1}^m \nabla f_i(x_k);$$

that is, one implements a full gradient descent update at each iteration, *in expectation*. Standard procedure is to sample indices from $\{1,2,\ldots,m\}$ uniformly, and the resulting convergence rate is limited by the *worst-case* Lipschitz constant associated with the component gradient functions. If however one has prior knowledge about the component Lipschitz constants, and has the liberty to draw indices proportionately to the associated Lipschitz constants, then the convergence rate of stochastic gradient can be improved so as to depend on the *average* Lipschitz constant among the components. This is in line with the principle of importance sampling: if ∇f_i has a larger Lipschitz constant, then this component is contributing more in content, and should be sampled with higher probability. We review some results of this kind in more detail below. For more details, see Section "Importance sampling in Stochastic Optimization".
2. **Compressive sensing:** Consider an orthonormal matrix $\Phi \in \mathbb{R}^{n\times n}$ (or $\Phi \in \mathbb{C}^{n\times n}$), along with a vector $x \in \mathbb{R}^n$. Then clearly

$$\Phi^* \Phi x = x;$$

moreover, if $\varphi_{i_k} \in \mathbb{R}^{1,n}$ is a randomly selected row from Φ, drawn such that row i is sampled with probability $p(i)$, then also

$$\mathbb{E}_p \left[\frac{1}{[p(i_k)]^2} \left(\varphi_{i_k}^* \varphi_{i_k} \right) \right] x = x.$$

Compressive sensing shows that if x is s-sparse, with $s \ll n$, then for certain orthonormal Φ, as few as $m \propto s \log^4(n)$ i.i.d. samples of the form $\langle \varphi_{i_k}, x \rangle$ can suffice to *exactly* recover x as the solution to a convex optimization program. For instance, such results hold if all of the rows of Φ are "equally important" (i.e., Φ

has uniformly bounded entries), and if rows are drawn i.i.d. uniformly from Φ. One may also incorporate importance sampling: if rows are drawn i.i.d. proportionately to their squared Euclidean norm, and if the *average* Euclidean row norm is small, then $m \propto s\log^4(n)$ i.i.d. samples still suffice for exact reconstruction. For more details, see Section "'Importance sampling in compressive sensing'.

3. **Low-rank matrix approximations:** Consider a matrix $M \in \mathbb{R}^{n_1 \times n_2}$ of rank $r \ll \min\{n_1, n_2\}$, and a subset $\Omega \subset [n_1] \times [n_2]$ of $|\Omega| = m$ revealed entries $M_{i,j}$. If the entries are revealed as i.i.d. draws where $\mathbf{Prob}[(i,j)] = p_{i,j}$, then $\mathbb{E}\left[\frac{1}{p_{i,j}} M_{i,j}\right] = M$. Importance sampling here corresponds to putting more weight $p_{i,j}$ on "important" entries in order to exactly recover M using fewer samples. We will see that if entries are drawn from a weighted distribution based on matrix *leverage scores*, then $m = r\log^2(\max\{n_1, n_2\})$ revealed entries suffices for M to be exactly recoverable as the solution to a convex optimization problem.

Importance sampling in Stochastic Optimization

Gradient descent is a standard method for solving unconstrained optimization problems of the form

$$\min_{x \in \mathbb{R}^n} F(x); \tag{3}$$

gradient descent proceeds as follows: initialize $x_0 \in \mathbb{R}^n$, and iterate along the direction of the negative gradient of F (the direction of "steepest descent") until convergence

$$x_{k+1} = x_k - \gamma_k \nabla F(x_k). \tag{4}$$

Here γ_k is the step-size which may change at every iteration. For optimization problems of very big size, however, even a full gradient computation of the form $\nabla F(x_k)$ can require substantial computational efforts and full gradient descent might not be feasible. This has motivated recent interest in random coordinate descent or stochastic gradient methods (see [3, 28, 29, 35, 36, 40], to name just a few), where one descends along gradient directions which are cheaper to compute. For example, suppose that F to be minimized is differentiable and admits a decomposition of the form

$$F(x) = \sum_{i=1}^{m} f_i(x). \tag{5}$$

Since $\nabla F(x) = \sum_{i=1}^m \nabla f_i(x)$, a full gradient computation involves computing all m gradients $\nabla f_i(x)$; still, one could hope to get *close* to the minimum, at a much smaller expense, by instead selecting a single index i_k at random from $\{1, 2, \ldots, m\}$ at each iteration. This is the principle behind *stochastic gradient* descent.

(5) **Stochastic Gradient (SG)**

Consider the minimization of $F : \mathbb{R}^n \to \mathbb{R}$ of the form $F(x) = \sum_{i=1}^m f_i(x)$.
Choose $x_0 \in \mathbb{R}^n$.
For $k \geq 1$ iterate until convergence criterion is met:

1. Choose $i_k \in [m]$ according to the rule $\mathbf{Prob}[i_k = k] = w(k)$
2. Update $x_{k+1} = x_k - \gamma \frac{1}{w(i_k)} \nabla f_{i_k}(x_k)$.

We have set the step-size γ to be constant for simplicity. Note that with the normalization in the update rule,

$$\begin{aligned} \mathbb{E}^{(w)}[x_{k+1}] &= x_k - \gamma \sum_{i=1}^m \nabla f_{i_k}(x_k) \\ &= x_k - \gamma \nabla F(x_k). \end{aligned} \tag{6}$$

Thus, we might hope for convergence *in expectation* of such stochastic iterations to the minimizer of (5) under similar conditions guaranteeing convergence of full gradient descent, namely, when F is convex (so that every minimizer is a global minimizer) and ∇F is Lipschitz continuous [30]. That is, we will assume

1. F is convex with convexity parameter $\mu = \mu(F) \geq 0$: for any x and y from $\mathbb{R}^n$ we have
$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{1}{2}\mu \|y - x\|^2. \tag{7}$$
When $\mu > 0$ strictly, we say that F is μ-strongly convex.
2. The component functions f_i are continuously differentiable and satisfy
$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|y - x\|, \qquad i = 1, 2, \ldots, m, \quad x, y \in \mathbb{R}^n. \tag{8}$$
We refer to L_i as the *Lipschitz constant* of ∇f_i.

The default sampling strategy in stochastic gradient methods is to sample uniformly, taking $w(i) = \frac{1}{m}$ in (5). In cases where the component functions f_i are only observed sequentially or in a streaming fashion, one does not have the freedom to choose a different sampling strategy. But if one *does* have such freedom, and has prior knowledge about the distribution of the Lipschitz constants L_i associated with the component function gradients, choosing probabilities $w(i) \propto L_i$ can significantly speed up the convergence rate of stochastic gradient. This is in line with the principle of importance sampling: if ∇f_i has a larger Lipschitz constant, it is contributing more in content, and should be sampled with higher probability. We review some results of this kind in more detail below.

Stochastic Gradient (SG) with Importance Sampling

For strongly convex functions, a central quantity in the analysis of stochastic descent is the *conditioning* of the problem, which is, roughly speaking, the ratio of the Lipschitz constant to the parameter of strong convexity. Recall that for a convex quadratic $F(x) = \frac{1}{2}x'Hx$, the Lipschitz constant of the gradient is given by the maximal eigenvalue of the Hessian H while the parameter of strong convexity is given by its minimal eigenvalue, and so in this case the conditioning reduces to the condition number of the Hessian matrix. In the general setting where $F(x) = \sum_{i=1}^m f_i(x)$ is strongly convex, the Hessian can vary with x, and the results will depend on the Lipschitz constants L_i of the ∇f_i and not only of the aggregate ∇F.

In short: with importance sampling, the convergence rate of stochastic descent is proportional to the *average conditioning* $\overline{L}/\mu = \frac{1}{m}\sum_{i=1}^m L_i/\mu$ of the problem; without importance sampling, the convergence rate must depend on the *uniform conditioning* $\sup_i L_i/\mu$. Thus, importance sampling has the highest potential impact if the Lipschitz constants are highly variable. This is made precise in the following theorem from [26], which in the case of uniform sampling, improves on a previous result of [2].

Theorem 1. *Let each f_i be convex where ∇f_i has Lipschitz constant L_i, with $L_i \leq \sup L$, and let $F(x) = \mathbb{E}f_i(x)$ be μ-strongly convex. Set $\sigma^2 = \mathbb{E}\|\nabla f_i(x_*)\|^2$, where $x_* = \text{argmin}_x F(x)$. Suppose that $\gamma \leq \frac{1}{\mu}$. Then the SG iterates in (5) satisfy:*

$$\mathbb{E}\|x_k - x_*\|^2 \leq \Big[1 - 2\gamma\mu(1-\gamma\sup L)\Big]^k \|x_0 - x_*\|^2 + \frac{\gamma\sigma^2}{\mu(1-\gamma\sup L)}. \tag{9}$$

where the expectation is with respect to the sampling of $\{i_k\}$ in (5).

The parameter σ^2 should be thought of as a 'residual' parameter measuring the extent to which the component functions f_i share a common minimizer. As a corollary of Theorem 1, if one pre-specifies a target accuracy $\varepsilon > 0$, then the optimal step-size $\gamma^* = \gamma^*(\varepsilon, \mu, \sigma^2, \sup L)$ is such that

$$k = 2\log(\varepsilon_0/\varepsilon)\left(\frac{\sup L}{\mu} + \frac{\sigma^2}{\mu^2\varepsilon}\right) \tag{10}$$

SG iterations suffice so that $\mathbb{E}\|x_k - x_*\|_2^2 \leq \varepsilon$,. See [26] for more details.

To see what this result implies for importance sampling, consider the stochastic gradient algorithm (5) with weights $w^{(k)}$. Then, when expectation is taken with respect to the sampling of $\{i_k\}$, we have $F(x) = \mathbb{E}f_i^{(w)}(x)$ where $f_i^{(w)} = \frac{1}{w^{(k)}}f_i$ has Lipschitz constant $L_i^{(w)} = \frac{1}{w(i)}L_i$. The supremum of $L_i^{(w)}$ is then given by:

$$\sup L_{(w)} = \sup_i L_i^{(w)} = \sup_i \frac{L_i}{w(i)}. \tag{11}$$

It is easy to verify that (11) is minimized by the weights

$$w(i) = \frac{L_i}{\overline{L}}, \quad \text{so that} \quad \sup L_{(w)} = \sup_i \frac{L_i}{L_i/\overline{L}} = \overline{L}. \tag{12}$$

Since μ is invariant to choice of weights, we find that in the "realizable" regime where $\sigma^2 = 0$, and hence $\sigma^2_{(w)} = 0$, then choosing the weights $w(i)$ as in (11) gives linear convergence with a linear dependence on the average conditioning $\overline{L}/\mu$, and a number of iterations,

$$k^{(w)} \propto \log(1/\varepsilon)\overline{L}/\mu,$$

to achieve a target accuracy ε. This strictly improves over the best possible results with uniform sampling, where the linear dependence is on the uniform conditioning $\sup L/\mu$ (see [26] for more details).

However, when $\sigma^2 > 0$, we get a potentially much *worse* scaling of the second term, by a factor of $\overline{L}/\inf L$:

$$\sigma^2_{(w)} = \mathbb{E}^{(w)}[\|\nabla f_i^{(w)}(x)\|_2^2] \leq \frac{\overline{L}}{\inf L}\sigma^2. \tag{13}$$

Fortunately, we can easily overcome this factor by sampling from a mixture of the uniform and fully weighted sampling, referred to as *partially biased sampling*. Using the weights

$$w(i) = \frac{1}{2}\frac{L_i}{\overline{L}} + \frac{1}{2}m,$$

we have

$$\sup L_{(w)} = \sup_i \frac{1}{\frac{1}{2} + \frac{1}{2}\cdot\frac{L_i}{\overline{L}}} L_i \leq 2\overline{L} \tag{14}$$

and

$$\sigma^2_{(w)} = \mathbb{E}\left[\frac{1}{\frac{1}{2} + \frac{1}{2}\cdot\frac{L_i}{\overline{L}}}\|\nabla f_i(x)\|_2^2\right] \leq 2\sigma^2. \tag{15}$$

In this sense, under the assumptions of Theorem 1, partially biased sampling will never be worse in terms of convergence rate than uniform sampling, up to a factor of 2, but can potentially have much better convergence.

Remark 1. An important example where all of these parameters have explicit forms is the *least squares problem*, where

$$F(x) = \frac{1}{2}\|Ax - b\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}(\langle a_i, x\rangle - b_i)^2, \tag{16}$$

with b an m-dimensional vector, A an $m \times n$ matrix with rows a_i, and $x_* = \arg\min_x \frac{1}{2}\|Ax - b\|_2^2$ is the least-squares solution. The Lipschitz constants of the components $f_i = \frac{m}{2}(\langle a_i, x\rangle - b_i)^2$ are $L_i = m\|a_i\|_2^2$, and the average Lipschitz constant is $\frac{1}{m}\sum_i L_i = \|A\|_F^2$ where $\|\cdot\|_F$ denotes the Frobenius norm. If A is full-rank and overdetermined, then F is strongly convex with strong convexity parameter $\mu = \|(A^TA)^{-1}\|_2^{-1}$, so that the average condition number is $\overline{L}/\mu = \|A\|_F^2\|(A^TA)^{-1}\|_2$.

Moreover, the residual is $\sigma^2 = m\sum_i \|a_i\|^2 |\langle a_i, x\rangle - b_i|^2$. Observe the bounds $\sigma^2 \leq n\|A\|_F^2 \sup_i |\langle a_i, x\rangle - b_i|^2$ and $\sigma^2 \leq m \sup_i \|a_i\|^2 \|Ax_* - b\|_2^2$.

Importance Sampling for SG in other regimes

Theorem 1 is stated for smooth and strongly convex objectives, and is particularly useful in the regime where the residual σ^2 is low, and the linear convergence term is dominant. But importance sampling can be incorporated into SG methods also in other regimes, and we now briefly survey some of these possibilities.

Smooth, Not Strongly Convex

When each component f_i is convex, non-negative, and has an L_i-Lipschitz gradient, but the objective $F(x)$ is not necessarily strongly convex, then after

$$k = O\left(\frac{(\sup L)\|x_*\|_2^2}{\varepsilon} \cdot \frac{F(x_*) + \varepsilon}{\varepsilon}\right) \tag{17}$$

iterations of SGD with an appropriately chosen step-size we will have $F(\bar{x}) \leq F(x_*) + \varepsilon$, where $\bar{x}$ is an appropriate averaging of the k iterates [43]. The relevant quantity here determining the iteration complexity is again $\sup L$. Furthermore, the dependence on the supremum is unavoidable and *cannot* be replaced with the average Lipschitz constant $\overline{L}$ [43]: if we sample gradients according to the uniform distribution, we must have a linear dependence on $\sup L$.

The only quantity in (17) that changes with a re-weighting is $\sup L$—all other quantities ($\|x_*\|_2^2$, $F(x_*)$, and the sub-optimality ε) are invariant to re-weightings. We can therefore replace the dependence on $\sup L$ with a dependence on $\sup L_{(w)}$ by using a weighted SGD as in (12). As we already calculated, the optimal weights are given by (12), and using them we have $\sup L_{(w)} = \overline{L}$. In this case, there is no need for partially biased sampling and we obtain that

$$k = O\left(\frac{\overline{L}\|x_*\|_2^2}{\varepsilon} \cdot \frac{F(x_*) + \varepsilon}{\varepsilon}\right) \tag{18}$$

iterations of weighed SGD updates (5) using the weights (12) suffice.

Non-Smooth Objectives

We now turn to non-smooth objectives, where the components f_i might not be smooth, but each component is G_i-Lipschitz. Roughly speaking, G_i is a bound on the first derivative (gradient) of f_i, while L_i is a bound on the second derivatives of f_i.

Here, the performance of SGD depends on the second moment $\overline{G^2} = \mathbb{E}[G_i^2]$. The precise iteration complexity depends on whether the objective is strongly convex or whether x_* is bounded, but in either case depends linearly on $\overline{G^2}$.

Using weighted SGD, we get linear dependence on:

$$\overline{G_{(w)}^2} = \mathbb{E}^{(w)}\left[(F_i^{(w)})^2\right] = \mathbb{E}^{(w)}\left[\frac{G_i^2}{w(i)^2}\right] = \mathbb{E}\left[\frac{G_i^2}{w(i)}\right], \tag{19}$$

where $F_i^{(w)} = G_i/w(i)$ is the Lipschitz constant of the scaled $f_i^{(w)}$. This is minimized by the weights $w(i) = G_i/\overline{G}$, where $\overline{G} = \mathbb{E}[G_i]$, yielding $\overline{G_{(w)}^2} = \overline{G}^2$. Using importance sampling, we reduce the dependence on $\overline{G^2}$ to a dependence on $\overline{G}^2$. It is helpful to recall that $\overline{G^2} = \overline{G}^2 + \mathrm{Var}[G_i]$. What we save is thus exactly the variance of the Lipschitz constants G_i. For more details, see [46].

Importance sampling in random coordinate descent

A related stochastic optimization problem is *randomized coordinate descent*, where one minimizes $F : \mathbb{R}^n \to \mathbb{R}$, not necessarily having the form $F(x) = \sum_{i=1}^m f_i(x)$, but still assumed to be strongly convex, by decomposing its gradient (5) into its *coordinate* directions

$$\nabla F(x) = \sum_{i=1}^{n} \nabla_i F(x)$$

and performing the stochastic updates:

1. Choose coordinate $i \in [n]$ according to rule $\mathbf{Prob}[i_k = k] = w(k)$
2. Update $x_{k+1} = x_k - \gamma \frac{1}{w(i_k)} \nabla_{i_k} F(x_k)$.

The motivation is that a coordinate directional derivative can be much simpler than computation of either function value, or a directional derivative along an *arbitrary* direction.

Actually, Theorem 1 can also be applied to this setting; its proof from [26] uses only that

$$\nabla F(x) = \mathbb{E}[\nabla f_i(x)], \tag{20}$$

and the fact that for, given any $x, y \in \mathbb{R}^n$,

$$\|\nabla f_i(x) - \nabla f_i(y)\|_2^2 \le L_i \langle x - y, \nabla f_i(x) - \nabla f_i(y)\rangle. \tag{21}$$

which follows from the assumption that f_i is smooth with Lipschitz continuous gradient by the so-called *co-coercivity Lemma*, see [26, Lemma A.1]. Note that (20) still

holds in the setting of randomized coordinate descent, and (21) holds if $F : \mathbb{R}^n \to \mathbb{R}$ has component-wise Lipschitz continuous gradient:

$$|\nabla_i F(x + he_i) - \nabla_i F(x)| \leq L_i |h|, \quad x \in \mathbb{R}^n,\ h \in \mathbb{R},\ i \in [n], \tag{22}$$

Under these assumptions, one may consider importance sampling for random coordinate descent with weights $w(k) = L_k / \sum_j L_j$, then we may apply Theorem 1 to get a linear convergence rate depending on $\overline{L}/\mu$ as opposed to $\sup L/\mu$. This is because coordinate descent falls into the *realizable* regime, as $\nabla_i F(x_*) = 0$ for each i, and hence also $\sigma^2 = \mathbb{E}\|(\nabla F)_i(x_*)\|^2 = 0$. Coordinate descent with importance sampling was considered before SG with importance sampling, originating in the works of [29] and [35]. One may consider the extension of randomized coordinate descent (8) to randomized *block* coordinate descent, descending in *blocks* of coordinates at a time. Then, the important Lipschitz constants are those associated with the *partial gradients* of F as opposed to the component-wise gradients [29].

Notes and extensions

Several aspects of importance sampling in stochastic optimization were not covered here, but we point out further results and references.

1. If the Lipschitz constants are not known a priori, then one could still consider doing importance sampling via *rejection sampling*, simulating sampling from the weighted distribution; this can be done by accepting samples with probability proportional to $L_i / \sup_j L_j$. The overall probability of accepting a sample is then $\overline{L}/\sup L_i$, introducing an additional factor of $\sup L_i/\overline{L}$, and thus again obtaining a linear dependence on $\sup L_i$. Thus, if we are only presented with samples from the uniform distribution, and the cost of obtaining the sample dominates the cost of taking the gradient step, we do not gain (but do not lose much either) from rejection sampling. We might still gain from rejection sampling if the cost of operating on a sample (calculating the actual gradient and taking a step according to it) dominates the cost of obtaining it and (a bound on) the Lipschitz constant.
2. All of the convergence results we stated in this section were with respect to the expected value. Nevertheless, all these rates extend to high probability results using Chebyshev's inequality. See [29] for more details.
3. Recently, several *hybrid* full-gradient/stochastic gradient methods have emerged which, as opposed to pure SG as in (5), have the advantage of progressively reducing the variance of the stochastic gradient with the iterations [19, 37, 41, 42], thus allowing convergence to the true minimizer. These algorithms can further be applied to the more general class of composite problems,

$$\text{minimize}_{x \in \mathbb{R}^n} \{P(x) = F(x) + R(x)\}, \tag{23}$$

where $F(x)$ is the average of many smooth component functions $f_i(x)$ whose gradients have Lipschitz constants L_i as in (5) and $R(x)$ is relatively simple but can

be non-differentiable. These algorithms have the added complexity of requiring a single pass over the data, all having complexity $O((n+\sup L/\mu)\log(1/\varepsilon))$. As shown in [45], importance sampling can also be applied in this more general setting to speed up convergence: sampling component functions proportional to their Lipschitz constants, this complexity bound becomes $O((n+\overline{L}/\mu)\log(1/\varepsilon))$.

4. An observation that is important not only for this chapter but also for the entire discussion on importance sampling is the computational cost of implementing a random counter, that is, given values $L_1, L_2, \ldots, L_m$, generate efficiently random integer numbers $i \in \{1,2,\ldots,m\}$ with probabilities

$$\mathbf{Prob}[i=k] = \frac{L_k}{\sum_{j=1}^{m} L_j}, \quad k=1,2,\ldots,m. \tag{24}$$

Using a tree search algorithm [29], such a counter can be implemented with $\log(m)$ operations, and by generating one random number.

Importance sampling in compressive sensing

Introduction

The emerging area of mathematical signal processing known as *compressive sensing* is based on the observation that a signal which allows for an approximately sparse representation in a suitable basis or dictionary can be recovered from relatively few linear measurements via convex optimization, provided these measurements are sufficiently *incoherent* with the basis in which the signal is sparse [8, 10, 38]. In this section we will see how importance sampling can be used to enhance the incoherence between measurements and signal basis, again, allowing for recovery from fewer linear measurements.

We illustrate the power of importance sampling through two examples: compressed sensing imaging and polynomial interpolation. In compressed sensing imaging, coherence-based sampling provides a theoretical justification for empirical studies [23, 24] pointing to variable-density sampling strategies for improved MRI compressive imaging. In polynomial interpolation, coherence-based sampling implies that sampling points drawn from the Chebyshev distribution are better suited for the recovery of polynomials and smooth functions than uniformly distributed sampling points, aligning with classical results on Lagrange interpolation [5].

Before continuing, let us fix some notation. A vector $x \in \mathbb{C}^N$ is called s-sparse if $\|x\|_0 = \#\{x_j : x_j \neq 0\} \leq s$, and the best s-term approximation of a vector $x \in \mathbb{C}^N$ is the s-sparse vector $x_s \in \mathbb{C}^N$ satisfying $x_s = \inf_{u:\|u\|_0 \leq s} \|x-u\|_p$. Clearly, $x_s = x$ if x is s-sparse. Informally, x is called compressible if $\|x-x_s\|$ decays quickly as s increases.

Incoherence in compressive sensing

Here we recall sparse recovery results for structured random sampling schemes corresponding to *bounded orthonormal systems*, of which the partial discrete Fourier transform is a special case. We refer the reader to [15] for an expository article including many references.

Definition 1 (Bounded orthonormal system (BOS)). Let $\mathscr{D}$ be a measurable subset of $\mathbb{R}^d$.

- A set of functions $\{\psi_j : \mathscr{D} \to \mathbb{C}, \quad j \in [N]\}$ is called an *orthonormal system* with respect to the probability measure ν if $\int_{\mathscr{D}} \bar{\psi}_j(u)\psi_k(u)d\nu(u) = \delta_{jk}$, where δ_{jk} denotes the Kronecker delta.
- Let μ be a probability measure on $\mathscr{D}$. A *random sample* of the orthonormal system $\{\psi_j\}$ is the random vector $(\psi_1(T), \ldots, \psi_N(T))$ that results from drawing a sampling point T from the measure μ.
- An orthonormal system is said to be *bounded* with bound K if $\sup_{j\in[N]} \|\psi_j\|_\infty \leq K$.

Suppose now that we have an orthonormal system $\{\psi_j\}_{j\in[N]}$ and m random sampling points $T_1, T_2, \ldots, T_m$ drawn independently from some probability measure μ. Here and throughout, we assume that the number of sampling points $m \ll N$. As shown in [15], if the system $\{\psi_j\}$ is *bounded*, and if the probability measure μ from which we sample points is the orthogonalization measure ν associated with the system, then the (underdetermined) structured random matrix $A : \mathbb{C}^N \to \mathbb{C}^m$ whose rows are the independent random samples will be well conditioned, satisfying the so-called *restricted isometry property* [11] with nearly order-optimal restricted isometry constants with high probability. Consequently, matrices associated with random samples of bounded orthonormal systems have nice sparse recovery properties.

Proposition 1 (Sparse recovery through BOS). *Consider the matrix $A \in \mathbb{C}^{m\times N}$ whose rows are independent random samples of an orthonormal system $\{\psi_j,\ j \in [N]\}$ with bound $\sup_{j\in[N]} \|\psi_j\|_\infty \leq K$, drawn from the orthogonalization measure ν associated with the system. If the number of random samples satisfies*

$$m \gtrsim K^2 s \log^3(s) \log(N), \tag{25}$$

for some $s \gtrsim \log(N)$, then the following holds with probability exceeding $1 - N^{-C\log^3(s)}$: For each $x \in \mathbb{C}^N$, given noisy measurements $y = Ax + \sqrt{m}\eta$ with $\|\eta\|_2 \leq \varepsilon$, the approximation

$$x^\# = \arg\min_{z\in\mathbb{C}^N} \|z\|_1 \text{ subject to } \|Az - y\|_2 \leq \sqrt{m}\varepsilon$$

satisfies the error guarantee $\|x - x^\#\|_2 \lesssim \frac{1}{\sqrt{s}}\|x - x_s\|_1 + \varepsilon$.

An important special case of such a matrix construction is the *subsampled discrete Fourier matrix*, constructed by sampling $m \ll N$ rows uniformly at random from

the unitary discrete Fourier matrix $\Psi \in \mathbb{C}^{N\times N}$ with entries $\psi_{j,k} = \frac{1}{\sqrt{N}} e^{i2\pi(j-1)(k-1)}$. Indeed, the system of complex exponentials $\psi_j(u) = e^{i2\pi(j-1)u}$, $j \in [N]$, is orthonormal with respect to the uniform measure over the discrete set $\mathscr{D} = \{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\}$, and is bounded with optimally small constant $K = 1$. In the discrete setting, we may speak of a more general procedure for forming matrix constructions adhering to the conditions of Proposition 1: given any two unitary matrices Φ and Ψ, the composite matrix $\Phi^*\Psi$ is also a unitary matrix, and this composite matrix will have uniformly bounded entries if the orthonormal bases (ϕ_j) and (ψ_k), corresponding to the rows of Φ and Ψ, respectively, are *mutually incoherent:*

$$\mu(\Phi, \Psi) := \sqrt{N} \sup_{1\le j,k\le N} |\langle \phi_j, \psi_k \rangle| \le K. \tag{26}$$

Indeed, if Φ and Ψ are mutually incoherent, then the rows of $B = \sqrt{N}\Psi^*\Phi$ constitute a bounded orthonormal system with respect to the uniform measure on $\mathscr{D} = \{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\}$. Proposition 1 then implies a sampling strategy for reconstructing signals $x \in \mathbb{C}^N$ with assumed sparse representation in the basis Ψ, that is $x = \Psi b$ and $b \approx b_s$ (the s-sparse vector corresponding to its best s-term approximation), from a few linear measurements: form a sensing matrix $A \in \mathbb{C}^{m\times N}$ by sampling rows i.i.d. uniformly from an incoherent basis Φ, collect measurements $y = Ax + \eta$, $\|\eta\|_2 \le \varepsilon$, and solve the ℓ_1 minimization program,

$$x^{\#} = \arg\min_{z\in\mathbb{C}^N} \|\Psi^* z\|_1 \text{ subject to } \|Az - y\|_2 \le \sqrt{m}\varepsilon.$$

This scenario is referred to as *incoherent sampling*.

Importance sampling via local coherences

Consider more generally the setting where we aim to compressively sense signals $x \in \mathbb{C}^N$ with assumed sparse representation in the orthonormal basis $\Psi \in \mathbb{C}^{N\times N}$, but our sensing matrix $A \in \mathbb{C}^{m\times N}$ can only consist of rows from some fixed orthonormal basis $\Phi \in \mathbb{C}^{N\times N}$ that is not necessarily incoherent with Ψ. In this setting, we ask: *Given a fixed sensing basis Ψ and sparsity basis Φ, how should we sample rows of Ψ in order to make the resulting system as incoherent as possible?* We will answer this question by introducing the concept of *local coherence* between two bases as described in [21, 32], whereby in the discrete setting the coherences of individual elements of the sensing basis are calculated and used to derive the sampling strategy.

The following result quantifies how regions of the sensing basis that are more coherent with the sparsity basis should be sampled with higher density: they should be given more "importance". The following is essentially a generalization of Theorem 2.1 in [32], but for completeness, we include a short self-contained proof.

Theorem 2 (Sparse recovery via local coherence sampling). *Consider a measurable set $\mathscr{D}$ and a system $\{\psi_j,\ j \in [N]\}$ that is orthonormal with respect to a measure ν on $\mathscr{D}$ which has square-integrable local coherence,*

$$\sup_{j\in[N]} |\psi_j(u)| \le \kappa(u), \qquad \int_{u\in\mathscr{D}} |\kappa(u)|^2 \nu(u) du = B. \tag{27}$$

We can define the probability measure $\mu(u) = \frac{1}{B}\kappa^2(u)\nu(u)$ *on* $\mathscr{D}$. *Draw* m *sampling points* $T_1, T_2, \dots, T_m$ *independently from the measure* μ, *and consider the matrix* $A \in \mathbb{C}^{m\times N}$ *whose rows are the random samples* $\psi_j(T_k), j \in [N]$. *Consider also the diagonal preconditioning matrix* $\mathscr{P} \in \mathbb{C}^{m\times m}$ *with entries* $p_{k,k} = 1/\mu(T_k)$. *If the number of sampling points*

$$m \gtrsim B^2 s \log^3(s) \log(N), \tag{28}$$

for some $s \gtrsim \log(N)$, *then the following holds with probability exceeding* $1 - N^{-C\log^3(s)}$.

For each $x \in \mathbb{C}^N$, *given noisy measurements* $y = Ax + \sqrt{m}\eta$ *with* $\|\mathscr{P}\eta\|_2 \le \sqrt{m}\varepsilon$, *the approximation*

$$x^{\#} = \arg\min_{z\in\mathbb{C}^N} \|z\|_1 \text{ subject to } \|\mathscr{P}Az - \mathscr{P}y\|_2 \le \sqrt{m}\varepsilon$$

satisfies the error guarantee

$$\|x - x^{\#}\|_2 \lesssim \frac{1}{\sqrt{s}}\|x - x_s\|_1 + \varepsilon.$$

The proof is a simple change-of-measure argument following the lines of standard importance sampling principle:

Proof. Consider the functions $Q_j(u) = \frac{\sqrt{B}}{\kappa(u)}\psi_j(u)$. The system $\{Q_j\}$ is bounded with $\sup_{j\in[N]} \|Q_j\|_\infty \le \sqrt{B}$, and this system is orthonormal on $\mathscr{D}$ with respect to the sampling measure μ:

$$\begin{aligned}
&\int_{u\in\mathscr{D}} \bar{Q}_j(u) Q_k(u) \mu(u) du \\
&= \int_{u\in\mathscr{D}} \left(\frac{1}{\kappa(u)}\bar{\psi}_j(u)\right)\left(\frac{1}{\kappa(u)}\psi_k(u)\right)\left(\kappa^2(u)\nu(u)\right) du \\
&= \int_{u\in\mathscr{D}} \bar{\psi}_j(u)\psi_k(u)\nu(u) du = \delta_{jk}.
\end{aligned} \tag{29}$$

Thus we may apply Proposition 1 to the system $\{Q_j\}$, noting that the matrix of random samples of the system $\{Q_j\}$ may be written as $\mathscr{P}A$.

In the discrete setting where $\{\psi_j\}_{j\in[N]}$ and $\{\phi_k\}$ are rows of unitary matrices Ψ and Φ, and ν is the uniform measure over the set $\mathscr{D} = \{0, \frac{1}{N}, \dots, \frac{N-1}{N}\}$, the integral in condition (27) reduces to a sum,

$$\sup_{k\in[N]} \sqrt{N}|\langle\psi_j, \phi_k\rangle| \le \kappa_j, \quad \frac{1}{N}\sum_{j=1}^{N} \kappa_j^2 = B. \tag{30}$$

This motivates the introduction of the local coherence of an orthonormal basis $\{\phi_j\}_{j=1}^N$ of $\mathbb{C}^N$ with respect to the orthonormal basis $\{\psi_k\}_{k=1}^N$ of $\mathbb{C}^N$:

Definition 2. The local coherence of an orthonormal basis $\{\phi_j\}_{j=1}^N$ of $\mathbb{C}^N$ with respect to the orthonormal basis $\{\psi_k\}_{k=1}^N$ of $\mathbb{C}^N$ is the function $\mu^{loc} = (\mu_j) \in \mathbb{R}^N$ defined coordinate-wise by

$$\mu_j = \sup_{1 \le k \le N} \sqrt{N} |\langle \varphi_j, \psi_k \rangle|.$$

We have the following corollary of Theorem 2.

Corollary 1. *Consider a pair of orthonormal basis (Φ, Ψ) with local coherences bounded by $\mu_j \le \kappa_j$. Let $s \ge 1$, and suppose that*

$$m \gtrsim s \left(\frac{1}{N} \sum_{j=1}^{N} \kappa_j^2 \right) \log^4(N).$$

Select m (possibly not distinct) rows of Φ^* independent and identically distributed from the multinomial distribution on $\{1,2,\ldots,N\}$ with weights $c\kappa_j^2$ to form the sensing matrix $A : \mathbb{C}^N \to \mathbb{C}^m$. Consider also the diagonal preconditioning matrix $\mathscr{P} \in \mathbb{C}^{m \times m}$ with entries $p_{k,k} = \frac{1}{\sqrt{c}\kappa_j}$. Then the following holds with probability exceeding $1 - N^{-C\log^3(s)}$: For each $x \in \mathbb{C}^N$, given measurements $y = Ax + \eta$, with $\|\mathscr{P}\eta\|_2 \le \sqrt{m}\varepsilon$, the approximation

$$x^{\#} = arg \min_{u \in \mathbb{C}^N} \|\Psi^* u\|_1 \text{ subject to } \|y - \mathscr{P}Au\|_2 \le \sqrt{m}\varepsilon$$

satisfies the error guarantee $\|x - x^{\#}\|_2 \lesssim \frac{1}{\sqrt{s}} \|\Psi^* x - (\Psi^* x)_s\|_1 + \varepsilon$.

Remark 2. Note that the local coherence not only influences the embedding dimension m, it also influences the sampling measure. Hence a priori, one cannot guarantee the optimal embedding dimension if one only has suboptimal bounds for the local coherence. That is why the sampling measure in Theorem 2 is defined via the (known) upper bounds κ and $\|\kappa\|_2$ rather than the (usually unknown) exact values μ_{loc} and $\|\mu_{loc}\|_2$, showing that local coherence sampling is *robust with respect to the sampling measure:* suboptimal bounds still lead to meaningful bounds on the embedding dimension.

We now present two applications where local-coherence sampling enables a sampling scheme with sparse recovery guarantees.

Remark 3. The $\log(N)^4$ factor in the required number of measurements, m, can be reduced to a single $\log(N)$ factor if one asks not for *uniform* sparse recovery (of the form "with high probability, this holds for all x") but rather a with-high probability result holding only for a particular x (of the form "for this x, recovery holds with high probability"). See [18] for more details.

Variable-density sampling for compressive sensing MRI

In Magnetic Resonance Imaging, after proper discretization, the unknown image (x_{j_1,j_2}) is a two-dimensional array in $\mathbb{R}^{n\times n}$, and allowable sensing measurements are two-dimensional Fourier transform measurements [1]:

$$\phi_{k_1,k_2} = \frac{1}{n}\sum_{j_1,j_2} x_{j_1,j_2} e^{2\pi i(k_1 j_1 + k_2 j_2)/n}, \quad -n/2+1 \le k_1,k_2 \le n/2.$$

Natural sparsity domains for images, such as discrete spatial differences, are not incoherent to the Fourier basis.

A number of empirical studies, including the very first papers on compressed sensing MRI, observed that image reconstructions from compressive frequency measurements could be significantly improved by variable-density sampling.

Note that lower frequencies are more coherent with wavelets and step functions than higher frequencies. In [21], the local coherence between the two-dimensional Fourier basis and bivariate Haar wavelet basis was calculated:

Proposition 2. *The local coherence between frequency ϕ_{k_1,k_2} and the bivariate Haar wavelet basis $\Psi = (\psi_I)$ can be bounded by*

$$\mu(\phi_{k_1,k_2},\Psi) \lesssim \frac{\sqrt{N}}{(|k_1+1|^2+|k_2+1|^2)^{1/2}}.$$

Note that this local coherence is *almost square integrable independent of discretization size n^2*, as

$$\frac{1}{N}\sum_{j=1}^{N}\mu_j^2 \lesssim \log(n).$$

Applying Corollary 1 to compressive MRI imaging, we then have

Corollary 2. *Let $n \in \mathbb{N}$. Let Ψ be the bivariate Haar wavelet basis and let $\Phi = (\phi_{k_1,k_2})$ be the two-dimensional discrete Fourier transform. Let $s \ge 1$, and suppose that $m \gtrsim s\log^5(N)$. Select m (possibly not distinct) frequencies (ϕ_{k_1,k_2}) independent and identically distributed from the multinomial distribution on $\{1,2,\ldots,N\}$ with weights proportional to the inverse squared Euclidean distance to the origin, $\frac{1}{(|k_1+1|^2+|k_2+1|^2)}$, and form the sensing matrix $A:\mathbb{C}^N \to \mathbb{C}^m$. Then the following holds with probability exceeding $1-N^{-C\log^3(s)}$: for each image $x \in \mathbb{C}^{n\times n}$, given measurements $y = Ax$, the approximation*

$$x^{\#} = arg\min_{u\in\mathbb{C}^{n\times n}} \|\Psi^* u\|_1 \text{ subject to } \|\mathscr{D}y - Au\|_2 \le \varepsilon$$

satisfies the error guarantee $\|x-x^{\#}\|_2 \lesssim \frac{1}{\sqrt{s}}\|\Psi^* x - (\Psi^* x)_s\|_1 + \varepsilon$.

[1] The unknown might also be higher-dimensional, and is often 3-dimensional, but the ideas are analogous and we focus on the 2D example for simplicity.

Remark 4. This result was generalized to multidimensional wavelet and Fourier bases (not just two dimensions as considered above), and to any Daubechies wavelet basis in [20].

Remark 5. One can prove similar guarantees as in (2) using *total variation minimization* reconstruction, see [21, 25].

Sparse orthogonal polynomial expansions

Here we consider the problem of recovering polynomials g from m sample values $g(x_1), g(x_2) \dots, g(x_m)$, with sampling points $x_\ell \in [-1,1]$ for $\ell = 1, \dots, m$. If the number of sampling points is less or equal to the degree of g, then in general such reconstruction is impossible due to dimension reasons. However, the situation becomes tractable if we make a sparsity assumption. In order to introduce a suitable notion of sparsity, we consider the orthonormal basis of Legendre polynomials.

Definition 3. The (orthonormal) Legendre polynomials $P_0, P_1, \dots, P_n, \dots$ are uniquely determined by the following conditions:

- $P_n(x)$ is a polynomial of precise degree n in which the coefficient of x^n is positive,
- the system $\{P_n\}_{n=0}^{\infty}$ is orthonormal with respect to the normalized Lesbegue measure on $[-1,1]$: $\frac{1}{2}\int_{-1}^{1} P_n(x)P_m(x)dx = \delta_{n,m}, \quad n,m = 0,1,2,\dots$

Since the interval $[-1,1]$ is symmetric, the Legendre polynomials satisfy $P_n(x) = (-1)^n P_n(-x)$. For more information see [44].

An arbitrary real-valued polynomial g of degree $N-1$ can be expanded in terms of Legendre polynomials,

$$g(x) = \sum_{j=0}^{N-1} c_j P_j(x), \quad x \in [-1,1]$$

with coefficient vector $c \in \mathbb{R}^N$. The vector is s-sparse if $\|c\|_0 \leq s$. Given a set of m sampling points $(x_1, x_2, \dots, x_m)$, the samples $y_k = g(x_k)$, $k = 1, \dots, m$, may be expressed concisely in terms of the coefficient vector according to

$$y = \boldsymbol{\Phi} c,$$

where $\phi_{k,j} = P_j(x_k)$. If the sampling points $x_1, \dots, x_m$ are random variables, then the matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m \times N}$ is exactly the sampling matrix corresponding to random samples from the Legendre system $\{P_j\}_{j=1}^{N}$. This is not a bounded orthonormal system, however, as the Legendre polynomials grow like

$$|P_n(x)| \leq (n + 1/2)^{1/2}, \quad -1 \leq x \leq 1.$$

Nevertheless the Legendre system does have bounded local coherence. A classic result from [44] follows.

Proposition 3. *For all $n > 0$ and for all $x \in [-1,1]$, $|P_n(x)| < \kappa(x) = 2\pi^{-1/2}(1-x^2)^{-1/4}$. Here, the constant $2\,\pi^{-1/2}$ cannot be replaced by a smaller one.*

Indeed, $\kappa(x)$ is a square integrable function proportional to the Chebyshev measure $\pi^{-1}(1-x^2)^{-1/2}$. We arrive at the following result for Legendre polynomial interpolation as a corollary of Theorem 2.

Corollary 3. *Let $x_1, \dots, x_m$ be chosen independently at random on $[-1,1]$ according to the Chebyshev measure $\pi^{-1}(1-x^2)^{-1/2}dx$. Let Ψ be the matrix with entries $\Psi_{k,j} = \sqrt{\pi/2}(1-x_k^2)^{1/4}P_n(x_k)$. Suppose that*

$$m \gtrsim s\log^3(N).$$

Consider the matrix $A \in \mathbb{C}^{m\times N}$ whose rows are independent random vectors $(\psi_j(X_k))$ drawn from the measure μ. If

$$m \gtrsim B^2 s\log^3(s)\log(N), \tag{31}$$

for some $s \gtrsim \log(N)$, then the following holds with probability exceeding $1 - N^{-C\log^3(s)}$. Let $\mathscr{D} \in \mathbb{C}^{m\times m}$ be the diagonal matrix with entries $d_{k,k} = \frac{1}{\mu(X_k)}$. For each $x \in \mathbb{C}^N$, given noisy measurements $y = Ax + \sqrt{m}\eta$ with $\|\mathscr{D}\eta\|_2 \le \sqrt{m}\varepsilon$, the approximation

$$x^{\#} = arg\min_{u\in\mathbb{C}^N} \|u\|_1 \text{ subject to } \|\mathscr{D}Au - \mathscr{D}y\|_2 \le \sqrt{m}\varepsilon$$

satisfies the error guarantee $\|x - x^{\#}\|_2 \lesssim \frac{1}{\sqrt{s}}\|x - x_s\|_1 + \varepsilon$ where x_s is the best s-term approximation to x.

In fact, more general theorems exist: the Chebyshev measure is a universal sampling strategy for interpolation with any set of orthogonal polynomials [32]. An extension to the setting of interpolation with spherical harmonics, and more generally, to the eigenfunctions corresponding to smooth compact manifolds, can be found in [6, 32], respectively. For extensive numerical illustrations comparing Chebyshev vs. uniform sampling, also for high-dimensional tensor-product polynomial expansions, we refer the reader to [18].

Structured sparse recovery

Often, the prior of sparsity can be refined, and additional *structure* of the support set is known. In the MRI example where one senses with Fourier measurements signals which are sparse in Wavelets, the sparsity level will be higher for higher-order wavelets. One may consider sampling strategies based on a more refined notion of local coherence – based not only on $\mu_j = \sup_{1\le k\le N}\sqrt{N}|\langle\phi_j,\psi_k\rangle|$, but also coherences of sub-blocks $\mu_{j,B_k} = \sup_{k\in B_k}\sqrt{N}|\langle\phi_j,\psi_k\rangle|$. For more information, we refer the reader to the survey article [1] and the references therein.

In fact, we also have more information about the sparsity structure in the setting of function interpolation. It is well known that the smoothness of a function is reflected in the rate of decay of its Fourier coefficients / orthonormal Legendre polynomial coefficients, and vice versa. Thus, smooth functions have directional sparsity in their orthonormal polynomial expansions: low-order and low-degree polynomials are more likely to contribute to the representation. Another way to account for directional sparsity is in the reconstruction method itself. A more general theory of sparse recovery involves *weighted* ℓ_1 minimization as a reconstruction strategy, which serves as a *weighted* sparse prior, and the incorporation of importance sampling there, can be found in [33].

One of the motivating applications of sparse orthogonal polynomial expansions is toward the setting of *Polynomial Chaos expansions* in the area of *Uncertainty Quantification* (UQ), which involves high-dimensional expensive random inputs and modeling the output as having approximately sparse expansion in a tensorized orthogonal polynomial expansion. As shown in [18], in high dimensions, local coherence sampling strategy will depend on how high is the *dimension* compared to the maximal *order* of orthogonal polynomial considered; for higher-order models, Chebyshev sampling is a good strategy; for low-order, high-dimensional problems, uniform sampling outperforms Chebyshev sampling. For a detailed overview and more results, we refer the reader to [18].

Importance sampling in low-rank matrix recovery

Low-rank matrix completion

The task of *low-rank matrix completion* concerns the recovery of a low-rank matrix from a subset of its revealed entries, and nuclear norm minimization has emerged as an effective surrogate for this combinatorial problem. In fact, nuclear norm minimization can recover an arbitrary $n \times n$ matrix of rank r from $\mathcal{O}(nr\log^2(n))$ revealed entries, provided that revealed entries are drawn proportionally to the local row and column coherences (closely related to leverage scores) of the underlying matrix. Matrix completion has been the subject of much recent study due to its application in myriad tasks: collaborative filtering, dimensionality reduction, clustering, non-negative matrix factorization and localization in sensor networks. Clearly, the problem is ill-posed in general; correspondingly, analytical work on the subject has focused on the joint development of algorithms, and sufficient conditions under which such algorithms are able to recover the matrix.

If the true matrix is M with entries M_{ij}, and the set of observed elements is Ω, this method guesses as the completion the optimum of the convex program:

$$\begin{aligned} \min_{X} \quad & \|X\|_* \\ \text{s.t.} \quad & X_{ij} = M_{ij} \text{ for } (i,j) \in \Omega. \end{aligned} \tag{32}$$

where the "nuclear norm" $\|\cdot\|_*$ of a matrix is the sum of its singular values[2]. Throughout, we use the standard notation $f(n) = \Theta(g(n))$ to mean that $cg(n) \leq f(n) \leq Cg(n)$ for some positive constants c, C.

We focus on the setting where matrix entries are revealed from an underlying probability distribution. To introduce the distribution of interest, we first need a definition.

Definition 4. For an $n_1 \times n_2$ real-valued matrix M of rank r with SVD given by $U\Sigma V^\top$, the **local coherences**[3] – μ_i for any row i, and ν_j for any column j - are defined by the following relations

$$\begin{aligned} \left\|U^\top e_i\right\| &= \sqrt{\frac{\mu_i r}{n_1}} \quad , \quad i = 1, \ldots, n_1 \\ \left\|V^\top e_j\right\| &= \sqrt{\frac{\nu_j r}{n_2}} \quad , \quad j = 1, \ldots, n_2. \end{aligned} \tag{33}$$

Note that the μ_i, ν_js are non-negative, and since U and V have orthonormal columns we always have $\sum_i \mu_i r/n_1 = \sum_j \nu_j r/n_2 = r$.

The following theorem is from [13].

Theorem 3. *Let $M = (M_{ij})$ be an $n_1 \times n_2$ matrix with local coherence parameters $\{\mu_i, \nu_j\}$, and suppose that its entries M_{ij} are observed only over a subset of elements $\Omega \subset [n_1] \times [n_2]$. There are universal constants $c_0, c_1, c_2 > 0$ such that if each element (i, j) is independently observed with probability p_{ij}, and p_{ij} satisfies*

$$\begin{aligned} p_{ij} &\geq \min\left\{ c_0 \frac{(\mu_i + \nu_j) r \log^2(n_1 + n_2)}{\min\{n_1, n_2\}} \ , \ 1 \right\}, \\ p_{ij} &\geq \frac{1}{\min\{n_1, n_2\}^{10}}, \end{aligned} \tag{34}$$

then M is the unique optimal solution to the nuclear norm minimization problem (32) with probability at least $1 - c_1(n_1 + n_2)^{-c_2}$.

We will refer to the sampling strategy (34) as *local coherence* sampling. Note that the expected number of observed entries is $\sum_{i,j} p_{ij}$, and this satisfies

$$\begin{aligned} \sum_{i,j} p_{ij} &\geq \max\left\{ c_0 \frac{r \log^2(n_1 + n_2)}{\min\{n_1, n_2\}} \sum_{i,j} (\mu_i + \nu_j), \sum_{i,j} \frac{1}{n^{10}} \right\} \\ &= 2c_0 \max\{n_1, n_2\} r \log^2(n_1 + n_2), \end{aligned}$$

[2] This becomes the trace norm for positive-definite matrices. It is now well recognized to be a convex surrogate for rank minimization.

[3] In the matrix sparsification literature [4, 14] and beyond, the quantities $\left\|U^\top e_i\right\|^2$ and $\left\|V^\top e_j\right\|^2$ are referred to as the *leverage scores* of M.

independent of the coherence, or indeed any other property, of the matrix. Hoeffding's inequality implies that the actual number of observed entries sharply concentrates around its expectation, leading to the following corollary:

Corollary 4. *Let $M = (M_{ij})$ be an $n_1 \times n_2$ matrix with local coherence parameters $\{\mu_i, \nu_j\}$. Draw a subset of its entries by local coherence sampling according to the procedure described in Theorem 3. There are universal constants $c_1', c_2' > 0$ such that the following holds with probability at least $1 - c_1'(n_1+n_2)^{-c_2'}$: the number m of revealed entries is bounded by*

$$m \leq 3c_0 \max\{n_1, n_2\} r \log^2(n_1+n_2),$$

and M is the unique optimal solution to the nuclear norm minimization program (32).

(A) Roughly speaking, the condition given in (34) ensures that entries in important rows/columns (indicated by large local coherences μ_i and ν_j) of the matrix should be observed more often. Note that Theorem 3 only stipulates that an *inequality* relation hold between p_{ij} and $\{\mu_i, \nu_j\}$. This allows for there to be some discrepancy between the sampling distribution and the local coherences. It also has the natural interpretation that the more the sampling distribution $\{p_{ij}\}$ is "aligned" to the local coherence pattern of the matrix, the fewer observations are needed.

(B) Sampling based on local coherences provides close to the optimal number of sampled elements required for exact recovery (when sampled with any distribution). In particular, recall that the number of degrees of freedom of an $n \times n$ matrix with rank r is $2nr(1 - r/2n)$. Hence, regardless how the entries are sampled, a minimum of $\Theta(nr)$ entries is required to recover the matrix. Theorem 3 matches this lower bound, with an additional $O(\log^2(n))$ factor.

(C) Theorem 3 is from [13] and improves on the first results of matrix completion [7, 9, 17, 34] which assumed uniform sampling and incoherence i.e. every $\mu_i \leq \mu_0$ and every $\nu_j \leq \mu_0$ – and an additional *joint incoherence parameter* μ_{str} defined by $\|UV^\top\|_\infty = \sqrt{\frac{r\mu_{str}}{n_1 n_2}}$. The proof of Theorem 3 involves an analysis based on bounds involving the *weighted* $\ell_{\infty,2}$ matrix norm, defined as the maximum of the appropriately weighted row and column norms of the matrix. This differs from previous approaches that use ℓ_∞ or unweighted $\ell_{\infty,2}$ bounds [12, 17]. In some sense, using the weighted $\ell_{\infty,2}$-type bounds is natural for the analysis of low-rank matrices, because the rank is a property of the rows and columns of the matrix rather than its individual entries, and the weighted norm captures the relative importance of the rows/columns.

(D) If the column space of M is incoherent with $\max_i \mu_i \leq \mu_0$ and the row space is arbitrary, then one can randomly pick $\Theta(\mu_0 r \log n)$ rows of M and observe all their entries, and compute the local coherences of the space spanned by these rows. These parameters will be equal to the ν_j's of M with high probability. Based on these values, we can perform non-uniform sampling according to (34) and *exactly* recover M. Note that this procedure does not require any prior knowledge about the

local coherences of M. It uses a total of $\Theta(\mu_0 rn \log^2 n)$ samples. This was observed in [22].

Theorem 3 has some interesting consequences, discussed in detail in [13] and outlined below.

- Theorem 3 can be turned on its head, and used to quantify the benefit of *weighted* nuclear norm minimization over standard nuclear norm minimization, and provide a strategy for choosing the weights in such problems given non-uniformly distributed samples so as to reduce the sampling complexity of weighted nuclear norm minimization to that of standard nuclear norm minimization. In particular, these results can provide exact recovery guarantees for weighted nuclear norm minimization as introduced in [16, 27, 39], thus providing theoretical justification for its good empirical performance.
- Numerical evidence suggests that a two-phase adaptive sampling strategy, which assumes no prior knowledge about the local coherences of the underlying matrix M, can perform on par with the optimal sampling strategy in completing coherent matrices, and significantly outperform uniform sampling. Specifically, [13] considers a two-phase sampling strategy whereby given a fixed budget of m samples, one first draws a fixed proportion of samples uniformly at random, and then draw the remaining samples according to the local coherence structure of the resulting sampled matrix.

References

1. B. Adcock, A. Hansen, B. Roman, The quest for optimal sampling: computationally efficient, structure-exploiting measurements for compressed sensing. arXiv preprint (2014)
2. F. Bach, E. Moulines, Non-asymptotic analysis of stochastic approximation algorithms for machine learning, in *Advances in Neural Information Processing Systems* (2011)
3. L. Bottou, Large-scale machine learning with stochastic gradient descent, in *Proceedings of COMPSTAT'2010*, pp. 177–186, 2010
4. C. Boutsidis, M. Mahoney, P. Drineas, An improved approximation algorithm for the column subset selection problem, in *Proceedings of the Symposium on Discrete Algorithms*, pp. 968–977, 2009
5. L. Brutman, Lebesgue functions for polynomial interpolation: a survey. Ann. Numer. Math. **4**, 111– 127 (1997)
6. N. Burq, S. Dyatlov, R. Ward, M. Zworski, Weighted eigenfunction estimates with applications to compressed sensing. SIAM J. Math. Anal. **44**(5), 3481–3501 (2012)
7. E. Candès, B. Recht, Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717–772 (2009)
8. E.J. Candès, T. Tao, Near optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
9. E. Candès, T. Tao, The power of convex relaxation: near-optimal matrix completion. IEEE Trans. Inf. Theory **56**(5), 2053–2080 (2010)
10. E.J. Candès, T. Tao, J. Romberg, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
11. E.J. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)

12. Y. Chen, Incoherence-optimal matrix completion. arXiv preprint arXiv:1310.0154 (2013)
13. Y. Chen, S. Bhojanapalli, S. Sanghavi, R. Ward, Coherent matrix completion, in *Proceedings of the 31st International Conference on Machine Learning*, pp. 674–682, 2014
14. P. Drineas, M. Magdon-Ismail, M. Mahoney, D. Woodruff, Fast approximation of matrix coherence and statistical leverage. J. Mach. Learn. Res. **13**, 3475–3506 (2012)
15. S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Springer, Berlin, 2013)
16. R. Foygel, R. Salakhutdinov, O. Shamir, N. Srebro, Learning with the weighted trace-norm under arbitrary sampling distributions. arXiv:1106.4251 (2011)
17. D. Gross, Recovering low-rank matrices from few coefficients in any basis. IEEE Trans. Inf. Theory **57**(3), 1548–1566 (2011)
18. J. Hampton, A. Doostan, Compressive sampling of polynomial chaos expansions: convergence analysis and sampling strategies. J. Comput. Phys. **280**, 363–386 (2015)
19. R. Johnson, T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction. Adv. Neural Inf. Process. Syst. **26**, 315–323 (2013)
20. A. Jones, B. Adcock, A. Hansen, Analyzing the structure of multidimensional compressed sensing problems through coherence. arXiv preprint (2014)
21. F. Krahmer, R. Ward, Stable and robust sampling strategies for compressive imaging. IEEE Trans. Image Process. **23**(2), 612–622 (2014)
22. A. Krishnamurthy, A. Singh, Low-rank matrix and tensor completion via adaptive sampling. arXiv preprint. arXiv:1304.4672v2 (2013)
23. M. Lustig, D. Donoho, J. Pauly, Sparse mri: the application of compressed sensing for rapid mri imaging. Magn. Reson. Med. **58**(6), 1182–1195 (2007)
24. M. Lustig, D. Donoho, J. Santos, J. Pauly, Compressed sensing mri. IEEE Signal Process. Mag. **25**(2), 72–82 (2008)
25. D. Needell, R. Ward, Stable image reconstruction using total variation minimization. SIAM J. Imag. Sci. **6**(2), 1035–1058 (2013)
26. D. Needell, N. Srebro, R. Ward, Stochastic gradient descent and the randomized kaczmarz algorithm. arXiv preprint. arXiv:1310.5715 (2013)
27. S. Negahban, M. Wainwright, Restricted strong convexity and weighted matrix completion: optimal bounds with noise. J. Mach. Learn. Res. **13**, 1665–1697 (2012)
28. A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro, Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
29. Y. Nesterov, Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**(2), 341–362 (2012)
30. J. Nocedal, S.J. Wright, *Conjugate Gradient Methods* (Springer, Berlin, 2006)
31. A.B. Owen, *Monte Carlo Theory, Methods and Examples* (2013)
32. H. Rauhut, R. Ward, Sparse Legendre expansions via ℓ_1-minimization. J. Approx. Theory **164**, 517–533 (2012)
33. H. Rauhut, R. Ward, Interpolation via weighted ℓ_1 minimization. arXiv preprint. arXiv:1308.0759 (2013)
34. B. Recht, A simpler approach to matrix completion. arXiv preprint. arXiv:0910.0651 (2009)
35. P. Richtárik, M. Takáč, Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Math. Program. **144**(1), 1–38 (2014)
36. H. Robbins, S. Monrow, A stochastic approximation method. Ann. Math. Stat. **22**(22), 400–407 (1951)
37. N.L. Roux, M. Schmidt, F. Bach, A stochastic gradient method with an exponential convergence rate for finite training sets. Adv. Neural Inf. Process. Syst. **25**, 2672–2680 (2012)
38. M. Rudelson, R. Vershynin, On sparse reconstruction from Fourier and Gaussian measurements. Commun. Pure Appl. Math. **61**, 1025–1045 (2008)
39. R. Salakhutdinov, N. Srebro, Collaborative filtering in a non-uniform world: learning with the weighted trace norm. arXiv preprint. arXiv:1002.2780 (2010)
40. S. Shalev-Shwartz, N. Srebro, Svm optimization: inverse dependence on training set size, in *Proceedings of the 25th International Conference on Machine Learning*, pp. 928–935, 2008

41. S. Shalev-Shwartz, T. Zhang, Proximal stochastic dual coordinate ascent. arXiv:1211.2772 (2012)
42. S. Shalev-Shwartz, T. Zhang, Stochastic dual coordinate ascent methods for regularized loss minimization. J. Mach. Learn. Res. **14**, 567–599 (2013)
43. N. Srebro, K. Sridharan, A. Tewari, Smoothness, low noise and fast rates, in *Advances in Neural Information Processing Systems* (2010)
44. G. Szegö, *Orthogonal Polynomials* (American Mathematical Society, Providence, RI, 1939)
45. L. Xiao, T. Zhang, A proximal stochastic gradient method with progressive variance reduction. SIAM J. Optim. **24**(4), 2057–2075 (2014)
46. P. Zhao, T. Zhang, Stochastic optimization with importance sampling. arXiv:1401.2753 (2014)

Part XV
Signal Processing and Sampling

This part contains four chapters devoted to applications of harmonic analysis to signal processing and sampling theory. The contributions are written by leading experts in this field from academia.

The first chapter of this part is by STEPHEN D. CASEY who develops a numerically attractive method to partition the time-frequency plane. He uses bounded partitions of unity to obtain adaptive coverings of the phase plane. The corresponding bounded adaptive partition of unity gives rise to an orthonormal basis for the Paley-Wiener space of bandlimited functions. Next the author develops an almost orthogonal system for the same space of bandlimited functions using the biorthogonal systems theory. The last part of this chapter develops a signal adaptive frame theory. The main application of this framework is analog-to-digital conversion of ultra-wideband signals and software-defined radios.

DAVID WALNUT, GÖTZ E. PFANDER, and THOMAS KAILATH review the identification theory of a class of time-varying linear systems. Specifically, the authors consider the class of sampling operators, namely the linear operators whose Kohn-Nirenberg symbols are bandlimited. The chapter begins with a very nice historical note on early time-frequency doubly dispersive communication channel systems (the RAKE receiver). Next the authors present the more recent results on sampling of operators using periodically weighted delta-trains and the Zak transform. The last part of the chapter introduces novel results on operator sampling in higher dimensions. The authors discuss also connections with stochastic operators and review implications to MIMO systems.

AKRAM ALDROUBI, ILYA KRISHTAL, and ERIC WEBER develop a novel mathematical framework for time-space sampling, called dynamical sampling. A system state is characterized by a space-dependent function that changes over time by the action of a family of known time-evolution operators. At each time instant only a compressed state measurement is observed. Additionally the measurement operator is allowed to vary over time. The authors develop in details the case of uniform space-subsampling for time-invariant systems. They obtain necessary and sufficient conditions for perfect reconstruction in such cases. The two key ingredients of this theory are the Poisson summation formula and a well-mixing property of the evolution operator. Next the authors extend this construction to a special case of space-nonuniform sampling. In the final sections of this chapter, the authors connect dynamical sampling to an unsupervised system identification problem followed by a brief discussion of extensions to an infinite dimensional setting.

In the last chapter of this part, ALIAKSEI SANDRYHAILA and JELENA KOVAČEVIĆ describe a signal processing framework for signals defined on weighted line graphs. Two signal processing methods are known in literature for graph-supported signals: one method is based on the adjacency matrix of the underlying graph, the other method is based on the graph Laplacian matrix. Following previous results of the first author, this chapter develops further the adjacency matrix-based approach. In particular they present definitions of fundamental concepts of signals, filters, z-transform, Fourier transform, frequency, and spectrum for the class of weighted line graphs. They also present some applications to signal representation and compression using fast algorithms.

Finite Dimensional Dynamical Sampling: An Overview

Akram Aldroubi, Ilya Krishtal, and Eric Weber

Abstract Dynamical sampling is an emerging paradigm for studying signals that evolve in time. In this chapter we present many of the available results pertaining to dynamical sampling in the finite dimensional setting. We also provide a brief survey of the latest results in the infinite dimensional setting.

Key words: Sampling and reconstruction, evolution systems, sampling schemes, system identification

Introduction

The typical sampling and reconstruction problem consists of recovering a function f from its samples $f(X) = \{f(x_j)\}_{x_j \in X}$. There are many situations in which the function f is an initial distribution that is evolving in time under the action of a family of evolution operators $\{A_t\}_{t \in [0,\infty)}$:

$$f_t(x) = (A_t f)(x). \tag{1}$$

The standard approaches to solving the reconstruction problem, however, are not designed to take into account this time dependency [4, 10, 11, 15, 17, 18, 20, 21,

A. Aldroubi (✉)
Vanderbilt University, Nashville, TN, USA
e-mail: akram.aldroubi@vanderbilt.edu

I. Krishtal
Northern Illinois University, DeKalb, IL, USA
e-mail: krishtal@math.niu.edu

E. Weber
Iowa State University, Ames, IA, USA
e-mail: esweber@iastate.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_9

and references therein]. As a result, using the standard techniques may lead to an unnecessary bloating of the sampling set $X = \{x_j\}$, create a deluge of data, and drive up the costs of data acquisition and processing. In some cases obtaining the samples at a sufficient rate at time $t = 0$ may not even be possible. Recently, there started to develop a new mathematical framework which allows one to utilize not only the spatial samples $f(X)$ but also the temporal samples $f_t(X_t)$ to recover f, A_t, and f_t, and, at the same time, keep the sampling procedure manageable. In this chapter we present an overview of this new mathematical framework, which we call *Dynamical Sampling* [1, 2, 6–9]. Various dynamical sampling problems exhibit features that are similar to many other fundamental problems: deconvolution [25, 26], filter banks [22, 24], sampling and reconstruction in shift-invariant spaces [3–6, 23], super-resolution [13, 19], etc. However, even in the most basic cases, the dynamical sampling problems are different and necessitate new theoretical and algorithmic techniques.

In general, we consider the problem of spatiotemporal sampling in which an initial state f of an evolution process $\{f_t\}_{t\geq 0}$ is to be recovered from a set of samples $\{f_t(X_t)\}_{t\in\mathscr{T}}$ at different time levels, i.e., $t \in \mathscr{T} = \{t_0 = 0, t_1, \ldots, t_N\}$. Typical evolution processes are driven by well-studied families of evolution operators as in (1). A common example is provided by diffusion and is modeled by the heat equation. Sampling is done by sensors or measurement devices that are placed at various locations and can be activated at different times. Clearly for the problem to be well posed (outside of a finite dimensional context), certain assumptions on f are necessary. A standard assumption (consistent with the nature of signals) is that f belongs to a reproducing Kernel Hilbert space (RKHS) such as a Paley-Wiener space or some other shift invariant spaces (SIS) V [16, 23]. The first general problem of dynamical sampling can be stated as follows

Problem 1 (Spaciotemporal trade-off). Assume $f \in V$ satisfies (1) for some family of evolution operators $\{A_t\}$, $t \geq 0$. Describe all spatiotemporal sampling sets $(\mathscr{X}, \mathscr{T}) = \{X_t,\ t \in \mathscr{T}\}$ such that any $f \in V$ can be stably recovered from the samples $f_t(X_t)$, $t \in \mathscr{T}$.

The name of the above problem [7] comes from the fact that in many cases it is possible to provide the same information about the initial state from a reduced number of devices activated more frequently. In Section "Time-space trade-off in dynamical sampling" we provide several examples illustrating this idea.

Another important problem arises when the evolution operators are themselves unknown (or partially unknown).

Problem 2 (Unsupervised system identification). Assume $f \in V$ satisfies (1) for an unknown family of evolution operators $\{A_t\}$, $t \geq 0$. Describe all spatiotemporal sampling sets $(\mathscr{X}, \mathscr{T}) = \{X_t,\ t \in \mathscr{T}\}$ and classes of evolution operator families such that the family $\{A_t\}$ or its key parameters can be identified from almost any $f \in V$.

In Section "Unsupervised system identification" we describe a few results which provide a solution to a few special cases of the above problem.

Time-space trade-off in dynamical sampling

In this section we discuss various instances of Problem 1 in a finite dimensional setting.

Let $x \in \mathbb{C}^d$ and A be a $d \times d$ invertible matrix with complex entries. We seek to recover vector x from subsampled versions of the vectors x, Ax, A^2x, etc. More precisely, we let $S(\Omega_n)$ be diagonal idempotent matrices so that $s_{ii} = 1$ if and only if $i \in \Omega_n \subseteq \{1,\dots,d\}$, and

$$y_n = S(\Omega_n)A^{n-1}x, \quad n = 1,\dots,N. \tag{2}$$

We would like to know under which conditions we can recover x from y_n, $n = 1,\dots,N$, or, in other words, what information about x, Ax, $\dots$, $A^{N-1}x$, we need in order to make the recovery possible. By $x \in \mathbb{C}^d$, we model an unknown spatial signal at time $t = 0$, and the matrix A represents an evolution operator so that $A^n x$ is the signal at time $t = n$. Then the vectors y_n, $n = 1,\dots,N$, give the samples of the evolving system at time $t = n-1$ at a (possibly) reduced number of locations (given by the sum of the ranks of the matrices $S(\Omega_n)$). We typically assume that $\operatorname{rank}(S(\Omega_n)) < d$, $n = 1,\dots,N$ so that at any time $t = n$ the signal is undersampled and cannot be recovered. This situation arises when there is a restriction on sampling locations or when we would like to keep at a minimum the information we need to sample and store. The latter would reduce the number of measuring devices and, thus, make the sampling process cheaper.

We write the problem in the following matrix form

$$\mathbf{y} = \mathbb{A}x, \tag{3}$$

where $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$ and $\mathbb{A} = \begin{pmatrix} S(\Omega_1) \\ S(\Omega_2)A \\ \vdots \\ S(\Omega_N)A^{N-1} \end{pmatrix}$ is an $Nd \times d$ matrix which we call the *dynamical sampling matrix*. The choice of the sets Ω_k, $k = 1,\dots,N$, will be referred to as the *dynamical sampling procedure*. Thus, the first dynamical sampling problem is to establish conditions under which this procedure is *admissible*, i.e., which would ensure that the matrix $\mathbb{A}$ has full rank d. In this case $\mathbb{A}$ has a left inverse and the recovery of x is possible.

The above linear algebraic formulation of the problem can be restated in terms of frame theory as follows: given a frame Φ for $\mathbb{C}^d$ that consists of all rows of matrices I, A, $\dots$, A^{N-1}, describe all subsets of Φ that are themselves frames for $\mathbb{C}^d$. A related problem is to describe all matrices A for which a fixed dynamical sampling procedure is admissible.

Example 1 (Sampling at one node). Assume that $\Omega_k = \{j\}$ for all $k = 1,\dots,d$, and some $j \in \{1,\dots,d\}$. In other words, we would like to recover the original signal x from its temporal samples at a single spatial location. One would expect this to be possible only if the system is "well mixed", and, in fact, in some sense this is

sufficient. To see this, let us assume that $A = UDU^*$ is positive definite and $U = (u_{jk})$ is the unitary that diagonalizes A so that D is a diagonal matrix with eigenvalues $\lambda_1, \ldots \lambda_d$. Then the reduced dynamical sampling matrix $\mathbb{A}_r$ obtained from $\mathbb{A}$ by eliminating the zero rows satisfies

$$\mathbb{A}_r = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \lambda_1 & \lambda_2 & \ldots & \lambda_d \\ \vdots & \vdots & \ldots & \vdots \\ \lambda_1^{d-1} & \lambda_2^{d-1} & \ldots & \lambda_d^{d-1} \end{pmatrix} \begin{pmatrix} u_{j1} & 0 & \ldots & 0 \\ 0 & u_{j2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & u_{jd} \end{pmatrix} U^*.$$

Since the first of the matrices in the above product is Vandermonde, $\mathbb{A}_r$ is invertible if and only if all eigenvalues of A are distinct and the jth row of the "mixing" matrix U has no zero entries.

The case of sampling at two nodes already presents a far less trivial problem. We will describe results related to this problem in Section "Non-uniform sampling in the general case". Before doing that we consider other, slightly simpler, special cases of the dynamical sampling problem.

Uniform subsampling in invariant evolution systems

In practice, one of the most important cases of the dynamical sampling problem is represented by (spatially) invariant evolution systems in which the matrix A is circular and the subsampling is uniform and independent of n. In this case, the matrix A represents the (circular) convolution operator with a fixed vector $a \in \mathbb{C}^d$ and $S(\Omega_n)$ is an operator of subsampling by some fixed factor $m \in \mathbb{N}$. Thus, $\Omega_n = \{mx : x = 0, \ldots, \frac{d}{m} - 1\}$ for $n = 1, \ldots, N$, where we assume $m|d$, and we will denote $S(\Omega_n) = S_m$. This special case will be referred to as the *uniform invariant dynamical sampling problem*. There, a vector $x \in \mathbb{C}^d$ representing the signal at time $t = 0$ is sampled only at a fraction $J = d/m$ of its components, and subsequently the vectors $A^{n-1}x$, $n = 2, \ldots, N$, are sampled at the same locations. It is not difficult to see that, in order to recover x we would need a minimum of m time levels so that $N \geq m$. Note that the number of sampling devices that are needed for measurements is reduced from d to J, but the devices have to be activated m times more frequently.

For technical reasons we let $N = m$, $d = 2K + 1$, and assume that J is an integer (so that d, m, and J are odd). Then the (k,k) entry of the matrix S_m equals 1 if m divides $K + 1 - k$ and is 0 otherwise. Clearly, in practice, any reasonable model can be tweaked to satisfy these conditions.

The following proposition is the key to the solution of the dynamical sampling problem in this special case. In its formulation, we shall use the notation $\hat{a} = \mathbf{F}_d a$ for the d-dimensional discrete Fourier transform (DFT) of a:

$$\hat{a}(k) = (\mathbf{F}_d a)(k) = \sum_{\ell=0}^{d-1} a(\ell) e^{-\frac{2\pi i k \ell}{d}}, \quad k = 1, \ldots, d.$$

The proof is based on the Poisson summation formula

$$(S_m z)^\wedge(k) = \frac{1}{m}\sum_{\ell=0}^{m-1} \hat{z}(k+J\ell), \quad k = 1,\dots,d,\ z \in \mathbb{C}^d. \tag{4}$$

The proof of this result is contained in [7]. A more general version of this proposition below is Theorem 2 in the following section, which we shall prove there.

Proposition 1 ([7]). *A uniform dynamical sampling procedure in an invariant problem is admissible if and only if the $J = d/m$ matrices*

$$\mathscr{A}_m(k) = \begin{pmatrix} 1 & 1 & \dots & 1 \\ \hat{a}(k) & \hat{a}(k+J) & \dots & \hat{a}(k+(m-1)J) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{a}^{(m-1)}(k) & \hat{a}^{(m-1)}(k+J) & \dots & \hat{a}^{(m-1)}(k+(m-1)J) \end{pmatrix}, \tag{5}$$

$k = 0,\dots,J-1$, are invertible.

Since each matrix $\mathscr{A}_m(k)$ in (5) is a Vandermonde matrix, it is invertible if and only if the values $\{\hat{a}(k+\ell J) : \ell = 0,\dots,m-1\}$ are distinct. If some of these values coincide, the signal x cannot be recovered unless we take extra spatial samples.

The procedure that allows one to prescribe which extra spatial samples may be taken is outlined in [7]. For example, when the kernel a is real symmetric and $\hat{a}$ is strictly monotonic on $\{0,\dots,K\}$, all matrices $\mathscr{A}_m(k)$, $k = 1,\dots,J-1$, are invertible but the matrix $\mathscr{A}_m(0)$ is not. Such kernels are realistic in applications because they correspond to isotropic symmetric processes such as diffusion in isotropic media.

The following result characterizes extra sampling sets which make the recovery possible in this case.

Theorem 1 ([7]). *Consider an invariant dynamical sampling problem with a real symmetric kernel a such that $\hat{a}$ is strictly monotonic on $\{0,\dots,K\}$. Then the uniform dynamical sampling procedure augmented by a set $\Omega_0 \subseteq \{1,\dots,d\}$ is admissible if and only if Ω_0 contains a set of cardinality $\frac{m-1}{2}$ such that no two of its elements are m-congruent or have a sum divisible by m.*

A natural choice of Ω_0 in the above theorem is

$$\Omega_0 = \left\{-K, -K+1, \dots, -K+\frac{m-1}{2}, K-\frac{m-1}{2}, \dots, K-1, K\right\}.$$

Alternatively, we may assume that $\operatorname{supp} x \subseteq \left[-K+\frac{m-1}{2}, K-\frac{m-1}{2}\right]$.

It can also be shown [7] that if the vector x is $(J-1)$-sparse, that is, has at most $J-1$ nonzero components, then it is completely recoverable via the uniform dynamical sampling procedure (without the extra samples) in an invariant dynamical sampling problem with a real symmetric kernel a such that $\hat{a}$ is strictly monotonic on $\{0,\dots,K\}$.

Generically, the number of extra samples ν needed for the recovery in the uniform case satisfies $\nu \ll d$, that is the oversampling factor is typically negligible. It is also clear from the Vandermonde structure of the matrices (5) that adding more time samples at the same locations provides no additional information about x, thus justifying our choice of $N = m$. On the other hand, in the presence of noise and once an appropriate set Ω_0 is chosen, additional time samples may be used to improve stability of the estimation of x. We refer to [7] for additional information and stability estimates for the uniform dynamical sampling procedure.

Nonuniform sampling in a special case

In this section we consider a more general case of Problem 1 which can be solved by a method that is very similar to the one outlined in the previous section. Here the subsampling is no longer uniform but the class of admissible evolution operators remains fairly small and depends heavily on the set of measurement locations Ω, or conversely determines what Ω must be.

As we hinted in Example 1 and remarks preceding Proposition 1 the key ingredients in solving some cases of Problem 1 are the Poisson summation formula (4) and the well-mixing property of the evolution operator A. Motivated by these observations we make the following definitions.

For $k \in \mathbb{N}$ we let $\mathbf{P}_k$ be the $k \times k$ matrix given by

$$\mathbf{P}_k = \frac{1}{k} \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}.$$

We say that a $d \times d$ matrix P is a *Poisson projection* if P is a block diagonal matrix of the form

$$P = \begin{pmatrix} \mathbf{P}_{k_1} & 0 & \dots & 0 \\ 0 & \mathbf{P}_{k_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{P}_{k_\ell} \end{pmatrix} \tag{6}$$

with $k_1 + k_2 + \ldots + k_\ell = d$. We shall denote by $\mathscr{P}$ the set of all Poisson projections.

Clearly, P in (6) is an orthogonal projection of rank ℓ. Hence, it is unitarily equivalent to a subsampling projection S_Ω with $|\Omega| = \ell$.

Definition 1. Let $P \in \mathscr{P}$ and U be a unitary matrix such that $P = US_\Omega U^*$ for some $\Omega \subseteq \{1, \ldots, d\}$. We say that a $d \times d$ matrix A is *well mixing* with respect to P and U if A is diagonalized by U^*, i.e., $A = U^*DU$ for some diagonal matrix

$$D = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_d \end{pmatrix}.$$

Given A, P, U, and Ω as in the above definition, an (A,P,U) *dynamical sampling problem* consists of finding a vector $x \in \mathbb{C}^d$ from the samples $S_\Omega x$, $S_\Omega Ax$, …, $S_\Omega A^N x$.

Theorem 2. *Assume that $P \in \mathscr{P}$, U is a unitary matrix such that $P = US_\Omega U^*$ for some $\Omega \subseteq \{1,\dots,d\}$, and A is well mixing with respect to P and U. Then the (A,P,U) dynamical sampling problem is solvable for $N = \max\{k_r, r = 1,\dots,\ell\} - 1$ if and only if the eigenvalues of A satisfy $\lambda_i \neq \lambda_j$ for all $\sum_{m=0}^{r-1} k_m < i < j \leq \sum_{m=1}^{r} k_m$, $r = 1,2,\dots,\ell$ with $k_0 = 0$.*

Proof. Observe that $S_\Omega A^n x = U^* P D^n U x$, $n \in \mathbb{N}$. Thus, the matrix $\mathbb{A}$ in (3) can be written in the form $U^* \mathscr{P} U$, where

$$\mathscr{P} = \begin{pmatrix} P \\ PD \\ \vdots \\ PD^N \end{pmatrix}.$$

Hence, the (A,P,U) problem is solvable if and only if the matrix $\mathscr{P}$ has a left inverse. The matrix $\mathscr{P}$ has the following as a $d \times d$ submatrix

$$\mathscr{P}_r = \begin{pmatrix} 1 & 1 & \dots & 1 & & & & & & & & \\ \lambda_1 & \lambda_2 & \dots & \lambda_{k_1} & & & & & & & & \\ \vdots & \vdots & \ddots & \vdots & & 0 & & \dots & & 0 & & \\ \lambda_1^{k_1-1} & \lambda_2^{k_1-1} & \dots & \lambda_{k_1}^{k_1-1} & & & & & & & & \\ & & & & 1 & 1 & \dots & 1 & & & & \\ & & & & \lambda_{k_1+1} & \lambda_{k_1+2} & \dots & \lambda_{k_1+k_2} & & & & \\ & 0 & & & \vdots & \vdots & \ddots & \vdots & \dots & & 0 & \\ & & & & \lambda_{k_1+1}^{k_2-1} & \lambda_{k_1+2}^{k_2-1} & \dots & \lambda_{k_1+k_2}^{k_2-1} & & & & \\ & \vdots & & & & \vdots & & & \ddots & & \vdots & \\ & & & & & & & & 1 & 1 & \dots & 1 \\ & & & & & & & & \lambda_{d-k_\ell+1} & \lambda_{d-k_\ell+2} & \dots & \lambda_d \\ & 0 & & & & 0 & & \dots & \vdots & \vdots & \ddots & \vdots \\ & & & & & & & & \lambda_{d-k_\ell+1}^{k_\ell-1} & \lambda_{d-k_\ell+2}^{k_\ell-1} & \dots & \lambda_d^{k_\ell-1} \end{pmatrix}$$

after rescaling the rows appropriately. The hypotheses regarding the λ_js guarantee that each Vandermonde block is invertible. Hence, $\mathscr{P}$ has full rank, and thus has a left inverse.

In Theorem 2, we began with a sampling set Ω and determined for which A the (A,P,U) dynamical sampling problem is solvable. The proof of Theorem 2 also gives a way of reversing this process, i.e., starting with A which satisfies the appropriate hypotheses, we can determine a sampling set Ω such that the (A,P,U) dynamical sampling problem is solvable. To do so, we take a closer look at a Poisson projection. If $P \in \mathscr{P}$, we can write P as a sum of rank one projections (or tensor products):

$$P = \sum_{r=1}^{\ell} \mathbf{v}_r \cdot \mathbf{v}_r^T, \tag{7}$$

where $\mathbf{v}_r$ is the column vector whose qth coordinate is $\frac{1}{\sqrt{k_r}}$ if $\sum_{s=0}^{r-1} k_s < q \leq \sum_{s=0}^{r} k_s$ and is 0 otherwise. Thus, $\mathbf{v}_r$ has the form $\frac{1}{\sqrt{k_r}} \begin{pmatrix} 0 \dots 0 & 1 \dots 1 & 0 \dots 0 \end{pmatrix}^T$. We will call a vector of this form a *Poisson vector*, and we will call $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_\ell\}$ a *Poisson ensemble* if they are Poisson vectors, and the corresponding sum of rank one projections gives a Poisson projection. Note that a Poisson ensemble is an orthonormal set.

Theorem 3. *Suppose A is a $d \times d$ matrix. Suppose also*

1. U is a unitary matrix such that $UAU^* = D$ (D diagonal);
2. the column vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_d\}$ of U contain a Poisson ensemble $\{\mathbf{u}_{j_1}, \dots, \mathbf{u}_{j_\ell}\}$;
3. P is the Poisson projection obtained from $\{\mathbf{u}_{j_1}, \dots, \mathbf{u}_{j_\ell}\}$ as in (7).

Then if we choose $\Omega = \{j_1, \dots, j_\ell\}$, we have that $A = U^*DU$ and $P = US_\Omega U^*$. In other words, A is well mixing with respect to P and U.

Proof. Let $\Omega = \{j_1, \dots, j_\ell\}$. Note that for the Poisson projection

$$P = \sum_{j \in \Omega} \mathbf{u}_j \cdot \mathbf{u}_j^T,$$

we have that $P\mathbf{u}_j = \mathbf{u}_j$ if $j \in \Omega$, and is 0 if $j \notin \Omega$. Therefore, the jth column of PU is $\mathbf{u}_j$ if $j \in \Omega$ and 0 otherwise, from which it follows that U^*PU is a diagonal matrix such that the jth diagonal entry is 1 if $j \in \Omega$ and 0 otherwise. Thus,

$$U^*PU = S_\Omega$$

as required by Theorem 2.

Corollary 1. *Suppose A is an invertible $d \times d$ matrix with distinct eigenvalues and satisfies the conditions of Theorem 3. Then the (A,P,U) dynamical sampling problem is solvable.*

Proof. We choose Ω as prescribed in Theorem 3 and apply Theorem 2.

Example 2. Suppose the diagonalizer U^* for the matrix A has the form

$$U = \begin{pmatrix} * & 0 & * & * & \frac{1}{\sqrt{3}} \\ * & 0 & * & * & \frac{1}{\sqrt{3}} \\ * & 0 & * & * & \frac{1}{\sqrt{3}} \\ * & \frac{1}{\sqrt{2}} & * & * & 0 \\ * & \frac{1}{\sqrt{2}} & * & * & 0 \end{pmatrix}.$$

Then we choose P as

$$P = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

We have

$$U^*PU = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and so $\Omega = \{2,5\}$.

Nonuniform sampling in the general case

In this subsection we present two results about the necessary and sufficient conditions for the solvability of Problem 1 in a general finite dimensional setting. The proofs of these results appear in [9].

Assume that a $d \times d$ matrix A is such that $A^* = B^{-1}DB$, where B is a $d \times d$ invertible matrix, D is a diagonal matrix of the form

$$D = \begin{pmatrix} \lambda_1 I_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 I_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k I_k \end{pmatrix}, \tag{8}$$

and I_k is an $\ell_k \times \ell_k$ identity matrix for some ℓ_k.

Thus, A^* is a diagonalizable matrix with eigenvalues $\{\lambda_1, \ldots, \lambda_k\}$, and with corresponding eigenvectors being the column vectors of B.

Here we use (arbitrary) irregular sampling sets $\Omega \subset \{1, \ldots, d\}$ and may use a different number of time samples at each point $i \in \Omega$. For each $i \in \Omega$ we denote by L_i the number of time samples we take. Our collected data for each sampling

point $i \in \Omega$ consists of $\{x(i), Ax(i), \ldots, A^{L_i}x(i)\}$. Our goal is to find necessary and sufficient conditions on Ω, L_i, and A for the recovery of x. We reorganize the data by letting $y_1(j) = x(j)$, $j \in \Omega_1 = \Omega$; $y_2(j) = Ax(j)$, $j \in \Omega_2$, $\Omega_2 = \{j \in \Omega : L_j \geq 1\}$; and, in general, $y_k(j) = A^{k-1}x(j)$, $j \in \Omega_k$, $\Omega_k = \{j \in \Omega : L_j \geq k-1\}$. It is clear that in this case we have $\Omega = \Omega_1 \supseteq \cdots \supseteq \Omega_{L_{\max}}$ where $L_{\max} = \max\{L_i : i \in \Omega\}$. Thus, we need to solve the system $\mathbf{y} = \mathbb{A}x$ described by (3) with the aforementioned Ω_j, $j = 1, \ldots, L_{\max}$. For this system to be solvable, the rows of $I, A, \ldots, A^{L_{\max}}$ corresponding to $\Omega_1, \ldots, \Omega_L$, must form a frame for $\mathbb{C}^d$. Equivalently, the columns of $I, A^*, \ldots, (A^*)^{L_{\max}}$ corresponding to $\Omega_1, \ldots, \Omega_{L_{\max}}$, must form a frame for $\mathbb{C}^d$. Letting b_i denote the column vector of B, and using the fact that frames of $\mathbb{C}^d$ remain frames if an invertible transformation is applied to them, we observe that solving (3) is equivalent to having the set $\{D^n b_i : i \in \Omega, n = 0, \ldots, L_i\}$ be a frame for $\mathbb{C}^d$. Let f_D^b be the minimal polynomial in D that annihilates b, and let r_i be the degree of $f_D^{b_i}$, $i \in \{1, \ldots, d\}$. In the following theorem we denote by P_j, $j = 1, \ldots, k$, the projection operators with $d \times d$ matrices given by

$$P_j = \begin{pmatrix} 0 & & & & \\ & 0 & & & \\ & & \ddots & & \\ & & & I_j & \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix}.$$

Theorem 4 ([9]). *Let $\{b_i : i \in \Omega\}$ be the column vectors of B corresponding to Ω. Then*

1. *If x can be recovered from $\mathbf{y}$ then $\{P_j b_i : i \in \Omega\}$ is a frame for $P_j(\mathbb{C}^d)$, $j = 1, \ldots k$.*
2. *If $\{P_j b_i : i \in \Omega\}$ is a frame for $P_j(\mathbb{C}^d)$, $j = 1, \ldots k$ and $L_i \geq r_{i-1}$ for every $i \in \Omega$, then x can be recovered from $\mathbf{y}$.*

If we fix $L \leq L_{\max}$ and let $\Omega_j = \Omega$ for every $j = 1, \ldots, L$, we obtain the following result.

Theorem 5 ([9]). *Let $\{b_i : i \in \Omega\}$ be the column vectors of B corresponding to Ω and let L be any fixed integer. Then x can be recovered from $\mathbf{y}$ if and only if $\{P_j b_1, \ldots, P_j b_m\}$ form a frame for $P_j(\mathbb{C}^d)$, $j = 1, \ldots k$, and $\{D^L b_i : i \in \Omega\} \subset span(E)$, where $E = \bigcup_{i \in \Omega} \{b_i, Db_i, \ldots, D^{L-1} b_i\}$.*

Note that conditions of Theorem 5 imply that if x can be recovered, the cardinality of Ω (minimal number of measurement devices) may not be smaller than the (largest) multiplicity of an eigenvalue of A. Applying Theorem 5 to Example 1, we can deduce that to recover x from sampling at one point j it is necessary and sufficient that $u_{ji} \neq 0$ for $i = 1, \ldots, d$ and that $L \geq d$, as we stated in the example. It is also not hard to see that Proposition 1 and Theorem 2 are special cases of Theorem 5.

We also note that similar results for non-diagonalizable matrices A can be found in [9].

Unsupervised system identification

In this section we discuss a few instances of Problem 2 in a finite dimensional setting. In particular, given the samples (2) with both x and A unknown, we would like to recover as much information about them as we can, given various priors. As in the previous section, we first discuss the results for invariant systems in case of uniform subsampling and then exhibit a more general result.

Filter recovery in the invariant uniform case.

Recall that in this setting the evolution operator A is given by a (circular) convolution matrix so that $Ax = a * x$, $a,x \in \mathbb{C}^d$. In the uniform sampling case, our problem consists of finding both a and x from partial observations

$$y_\ell = S_m A^\ell x,\ \ell = 0,\dots,N, \tag{9}$$

of the evolving signal x, where m is, as before, an odd integer that divides d and

$$S_m : \ell^2(\mathbb{Z}_d) \to \ell^2(\mathbb{Z}_d), \quad (S_m x)(n) = \delta_{(n \mod m),0} x(n), \tag{10}$$

is an operator of subsampling by a factor of m. Also as before, for computational simplicity we assume that $J = d/m$ is odd.

Theorem 6 ([6]). *Assume also that the unknown evolution filter $a \in \mathbb{R}^d$ is such that its Fourier transform $\hat{a} \in \mathbb{C}^d$ is nonvanishing, real, and strictly decreasing on $\{0,\dots,\frac{m-1}{2}\}$. Then for almost every $x \in \mathbb{C}^d$ the filter a can be recovered from the measurements y_ℓ defined in (9) with $N \geq 2m-1$.*

The proof is based on a nonlinear method which is a generalization of the classical Prony's method [12] for finding an s-sparse vector from $2s$ of its consecutive Fourier coefficients. Our method allows us to first find the spectrum of A, i.e., the range of $\hat{a}$ and then use other assumptions to recover $\hat{a}$ completely. In the next subsection we provide a result about the spectrum recovery in a more general case.

Spectrum recovery of general evolution operators

As in Subsection "Non-uniform sampling in the general case", we let $b = b_i$ be the ith column vector of B where $A^* = B^{-1}DB$, and let r_i be the degree of the minimal polynomial in D that annihilates $b = b_i$.

Theorem 7 ([6]). *Let b_i be the ith column vector of B for some $i = 1,\dots,d$ and let $\Lambda = \{j : P_j b_i \neq 0\}$. Then for almost every $x \in \mathbb{C}^d$ the subset $\Theta = \{\lambda_j : j \in \Lambda\}$ of the*

spectrum of A can be recovered from the measurements $\{A^k x(i) : k = 0, \ldots, 2r_i - 1\}$. *In particular, if the sampling set* Ω *is such that* $\{P_j b_i : i \in \Omega\}$ *is a frame for* $P_j(\mathbb{C}^d)$, *then the spectrum of A can be recovered from* $\{A^k x(i) : i \in \Omega, k = 0, \ldots, 2r_i - 1\}$ *for almost every* $x \in \mathbb{C}^d$.

For the case in which we stop at a fixed time level L with $L < \max\{r_i - 1 : i \in \Omega\}$ we can still recover the spectrum of A if the conditions of Theorem 5 are satisfied:

Theorem 8 ([6]). *Let* $\{b_i : i \in \Omega\}$ *be the column vectors of B corresponding to* Ω *and let L be any fixed integer. Assume that* $\{P_j b_1, \ldots, P_j b_m\}$ *form a frame for* $P_j(\mathbb{C}^d)$, $j = 1, \ldots k$, *and* $\{D^L b_i : i \in \Omega\} \subset span(E)$, *where* $E = \bigcup_{i \in \Omega} \{b_i, Db_i, \ldots, D^{L-1} b_i\}$. *Then the spectrum of A can be recovered from* $\{A^k x(i) : i \in \Omega, k = 0, \ldots, (|\Omega| + 1)L - 1\}$ *for almost every* $x \in \mathbb{C}^d$.

Dynamical sampling in the infinite dimensional case

In this section we provide a brief account of other available research on dynamical sampling. Apart from the finite dimensional theory outlined in the previous sections, several infinite dimensional results have also been worked out. The $\ell^2(\mathbb{Z})$ theory, for example, parallels that of the finite dimensional case. In this case, $a \in \ell^2(\mathbb{Z})$ is the kernel of an evolution system so that the signal $x \in \ell^2(\mathbb{Z})$ at time $t = n$ is given by $A^n x = (\underbrace{a * \ldots * a}_{n}) * x$. If $S_m : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ denotes the operator of subsampling by a factor of m, $(S_m z)(k) = z(mk)$, and $S_{mn} T_c$ represents shifting by c followed by sampling by mn for some positive integer n, then the analog of Theorem 1 becomes

Theorem 9 ([8, 14]). *Suppose* $\hat{a}$ *is real, symmetric, continuous, and strictly decreasing on* $[0, \frac{1}{2}]$, n *is odd, and* $\Omega = \{1, \ldots, \frac{m-1}{2}\}$. *Then, for any* $N \geq m - 1$, *any* $x \in \ell^2(\mathbb{Z})$ *can be recovered in a stable way from the samples* $\{S_m x, S_m Ax, \ldots, S_m A^N x\}$ *and the additional samples given by either* $\{S_{mn} T_c x\}_{c \in \Omega}$ *or* $\{S_{mn} T_c x, S_{mn} T_c Ax, \ldots, S_{mn} T_c A^N x\}_{c \notin m\mathbb{Z}}$.

In the above theorem n can be taken arbitrarily large so that the oversampling factor is negligible just like the finite dimensional case. However, taking larger and larger n adversely affects the stability of the reconstruction. Estimates on the stability can be found in [7] in the finite dimensional case and [8] in the infinite dimensional case.

The $\ell^2(\mathbb{Z})$ theory leads to dynamical sampling in shift-invariant spaces which is an important setting for any sampling theory. Specifically, for an appropriate $\phi \in L^2(\mathbb{R})$, a principal shift invariant space (PSIS) $V(\phi)$ is the space defined by

$$V(\phi) = \left\{ \sum_{k \in \mathbb{Z}} c_k \phi(\cdot - k) : (c_k)_{k \in \mathbb{Z}} \in \ell^2(\mathbb{Z}) \right\}. \tag{11}$$

For the case in which $a * \phi \in V(\phi)$, the theory for $\ell^2(\mathbb{Z})$ can be applied directly via an isomorphism. For example, when $\phi = sinc$ then $V(\phi)$ is the Paley-Wiener space and the ℓ^2 theory applies directly. A characterization of the condition $a * \phi \in V(\phi)$ appears in [2]. The same paper also provides results for the dynamical sampling when $a * \phi \notin V(\phi)$. Similar results are also proved in the case of the so-called hybrid shift-invariant spaces in [1].

Acknowledgements The research in this chapter is funded by the collaborative NSF ATD grant DMS-1322127 and DMS-1322099. The authors would like to thank the organizers of the FFT at the University of Maryland for the opportunity to present our research to a wide audience of mathematicians and engineers. We are also grateful to S. J. Rose for his continued involvement in our projects.

References

1. R. Aceska, S. Tang, Dynamical sampling in hybrid shift invariant spaces, in *Operator Methods in Wavelets, Tilings, and Frames*, ed. by V. Furst, K.A. Kornelson, E.S. Weber. Contemporary Mathematics, vol. 626 (American Mathematical Society, Providence, RI, 2014)
2. R. Aceska, A. Aldroubi, J. Davis, A. Petrosyan, Dynamical sampling in shift invariant spaces, in *Commutative and Noncommutative Harmonic Analysis and Applications*, ed. by A. Mayeli, A. Iosevich, P.E.T. Jorgensen, G. Ólafsson. Contemporary Mathematics, vol. 603 (American Mathematical Society, Providence, RI, 2013), pp. 139–148
3. E. Acosta-Reyes, A. Aldroubi, I. Krishtal, On stability of sampling-reconstruction models. Adv. Comput. Math. **31**, 5–34 (2009)
4. A. Aldroubi, K. Gröchenig, Nonuniform sampling and reconstruction in shift-invariant spaces. SIAM Rev. **43**, 585–620 (2001) (electronic)
5. A. Aldroubi, I. Krishtal, Robustness of sampling and reconstruction and Beurling-Landau-type theorems for shift-invariant spaces. Appl. Comput. Harmon. Anal. **20**, 250–260 (2006)
6. A. Aldroubi, I. Krishtal, Krylov subspace methods in dynamical sampling (2015). ArXiv:1412.1538
7. A. Aldroubi, J. Davis, I. Krishtal, Dynamical sampling: time-space trade-off. Appl. Comput. Harmon. Anal. **34**, 495–503 (2013)
8. A. Aldroubi, J. Davis, I. Krishtal, Exact reconstruction of spatially undersampled signals in evolutionary systems. J. Fourier Anal. Appl. **21**(1), 11–31 (2015). doi:10.1007/s00041-014-9359-9. ArXiv:1312.3203
9. A. Aldroubi, C. Cabrelli, U. Molter, S. Tang, Dynamical sampling (2015). ArXiv:1409.8333
10. R.F. Bass, K. Gröchenig, Relevant sampling of band-limited functions. Illinois J. Math. **57**, 43–58 (2013)
11. J.J. Benedetto, P.J.S.G. Ferreira (eds.), *Modern Sampling Theory*. Applied and Numerical Harmonic Analysis (Birkhäuser Boston Inc., Boston, 2001)
12. T. Blu, P.-L. Dragotti, M. Vetterli, P. Marziliano, L. Coulot, Sparse sampling of signal innovations. IEEE Signal Process. Mag. **25**, 31–40 (2008)
13. E.J. Candès, C. Fernandez-Granda, Super-resolution from noisy data. J. Fourier Anal. Appl. **19**, 1229–1254 (2013)
14. J. Davis, Dynamical sampling with a forcing term, in *Operator Methods in Wavelets, Tilings, and Frames*, ed. by V. Furst, K.A. Kornelson, E.S. Weber. Contemporary Mathematics, vol. 626 (American Mathematical Society, Providence, RI, 2014)

15. A.G. Garcia, J.M. Kim, K.H. Kwon, G.J. Yoon, Multi-channel sampling on shift-invariant spaces with frame generators. Int. J. Wavelets Multiresolution Inf. Process. **10**, 1250003, 20 pp (2012)
16. D. Han, M.Z. Nashed, Q. Sun, Sampling expansions in reproducing kernel Hilbert and Banach spaces. Numer. Funct. Anal. Optim. **30**, 971–987 (2009)
17. J.A. Hogan, J.D. Lakey, *Duration and Bandwidth Limiting*. Applied and Numerical Harmonic Analysis (Birkhäuser/Springer, New York, 2012)
18. P.E.T. Jorgensen, A sampling theory for infinite weighted graphs. Opuscula Math. **31**, 209–236 (2011)
19. M. Liang, J. Du, H. Liu, Spatiotemporal super-resolution reconstruction based on robust optical flow and Zernike moment for video sequences. Math. Probl. Eng. 14 pp. (2013). Art. ID 745752
20. Y. Lyubarskiĭ, W.R. Madych, The recovery of irregularly sampled band limited functions via tempered splines. J. Funct. Anal. **125**, 201–222 (1994)
21. M.Z. Nashed, Q. Sun, Sampling and reconstruction of signals in a reproducing kernel subspace of $L^p(\mathbb{R}^d)$. J. Funct. Anal. **258**, 2422–2452 (2010)
22. A. Papoulis, Generalized sampling expansion, in *IEEE Transactions on Circuits and Systems, CAS-24* (1977), pp. 652–654
23. Q. Sun, Local reconstruction for sampling in shift-invariant spaces. Adv. Comput. Math. **32**, 335–352 (2010)
24. P.P. Vaidyanathan, V.C. Liu, Classical sampling theorems in the context of multirate and polyphase digital filter bank structures. IEEE Trans. Acoust. Speech Signal Process. **36**, 1480–1495 (1988)
25. N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series. With Engineering Applications* (The Technology Press of the Massachusetts Institute of Technology, Cambridge, 1949)
26. F. Xue, F. Luisier, T. Blu, Multi-Wiener SURE-LET deconvolution. IEEE Trans. Image Process. **22**, 1954–1968 (2013)

Signal Processing on Weighted Line Graphs

Aliaksei Sandryhaila and Jelena Kovačević

Abstract This chapter describes a signal processing framework for signals that are represented, or indexed, by weighted line graphs, which are a generalization of directed line graphs used for representation of time signals in classical signal processing theory. The presented framework is based on the theory of discrete signal processing on graphs and on algebraic signal processing theory. It defines fundamental signal processing concepts, such as signals and filters, z-transform, frequency and spectrum, Fourier transform and others, in a principled way. The framework also illustrates a strong connection between signal processing on weighted line graphs and signal representation based on orthogonal polynomials.

Key words: Signal processing on graphs, Algebraic signal processing, Orthogonal polynomials, Graph filter, Graph Fourier transform, Graph frequency

Introduction

Classical discrete signal processing (DSP) theory is based on a set of fundamental concepts that include signals, shift, filters, z-transform, convolution, spectrum, frequency, and Fourier transform [18, 36]. DSP assumes, sometimes implicitly, that signals are either infinite or periodically extended finite time series. As such, they can be visualized with graphs shown in Fig. 1. These are directed line graphs, where

A. Sandryhaila
Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: asandryh@andrew.cmu.edu

J. Kovačević (✉)
Electrical and Computer Engineering, Biomedical Engineering,
Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: jelenak@cmu.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_10

each node indicates a point in time and edge direction corresponds to the time flow from past to future. These graphs are unweighted, that is, all edges have weight 1.

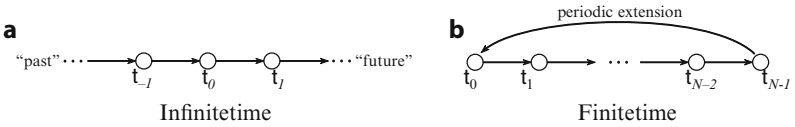


Fig. 1 Graph representations for infinite and finite discrete time signals.

In many applications, however, it is convenient and useful to represent signals using graphs. These signals cannot be viewed as time series and do not reside on directed unweighted line graphs. Examples include processing of measurements collected by networks of sensors; analysis of information and interactions in social and economic networks; research in collaborative activities, such as paper co-authorship and citations; topics and relevance of documents in the World Wide Web; interactions in molecular and gene regulatory networks; and many others.

To analyze and process such signals, a theory of *signal processing on graphs* has been developed that extends DSP concepts and techniques to signals represented by graphs. Presently, two main approaches to signal processing on graphs are based on the adjacency matrix of the underlying graph [27, 28] or the graph Laplacian matrix (see [33] and the references therein). Both frameworks define fundamental signal processing concepts on graphs, but the difference in their foundation leads to different definitions and techniques for signal analysis and processing. Moreover, due to the properties of the graph Laplacian, the latter approach is restricted to undirected graphs with nonnegative real weights.

The adjacency matrix-based approach was proposed in [27, 28]. Called *discrete signal processing on graphs* (DSP_G), it has been motivated by the algebraic signal processing theory [20–22]. Algebraic signal processing is a formal approach to signal processing that uses an algebraic representation of signals and filters as polynomials to derive fundamental signal processing concepts. This framework has been used for discovery of fast computational algorithms for discrete signal transforms [19, 23, 30]. It was extended to multidimensional signals and nearest neighbor graphs [24, 31] and used for signal compression [26, 29, 32]. The DSP_G framework generalizes and extends the algebraic approach to signals represented with arbitrary graphs. It uses the weighted adjacency matrix of the representation graph as the basic shift operator and develops appropriate concepts of z-transform, impulse and frequency response, filtering, convolution, and Fourier transform.

In this chapter, we illustrate the DSP_G framework by instantiating it for a special family of weighted line graphs, as shown in Fig. 2. These graphs generalize the representation graphs for discrete time signals in Fig. 1. We discuss how DSP_G framework leads to appropriate definitions of fundamental signal processing concepts for these graphs and demonstrate that it naturally connects them to signal representation with orthogonal polynomials.

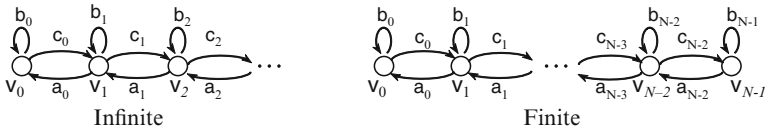


Fig. 2 Weighted line graphs.

Weighted Line Graphs

Weighted line graphs, such as those shown in Fig. 2, are infinite or finite graphs in which vertices are connected sequentially, so that each vertex v_n can only be connected to its neighbors v_{n-1} and v_{n+1} as well as to itself. For the clarity of discussion, we focus on finite weighted line graphs, such as those shown in Fig. 2(b). The signal model discussed here can be extended to infinite weighted line graphs using results from [31].

Consider a finite weighted line graph $G = (\mathscr{V}, \mathbf{A})$, where $\mathscr{V} = \{v_0, \ldots, v_{N-1}\}$ is a set of N vertices and $\mathbf{A}$ is a weighted adjacency matrix. The value $\mathbf{A}_{n,m}$ describes the weight of the edge from v_m to v_n. In particular, in finite weighted line graphs, each vertex v_n is connected to its neighbors v_{n-1} and v_{n+1} by edges with weights a_{n-1} and c_n, respectively, as well as to itself by an edge with weight b_n. Hence, the adjacency matrix of this graph has the form

$$\mathbf{A} = \begin{pmatrix} b_0 & a_0 & & & \\ c_0 & b_1 & \ddots & & \\ & \ddots & \ddots & a_{N-2} & \\ & & c_{N-2} & b_{N-1} & \end{pmatrix}. \tag{1}$$

We assume that $a_n, b_n, c_n \in \mathbb{R}$ and $a_n \neq 0$, $c_n \neq 0$ for all n.

Matrices of the form (1) are strongly related to a special class of polynomials called *orthogonal polynomials*. In particular, consider a set of polynomials $p_n(x)$, $n \geq 0$, that satisfy the three-term recurrence

$$x \cdot p_n(x) = a_{n-1} p_{n-1}(x) + b_n p_n(x) + c_n p_{n+1}(x), \tag{2}$$

with initial conditions $p_0(x) = 1$ and $p_{-1}(x) = 0$, where $a_n, b_n, c_n \in \mathbb{R}$ and $a_n c_n > 0$ for all values of n. There exists a real interval $\mathscr{I} \subseteq \mathbb{R}$ and a real-valued weight function $\mu(x)$ nonnegative on $\mathscr{I}$, such that $p_n(x)$ satisfy

$$\int_{\mathscr{I}} p_n(x) p_m(x) \mu(x) dx = \mu_n \delta_{n-m}, \tag{3}$$

that is, these are *orthogonal polynomials* defined by the recurrence (3). A thorough discussion of orthogonal polynomials can be found in [4, 35].

Orthogonal polynomials possess a number of important properties. Each polynomial $p_n(x)$ has exactly n real distinct roots $\lambda_0, \ldots, \lambda_{n-1}$ that lie within the

interval $\mathscr{I}$, that is, $\lambda_k \in \mathscr{I}$ for all $0 \leq k < n$. Hence, orthogonal polynomials satisfy $\deg p_n(x) = n$ and are linearly independent. Furthermore, if the weights a_n and b_n in recurrence (2) are the same as in the adjacency matrix (1), then the polynomial $p_N(x)$ is equal (up to a scalar factor) to the characteristic polynomial of $\mathbf{A}$, and the roots of $p_N(x)$ are the eigenvalues of $\mathbf{A}$. Hence, the adjacency matrix (1) of a weighted line graph has N distinct, simple eigenvalues and its eigendecomposition is [7]

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}, \tag{4}$$

where $\mathbf{\Lambda} = \mathbf{diag}(\lambda_0, \ldots, \lambda_{N-1})$ is a diagonal matrix of eigenvalues and $\mathbf{V}$ is a non-singular eigenvector matrix, so that the nth column of $\mathbf{V}$ is the eigenvector corresponding to the nth eigenvalue.

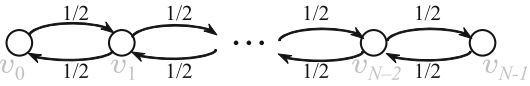


Fig. 3 Finite weighted line graph that corresponds to Chebyshev polynomials of the second kind.

As an example, consider the weighted line graph in Fig. 3 with edge weights $a_n = c_n = 1/2$ and $b_n = 0$. Orthogonal polynomials that correspond to this graph are *Chebyshev polynomials of the second kind*, one of the most well-known and widely used family of orthogonal polynomials [14]. They are denoted by $U_n(x)$ and satisfy the recurrence

$$x \cdot U_n(x) = \frac{1}{2}U_{n-1}(x) + \frac{1}{2}U_{n+1}(x), \tag{5}$$

with initial conditions $U_0(x) = 1$ and $U_{-1}(x) = 0$. Chebyshev polynomials of the second kind are orthogonal over the interval $\mathscr{I} = [-1, 1]$ with respect to the weight function $\mu(x) = \sqrt{1-x^2}$.

Signal Model

In this section, we define the model for signals represented by weighted line graphs. We do so by introducing appropriate concepts of the DSP_G framework and instantiating them for this class of graphs. A complete introduction to the DSP_G theory can be found in [27, 28].

Graph Signals

Signal processing on graphs is concerned with the analysis and processing of signals, in which signal values can be connected to each other according to some relation, such as physical proximity in sensor networks or friendship between

individuals in a social network. This relation is expressed though a graph $G = (\mathscr{V}, \mathbf{A})$ with vertices $\mathscr{V} = \{v_0, \ldots, v_{N-1}\}$ and a weighted adjacency matrix $\mathbf{A}$. Each signal value corresponds to vertex v_n, that is, it is *indexed* by v_n; and each weight $\mathbf{A}_{n,m}$ of a directed edge from v_m to v_n reflects the degree of relation between the mth and nth signal values. In particular, for weighted line graphs the adjacency matrix is given by (1).

A *graph signal* represented by a weighted line graph is a mapping

$$\begin{aligned} \mathbf{s} : \mathscr{V} &\to \mathbb{C}, \\ v_n &\mapsto s_n. \end{aligned} \tag{6}$$

Although a graph signal (6) can also be written as a complex-valued vector

$$\mathbf{s} = \begin{bmatrix} s_0 & s_1 & \ldots & s_{N-1} \end{bmatrix}^T \in \mathbb{C}^N,$$

one should view graph signals not merely as vectors, but as signal values s_n indexed by vertices v_n of the associated graph, as defined by (6).

Graph Filters

In general, a filter is a system that takes a signal as input, processes it, and produces another signal as output. Hence, DSP_G defines a *graph filter* as a system $\mathbf{H}(\cdot)$ that takes a graph signal $\mathbf{s}$ as an input and outputs a graph signal $\widetilde{\mathbf{s}} = \mathbf{H}(\mathbf{s})$.

The same way that time shift, or delay, is the basic filter in DSP, a basic non-trivial filter defined on a weighted line graph $G = (\mathscr{V}, \mathbf{A})$ is the *graph shift*. It is a local operation that replaces a signal sample s_n at vertex v_n with the linear combination of values at the neighbors of vertex v_n weighted by the corresponding edge weights:

$$\widetilde{s}_n = \sum_{m=0}^{N-1} \mathbf{A}_{n,m} s_m = c_{n-1} s_{n-1} + b_n s_n + a_n s_{n+1}. \tag{7}$$

Using the adjacency matrix (1), the output of the graph shift for a weighted line graph can also be written as the product of the input signal (6) with the adjacency matrix of the graph:

$$\widetilde{\mathbf{s}} = \begin{bmatrix} \widetilde{s}_0 & \ldots & \widetilde{s}_{N-1} \end{bmatrix}^T = \mathbf{A}\mathbf{s}. \tag{8}$$

All linear, shift-invariant[1] graph filters on weighted line graphs are polynomials in the adjacency matrix (1) of the form [27]

$$h(\mathbf{A}) = h_0 \mathbf{I} + h_1 \mathbf{A} + \ldots + h_L \mathbf{A}^L. \tag{9}$$

[1] Filters are *linear* if for a linear combination of inputs they produce the same linear combination of outputs. Filters are *shift-invariant* if the result of consecutive processing of a signal by multiple graph filters does not depend on the order of processing; that is, shift-invariant filters commute with each other.

The output of the filter (9) is the signal

$$\widetilde{\mathbf{s}} = \mathbf{H}(\mathbf{s}) = h(\mathbf{A})\mathbf{s}.$$

Linear, shift-invariant graph filters possess a number of useful properties. They have at most $L \leq N_{\mathbf{A}}$ taps h_ℓ, where $N_{\mathbf{A}} = \deg m_{\mathbf{A}}(x)$ is the degree of the minimal polynomial[2] $m_{\mathbf{A}}(x)$ of $\mathbf{A}$. If a graph filter (9) is invertible, that is, matrix $h(\mathbf{A})$ is non-singular, then its inverse is also a graph filter $g(\mathbf{A}) = h(\mathbf{A})^{-1}$ on the same graph $G = (\mathscr{V}, \mathbf{A})$. Finally, the space of graph filters is an *algebra*, that is, a vector space that is simultaneously a ring.

Graph z-transform

In DSP, z-transform provides a generalization to the Fourier transform as well as a means to express filtering through multiplication of series or polynomials in the time shift z^{-1} [18, 36].

DSP_G extends the concepts of the z-transform to signals and filters on graphs. The graph z-transform of the signal (6) is defined as a mapping

$$\mathbf{s} = (s_0, \ldots, s_{N-1})^T \mapsto s(x) = \sum_{n=0}^{N-1} s_n q_n(x). \tag{10}$$

Here, x stands for z^{-1}, and polynomials $q_0(x), \ldots, q_{N-1}(x)$ are linearly independent polynomials of degree at most $N-1$. Their exact structure and computation is discussed in Theorem 6 in [27].

The z-transform for graph filters is defined by the mapping $\mathbf{A} \mapsto x$, or

$$h(\mathbf{A}) \mapsto h(x). \tag{11}$$

Filtering using z-transform is performed through multiplication of z-transforms modulo characteristic polynomial[3] $p_{\mathbf{A}}(x)$ of $\mathbf{A}$. Namely, if $\widetilde{\mathbf{s}} = h(\mathbf{A})\mathbf{s}$ is the output signal of the filter $h(\mathbf{A})$, then its z-transform is given by the product

$$\widetilde{\mathbf{s}} \mapsto \widetilde{s}(x) = \sum_{n=0}^{N-1} \widetilde{s}_n q_n(x) = h(x)s(x) \mod p_{\mathbf{A}}(x). \tag{12}$$

It follows from Theorem 6 in [27] that for weighted line graphs, the basis polynomials $q_n(x)$ in the signal z-transform (10) satisfy the following property: the vector

$$\big(q_0(\lambda_m) \ \ldots \ q_{N-1}(\lambda_m)\big)^T \tag{13}$$

[2] The minimal polynomial of $\mathbf{A}$ is the unique monic polynomial of the smallest degree that annihilates $\mathbf{A}$, that is, $m_{\mathbf{A}}(\mathbf{A}) = 0$ [7].

[3] The *characteristic polynomial* of a matrix $\mathbf{A}$ is defined as $p_{\mathbf{A}}(x) = \det(x\mathbf{I} - \mathbf{A}) = \prod_{n=0}^{N-1}(x - \lambda_n)$ [7].

of polynomials $q_n(x)$ evaluated at the eigenvalue λ_m is exactly the eigenvector of $\mathbf{A}^T$ that corresponds to λ_m. Hence, vector (13) satisfies

$$\mathbf{A}^T \left(q_0(\lambda_m) \ \dots \ q_{N-1}(\lambda_m)\right)^T = \lambda_m \left(q_0(\lambda_m) \ \dots \ q_{N-1}(\lambda_m)\right)^T,$$

which, given the structure (1) of matrix $\mathbf{A}$, is equivalent to

$$\begin{aligned} b_0 q_0(\lambda_m) + c_0 q_1(\lambda_m) &= \lambda_m q_0(\lambda_m) \\ a_0 q_0(\lambda_m) b_1 q_1(\lambda_m) + c_1 q_2(\lambda_m) &= \lambda_m q_1(\lambda_m) \\ &\vdots \\ a_{N-3} q_{N-3}(\lambda_m) b_{N-2} q_{N-2}(\lambda_m) + c_{N-2} q_{N-2}(\lambda_m) &= \lambda_m q_{N-2}(\lambda_m) \\ a_{N-2} q_{N-2}(\lambda_m) + b_{N-1} q_{N-1}(\lambda_m) &= \lambda_m q_{N-1}(\lambda_m). \end{aligned} \tag{14}$$

Comparing the system of equations (14) to the recurrence (2) for orthogonal polynomials $p_n(x)$ and recalling from the discussion in Section that λ_m is a root of the orthogonal polynomial $p_N(x)$, we conclude that $q_n(x) = p_n(x)$ for $0 \leq n < N$.

Hence, the graph z-transform for weighted line graphs with adjacency matrices (1) has the form

$$\mathbf{s} = (s_0, \dots, s_{N-1})^T \mapsto s(x) = \sum_{n=0}^{N-1} s_n p_n(x), \tag{15}$$

where $p_n(x)$ are orthogonal polynomials that satisfy the recurrence (2). Since the characteristic polynomial of $\mathbf{A}$ in (1) is $p_N(x)$, we also conclude that filtering (12) using z-transforms on weighted line graphs is performed as

$$\widetilde{\mathbf{s}} \mapsto \widetilde{s}(x) = \sum_{n=0}^{N-1} \widetilde{s}_n p_n(x) = h(x)s(x) \mod p_N(x). \tag{16}$$

Frequency Analysis

In this section, we continue to build the signal processing framework on weighted line graphs. We introduce the DSP_G notions of frequency, spectrum, Fourier transform, and signal variation and instantiating them for weighted line graphs. These concepts are defined and discussed in detail in [27, 28].

Graph Fourier transform

In finite-time signal processing, a Fourier transform is the decomposition of a signal into a *Fourier basis* of signals that are invariant to filtering [18, 36]. In DSP_G,

a graph Fourier basis is given by the Jordan basis of the adjacency matrix $\mathbf{A}$. Since weighted line graphs have *diagonalizable* adjacency matrices (4), for them the Jordan decomposition coincides with the eigendecomposition and it suffices to discuss the graph Fourier transforms in terms of the eigenvectors of $\mathbf{A}$.

Given the eigendecomposition (4), the graph Fourier basis is given by the eigenvectors of $\mathbf{A}$. Hence, the *graph Fourier transform* of a graph signal $\mathbf{s}$ is defined as

$$\widehat{\mathbf{s}} = \mathbf{F}\mathbf{s} = \mathbf{V}^{-1}\mathbf{s}, \tag{17}$$

where $\mathbf{F} = \mathbf{V}^{-1}$ denotes the graph Fourier transform matrix. The values $\widehat{s}_n$ in (17) are called the *spectrum* of the signal $\mathbf{s}$. The *inverse graph Fourier transform*, given by

$$\mathbf{s} = \mathbf{F}^{-1}\widehat{\mathbf{s}} = \mathbf{V}\widehat{\mathbf{s}},$$

reconstructs the signal from its spectrum.

The graph Fourier transform on weighted line graphs has a number of useful properties. Recall that the vector (13) is the eigenvector of $\mathbf{A}^T$ with the eigenvalue λ_m. Hence, the eigendecomposition (4) of $\mathbf{A}$ is

$$\mathbf{A} = \left(\mathbf{A}^T\right)^T = \left(\mathbf{P}_{p,\lambda}^T \mathbf{\Lambda} \left(\mathbf{P}_{p,\lambda}^T\right)^{-1}\right)^T = \mathbf{P}_{p,\lambda}^{-1} \mathbf{\Lambda} \mathbf{P}_{p,\lambda},$$

where $\mathbf{\Lambda}$ is the diagonal eigenvalue matrix from (4) and $\mathbf{P}_{p,\lambda}$ is the *polynomial transform* matrix

$$\mathbf{P}_{p,\lambda} = \begin{pmatrix} p_0(\lambda_0) & \dots & p_{N-1}(\lambda_0) \\ \vdots & \vdots & \vdots \\ p_0(\lambda_{N-1}) & \dots & p_{N-1}(\lambda_{N-1}) \end{pmatrix}. \tag{18}$$

The (m,k)th element of $\mathbf{P}_{p,\lambda}$ is the kth orthogonal polynomial $p_k(x)$ evaluated at the mth eigenvalue λ_m.

Hence, the graph Fourier transform (17) on weighted line graphs is given by the polynomial transform (18):

$$\mathbf{F} = \mathbf{P}_{p,\lambda}. \tag{19}$$

The inverse Fourier transform matrix, in turn, is given by $\mathbf{F}^{-1} = \mathbf{P}_{p,\lambda}^{-1}$. As Theorem 4 in [31] demonstrates, the Fourier transform matrix is "almost" orthogonal, which may simplify the computation of its inverse. Here, we restate this theorem for convenience.

Theorem 1. *Define* $\eta_m = \prod_{k=0}^{m-1}(c_k/a_k)$ *and diagonal matrices*

$$\mathbf{D} = c_{N-1}\eta_{N-1} \begin{pmatrix} p_{N-1}(\lambda_0)p'_N(\lambda_0) & & \\ & \ddots & \\ & & p_{N-1}(\lambda_{N-1})p'_N(\lambda_{N-1}) \end{pmatrix}$$

$$\mathbf{E} = \begin{pmatrix} \eta_0 & & \\ & \ddots & \\ & & \eta_{N-1} \end{pmatrix},$$

where $p'_N(x)$ denotes the derivative of the Nth orthogonal polynomial $p_N(x)$. Then the matrix

$$\mathbf{D}^{-1/2}\mathbf{P}_{p,\lambda}\mathbf{E}^{1/2} \tag{20}$$

is orthogonal.

As an example, consider the graph in Fig. 3 and the corresponding Chebyshev polynomials $U_n(x)$ defined by (5). The roots of the Nth polynomial $U_N(x)$ are

$$\lambda_m = \cos\frac{\pi m}{N+1} \tag{21}$$

and Chebyshev polynomials satisfy the property [14]

$$U_n\left(\cos\frac{\pi m}{N+1}\right) = \frac{\sin\left(\pi m(n+1)/(N+1)\right)}{\sin\left(\pi m/(N+1)\right)},$$

we obtain that the graph Fourier transform (18) associated with the weighted line graph in Fig. 3 is the discrete sine transform of type I [19, 22]. Moreover, for this graph we obtain $\eta_m = \prod_{k=0}^{m-1}(c_k/a_k) = 1$, and the nth diagonal elements of the corresponding matrices $\mathbf{D}$ and $\mathbf{E}$ are, respectively, $U_{N-1}(\lambda_n)U'_N(\lambda_n)/2$ and 1, that is, $\mathbf{E}$ is an identity matrix.

Frequency Response

The graph Fourier transform also characterizes the effect of the filter on the frequency content of an input signal. It follows from (4), (9) and (17) that

$$\widetilde{\mathbf{s}} = h(\mathbf{A})\mathbf{s} = \mathbf{F}^{-1}h(\mathbf{\Lambda})\mathbf{F}\mathbf{s} \;\Leftrightarrow\; \mathbf{F}\widetilde{\mathbf{s}} = h(\mathbf{\Lambda})\widehat{\mathbf{s}}. \tag{22}$$

Thus, the spectrum of the output signal is the spectrum of the input signal multiplied by the diagonal matrix $h(\mathbf{\Lambda})$. This matrix is called the *graph frequency response* of the filter $h(\mathbf{A})$.

Note that (22) extends the *convolution theorem* from DSP [18] to graphs: filtering a signal on a graph is equivalent in the frequency domain to multiplying the signal's spectrum by the frequency response of the filter.

Signal Variation on Graphs

In DSP, signal frequencies are often described as "low" or "high". For time signals, these concepts have a simple interpretation, since their frequency contents are described by complex or real sinusoids that oscillate at different rates [9]. Frequency components oscillating with low or high rates correspond, respectively, to low or high frequencies.

On graphs, frequency components are characterized based on how much they vary with respect to the underlying graph, that is, how much they change from a vertex to its neighbors. This characteristic is quantified by the *total variation on graphs* defined as

$$TV_G(\mathbf{s}) = ||\mathbf{s} - \mathbf{A}^{\text{norm}}\mathbf{s}||_1 , \tag{23}$$

where

$$\mathbf{A}^{\text{norm}} = \frac{1}{|\lambda_{\max}|}\mathbf{A} \tag{24}$$

is the graph shift matrix normalized by the eigenvalue of $\mathbf{A}$ with the largest magnitude, that is $|\lambda_{\max}| \geq |\lambda_n|$ for all $0 \leq n \leq N-1$.

Weighted line graphs have real spectra of simple, distinct eigenvalues, as discussed in Section . In this case, the ordering of frequencies from the lowest to the highest is opposite to the eigenvalue ordering [28]: if eigenvalues are ordered as $\lambda_0 < \lambda_1 < \ldots < \lambda_{N-1}$, then λ_{N-1} corresponds to the lowest frequency and λ_0 corresponds to the highest one.

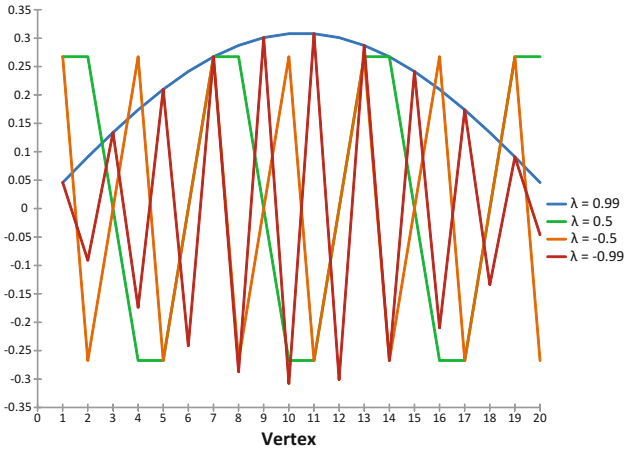


Fig. 4 Frequency components for the graph in Fig. 3 with $N = 20$ vertices that correspond to different frequencies λ. Larger values of λ represent "lower" graph frequencies that are smoother with respect to the underlying graph, and smaller values of λ represent "higher" graph frequencies.

As an illustration, consider again the graph in Fig. 3. The eigenvalues (21) of its adjacency matrix are real numbers in the interval $[-1, 1]$. Fig. 4 shows several

frequency components for the graph with $N = 20$ vertices. As can be immediately observed, the low-frequency components change significantly less between connected vertices than the high-frequency components, supporting the intuition behind the definition of signal variation on graphs.

Applications

In this section, we discuss potential applications for signal processing framework for weighted line graphs. Some of the orthogonal polynomials mentioned in this section, such as Chebyshev, Hermite, and Laguerre ones, are well known [4, 14, 35]; others can be constructed with the recurrence (2) using application-specific coefficients a_n, b_n and c_n.

Signal representation

One motivation for novel signal models is their suitability for the description and analysis of certain classes of signals. The advantage of proposed models can be related to the improved characterization of signal properties and efficient signal processing tools they produce.

Bases of orthogonal polynomials for the representation of diverse functions have been proposed in numerous studies, some of which we discuss below. Two common applications are the representation of continuous functions using orthogonal polynomials and the representation of discrete functions using sampled orthogonal polynomials. The properties of these bases, including the calculation of projection coefficients and approximation error analysis, are discussed in [4, 5, 35]. Furthermore, for different functions, the choice of orthogonal polynomials may not be obvious and may depend on the application.

The projection coefficients in both cases can be viewed and manipulated as signals residing on infinite or finite weighted line graphs. Besides deeper insight, this framework offers additional tools for the processing of signals based on orthogonal polynomials. For instance, since the framework defines the concepts of filtering, spectrum, and frequency response, one can potentially construct frequency-selective filters, or even filter banks, to facilitate signal analysis and processing [25].

Below, we discuss several applications that are based on the expansion of finite discrete functions into sampled orthogonal polynomials.

Electrocardiographic signal processing

Hermite polynomials, denoted as $H_n(x)$, satisfy the recurrence (2) with coefficients $a_n = n+1, b_n = 0$, and $c_n = 1/2$. They are orthogonal over the entire real line $\mathscr{I} = \mathbb{R}$ with respect to the weight function $\mu(x) = e^{-x^2}$.

The expansion of signals into continuous and sampled Hermite polynomials has been proposed for image analysis and processing [12, 13] and electrophysiological signal compression [6, 8, 29, 32, 34]. In particular, an interesting observation resulting from [29, 32] was that sampling the electrocardiographic signals at time points proportional to the roots of Hermite polynomials is more efficient than sampling at equal time intervals. Moreover, the proposed expansion led to a significantly improved compression algorithm for electrocardiographic signals.

Speech signal analysis

Similarly to the above expansion of signals into Hermite polynomials, signals can also be expanded into Laguerre polynomials. These polynomials, denoted as $L_n(x)$, satisfy the recurrence (2) with coefficients $a_n = -(n+1)$, $b_n = 2n+1$, and $c_n = -(n+1)$. They are orthogonal over the semi-infinite interval $\mathscr{I} = \mathbb{R}_+$ with respect to the weight function $\mu(x) = e^{-x}$.

Speech coding and a representation of exponentially decaying signals using sampled Laguerre polynomials has been studied in [10, 11]. The analysis of the corresponding signal models may offer valuable insights for improved analysis and efficient processing of these classes of signals.

Image compression

Image compression is an extensive research area in signal processing (see [2] and the references therein). Compression of multiple images with similar structure, such as collections of faces, handwritten digits, etc., using weighted line graphs and associated signal models was considered in [26]. Since images are finite discrete two-dimensional signals that reside on rectangular lattices, these lattices can be represented as a tensor product of two 1-D weighted line graphs [3, 20]. For both graphs, the coefficients a_k, b_k, c_k in (2), can be obtained by solving an ℓ_2-minimization problem, and are dependent on images of interest.

Correlation analysis of Gauss-Markov random fields

Signal models for weighted line graphs can also be relevant to the analysis of Gauss-Markov random fields [15, 16, 20]. Consider N random variables $\xi_0, \ldots, \xi_{N-1}$ that satisfy the difference equation

$$\xi_n = v_{n-1}\xi_{n-1} + u_n\xi_n + v_n\xi_{n+1} + \nu_n, \tag{25}$$

where ν_n is a zero-mean Gaussian noise, and $v_n, u_n \in \mathbb{R}$ are real-valued coefficients. The set $\{\xi_n\}_{0 \le n < N}$ is called a *first-order Gauss-Markov random field* defined on the finite lattice $0 \le n < N$. We assume zero (Dirichlet) boundary conditions $\xi_{-1} = 0$ and $\xi_N = 0$.

The Karhunen-Loève transform (KLT), described by the eigenvector matrix of the covariance matrix $\boldsymbol{\Sigma}$, decorrelates the signal $\mathbf{s} = \left(\xi_0 \ldots, \xi_{N-1}\right)^T$. Under certain

conditions, it is considered to be the optimal transform for signal compression; however, there is no general efficient algorithm to compute this transform [1].

As demonstrated in [15, 16], the inverse of the covariance matrix $\boldsymbol{\Sigma}$ for the above Gauss-Markov random field is

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} u_0 & v_0 & & & \\ v_0 & u_1 & \ddots & & \\ & \ddots & \ddots & v_{N-2} & \\ & & v_{N-2} & u_{N-1} & \end{pmatrix}.$$

We can set the values of coefficients in the recurrence (2) to $a_n = c_n = v_n$ and $b_n = u_n$, and construct the corresponding family of orthogonal polynomials. In this case, the orthogonalized graph Fourier transform (20) is precisely the KLT for the above random field [31]. This result implies that an instantiation of the random variables $\xi_0, \ldots, \xi_{N-1}$ can be viewed, analyzed, and processed as Fourier transform coefficients in the constructed signal model.

Fast signal transforms

Invertible linear transforms are widely used in signal processing. Examples include discrete Fourier transform, discrete cosine and sine transforms, discrete wavelet transform, KLT, and many others.

Efficient and fast implementations of these transforms is an important research problem that can be addressed using multiple approaches. One of them is based on the recognition of a transform as a polynomial transform (18) for an appropriate signal model. In this case, a decomposition of the model into a combination of simpler models corresponds to a factorization of the transform into a series of simpler transforms that may yield efficient and fast computational algorithms. The general theory of this approach has been discussed in [23, 30, 37]; early work on using polynomial transforms to derive algorithms for this was done in [17].

Fast algorithms for discrete cosine and sine transforms

It was demonstrated in [22, 23] that the discrete cosine and sine transforms are Fourier transforms for the 1-D space signal model, which is a signal model for signals residing on graphs with edge weights $a_n = c_n = 1/2$, such as the graph in Fig. 3, with possible exceptions for boundary vertices. These graphs are associated Chebyshev polynomials of four kinds. By exploiting the structure of the underlying signal models, a large number of fast algorithms for discrete cosine and sine transforms was derived that require significantly fewer operations than direct computations of these transforms.

Conclusion

We have described a signal processing framework for signals that are represented by weighted line graphs—a generalization of directed line graphs used for representation of time signals in the classical DSP theory. The presented framework is built on the theory of discrete signal processing on graphs and on algebraic signal processing theory. We presented definitions of fundamental signal processing concepts, such as signals and filters, z-transform, frequency and spectrum, Fourier transform and others, for the class of weighted line graphs. We illustrated a connection between signal processing on weighted line graphs and representation of signals using on orthogonal polynomials. We also discussed potential applications of the presented framework to signal representation and compression and to design of fast algorithms for certain linear transforms.

References

1. H.C. Andrews, Multidimensional rotations in feature selection. IEEE Trans. Comput. **20**(9),1045–1051 (1971)
2. A. Bovik, *Handbook of Image and Video Processing*, 2nd edn. (Academic, Amsterdam, 2005)
3. D.E. Dudgeon, R.M. Mersereau, Multidimensional digital signal processing, in *Prentice-Hall Signal Processing Series* (Englewood Cliffs, Prentice-Hall, 1984)
4. W. Gautschi, *Orthogonal Polynomials: Computation and Approximation* (Oxford University Press, New York, 2004)
5. A. Jerri, *Integral and Discrete Transforms with Applications and Error Analysis* (CRC Press, New York, NY 1992)
6. P. Laguna, R. Jané, S. Olmos, N.V. Thakor, H. Rix, P. Caminal, Adaptive estimation of QRS complex wave features of ECG signal by the Hermite model. J. Med. Biol. Eng. Comput. **34**(1), 58–68 (1996)
7. P. Lancaster, M. Tismenetsky, *The Theory of Matrices*, 2nd edn. (Academic, San Diego, 1985)
8. L.R. Lo Conte, R. Merletti, G.V. Sandri, Hermite expansion of compact support waveforms: applications to myoelectric signals. IEEE Trans. Biomed. Eng. **41**(12), 1147–1159 (1994)
9. S. Mallat, *A Wavelet Tour of Signal Processing*, 3rd edn. (Academic, Burlington, 2008)
10. G. Mandyam, N. Ahmed, The discrete Laguerre transform: derivation and applications. IEEE Trans. Signal Process. **44**(12), 2925–2931 (1996)
11. G. Mandyam, N. Ahmed, N. Magotra, Application of the discrete Laguerre transform to speech coding, in *Proceedings of Asilomar Conference on Signals, Systems, Computers* (1995), pp. 1225–1228
12. J.-B. Martens, The Hermite transform - theory. IEEE Trans. Acoust. Speech Signal Process. **38**(9), 1595–1605 (1990)
13. J.-B. Martens, The Hermite transform - applications. IEEE Trans. Acoust. Speech Signal Process. **38**(9), 1607–1618 (1990)
14. J.C. Mason, D.C. Handscomb, *Chebyshev Polynomials* (Chapman and Hall/CRC, Boca Raton, 2002)
15. J.M.F. Moura, N. Balram, Recursive structure of noncausal Gauss Markov random fields. IEEE Trans. Inf. Theory **38**(2), 334–354 (1992)
16. J.M.F. Moura, M.G.S. Bruno, DCT/DST and Gauss-Markov fields: conditions for equivalence. IEEE Trans. Signal Process. **46**(9), 2571–2574 (1998)

17. H.J. Nussbaumer, *Fast Fourier Transformation and Convolution Algorithms*, 2nd edn. (Springer, Berlin, 1982)
18. A.V. Oppenheim, R.W. Schafer, J.R. Buck, *Discrete-Time Signal Processing*, 2nd edn. (Prentice Hall, Upper Saddle River, 1999)
19. M. Püschel, J.M.F. Moura, The algebraic approach to the discrete cosine and sine transforms and their fast algorithms. SIAM J. Comput. **32**(5), 1280–1316 (2003)
20. M. Püschel, J.M.F. Moura, Algebraic signal processing theory (2005), http://arxiv.org/abs/cs.IT/0612077
21. M. Püschel, J.M.F. Moura, Algebraic signal processing theory: foundation and 1-D time. IEEE Trans. Signal Process. **56**(8), 3572–3585 (2008)
22. M. Püschel, J.M.F. Moura, Algebraic signal processing theory: 1-D space. IEEE Trans. Signal Process. **56**(8), 3586–3599 (2008)
23. M. Püschel, J.M.F. Moura, Algebraic signal processing theory: Cooley-Tukey type algorithms for DCTs and DSTs. IEEE Trans. Signal Process. **56**(4), 1502–1521 (2008)
24. M. Püschel, M. Rötteler, Algebraic signal processing theory: 2-D hexagonal spatial lattice. IEEE Trans. Image Process. **16**(6), 1506–1521 (2007)
25. A. Sandryhaila, Algebraic signal processing: modeling and subband analysis. Thesis, Carnegie Mellon University, Pittsburgh (2010)
26. A. Sandryhaila, J.M.F. Moura, Nearest-neighbor image model. *Proceedings of IEEE International Conference on Image Processing* (2012), pp. 2521–2524
27. A. Sandryhaila, J.M.F. Moura, Discrete signal processing on graphs. IEEE Trans. Signal Process. **61**(7), 1644–1656 (2013)
28. A. Sandryhaila, J.M.F. Moura. Discrete signal processing on graphs: frequency analysis. IEEE Trans. Signal Process **62**(12), 3042–3054 (2014)
29. A. Sandryhaila, J. Kovačević, M. Püschel, Compression of QRS complexes using Hermite expansion. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (2011), pp. 581–584
30. A. Sandryhaila, J. Kovačević, M. Püschel, Algebraic signal processing theory: Cooley-Tukey type algorithms for polynomial transforms based on induction. SIAM J. Matrix Anal. Appl. **32**(2), 364–384 (2011)
31. A. Sandryhaila, J. Kovačević, M. Püschel, Algebraic signal processing theory: 1-D nearest-neighbor models. IEEE Trans. Signal Process. **60**(5), 2247–2259 (2012)
32. A. Sandryhaila, S. Saba, M. Püschel, J. Kovačević, Efficient compression of QRS complexes using Hermite expansion. IEEE Trans. Signal Process. **60**(2), 947–955 (2012)
33. D.I. Shuman, S.K. Narang, P. Frossard, A. Ortega, P. Vandergheynst, The emerging field of signal processing on graphs. IEEE Signal Process. Mag. **30**(3), 83–98 (2013)
34. L. Sörnmo, P.O. Börjesson, P. Nygards, O. Pahlm, A method for evaluation of QRS shape features using a mathematical model for the ECG. IEEE Trans. Biomed. Eng. **BME-28**(10), 713–717 (1981)
35. G. Szegö, *Orthogonal Polynomials*, 3rd edition. American Mathematical Society Colloquium Publications, USA (1967)
36. M. Vetterli, J. Kovačević, V.K. Goyal, *Foundations of Signal Processing* (Cambridge University Press, Cambridge, 2014)
37. Y. Voronenko, M. Püschel, Algebraic signal processing theory: Cooley-Tukey type algorithms for teal DFTs. IEEE Trans. Signal Process. **57**(1), 205–222 (2009)

Adaptive Signal Processing

Stephen D. Casey

Abstract Adaptive frequency band (AFB) and ultra-wideband (UWB) systems require either rapidly changing or very high sampling rates. Conventional analog-to-digital devices are nonadaptive and have limited dynamic range. We investigate AFB and UWB sampling via a basis projection method. The method decomposes the signal into a basis over time segments via a continuous-time inner product operation and then samples the basis coefficients in parallel. The signal may then be reconstructed from the basis coefficients to recover the signal in the time domain. The overarching goal of the theory developed in this chapter is to develop a computable atomic decomposition of time-frequency space. The idea is to come up with a way of nonuniformly tiling time and frequency so that if the signal has a burst of high-frequency information, we tile quickly and efficiently in time and broadly in frequency, whereas if the signal has a relatively low-frequency segment, we can tile broadly in time and efficiently in frequency. Computability is key; systems are designed so that they can be constructed using splines and implemented in circuitry.

Key words: interpolation theory, Shannon Sampling, irregular sampling, Gabor systems, wavelets, splines, frame theory.

1 Introduction

Adaptive frequency band (AFB) and ultra-wideband (UWB) systems, requiring either rapidly changing or very high sampling rates, stress classical sampling approaches. At UWB rates, conventional analog-to-digital devices have limited dynamic range and exhibit undesired nonlinear effects such as timing jitter.

S.D. Casey (✉)
Department of Mathematics and Statistics, American University,
Washington, DC 20016-8050, USA
e-mail: scasey@american.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_11

Increased sampling speed leads to less accurate devices that have lower precision in numerical representation. This motivates alternative sampling schemes that use mixed-signal approaches, coupling analog processing with parallel sampling, to provide improved sampling accuracy and parallel data streams amenable to more efficient (parallel) digital computation. Wideband problems continue to hit barriers in sample and hold architectures and analog-to-digital conversion, especially with regard to energy. Digital circuitry has provided dramatically enhanced digital signal processing operation speeds, but there has not been a corresponding dramatic energy capacity increase in batteries to operate these circuits; there is no Moore's Law for batteries or analog-to-digital conversion.

This chapter presents a different approach to sampling developed to address the challenges of AFB and UWB signals. We investigate sampling for these classes of signals by a basis projection method. The method represents a change of view[1] in sampling, from that of a stationary view of a signal used in classical sampling to an "adaptive windowed stationary" view. This viewpoint, in the AFB case, gives that the time and frequency space "tile" occupied by the signal changes in time. The windows give us the tools to partition time-frequency so that the signal can be sampled efficiently. The UWB case takes advantage of the windowing to partition the signal uniformly but also quickly and efficiently. With the blocks, the signal can be sampled in parallel.

The method was introduced as a means of UWB parallel sampling by Hoyos *et. al.* [17] and applied to UWB communications systems [5, 8, 18–20]. The method first systematically windows the signal in time. It then decomposes the windowed signal into a basis over time segments via a continuous-time inner product operation and then samples the basis coefficients in parallel. The signal may then be reconstructed from the basis coefficients to recover time domain samples, or further processing may be carried out in the new domain [8, 17]. We address several issues associated with the basis expansion and sampling procedure, including windowing systems, choice of basis, truncation error, rate of convergence, and segmentation of the signal. We develop the theory in truncated and overlapping domains, using the theory of splines to get smoothness in time and decay in frequency. We employ the theory of lapped orthogonal transforms to preserve the orthogonality of basis elements in the overlapping regions. We compute exact truncation error bounds and compare the method with traditional sampling. We close by putting the theory in the context of general methods for time-frequency analysis.

1.1 Preliminary Definitions

In this chapter, all functions considered are absolutely and square integrable functions on the real line $(f \in L^1 \cap L^2(\mathbb{R}))$, unless noted otherwise. Likewise, all integrals are assumed to be over the whole domain (either $\mathbb{R}$ or $\mathbb{C}$ depending on the

[1] Meyer [28] gives an excellent overview of these different points in Chapter 1 of his book.

context) unless noted otherwise. References for the material on harmonic analysis and sampling include Benedetto [1], Dym and McKean [12], Grafakos [13], Higgins [15], Hömander [16], Jerri [22], Körner [24], Levin [25], Meyer [28], Papoulis [31, 32], and Young [41]. Background for the work on splines can be found in Nürnberger [29], Prenter [33], and Schoenberg [34].

Fourier series and Fourier transforms play key roles in our work. Their definitions, from [1] and [12], follow. Let $\exp(\cdot) = e^{(\cdot)}$.

Definition 1 (Fourier Series). Let f be a periodic, integrable function on $\mathbb{R}$, with period 2Φ, i.e., $f \in L^1(\mathbb{T}_{2\Phi})$. The Fourier coefficients of f, $\widehat{f}[n]$, are defined by

$$\widehat{f}[n] = \frac{1}{2\Phi} \int_{-\Phi}^{\Phi} f(t) \exp(-i\pi nt/\Phi)\, dt\,.$$

If $\{\widehat{f}[n]\}$ is absolutely summable ($\{\widehat{f}[n]\} \in l^1$), then the Fourier series of f is

$$f(t) = \sum_{n\in\mathbb{Z}} \widehat{f}[n] \exp(i\pi nt/\Phi)\,.$$

For $f \in L^1$, the Fourier transform $\widehat{f}(\omega)$ is given as follows.

Definition 2 (Fourier Transform and Inversion Formulae). Let f be a function in L^1. The Fourier transform of f is defined as

$$\widehat{f}(\omega) = \int_{\mathbb{R}} f(t) e^{-2\pi i t\omega} dt$$

for $t \in \mathbb{R}$ (time), $\omega \in \widehat{\mathbb{R}}$ (frequency). The inversion formula, for $\widehat{f} \in L^1(\widehat{\mathbb{R}})$, is

$$f(t) = (\widehat{f})^{\vee}(t) = \int_{\widehat{\mathbb{R}}} \widehat{f}(\omega) e^{2\pi i\omega t} d\omega.$$

Formally, we can think of the transform and the coefficient integral as *analysis*, and the inverse transform and series as *synthesis*. The choice to have 2π in the exponent simplifies certain expressions, e.g., for $f, g \in L^1 \cap L^2(\mathbb{R})$, $\widehat{f}, \widehat{g} \in L^1 \cap L^2(\widehat{\mathbb{R}})$, we have the *Parseval – Plancherel* equations –

$$\|f\|_{L^2(\mathbb{R})} = \|\widehat{f}\|_{L^2(\widehat{\mathbb{R}})} \text{ and } \langle f, g\rangle = \langle \widehat{f}, \widehat{g}\rangle\,.$$

We extend the transform from $L^1 \cap L^2$ to L^2 via a density argument. We also need to define the *periodization* of a function of finite support.

Definition 3 (Periodization). Let $T > 0$ and let $g(t)$ be a function such that $\operatorname{supp} g \subseteq [0, T]$. The *T-periodization* of g is $[g]^{\circ}(t) = \sum_{n=-\infty}^{\infty} g(t - nT)$.

1.2 W-K-S Sampling

Classical sampling theory applies to square integrable bandlimited functions. A function that is both Ω bandlimited and L^2 has several smoothness and growth properties given in the Paley-Wiener Theorem (see, e.g., [12]). We denote this class of functions by $\mathbb{PW}_\Omega$. The Whittaker-Kotel'nikov-Shannon (W-K-S) Sampling Theorem [23, 35, 39, 40] applies to functions in $\mathbb{PW}_\Omega$.

Definition 4 (Paley-Wiener Space $\mathbb{PW}_\Omega$).

$$\mathbb{PW}_\Omega = \{f : f, \widehat{f} \in L^2, \text{supp}(\widehat{f}) \subset [-\Omega, \Omega]\}.$$

Theorem 1 (W-K-S Sampling Theorem). *Let* $f \in \mathbb{PW}_\Omega$, $\text{sinc}_T(t) = \frac{\sin(\frac{\pi}{T}t)}{\pi t}$ *and* $\delta_{nT}(t) = \delta(t - nT)$.

a.) *If* $T \leq 1/2\Omega$, *then for all* $t \in \mathbb{R}$,

$$f(t) = T \sum_{n\in\mathbb{Z}} f(nT) \frac{\sin(\frac{\pi}{T}(t - nT))}{\pi(t - nT)} = T \left(\left[\sum_{n\in\mathbb{Z}} \delta_{nT} \right] f \right) * \underset{T}{\text{sinc}}(t). \tag{1}$$

b.) *If* $T \leq 1/2\Omega$ *and* $f(nT) = 0$ *for all* $n \in \mathbb{Z}$, *then* $f \equiv 0$.

A beautiful way to prove the W-K-S Sampling Theorem is to use the Poisson Summation Formula (PSF). Let $T > 0$ and for $f \in L^1([0,T))$, let $[f]^\circ(t) = \sum_{n\in\mathbb{Z}} f(t - nT)$ be the T-periodization of f. We can then expand $[f]^\circ(t)$ in a Fourier series. The sequence of Fourier coefficients of this T-periodic function are given by $\widehat{(f)^\circ}[n] = \frac{1}{T}\widehat{f}\left(-\frac{n}{T}\right)$. We have

$$T \sum_{n\in\mathbb{Z}} f(t + nT) = \sum_{n\in\mathbb{Z}} \widehat{f}(n/T) e^{2\pi i n t/T}. \tag{PSF1}$$

Therefore,

$$T \sum_{n\in\mathbb{Z}} f(nT) = \sum_{n\in\mathbb{Z}} \widehat{f}(n/T).$$

Thus, the Poisson Summation Formula allows us to compute the Fourier series of $[f]^\circ$ in terms of the Fourier transform of f at equally spaced points. This extends to the Schwartz class of distributions as

$$\widehat{\sum_{n\in\mathbb{Z}} \delta_{nT}} = \frac{1}{T} \sum_{n\in\mathbb{Z}} \delta_{n/T}. \tag{PSF2}$$

If $f \in \mathbb{PW}_\Omega$ and $T \leq 1/2\Omega$,

$$\widehat{f}(\omega) = \left(\sum_{n\in\mathbb{Z}} \widehat{f}(\omega - \frac{n}{T}) \right) \cdot \chi_{[-1/2T, 1/2T)}(\omega).$$

But, by computing transforms and applying PSF2,

$$\widehat{f}(\omega) = \left(\sum_{n\in\mathbb{Z}} \widehat{f}(\omega - \frac{n}{T})\right) \cdot \chi_{[-1/2T,1/2T)}(\omega) = \left(\sum_{n\in\mathbb{Z}} \left[\delta_{n/T}\right] \widehat{f}\right) \cdot \chi_{[-1/2T,1/2T)}(\omega)$$

if and only if

$$f(t) = T\left(\left[\sum_{n\in\mathbb{Z}} \delta_{nT}\right] f\right) * \operatorname*{sinc}_T(t)$$

(see [1], pp. 254–257). An additional bonus to this derivation is that it gives a direct method for analyzing reconstruction errors.

Complete reconstruction requires samples over all time. If only a finite number of the samples are used, we get *truncation error* $\mathscr{E}_N$. This can be computed for uniform truncation as follows. Define

$$f_N(t) = T\sum_{n=-N}^{N} f(nT)\frac{\sin(\frac{\pi}{T}(t-nT))}{\pi(t-nT)}.$$

The L^2 truncation error for f is $E_N = \|f - f_N\|_2^2 = T\sum_{|n|>N} |f(nT)|^2$ (see [32], p. 142)[2]. The function $f - f_N$ is bandlimited with finite energy. Therefore, we have the pointwise estimate

$$\mathscr{E}_N = \sup|f(t) - f_N(t)| \le (TE_N)^{1/2}$$

(see [32], p. 142). More sophisticated analysis is needed for non-uniform truncation and/or missing sample blocks (see [15], Chapter 11, and/or [22], Section 4).

The sampling rate $1/2\Omega$ is called the *Nyquist rate*. Sampling sub-Nyquist results in *aliasing error* $\mathscr{E}_A$, described in the following. If f has bandlimit Ω, and we sample at rate $T > 1/2\Omega$, high frequencies of one block of $e^{2\pi nt/T} f(t)$ intersect with low frequencies of the next block $e^{2\pi(n+1)t/T} f(t)$. Aliasing results in a stroboscopic effect [1, pg. 258], an effect which is visualized as jumps in the output signal. The high and low frequencies of adjacent blocks alias each other. To analyze aliasing error, we compute the pointwise estimate. For simplicity, assume $f \in \mathbb{PW}_1$. If $T = \frac{1}{2}$, applying PSF1 and integrating gives

$$\int_{-1/2}^{1/2} [\widehat{f}]^\circ(\omega) e^{2\pi i t\omega}\, d\omega = \sum_{n\in\mathbb{Z}} f(n) \int_{-1/2}^{1/2} e^{2\pi i(t-n)\omega}\, d\omega = \sum_{n\in\mathbb{Z}} f(n)\frac{\sin(\pi(t-n)}{\pi(t-n)}.$$

If $T > \frac{1}{2}$, then

$$\int_{-1/2}^{1/2} [\widehat{f}]^\circ(\omega) e^{2\pi i t\omega}\, d\omega = \sum_{n\in\mathbb{Z}} \int_{-1/2}^{1/2} \widehat{f}(\omega+n) e^{2\pi i t\omega}\, d\omega$$

$$= \sum_{n\in\mathbb{Z}} \int_{n-1/2}^{n+1/2} \widehat{f}(u) e^{2\pi i t(u-n)}\, du = \sum_{n\in\mathbb{Z}} e^{2\pi i t(-n)} \int_{n-1/2}^{n+1/2} \widehat{f}(u) e^{2\pi i t u}\, du.$$

[2] Computations throughout the chapter have been adjusted to compensate for our definition of the Fourier transform.

Now,

$$f(t) = \sum_{n\in\mathbb{Z}} \int_{n-1/2}^{n+1/2} \widehat{f}(u) e^{2\pi i t u}\, du\,.$$

Thus,

$$\begin{aligned}
\mathscr{E}_A &= \sup\left| f(t) - \int_{-1/2}^{1/2} [\widehat{f}]^{\circ}(\omega) e^{2\pi i t\omega}\, d\omega \right| \\
&= \sup\left| \sum_{n\neq 0} \left(1 - e^{2\pi i t(-n)}\right) \int_{n-1/2}^{n+1/2} \widehat{f}(u) e^{2\pi i t(u)}\, du \right| \\
&\leq 2 \sup\left[\sum_{n\neq 0} \int_{n-1/2}^{n+1/2} |\widehat{f}(u)|\, du \right] = 2 \int_{|u|\geq 1/2} |\widehat{f}(u)| du\,.
\end{aligned}$$

The constant 2 is sharp. An analysis of this error bound in terms of operators can be found in Chapter 11 of [15].

If sample values are not measured at intended points, we can get *jitter error* $\mathscr{E}_J$. Let $\{\varepsilon_n\}$ denote the error in the nth sample point. First we note that if $f \in \mathbb{PW}(1)$, then, by *Kadec's 1/4 Theorem*, the set $\{n \pm \varepsilon_n\}_{n\in\mathbb{Z}}$ is a stable sampling set if $|\varepsilon_n| < 1/4$. Moreover, this bound is sharp. Jitter error is given by

$$\mathscr{E}_J = \sup\left| f(t) - T\left(\left[\sum_{n\in\mathbb{Z}} \delta_{nT\pm\varepsilon_n} \right] f \right) * \operatorname*{sinc}_T(t) \right|.$$

If we assume $|\varepsilon_n| \leq J \leq \min\{1/(4\Omega), e^{-1/2}\}$, then [15] (Chapter 11) shows that $\mathscr{E}_J \leq KJ\log(1/J)$, where K is a constant expressed in terms of the sup norm of f and f'.

2 Sampling via Projection

Adaptive frequency band and ultra-wideband systems require either rapidly changing or very high sampling rates. These rates stress signal reconstruction in a variety of ways. Clearly, sub-Nyquist sampling creates aliasing error, but error would also show up in truncation, jitter and amplitude, as computation is stressed. The W-K-S Sampling Theorem does not have a way to accurately reconstruct the signal for sub-Nyquist samples nor adjust the sampling rate for variable bandwidth signals. We can think of this as follows. Truncation loses the energy in the lost samples, aliasing introduces ambiguous information in the signal and extremely high sampling increases the likelihood of jitter error. In fact, perturbations of sampling sets of ultra-wideband signals can result in unstable sampling sets. Developing a sampling theory for adaptive frequency band and ultra-wideband systems is the motivation for the methods in this chapter. Two of the key items needed for this approach are quick

and accurate computations of Fourier coefficients, which are computed in parallel [8, 18–20], and effective adaptive windowing systems [5], described in section 3.

2.1 An Introduction to the Projection Method

We start with a few "back of the envelope computations". Let χ_S denote the characteristic (or indicator) function of the set S. Let f be a signal of finite energy in the Paley-Wiener class $\mathbb{PW}_\Omega$. For a block of time T, let $f(t) = \sum_{k\in\mathbb{Z}} f(t)\chi_{[(k)T,(k+1)T]}(t)$. If we take a given block $f_k(t) = f(t)\chi_{[(k)T,(k+1)T]}(t)$, we can T-periodically continue the function, getting

$$[f_k]^\circ(t) = [f(t)\chi_{[(k)T,(k+1)T]}(t)]^\circ .$$

Expanding $(f_k)^\circ(t)$ in a Fourier series, we get

$$[f_k]^\circ(t) = \sum_{n\in\mathbb{Z}} \widehat{[f_k]^\circ}[n]\exp(2\pi int/T) .$$

The original function f is Ω bandlimited; however, the truncated block functions f_k are not. Using the original Ω bandlimit gives us a lower bound on the number of nonzero Fourier coefficients $\widehat{[f_k]^\circ}[n]$ as follows. We have $\frac{n}{T} \leq \Omega$,i.e., ,$n \leq T\cdot\Omega$. So, choose $N = \lceil T\cdot\Omega\rceil$, where $\lceil\cdot\rceil$ denotes the ceiling function. For this choice of N, we compute

$$\begin{aligned} f(t) &= \sum_{k\in\mathbb{Z}} f(t)\chi_{[(k)T,(k+1)T]}(t) = \sum_{k\in\mathbb{Z}} \left[[f_k]^\circ(t)\right]\chi_{[(k)T,(k+1)T]}(t) \\ &\approx \sum_{k\in\mathbb{Z}} \left[\sum_{n=-N}^{n=N} \widehat{[f_k]^\circ}[n]\exp(2\pi int/T)\right]\chi_{[(k)T,(k+1)T]}(t) . \end{aligned}$$

Note that for this choice of the standard (sines, cosines) basis, we can, for a fixed value of N, adjust to a large bandwidth Ω by choosing small time blocks T. Also, after a given set of time blocks, we can deal with an increase or a decrease in bandwidth Ω by again adjusting the time blocks, e.g., given an increase in Ω, decrease T, and vice versa. There is, of course, a price to be paid. The quality of the signal, as expressed in the accuracy the representation of f, depends on N, Ω, and T. The basic projection formula is given as follows.

Proposition 1. *Let $f \in \mathbb{PW}_\Omega$ and let T be a fixed block of time. Then, for $N = \lceil T\cdot\Omega\rceil$,*

$$f(t) \approx f_p(t) = \sum_{k\in\mathbb{Z}} \left[\sum_{n=-N}^{N} \widehat{[f_k]^\circ}[n]\exp(2\pi int/T)\right]\chi_{[kT,(k+1)T]}(t). \tag{2}$$

It is now evident how this method approximates the signal. Unlike the Shannon method which examined the function at specific points, then used those individual points to recreate the curve, the projection method breaks the signal into time blocks and then approximates their respective periodic expansions with a Fourier series. This process allows the system to individually evaluate each piece and base its calculation on the needed bandwidth. The individual Fourier series are then summed, recreating a close approximation of the original signal. It is important to note that instead of fixing T, the method allows us to fix one of the parameters (N, Ω, or T) while allowing the other two to fluctuate. From the design point of view, the easiest and most practical parameter to fix is N. For situations in which the bandwidth does not need flexibility, it is possible to fix Ω and T by the equation $N = \lceil T \cdot \Omega \rceil$. For fixed N, to increase bandwidth Ω, decrease the time blocks T.

The projection method can adapt to changes in the signal. Suppose that the signal $f(t)$ has a bandlimit $\Omega(t)$ which changes with time. This change effects the time blocking $\tau(t)$ and the number of basis elements $N(t)$. This, of course, makes the analysis more complicated, but is at the heart of the advantage the projection method has over conventional methods. During a given $\tau(t)$, let $\overline{\Omega}(t) = \max\{\Omega(t) : t \in \tau(t)\}$. For a signal f that is $\Omega(t)$ band-limited, we can estimate the value of n for which $\widehat{[f_k]^\circ}[n]$ is nonzero. At minimum, $\widehat{[f_k]^\circ}[n]$ is non-zero if $\frac{n}{\tau(t)} \leq \overline{\Omega}(t)$, or equivalently, $n \leq \tau(t) \cdot \overline{\Omega}(t)$. Let $N(t) = \lceil \tau(t) \cdot \overline{\Omega}(t) \rceil$. For this choice of $N(t)$, we get the basic adaptive projection formula.

Proposition 2. *Let f, $\widehat{f} \in L^2(\mathbb{R})$ and f have a variable but bounded bandlimit $\Omega(t)$. Let $\tau(t)$ be an adaptive block of time. Given $\tau(t)$, let $\overline{\Omega}(t) = \max\{\Omega(t) : t \in \tau(t)\}$. Then, for $N(t) = \lceil \tau(t) \cdot \overline{\Omega}(t) \rceil$, $f(t) \approx f_{\mathscr{P}}(t)$, where*

$$f_{\mathscr{P}}(t) = \sum_{k \in \mathbb{Z}} \left[\sum_{n=-N(t)}^{N(t)} \widehat{[f_k]^\circ}[n] e^{(2\pi i n t/\tau)} \right] \chi_{[k\tau,(k+1)\tau]}(t). \tag{3}$$

The projection method also adapts to general orthonormal systems, much as Kramer-Weiss extends sampling to general orthonormal bases [22]. Given a function f such that $f \in \mathbb{PW}_\Omega$, let T be a fixed time block. Define $f(t)$ and $f_k(t)$ as in the beginning of the computation above. Now, let $\{\varphi_n\}$ be a general orthonormal system for $L^2[0,T]$, and let $\{\varphi_{n,k}(t)\} = \{\varphi_n(t - kT)\}$. Since $f \in \mathbb{PW}_\Omega$, there exists $N = N(T,\Omega)$ such that $\widehat{f_k}[n] = \langle f, \varphi_{n,k} \rangle = 0$ for all $n > N$. In fact, let $N = \max_n \langle f, \varphi_{n,k} \rangle \neq 0$. Expanding in a Fourier series relative to $\{\varphi_{n,k}\}$ gives $f_k(t) = \sum_{n \in \mathbb{Z}} \widehat{f_k}[n] \varphi_{n,k}(t)$, where $\widehat{f_k}[n] = \langle f_k, \varphi_{n,k} \rangle$. Summing over all blocks gives the following.

Proposition 3. *Let $\{\varphi_n\}$ be a general orthonormal system for $L^2[0,T]$ and let $\{\varphi_{n,k}(t)\} = \{\varphi_n(t - kT)\}$. Let $N = N(T,\Omega)$ be such that $\widehat{f_k}[n] = 0$ for all $n > N$. Therefore, $f(t) \approx f_{\mathscr{P}}(t)$, where*

$$f_{\mathscr{P}}(t) = \sum_{k=-\infty}^{\infty} \left[\sum_{n=-N}^{N} \langle f_k, \varphi_{n,k} \rangle \varphi_{n,k}(t) \right] \chi_{[kT,(k+1)T]}(t). \tag{4}$$

Given characteristics of the class of input signals, the choice of basis functions used in the projection method can be tailored to optimal representation of the signal or a desired characteristic in the signal.

2.2 Error Analysis

To compute truncation error, we calculate the Fourier transform of both sides of the projection formula. Let $f \in \mathbb{PW}_{\Omega}$, so $f \in L^2$ and Ω bandlimited. For $N = \lceil T \cdot \Omega \rceil$, we have $f(t) \approx f_p(t)$, where

$$f_p(t) = \sum_{k\in\mathbb{Z}} \left[\sum_{n=-N}^{N} \widehat{(f_k)^{\circ}}[n] \exp(2\pi int/T) \right] \chi_{[kT,(k+1)T]}(t).$$

Taking the transform of both sides of this last equation and evoking the relationship between the transform and convolution gives

$$\widehat{f_p}(\omega) = \sum_{k\in\mathbb{Z}} \left[\sum_{n=-N}^{N} [\widehat{(f_k)^{\circ}}[n] (\exp(2\pi int/T))\widehat{\ }] * \left[\chi_{[kT,(k+1)T]}(t) \right] \widehat{\ }(\omega) \right].$$

Performing the indicated transforms results in

$$\widehat{f_p}(\omega) = \sum_{k\in\mathbb{Z}} \left[\sum_{n=-N}^{N} \widehat{(f_k)^{\circ}}[n] \left(\delta(\omega - \frac{n}{T}) \right) \right] * \exp(2\pi i(k-(1/2))T\omega) \frac{\sin(\pi T\omega)}{\pi\omega}.$$

Applying $\left(\delta(\omega - \frac{n}{T})\right)$, we get the following.

Proposition 4. *Let $f \in \mathbb{PW}_{\Omega}$ and let T be a fixed block of time. Then, for $N = \lceil T \cdot \Omega \rceil$,*

$$\widehat{f_p}(\omega) = \sum_{k\in\mathbb{Z}} \left[\sum_{n=-N}^{N} \widehat{(f_k)^{\circ}}[n] \exp(2\pi i(k-(1/2))T)(\omega - \frac{n}{T}) \left(\frac{\sin(\pi(\omega - \frac{n}{T}))}{\pi(\omega - \frac{n}{T})} \right) \right]. \tag{5}$$

The sharp truncations $f \cdot \chi_{[kT,(k+1)T]}$ introduce high-frequency modulation terms into the signal. These terms are the primary source of error in the projection method and the motivation for developing the windowing system discussed in Section 3.

The number N in the projection formula is the number of Fourier series components chosen in the sum on each block. In order to ensure maximum utility from the formula, the difference between the infinitely summed series and the truncated series must be made a minimum. We calculate the mean square error as a truncation error on the number of Fourier coefficients used to represent a given block f_k. For a fixed N, the mean square error is

$$E_N^2 = \|f_k - f_{k,N}\|_2^2 = \|\widehat{f_k} - \widehat{f_{k,N}}\|_2^2.$$

Computing and then simplifying gives

$$E_N^2 = \frac{1}{T}\int_{kT}^{(k+1)T} |f_k{}^{\circ}(t) - \sum_{|n|\le N} \widehat{f_k}[n]\exp(2\pi itn/T)|^2 dt$$
$$= \frac{1}{T}\int_{kT}^{(k+1)T} |\sum_{|n|>N} \widehat{f_k}[n]\exp(2\pi itn/T)|^2 dt = \sum_{|n|>N} |\widehat{f_k}[n]|^2 .$$

Truncation error perpetuates over all the blocks.

Given a general orthonormal system $\{\varphi_j\}$ for $L^2[0,T]$, we can create an orthonormal system for $L^2(\mathbb{R})$ by translating, getting $\{\varphi_{n,k}(t)\} = \{\varphi_n(t-kT)\}$. We can then analyze error generated by projection as

$$\mathscr{E}_{\mathscr{P}} = \|f(t) - \sum_{k\in\mathbb{Z}}\left[\sum_{j=-N}^{N} \langle f_k, \varphi_{n,k}\rangle \varphi_{n,k}(t)\right]\chi_{[kT,(k+1)T]}(t)\|_2$$
$$= \|\sum_{k\in\mathbb{Z}}\left[\sum_{|n|>N} \langle f_k, \varphi_{n,k}\rangle \varphi_{n,k}(t)\right]\chi_{[kT,(k+1)T]}(t)\|_2 .$$

The sharp cut-offs $\chi_{[kT,(k+1)T]}$ have a decay of only $\mathscr{O}(1/\omega)$ in frequency[3]. The windowing systems introduced in the next section greatly diminish this error.

3 Bounded Adaptive Partitions

This section introduces methods for segmenting time-frequency ($\mathbb{R} - \hat{\mathbb{R}}$) space that are developed to minimize the "ringing" in the signal introduced by sharp cutoffs. We develop windowing systems have variable partitioning length, variable roll-off, and variable smoothness. Three types of systems are constructed. The first systems preserve orthogonality of any orthonormal systems between adjacent blocks. These are used to develop tiling systems which cut up time into segments of possibly varying length, where the length is determined by signal bandwidth. The techniques developed give control over smoothness in time and corresponding decay in frequency. We also develop our systems so that the orthogonality of bases in adjacent and possible overlapping blocks is preserved. The "Achilles heel" of these systems is that they are difficult to compute. We then construct smooth Bounded Adaptive Partitions of unity using B-splines. These systems give flexible adaptive partitions of unity of variable smoothness. Finally, we use the concept of *almost orthogonality* (Cotlar, Knapp and Stein) [13] and our B-spline techniques to create almost orthogonal windowing systems that are more computable/constructible than the orthogonality preserving systems.

[3] Given two functions f and g, we say that $f = \mathscr{O}(g)$ if there exists a positive constant K such that $f(\omega) < Kg(\omega)$ for ω sufficiently large.

Our first windowing system uses sine, cosine, and linear functions. This was created because it is relatively easy to implement, cuts down on frequency error and preserves orthogonality. Consider a signal block of length $T+2r$ centered at the origin. Let $0 < r \ll T/2$. Define $\mathrm{Cap}(t)$ as follows:

$$\mathrm{Cap}(t) = \begin{cases} 0 & |t| \geq T/2+r \\ 1 & |t| \leq T/2-r \\ \sin(\frac{\pi}{4r}(t+(T/2+r))) & -T/2-r < t < -T/2+r \\ \cos(\frac{\pi}{4r}(t-(T/2-r))) & T/2-r < t < T/2+r\,. \end{cases} \tag{6}$$

Given Cap, we form a system $\{\mathrm{Cap}_k(t)\}$ such that $\mathrm{supp}(\mathrm{Cap}_k(t)) \subseteq [kT-r,(k+1)T+r]$ for all k. Note that the Cap window has several properties that make it a good window for our purposes. It has a partition property in that it windows the signal in $[-T/2-r,T/2+r]$ and is identically 1 on $[-T/2+r,T/2-r]$. It has a continuous roll-off at the endpoints. Finally, it has the property that for all $t \in \mathbb{R}$

$$\sum_k [\mathrm{Cap}_k(t)]^2 \equiv 1\,.$$

This last condition is needed to preserve the orthogonality of basis elements between adjacent blocks. Additionally, it has $\mathscr{O}(1/\omega^2)$ decay in frequency space, and, when one time block is ramping down, the adjacent block is ramping up at exactly the same rate. The system using overlapping Cap functions has the advantage of $\mathscr{O}(1/\omega^2)$ decay in frequency. We have that, for example, letting $T=2$ and $r=1$

$$\widehat{\mathrm{Cap}}(\omega) = \left[\frac{\sin(2\pi\omega)+4\omega\cos(4\pi\omega)}{\pi\omega(16\omega^2-1)}\right]. \tag{7}$$

Again let $[f]^\circ$ be the $T+2r$ periodization of f. Because both $[f]^\circ$ and Cap have absolutely converging Fourier series,

$$\widehat{[f\cdot\mathrm{Cap}]^\circ}[n] = \sum_m \widehat{[f]^\circ}[n-m]\widehat{\mathrm{Cap}}[m] = \widehat{[f]^\circ} * \widehat{\mathrm{Cap}}[n]\,.$$

3.1 Orthogonality Preserving Systems

The theory of splines and some techniques from ordinary differential equations give us the tools to generalize this system. The idea is to cut up the time domain into perfectly aligned segments so that there is no loss of information. We want the systems to be smooth, so as to provide control over decay in frequency, have variable cutoff functions for flexibility in design, and adaptive, so as to adjust accordingly to changes in frequency. We also want to develop our systems so that the orthogonality of bases in adjacent and possible overlapping blocks is preserved.

Definition 5 (ON Window System). Let $0 < r \ll T$. An *ON Window System* is a set of functions $\{\mathbb{W}_k(t)\}$ such that for all $k \in \mathbb{Z}$

$$\begin{aligned}
&(i.)\ \ \mathrm{supp}(\mathbb{W}_k(t)) \subseteq [kT-r,(k+1)T+r]\,,\\
&(ii.)\ \ \mathbb{W}_k(t) \equiv 1 \text{ for } t \in [kT+r,(k+1)T-r]\,,\\
&(iii.)\ \ \mathbb{W}_k((kT+T/2)-t) = \mathbb{W}_k(t-(kT+T/2)) \text{ for } t \in [0,T/2+r]\,,\\
&(iv.)\ \ \sum [\mathbb{W}_k(t)]^2 \equiv 1\,,\\
&(v.)\ \ \{\widehat{[\mathbb{W}_k]^\circ}[n]\} \text{ is absolutely summable, i.e., } \{\widehat{[\mathbb{W}_k]^\circ}[n]\} \in l^1\,.
\end{aligned} \tag{8}$$

Conditions $(i.)$ and $(ii.)$ are partition properties in that they give an exact snapshot of the input function f on $[kT+r,(k+1)T-r]$, with smooth roll-off at the edges. Condition $(iii.)$ is needed to preserve orthogonality between adjacent blocks. Condition $(iv.)$ simplifies computations, and condition $(v.)$ is needed for the computation of Fourier coefficients.

We will generate our systems by translations and dilations of a given window $\mathbb{W}_I$, where $\mathrm{supp}(\mathbb{W}_I) = [-T/2-r,T/2+r]$. Condition $(v.)$ gives, for $f \in \mathbb{PW}_\Omega$ and $\{\mathbb{W}_k(t)\}$ an ON Window System with generating window $\mathbb{W}_I$, that

$$\frac{1}{T+2r}\int_{-T/2-r}^{T/2+r} [f\cdot\mathbb{W}_I]^\circ(t)\exp(-2\pi int/[T+2r])\,dt = \widehat{[f]^\circ} * \widehat{\mathbb{W}_I}[n]\,. \tag{9}$$

Examples:

- $\{\mathbb{W}_k(t)\} = \bigcup_{k\in\mathbb{Z}} \chi_{[(k)T,(k+1)T]}(t)\,.$
- $\{\mathbb{W}_k(t)\} = \bigcup_{k\in\mathbb{Z}} \mathrm{Cap}_{[(k)T-r,(k+1)T+r]}(t)\,.$

The first example has jump discontinuities at all segment boundaries and has $\mathscr{O}(1/\omega)$ decay in frequency. The second is continuous but not differentiable and has overlaps at segment boundaries. This system has $\mathscr{O}(1/\omega^2)$ decay in frequency.

The general window function $\mathbb{W}_I$ is k-times differentiable, has $\mathrm{supp}(\mathbb{W}_I) = [-T/2-r,T/2+r]$, and has values

$$\mathbb{W}_I = \begin{cases} 0 & |t| \geq T/2+r \\ 1 & |t| \leq T/2-r \\ \rho(\pm t) & T/2-r < |t| < T/2+r\,. \end{cases} \tag{10}$$

We solve for $\rho(t)$ by solving the Hermite interpolation problem

$$\begin{cases} (a.)\ \rho(T/2-r) = 1\,, \\ (b.)\ \rho^{(n)}(T/2-r) = 0\,, n = 1,2,\ldots,k\,, \\ (c.)\ \rho^{(n)}(T/2+r) = 0\,, n = 0,1,2,\ldots,k\,, \end{cases}$$

with the conditions that $\rho \in C^k$ and

$$[\rho(t)]^2 + [\rho(-t)]^2 = 1 \text{ for } [T/2-r] \leq |t| \leq [T/2+r]\,. \tag{11}$$

We refer to (11) as the *sin − cos condition*. It directs us to get solutions expressed in terms of $\sin(t)$ and $\cos(t)$. Therefore, a way to solve this interpolation problem is the method of undetermined coefficients. We demonstrate by solving for ρ so that we have a C^1 window. Start by assuming that

$$\rho(t) = A\sin(B[T/2-t]) + C\,,\, T/2 \le t \le T/2 + r\,.$$

Since ρ is C^1, $\rho'(T/2+r) = 0$, and so $AB\cos(B[r]) = 0$, giving $B = \pi/2r$. Window condition $(iv.)$ gives that $2[\rho(t/2)]^2 = 1$, and so $C = \sqrt{2}/2$. Finally, $\rho(T/2+r) = 0$, and so $A = -\sqrt{2}/2$.

To extend ρ onto $T/2 - r \le t \le T/2$, we again use window condition $(iv.)$, getting

$$\rho(t) = \sqrt{\left[1 - \frac{1}{2}\left[1 - \sin(\frac{\pi}{2r}(\frac{T}{2} - t))\right]^2\right]}\,,\, \frac{T}{2} - r \le t \le \frac{T}{2}\,.$$

Finally, we use window conditions $(ii.)$ and $(iii.)$, getting

$$\rho(t) = \begin{cases} \frac{\sqrt{2}}{2}\left[1 - \sin(\frac{\pi}{2r}(\frac{T}{2} - t))\right] & \frac{-T}{2} - r \le t \le \frac{-T}{2} \\ \sqrt{\left[1 - \frac{1}{2}\left[1 - \sin(\frac{\pi}{2r}(t - \frac{T}{2}))\right]^2\right]} & \frac{-T}{2} \le t \le \frac{-T}{2} + r \\ 1 & \frac{-T}{2} + r < t < \frac{T}{2} - r \\ \sqrt{\left[1 - \frac{1}{2}\left[1 - \sin(\frac{\pi}{2r}(\frac{T}{2} - t))\right]^2\right]} & \frac{T}{2} - r \le t \le \frac{T}{2} \\ \frac{1}{\sqrt{2}}\left[1 - \sin(\frac{\pi}{2r}(t - \frac{T}{2}))\right] & \frac{T}{2} \le t \le \frac{T}{2} + r\,. \end{cases} \tag{12}$$

With each degree of smoothness, we get an additional degree of decay in frequency.

3.2 Orthogonality Between Blocks

An ON Window System $\{\mathbb{W}_k(t)\}$ preserves orthogonality of basis element of overlapping blocks. Because of the partition properties of these systems, we need only check orthogonality of adjacent overlapping blocks. Our construction involves the folding technique developed by Malvar [26, 27], Coifman and Meyer [9] and Jawerth and Sweldens [21]. We develop our systems constructively by using spline theory. The best way to think about the construction is to visualize how one would do the extension for a system of sines and cosines. We would extend the odd reflections about the left endpoint and the even reflections about the right.

Let $\{\varphi_j(t)\}$ be an orthonormal basis for $L^2[-T/2,T/2]$. Define

$$\widetilde{\varphi}_j(t) = \begin{cases} 0 & |t| \geq T/2+r \\ \varphi_j(t) & |t| \leq T/2 \\ -\varphi_j(-T-t) & -T/2-r < t < -T/2 \\ \varphi_j(T-t) & T/2 < t < T/2+r\,. \end{cases} \tag{13}$$

Theorem 2 (The Orthogonality of Overlapping Blocks). *The collection* $\{\Psi_{k,j}\} = \{\mathbb{W}_k\widetilde{\varphi}_j(t)\}$ *forms an orthonormal basis for* $L^2(\mathbb{R})$.

Proof. Since $\mathbb{W}_I \in L^2[-T/2-r,T/2+r]$,

$$\|\Psi_{k,j}\|_2 = \|\mathbb{W}_I\|_2\|\widetilde{\varphi}_j\|_2 < \infty\,.$$

We want to show that $\langle \Psi_{k,j}, \Psi_{m,n}\rangle = \delta_{k,m}\cdot\delta_{j,n}$. The partitioning properties of the windows give that we need only check overlapping and adjacent windows. If $k=m$, we can consider the window centered at the origin and the basis $\widetilde{\varphi}_j$. We want to show $\langle \mathbb{W}_I\widetilde{\varphi}_i, \mathbb{W}_I\widetilde{\varphi}_j\rangle = \delta_{i,j}$. Computing, we have

$$\begin{aligned} \langle \mathbb{W}_I\widetilde{\varphi}_i, \mathbb{W}_I\widetilde{\varphi}_j\rangle &= \int_{-T/2-r}^{-T/2} (\mathbb{W}_I(t))^2\varphi_i(-T-t)\varphi_j(-T-t)\,dt \\ &+ \int_{-T/2}^{-T/2+r} ((\mathbb{W}_I(t))^2-1)\varphi_i(t)\varphi_j(t)\,dt \\ &+ \int_{-T/2+r}^{T/2-r} \varphi_i(t)\varphi_j(t)\,dt \\ &+ \int_{T/2-r}^{T/2} ((\mathbb{W}_I(t))^2-1)\varphi_i(t)\varphi_j(t)\,dt \\ &+ \int_{T/2}^{T/2+r} (\mathbb{W}_I(t))^2\varphi_i(T-t)\varphi_j(T-t)\,dt\,. \end{aligned} \tag{14}$$

Since $\{\varphi_j\}$ in an orthonormal basis, the third integral equals 1. We apply the linear change of variables $t=-T/2-\tau$ to the first integral and $t=-T/2+\tau$ to the second integral. We then add these two integrals together to get

$$\int_0^r [(\mathbb{W}_I(T/2-\tau))^2 + (\mathbb{W}_I(\tau-T/2))^2 - 1]\varphi_i(-T/2+\tau)\varphi_j(-T/2+\tau)\,d\tau\,.$$

Conditions (*iii.*) and (*iv.*) of our windowing system give that the expression

$$[(\mathbb{W}_I(T/2-\tau))^2 + (\mathbb{W}_I(\tau-T/2))^2 - 1]$$

equals zero, and therefore this integral equals zero. Applying the linear change of variables $t=T/2-\tau$ to the fourth integral and $t=T/2+\tau$ to the fifth integral gives that these two integrals also sum to zero by essentially the same argument.

We now need to verify that $\langle \mathbb{W}_k\widetilde{\varphi}_i, \mathbb{W}_l\widetilde{\varphi}_j\rangle = \delta_{k,l}\cdot\delta_{i,j}$. The partitioning properties of the windows give that we need only check adjacent windows. The symmetry of

our construction allows us to check $\mathbb{W}_{-1}$ and $\mathbb{W}_0$, where we need to only check the overlapping region $t \in [-r, r]$. We have that

$$\begin{aligned}\left\langle \mathbb{W}_{-1}\widetilde{\varphi}_i, \mathbb{W}_0\widetilde{\varphi}_j \right\rangle &= 0 + \int_{-r}^{0} (\mathbb{W}_{-1}(t))\varphi_i(t)(\mathbb{W}_0(t))(-\varphi_j(-t))\,dt \\ &\quad + \int_0^r (\mathbb{W}_{-1}(t))\varphi_i(-t)(\mathbb{W}_0(t))\varphi_j(t)\,dt\,. \end{aligned} \tag{15}$$

Applying the linear change of variables $t = -\tau$ to the first integral and substituting the variable τ and adding gives

$$\int_0^r [-\mathbb{W}_{-1}(-\tau)\mathbb{W}_0(-\tau) + \mathbb{W}_{-1}(\tau)\mathbb{W}_0(\tau)]\varphi_i(-\tau)\varphi_j(\tau)\,d\tau\,.$$

Condition (*iii.*) of our windowing system gives that the expression

$$[-\mathbb{W}_{-1}(-\tau)\mathbb{W}_0(-\tau) + \mathbb{W}_{-1}(\tau)\mathbb{W}_0(\tau)]$$

equals zero, and so the integral equals zero. Combining these two computations shows that

$$\left\langle \Psi_{k,j}, \Psi_{m,n} \right\rangle = \delta_{k,m} \cdot \delta_{j,n}\,.$$

To finish we have to show that $\{\Psi_{k,j}\}$ spans $L^2(\mathbb{R})$. Given any function $f \in L^2$, consider the windowed element $f_k(t) = \mathbb{W}_k(t) \cdot f(t)$. We first consider the expansion in the window $\mathbb{W}_I$ symmetric to the origin. Let $f_I(t) = \mathbb{W}_I(t) \cdot f(t)$. We have that $\{\varphi_j(t)\}$ is an orthonormal basis for $L^2[-T/2, T/2]$. Given f_I, define

$$\bar{f}_I(t) = \begin{cases} 0 & |t| \geq T/2 + r \\ f_I(t) & |t| \leq T/2 \\ f_I(t) - f_I(-T-t) & -T/2 - r < t < -T/2 \\ f_I(t) + f_I(T-t) & T/2 < t < T/2 + r\,. \end{cases} \tag{16}$$

Since $\bar{f}_I \in L^2[-T/2, T/2]$, we may expand it as

$$\sum_{j=1}^{\infty} \left\langle \bar{f}_I, \varphi_j \right\rangle \varphi_j(t)\,.$$

To extend this to $L^2[-T/2 - r, T/2 + r]$, we expand using $\{\widetilde{\varphi}_j(t)\}$, getting

$$\widetilde{\bar{f}}_I = \sum_{j=1}^{\infty} \left\langle \bar{f}_I, \varphi_j \right\rangle \widetilde{\varphi}_j(t)\,. \tag{17}$$

where

$$\widetilde{\bar{f}}_I(t) = \begin{cases} 0 & |t| \geq T/2 + r \\ f_I(t) & |t| \leq T/2 \\ f_I(t) - f_I(-T-t) & -T/2 - r < t < -T/2 \\ f_I(t) + f_I(T-t) & T/2 < t < T/2 + r\,. \end{cases} \tag{18}$$

This construction preserves orthogonality between adjacent blocks.

To finish, let f be any function in L^2. Consider the windowed element $f_k(t) = \mathbb{W}_k(t) \cdot f(t)$. Repeat the construction above for this window. This shows that for fixed k, $\{\Psi_{k,j}\}$ spans $L^2([kT - r, (k+1)T + r])$ and preserves orthogonality between adjacent blocks on either side. Summing over all $k \in \mathbb{Z}$ gives that $\{\Psi_{k,j}\}$ is an orthonormal basis for $L^2(\mathbb{R})$. □

Recall that an operator U is *unitary* if its transpose is its inverse, i.e., $U^* = U^{-1}$. The folding operation used in Theorem 2 can be written down as a unitary operator. Jawerth and Sweldens [21] wrote the operator down as a product of operators. Fix a point α in $\mathbb{R}$, and define the reflection, or mirror in α as $\mathscr{M}_\alpha f(t) = f(2\alpha - t)$. Let $\chi^l_\alpha = \chi_{(-\infty,\alpha]}$ and $\chi^r_\alpha = \chi_{(\alpha,\infty)}$ be the left and right cutoff functions, and let ρ^u_α and ρ^d_α be the up and down ramp functions. In terms of these operators, $\mathscr{M}_0 \mathbb{W}_I(t) = \mathbb{W}_I(t)$ and $\mathscr{M}_0 \rho^u(t) = \rho^d(t)$.

Definition 6 (Folding). The *folding operation* about a point α is given by

$$\mathscr{F}_\alpha = \chi^l_\alpha(1 + \mathscr{M}_\alpha)\rho^u_\alpha + \chi^r_\alpha(1 - \mathscr{M}_\alpha)\rho^d_\alpha . \tag{19}$$

Lemma 1. *The folding operator is unitary if and only if the* sin-cos *condition holds, i.e.,*

$$(\rho^u_\alpha)^2 + (\rho^d_\alpha)^2 = 1 . \tag{20}$$

Proof. The adjoint of $\mathscr{F}_\alpha$ is given by

$$\mathscr{F}^*_\alpha = \chi^l_\alpha(1 - \mathscr{M}_\alpha)\rho^u_\alpha + \chi^r_\alpha(1 + \mathscr{M}_\alpha)\rho^d_\alpha = \rho^u_\alpha(1 + \mathscr{M}_\alpha)\chi^l_\alpha + \rho^d_\alpha(1 - \mathscr{M}_\alpha)\chi^r_\alpha .$$

Thus,

$$\begin{aligned}
\mathscr{F}^*_\alpha \mathscr{F}_\alpha &= \rho^u_\alpha(1 + \mathscr{M}_\alpha)\chi^l_\alpha(1 + \mathscr{M}_\alpha)\rho^u_\alpha + \rho^d_\alpha(1 - \mathscr{M}_\alpha)\chi^r_\alpha(1 - \mathscr{M}_\alpha)\rho^d_\alpha \\
&= (\rho^u_\alpha)^2[\chi^l_\alpha + \chi^r_\alpha] + \rho^u_\alpha \chi^r_\alpha \rho^d_\alpha \mathscr{M}_\alpha + \rho^u_\alpha \chi^l_\alpha \rho^d_\alpha \mathscr{M}_\alpha \\
&\quad + (\rho^d_\alpha)^2[\chi^l_\alpha + \chi^r_\alpha] - \rho^u_\alpha \chi^r_\alpha \rho^d_\alpha \mathscr{M}_\alpha - \rho^u_\alpha \chi^l_\alpha \rho^d_\alpha \mathscr{M}_\alpha \\
&= \mathscr{F}_\alpha \mathscr{F}^*_\alpha .
\end{aligned}$$

□

We now have the results to state two of the main theorems of the chapter. Given characteristics of the class of input signals, the choice of basis functions used in the projection method can be tailored to optimal representation of the signal or a desired characteristic in the signal.

Theorem 3 (Projection Formula for ON Windowing). *Let $\{\mathbb{W}_k(t)\}$ be an ON Window System, and let $\{\Psi_{k,j}\}$ be an orthonormal basis that preserves orthogonality between adjacent windows. Let $f \in \mathbb{PW}_\Omega$ and $N = N(T, \Omega)$ be such that $\langle f, \Psi_{k,n} \rangle = 0$ for all $n > N$ and all k. Then, $f(t) \approx f_{\mathscr{P}}(t)$, where*

$$f_{\mathscr{P}}(t) = \sum_{k \in \mathbb{Z}} \left[\sum_{n=-N}^{N} \langle f, \Psi_{k,n} \rangle \Psi_{k,n}(t) \right] . \tag{21}$$

Given the flexibility of our windowing system, we can also formulate an adaptive projection system for the ON Window Systems.

Theorem 4 (Adaptive Projection Formula for ON Windowing). *Let f, $\widehat{f} \in L^2(\mathbb{R})$ and f have a variable but bounded bandlimit $\Omega(t)$. Let $\tau(t)$ be an adaptive block of time. Let $\{\mathbb{W}_k(t)\}$ be an ON Window System with window size $\tau(t)+2r$ on the kth block, and let $\{\Psi_{k,j}\}$ be an orthonormal basis that preserves orthogonality between adjacent windows. Given $\tau(t)$, let $\overline{\Omega}(t) = \max\{\Omega(t) : t \in \tau(t)\}$. Let $N(t) = N(\tau(t), \Omega(t))$ be such that $\langle f, \Psi_{k,n}\rangle = 0$. Then, $f(t) \approx f_{\mathscr{P}}(t)$, where*

$$f_{\mathscr{P}}(t) = \sum_{k\in\mathbb{Z}} \left[\sum_{n=-N(t)}^{N(t)} \langle f, \Psi_{k,n}\rangle \Psi_{k,n}(t) \right]. \tag{22}$$

Examples:
Let $\mathscr{T}_\alpha$ be the translation operator, i.e., $\mathscr{T}_\alpha[f](t) = f(t-\alpha)$. All of the basis elements are presented in the interval $[T/2-r, T/2+r]$ centered at the origin. Therefore, the operator $\tau_{[(k)T+T/2]}$ will place the basis in the interval $[(k)T-r, (k+1)T+r]$. In the following, $\mathbb{W}_I(t)$ is the window centered at the origin, and φ_j is a basis element in that window.

- $\{\Psi_{k,j}\} = \{\mathscr{T}_{[(k)T+T/2]}[\mathbb{W}_I \varphi_j](t)\}$, where $\mathbb{W}_I(t) = \chi_{[-T/2,T/2]}(t)$ and

$$\varphi_j(t) = \exp(i\frac{2\pi j}{T}(t-T/2)).$$

- $\{\Psi_{k,j}\} = \{\mathscr{T}_{[(k)T+T/2]}[\mathbb{W}_I \widetilde{\varphi}_j](t)\}$, where $\mathbb{W}_I(t) = \mathrm{Cap}(t)$ and

$$\varphi_j(t) = \sqrt{\frac{2}{T}} \sin\left(\pi(j+1/2)\frac{(t+T/2)}{T}\right).$$

The first example has jump discontinuities at all segment boundaries and has $\mathscr{O}(1/\omega)$ decay in frequency. Note, as there is no overlap, basis elements are not folded. The second is continuous but not differentiable and has overlaps at segment boundaries. This system has $\mathscr{O}(1/\omega^2)$ decay in frequency.

The development of a C^1 system involves solving a Hermite interpolation problem for not only the window but also the folded basis elements. Using undetermined coefficients we solve for ρ so that the window is C^1, getting

$$\rho(t) = \begin{cases} \sqrt{\left[1-\frac{1}{2}\left[1-\sin(\frac{\pi}{2r}(\frac{T}{2}-t))\right]^2\right]} & \frac{T}{2}-r \le t \le \frac{T}{2}. \\ \frac{1}{\sqrt{2}}\left[1-\sin(\frac{\pi}{2r}(t-\frac{T}{2}))\right] & \frac{T}{2} \le t \le \frac{T}{2}+r. \end{cases}$$

We then use the same technique to solve for C^1 folded basis elements $\{\widetilde{\varphi}_j\}$. The constraints that make C^1 folded basis elements are

$$\begin{cases} (a.)\ \varphi_j(-T/2) = 0 \\ (b.)\ \varphi_j{}'(-T/2) \text{ exists} \\ (c.)\ \varphi_j{}'(T/2) = 0 \end{cases} \tag{23}$$

Constraint (23) directs us to get solutions expressed in terms of $\sin(t)$ and $\cos(t)$. Solving (23) for φ_j, we get

$$\varphi_j(t) = \sqrt{\frac{2}{T}} \sin\left(\pi(j+1/2)\frac{(t+T/2)}{T}\right). \tag{24}$$

Example: A C^1 system – $\{\Psi_{k,j}\} = \{\mathscr{T}_{[(k)T+T/2]}[\mathbb{W}_I\widetilde{\varphi}_j](t)\}$, where

$$\mathbb{W}_I = \begin{cases} 0 & |t| \geq T/2 + r \\ 1 & |t| \leq T/2 - r \\ \rho(\pm t) & T/2 - r < |t| < T/2 + r, \end{cases}$$

with

$$\rho(t) = \begin{cases} \sqrt{\left[1 - \frac{1}{2}\left[1 - \sin(\frac{\pi}{2r}(\frac{T}{2} - t))\right]^2\right]} & \frac{T}{2} - r \leq t \leq \frac{T}{2} \\ \frac{1}{\sqrt{2}}\left[1 - \sin(\frac{\pi}{2r}(t - \frac{T}{2}))\right] & \frac{T}{2} \leq t \leq \frac{T}{2} + r \end{cases}$$

and

$$\varphi_j(t) = \sqrt{\frac{2}{T}} \sin\left(\pi(j+1/2)\frac{(t+T/2)}{T}\right).$$

The computations become increasingly complicated as the parameter k increases. This is the motivation for creating "almost orthogonal" windowing systems using B-spline constructions. These B-spline constructions allow for a direct computation of the Fourier coefficients.

The analysis of the error generated by the projection method involves looking at the decay rates of the Fourier coefficients. If we are working with the standard basis, for $f \in C(\mathbb{T}_{2\Phi})$, we can define the modulus of continuity as

$$\mu(\delta) = \sup_{|x-y| \leq \delta} |f(x) - f(y)|$$

and have that

$$|\widehat{f}[n]| \leq \frac{1}{2}\mu(1/n).$$

We say that f satisfies a Hölder condition with exponent α if there exists a constant K such that

$$|f(x+\delta) - f(x)| \leq K\delta^{\alpha}.$$

If f is k-times continuously differentiable and f^k satisfies a Hölder condition with exponent α, then there exists a constant K such that

$$|\widehat{f}[n]| \leq K \frac{1}{n^{k+\alpha}}.$$

The sharp cutoffs $\chi_{[kT,(k+1)T]}$ have a decay of only $\mathscr{O}(1/\omega)$ in frequency. We designed the ON windowing systems so that the windows have decay $\mathscr{O}(1/(\omega)^{k+2})$ in frequency. Thus makes the error on each block summable.

We assume $\mathbb{W}_k$ is C^k. Therefore, $\widehat{\mathbb{W}_k}(\omega) = \mathscr{O}(1/(\omega)^{k+2})$. We will analyze the error $\mathscr{E}_{k_{\mathscr{P}}}$ on a given block. Let $M = \|(f \cdot \mathbb{W}_k)\|_2$. Then,

$$\begin{aligned}\mathscr{E}_{k_{\mathscr{P}}} &= \|(f(t) \cdot \mathbb{W}_k) - \left[\sum_{n=-N}^{N} \langle f, \Psi_{n,k}\rangle \Psi_{n,k}(t)\right] \mathbb{W}_k(t)\|_2 \\ &= \| \sum_{|n|>N} \langle f, \Psi_{n,k}\rangle \Psi_{n,k}(t) \mathbb{W}_k(t)\|_2 \leq \left[\sum_{|n|>N} \frac{M}{n^{k+2}}\right].\end{aligned}$$

3.3 Partition of Unity Systems

We can construct partition of unity windowing systems using similar techniques as those used in ON Window Systems. The theory of B-splines gives us the tools to create these systems.

The most straightforward system is created by $\{\chi_{[(k)T,(k+1)T]}(t)\}$ for $k \in \mathbb{Z}$. A second example can be developed by studying the de la Vallée-Poussin kernel used in Fourier series (see [24]). Consider a signal block of length $T + 2r$ at the origin. Let $0 < r \ll T/2$. Let

$$\begin{aligned}&\mathrm{Tri}_L(t) = \max\{[((T/(4r)) + r) - |t|/(2r)], 0\}, \\ &\mathrm{Tri}_S(t) = \max\{[((T/(4r)) + r - 1) - |t|/(2r)], 0\} \text{ and} \\ &\mathrm{Trap}_{[-T/2-r,T/2+r]}(t) = \mathrm{Tri}_L(t) - \mathrm{Tri}_S(t).\end{aligned} \tag{25}$$

The Trap function has perfect overlay in the time domain and $\mathscr{O}(1/\omega^2)$ decay in frequency space. When one time block is ramping down, the adjacent block is ramping up at exactly the same rate. The system using overlapping Trap functions has the advantage of $\mathscr{O}(1/\omega^2)$ decay in frequency. Let $\beta_L = \sqrt{T/(4r) + r}$, $\alpha_L = T/(4r) + r/2$, $\beta_S = \sqrt{T/(4r) + r - 1}$, $\alpha_S = T/(4r) - r/2$. The Fourier transform of Trap is

$$\widehat{\mathrm{Trap}}(\omega) = \left[(\beta_L)\frac{\sin(2\pi\alpha_L\omega)}{\pi\omega}\right]^2 - \left[(\beta_S)\frac{\sin(2\pi\alpha_S\omega)}{\pi\omega}\right]^2. \tag{26}$$

Definition 7 (Bounded Adaptive Partition of Unity). Let $0 < r \ll T$. A *Bounded Adaptive Partition of Unity* is a set of functions $\{\mathscr{B}_k(t)\}$ such that

$$
\begin{aligned}
&(i.)\ \ \mathrm{supp}(\mathscr{B}_k(t)) \subseteq [kT-r,(k+1)T+r] \text{ for all } k\,,\\
&(ii.)\ \ \mathscr{B}_k(t) \equiv 1 \text{ for } t \in [kT+r,(k+1)T-r] \text{ for all } k\,,\\
&(iii.)\ \ \sum \mathscr{B}_k(t) \equiv 1\,,\\
&(iv.)\ \ \{\widehat{[\mathscr{B}_k]^\circ}[n]\} \text{ is absolutely summable, i.e. } \{\widehat{[\mathscr{B}_k]^\circ}[n]\} \in l^1\,. \qquad (27)
\end{aligned}
$$

Conditions $(i.),(ii.)$, and $(iii.)$ make $\{\mathscr{B}_k(t)\}$ a bounded partition of unity. Condition $(iv.)$ is needed for the computation of Fourier coefficients. We have that

$$
\frac{1}{T+2r}\int_{-T/2-r}^{T/2+r} [f\cdot\mathscr{B}_I]^\circ(t)\exp(-2\pi int/[T+2r])\,dt = \widehat{[f]^\circ} * \widehat{\mathscr{B}_I}[n]\,. \qquad (28)
$$

Examples:

- $\{\mathscr{B}_k(t)\} = \bigcup_{k\in\mathbb{Z}} \chi_{[(k)T,(k+1)T]}(t)$
- $\{\mathscr{B}_k(t)\} = \bigcup_{k\in\mathbb{Z}} \mathrm{Trap}_{[(k)T-r,(k+1)T+r]}(t)\,.$

The first example has jump discontinuities at all segment boundaries and has $\mathscr{O}(1/\omega)$ decay in frequency. The second is continuous but not differentiable and has overlaps at segment boundaries. This system has $\mathscr{O}(1/\omega^2)$ decay in frequency.

We will generate our general systems by translations and dilations of a given window $\mathscr{B}_I$, where $\mathrm{supp}(\mathscr{B}_I) = [-T/2-r,T/2+r]$. The generating window function $\mathscr{B}_I$ is k-times differentiable, has $\mathrm{supp}(\mathscr{B}_I) = [-T/2-r,T/2+r]$, and has values

$$
\mathscr{B}_I = \begin{cases} 0 & |t| \geq T/2+r \\ 1 & |t| \leq T/2-r \\ \rho(\pm t) & T/2-r<|t|<T/2+r. \end{cases} \qquad (29)
$$

We solve for $\rho(t)$ by solving the Hermite interpolation problem

$$
\begin{cases} (a.)\ \rho(T/2-r) = 1 \\ (b.)\ \rho^{(n)}(T/2-r) = 0\,, n = 1,2,\ldots,k \\ (c.)\ \rho^{(n)}(T/2+r) = 0\,, n = 0,1,2,\ldots,k\,, \end{cases}
$$

with the conditions that $\rho \in C^k$ and

$$
[\rho(t)] + [\rho(-t)] = 1 \text{ for } t \in [T/2-r,T/2+r]\,. \qquad (30)
$$

We use B-splines as our cardinal functions. Let $0<\alpha\ll\beta$ and consider $\chi_{[-\alpha,\alpha]}$. We want the n-fold convolution of $\chi_{[\alpha,\alpha]}$ to fit in the interval $[-\beta,\beta]$. Then, we choose α so that $0<n\alpha<\beta$, and let

$$
\Psi(t) = \underbrace{\chi_{[-\alpha,\alpha]} * \chi_{[-\alpha,\alpha]} * \cdots * \chi_{[-\alpha,\alpha]}(t)}_{n-times}\,.
$$

The β-periodic continuation of this function, $\Psi^{\circ}(t)$, has the Fourier series expansion

$$\sum_{k\neq 0} \frac{\alpha}{n\beta}\left[\frac{\sin(\pi k\alpha/n\beta)}{2\pi k\alpha/n\beta}\right]^n \exp(\pi ikt/\beta)\,.$$

The C^k solution for ρ is given by a theorem of Schoenberg (see [34], pp. 7–8). Schoenberg solved the Hermite interpolation problem with endpoints -1 and 1. An interpolant that minimizes the Chebyshev norm is called the *perfect spline*. The perfect spline $S(t)$ for Hermite problem with endpoints -1 and 1 such that

$$S(1)=1\ ,\ S^{(n)}(1)=0\,,n=1,2,\ldots,k\ ,\ S^{(n)}(-1)=0\,,n=0,1,2,\ldots,k$$

is given by the integral of the function

$$M(x)=(-1)^n\sum_{j=0}^{k}\frac{\Psi(t-t_j)}{\phi'(t_j)}\,,$$

where Ψ is the $k-1$ convolution of characteristic functions, the knot points are $t_j=-\cos(\frac{\pi j}{n}), j=0,1,\ldots,n$, and $\phi(t)=\prod_{j=0}^{k}(t-t_j)$. Given these knots, we have to choose α sufficiently large, e.g., $\alpha>1$. If k is even, the midpoint occurs at the $k/2$ knot point. If k is odd, the midpoint occurs at the midpoint between the $k/2$ and $(k+1)/2$ knot points. We then have that

$$\rho(t)=S\circ\ell(t)\,,\ \text{where}\ \ell(t)=-\frac{1}{r}t+\frac{T}{2r}\,.$$

For this ρ, and for

$$\mathscr{B}_I=\begin{cases}0 & |t|\geq T/2+r\\ 1 & |t|\leq T/2-r\\ \rho(\pm t) & T/2-r<|t|<T/2+r\end{cases}$$

we have that $\widehat{\mathscr{B}_I}(\omega)$ is given by the antiderivative of a linear combination of functions of the form

$$\left[\frac{\sin(\pi k\alpha\omega/nT)}{2\pi k\alpha\omega/nT}\right]^n,$$

and therefore has decay $\mathscr{O}(1/\omega^{n+1})$ in frequency.

3.4 Almost Orthogonal Systems

The partition of unity systems do *not* preserve orthogonality between blocks. However, they are easier to compute in both time and frequency. Therefore, these systems can be used to approximate the Cap system with B-splines. We get windowing

systems that nearly preserve orthogonality. Each added degree of smoothness in time adds to the degree of decay in frequency.

Cotlar, Knapp and Stein introduced *almost orthogonality* via operator inequalities (see [13]). The concept allows us to create windowing systems that are more computable/constructible such as the Bounded Adaptive Partition of Unity Systems $\{\mathscr{B}_k(t)\}$ with the orthogonality preservation of the ON Window Systems $\{\mathbb{W}_k(t)\}$.

Definition 8 (Almost ON System). Let $0 < r \ll T$. An *Almost ON System* for adaptive and ultra-wideband sampling is a set of functions $\{\mathbb{A}_k(t)\}$ for which there exists $\delta, 0 \leq \delta < 1/2$ such that

$$\begin{aligned}
&(i.)\ \ \mathrm{supp}(\mathbb{A}_k(t)) \subseteq [kT - r, (k+1)T + r] \text{ for all } k\,,\\
&(ii.)\ \ \mathbb{A}_k(t) \equiv 1 \text{ for } t \in [kT + r, (k+1)T - r] \text{ for all } k\,,\\
&(iii.)\ \ \mathbb{A}_k((kT + T/2) - t) = \mathbb{A}_k(t - (kT + T/2)), t \in [0, T/2 + r]\,,\\
&(iv.)\ \ 1 - \delta \leq [\mathbb{A}_k(t))]^2 + [\mathbb{A}_{k+1}(t))]^2 \leq 1 + \delta \text{ for } t \in [kT, (k+1)T]\,,\\
&(v.)\ \ \{\widehat{\mathbb{A}_k^{\circ}}[n]\} \in l^1\,.
\end{aligned} \tag{31}$$

We start with $\mathrm{Cap}(t)$, where

$$\mathrm{Cap}(t) = \begin{cases} 0 & |t| \geq T/2 + r \\ 1 & |t| \leq T/2 - r \\ \sin(\frac{\pi}{4r}(t + (T/2 + r))) & -T/2 - r < t < -T/2 + r \\ \cos(\frac{\pi}{4r}(t - (T/2 - r))) & T/2 - r < t < T/2 + r\,. \end{cases}$$

Let $\Delta_{(T,r)} = \frac{T+2r}{m}$. By placing equidistant knot points

$$-T/2 - r = x_0, -T/2 - r + \Delta_{(T,r)} = x_1, \ldots, T/2 + r = x_m,$$

we can construct C^{m-1} polynomial splines S_{m+1} approximating

$$\mathrm{Cap}(t) \text{ in } [(-T/2 - r), (T/2 + r)]\,.$$

A theorem of Curry and Schoenberg gives that the set of B-splines

$$\{B_{-(m+1)}^{(m+1)}, \ldots, B_k^{(m+1)}\}$$

forms a basis for the space of degree $m+1$ polynomial splines S_{m+1} (see [29], pp. 98–99). Therefore,

$$\mathrm{Cap}(t) \approx \sum_{i=-(m+1)}^{k} a_i B_i^{(m+1)}\,.$$

Let

$$\delta = \left\| \sum_{i=-(m+1)}^{k} a_i B_i^{(m+1)} - \mathrm{Cap}(t) \right\|_{\infty}\,.$$

Then, $\delta < 1/2$, with the largest value for the piecewise linear spline approximation. Moreover, $\delta \longrightarrow 0$ as m and k increase. Thus, we get computable windowing systems that nearly preserve orthogonality. Each added degree of smoothness in time adds to the degree of decay in frequency.

4 Biorthogonal Constructions

The collection $\{\Psi_{k,j}\} = \{\mathbb{W}_k \widetilde{\varphi}_j(t)\}$ forms an orthonormal basis for $L^2(\mathbb{R})$. In this section, we develop the biorthogonal basis to $\{\Psi_{k,j}\}$.

Let $\mathbb{H}$ be a Hilbert space with norm $\|\cdot\|$. We say that a sequence of vectors $\{x_n\}$ in $\mathbb{H}$ is *complete* if given and $x \in \mathbb{H}$ such that $\langle x, x_n \rangle = 0$ implies that $x = 0$. A sequence $\{x_n\}$ is *minimal* if every element in the sequence lies outside of the closed linear span of the other elements. A sequence that is both minimal and complete is called *exact*. Clearly, a basis is exact.

Definition 9 (Biorthogonal). A sequence $\{y_m\}_{m=1}^{\infty}$ in a Hilbert space $\mathbb{H}$ is *biorthogonal* to a sequence $\{x_n\}_{n=1}^{\infty}$ if

$$\langle x_n, y_m \rangle = \delta_{n,m}.$$

By the Hahn-Banach theorem, a given $\{x_n\}$ will have a biorthogonal sequence $\{y_m\}$ if and only if $\{x_n\}$ is minimal. Thus, a basis $\mathscr{B} = \{x_n\}_{n=1}^{\infty}$ for $\mathbb{H}$ possesses a biorthogonal basis $\mathscr{B}^* = \{y_m\}_{m=1}^{\infty}$, Also, there exists M such that for all n

$$1 \leq \|x_n\| \|y_n\| \leq M$$

(see [41], pp. 19–20). Two bases $\mathscr{A} = \{x_n\}_{n=1}^{\infty}$ and $\mathscr{B} = \{y_m\}_{m=1}^{\infty}$ are said to be *equivalent* if

$$\sum_n c_n x_n \text{ is convergent if and only if } \sum_n c_n y_n \text{ is convergent.}$$

Equivalent bases have equivalent biorthogonal bases.

A sequence $\{x_n\} \in \mathbb{H}$ is called a *Bessel sequence* if there is a constant B such that for all $x \in \mathbb{H}$, $\sum_n |\langle x, x_n \rangle|^2 \leq B\|x\|^2$. A *Riesz basis* $\mathscr{B} = \{x_n\}_{n=1}^{\infty}$ for $\mathbb{H}$ is a bounded basis. It is also *unconditional* in that for all $x \in \mathbb{H}$, $x = \sum_n \langle x, x_n \rangle x_n$ converges unconditionally, i.e., regardless of the order in which the terms are summed. There are many characterizations of Riesz bases. A set $\mathscr{B}$ is a Riesz basis if and only if it is equivalent to $\mathscr{E}$, an orthonormal basis for $\mathbb{H}$. Also, $\mathscr{B}$ is a Riesz basis if and only if there exists $A, B > 0$ such that

$$A\|x\|^2 \leq \sum_n |\langle x, x_n \rangle|^2 \leq B\|x\|^2$$

(see [41], pp. 26–30). If $\mathscr{B}^* = \{y_m\}_{m=1}^{\infty}$ is biorthogonal to $\mathscr{B}$, then $\mathscr{B}^*$ is a basis and, for every $x \in \mathbb{H}$,

$$x = \sum_n \langle x, y_n \rangle x_n = \sum_n \langle x, x_n \rangle y_n$$

where both sums converge unconditionally.

Given a Riesz basis $\mathscr{B} = \{x_n\}_{n=1}^{\infty}$ with bounds $A, B > 0$ such that

$$A\|x\|^2 \leq \sum_n |\langle x, x_n \rangle|^2 \leq B\|x\|^2,$$

there exists a biorthogonal Riesz basis $\mathscr{B}^* = \{y_m\}_{m=1}^{\infty}$ with bounds $B^{-1}, A^{-1} > 0$ such that

$$B^{-1}\|x\|^2 \leq \sum_m |\langle x, y_m \rangle|^2 \leq A^{-1}\|x\|^2.$$

The collection $\{\Psi_{k,j}\} = \{\mathbb{W}_k \widetilde{\varphi}_j(t)\}$ forms an orthonormal basis for $L^2(\mathbb{R})$. Therefore, it has a biorthogonal Riesz basis $\{\Psi^*_{k,j}\}$. The basis is uniquely determined by the biorthogonality relationship

$$\langle \{\Psi_{m,n}\}, \{\Psi^*_{k,j}\} \rangle = \delta_{m,k} \cdot \delta_{n,j}.$$

Theorem 5 (Biorthogonal Basis). *The Riesz basis $\{\Psi_{k,j}\} = \{\mathbb{W}_k \widetilde{\varphi}_j(t)\}$ has a unique biorthogonal Riesz basis $\{\Psi^*_{k,j}\}$, with uniqueness given by the biorthogonality relationship*

$$\langle \{\Psi_{m,n}\}, \{\Psi^*_{k,j}\} \rangle = \delta_{m,k} \cdot \delta_{n,j}.$$

The basis $\{\Psi^*_{k,j}\}$ is given by

$$\{\Psi_{k,j}\} = \{\widetilde{\mathbb{W}}_k \widetilde{\varphi}_j(t)\},$$

where $\widetilde{\mathbb{W}}_k$ is the translation of the window

$$\widetilde{\mathbb{W}}_I = \begin{cases} 0 & |t| \geq T/2 + r, \\ 1 & |t| \leq T/2 - r, \\ \frac{\rho_0^u(t)}{\rho_0^{u\,2} + \rho_0^{d\,2}} & -T/2 - r < t < -T/2 + r, \\ \frac{\rho_0^d(t)}{\rho_0^{u\,2} + \rho_0^{d\,2}} & T/2 - r < t < T/2 + r. \end{cases} \tag{32}$$

Proof. Recall that the folding operator about a point α is given by

$$\mathscr{F}_\alpha = \chi^l_\alpha (1 + \mathscr{M}_\alpha)\rho^u_\alpha + \chi^r_\alpha (1 - \mathscr{M}_\alpha)\rho^d_\alpha.$$

Lemma 2. *Assume $0 < m \leq 1 \leq M$, and that*

$$m \leq (\rho^u_\alpha)^2 + (\rho^d_\alpha)^2 \leq M. \tag{33}$$

Then for

$$m = \min_t \sqrt{(\rho_\alpha^u)^2 + (\rho_\alpha^d)^2}\,, M = \max_t \sqrt{(\rho_\alpha^u)^2 + (\rho_\alpha^d)^2}\,, \tag{34}$$

$$m\|f\|_2^2 \le \|\mathscr{F}_\alpha f\|_2^2 \le M\|f\|_2^2\,. \tag{35}$$

Proof of Lemma.

$$\int_{\mathbb{R}} |\mathscr{F}_\alpha f|^2 dt = \int_{\mathbb{R}} (\rho_\alpha^u)^2 + (\rho_\alpha^d)^2 |f|^2 dt\,.$$

The result follows by estimating the integral. □

Lemma 3. *The inverse of the folding operator is again a folding operator.*

Proof of Lemma. Let $g = \mathscr{F}_\alpha f$. Writing this in matrix form,

$$\chi_\alpha^l \begin{bmatrix} g \\ \mathscr{M}_\alpha g \end{bmatrix} = \underbrace{\begin{bmatrix} \rho_\alpha^u & \rho_\alpha^d \\ -\rho_\alpha^d & \rho_\alpha^u \end{bmatrix}}_{(*)} \chi_\alpha^l \begin{bmatrix} f \\ \mathscr{M}_\alpha f \end{bmatrix}. \tag{36}$$

Inverting $(*)$, we have that

$$\widetilde{\rho_\alpha^u} = \frac{\rho_\alpha^u}{\rho_\alpha^{u\,2} + \rho_\alpha^{u\,2}}\,, \widetilde{\rho_\alpha^d} = \frac{\rho_\alpha^u}{\rho_\alpha^{u\,2} + \rho_\alpha^{u\,2}}\,, \tag{37}$$

and that $\widetilde{\mathscr{F}_\alpha} g = f$

$$\widetilde{\mathscr{F}_\alpha} = \chi_\alpha^l (1 + \mathscr{M}_\alpha)\widetilde{\rho_\alpha^u} + \chi_\alpha^r (1 - \mathscr{M}_\alpha)\widetilde{\rho_\alpha^d}\,. \tag{38}$$

□

The result follows by combining the two lemmas. □

5 Signal Adaptive Frame Theory

The theory of frames gives us the mathematical structure in which to express sampling via the projection method. All nonuniform sampling schemes could be expressed in terms of the language of frames. The work we discuss in this section is preliminary and will be developed in subsequent papers.

5.1 Frame Theory

The concept of a frame goes back to the work of Duffin and Schaeffer [11].

Definition 10 (Frame). A sequence of elements $\mathscr{F} = \{f_n\}_{n\in\mathbb{Z}}$ in a Hilbert space $\mathbb{H}$ is a *frame* if there exist constants A and B such that

$$A\|f\|^2 \leq \sum_{n\in\mathbb{Z}} |\langle f, f_n\rangle|^2 \leq B\|f\|^2 .$$

A and B are called the *upper and lower frame bounds*, respectively. Also, $\widetilde{A} = \sup\{A\}$, $\widetilde{B} = \inf\{B\}$ are called the *optimal frame bounds*. A frame is called **tight** if $A = B$, and *normalized tight* if $A = B = 1$. A frame is called **exact** if it ceases to be a frame when any of its elements are removed. Thus, an ON basis $\mathscr{E}$ for $\mathbb{H}$ is an exact normalized tight frame for $\mathbb{H}$.

5.2 $\mathbb{W}$, $\mathscr{B}$, and $\mathbb{A}$ Frames

The windowing systems above allow us to develop *Signal Adaptive Frame Theory*. The idea is as follows. If we work with the ON windowing system $\{\mathbb{W}_k(t)\}$, let $\{\Psi_{k,j}\}$ be an orthonormal basis that preserves orthogonality between adjacent windows. Let $f \in \mathbb{PW}_\Omega$ and $N = N(T,\Omega)$ be such that $\langle f, \Psi_n\rangle = 0$ for all $n > N$. Then,

$$f(t) = \sum_{k\in\mathbb{Z}} \left[\sum_{n\in\mathbb{Z}} \langle f, \Psi_{n,k}\rangle \Psi_{n,k}(t) \right]. \tag{39}$$

This also gives

$$\|f\|_2^2 = \sum_{k\in\mathbb{Z}} \left[\sum_{n\in\mathbb{Z}} |\langle f, \Psi_{n,k}\rangle|^2 \right]. \tag{40}$$

Given the flexibility of our windowing system, we can also formulate an adaptive projection system for the ON Window Systems. Let f, $\widehat{f} \in L^2(\mathbb{R})$ and f have a variable but bounded bandlimit $\Omega(t)$. Let $\tau(t)$ be an adaptive block of time. Let $\{\mathbb{W}_k(t)\}$ be an ON Window System with window size $\tau(t) + 2r$ on the kth block, and let $\{\Psi_{k,j}\}$ be an orthonormal basis that preserves orthogonality between adjacent windows.

Given $\tau(t)$, let $\overline{\Omega}(t) = \max\{\Omega(t) : t \in \tau(t)\}$. Let $N(t) = N(\tau(t), \Omega(t))$ be such that $\langle f, \Psi_{n,k}\rangle = 0$. Then,

$$f(t) = \sum_{k\in\mathbb{Z}} \left[\sum_{n\in\mathbb{Z}} \langle f, \Psi_{n,k}\rangle \Psi_{n,k}(t) \right]. \tag{41}$$

Again we have

$$\|f\|_2^2 = \sum_{k\in\mathbb{Z}} \left[\sum_{n\in\mathbb{Z}} |\langle f, \Psi_{n,k}\rangle|^2 \right]. \tag{42}$$

In both of these cases, given that $\{\Psi_{k,j}\} = \{\mathbb{W}_k \widetilde{\varphi}_j(t)\}$ is an orthonormal basis for $L^2(\mathbb{R})$, we have a representation of a given function f in L^2. The set

$\{\Psi_{k,j}\} = \{\mathbb{W}_k \widetilde{\varphi}_j(t)\}$ is an exact normalized tight frame for L^2. The restriction that these basis elements present is computability. They become increasing difficult to compute as the smoothness in time/decay in frequency increases.

A way around this is to connect the *Bounded Adaptive Partition of Unity* systems $\{\mathscr{B}_k(t)\}$ to frame theory. The ideas behind this connection go back to the curvelet work of Candès and Donoho. The paper of Borup and Neilsen [2] gives a nice overview of this connection, and we will refer to that paper for the background from which we develop our approach. The set $\{\mathscr{B}_k(t)\}$ forms an *admissible* cover in that they form a partition of unity and have overlap with only their immediate neighbors.

For each window $\mathscr{B}_k(t)$, let $\phi_{n,k}(t)$ be the shifted $\exp[\pi i t T/n]$ centered in the window. Then, define

$$\Phi_{n,k} = \mathscr{B}_k(t)\phi_{n,k}(t)\,.$$

Given an $f \in L^2$ we can write

$$f(t) \approx \sum_{k\in\mathbb{Z}} \left[\sum_{n\in\mathbb{Z}} \langle f, \Phi_{n,k}\rangle \Phi_{n,k}(t) \right]. \tag{43}$$

For this system we can compute

$$A\|f\|_2^2 \le \sum_{k\in\mathbb{Z}} \left[\sum_{n\in\mathbb{Z}} |\langle f, \Phi_{n,k}\rangle|^2 \right] \le B\|f\|_2^2\,.$$

The signal will be underrepresented on some blocks, overrepresented on others. This is a function of how much of the signal is concentrated in the overlap regions. The frame bounds will be tightened for the almost orthogonal windowing systems. The fact that the almost ON windows $\{\mathbb{A}_k(t)\}$ approximate the ON windowing system will result in approximating the expansion of the signal contained in the overlapping region in an ON basis. The closer the approximation, the better the frame bounds. Developing these signal adaptive frames, their bounds and the associated frame operators will be a major point of emphasis in future work.

6 Remarks on Applications

Despite extensive advances in integrated circuit design and fabrication processes, wideband problems continue to hit barriers in sample and hold architectures and analog-to-digital conversion (ADC). ADC signal-to-noise and distortion ratio, the effective number of resolution bits, declines with sampling rate due to timing jitter, circuit imperfections, and electronic noise. ADC performance (speed and total integrated noise) can be improved to some extent, e.g., by cooling and therefore lowering the system temperature. However, the energy cost may be significant, and this presents a major hurdle for implementation in miniaturized devices. Digital circuitry has provided dramatically enhanced digital signal processing operation speeds, but there has not been a corresponding dramatic energy capacity increase in batteries to operate these circuits; there is no Moore's Law for batteries or ADCs.

A growing number of applications face this challenge, such as miniature and hand-held devices for communications, robotics, and micro- aerial vehicles. Very wideband sensor bandwidths are desired for dynamic spectrum access and cognitive radio, radar, and ultra-wideband systems. Multi-channel and multi-sensor systems, array processing and beamforming, multi-spectral imaging, and vision systems compound the issue. All of these rely on analog sensing and a digital interface, perhaps with feedback. This motivates mixed-signal circuit designs that tightly couple the analog and digital portions and operate with parallel reduced bandwidth paths to relax ADC requirements. The goal of such wideband integrated circuit designs is to achieve good trade-offs in dynamic range, bandwidth, and parallelization, while maintaining low energy consumption. This requires a careful balance of analog and digital functionality. We address this in [8].

From a signal processing perspective, we have approached this problem by implementing an appropriate signal decomposition in the analog portion that provides parallel outputs for integrated digital conversion and processing. This naturally leads to an architecture with windowed time segmentation and parallel analog basis expansion. In this chapter we viewed this from the sampling theory perspective, including segmentation and window design, achieving orthogonality between segments, basis expansion and choice of basis, signal filtering, and reconstruction. The approach we have developed in this chapter is tailored toward strong connections to circuit design considerations and applications.

Acknowledgments: The author's research was partially supported by US Army Research Office Scientific Services program, administered by Battelle (TCN 06150, Contract DAAD19-02-D-0001) and US Air Force Office of Scientific Research Grant Number FA9550-12-1-0430.

References

1. J.J. Benedetto, *Harmonic Analysis and Applications* (CRC Press, Boca Raton, 1997)
2. L. Borup, M. Neilsen, Frame decomposition of decomposition spaces. J. Four. Anal. Appl. **13**(1), 39–70 (2007)
3. W.L. Briggs, V.E. Henson, *The DFT: An Owner's Manual for the Discrete Fourier Transform* (SIAM, Philadelphia, 1995)
4. S.D. Casey, Two problems from industry and their solutions via harmonic and complex analysis. J. Appl. Funct. Anal. **2**(4), 427–460 (2007)
5. S.D. Casey, Windowing systems for time-frequency analysis. Sampl. Theory Signal Image Process. **11**(2–3), 221–251 (2012)
6. S.D. Casey, D.F. Walnut, Systems of convolution equations, deconvolution, Shannon sampling, and the wavelet and Gabor transforms. SIAM Rev. **36**(4), 537–577 (1994)
7. S.D. Casey, D.F. Walnut, Residue and sampling techniques in deconvolution, in *Modern Sampling Theory: Mathematics and Applications*, ed. by P. Ferreira, J. Benedetto. Birkhauser Research Monographs (Birkhauser, Boston, 2001), pp. 193–217
8. S.D. Casey, S. Hoyos, B.M. Sadler, Wideband sampling via windowed signal segmentation and projection. *Proc. IEEE* (2015), 28 pp

9. R. Coifman, Y. Meyer, Remarques sur l'analyse de Fourier a fenetre. C. R. Acad. Sci. Paris **312**, 259–261 (1991)
10. I. Daubechies, Ten lectures on wavelets, in *CBMS–NSF Conference Series in Applied Mathematics*, vol. 61 (SIAM, Philadelphia, 1992)
11. R.J. Duffin, A.C. Schaeffer, A class of non-harmonic Fourier series. Trans. Am. Math. Soc. **72**, 341–366 (1952)
12. H. Dym, H.P. McKean, *Fourier Series and Integrals* (Academic Press, Orlando, 1972)
13. L. Grafakos, *Classical and Modern Fourier Analysis* (Pearson Education, Upper Saddle River, 2004)
14. K. Gröchenig, *Foundations of Time-Frequency Analysis* (Birkhäuser, Boston, 2000)
15. J.R. Higgins, *Sampling Theory in Fourier and Signal Analysis: Foundations* (Clarendon Press, Oxford, 1996)
16. L. Hörmander, *The Analysis of Linear Partial Differential Operators I. Distribution Theory and Fourier Analysis*, 2nd edn. (Springer, New York, 1990)
17. S. Hoyos, B.M. Sadler, Ultra wideband analog-to-digital conversion via signal expansion. IEEE Trans. Veh. Technol. Invited Special Section on UWB Wireless Communications **54**(5), 1609–1622 (2005)
18. S. Hoyos, B.M. Sadler, Frequency domain implementation of the transmitted-reference ultra-wideband receiver. IEEE Trans. Microwave Theory Tech. Special Issue on Ultra-Wideband **54**(4), 1745–1753 (2006)
19. S. Hoyos, B.M. Sadler, UWB mixed-signal transform-domain direct-sequence receiver. IEEE Trans. Wirel. Commun. **6**(8), 3038–3046 (2007)
20. S. Hoyos, B.M. Sadler, G. Arce, Broadband multicarrier communication receiver based on analog to digital conversion in the frequency domain. IEEE Trans. Wirel. Commun. **5**(3), 652–661 (2006)
21. B. Jawerth, W. Sweldens, Biorthogonal smooth local trigonometric bases. J. Four. Anal. Appl. **2**(2), 109–133 (1995)
22. A.J. Jerri, The Shannon sampling theorem - its various extensions and applications: a tutorial review. Proc. IEEE **65**(11), 1565–1596 (1977)
23. V.A. Kotel'nikov, On the transmission capacity of 'ether' and wire in electrocommunications, in *Izd, Red. Upr. Svyazi RKKA*, Moscow (1933)
24. T.W. Körner, *Fourier Analysis* (Cambridge University Press, Cambridge, 1988)
25. B.Ya. Levin, *Lectures on Entire Functions* (American Mathematical Society, Providence, 1996)
26. H.S. Malvar, *Signal Processing with Lapped Transforms* (Artech House, Norwood, 1992)
27. H.S. Malvar, Biorthogonal and nonuniform lapped transforms for transform coding with reduced blocking and ringing artifacts. IEEE Trans. Signal Process. **46**(4), 1043–1053 (1998)
28. Y. Meyer, *Wavelets: Algorithms and Applications*, translated by R.D. Ryan (SIAM, Philadelphia, 1993)
29. N. Nürnberger, *Approximation by Spline Functions* (Springer, New York, 1989)
30. H. Nyquist, Certain topics in telegraph transmission theory. AIEE Trans. **47**, 617–644 (1928)
31. A. Papoulis, *The Fourier Integral and Its Applications* (McGraw-Hill, New York, 1962)
32. A. Papoulis, *Signal Analysis* (McGraw-Hill, New York, 1977)
33. P.M. Prenter, *Splines and Variational Methods* (Wiley, New York, 1989)
34. I.J. Schoenberg, *Cardinal Spline Interpolation*. CBMS–NSF Conference Series in Applied Mathematics, vol. 12 (SIAM, Philadelphia, 1973)
35. C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
36. C.E. Shannon, Communications in the presence of noise. Proc. IRE. **37**, 10–21 (1949)
37. N. Wiener, *The Fourier Integral and Certain of its Applications* (MIT Press, Cambridge, 1933)
38. N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series* (MIT Press, Cambridge, 1949)

39. E.T. Whittaker, On the functions which are represented by the expansions of the interpolation theory. Proc. R. Soc. Edinb. **35**, 181–194 (1915)
40. J.M. Whittaker, *Interpolatory Function Theory* (Cambridge University Press, Cambridge, 1935)
41. R. Young, *An Introduction to Nonharmonic Fourier Series* (Academic Press, New York, 1980)

Cornerstones of Sampling of Operator Theory

David Walnut, Götz E. Pfander, and Thomas Kailath

Abstract This paper reviews some results on the identifiability of classes of operators whose Kohn-Nirenberg symbols are band-limited (called *band-limited operators*), which we refer to as *sampling of operators*. We trace the motivation and history of the subject back to the original work of the third-named author in the late 1950s and early 1960s, and to the innovations in spread-spectrum communications that preceded that work. We give a brief overview of the NOMAC (Noise Modulation and Correlation) and Rake receivers, which were early implementations of spread-spectrum multi-path wireless communication systems. We examine in detail the original proof of the third-named author characterizing identifiability of channels in terms of the maximum time and Doppler spread of the channel, and do the same for the subsequent generalization of that work by Bello. The mathematical limitations inherent in the proofs of Bello and the third author are removed by using mathematical tools unavailable at the time. We survey more recent advances in sampling of operators and discuss the implications of the use of periodically weighted delta-trains as identifiers for operator classes that satisfy Bello's criterion

The inversion of the traditional alphabetical ordering of authors is at the suggestion of the third author, who desires that those at the end of the alphabet get some recognition.

D. Walnut (✉)
George Mason University, Fairfax, VA, USA
e-mail: dwalnut@gmu.edu

G.E. Pfander
Jacobs University, Bremen, Germany
e-mail: g.pfander@jacobs-university.de

T. Kailath
Stanford University, Stanford, CA, USA
e-mail: kailath@stanford.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_12

for identifiability, leading to new insights into the theory of finite-dimensional Gabor systems. We present novel results on operator sampling in higher dimensions, and review implications and generalizations of the results to stochastic operators, MIMO systems, and operators with unknown spreading domains.

Key words: Sampling, Gabor frame, Delta trains, Kohn-Nirenberg symbol, Spreading function, Operator Paley-Wiener space, Operator identification, Operator sampling, Channel identification, Channel measurement, Rake receiver, Time-variant filters, Band-limited operators, Gabor matrices, Stochastic operator, Compressive sensing, Matrix probing, MIMO channel

Introduction

The problem of identification of a time-variant communication channel arose in the 1950s as the problem of secure long-range wireless communications became increasingly important due to the geopolitical situation at the time. Some of the theoretical and practical advances made then are described in this paper, and more recent advances extending the theory to more general operators, and onto a more rigorous mathematical footing, known as *sampling of operators* are developed here as well.

The launching point for the theory of operator sampling is the early work of the third-named author in his Master's thesis at MIT, entitled "Sampling models for linear time-variant filters" [19], see also [22, 23], and [21] in which he reviews the identification problem for time-variant channels. The third named author as well as Bello in subsequent work [6] were attempting to understand and describe the theoretical limits of identifiability of time-variant communication channels. Section "Historical Remarks" of this paper describes in some detail their work and explores some of the important mathematical challenges they faced. In Section "Operator Sampling", we describe the more recently developed framework of operator sampling. Results addressing the problem considered by Bello are based on insights on finite dimensional Gabor systems which are presented in Section "Linear Independence Properties of Gabor Frames". Malikiosis's recent result [32] allows for the generalization of those results to a higher-dimensional setting, these are stated and proven in Section "Generalizations of operator sampling to higher dimensions". We conclude the paper in Section "Further results on operator sampling" with a short summary of the sampling of operators literature, that is, of results presented in detail elsewhere.

Historical Remarks

The Cold War Origins of the Rake System

In 1958, Price and Green published *A Communication Technique for Multi-path Channels* in Proc. IRE [56], in which they describe a communication system called

Rake, designed to solve the *multi-path problem*. When a wireless transmitter does not have line-of-sight with the receiver, the transmitted signal is reflected possibly multiple times before reaching the receiver. Reflection by stationary objects such as the ground or buildings introduces random time delays to the signal, and reflection or refraction by moving objects such as clouds, the troposphere, ionosphere, or a moving vehicle produce random frequency or Doppler shifts in the signal as well. Due to scattering and absorption, the reflected signals are randomly amplitude-attenuated too. The problem is to recover the transmitted signal as accurately as possible from the superposition of time-frequency-shifted and randomly amplitude-attenuated versions of it. Since the location and velocities of the reflecting objects change with time, the effects of the unknown, time-variant channel must be estimated and compensated for.

Price and Green's paper [56] was the disclosure in the literature of a long-distance system of wide-band or spread-spectrum communications that had been developed in response to strategic needs related to the Cold War. This fascinating story has been described in several articles by those directly involved [12, 55, 57, 58]. We present a summary of those remarks and of the Rake system below. The goal is to motivate the original work of the third-named author on which the theory of operator sampling is based.

In the years following World War II, the Soviet Union was exercising its power in Eastern Europe with a major point of contention being Berlin, which the Soviets blockaded in the late 1940s. This made secure communication with Berlin a top priority. As Paul Green describes it,

> [T]he Battle of Berlin was raging, the Russians having isolated the city physically on land, so that the Berlin Airlift was resorted to, and nobody knew when all the communication links would begin to be subjected to heavy Soviet jamming. [12]

By 1950, with a shooting war in Korea about to begin, the Army Signal Corps approached researchers at MIT to develop secure, and reliable wireless communication with the opposite ends of the world. According to Green,

> It is difficult today to recall the fearful excitement of those times. The Russians were thought to be 12 feet high in anything having to do with applying mathematics to communication problems ("all Russians were Komogorovs or Kotelnikovs")....[T]here was a huge backlog of unexploited theory lying around, and people were beginning to build digital equipment with the unheard of complexity of a hundred or so vacuum tube-based bits (!). And the money flowed. [12]

The effort was called Project Lincoln (precursor to Lincoln Laboratory). The researchers were confronted by two main problems: 1) making a communications system robust to noise and deliberate jamming, and 2) enabling good signal recovery from multiple paths.

Spread Spectrum communications and NOMAC

The technique chosen to address the first problem is an application of the notion, already well understood and used by that time, that combatting distortions

from noise and jamming can be achieved by spreading the signal over a wide frequency band. The idea of spreading the spectrum had been around for a long time [51, 55, 63] and can be found even in a now famous Hedy Lamarr-George Antheil patent of 1942 [33, 55], which introduced the concept later called "frequency hopping". The system called NOMAC (Noise Modulation and Correlation) was developed in the early 1950s and used noise-like (pseudo-noise or PN) signals to achieve spectrum spreading. Detailed discussion of its history can be found in [12, 55, 64].

The huge backlog of "unexploited theory" mentioned above included the recent work of Claude Shannon on communication theory [61], of Norbert Wiener on correlation functions and least mean squares prediction and filtering [65], and recent applications of statistical decision theory to detection problems in radar and communications.

The communication problem addressed by NOMAC was to encode data represented by a string of ones and zeros into analog signals that could be electromagnetically transmitted over a noisy communication channel in a way that foiled "jamming" by enemies. The analog signals $x_1(\cdot)$ and $x_0(\cdot)$, commonly called Mark and Space, associated with the data digits 1 and 0, were chosen to be waveforms of approximate bandwidth B, and with small cross correlation. The target application was 60 wpm teletype, with 22 msec per digit (called a baud), which corresponds to a transmission rate of $1/0.022\,\text{sec} = 45\,\text{Hz}$. The transmitted signals were chosen to have a bandwidth of 10 KHz, which was therefore expected to yield a "jamming suppression ratio" of $10,000/45 = 220$, or 23 db [12, 64]. The jamming ratio is often called the "correlation gain", because the receiver structure involves cross correlation of the received signal with each of the possible transmitted signals. If the correlation with the signal $x_1(\cdot)$ is larger than the one with the signal $x_0(\cdot)$, then it is decided that the transmitted signal corresponded to the digit 1. This scheme can be shown to be optimum in the sense of minimum probability of error provided that the transmitted signals are not distorted by the communications channel and that the receiver noise is white Gaussian noise (see, for example, [16]). The protection against jamming is because unless the jammer has good knowledge of the noise-like transmitted signals, any jamming signals would just appear as additional noise at the output of the crosscorrelations.

More details on the nontrivial ideas required for building a practical system can be found in the references. We may mention that the key ideas arose from three classified MIT dissertations by Basore [4], Pankowski [34], and Green [10], in fact, documents on NOMAC remained classified until 1961 [12].

A transcontinental experiment was run on a NOMAC system, but was found to have very poor performance because of the presence of multiple paths; the signals arriving at the receiver by these different paths sometimes interfere destructively. This is the phenomenon of "fading", which causes self-jamming of the system. Some improvement was achieved by adding additional circuitry and the receiver to separately identify and track the two strongest signals and combine them after phase correction; this use of time and space diversity enabled a correlation gain of 17 db, 6 db short of the expected performance. It was determined that this loss was

because of the neglected weaker paths, of which there could be as many as 20 or 30. So attention turned to a system that would allow the use of all the different paths.

The RAKE system

One conceptual basis for this new system was provided by the doctoral thesis of Robert Price [52], the main results of which were published in 1956 [53]. In a channel with severe multi-path the signal at the receiver is composed of large number of signals of different amplitudes and phases and so Price made the assumption that the received "signal" was a Gaussian random process. He studied the problem of choosing between the hypothesis

$$H_i : w(\cdot) = Ax_i(\cdot) + n(\cdot), \quad i = 0, 1,$$

where the random time variant linear communication channel A is such that the $\{Ax_i(\cdot)\}$ are Gaussian processes. In this case, the earlier cross correlation detection scheme makes no sense, because the "signal" arriving at the receiver is not deterministic but is a sample function of a random process, which is not available to the receiver because it is corrupted by the additive noise. Price worked out the optimum detection scheme and then ingeniously interpreted the mathematical formulas to conclude that the new receiver forms least mean-square estimates of the $\{Ax_i(\cdot)\}$ and then crosscorrelates the $w(\cdot)$ against these estimates. In practice of course, one does not have enough statistical information to form these estimates and therefore more heuristic estimates are used and this was done in the actual system that was built. The main heuristic, from Wiener's least mean-square smoothing filter solution and earlier insights, is that one should give greater weight to paths with higher signal-to-noise ratio.

So Price and Green devised a new receiver structure comprised of a delay line of length 3 ms intervals (the maximum expected time spread in their channel), with 30 taps spaced every $1/10$ Khz, or $100\,\mu$s. This would enable the capture of all the multi-path signals in the channel. Then the tap gains were made proportional to the strength of the signal received at that tap. Since a Mark/Space decision was only needed every 22 ms (for the transmission rate of 60 wpm), and since the fading rate of the channel was slow enough that the channel characteristics remain constant over even longer than 22 ms, tap gains could be averaged over several 3 ms intervals. The new system was called "Rake", because the delay line structure resembled that in a typical garden rake!

Trials showed that this scheme worked well enough to recover the 6 db loss experienced by the NOMAC system. The system was put into production and was successfully used for jam-proof communications between Washington DC and Berlin during the "Berlin crisis" in the early 60s.

HF communications is no longer very significant, but the Rake receiver has found application in a variety of problems such as sonar, the detection of underground

nuclear explosions, and planetary radar astronomy (pioneered by Price and Green, [11, 54]) and currently it is much used in mobile wireless communications. It is interesting to note that the eight racks of equipment needed to build the Rake system in the 1960s is now captured in a small integrated circuit chip in a smart phone!

However the fact that the Rake system did not perform satisfactorily when the fading rates of the communication channel were not very slow led MIT professor John Wozencraft, (who had been part of the Rake project team at Lincoln Lab) to suggest in 1957 (even before the open 1958 publication of the Rake system) to his new graduate student Thomas Kailath a fundamental study of linear time-variant communication channels and their identifiability for his Masters thesis. While time-variant linear systems had begun to be studied at least as early as 1950 (notably by Zadeh [66]), in communication systems there are certain additional constraints, notably limits on the bandwidths of the input signal and the duration of the channel memory. So a more detailed study was deemed to be worthwhile.

Kailath's Time-Variant Channel Identification Condition

In the paper [19], the author considers the problem of measuring a channel whose characteristics vary rapidly with time. He considers the dependence of any theoretical channel estimation scheme on how rapidly the channel characteristics change and concludes that there are theoretical limits on the ability to identify a rapidly changing channel. He models the channel A as a linear time-variant filter and defines

> $A(\lambda,t) =$ response of A, measured at time t to a unit impulse input at time $t-\lambda$.

$A(\lambda,t)$ is one form of the time-variant impulse response of the linear channel that emphasizes the role of the "age" variable λ. The channel response to an input signal $x(\cdot)$ is

$$Ax(t) = \int A(\lambda,t)x(t-\lambda)\,d\lambda.$$

An impulse response $A(\lambda,t) = A(\lambda)$ represents a time-invariant filter. Further, the author states

> Therefore the rate of variation of $A(\lambda,t)$ with t, for fixed λ, is a measure of the rate of variation of the filter. It is convenient to measure this variation in the frequency domain by defining a function $\mathcal{A}$
>
> $$\mathcal{A}(\lambda,f) = \int_{-\infty}^{\infty} A(\lambda,t)e^{-2\pi ift}dt$$

Then he defines

$$B = \max_{\lambda}[b-a, \text{ where } \mathcal{A}(\lambda,f) = 0 \text{ for } f \notin [a,b]\,].$$

While symmetric support is assumed in the paper, this definition makes clear that non-rectangular regions of support are already in view. Additionally, he defines the memory as the maximum time-delay spread in response to an impulse of the channel as

$$L = \max_t [\min_{\lambda'} \text{ such that } A(\lambda, t) = 0,\ \lambda \geq \lambda'].$$

In short, the assumption in the continuation of the paper is that

$$\operatorname{supp} \mathcal{A}(\lambda, f) \subseteq [0, L] \times [-W, W]$$

where $W = B/2$. The function $\mathcal{A}(\lambda, f)$ is often called the *spreading function* of the channel. He then asks under what assumptions on L and $B = 2W$ can such a channel be measured? In the context of the Rake system, this translates to the question of whether there are limits on the rate of variation of the filter that can assure that the measurement filter can be presumed to be effective.

The author's assertion is that as long as $BL \leq 1$, then a "simple measurement scheme" is sufficient.

> We have assumed that the bandwidth of any "tap function", $A_\lambda(\cdot)\,[= A(\lambda, \cdot)]$, is limited to a frequency region of width B, say a low-pass region $(-W, W)$ for which $B = 2W$. Such band-limited taps are determined according to the Sampling theorem, by their values at the instants $i/2W$, $i = 0, \pm 1, \pm 2, \ldots$.
>
> If the memory, L, of the filter, $A(\lambda, t)$ is less than $1/2W$ these values are easily determined: we put in unit impulses to $A(\lambda, t)$ at instants $0,\ 1/2W,\ 2/2W, \ldots, T$, and read off from the responses the desired values of the impulse response $A(\lambda, t)$. [...] If $L \leq 1/2W$, the responses to the different input impulses do not interfere with one another and the above values can be unambiguously determined.

In other words, sufficiently dense samples of the tap functions can be obtained by sending an impulse train $\sum_n \delta_{n/2W}$ through the channel. Indeed,

$$A\big(\sum_n \delta_{n/2W}\big)(t) = \sum_n \int A(\lambda, t)\, \delta_{n/2W}(t - \lambda)\, d\lambda = \sum_n A(t - n/2W, t).$$

Evaluating the operator response at $t = \lambda_0 + n_0/2W$, $n_0 \in \mathbb{Z}$, we obtain

$$\begin{aligned} A\big(\sum_n \delta_{n/2W}\big)(\lambda_0 + n_0/2W) &= \sum_n A(\lambda_0 + (n_0 - n)/2W, \lambda_0 + n_0/2W) \\ &= A(\lambda_0, \lambda_0 + n_0/2W) \end{aligned}$$

since $L \leq 1/2W$ implies that $A(\lambda_0 + (n_0 - n)/2W, \lambda_0 + n_0/2W) = 0$ if $n \neq n_0$. In short, for each λ, the samples $A(\lambda, \lambda + n/2W)$ for $n \in \mathbb{Z}$ can be recovered.

The described Kailath sounding procedure is depicted in Figure 1. In this visualization, we plot the kernel $\kappa(s, t) = A(t - s, t)$ of the operator A, that is,

$$Ax(t) = \int A(\lambda, t)\, x(t - \lambda)\, d\lambda = \int A(t - s, t)\, x(s)\, ds = \int \kappa(t, s)\, x(s)\, ds.$$

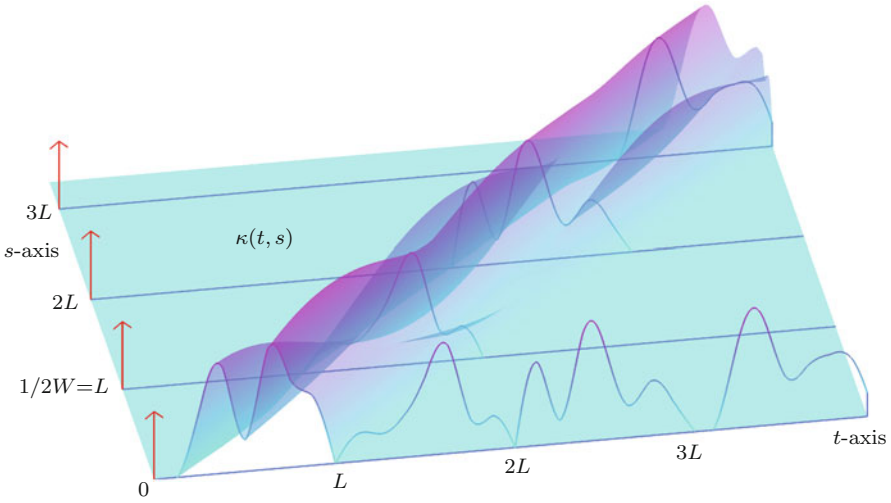


Fig. 1 Kailath sounding of A with $\operatorname{supp} \mathcal{A}(\lambda, f) \subseteq [0, L] \times [-W, W]$ and $L = 1/2W$. The kernel $\kappa(t,s)$ is displayed on the (t,s) plane, the impulse train $\sum_n \delta_{n/2W}(s)$ on the s-axis, and the output signal $Ax(t) = A\big(\sum_n \delta_{n/2W}\big)(t) = \sum_n A(t - n/2W, t) = \sum_n \kappa(t, n/2W)$. The sample values of the tab functions $A_\lambda(t) = A(\lambda, t) = \kappa(t, t-\lambda)$ can be read off $Ax(t)$.

Necessity of Kailath's Condition for Channel Identification

For the "simple measurement scheme" to work, $BL \leq 1$ is sufficient but could be restrictive.

> We need, therefore, to devise more sophisticated measurement schemes. However, we have not pursued this question very far because for a certain class of channels we can show that the condition
>
> $$L \leq 1/2W, \text{ i.e. }, BL \leq 1$$
>
> is necessary as well as sufficient for unambiguous measurement of $A(\lambda, t)$. The class of channels is obtained as follows: We first assume that there is a bandwidth constraint on the possible input signals to $A(\lambda, t)$, in that the signals are restricted to $(-W_i, W_i)$ in frequency. We can now determine a filter $A_{W_i}(\lambda, t)$ that is equivalent to $A(\lambda, t)$ over the bandwidth $(-W_i, W_i)$, and find necessary and sufficient conditions for unambiguous measurement of $A_{W_i}(\lambda, t)$. If we now let $W_i \to \infty$, this condition reduces to condition (1), viz: $L \leq 1/2W$. Therefore, condition (1) is valid for all filters $A(\lambda, t)$ that may be obtained as the limit of band-limited channels. This class includes almost all filters of physical interest. The argument is worked out in detail in Ref. 6 [1] but we give a brief outline here.

The class of operators in view here can be described as limits (in some unspecified sense) of operators whose impulse response $A(\lambda, t)$ is bandlimited to $[-W_i, W_i]$ in λ for each t and periodic with period $T > 0$ in t for each λ. Here, T is assumed to have some value larger than the maximum time over which the channel will be operated. We could take it as the duration of the input signal to the channel.

[1] Ref. 6 is [19].

The restriction to input signals bandlimited to $(-W_i, W_i)$ indicates that it suffices to know the values of $A(\lambda,t)$ or $\mathcal{A}(\lambda,f)$ for a finite set of values of λ: $\lambda = 0, 1/2W_i, 2/2W_i, \ldots, L$, assuming for simplicity that L is a multiple of $1/2W_i$. Therefore, we can write

$$A(\lambda,t) = \sum_n A(n/2W_i,t)\,\mathrm{sinc}_{W_i}(\lambda - n/2W_i),$$

where $\mathrm{sinc}_{W_i}(t) = \sin(2\pi W_i t)/(2\pi W_i t)$ so that as $W_i \to \infty$, $\mathrm{sinc}_{W_i}(t)$ becomes more concentrated at the origin.

Also, T-periodicity in t allows us to write

$$A(\lambda,t) = \sum_k A(\lambda,k/T)\,e^{2\pi ikt/T},$$

so that combining gives

$$A(\lambda,t) = \sum_n \sum_k A(n/2W_i,k/T)\,\mathrm{sinc}_{W_i}(\lambda - n/2W_i)\,e^{2\pi ikt/T}.$$

Based on the restriction to bandlimited input signals which are T periodic, we have obtained a representation of A which is neither compactly supported in λ nor bandlimited in t. However, the original restriction that

$$\mathrm{supp}\,\mathcal{A}(\lambda,f) \subseteq [0,L] \times [-W,W]$$

motivates the assumption that we are working with finite sums, viz.

$$A(\lambda,t) = \sum_{n/2W_i \in [0,L]} \sum_{k/t \in [-W,W]} A(n/2W_i,k/T)\,\mathrm{sinc}_{W_i}(\lambda - n/2W_i)\,e^{2\pi ikt/T}.$$

This is how the author obtains the estimate that there are at most $(2W_iL+1)(2WT+1)$ degrees of freedom in any impulse response A in the given class.

For any input signal $x(t)$ bandlimited to $[-W_i, W_i]$, the output will be bandlimited to $[-W - W_i, W + W_i]$. Specifically,

$$\begin{aligned} Ax(t) &= \int A(\lambda,t)\,x(t-\lambda)\,d\lambda \\ &= \sum_{n/2W_i \in [0,L]} \sum_{k/T \in [-W,W]} A(n/2W_i,k/T)\,e^{2\pi ikt/T} \\ &\qquad \int x(t-\lambda)\,\mathrm{sinc}_{W_i}(\lambda - n/2W_i)\,d\lambda \\ &= \sum_{n/2W_i \in [0,L]} \sum_{k/T \in [-W,W]} A(n/2W_i,k/T)\,e^{2\pi ikt/T} \\ &\qquad (x * \mathrm{sinc}_{W_i})(t - n/2W_i). \end{aligned}$$

Since $e^{2\pi ikt/T}\,(x * \mathrm{sinc}_{W_i})(t - n/2W_i)$ is bandlimited to $[-W_i, W_i] + (k/T)$ for $k/T \in [-W, W]$, it follows that $Ax(t)$ is bandlimited to $[-W - W_i, W + W_i]$.

If we restrict our attention to signals $x(t)$ time-limited to $[0,T]$, the output signal $Ax(t)$ will have duration $T+L$, and $Ax(\cdot)$ will be completely determined by its samples at $\frac{n}{2(W+W_i)} \in [0,T+L]$, from which we can identify $2(T+L)(W+W_i)+1$ degrees of freedom.

In order for identification to be possible, the number of degrees of freedom of the output signal must be at least as large as the number of degrees of freedom of the operator, i.e.

$$\begin{aligned} 2W_iT + 2W_iL + 2WT + 2WL + 1 = \\ 2(T+L)(W_i+W)+1 &\geq (2WT+1)(2W_iL+1) \\ &= 2WT + 2W_iL + 1 + 4W_iWTL \end{aligned}$$

which reduces ultimately to

$$\frac{1}{1-1/(2W_iT)} \geq 2WL = BL.$$

That is, BL needs to be strictly smaller than 1 in the approximation while $BL=1$ may work in the limiting case $W_i \to \infty$ (and/or $T \to \infty$).

This result got a lot of attention because it corresponded with experimental evidence that Rake did not function well when the condition $BL<1$ was violated. It led to the designation of "underspread" and "overspread" channels for which BL was less than or greater than 1.

Some Remarks on Kailath's Results

This simple argument is surprising, particularly in light of the fact that the author obtained a deep result in time-frequency analysis with none of the tools of modern time-frequency analysis at his disposal. He very deftly uses the extremely useful engineering "fiction" that the dimension of the space of signals essentially bandlimited to $[-W,W]$ and time-limited to $[0,T]$ is approximately $2WT$. The then recent papers of Landau, Slepian, and Pollak [28, 29], which are mentioned explicitly in [19], provided a rigorous mathematical framework for understanding the phenomenon of essentially simultaneous band- and time-limiting. While the existence of these results lent considerable mathematical heft to the argument, they were not incorporated into a fully airtight mathematical proof of his theorem.

> In the proof we have used a degrees-of-freedom argument based on the sampling theorem which assumes strictly bandlimited functions. This is an unrealistic assumption for physical processes. It is more reasonable to call a process band (or time) limited if some large fraction of its energy, say 95%, is contained within a finite frequency (or time) region. Recent work by Landau and Slepian has shown the concept of approximately $2TW$ degrees of freedom holds even in such cases. This leads us to believe that our proof of the necessity of the $BL \leq 1$ condition is not merely a consequence of the special properties of strictly bandlimited functions. It would be valuable to find an alternative method of proof.

While Kailath's Theorem is stated for channel operators whose spreading functions are supported in a rectangle, it is clear that the later work of Bello [6] was anticipated and more general regions were in view. This is stated explicitly.

> We have not discussed how the bandwidth, B is to be defined. There are several possibilities: we might take the nonzero f-region of $\mathcal{A}(\lambda, f)$; or use a "counting" argument. We could proceed similarly for the definition of L. As a result of these several possibilities, the value 1, of the threshold in the condition $BL \leq 1$ should be considered only as an order of magnitude value.
>
> ...constant and predictable variations in B and L, due for example to known Doppler shifts or time displacements, would yield large values for the absolute values of the time and frequency spreadings. However such predictable variations should be subtracted out before the BL product is computed; *what appears to be important is the area covered in the time- and frequency-spreading plane rather than the absolute values of B and L.* (emphasis added)

The reference to "counting" as a definition of bandwidth clearly indicates that essentially arbitrary regions of support for the operator spreading function were in view here, and that a necessity argument relying on degrees of freedom and not the shape of the spreading region was anticipated. The third-named author did not pursue the measurement problem studied in his MS thesis because he went on in his PhD dissertation to study the optimum (in the sense of minimum probability of error) detector scheme of which Rake is an intelligent engineering approximation. See [20, 21, 23].

The mathematical limitations of the necessity proof in [19] can be removed by addressing the identification problem directly as a problem on infinite-dimensional space rather than relying on finite-dimensional approximations to the channel. This approach also avoids the problem of dealing with simultaneously time and frequency-limited functions. In this way, the proof can be made completely mathematically rigorous. This approach is described in Section "Kailath's necessity proof and operator identification".

Bello's time-variant Channel Identification Condition

Kailath's Theorem was generalized by Bello in [6] along the lines anticipated in [19]. Bello's argument follows that of [19] in its broad outlines but with some significant differences. Bello clearly anticipates some of the technical difficulties that have been solved more recently by the authors and others and which have led to the general theory of operator sampling.

Continuing with the notation of this section, Bello considers channels with spreading function $\mathcal{A}(\lambda, f)$ supported in a rectangle $[0, L] \times [-W, W]$. If L and W are all that is known about the channel, then Kailath's criterion for measurability requires that $2WL \leq 1$. Bello considers channels for which $2WL$ may be greater than 1 but for which

$$S_A = |\operatorname{supp} \mathcal{A}(\lambda, f)| \leq 1$$

and argues that this is the most appropriate criterion to assess measurability of the channel modeled by A.

In order to describe Bello's proof we will fix parameters $T \gg L$ and $W_i \gg W$ and following the assumptions earlier in this section, assume that inputs to the

channel are time-limited to $[0,T]$ and (approximately) bandlimited to $[-W_i,W_i]$. Under this assumption, Bello considers the spreading function of the channel to be approximated by a superposition of point scatterers, viz.

$$\mathcal{A}(\lambda,f)=\sum_n\sum_k A_{n,k}\,\delta(f-(k/T))\,\delta(\lambda-(n/2W_i)).$$

Hence the response of the channel to an input $x(\cdot)$ is given by

$$\begin{aligned} Ax(t) &= \iint x(t-\lambda)\,e^{2\pi i f(t-\lambda)}\mathcal{A}(\lambda,f)\,d\lambda\,df \\ &= \sum_n\sum_k A_{n,k}\,x(t-(n/2W_i))\,e^{2\pi i(k/T)(t-(n/2W_i))}. \end{aligned} \tag{1}$$

Note that this is a continuous-time Gabor expansion with window function $x(\cdot)$ (see, e.g., [13]). By standard density results in Gabor theory, the collection of functions $\{x(t-(n/2W_i))\,e^{2\pi i(k/T)(t-(n/2W_i))}\}$ is overcomplete as soon as $2TW_i>1$. Consequently, without further discretization, the coefficients $A_{n,k}$ are in principle unrecoverable. Taking into consideration support constraints on $\mathcal{A}$, we assume that the sums are finite, viz.

$$\left(\frac{n}{2W_i},\frac{k}{T}\right)\in\operatorname{supp}\mathcal{A}.$$

Hence determining the channel characteristics amounts to finding $A_{n,k}$ for those pairs (n,k). It should be noted that for a given spreading function $\mathcal{A}(\lambda,f)$ for which $\operatorname{supp}\mathcal{A}$ is a Lebesgue measurable set, given $\varepsilon>0$, there exist T and W_i sufficiently large that the number of such (n,k) is no more than $2S_AW_iT(1+\varepsilon)$. On the other hand, for a given T and W_i, there exist spreading functions $\mathcal{A}(\lambda,f)$ with arbitrarily small non-convex S_A for which the number of nonzero coefficients $A_{n,k}$ can be large. For example, given T and W_i, S_A could consist of rectangles centered on the points $(n/(2W_i),k/T)$ with arbitrarily small total area.

By sampling, (1) reduces to a discrete, bi-infinite linear system, viz.

$$Ax\left(\frac{p}{2W_i}\right)=\sum_n\sum_k A_{n,k}\,x\left(\frac{p-n}{2W_i}\right)e^{2\pi i\frac{k}{T}\left(\frac{p-n}{2W_i}\right)} \tag{2}$$

for $p\in\mathbb{Z}$. Note that (2) is the expansion of a vector in a discrete Gabor system on $\ell^2(\mathbb{Z})$, a fact not mentioned by Bello, and of which he was apparently unaware. Specifically, defining the translation operator $\mathcal{T}$ and the modulation operator $\mathcal{M}$ on ℓ^2 by

$$\mathcal{T}x(n)=x(n-1),\qquad\text{and}\qquad\mathcal{M}x(n)=e^{\pi in/(TW_i)}x(n), \tag{3}$$

(2) can be rewritten as

$$Ax\left(\frac{p}{2W_i}\right)=\sum_n\sum_k(\mathcal{T}^n\mathcal{M}^kx)(p)A_{n,k}. \tag{4}$$

Since there are only finitely many nonzero unknowns in this system, Bello's analysis proceeds by looking at finite sections of (4) and counting degrees of freedom.

Necessity. Following the lines of the necessity argument in [19], we note that there are at least $2(T+L)(W+W_i)$ degrees of freedom in the output vector $Ax(t)$, that is, at least that many independent samples of the form $Ax(p/2W_i)$, and as observed above, no more than $2S_AW_iT(1+\varepsilon)$ nonzero unknowns $A_{n,k}$. Therefore, in order for the $A_{n,k}$ to be determined in principle, it must be true that

$$2W_iT(1+\varepsilon)S_A \leq 2(T+L)(W+W_i)$$

or

$$S_A \leq \frac{(T+L)(W+W_i)}{W_iT(1+\varepsilon)}.$$

Letting $T, W_i \to \infty$ and $\varepsilon \to 0$, we arrive at $S_A \leq 1$.

Sufficiency. Considering a section of the system (4) based on the assumption that $\operatorname{supp}\mathcal{A} \subseteq [0,L] \times [-W,W]$, the system has approximately $2W_i(T+L)$ equations in $(2W_iT)(2WL)$ unknowns. Since L and $2W$ are simply the dimensions of a rectangle that encloses the support of $\mathcal{A}$, $2WL$ may be quite large and independent of S_A. Hence the system will not in general be solvable. However by assuming that $S_A < 1$, only approximately $S_A(2W_iT)$ of the $A_{n,k}$ do not vanish and the system reduces to one in which the number of equations is roughly equal to the number of unknowns. In this case it would be possible to solve (4) as long as the collection of appropriately truncated vectors $\{\mathcal{T}^n\mathcal{M}^kx\colon A_{n,k} \neq 0\}$ forms a linearly independent set for some vector x.

In his paper, Bello was dealing with independence properties of discrete Gabor systems apparently without realizing it, or at least without stating it explicitly. Indeed, he argues in several different ways that a vector x that produces a linearly independent set should exist, and intriguingly suggests that a vector consisting of ± 1 should exist with the property that the Grammian of the Gabor matrix corresponding to the section of (4) being considered is diagonally dominant.

The setup chosen below to prove Bello's assertion leads to the consideration of a matrix whose columns stem from a Gabor system on a finite-dimensional space, not on a sequence space.

Operator Sampling

The first key contribution of operator sampling is the use of frame theory and time-frequency analysis to remove assumptions of simultaneous band- and time-limiting, and also to deal with the infinite number of degrees of freedom in a functional analytic setting (Section "Operator classes and operator identification"). A second key insight is the development of a "simple measurement scheme" of the type used by the third-named author but that allows for the difficulties identified by Bello to be resolved. This insight is the use of periodically weighted delta-trains as measurement functions for a channel. Such measurement functions have three distinct advantages.

First, they allow for the channel model to be essentially arbitrary and clarify the reduction of the operator identification problem to a finite-dimensional setting without imposing a finite dimensional model that approximates the channel. Second, it combines the naturalness of the simple measurement scheme described earlier with the flexibility of Bello's idea for measuring channels with arbitrary spreading support. Third, it establishes a connection between identification of channels and finite-dimensional Gabor systems and allows us to determine windowing vectors with appropriate independence properties.

In Section "Operator classes and operator identification", we introduce some operator-theoretic descriptions of some of the operator classes that we are able to identify, and discuss briefly different ways of representing such operators. Such a discussion is beneficial in several ways. First, it contains a precise definition of identifiability, which comes into play when considering the generalization of the necessity condition for so-called overspread channels (Section "Kailath's necessity proof and operator identification"). Second, we can extend the necessity condition to a very large class of inputs. In other words, we can assert that in a very general sense, no input can identify an overspread channel. Third, it allows us to include both convolution operators and multiplication operators (for which the spreading functions are distributions) in the operator sampling theory. The identification of multiplication operators via operator sampling reduces to the classical sampling formula, thereby showing that classical sampling is a special case of operator sampling. In Section "Kailath's necessity proof and operator identification" we present a natural formalization of the original necessity proof of [19] (Section "Necessity of Kailath's Condition for Channel Identification") to the infinite-dimensional setting, which involves an interpretation of the notion of an under-determined system to that setting. Finally, in Section "Identification of operator Paley-Wiener spaces by periodically weighted delta-trains" we present the scheme given first in [41, 45] for the identification of operator classes using periodically weighted delta trains and techniques from modern time-frequency analysis.

Operator classes and operator identification

We formally consider an arbitrary operator as a *pseudodifferential operator* represented by

$$Hf(x) = \int \sigma_H(x,\xi)\widehat{f}(\xi)\, e^{2\pi i x\xi}\, d\xi, \tag{5}$$

where $\sigma_H(x,\xi) \in L^2(\mathbb{R}^2)$ is the *Kohn-Nirenberg* (KN) symbol of H. The *spreading function* $\eta_H(t,\nu)$ of the operator H is the *symplectic Fourier transform* of the KN symbol, viz.

$$\eta_H(t,\nu) = \iint \sigma_H(x,\xi)\, e^{-2\pi i(\nu x - \xi t)}\, dx\, d\xi \tag{6}$$

and we have the representation

$$Hf(x) = \iint \eta_H(t,\nu)\, \mathcal{T}_t \mathcal{M}_\nu f(x)\, d\nu\, dt \tag{7}$$

where $\mathcal{T}_t f(x) = f(x-t)$ is the *time-shift operator* and $\mathcal{M}_\nu f(x) = e^{2\pi i \nu x} f(x)$ is the *frequency-shift operator*.

This is identical to the representation given in [19] where $\eta_H(t,\nu) = \mathcal{A}(\nu,t)$, see Section "Kailath's Time-Variant Channel Identification Condition".

To see more clearly where the spreading function arises in the context of communication theory, we can define the *impulse response* of the channel modeled by H, denoted $h_H(x,t)$, by

$$Hf(x) = \int h_H(x,t)\, f(x-t)\, dt.$$

Note that if h_H were independent of x, then H would be a convolution operator and hence a model for a time-invariant channel. In fact, with $\kappa_H(x,t)$ being the *kernel* of the operator H,

$$\begin{aligned} Hf(x) &= \int \kappa_H(x,t)\, f(t)\, dt && (8)\\ &= \int h_H(x,t)\, f(x-t)\, dt && (9)\\ &= \iint \eta_H(t,\nu)\, e^{2\pi i \nu(x-t)}\, f(x-t)\, d\nu\, dt && (10)\\ &= \int \sigma_H(x,\xi)\, \widehat{f}(\xi)\, e^{2\pi i x \xi}\, d\xi, && (11) \end{aligned}$$

where

$$\begin{aligned} h_H(x,t) &= \kappa_H(x, x-t)\\ &= \int \sigma_H(x,\xi)\, e^{2\pi i \xi t}\, d\xi,\\ &= \int \eta_H(t,\nu)\, e^{2\pi i \nu(x-t)}\, d\nu. \end{aligned} \tag{12}$$

With this interpretation, the maximum support of $\eta_H(t,\nu)$ in the first variable corresponds to the maximum spread of a delta impulse sent through the channel and the maximum support of $\eta_H(t,\nu)$ in the second variable corresponds to the maximum spread of a pure frequency sent through the channel.

Since we are interested in operators whose spreading functions have small support, it is natural to define the following operator classes, called *operator Paley-Wiener spaces* (see [38]).

Definition 1. For $S \subseteq \mathbb{R}^2$, we define the operator Paley-Wiener spaces $OPW(S)$ by

$$OPW(S) = \{H \in \mathcal{L}(L^2(\mathbb{R}), L^2(\mathbb{R})) :\ \operatorname{supp} \eta_H \subseteq S,\ \|\sigma_H\|_{L^2} < \infty\}.$$

Remark 1. In [38, 42], the spaces $OPW^{p,q}(S)$, $1 \le p, q < \infty$, were considered, where L^2-membership of σ_H is replaced by

$$\|\sigma_H\|_{L^{p,q}} = \left(\int \left(\int |\sigma_H(x,\xi)|^q d\xi \right)^{p/q} dx \right)^{1/p}$$

with the usual adjustments made when either $p = \infty$ or $q = \infty$. $OPW^{p,q}(S)$ is a Banach space with respect to the norm $\|H\|_{OPW^{p,q}} = \|\sigma_H\|_{L^{p,q}}$. Note that if S is bounded, then $OPW^{\infty,\infty}(S)$ consists of all bounded operators whose spreading function is supported on S. In fact, the operator norm is then equivalent to the $OPW^{\infty,\infty}(S)$ norm, where the constants depend on S [26].

The general definition is beneficial since it also allows the inclusion of convolution operators with kernels whose Fourier transforms lie in $L^q(\mathbb{R})$ ($OPW^{\infty,q}(\mathbb{R})$) and multiplication operators whose multiplier is in $L^p(\mathbb{R})$ ($OPW^{p,\infty}(\mathbb{R})$).

The goal of operator identification is to find an input signal g such that each operator H in a given class is completely and stably determined by Hg. In other words, we ask that the operator $H \mapsto Hg$ be continuous and bounded below on its domain. In our setting, this translates to the existence of $c_1, c_2 > 0$ such that

$$c_1 \|\sigma_H\|_{L^2} \le \|Hg\|_{L^2} \le c_2 \|\sigma_H\|_{L^2}, \quad H \in OPW(S). \tag{13}$$

This definition of identifiability of operators originated in [24]. Note that (13) implies that the mapping $H \mapsto Hg$ is *injective*, that is, that $Hg = 0$ implies that $H \equiv 0$, but is not equivalent to it. The inequality (13) adds to injectivity the assertion that H is also stably determined by Hg in the sense that a small change in the output Hg would correspond to a small change in the operator H. Such stability is also necessary for the existence of an algorithm that will reliably recover H from Hg. In this scheme, g is referred to as an *identifier* for the operator class $OPW(S)$ and if (13) holds, we say that *operator identification* is possible.

In trying to find an explicit expression for an identifier, we use as a starting point the "simple measurement scheme" of [19], in which g is a delta train, viz. $g = \sum_n \delta_{nT}$ for some $T > 0$. In the framework of operator identification the channel measurement criterion in [19] takes the following form [24, 38, 41].

Theorem 1. *For $H \in OPW\big([0,T]\times[-\Omega/2,\Omega/2]\big)$ with $T\Omega \le 1$, we have*

$$\|H \sum_{k\in\mathbb{Z}} \delta_{kT}\|_{L^2(\mathbb{R})} = T\|\sigma_H\|_{L^2},$$

and H can be reconstructed by means of

$$\kappa_H(x+t,x) = \chi_{[0,T]}(t) \sum_{n\in\mathbb{Z}} \big(H \sum_{k\in\mathbb{Z}} \delta_{kT}\big)(t+nT) \frac{\sin(\pi T(x-n))}{\pi T(x-n)} \tag{14}$$

where $\chi_{[0,T]}(t) = 1$ for $t \in [0,T]$ and 0 elsewhere and with convergence in the L^2 norm and uniformly in x for every t.

As was observed earlier, the key feature of this scheme is that the spacing of the deltas in the identifier is sufficiently large so as to allow the response of the channel to a given delta to "die out" before the next delta is sent. In other words, the parameter T must exceed the time-spread of the channel. On the other hand, the rate of change of the channel, as measured by its bandwidth Ω, must be small enough that its impulse response can be recovered from "samples" of the channel taken T time units apart. In particular, the samples of the impulse response T units apart can be easily determined from the output. In the general case considered by Bello, in which the spreading support of the operator is not contained in a rectangle of unit area, this intuition breaks down.

Specifically, suppose that we consider the operator class $OPW(S)$ where $S \subseteq [0,T_0] \times [-\Omega_0/2, \Omega_0/2]$ and $T_0\Omega_0 \gg 1$ but where $|S| < 1$. Then sounding the channel with a delta train of the form $g = \sum_n \delta_{nT_0}$ would severely *undersample* the impulse response function. Simply increasing the sampling rate, however, would produce overlap in the responses of the channel to deltas close to each other. An approach to the undersampling problem in the literature of classical sampling theory is to sample at the low rate transformed versions of the function, chosen so that the interference of the several undersampled functions can be dealt with. This idea has its most classical expression in the Generalized Sampling scheme of Papoulis [35]. Choosing shifts and constant multiples of our delta train results in an identifier of the form $g = \sum_n c_n \delta_{nT}$ where the weights (c_n) have period P (for some $P \in \mathbb{N}$) and $T > 0$ satisfies $PT > T_0$.

If g is discretely supported (for example, a periodically-weighted delta-train), then we refer to operator identification as *operator sampling*. The utility of periodically weighted delta trains for operator identification is a cornerstone of operator sampling and has far-reaching implications culminating in the developments outlined in Sections "Generalizations of operator sampling to higher dimensions" and "Further results on operator sampling".

Kailath's necessity proof and operator identification

In Section "Necessity of Kailath's Condition for Channel Identification" we presented the proof of the necessity of the condition $BL \leq 1$ for channel identification as given in [19]. The argument consisted of finding a finite-dimensional approximation of the channel H, and then showing that, given any putative identifier g, the number of degrees of freedom present in the output Hg must be at least as large as the number of degrees of freedom in the channel itself. For this to be true in any finite-dimensional setting, we must have $BL < 1$ and so in the limit we require $BL \leq 1$. In essence, if $BL > 1$, we have a linear system with fewer equations than unknowns which necessarily has a nontrivial nullspace. The generalization of this notion to the infinite-dimensional setting is the basis of the necessity proof that appears in [24]. In this section, we present an outline of that proof, and show how the natural tool for this purpose once again comes from time-frequency analysis.

To see the idea of the proof, assume that $BL > 1$ and for simplicity let $S = [-\frac{L}{2}, \frac{L}{2}] \times [-\frac{B}{2}, \frac{B}{2}]$. The goal is to show that for any sounding signal s in an appropriately large space of distributions[2], the operator $\Phi_s \colon OPW(S) \longrightarrow L^2(\mathbb{R})$, $H \mapsto Hs$, is not stable, that is, it does not possess a lower bound in the inequality (13).

First, define the operator $E \colon l_0(\mathbb{Z}^2) \longrightarrow OPW(S)$, where $l_0(\mathbb{Z}^2)$ is the space of finite sequences equipped with the l^2 norm, by

$$E(\sigma) = E(\{\sigma_{k,l}\}) = \sum_{k,l} \sigma_{k,l} \mathcal{M}_{k\lambda/L} \mathcal{T}_{l\lambda/B} P \mathcal{T}_{-l\lambda/B} \mathcal{M}_{-k\lambda/L}$$

where $1 < \lambda$ is chosen so that $1 < \lambda^4 < BL$ and where P is a time-frequency localization operator whose spreading function $\eta_P(t, \nu)$ is infinitely differentiable, supported in S, and identically one on $[-\frac{L}{2\lambda}, \frac{L}{2\lambda}] \times [-\frac{B}{2\lambda}, \frac{B}{2\lambda}]$. It is easily seen that the operator E is well defined and has spreading function

$$\eta_{E(\sigma)}(t, \nu) = \eta_P(t, \nu) \sum_{k,l} \sigma_{k,l} \, e^{2\pi i (k\lambda t/L - l\lambda \nu/B)}.$$

By construction, it follows that for some constant c_1, $\|E(\sigma)\|_{OPW(S)} \geq c_1 \|\sigma\|_{l^2(\mathbb{Z}^2)}$, for all σ, and that for any distribution s, Ps decays rapidly in time and in frequency.

Next define the Gabor analysis operator $C_g \colon L^2(\mathbb{R}) \longrightarrow l^2(\mathbb{Z}^2)$ by

$$C_g(s) = \{\langle s, \mathcal{M}_{k\lambda^2/L} \mathcal{T}_{l\lambda^2/B} g \rangle\}_{k,l \in \mathbb{Z}}$$

where $g(x) = e^{-\pi x^2}$. A well-known theorem in Gabor theory asserts that $\{\mathcal{M}_{k\alpha} \mathcal{T}_{l\beta} g\}_{k,l \in \mathbb{Z}}$ is a Gabor frame for $L^2(\mathbb{R})$ for every $\alpha\beta < 1$ [31, 59, 60]. Consequently C_g satisfies, for some $c_2 > 0$, $\|C_g(s)\|_{l^2(\mathbb{Z}^2)} \geq c_2 \|s\|_{L^2(\mathbb{R})}$ for all s, since $\lambda^2/L \cdot \lambda^2/B = \lambda^4/BL < 1$.

For any s, consider the composition operator

$$C_g \circ \Phi_s \circ E \colon l_0(\mathbb{Z}^2) \longrightarrow l^2(\mathbb{Z}^2).$$

The crux of the proof lies in showing that this composition operator is not stable, that is, it does not have a lower bound. Since C_g and E are both bounded below, it follows that Φ_s cannot be stable. Since $s \in S_0'(\mathbb{R})$ was arbitrary, this completes the proof.

To complete this final step we examine the canonical bi-infinite matrix representation of the above defined composition of operators, that is, the matrix $M = (m_{k',l',k,l})$ that satisfies

$$(C_g \circ \Phi_s \circ E(\sigma))_{k',l'} = \sum_{k,l} m_{k',l',k,l} \, \sigma_{k,l}.$$

[2] $S_0'(\mathbb{R})$, the dual space of the Feichtinger algebra $S_0(\mathbb{R})$ [13], or $\mathcal{S}'(\mathbb{R})$, the space of tempered distributions [42]. These spaces are large enough to contain weighted infinite sums of delta distributions.

It can be shown that M has the property that for some rapidly decreasing function $w(x)$,

$$|m_{k',l',k,l}| \leq w(\max\{|\lambda k' - k|, |\lambda l' - l|\}). \tag{15}$$

The proof is completed by the following Lemma. Its proof can be found in [24] and generalizations can be found in [37].

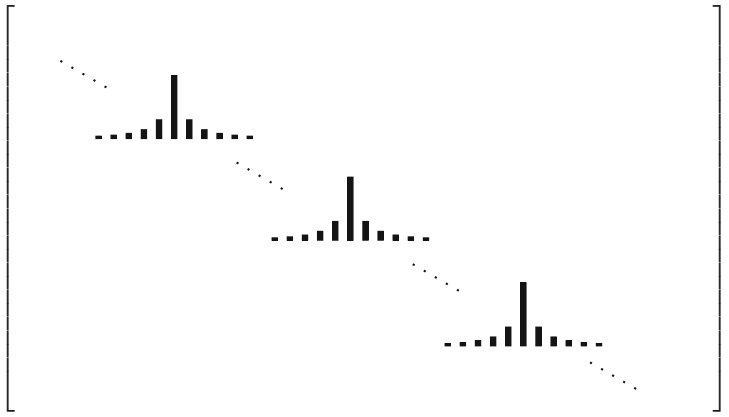


Fig. 2 A $1/\lambda-$slanted matrix M. The matrix is dominated by entries on a slanted diagonal of slope $1/\lambda$.

Lemma 1. *Given* $M = (m_{j',j})_{j',j\in\mathbb{Z}^2}$. *If there exists a monotonically decreasing function* $w\colon R_0^+ \longrightarrow R_0^+$ *with* $w = O(x^{-2-\delta})$, $\delta > 0$, *and constants* $\lambda > 1$ *and* $K_0 > 0$ *with* $|m_{j',j}| < w(\|\lambda j' - j\|_\infty)$ *for* $\|\lambda j' - j\|_\infty > K_0$, *then M is not stable.*

Intuitively, this result asserts that a bi-infinite matrix whose entries decay rapidly away from a skew diagonal behaves like a finite matrix with more rows than columns (see Figure 2). Such a matrix will always have a nontrivial nullspace. In the case of an infinite matrix what can be shown is that at best its inverse will be unbounded.

We can make a more direct connection from this proof to the original necessity argument in [19] in the following way. If we restrict our attention to sequences $\{\sigma_{k,l}\}$ with a fixed finite support of size say N, then the image of this subspace of sequence space under the mapping E is an N-dimensional subspace of $OPW(S)$. The operator P is essentially a time-frequency localization operator. This fact is established in [24] and follows from the rapid decay of the Fourier transform of η_P. Since η_P itself is concentrated on a rectangle of area BL/λ^2, its Fourier transform will be concentrated on a rectangle of area λ^2/BL. From this it follows that for σ as described above, the operator $E(\sigma)$ essentially localizes a function to a region in the time-frequency plane of area $N(\lambda^2/BL)$.

Considering now the Gabor analysis operator C_g, we observe that the Gaussian $g(x)$ essentially occupies a time-frequency cell of area 1, and that this function is

shifted in the time-frequency plane by integer multiples of $(\lambda^2/B, \lambda^2/L)$. Hence to "cover" a region in the time-frequency plane of area $N(\lambda^2/BL)$ would require only about

$$\frac{N(\lambda^2/BL)}{\lambda^4/BL} = \frac{N}{\lambda^2}$$

time-frequency shifts. So roughly speaking, in order to resolve N degrees of freedom in the operator $E(\sigma_{k,l})$, we have only $N/\lambda^2 < N$ degrees of freedom in the output of the operator $E(\sigma_{k,l})s$.

Identification of operator Paley-Wiener spaces by periodically weighted delta-trains

Theorem 1 is based on arguments outlined in Section "Kailath's Time-Variant Channel Identification Condition" and applies only to $OPW(S)$ if S is contained in a rectangle of area less than or equal to one. In the following, we will develop the tools that allow us to identify $OPW(S)$ for any compact set S of Lebesgue measure less than one.

In our approach we discretize the channel by covering the spreading support S with small rectangles of fixed sidelength, which we refer to as a *rectification* of S. As long as the measure of S is less than one, it is possible to do this in such a way that the total area of the rectangles is also less than one. This idea seems to bear some similarity to Bello's philosophy of sampling the spreading function on a fixed grid but with one fundamental difference. Bello's approach is based on replacing t and x by samples, thereby approximating the channel. For a better approximation, sampling on a finer grid is necessary, which results in a larger system of equations that must be solved. In our approach, as soon as the total area of the rectification is less than one, the operator modeling the channel is completely determined by the discrete model. Once this is achieved, identification of the channel reduces to solving a single linear system of equations at each point (Figure 3).

Given parameters $T > 0$ and $P \in \mathbb{N}$, we assume that S is rectified by rectangles of size $T \times \Omega$, where $\Omega = 1/(TP)$, such that the total area of the rectangles is less than one. Given a period-P sequence $c = (c_n)_{n\in\mathbb{Z}}$, we then define the *periodically weighted delta-train* g by $g = \sum_{n\in\mathbb{Z}} c_n \delta_{nT}$. The goal of this subsection is to describe the scheme by which a linear system of P equations in a priori P^2 unknowns can be derived by which an operator $H \in OPW(S)$ can be completely determined by $Hg(x)$. In this sense, the "degrees of freedom" in the operator class $OPW(S)$, and that of the output function $Hg(x)$ are precisely defined and can be effectively compared (Figure 4).

The basic tool of time-frequency analysis that makes this possible is the *Zak transform* (see [13]).

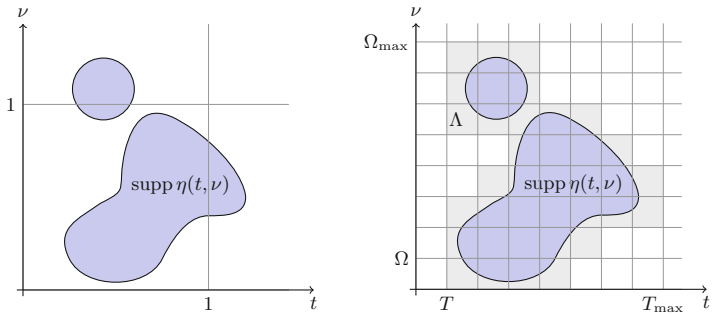


Fig. 3 A set not satisfying Kailath's condition is rectified with $1/(T\Omega) = P \in \mathbb{N}$, the rectification has area ≤ 1, $\Omega_{\max} \leq 1/T$, and $T_{\max} \leq 1/\Omega$.

Definition 2. The non-normalized Zak Transform is defined for $f \in \mathcal{S}(\mathbb{R})$[3], and $a > 0$ by

$$Z_a f(t,\nu) = \sum_{n\in\mathbb{Z}} f(t-an)\, e^{2\pi i a n \nu}.$$

$Z_a f(t,\nu)$ satisfies the quasi-periodicity relations

$$Z_a f(t+a,\nu) = e^{2\pi i a \nu}\, Z_a f(t,\nu)$$

and

$$Z_a f(t,\nu+1/a) = Z_a f(t,\nu).$$

$\sqrt{a}\, Z_a$ can be extended to a unitary operator from $L^2(\mathbb{R})$ onto $L^2([0,a]\times[0,1/a])$.

A somewhat involved but elementary calculation yields the following (see [44, 46] and Section "Proof of Lemma 2").

Lemma 2. *Let $T > 0$, $P \in \mathbb{N}$, $c = (c_n)$, and g be given as above. Then for all $(t,\nu) \in \mathbb{R}^2$, and $p = 0, 1, \ldots, P-1$,*

$$\begin{aligned} &e^{-2\pi i \nu T p}\, (Z_{TP} \circ H) g(t+Tp,\nu) \\ &= \Omega \sum_{q,m=0}^{P-1} (T^q M^m c)_p\, e^{-2\pi i \nu T q}\, \eta_H^Q(t+Tq, \nu+m/TP). \end{aligned} \tag{16}$$

Here $\mathcal{T}$ and $\mathcal{M}$ are the translation and modulation operators given in Definition 3, and $\eta_H^Q(t,\nu)$ is the *quasiperiodization* of η_H,

$$\eta_H^Q(t,\nu) = \sum_k \sum_\ell \eta_H(t+kTP, \nu+\ell/T)\, e^{-2\pi i \nu k T P} \tag{17}$$

whenever the sum is defined (Figure 4).

Under the additional simplifying assumption that the spreading function $\eta_H(t,\nu)$ is supported in the large rectangle $[0,TP]\times[0,1/T]$, and by restricting (16) to the rectangle $[0,T]\times[0,1/(TP)]$, we arrive at the $P\times P^2$ linear system

[3] $\mathcal{S}(\mathbb{R})$ denotes the Schwartz class of infinitely differentiable, rapidly decreasing functions.

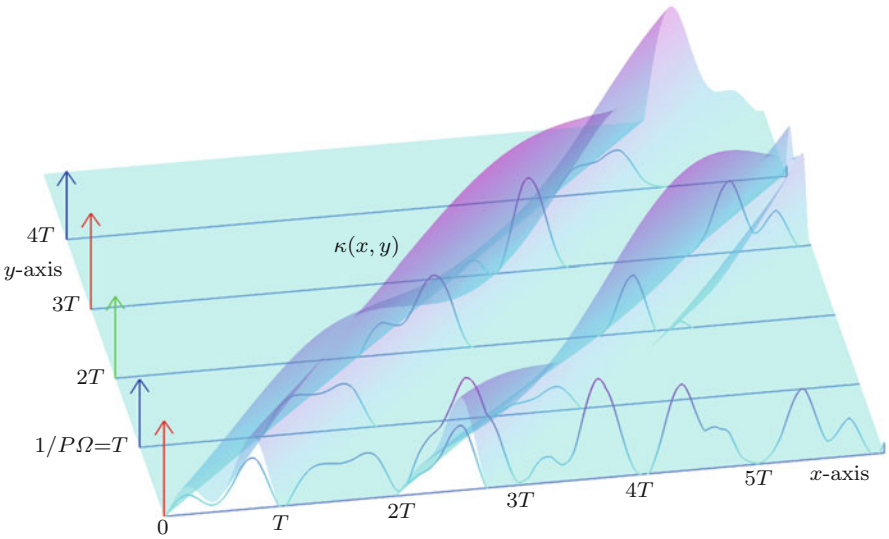


Fig. 4 Channel sounding of $OPW([0,2/3]\times[-1/4,1/4]\cup[4/3,2]\times[-1/2,1/2])$ using a P-periodically weighted delta train g. The kernel $\kappa(x,y)$ takes values on the (x,y)-plane, the sounding signal g, a weighted impulse train, is defined on the y-axis, and the output signal $Hg(x)=\int\kappa(x,y)g(y)dy$ is displayed on the x-axis. Here, the sample values of the tab functions $h(x,t)=\kappa(x,t-x)$ are not easily read of the response $Hg(x)$ as, for example, for $x\in[2T,3T]=[4/3,2]$ we have $Hg(x)=0.7\kappa(x,0)+0.6\kappa(x,2T)=0.7h(x,x)+.6h(x,2T-x)$. In detail, we have $g=\ldots+0.7\delta_{-2}+0.5\delta_{-4/3}+0.6\delta_{-2/3}+0.7\delta_0+0.5\delta_{2/3}+0.6\delta_{4/3}+0.7\delta_2+0.5\delta_{8/3}+\ldots$, so $P=3$, $T=2/3$, $\Omega=1/PT=1/2$, $c_n=0.7$ if $n \mod 3=0$, $c_n=0.5$ if $n \mod 3=1$, $c_n=0.6$ if $n \mod 3=2$.

$$\mathbf{Z}_{Hg}(t,\nu)_p=\sum_{q,m=0}^{P-1}G(c)_{p,(q,m)}\,\boldsymbol{\eta}_H(t,\nu)_{(q,m)} \tag{18}$$

where

$$\mathbf{Z}_{Hg}(t,\nu)_p=(Z_{TP}\circ H)g(t+pT,\nu)\,e^{-2\pi i\nu pT}, \tag{19}$$

$$\boldsymbol{\eta}_H(t,\nu)_{(q,m)}=\Omega\,\eta_H(t+qT,\nu+m/TP)\,e^{-2\pi i\nu qT}\,e^{-2\pi iqm/P}, \tag{20}$$

and where $G(c)$ is a finite Gabor system matrix (23). If (18) can be solved for each $(t,\nu)\in[0,T]\times[0,1/(TP)]$, then the spreading function for an operator H can be completely determined by its response to the periodically weighted delta-train g.

As anticipated by Bello, two issues now become relevant. (1) We require that $\operatorname{supp}\eta_H$ occupy no more than P of the shifted rectangles $[0,T]\times[0,1/(TP)]+(qT,k/(TP))$, so that (18) has at least as many equations as unknowns. This forces $|\operatorname{supp}\eta_H|\le 1$. (2) We require that c be chosen in such a way that the $P\times P$ system formed by removing the columns of $G(c)$ corresponding to vanishing components of $\boldsymbol{\eta}_H$ is invertible. That such c exist is a fundamental cornerstone of operator sampling and is the subject of the next section.

Based on the existence of c such that any set of P columns of $G(c)$ form a linearly independent set, we can prove the following [45].

Theorem 2. *For $S \subseteq (0,\infty)\times\mathbb{R}$ compact with $|S| < 1$, there exists $T > 0$ and $P \in \mathbb{N}$, and a period-P sequence $c = (c_n)$ such that $g = \sum_n c_n \delta_{nT}$ identifies $OPW(S)$. In particular, there exist period-P sequences $b_j = (b_{j,k})$, and integers $0 \le q_j, m_j \le P-1$, for $0 \le j \le P-1$ such that*

$$h(x,t) = e^{-\pi i t/T} \sum_k \sum_{j=0}^{P-1} \big[b_{j,k} H g(t-(q_j-k)T) \\ e^{2\pi i m_j (x-t)/PT}\, \phi((x-t)+(q_j-k)T)\, r(t-q_jT) \big] \tag{21}$$

where $r, \phi \in \mathcal{S}(\mathbb{R})$ satisfy

$$\sum_{k\in\mathbb{Z}} r(t+kT) = 1 = \sum_{n\in\mathbb{Z}} \widehat{\phi}(\gamma + n/PT), \tag{22}$$

where $r(t)\widehat{\phi}(\gamma)$ is supported in a neighborhood of $[0,T]\times[0,1/PT]$, and where the sum in (21) converges unconditionally in L^2 and for each t uniformly in x.

Equation (21) is a generalization of (14) which is easily seen by choosing $\phi(x) = \sin(\pi PTx)/(\pi PTx)$ and $r(t)$ to be the characteristic function of $[0,T)$.

Linear Independence Properties of Gabor Frames

Finite Gabor Frames

Definition 3. Given $P \in \mathbb{N}$, let $\omega = e^{2\pi i/P}$ and define the *translation operator* $\mathcal{T}$ on $(x_0, \ldots, x_{P-1}) \in \mathbb{C}^P$ by

$$\mathcal{T}x = (x_{P-1}, x_0, x_1, \ldots, x_{P-2}),$$

and the *modulation operator* $\mathcal{M}$ on $\mathbb{C}^P$ by

$$\mathcal{M}x = (\omega^0 x_0, \omega^1 x_1, \ldots, \omega^{P-1} x_{P-1}).$$

Given a vector $c \in \mathbb{C}^P$ the *finite Gabor system with window c* is the collection $\{\mathcal{T}^q \mathcal{M}^p c\}_{q,p=0}^{P-1}$. Define the *full Gabor system matrix* $G(c)$ to be the $P \times P^2$ matrix

$$G(c) = [\, D_0 W_P \,|\, D_1 W_P \,|\, \cdots \,|\, D_{P-1} W_P \,] \tag{23}$$

where D_k is the diagonal matrix with diagonal

$$\mathcal{T}^k c = (c_{P-k}, \ldots, c_{P-1}, c_0, \ldots, c_{P-k-1}),$$

and W_P is the $P \times P$ Fourier matrix $W_P = (e^{2\pi i nm/P})_{n,m=0}^{P-1}$.

Remark 2. (1) For $0 \le q, p \le P-1$, the $(q+1)$st column of the submatrix $D_p W_P$ is the vector $\mathcal{M}^p \mathcal{T}^q c$ where the operators $\mathcal{M}$ and $\mathcal{T}$ are as in Definition 3. This means that each column of the matrix $G(c)$ is a unimodular constant multiple of an element of the finite Gabor system with window c, namely $\{e^{-2\pi ipq/P} \mathcal{T}^q \mathcal{M}^p c\}_{q,p=0}^{P-1}$.

(2) Note that the finite Gabor system defined above consists of P^2 vectors in $\mathbb{C}^P$ which form an overcomplete tight frame for $\mathbb{C}^P$ [30]. For details on Gabor frames in finite dimensions, see [9, 27, 30] and the overview article [39].

(3) Note that we are abusing notation slightly by identifying a vector $c \in \mathbb{C}^P$ with a P-periodic sequence $c = (c_n)$ in the obvious way.

Definition 4. [8] The *Spark* of an $M \times N$ matrix F is the size of the smallest linearly dependent subset of columns, i.e.,

$$Spark(F) = \min\{\|x\|_0 \colon Fx = 0, \ \ x \neq 0\}$$

where $\|x\|_0$ is the number of nonzero components of the vector x. If $Spark(F) = M+1$, then F is said to have *full Spark*. $Spark(F) = k$ implies that any collection of fewer than k columns of F is linearly independent.

Finite Gabor frames are generically full Spark

The existence of Gabor matrices with full Spark has been addressed in [30, 32]. The results in these two papers are as follows.

Theorem 3. *[30] If $P \in \mathbb{N}$ is prime, then there exists a dense, open subset of $c \in \mathbb{C}^P$ such that every minor of the Gabor system matrix $G(c)$ is nonzero. In particular, for such c, $G(c)$ has full Spark.*

Theorem 4. *[32] For every $P \in \mathbb{N}$, there exists a dense, open subset of $c \in \mathbb{C}^P$ such that the Gabor system matrix $G(c)$ has full Spark.*

The goal of this subsection is to outline the proof of Theorems 3 and 4. We will adopt some of the following notation and terminology of [32].

Let $P \in \mathbb{N}$ and let M be an $P \times P$ submatrix of $G(c)$. For $0 \le \kappa < P$ let ℓ_κ be the number of columns of M chosen from the submatrix $D_\kappa W_P$ of (23). While the vector $\ell = (\ell_\kappa)_{\kappa=0}^{P-1}$ does not determine M uniquely, it describes the matrix M sufficiently well for our purposes. Define M_κ to be the $P \times \ell_\kappa$ matrix consisting of those columns of M chosen from $D_\kappa W_P$. Given the *ordered partition* $B = (B_0, B_1, \ldots, B_{P-1})$ where $\{B_0, B_1, \ldots, B_{P-1}\}$ forms a partition of $\{0, \ldots, P-1\}$, and where for each $0 \le \kappa < P$, $|B_\kappa| = \ell_\kappa$, let $M_\kappa(B_\kappa)$ be the $\ell_\kappa \times \ell_\kappa$ submatrix of M_κ whose rows belong to B_κ. Then $\det(M) = \sum \prod_{\kappa=0}^{P-1} \det(M_\kappa(B_\kappa))$ where the sum is taken over all such ordered partitions B. This formula is called the *Lagrange expansion* of the determinant.

Each ordered partition B corresponds to a permutation on $\mathbb{Z}_P$ as follows. Define the *trivial partition* $A = (A_0, A_1, \ldots, A_{P-1})$ by

$$A_j = \{\sum_{i=0}^{j-1} \ell_i, (\sum_{i=0}^{j-1} \ell_i) + 1, \ldots, (\sum_{i=0}^{j} \ell_i) - 1\}$$

so that $A_0 = [0, \ell_0 - 1]$, $A_1 = [\ell_0, \ell_0 + \ell_1 + 1]$, $\ldots$, $A_{P-1} = [\ell_0 + \cdots + \ell_{P-2}, P-1]$. Then given $B = (B_0, B_1, \ldots, B_{P-1})$ there is a permutation $\sigma \in S_P$ such that $\sigma(A_\kappa) = B_\kappa$ for all κ. This σ is unique up to permutations that preserve A, that is, up to $\tau \in S_P$ such that $\tau(A_\kappa) = A_\kappa$ for all κ. Call such a permutation *trivial* and denote by Γ the subgroup of S_P consisting of all trivial permutations. Then the ordered partitions B of $\mathbb{Z}_P$ can be indexed by equivalence classes of permutations $\sigma \in S_P/\Gamma$.

The key observation is that $\det(M)$ is a homogeneous polynomial in the P variables $c_0, c_1, \ldots, c_{P-1}$ and we can write

$$\det(M) = \sum_{\sigma \in S_P/\Gamma} a_\sigma C^\sigma \tag{24}$$

where the monomial C^σ is given by

$$C^\sigma = \prod_{\kappa=0}^{P-1} \prod_{j \in \sigma(A_\kappa)} c_{(j-\kappa)(mod\ P)}.$$

If it can be shown that this polynomial does not vanish identically, then we can choose a dense, open subset of $c \in \mathbb{C}^P$ for which $\det(M) \neq 0$. Since there are only finitely many $P \times P$ submatrices of $G(c)$ it follows that there is a dense, open subset of c for which $\det(M) \neq 0$ for all M, and we conclude that, for these c, $G(c)$ has full Spark.

Following [32], we say that a monomial C^{σ_0} *appears uniquely* in (24) if for every $\sigma \in S_P/\Gamma$ such that $\sigma \neq \sigma_0$, $C^\sigma \neq C^{\sigma_0}$. Therefore, in order to show that the polynomial (24) does not vanish identically, it is sufficient to show that (1) there is a monomial C^σ that appears uniquely in (24) and (2) the coefficient a_σ of this monomial does not vanish.

Obviously, whether or not (24) vanishes identically does not depend on how the variables c_i are labelled. More specifically, if the variables are renamed by a cyclical shift of the indices, viz., $c_i \mapsto c_{(i+\gamma)mod\ P}$ for some $0 \leq \gamma < P$, then

$$\det(M)(c_{\gamma+1}, \ldots, c_{P-1}, c_0, \ldots, c_\gamma) = \pm \det(M')(c_0, \ldots, c_{P-1})$$

where M' is a $P \times P$ submatrix described by the vector

$$\ell' = (\ell_{\gamma+1}, \ldots, \ell_{P-1}, \ell_0, \ldots, \ell_\gamma).$$

The lowest index monomial

In [30], a monomial referred to in [32] as the *lowest index (LI) monomial* is defined that has the required properties when P is prime. In order to see this, note first that each coefficient a_σ in the sum (24) is the product of minors of the Fourier matrix

W_P and since P is prime, Chebotarev's Theorem says that such minors do not vanish [62]. More specifically,

$$a_\sigma C^\sigma = \pm \prod_{\kappa=0}^{P-1} \det(M_\kappa(\sigma(A_\kappa)))$$

and for each κ, the columns of M_κ are columns of W_P where each row has been multiplied by the same variable c_j and $M_\kappa(\sigma(A_\kappa))$ is a square matrix formed by choosing ℓ_κ rows of M_κ. Hence for each κ, $\det(M_\kappa(\sigma(A_\kappa)))$ is a monomial in c with coefficients a constant multiple of a minor of W_P. Since a_σ is the product of those minors, it does not vanish.

Note moreover that each submatrix $M_\kappa(\sigma(A_\kappa))$ is an $\ell_\kappa \times \ell_\kappa$ matrix, so that $\det(M_\kappa(\sigma(A_\kappa)))$ is the sum of a multiple of the product of $\ell_\kappa!$ diagonals of $M_\kappa(\sigma(A_\kappa))$. Hence $a_\sigma C^\sigma$ is the sum of multiples of the product of $\prod_{\kappa=0}^{P-1} \ell_\kappa!$ generalized diagonals of M.

We define the LI monomial as in [30] as follows. If M is 1×1, then $\det(M)$ is a multiple of a single variable c_j and we define the LI monomial, p_M by $p_M = c_j$. If M is $d \times d$, let c_j be the variable of lowest index appearing in M. Choose any entry of M in which c_j appears, eliminate the row and column containing that entry, and call the remaining $(d-1) \times (d-1)$ matrix M'. Define $p_M = c_j p_{M'}$. It is easy to see that the monomial p_M is independent of the entry of M chosen at each step. In order to show that the LI monomial appears uniquely in (24), we observe as in [30] that the number of diagonals in $\det(M)$ that correspond to the LI monomial is $\prod_{\kappa=0}^{P-1} \ell_\kappa!$. Since this is also the number of generalized diagonals appearing in the calculation of each $\det(M_\kappa(\sigma(A_\kappa)))$, it follows that this monomial appears only once. The details of the argument can be found in Section "Proof of Theorem 3". Note that because P is prime, this argument is valid no matter how large the matrix M is. In other words, M does not have to be a $P \times P$ submatrix in order for the result to hold. Consequently, given $k < P$ and M an arbitrary $P \times k$ submatrix of $G(c)$, we can form the $k \times k$ matrix M_0 by choosing k rows of M in such a way that the LI monomial of M_0 contains at most only the variables $c_0, \ldots, c_{k-1}$. This observation leads to the following theorem for matrices with arbitrary Spark.

Theorem 5. *[46] If $P \in \mathbb{N}$ is prime, and $0 < k < P$, there exists an open, dense subset of $c \in \mathcal{C}^k \times \{0\} \subseteq \mathbb{C}^P$ with the property that $Spark(G(c)) = k+1$.*

This result has implications for relating the capacity of a time-variant communication channel to the area of the spreading support, see [46].

The consecutive index monomial

In [32], a monomial referred to as the *consecutive index (CI) monomial* is defined that has the required properties for any $P \in \mathbb{N}$. The CI monomial, C^I, is defined as the monomial corresponding to the identity permutation in S_P/Γ, that is, to the equivalence class of the trivial partition $A = (A_0, A_1, \ldots, A_{P-1})$. Hence

$$C^I = \prod_{\kappa=0}^{P-1} \prod_{j \in A_\kappa} c_{(j-\kappa) mod\ P}.$$

For each κ, the monomial appearing in $\det(M_\kappa(A_\kappa))$, $\prod_{j\in A_\kappa} c_{(j-\kappa) mod\ P}$, consists of a product of ℓ_k variables c_j with consecutive indices modulo P.

That $a_I \neq 0$ follows from the observation that for each κ, $\det(M_\kappa(A_\kappa))$ is a monomial whose coefficient is a nonzero multiple of a Vandermonde determinant and hence does not vanish (for details, see [32]). The proof that C^I appears uniquely in (24) amounts to showing that, with respect to an appropriate cyclical renaming of the variables c_i, the CI monomial uniquely minimizes the quantity $\Lambda(C^\sigma) = \sum_{i=0}^{P-1} i^2\, \alpha_i$, where α_i is the exponent of the variable c_i in C^σ. An abbreviated version of the proof of this result as it appears in [32] is given in Section "Proof of Theorem 4".

As a final observation, we quote the following corollary that provides an explicit construction of a unimodular vector c such that $G(c)$ has full Spark.

Corollary 1. *[32] Let $\zeta = e^{2\pi i/(P-1)^4}$ or any other primitive root of unity of order $(P-1)^4$ where $P \geq 4$. Then the vector*

$$c = (1, \zeta, \zeta^4, \zeta^9, \ldots, \zeta^{(P-1)^2})$$

generates a Gabor frame for which $G(c)$ has full Spark.

Generalizations of operator sampling to higher dimensions

The operator representations (5), (6), and (7) hold verbatim for higher dimensional variables $x, \xi, t, \nu \in \mathbb{R}^d$. In this section, we address the identifiability of

$$OPW(S) = \{H \in \mathcal{L}(L^2(\mathbb{R}^d), L^2(\mathbb{R}^d)) \colon \operatorname{supp} \mathcal{F}_s \sigma_H \subseteq S, \|\sigma_H\|_{L^2} < \infty\}$$

where $S \subseteq \mathbb{R}^{2d}$.

Looking at the components of the multidimensional variables separately, Theorem 1 easily generalizes as follows.

Theorem 6. *For $H \in OPW\left(\prod_{\ell=1}^d [0, T_\ell] \times \prod_{\ell=1}^d [-\Omega_\ell/2, \Omega_\ell/2]\right)$ with $T_\ell \Omega_\ell \leq 1$, $\ell = 1, \ldots, d$, we have*

$$\|H \sum_{k_1 \in \mathbb{Z}} \cdots \sum_{k_d \in \mathbb{Z}} \delta_{(k_1 T_1, \ldots, k_d T_d)}\|_{L^2(\mathbb{R})} = T_1 \ldots T_d \|\sigma_H\|_{L^2},$$

and H can be reconstructed by means of

$$\kappa_H(x+t,x) = \chi_{\prod_{\ell=1}^d [0,T_\ell]}(t) \sum_{n_1\in\mathbb{Z}} \cdots \sum_{n_d\in\mathbb{Z}} \Big(H \sum_{k_1\in\mathbb{Z}} \cdots \sum_{k_d\in\mathbb{Z}} \delta_{(k_1T_1,\ldots,k_dT_d)}\Big)(t+(n_1T_1,\ldots,n_dT_d) \frac{\sin(\pi T_1(x_1-n_1))}{\pi T_1(x_1-n_1)} \cdots \frac{\sin(\pi T_d(x_d-n_d))}{\pi T_d(x_d-n_d)}$$

with convergence in the L^2 norm.

In the following, we address the situation where S is not contained in a set $\prod_{\ell=1}^d [0,T_\ell] \times \prod_{\ell=1}^d [-\Omega_\ell/2, \Omega_\ell/2]$ with $T_\ell\Omega_\ell \le 1$, $\ell = 1,\ldots,d$. For example, $S = [0,1]\times[0,2]\times[0,\frac{1}{4}]\times[0,1] \subseteq \mathbb{R}^4$ of volume $\frac{1}{2}$ is not covered by Theorem 6.

To give a higher dimensional variant of Theorem 2, we shall denote pointwise products of finite and infinite length vectors k and T by $k\star T$, that is, $k\star T = (k_1T_1,\ldots,k_dT_d)$ for $k,T\in\mathbb{C}^d$. Similarly, $k/T = (k_1/T_1,\ldots,k_d/T_d)$.

Theorem 7. *If $S \subseteq (0,\infty)^d \times \mathbb{R}^d$ is compact with $|S| < 1$, then $OPW(S)$ is identifiable. Specifically, there exist $T_1,\ldots,T_d > 0$ and pairwise relatively prime natural numbers $P_1,\ldots,P_d$ such that*

$$S \subseteq \prod_{\ell=1}^d [0,P_\ell T_\ell] \times \prod_{\ell=1}^d [-1/(2T_\ell), 1/(2T_\ell)],$$

and a sequence $c = (c_n) \in \ell^\infty(\mathbb{Z}^d)$ which is P_ℓ periodic in the ℓ-th component n_ℓ such that $g = \sum_{n\in\mathbb{Z}^d} c_n \delta_{n\star T}$ identifies $OPW(S)$. In fact, for such g there exists for each $j \in J = \prod_{\ell=1}^d \{0,1,\ldots,P_\ell-1\}$ a sequence $b_j = (b_{j,k})$ which is P_ℓ periodic in k_ℓ and $2d$-tuples $(q_j,m_j) \in J\times J$ with

$$h(x,t) = e^{-\pi i \Sigma_{\ell=1}^d t_\ell/T_\ell} \sum_{k\in\mathbb{Z}^d} \sum_{j\in J} \big[b_{j,k} Hg(t-(q_j-k)\star T) e^{2\pi i m_j\cdot((x-t)/P\star T)} \phi((x-t)+(q_j-k)\star T)\, r(t-q_j\star T)\big]. \tag{25}$$

The functions $r,\phi \in \mathcal{S}(\mathbb{R}^d)$ are assumed to satisfy

$$\sum_{k\in\mathbb{Z}^d} r(t+k\star T) = 1 = \sum_{n\in\mathbb{Z}^d} \widehat{\phi}(\gamma+(n/(P\star T)), \tag{26}$$

and $r(t)\widehat{\phi}(\gamma)$ is supported in a neighborhood of $\prod_{\ell=1}^d [0,T_\ell]\times\prod_{\ell=1}^d [0,1/P_\ell T_\ell]$. The sum in (25) converges unconditionally in L^2 and for each t uniformly in x.

This result follows from adjusting the proof of Theorem 7 to the higher dimensional setting. For example, it will employ the Zak transform

$$Z_{T\star P} f(t,\nu) = \sum_{n\in\mathbb{Z}^d} f(t-n\star P\star T)\, e^{2\pi i\nu\cdot(P\star T)},$$

where $P = (P_1, \ldots, P_d)$. We are then led again to a system of linear equations of the form

$$\mathbf{Z}_{Hg}(t,\nu)_p = \sum_{q\in J}\sum_{m\in J} G(c)_{p,(q,m)}\, \boldsymbol{\eta}_H(t,\nu)_{(q,m)} \tag{27}$$

where as before

$$\mathbf{Z}_{Hg}(t,\nu)_p = (Z_{T\star P}\circ H)g(t+p\star T,\nu)\, e^{-2\pi i \nu p\star T},$$
$$\boldsymbol{\eta}_H(t,\nu)_{(q,m)} = (T_1P_1\ldots T_dP_d)^{-1}\,\eta_H(t+q\star T,\nu+(m/(T\star P))$$
$$e^{-2\pi i\nu\cdot(q\star T)}\, e^{-2\pi i q\cdot(m/P)},$$

and where $G(c)$ is now a multidimensional finite Gabor system matrix similar to (23).

In order to show that the spreading function for operator H can be completely determined by its response to the periodically weighted d-dimensional delta-train g, we need to show that (27) can be solved for each $(t,\nu) \in \prod_{\ell=1}^d [0,T_\ell]\times\prod_{\ell=1}^d [0,1/(T_\ell P_\ell)]$ if $c \in \mathbb{C}^{P_1\times\ldots\times P_d}$ is chosen appropriately.

To see that a choice of c is possible, observe that the product group $\mathbb{Z}_{P_1}\times\ldots\times\mathbb{Z}_{P_d}$ is isomorphic to the cyclic group $\mathbb{Z}_{P_1\cdot\ldots\cdot P_d}$ since the P_ℓ are chosen pairwise relatively prime. Theorem 4 applied to the cyclic group $\mathbb{Z}_{P_1\cdot\ldots\cdot P_d}$ guarantees the existence of $\widetilde{c}\in\mathbb{C}^{P_1\cdot\ldots\cdot P_d}$ so that the Gabor system matrix $G(\widetilde{c})$ is full spark. We can now define $c\in\mathbb{C}^{P_1\times\ldots\times P_d}$ by setting

$$c_{n_1,\ldots,n_d} = \widetilde{c}_{n_1+n_2P_1+n_3P_1P_2+\ldots+n_dP_1\ldots P_{d-1}}, \quad n=(n_1,\ldots,n_d)\in J$$

and observe that $G(c)$ is simply a rearrangement of $G(\widetilde{c})$, hence, $G(c)$ is full spark.

Further results on operator sampling

The results discussed in this paper are discussed in detail in [6, 22, 24, 38, 41] and [46]. The last listed article contains the most extensive collection of operator reconstruction formulas, including extensions to some $OPW(S)$ with S unbounded. Moreover, some hints on how to use parallelograms to rectify a set S for operator sampling efficiently are given.

A central result in [46] is the classification of all spaces $OPW(S)$ that are identifiable for a given $g=\sum_{n\in\mathbb{Z}} c_n\delta_{nT}$ for c_n being P-periodic.

The papers [38, 42] address some functional analytic challenges in operator sampling, and [26] focuses on the question of operator identification if we are restricted to using more realizable identifiers, for example, truncated and modified versions of g, namely, $\widetilde{g}(t)=\sum_{n=0}^N c_n\varphi(t-nT)$. The problem of recovering parametric classes of operators in $OPW(S)$ is discussed in [2, 3].

In the following, we briefly review literature that address some other directions in operator sampling.

Multiple Input Multiple Output

A Multiple Input Multiput Output (MIMO) channel $\mathbf{H}$ with N transmitters and M receivers can be modeled by an $N \times M$ matrix whose entries are time-varying channel operators $H_{mn} \in OPW(S_{mn})$. For simplicity, we write $\mathbf{H} \in OPW(\mathbf{S})$. Assuming that the operators H_{mn} are independent, a sufficient criterion for identifiability is given by $\sum_{n=1}^{N} |S_{mn}| \leq 1$ for $m = 1, \ldots, M$. Conversely, if for a single m, $\sum_{n=1}^{N} |S_{mn}| > 1$, then $OPW(\mathbf{S})$ is not identifiable by any collection $s_1, \ldots, s_N$ of input signals [36, 43].

A somewhat dual setup was considered in [18]. Namely, a Single Input Single Output (SISO) channel with S being large, say $S = [0,M] \times [-N/2, N/2]$ with $N, M \geq 2$. As illustrated above, $OPW([0,M] \times [-N/2, N/2])$ is not identifiable, but if we are allowed to use MN (infinite duration) input signals $g_1, \ldots, g_{MN}$, then $H \in OPW([0,M] \times [-N/2, N/2])$ can be recovered from the MN outputs $Hg_1, \ldots, Hg_{MN}$.

Irregular Sampling of Operators

The identifier $g = \sum_{n \in \mathbb{Z}} c_n \delta_{nT}$ is supported on the lattice $T\mathbb{Z}$ in $\mathbb{R}$. In general, for stable operator identification, choosing a discretely supported identifier is reasonable, indeed, in [26] it is shown that identification for $OPW(S)$ in full requires the use of identifiers that neither decay in time nor in frequency. (Recovery of the characteristics of H during a fixed transmission band and a fixed transmission interval can be indeed recovered when using Schwartz class identifiers [26].)

In irregular operator sampling, we consider identifiers of the form $g = \sum_{n \in \mathbb{Z}} c_n \delta_{\lambda_n}$ for nodes λ_n that are not necessarily contained in a lattice. If such g identifies $OPW(S)$, then we refer to $\operatorname{supp} g = \{\lambda_n\}$ as a sampling set for $OPW(S)$, and similarly, the *sampling rate* of g is defined to be

$$D(g) = D(\operatorname{supp} g) = D(\Lambda) = \lim_{r \to \infty} \frac{n^-(r)}{r}$$

where

$$n^-(r) = \inf_{x \in \mathbb{R}} \#\{\Lambda \cap [x, x+r]\}$$

assuming that the limit exists [18, 46].

To illustrate a striking difference between irregular sampling of functions and operators, note that $\mathbb{Z}$ is a sampling set for $OPW([0,1] \times [-\frac{1}{2}, \frac{1}{2}])$ as well as for the Paley Wiener space $PW([-\frac{1}{2}, \frac{1}{2}])$, but the distribution $g = c_0 \delta_{\lambda_0} + \sum_{n \in \mathbb{Z} \setminus \{0\}} c_n \delta_n$ does not identify $OPW([0,1] \times [-\frac{1}{2}, \frac{1}{2}])$, regardless of our choice of c_n and $\lambda_0 \neq 0$. This shows that, for example, Kadec's $\frac{1}{4}$th theorem does not generalize to the operator setting [18].

In [46] we give with $D(g) = D(\Lambda) \geq B(S)$ a necessary condition on the (operator) sampling rate based on the bandwidth $B(S)$ of $OPW(S)$ which is defined as

$$B(S) = \sup_{t\in\mathbb{R}} |\operatorname{supp} \eta(t,\nu)| = \left\| \int_{\mathbb{R}} \chi_S(\cdot,\nu)\, d\nu \right\|_\infty . \tag{28}$$

Here, χ_S denotes the characteristic function of S. This quantity can be interpreted as the maximum vertical extent of S and takes into account gaps in S. Moreover, in [46] we discuss the goal of constructing $\{\lambda_n\}$ of small density, and/or large gaps in order to reserve time-slots for information transmission. Results in this direction can be interpreted as giving bounds on the capacity of a time-variant channel in $OPW(S)$ in terms of $|S|$ [46].

Finally, we give in [46] an example of an operator class $OPW(S)$ that cannot be identified by any identifier of the form $g = \sum_{n\in\mathbb{Z}} c_n \delta_{nT}$ with $T > 0$ and periodic c_n, but requires coefficients that form a bounded but non-periodic sequence. In this case, S is a parallelogram and $B(S) = D(g)$ (see Figure 5)

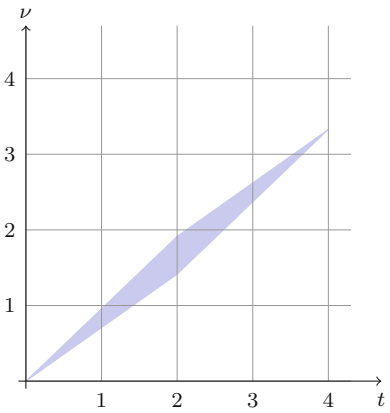


Fig. 5 The operator class $OPW(S)$ with $S = (2,\ 2\,;\ \sqrt{2},\ \sqrt{2}+1/2)[0,1]^2$ whose area equals 1 and bandwidth equals $1/2$ is identifiable by a (non-periodically) weighted delta train with sampling density $1/2$. It is not identifiable using a periodically weighted delta train.

Sampling of $OPW(S)$ with unknown S

In some applications, it is justified to assume that the set S has small area, but its shape and location are unknown. If further S satisfies some basic geometric assumptions that guarantee that S is contained in $[0,TP]\times[-1/2T,1/2T]$ and only meets few rectangles of the rectification grid $[kT,(k+1)T]\times[q/TP,(q+1)/TP]$, then recovery of S and, hence, an operator in $OPW(S)$ is possible [15, 46].

The independently obtained results in [15, 46] employ the same identifiers $g = \sum_{n\in\mathbb{Z}} c_n \delta_{\lambda_n}$ as introduced above. Operator identification is therefore again reduced to solving (18), that is, the system of P linear equations

$$\mathbf{Z}(t,\nu) = G(c)\boldsymbol{\eta}(t,\nu) \tag{29}$$

for the vector $\boldsymbol{\eta}(t,\nu) \in \mathbb{C}^{P^2}$ for $(t,\nu) \in [0,T] \times [-1/2TP, 1/2TP]$. While the zero components of $\boldsymbol{\eta}(t,\nu)$ are not known, the vector is known to be very sparse. Hence, for fixed (t,ν), we can use the fact that $G(c)$ is full spark and recover $\boldsymbol{\eta}(t,\nu)$ if it has at most $P/2$ nonzero entries. Indeed, assume $\boldsymbol{\eta}(t,\nu)$ and $\widetilde{\boldsymbol{\eta}}(t,\nu)$ solve (29) and both have at most $P/2$ nonzero entries. Then $\boldsymbol{\eta}(t,\nu) - \widetilde{\boldsymbol{\eta}}(t,\nu)$ has at most P nonzero entries and the fact that $G(c)$ is full spark indicates that $G(c)(\boldsymbol{\eta}(t,\nu) - \widetilde{\boldsymbol{\eta}}(t,\nu)) = 0$ implies $\boldsymbol{\eta}(t,\nu) - \widetilde{\boldsymbol{\eta}}(t,\nu) = 0$.

Clearly, under the geometric assumptions alluded to above, the criterion that at most $P/2$ rectangles in the grid are met can be translated to the unknown area of S has measure less than or equal to 1/2.

In [15], this area 1/2 criterion is improved by showing that H can be identified whenever at most $P-1$ rectangles in the rectification grid are met by S. This result is achieved by using a joint sparsity argument, based on the assumption that for all (t,ν), the same cells are active.

Alternatively, the "area 1/2" result can be strengthened by not assuming that for all (t,ν), the same cells are active. This requires solving (29), for $\boldsymbol{\eta}(t,\nu)$ sparse, for each considered (t,ν) independently, see Figure 6 and [46].

It must be added though, that solving (29) for $\boldsymbol{\eta}(t,\nu)$ being $P/2$ sparse is not possible for moderately sized P, for example for $P > 15$. If we further reduce the number of active boxes, then compressive sensing algorithms such as Basis Pursuit and Orthogonal Matching Pursuit become available, as is discussed in the following section.

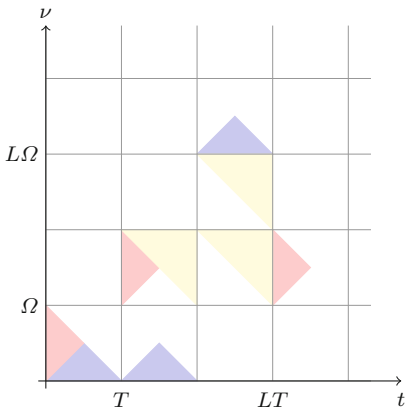


Fig. 6 For S the union of the colored sets, $OPW(S)$ is identifiable even though $7 > 3$ boxes are active, implying that S cannot be rectified with $P = 3$ and $T = 1$ (see Section "Identification of operator Paley-Wiener spaces by periodically weighted delta-trains"). Recovering η from Hg requires solving three systems of linear equations, one to recover η on the yellow support set, one to recover η on the red support set, and one to recover η on the blue support set. The reconstruction formula (21) does not apply for this set S.

Finite dimensional operator identification and compressive sensing

Operator sampling in the finite dimensional setting translates into the following matrix probing problem [5, 7, 49]. For a class of matrices $\mathcal{M} \in \mathbb{C}^{P\times P}$, find $c \in \mathbb{C}^P$ so that we can recover $M \in \mathcal{M}$ from Mc (Figure 7).

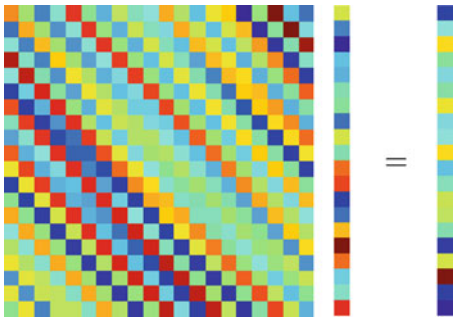


Fig. 7 The matrix probing problem: find c so that the map $\mathcal{M} \longrightarrow \mathbb{C}^P$, $M \mapsto Mc$ is injective and therefore invertible.

The classes of operator considered here are of the form $M_{\boldsymbol{\eta}} = \sum_\lambda \boldsymbol{\eta}_\lambda B_\lambda$ with $B_\lambda = B_{p,q} = \mathcal{T}^p\mathcal{M}^q$, and the matrix identification problem is reduced to solving

$$\boldsymbol{Z} = M_{\boldsymbol{\eta}}c = \sum_{p,q=0}^{P-1} \boldsymbol{\eta}_{p,q}\big(\mathcal{T}^p\mathcal{M}^q c\big) = G(c)\boldsymbol{\eta}, \tag{30}$$

where c is chosen appropriately; this is just (29) with the dependence on (t,ν) removed.

If $\boldsymbol{\eta}$ is assumed to be k-sparse, (Figure 8) we arrive at the classical compressive sensing problem with measurement matrix $G(c) \in \mathbb{C}^{P\times P^2}$ which depends on $c = (c_0, c_1, \ldots, c_{P-1})$. To achieve recovery guarantees for Basis Pursuit and Orthogonal Matching Pursuit, averaging arguments have to be used that yield results on the expected qualities of $G(c)$. This problem was discussed in [40, 49, 50] as well as, in slightly different terms, in [1, 17] (Figure 9). The strongest results were achieved in [25] by estimating Restricted Isometry Constants for c being a Steinhaus sequence. These results show that with high probability, $G(c)$ has the property that Basis Pursuit recovers $\boldsymbol{\eta}$ from $G(c)\boldsymbol{\eta}$ for every k sparse $\boldsymbol{\eta}$ as long as $k \le CP/\log^2 P$ for some universal constant C.

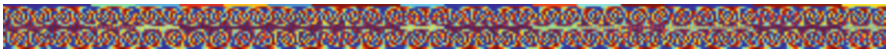


Fig. 8 Time-frequency structured measurement matrix $G(c)$ with c randomly chosen.

Stochastic operators and channel estimation

It is common that models of wireless channels and radar environments take the stochastic nature of the medium into account. In such models, the spreading function $\eta(t,\nu)$ (and therefore the operator's kernel and Kohn–Nirenberg symbol) are random processes, and the operator is split into the sum of its deterministic portion, representing the mean behavior of the channel, and its zero-mean stochastic portion that represents the noise and the environment.

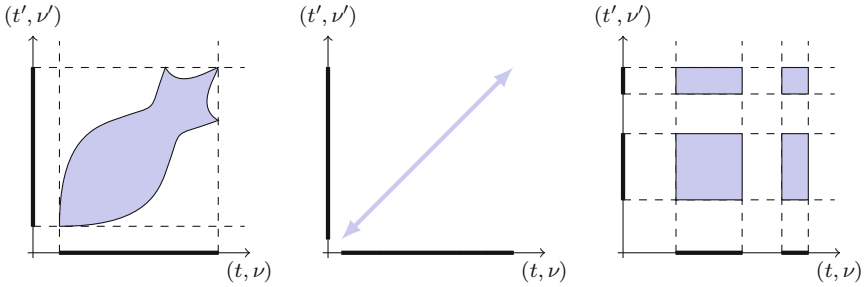


Fig. 9 Support sets of autocorrelation functions, the general case, the WSSUS case, and the tensor case.

The detailed analysis of the stochastic case was carried out in [47, 48]. One of the foci of these works lies in the goal of determining the second-order statistics of the (zero mean) stochastic process $\eta(\tau,\nu)$, that is, its so-called covariance function $R(\tau,\nu,\tau',\nu') = \mathbb{E}\{\eta(\tau,\nu)\,\overline{\eta(\tau',\nu')}\}$. In [47, 48], it was shown that a necessary but not sufficient condition for the identifiability of $R\eta(\tau,\nu,\tau',\nu')$ from the output covariance $A(t,t') = \mathbb{E}\{Hg(t)\,\overline{Hg(t')}\}$ is that $R(\tau,\nu,\tau',\nu')$ is supported on a bounded set of 4-dimensional volume less than or equal to one. Unfortunately, for some sets $S \subseteq \mathbb{R}^4$ of arbitrary small measure, the respective stochastic operator Paley–Wiener space $StOPW(S)$ of operators with $R\eta$ supported on S is not identifiable; this is a striking difference to the deterministic setup where the geometry of S does not play a role at all.

In [67, 68] the special case of *wide-sense stationary operators with uncorrelated scattering*, or WSSUS operators is considered. These operators are characterized by the property that

$$R\eta(t,\nu,t',\nu') = C_\eta(t,\nu)\,\delta(t-t')\,\delta(\nu-\nu').$$

The function $C_\eta(t,\nu)$ is then called the *scattering function* of H. Our results on the identifiability of stochastic operator classes allowed for the construction of two estimators for scattering functions [67, 68]. The estimator given in [67] is applicable, whenever the scattering function of H has bounded support. Note that the autocorrelation of a WSSUS operator is supported on a two-dimensional plane in $\mathbb{R}^4$ which

therefore has 4D volume 0, a fact that allows us to lift commonly assumed restrictions on the size of the 2D area of the support of the scattering function.

For details, formal definitions of identifiability and detailed statements of results we refer to the papers [47, 48, 67, 68].

Appendix: Proofs of Theorems

Proof of Lemma 2

In order to see how the time-frequency shifts of c arise, we will briefly outline the calculation that leads to (16). It can be seen by direct calculation using the representation given by (7), that if $g = \sum_n \delta_{nTP}$ then $\langle Hg, s\rangle = \langle \eta_H, Z_{TP}s\rangle$ for all $s \in \mathcal{S}(\mathbb{R})$ where the bracket on the left is the L^2 inner product on $\mathbb{R}$ and that on the right the L^2 inner product on the rectangle $[0,TP]\times[0,1/(TP)]$. Periodizing the integral on the left gives

$$\langle \eta_H, Z_{TP}s\rangle = \int_0^{1/(TP)} \int_0^{TP} \sum_k \sum_m \eta_H(t+kTP, \nu+m/(TP)) \\ e^{-2\pi i\nu kTP}\overline{Z_{TP}s(t,\nu)}\, dt\, d\nu.$$

Since this holds for every $s \in \mathcal{S}(\mathbb{R})$, we conclude that

$$(Z_{TP}\circ H)g(t,\nu) \\ = 1/(TP) \sum_k \sum_m \eta_H(t+kTP, \nu+m/(TP))\, e^{-2\pi i\nu kTP}.$$

Given $g = \sum_{n\in\mathbb{Z}} c_n\, \delta_{nT}$, for a period-$P$ sequence $c = (c_n)$, and letting $n = mP - q$ for $m \in \mathbb{Z}$ and $0 \le q < P$, we obtain

$$g = \sum c_n\, \delta_{nT} = \sum_{q=0}^{P-1} \sum_{m\in\mathbb{Z}} c_{mP-q}\, \delta_{mPT-qT} \\ = \sum_{q=0}^{P-1} c_{-q}\mathcal{T}_{-qT}\left(\sum_{m\in\mathbb{Z}} \delta_{mPT}\right).$$

Since for $\alpha \in \mathbb{R}$, the spreading function of $H\circ\mathcal{T}_\alpha$ is $\eta_H(t-\alpha,\nu)\, e^{2\pi i\nu\alpha}$, we arrive at

$$(Z_{TP}\circ H)g(t,\nu) \\ = 1/(TP) \sum_{q=0}^{P-1} c_{-q} \sum_k \sum_m \eta_H(t+kTP+qT, \nu+m/(TP)) \\ e^{-2\pi i(\nu+m/(TP))qT}\, e^{-2\pi i\nu kTP}. \tag{31}$$

Letting $m = jP + \ell$ for $j \in \mathbb{Z}$ and $0 \le \ell < P$, we obtain

$$\begin{aligned}
&(Z_{TP} \circ H) g(t, \nu) \\
&\quad = 1/(TP) \sum_{q=0}^{P-1} c_{-q} \sum_k \sum_j \sum_{\ell=0}^{P-1} \eta_H(t + kTP + qT, \nu + j/T + \ell/(TP)) \\
&\qquad\qquad e^{-2\pi i \nu q T} e^{-2\pi i \ell q/P} e^{-2\pi i \nu k T P} \\
&\quad = 1/(TP) \sum_{q=0}^{P-1} \sum_{\ell=0}^{P-1} \left(c_{-q} e^{-2\pi i \ell q/P}\right) e^{-2\pi i \nu q T} \eta_H^Q(t + Tq, \nu + \ell/TP).
\end{aligned}$$

Finally, replacing t by $t + pT$ for $p = 0, 1, \ldots, P-1$, and changing indices by replacing q by $q - p$, we obtain

$$\begin{aligned}
&(Z_{TP} \circ H) g(t + pT, \nu) \\
&= 1/(TP) \sum_{q=0}^{P-1} \sum_{\ell=0}^{P-1} \left(c_{-q} e^{-2\pi i \ell q/P}\right) e^{-2\pi i \nu q T} \eta_H^Q(t + (q+p)T, \nu + \ell/TP) \\
&= 1/(TP) \sum_{q=0}^{P-1} \sum_{\ell=0}^{P-1} \left(c_{-(q-p)} e^{-2\pi i \ell (q-p)/P}\right) \\
&\qquad\qquad e^{-2\pi i \nu (q-p) T} \eta_H^Q(t + qT, \nu + \ell/TP).
\end{aligned}$$

The observation that $(\mathcal{T}^q \mathcal{M}^m c)_p = c_{p-q} e^{2\pi i m(p-q)/P}$ completes the proof.

Proof of Theorem 3

To see why this is true, define $\mu(M)$ to be the number of diagonals of M whose product is a multiple of p_M, and proceed by induction on the size of the matrix M. If M is 1×1, then the result is obvious. Suppose that M is $n \times n$ and that it is described by the vector $\ell = (\ell_0, \ldots, \ell_{P-1})$. Assuming without loss of generality that the variable of smallest index in p_M with a nonzero exponent is c_0, there is a row of M in which the variable c_0 appears ℓ_j times for some index j. Choose one of these terms and delete the row and column in which it appears. Call the remaining matrix M'. The vector ℓ describing M' is $(\ell_0, \ldots, \ell_{j-1}, \ell_j - 1, \ell_{j+1}, \ldots, \ell_{P-1})$, and is independent of which term was chosen from the given row to form M'. By the construction of the LI monomial, $p_M = c_0 p_{M'}$ and by the induction hypothesis

$$\mu(M') = \ell_0! \cdots \ell_{j-1}! (\ell_j - 1)! \ell_{j+1}! \cdots \ell_{P-1}!.$$

Since there are ℓ_j ways to choose a term from the given row to produce M' we have that

$$\mu(M) = \ell_j \mu(M') = \ell_0! \cdots \ell_{j-1}! \ell_j (\ell_j - 1)! \ell_{j+1}! \cdots \ell_{P-1}! = \prod_{\kappa=0}^{P-1} \ell_\kappa!$$

which was to be proved.

Since each term $a_\sigma C^\sigma$ in (24) is made up of a sum of precisely this many terms, it follows that exactly one of these terms is a multiple of the LI monomial. Alternatively, we can think of the LI monomial as the one corresponding to the $\sigma \in S_P/\Gamma$ that minimizes the functional $\Lambda_0(C^\sigma) = \sum_{i=0}^{L-1} i^2 H(\alpha_i)$ where α_i is the exponent of c_i in C^σ and where $H(\alpha_i) = 0$ if $\alpha_i = 0$ and 1 otherwise.

Because by Chebotarev's Theorem, $a_\sigma \neq 0$ for all σ the proof works for any square submatrix M, no matter what size. This gives us Theorem 3.

Proof of Theorem 4

We first need to assert the existence of a cyclical renumbering of the variables such that with respect to the new trivial partition $A' = (A'_\kappa)_{\kappa=0}^{P-1}$, the CI monomial is given by

$$C^I = \prod_{\kappa=0}^{P-1} \prod_{j \in A'_\kappa} c_{j-\kappa}$$

in other words, if $j \in A'_\kappa$ then $0 \leq j - \kappa < P$. Note first that since $\min(A'_\kappa) = \sum_{i=0}^{\kappa-1} \ell'_i$ for all κ, $j \in A'_\kappa$ implies that $j \geq \sum_{i=0}^{\kappa-1} \ell'_i$. Therefore, it will suffice to find a $0 \leq \gamma < P$ such that for all κ, $\sum_{i=0}^{\kappa-1} \ell'_i - \kappa \geq 0$ so that $j - \kappa \geq \sum_{i=0}^{\kappa-1} \ell'_i - \kappa \geq 0$.

Let $0 \leq \gamma < P$ be such that the quantity $\sum_{i=0}^{\gamma-1} \ell_i - \gamma$ is minimized, let

$$\ell' = (\ell'_i)_{i=0}^{L-1} = (\ell_{(i+\gamma) mod\ P})_{i=0}^{P-1},$$

and let $A' = (A'_\kappa)_{\kappa=0}^{P-1}$ be the corresponding trivial partition. Now fix κ and assume that $\kappa + \gamma \leq P$. Then

$$\begin{aligned}
\sum_{i=0}^{\kappa-1} \ell'_i - \kappa &= \sum_{i=0}^{\kappa-1} \ell_{(i+\gamma)} - \kappa \\
&= \left(\sum_{i=0}^{\kappa+\gamma-1} \ell_i - (\kappa+\gamma) \right) - \left(\sum_{i=0}^{\gamma-1} \ell_i - \gamma \right) \\
&\geq 0
\end{aligned}$$

since the second term in the difference is minimal. If $\kappa + \gamma \geq P + 1$, then remembering that $\sum_{i=0}^{P-1} \ell_i = L$

$$\begin{aligned}
\sum_{i=0}^{\kappa-1} \ell'_i - \kappa &= \sum_{i=0}^{\kappa-1} \ell_{(i+\gamma) mod\ P} - \kappa \\
&= \sum_{i=\gamma}^{P-1} \ell_i + \sum_{i=0}^{\kappa+\gamma-P-1} \ell_i - \kappa
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=0}^{P-1} \ell_i - \sum_{i=0}^{\gamma-1} \ell_i + \sum_{i=0}^{\kappa+\gamma-P-1} \ell_i - \kappa \\
&= \left(\sum_{i=0}^{(\kappa+\gamma-P)-1} \ell_i - (\kappa+\gamma-P) \right) - \left(\sum_{i=0}^{\gamma-1} \ell_i - \gamma \right) \\
&\geq 0.
\end{aligned}
$$

In order to complete the proof, we must show that $\Lambda(C^\sigma) \geq \Lambda(C^I)$ for all $\sigma \in S_P/\Gamma$ with equality holding if and only if σ is trivial. This will follow by direct calculation together with the following lemma which follows from a classical result on rearrangements of series [14, Theorems 368, 369]. This result is Lemma 3.3 in [32].

First, however, we adopt the following notation. For $0 \leq n < P$, let $b_n = \kappa$ if $n \in A_\kappa$. With this notation, given $\sigma \in S_P/\Gamma$,

$$C^\sigma = \prod_{n=0}^{P-1} c_{(\sigma(n)-b_n) \bmod P}$$

and under the above assumptions,

$$C^I = \prod_{n=0}^{P-1} c_{(n-b_n)}.$$

Moreover,

$$
\begin{aligned}
\Lambda(C^\sigma) &= \sum_{i=0}^{P-1} i^2 \alpha_i \\
&= \sum_{i=0}^{P-1} i^2 \left(\#\{n\colon (\sigma(n)-b_n) \bmod P = i\} \right) \\
&= \sum_{i=0}^{P-1} \left((\sigma(n)-b_n) \bmod P \right)^2.
\end{aligned}
$$

Lemma 3. *Given two finite sequences of real numbers (α_n) and (β_n) defined up to rearrangement, the sum*

$$\sum_n \alpha_n \beta_n$$

is maximized when α and β are both monotonically increasing or monotonically decreasing. Moreover, if for every rearrangement α' of α,

$$\sum_n \alpha'_n \beta_n \leq \sum_n \alpha_n \beta_n$$

then α and β are similarly ordered*, that is, for every j, k,*

$$(\alpha_j - \alpha_k)(\beta_j - \beta_k) \geq 0.$$

In particular, for every $\sigma \in S_P$,

$$\sum_{n=0}^{P-1} n\, b_n \geq \sum_{n=0}^{P-1} \sigma(n)\, b_n$$

with equality holding if and only if σ *is trivial.*

Proof. The first part of the lemma is simply a restatement of Theorems 368 and 369 of [14]. To prove the second part, note first that b_n is a non-decreasing sequence and in particular is constant on each A_κ. Theorem 368 in [14] states that a sum of the form $\sum_{n=0}^{P-1} \sigma(n)\, b_n$ is maximized when $\sigma(n)$ is monotonically increasing, which proves the given inequality. Since b_n is constant on each A_κ, it follows that if σ is trivial, then we have equality.

If σ is not trivial, then we will show that the sequences $\sigma(n)$ and b_n are not similarly ordered. Letting κ be the minimal index such that A_κ is not left invariant by σ, there exists $m \in A_\kappa$ such that $\sigma(m) \in A_\mu$ for some $\mu > \kappa$, and for some $\lambda > \kappa$ there exists $k \in A_\lambda$ such that $\sigma(k) \in A_\kappa$. Therefore, $b_m = \kappa < \lambda = b_k$ but since $\mu > \kappa$, $\sigma(m) > \sigma(k)$, and so $\sigma(n)$ and b_n are not similarly ordered.

In order to complete the proof, define $\mathcal{C}_1, \mathcal{C}_2 \subseteq \{0, \ldots, P-1\}$ by $n \in \mathcal{C}_1$ if $0 \leq \sigma(n) - b_n < P$, and $n \in \mathcal{C}_2$ if $-P+1 \geq \sigma(n) - b_n < 0$ (note that always $|\sigma(n) - b_n| < P$) so that when $n \in C_2$, $(\sigma(n) - b_n)\ mod\ P = \sigma(n) - b_n + P$. Let $\sigma'(n) = \sigma(n)$ if $n \in \mathcal{C}_1$ and $\sigma(n) + P$ if $n \in \mathcal{C}_2$, and let $(a_n)_{n=0}^{P-1}$ be an increasing sequence enumerating the set $\sigma(\mathcal{C}_1) \cup (\sigma(\mathcal{C}_2) + P)$. Therefore,

$$\begin{aligned}
\Lambda(C^\sigma) - \Lambda(C^I) &= \sum_{n=0}^{P-1} (\sigma'(n) - b_n)^2 - \sum_{n=0}^{P-1} (n - b_n)^2 \\
&= \left[\sum_{n=0}^{P-1} (\sigma'(n) - b_n)^2 - \sum_{n=0}^{P-1} (a_n - b_n)^2 \right] \\
&\qquad + \left[\sum_{n=0}^{P-1} (a_n - b_n)^2 - \sum_{n=0}^{P-1} (n - b_n)^2 \right] \\
&= 2 \left[\sum_{n=0}^{P-1} a_n b_n - \sigma'(n) b_n \right] + \left[\sum_{n=0}^{P-1} (a_n - b_n)^2 - (n - b_n)^2 \right] \\
&= I + II.
\end{aligned}$$

Since a_n is increasing, $I \geq 0$ by Lemma 3, and since $a_n \geq n$ for all n, $(a_n - b_n) \geq (n - b_n) \geq 0$ so that $(a_n - b_n)^2 \geq (n - b_n)^2$ and hence $II \geq 0$. It remains to show that equality holds only if σ is trivial. If $\Lambda(C^\sigma) = \Lambda(C^I)$, then $I = II = 0$. Since $II = 0$, $\mathcal{C}_2 = \emptyset$ for if $a_n \in \sigma(\mathcal{C}_2) + P$ then $a_n > n$ and we would have $II > 0$. Since $\mathcal{C}_2 = \emptyset$, $\sigma'(n) = \sigma(n)$ so that

$$\begin{aligned} 0 &= \Lambda(C^{\sigma}) - \Lambda(C^{I}) \\ &= \sum_{n=0}^{P-1} (\sigma(n) - b_n)^2 - \sum_{n=0}^{P-1} (n - b_n)^2 \\ &= 2 \sum_{n=0}^{P-1} (n b_n - \sigma(n) b_n) \end{aligned}$$

which by Lemma 3 implies that σ is trivial. The proof is complete.

References

1. L. Applebaum, S.D. Howard, S. Searle, R. Calderbank, Chirp sensing codes: deterministic compressed sensing measurements for fast recovery. Appl. Comput. Harmon. Anal. **26**(2), 283–290 (2008)
2. W.U. Bajwa, K. Gedalyahu, Y.C. Eldar, Identification of parametric underspread linear systems and super-resolution radar. IEEE Trans. Signal Process. **59**(6), 2548–2561 (2011)
3. W.U. Bajwa, K. Gedalyahu, Y.C. Eldar, On the identification of parametric underspread linear systems, in *EEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2011), pp. 4088–4091
4. B.L. Basore, Noise-like signals and their detection by correlation. Thesis (Sc. D.)-Massachusetts Institute of Technology (1952). http://hdl.handle.net/1721.1/34322
5. R. Belanger-Rioux, L. Demanet, Compressed absorbing boundary conditions via matrix probing. (2014). arXiv.org/abs/1401.4421
6. P.A. Bello. Measurement of random time-variant linear channels. IEEE Trans. Commun. **15**, 469–475 (1969)
7. J. Chiu, L. Demanet, Matrix probing and its conditioning. SIAM J. Numer. Anal. **50**(1), 171–193 (2012)
8. D. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. Proc. Natl. Acad. Sci. **100**(5), 2197–2202 (2003)
9. H.G. Feichtinger, W. Kozek, F. Luef, Gabor analysis over finite abelian groups. Appl. Comput. Harmon. Anal. **26**(2), 230–248 (2009)
10. P.E. Green, Correlation detection using stored signals. Thesis (Sc. D.)-Massachusetts Institute of Technology (1953). http://hdl.handle.net/1721.1/34880
11. P.E. Green, Radar measurements of target scattering properties, in *Radar Astronomy*, ed. by J.V. Evans, T. Hagfors (McGraw-Hill, New York, 1968), pp. 1–78
12. P. Green, Early Spread-Spectrum and automatic equalization – NOMAC and Rake. in *IEEE GLOBECOM 2008: Global Telecommunications Conference, 2008* (2008), pp. 1–5
13. K. Gröchenig, *Foundations of Time-Frequency Analysis*. Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, 2001)
14. G.L. Hardy, J.E. Littlewood, G. Pólya, *Inequalities*, 2nd edn. (Cambridge University Press, Cambridge, 1952)
15. R. Heckel, H. Bölcskei, Identification of sparse linear operators. IEEE Trans. Inf. Theory **59**(12), 7985–8000 (2013)
16. C.W. Helstrom, *Statistical Theory of Signal Detection* (Pergamon Press, London, 1960)
17. M.A. Herman, T. Strohmer, High-resolution radar via compressed sensing. *IEEE Trans. Signal Process.* **57**(6), 2275–2284 (2009)
18. Y.M. Hong, G.E.Pfander, Irregular and multi-channel sampling of operators. Appl. Comput. Harmon. Anal. **29**(2), 214–231 (2010)
19. T. Kailath, Sampling models for linear time-variant filters. Technical Report 352, Massachusetts Institute of Technology, Research Laboratory of Electronics (1959)

20. T. Kailath, Correlation detection of signals perturbed by a random channel. IRE Trans. Inf. Theory **6**(3), 361–366 (1960)
21. T. Kailath, Communication via randomly varying channels. Thesis (Sc. D.)–Massachusetts Institute of Technology, Dept. of Electrical Engineering (1961). http://hdl.handle.net/1721.1/11319
22. T. Kailath, Measurements on time–variant communication channels. IEEE Trans. Inf. Theory **8**(5), 229–236 (1962)
23. T. Kailath, Time–variant communication channels. IEEE Trans. Inf. Theory. Inf. Theory. Prog. Rep. 1960–1963 **9**, 233–237 (1963)
24. W. Kozek, G.E. Pfander, Identification of operators with bandlimited symbols. SIAM J. Math. Anal. **37**(3), 867–888 (2006)
25. F. Krahmer, S. Mendelson, H. Rauhut, Suprema of chaos processes and the restricted isometry property. Commun. Pure Appl. Math. **67**(11), 1877–1904 (2014)
26. F. Krahmer, G.E. Pfander, Local sampling and approximation of operators with bandlimited Kohn-Nirenberg symbols. Construct. Approx. **39**(3), 541–572 (2014)
27. F. Krahmer, G.E. Pfander, P. Rashkov, Uncertainty in time-frequency representations on finite abelian groups and applications. Appl. Comput. Harmon. Anal. **25**(2), 209–225 (2008)
28. H.J. Landau, H.O. Pollak, Prolate spheroidal wave functions, Fourier analysis and uncertainty– I. Bell Syst. Tech. J. **40**, 43–64 (1961)
29. H.J. Landau, D. Slepian, H.O. Pollak, Prolate spheroidal wave functions, fourier analysis and uncertainty–II. Bell Syst. Tech. J. **40**(1), 65–84 (1961)
30. J. Lawrence, G.E. Pfander, D.F. Walnut, Linear independence of Gabor systems in finite dimensional vector spaces. J. Fourier Anal. Appl. **11**(6), 715–726 (2005)
31. Y.I. Lyubarskii, Frames in the Bargmann space of entire functions. Adv. Sov. Math. **429**, 107–113 (1992)
32. R.-D. Malikiosis, A note on Gabor frames in finite dimensions. Appl. Comput. Harmon. Anal. **38**(2), 318–330 (2015)
33. H.K. Markey, G. Antheil, *Secret Communication System, US 2292387* (11 August 1942)
34. B.J. Pankowski, Multiplexing a radio teletype system using a random carrier and correlation detection. Thesis (Sc. M.)-Massachusetts Institute of Technology (1952)
35. A. Papoulis, Generalized sampling expansion. IEEE Trans. Circuits Syst. **24**(11), 652–654 (1977)
36. G.E. Pfander, Measurement of time-varying multiple-input multiple-output channels. Appl. Comput. Harmon. Anal. **24**(3), 393–401 (2008)
37. G.E. Pfander, On the invertibility of rectangular bi-infinite matrices and applications in time-frequency analysis, Linear Algebra Appl. **429**(1), 331–345 (2008)
38. G.E. Pfander, Sampling of operators. http://arxiv.org/abs/1010.6165, preprint (2010)
39. G.E. Pfander, Gabor frames in finite dimensions. in *Finite Frames: Theory and Applications*, ed. by P.G. Casazza, G. Kutyniok (Springer, New York, 2013)
40. G.E. Pfander, H. Rauhut, Sparsity in time–frequency representations. J. Fourier Anal. Appl. **16**(2), 233–260 (2010)
41. G.E. Pfander, D. Walnut, Measurement of time–variant channels. IEEE Trans. Inf. Theory **52**(11), 4808–4820 (2006)
42. G.E. Pfander, D. Walnut, Operator identifcation and Feichtinger's algebra. Sampling Theory Signal Image Process. **5**(2), 151–168 (2006)
43. G.E. Pfander, D. Walnut, On the sampling of functions and operators with an application to Multiple–Input Multiple–Output channel identification, in *Proceedings SPIE Vol. 6701, Wavelets XII*, ed. by D. Van De Ville, V.K. Goyal, M. Papadakis (2007), pp. 67010T-1–67010T-14
44. G.E. Pfander, D. Walnut, Operator identification and sampling, in *Proceedings of the Conference on Sampling Theory and Applications* (Aix-Marseilles Université, Marseilles, 2009)
45. G.E. Pfander, D. Walnut, Sparse finite Gabor frames for operator sampling. *Proceedings of the Conference on Sampling Theory and Applications* (Jacobs University, Bremen, 2013)
46. G.E. Pfander, D. Walnut, Sampling and reconstruction of operators. IEEE Trans. Inf. Theory (2014, submitted). arXiv.org/abs/1503.00628

47. G.E. Pfander, P. Zheltov, Identification of stochastic operators. Appl. Comput. Harmon. Anal. **36**(2), 256–279 (2014)
48. G.E. Pfander, P. Zheltov, Sampling of stochastic operators. IEEE Trans. Inf. Theory **60**(4), 2359–2372 (2014)
49. G.E. Pfander, H. Rauhut, J. Tanner, Identification of matrices having a sparse representation. IEEE Trans. Signal Process. **56**(11), 5376–5388 (2008)
50. G.E. Pfander, H. Rauhut, J.A. Tropp, The restricted isometry property for time-frequency structured random matrices. Probab. Theory Relat. Fields **156**(3–4), 707–737 (2013)
51. R.I. Pickholtz, D.L. Schilling, L.B. Milstein, Theory of spread-spectrum communications – a tutorial. IEEE Trans. Commun. **30**, 855–884 (1982)
52. R. Price, Statistical theory applied to communication through multipath disturbances. RLE Tech. Rpt. No. 266, Lincoln Laboratory Tech. Rpt. No. 34 (1953)
53. R. Price, Optimum detection of random signals in noise, with applications to scatter-multipath communication, I. IRE Trans. Inf. Theory **IT-2**, 125–135 (1956)
54. R. Price, Detectors for Radar astronomy, in *Radar Astronomy*, ed. by J.V. Evans, T. Hagfors (McGraw-Hill, New York, 1968), pp. 547–614
55. R. Price, Further notes and anecdotes on spread-spectrum origins. IEEE Trans. Commun. **31**, 85–97 (1982)
56. R. Price, P. Green, A communication technique for multipath channels. Proc. IRE **46**, 555–570 (1958)
57. R.A. Scholtz, The origins of spread-spectrum communications., IEEE Trans. Commun. **30**, 822–854 (1982)
58. R.A. Scholtz, Notes on spread-spectrum history. IEEE Trans. Commun. **31**, 82–84 (1983)
59. K. Seip, Density theorems for sampling and interpolation in the Bargmann-Fock space. I. J. Reine Angew. Math. **429**, 91–106 (1992)
60. K. Seip, R. Wallstén, Density theorems for sampling and interpolation in the Bargmann-Fock space. II. J. Reine Angew. Math. **429**, 107–113 (1992)
61. C. Shannon, Communication in the presence of noise. Proc. IRE **37**(1), 10–21 (1949)
62. P. Stevenhagen, H.W. Lenstra Jr., Chebotarëv and his density theorem. Math. Intell. **18**(2), 26–37 (1996)
63. G. Turin, Introduction to spread-spectrum antimultipath techniques and their application to urban digital radio. Proc. IEEE **68**, 328–353 (1980)
64. W. Ward, The NOMAC and rake systems. Lincoln Lab. J. **5**, 351–366 (1992)
65. N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series, with Engineering Applications* (M.I.T. Press, Cambridge, 1949)
66. L.A. Zadeh, Frequency analysis of variable networks. Proc. IRE **67**, 291–299 (1950)
67. P. Zheltov, G.E. Pfander, Estimation of overspread scattering functions. IEEE Trans. Signal Process. **63**(10), 2451–2463 (2015)
68. P. Zheltov, G.E. Pfander, O. Oktay, Reconstruction of the scattering function of overspread radar targets. IET Signal Process. **8**(9), 1018–1024 (2014)

Part XVI
Spectral Analysis and Correlations

Part XVI is concerned with spectral analysis broadly construed. Spectral analysis and the study of signals/functions correlations are very intertwined topics and are partially rooted in harmonic analysis with pioneering contributions from N. Wiener, and A. Khinchin. In addition, these ever-present topics are also part of an important class of tools used in a number of areas of sciences and engineering. While some of the chapters in this part can be directly related to either spectral analysis or correlation analysis, others are only loosely related. But in either case, these two topics lie in the background of most of the following chapters. In the first chapter, HAMED FIROUZI, DENNIS WEI, and ALFRED O. HERO III consider a spectral analysis of the correlation of stationary multivariate Gaussian time series. This problem focuses on identifying those time series that are highly correlated with a specified number of other time series. They use an independent correlation analysis method in the Fourier domain. This allows them to handle the computational complexity usually associated with the analysis of high-dimensional time series.

In the second chapter of this part, R. A. BAILEY , PERSI DIACONIS , DANIEL N. ROCKMORE, and CHRIS ROWLEY introduce a spectral analysis approach that generalizes the classical analysis of variance (ANOVA)-based techniques used for studying data from designed experiments. Designed experiments are widely used in many fields, and this chapter offers a very self-contained introduction to the spectral analysis of the data obtained from such experiments. In particular, using the representation theory of certain groups related to the designed experiments, the chapter considers in details various examples.

In the third chapter, GAURAV THAKUR surveys recent developments related to the Synchrosqueezing transform. This is a nonstationary time-frequency method used to analyze complex signals in terms of their time-varying oscillatory components. Moreover, the Synchrosqueezing transform can be viewed as a sparse and invertible time-frequency reassignment technique that leads to the recovery of the signal. The chapter focusses on the theory and the stability properties of Synchrosqueezing, while indicating some applications in areas such as cardiology, climate science, and economics.

In the fourth chapter of this part, PABLO SPRECHMANN, ALEX M. BRONSTEIN, and GUILLERMO SAPIRO start with an overview of nonnegative matrix factorization (NMF) algorithms for solving source separation problems. These (difficult) problems with a number of applications in areas such as mobile telephony are prevalent in signal (audio) processing and often involve extracting or enhancing an audio signal recorded in a noisy environment. The main contribution is the introduction of an alternative supervised training technique in the NMF algorithms to solve the aforementioned problem. This new methodology is cast as an optimization problem solved using stochastic gradient descent.

Spectral Correlation Hub Screening of Multivariate Time Series

Hamed Firouzi, Dennis Wei, and Alfred O. Hero III

Abstract This chapter discusses correlation analysis of stationary multivariate Gaussian time series in the spectral or Fourier domain. The goal is to identify the hub time series, i.e., those that are highly correlated with a specified number of other time series. We show that Fourier components of the time series at different frequencies are asymptotically statistically independent. This property permits independent correlation analysis at each frequency, alleviating the computational and statistical challenges of high-dimensional time series. To detect correlation hubs at each frequency, an existing correlation screening method is extended to the complex numbers to accommodate complex-valued Fourier components. We characterize the number of hub discoveries at specified correlation and degree thresholds in the regime of increasing dimension and fixed sample size. The theory specifies appropriate thresholds to apply to sample correlation matrices to detect hubs and also allows statistical significance to be attributed to hub discoveries. Numerical results illustrate the accuracy of the theory and the usefulness of the proposed spectral framework.

Key words: Complex-valued correlation screening, Spectral correlation analysis, Gaussian stationary processes, Hub screening, Correlation graph, Correlation network, Spatiotemporal analysis of multivariate time series, High-dimensional data analysis

H. Firouzi • D. Wei • A.O. Hero III (✉)
Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI, USA
e-mail: firouzi@umich.edu; dlwei@eecs.umich.edu; hero@eecs.umich.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_13

Introduction

Correlation analysis of multivariate time series is important in many applications such as wireless sensor networks, computer networks, neuroimaging, and finance [1–5]. This chapter focuses on the problem of detecting *hub* time series, ones that have a high degree of interaction with other time series as measured by correlation or partial correlation. Detection of hubs can lead to reduced computational and/or sampling costs. For example, in wireless sensor networks, the identification of hub nodes can be useful for reducing power usage and adding or removing sensors from the network [6, 7]. Hub detection can also give new insights about underlying structure in the data set. In neuroimaging for instance, studies have consistently shown the existence of highly connected hubs in brain graphs (connectomes) [8]. In finance, a hub might indicate a vulnerable financial instrument or a sector whose collapse could have a major effect on the market [9].

Correlation analysis becomes challenging for multivariate time series when the dimension p of the time series, i.e., the number of scalar time series, and the number of time samples N are large [4]. A naive approach is to treat the time series as a set of independent samples of a p-dimensional random vector and estimate the associated covariance or correlation matrix, but this approach completely ignores temporal correlations as it only considers dependences at the same time instant and not between different time instants. The work in [10] accounts for temporal correlations by quantifying their effect on convergence rates in covariance and precision matrix estimation; however, only correlations at the same time instant are estimated. A more general approach is to consider all correlations between any two time instants of any two series within a window of $n \leq N$ consecutive samples, where the previous case corresponds to $n = 1$. However, in general this would entail the estimation of an $np \times np$ correlation matrix from a reduced sample of size $m = N/n$, which can be computationally costly as well as statistically problematic.

In this chapter, we propose *spectral* correlation analysis as a method of overcoming the issues discussed above. As before, the time series are divided into m temporal segments of n consecutive samples, but instead of estimating temporal correlations directly, the method performs analysis on the Discrete Fourier Transforms (DFT) of the time series. We prove in Theorem 1 that for stationary, jointly Gaussian time series under the mild condition of absolute summability of the auto- and cross-correlation functions, different Fourier components (frequencies) become asymptotically independent of each other as the DFT length n increases. This property of stationary Gaussian processes allows us to focus on the $p \times p$ correlations at each frequency separately without having to consider correlations between different frequencies. Moreover, spectral analysis isolates correlations at specific frequencies or timescales, potentially leading to greater insight. To make aggregate inferences based on all frequencies, straightforward procedures for multiple inference can be used as described in Section "Application to Spectral Screening of Multivariate Gaussian Time Series".

The spectral approach reduces the detection of hub time series to the independent detection of hubs at each frequency. However, in exchange for achieving spectral

resolution, the sample size is reduced by the factor n, from N to $m = N/n$. To confidently detect hubs in this high-dimensional, low-sample regime (large p, small m), as well as to accommodate complex-valued DFTs, we develop a method that we call *complex-valued (partial) correlation screening*. This is a generalization of the correlation and partial correlation screening method of [9, 11, 12] to complex-valued random variables. For each frequency, the method computes the sample (partial) correlation matrix of the DFT components of the p time series. Highly correlated variables (hubs) are then identified by thresholding the sample correlation matrix at a level ρ and screening for rows (or columns) with a specified number δ of nonzero entries.

We characterize the behavior of complex-valued correlation screening in the high-dimensional regime of large p and fixed sample size m. Specifically, Theorem 2 and Corollary 2 give asymptotic expressions in the limit $p \to \infty$ for the mean number of hubs detected at thresholds ρ, δ, and the probability of discovering at least one such hub. Bounds on the rates of convergence are also provided. These results show that the number of hub discoveries undergoes a phase transition as ρ decreases from 1, from almost no discoveries to the maximum number, p. An expression (33) for the critical threshold $\rho_{c,\delta}$ is derived to guide the selection of ρ under different settings of p, m, and δ. Furthermore, given a null hypothesis that the population correlation matrix is sufficiently sparse, the expressions in Corollary 2 become independent of the underlying probability distribution and can thus be easily evaluated. This allows the statistical significance of a hub discovery to be quantified, specifically in the form of a p-value under the null hypothesis. We note that our results on complex-valued correlation screening apply more generally than to spectral correlation analysis and thus may be of independent interest.

The remainder of the chapter is organized as follows. Section "Spectral Representation of Multivariate Time Series" presents notation and definitions for multivariate time series and establishes the asymptotic independence of spectral components. Section "Complex-Valued Correlation Hub Screening" describes complex-valued correlation screening and characterizes its properties in terms of numbers of hub discoveries and phase transitions. Section "Application to Spectral Screening of Multivariate Gaussian Time Series" discusses the application of complex-valued correlation screening to the spectra of multivariate time series. Finally, Section "Experimental Results" illustrates the applicability of the proposed framework through simulation analysis.

Notation

A triplet $(\Omega, \mathscr{F}, \mathbb{P})$ represents a probability space with sample space Ω, σ-algebra of events $\mathscr{F}$, and probability measure $\mathbb{P}$. For an event $A \in \mathscr{F}$, $\mathbb{P}(A)$ represents the probability of A. Scalar random variables and their realizations are denoted with upper case and lower case letters, respectively. Random vectors and their realizations are denoted with bold upper case and bold lower case letters. The expectation operator is denoted as $\mathbb{E}$. For a random variable X, the cumulative probability distribution

(cdf) of X is defined as $F_X(x) = \mathbb{P}(X \leq x)$. For an absolutely continuous cdf $F_X(.)$ the probability density function (pdf) is defined as $f_X(x) = dF_X(x)/dx$. The cdf and pdf are defined similarly for random vectors. Moreover, we follow the definitions in [13] for conditional probabilities, conditional expectations, and conditional densities.

For a complex number $z = a + b\sqrt{-1} \in \mathbb{C}$, $\mathrm{Re}(z) = a$ and $\mathrm{Im}(z) = b$ represent the real and imaginary parts of z, respectively. A complex-valued random variable is composed of two real-valued random variables as its real and imaginary parts. A complex-valued Gaussian variable has real and imaginary parts that are Gaussian. A complex-valued (Gaussian) random vector is a vector whose entries are complex-valued (Gaussian) random variables. The covariance of a p-dimensional complex-valued random vector $\mathbf{Y}$ and a q-dimensional complex-valued random vector $\mathbf{Z}$ is a $p \times q$ matrix defined as

$$\mathrm{cov}(\mathbf{Y}, \mathbf{Z}) = \mathbb{E}\left[(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])^H\right],$$

where ^{H} denotes the Hermitian transpose. We write $\mathrm{cov}(\mathbf{Y})$ for $\mathrm{cov}(\mathbf{Y}, \mathbf{Y})$ and $\mathrm{var}(Y) = \mathrm{cov}(Y, Y)$ for the variance of a scalar random variable Y. The correlation coefficient between random variables Y and Z is defined as

$$\mathrm{cor}(Y, Z) = \frac{\mathrm{cov}(Y, Z)}{\sqrt{\mathrm{var}(Y)\mathrm{var}(Z)}}.$$

Matrices are also denoted by bold upper case letters. In most cases the distinction between matrices and random vectors will be clear from the context. For a matrix $\mathbf{A}$ we represent the (i, j)th entry of $\mathbf{A}$ by a_{ij}. Also $\mathbf{D_A}$ represents the diagonal matrix that is obtained by zeroing out all but the diagonal entries of $\mathbf{A}$.

Spectral Representation of Multivariate Time Series

Definitions

Let $\mathbf{X}(k) = [X^{(1)}(k), X^{(2)}(k), \cdots X^{(p)}(k)]$, $k \in \mathbb{Z}$, be a multivariate time series with time index k. We assume that the time series $X^{(1)}, X^{(2)}, \cdots X^{(p)}$ are second-order stationary random processes, i.e.,

$$\mathbb{E}[X^{(i)}(k)] = \mathbb{E}[X^{(i)}(k + \Delta)] \tag{1}$$

and

$$\mathrm{cov}[X^{(i)}(k), X^{(j)}(l)] = \mathrm{cov}[X^{(i)}(k + \Delta), X^{(j)}(l + \Delta)] \tag{2}$$

for any integer time shift Δ.

For $1 \le i \le p$, let $\mathbf{X}^{(i)} = [X^{(i)}(k), \cdots, X^{(i)}(k+n-1)]$ denote any vector of n consecutive samples of time series $X^{(i)}$. The n-point Discrete Fourier Transform (DFT) of $\mathbf{X}^{(i)}$ is denoted by $\mathbf{Y}^{(i)} = [Y^{(i)}(0), \cdots, Y^{(i)}(n-1)]$ and defined by

$$\mathbf{Y}^{(i)} = \mathbf{W}\mathbf{X}^{(i)}, \quad 1 \le i \le p$$

in which $\mathbf{W}$ is the DFT matrix:

$$\mathbf{W} = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & \omega & \cdots & \omega^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n-1} & \cdots & \omega^{(n-1)^2} \end{bmatrix},$$

where $\omega = e^{-2\pi\sqrt{-1}/n}$.

We denote the $n \times n$ population covariance matrix of $\mathbf{X}^{(i)}$ as $\mathbf{C}^{(i,i)} = [c_{kl}^{(i,i)}]_{1 \le k,l \le n}$ and the $n \times n$ population cross-covariance matrix between $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$ as $\mathbf{C}^{(i,j)} = [c_{kl}^{(i,j)}]_{1 \le k,l \le n}$ for $i \ne j$. The translation invariance properties (1) and (2) imply that $\mathbf{C}^{(i,i)}$ and $\mathbf{C}^{(i,j)}$ are Toeplitz matrices. Therefore, $c_{kl}^{(i,i)}$ and $c_{kl}^{(i,j)}$ depend on k and l only through the quantity $k-l$. Representing the (k,l)th entry of a Toeplitz matrix $\mathbf{T}$ by $t(k-l)$, we write

$$c_{kl}^{(i,i)} = c^{(i,i)}(k-l) \quad \text{and} \quad c_{kl}^{(i,j)} = c^{(i,j)}(k-l),$$

where $k-l$ takes values from $1-n$ to $n-1$. In addition, $\mathbf{C}^{(i,i)}$ is symmetric.

Asymptotic Independence of Spectral Components

The following theorem states that for stationary time series, DFT components at different spectral indices (i.e., frequencies) are asymptotically uncorrelated under the condition that the auto-covariance and cross-covariance functions are absolutely summable. This theorem follows directly from the spectral theory of large Toeplitz matrices, see, for example, [14] and [15]. However, for the benefit of the reader we give a self-contained proof of the theorem.

Theorem 1. *Assume* $\lim_{n\to\infty} \sum_{t=0}^{n-1} |c^{(i,j)}(t)| = M^{(i,j)} < \infty$ *for all* $1 \le i,j \le p$. *Define* $\mathrm{err}^{(i,j)}(n) = M^{(i,j)} - \sum_{m'=0}^{n-1} |c^{(i,j)}(m')|$ *and* $\mathrm{avg}^{(i,j)}(n) = \frac{1}{n}\sum_{m'=0}^{n-1} \mathrm{err}^{(i,j)}(m')$. *Then, for* $k \ne l$, *we have*

$$\mathrm{cor}\left(Y^{(i)}(k), Y^{(j)}(l)\right) = O(\max\{1/n, \mathrm{avg}^{(i,j)}(n)\}).$$

In other words $Y^{(i)}(k)$ *and* $Y^{(j)}(l)$ *are asymptotically uncorrelated as* $n \to \infty$.

Proof. Without loss of generality we assume that the time series have zero mean (i.e., $\mathbb{E}[X^{(i)}(k)] = 0, 1 \le i \le p, 0 \le k \le n-1$). We first establish a representation of $\mathbb{E}[Z^{(i)}(k)Z^{(j)}(l)^*]$ for general linear functionals:

$$Z^{(i)}(k) = \sum_{m'=0}^{n-1} g_k(m')X^{(i)}(m'),$$

in which $g_k(.)$ is an arbitrary complex sequence for $0 \le k \le n-1$. We have

$$\begin{aligned}
&\mathbb{E}[Z^{(i)}(k)Z^{(j)}(l)^*] \\
&= \mathbb{E}\left[\left(\sum_{m'=0}^{n-1} g_k(m')X^{(i)}(m')\right)\left(\sum_{n'=0}^{n-1} g_l(n')X^{(j)}(n')\right)^*\right] \\
&= \sum_{m'=0}^{n-1} g_k(m') \sum_{n'=0}^{n-1} g_l(n')^* \mathbb{E}[X^{(i)}(m')X^{(j)}(n')^*] \\
&= \sum_{m'=0}^{n-1} g_k(m') \sum_{n'=0}^{n-1} g_l(n')^* c_{m'n'}^{(i,j)} \qquad (3)
\end{aligned}$$

Now for a Toeplitz matrix $\mathbf{T}$, define the circulant matrix $\mathbf{D_T}$ as

$$\mathbf{D_T} = \begin{bmatrix} t(0) & t(-1)+t(n-1) & \cdots & t(1-n)+t(1) \\ t(1)+t(1-n) & t(0) & \cdots & t(2-n)+t(2) \\ \vdots & \vdots & \ddots & \vdots \\ t(n-2)+t(-2) & t(n-3)+t(-3) & \cdots & t(-1)+t(n-1) \\ t(n-1)+t(-1) & t(n-2)+t(-2) & \cdots & t(0) \end{bmatrix}$$

We can write

$$\mathbf{C}^{(i,j)} = \mathbf{D}_{\mathbf{C}^{(i,j)}} + \mathbf{E}^{(i,j)}$$

for some Toeplitz matrix $\mathbf{E}^{(i,j)}$. Thus, $c^{(i,j)}(m'-n') = d^{(i,j)}(m'-n') + e^{(i,j)}(m'-n')$ where $d^{(i,j)}(m'-n')$ and $e^{(i,j)}(m'-n')$ are the (m',n') entries of $\mathbf{D}_{\mathbf{C}^{(i,j)}}$ and $\mathbf{E}^{(i,j)}$, respectively. Therefore, (3) can be written as

$$\sum_{m'=0}^{n-1} g_k(m') \sum_{n'=0}^{n-1} g_l(n')^* d^{(i,j)}(m'-n') + \sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} g_k(m')g_l(n')^* e^{(i,j)}(m'-n')$$

The first term can be written as:

$$\sum_{m'=0}^{n-1} g_k(m') \left(g_l^* \circledast d^{(i,j)}\right)(m') = \sum_{m'=0}^{n-1} g_k(m') v_l^{(i,j)}(m')$$

where we have recognized $v_l^{(i,j)}(m') = g_l^* \circledast d^{(i,j)}$ as the circular convolution of $g_l^*(.)$ and $d^{(i,j)}(.)$ [16]. Let $G_k(.)$ and $D^{(i,j)}(.)$ be the DFT of $g_k(.)$ and $d^{(i,j)}(.)$, respectively. By Plancherel's theorem [17] we have

$$\begin{aligned}\sum_{m'=0}^{n-1} g_k(m') v_l^{(i,j)}(m') &= \sum_{m'=0}^{n-1} g_k(m') \left(v_l^{(i,j)}(m')^*\right)^* \\ &= \sum_{m'=0}^{n-1} G_k(m') \left(G_l(m') D^{(i,j)}(-m')^*\right)^* \\ &= \sum_{m'=0}^{n-1} G_k(m') G_l(m')^* D^{(i,j)}(-m'). \end{aligned} \tag{4}$$

Now let $g_k(m') = \omega^{km'}/\sqrt{n}$ for $0 \le k, m' \le n-1$. For this choice of $g_k(.)$ we have $G_k(m') = 0$ for all $m' \ne n-k$ and $G_k(n-k) = 1$. Hence, for $k \ne l$ the quantity (4) becomes 0. Therefore, using the representation $\mathbf{E}^{(i,j)} = \mathbf{C}^{(i,j)} - \mathbf{D}_{\mathbf{C}^{(i,j)}}$ we have

$$\begin{aligned}|\mathrm{cov}\left(Y^{(i)}(k), Y^{(j)}(l)\right)| &= |\mathbb{E}[Y^{(i)}(k) Y^{(j)}(l)^*]| \\ &= |\sum_{m'=0}^{n-1} \sum_{n'=0}^{n-1} g_k(m') g_l(n')^* e^{(i,j)}(m'-n')| \\ &\le \frac{1}{n} \sum_{m'=0}^{n-1} \sum_{n'=0}^{n-1} |e^{(i,j)}(m'-n')| \\ &= \frac{2}{n} \sum_{m'=0}^{n-1} m' |c^{(i,j)}(m')|, \end{aligned} \tag{5}$$

in which the last equation is due to the fact that $|c^{(i,j)}(-m')| = |c^{(i,j)}(m')|$.

Now using (4) and (5) we obtain expressions for $\mathrm{var}\left(Y^{(i)}(k)\right)$ and $\mathrm{var}\left(Y^{(j)}(l)\right)$. Letting $j = i$ and $l = k$ in (4) and (5) gives

$$\begin{aligned}&\mathrm{var}\left(Y^{(i)}(k)\right) = \mathrm{cov}\left(Y^{(i)}(k), Y^{(i)}(k)\right) \\ &= \sum_{m'=0}^{n-1} G_k(m') G_k(m')^* D^{(i,i)}(-m') + \sum_{m'=0}^{n-1} \sum_{n'=0}^{n-1} g_k(m') g_k(n')^* e^{(i,i)}(m'-n') \\ &= n . \frac{1}{\sqrt{n}} . \frac{1}{\sqrt{n}} D^{(i,i)}(k) + \sum_{m'=0}^{n-1} \sum_{n'=0}^{n-1} g_k(m') g_k(n')^* e^{(i,i)}(m'-n') \\ &= D^{(i,i)}(k) + \sum_{m'=0}^{n-1} \sum_{n'=0}^{n-1} g_k(m') g_k(n')^* e^{(i,i)}(m'-n'), \end{aligned} \tag{6}$$

in which the magnitude of the summation term is bounded as

$$\begin{aligned} &|\sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} g_k(m')g_k(n')^* e^{(i,i)}(m'-n')| \\ &\le \frac{1}{n}\sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1}|e^{(i,i)}(m'-n')| \\ &= \frac{2}{n}\sum_{m'=0}^{n-1} m'|c^{(i,i)}(m')|. \end{aligned} \tag{7}$$

Similarly,

$$\mathrm{var}\left(Y^{(j)}(l)\right) = D^{(j,j)}(l) + \sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} g_l(m')g_l(n')^* e^{(j,j)}(m'-n'), \tag{8}$$

in which

$$\begin{aligned} &|\sum_{m'=0}^{n-1}\sum_{n'=0}^{n-1} g_l(m')g_l(n')^* e^{(j,j)}(m'-n')| \\ &\le \frac{2}{n}\sum_{m'=0}^{n-1} m'|c^{(j,j)}(m')|. \end{aligned} \tag{9}$$

To complete the proof the following lemma is needed.

Lemma 1. *If* $\{a_{m'}\}_{m'=0}^{\infty}$ *is a sequence of nonnegative numbers such that* $\sum_{m'=0}^{\infty} a_{m'} = M < \infty$. *Define* $\mathrm{err}(n) = M - \sum_{m'=0}^{n-1} a_{m'}$ *and* $\mathrm{avg}(n) = \frac{1}{n}\sum_{m'=0}^{n-1}\mathrm{err}(m')$. *Then,* $|\frac{1}{n}\sum_{m'=0}^{n-1} m'a_{m'}| \le M/n + \mathrm{err}(n) + \mathrm{avg}(n)$.

Proof. Let $S_0 = 0$ and for $n \ge 1$ define $S_n = \sum_{m'=0}^{n-1} a_{m'}$. We have

$$\sum_{m'=0}^{n-1} m a_{m'} = (n-1)S_n - (S_0 + S_1 + \ldots + S_{n-1}).$$

Therefore,

$$\frac{1}{n}\sum_{m'=0}^{n-1} m'a_{m'} = \frac{n-1}{n}S_{n-1} - \frac{1}{n}\sum_{m'=0}^{n-1} S_{m'}.$$

Since $M - M/n - \mathrm{err}(n) \le \frac{n-1}{n}S_{n-1} \le M$ and $M - \mathrm{avg}(n) \le \frac{1}{n}\sum_{m'=0}^{n-1} S_{m'} \le M$, using the triangle inequality the result follows. □

Now let $a_{m'} = |c^{(i,j)}(m')|$. By assumption $\lim_{n\to\infty}\sum_{m'=0}^{n-1} a_{m'} = M^{(i,j)} < \infty$. Therefore, Lemma 1 along with (5) concludes

$$\mathrm{cov}\left(Y^{(i)}(k), Y^{(j)}(l)\right) = O(\max\{1/n, \mathrm{err}^{(\mathrm{i,j})}(n), \mathrm{avg}^{(\mathrm{i,j})}(n)\}). \tag{10}$$

$\text{err}^{(i,j)}(n)$ is a decreasing function of n. Therefore, $\text{avg}^{(i,j)}(n) \geq \text{err}^{(i,j)}(n)$, for $n \geq 1$. Hence

$$\text{cov}\left(Y^{(i)}(k), Y^{(j)}(l)\right) = O(\max\{1/n, \text{avg}^{(i,j)}(n)\}).$$

Similarly using Lemma 1 along with (6), (7), (8), and (9) we obtain

$$|\text{var}\left(Y^{(i)}(k)\right) - D^{(i,i)}(k)| = O(\max\{1/n, \text{avg}^{(i,i)}(n)\}) \tag{11}$$

and

$$|\text{var}\left(Y^{(j)}(l)\right) - D^{(j,j)}(l)| = O(\max\{1/n, \text{avg}^{(j,j)}(n)\}). \tag{12}$$

Using the definition

$$\text{cor}\left(Y^{(i)}(k), Y^{(j)}(l)\right) = \frac{\text{cov}\left(Y^{(i)}(k), Y^{(j)}(l)\right)}{\sqrt{\text{var}\left(Y^{(i)}(k)\right)}\sqrt{\text{var}\left(Y^{(j)}(l)\right)}}$$

and the fact that as $n \to \infty$, $D^{(i,i)}(k)$ and $D^{(j,j)}(l)$ converge to constants $\mathbf{C}^{(i,i)}(k)$ and $\mathbf{C}^{(j,j)}(l)$, respectively, equations (10), (11), and (12) conclude

$$\text{cor}\left(Y^{(i)}(k), Y^{(j)}(l)\right) = O(\max\{1/n, \text{avg}^{(i,j)}(n)\}).$$

□

As an example we apply Theorem 1 to a scalar auto-regressive (AR) process $X(k)$ specified by

$$X(k) = \sum_{l=1}^{L} \varphi_l X(k-l) + \varepsilon(k),$$

in which φ_l are real-valued coefficients and $\varepsilon(.)$ is a stationary process with no temporal correlation. The auto-covariance function of an AR process can be written as [18]

$$c(t) = \sum_{l=1}^{L} \alpha_l r_l^{|t|},$$

in which $r_1, \ldots, r_l$ are the roots of the polynomial $\beta(x) = x^L - \sum_{l=1}^{L} \varphi_l x^{L-l}$. It is known that for a stationary AR process, $|r_l| < 1$ for all $1 \leq l \leq L$ [18]. Therefore, using the definition of err(.) we have

$$\begin{aligned}\text{err}(n) &= \sum_{t=n}^{\infty} |c(t)| = \sum_{t=n}^{\infty} |\sum_{l=1}^{L} \alpha_l r_l^t| \leq \sum_{l=1}^{L} |\alpha_l| \sum_{t=n}^{\infty} |r_l|^t \\ &= \sum_{l=1}^{L} |\alpha_l| \frac{|r_l|^n}{1-|r_l|} \leq C\zeta^n,\end{aligned}$$

in which $C = \sum_{l=1}^{L} |\alpha_l|/(1-|r_l|)$ and $\zeta = \max_{1\le l\le L}|r_l| < 1$. Hence,

$$\text{avg}(n) = \frac{1}{n}\sum_{m'=0}^{n-1} \text{err}(m') \le \frac{1}{n}\sum_{m'=0}^{n-1} C\zeta^{m'} \le \frac{C}{n(1-\zeta)}.$$

Therefore, Theorem 1 concludes

$$\text{cor}\,(Y(k), Y(l)) = O(1/n), \quad k \neq l,$$

where $Y(.)$ represents the n-point DFT of the AR process $X(.)$.

In the sequel, we assume that the time series $\mathbf{X}$ is multivariate Gaussian, i.e., $X^{(1)}, \ldots, X^{(p)}$ are jointly Gaussian processes. It follows that the DFT components $Y^{(i)}(k)$ are jointly (complex) Gaussian as linear functionals of $\mathbf{X}$. Theorem 1 then immediately implies asymptotic independence of DFT components through a well-known property of jointly Gaussian random variables.

Corollary 1. *Assume that the time series* $\mathbf{X}$ *is multivariate Gaussian. Under the absolute summability conditions in Theorem 1, the DFT components* $Y^{(i)}(k)$ *and* $Y^{(j)}(l)$ *are asymptotically independent for* $k \neq l$ *and* $n \to \infty$.

Corollary 1 implies that for large n, correlation analysis of the time series $\mathbf{X}$ can be done independently on each frequency in the spectral domain. This reduces the problem of screening for hub time series to screening for hub variables among the p DFT components at a given frequency. A procedure for the latter problem and a corresponding theory are described next.

Complex-Valued Correlation Hub Screening

This section discusses complex-valued correlation hub screening, a generalization of real-valued correlation screening in [9, 11], for identifying highly correlated components of a complex-valued random vector from its sample values. The method is applied to multivariate time series in Section "Application to Spectral Screening of Multivariate Gaussian Time Series" to discover correlation hubs among the spectral components at each frequency. Sections "Statistical Model" and "Screening Procedure" describe the underlying statistical model and the screening procedure. Sections "U-score Representation of Correlation Matrices" and "Properties of U-scores" provide background on the U-score representation of correlation matrices and associated definitions and properties. Section "Number of Hub Discoveries in the High-Dimensional Limit" contains the main theoretical result characterizing the number of hub discoveries in the high-dimensional regime, while Section "Phase Transitions and Critical Threshold" elaborates on the phenomenon of phase transitions in the number of discoveries.

Statistical Model

We use the generic notation $\mathbf{Z} = [Z_1, Z_2, \cdots, Z_p]^T$ in this section to refer to a complex-valued random vector. The mean of $\mathbf{Z}$ is denoted as $\boldsymbol{\mu}$ and its $p \times p$ nonsingular covariance matrix is denoted as $\boldsymbol{\Sigma}$. We assume that the vector $\mathbf{Z}$ follows a complex elliptically contoured distribution with pdf $f_{\mathbf{Z}}(\mathbf{z}) = g\left((\mathbf{z}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\right)$, in which $g : \mathbb{R}^{\geq 0} \to \mathbb{R}^{>0}$ is an integrable and strictly decreasing function [19]. This assumption generalizes the Gaussian assumption made in Section "Spectral Representation of Multivariate Time Series", as the Gaussian distribution is one example of an elliptically contoured distribution.

In correlation hub screening, the quantities of interest are the correlation matrix and partial correlation matrix associated with $\mathbf{Z}$. These are defined as $\boldsymbol{\Gamma} = \mathbf{D}_{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \boldsymbol{\Sigma} \mathbf{D}_{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$ and $\boldsymbol{\Omega} = \mathbf{D}_{\boldsymbol{\Sigma}^{-1}}^{-\frac{1}{2}} \boldsymbol{\Sigma}^{-1} \mathbf{D}_{\boldsymbol{\Sigma}^{-1}}^{-\frac{1}{2}}$, respectively. Note that $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}$ are normalized matrices with unit diagonals.

Screening Procedure

The goal of correlation hub screening is to identify highly correlated components of the random vector $\mathbf{Z}$ from its sample realizations. Assume that m samples $\mathbf{z}_1, \ldots, \mathbf{z}_m \in \mathbb{R}^p$ of $\mathbf{Z}$ are available. To simplify the development of the theory, the samples are assumed to be independent and identically distributed (i.i.d.) although the theory also applies to dependent samples.

We compute sample correlation and partial correlation matrices from the samples $\mathbf{z}_1, \ldots, \mathbf{z}_m$ as surrogates for the unknown population correlation matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}$ in Section "Statistical Model". First define the $p \times p$ sample covariance matrix $\mathbf{S}$ as $\mathbf{S} = \frac{1}{m-1} \sum_{i=1}^{m} (\mathbf{z}_i - \overline{\mathbf{z}})(\mathbf{z}_i - \overline{\mathbf{z}})^H$, where $\overline{\mathbf{z}}$ is the sample mean, the average of $\mathbf{z}_1, \ldots, \mathbf{z}_m$. The sample correlation and sample partial correlation matrices are then defined as $\mathbf{R} = \mathbf{D}_{\mathbf{S}}^{-\frac{1}{2}} \mathbf{S} \mathbf{D}_{\mathbf{S}}^{-\frac{1}{2}}$ and $\mathbf{P} = \mathbf{D}_{\mathbf{R}^\dagger}^{-\frac{1}{2}} \mathbf{R}^\dagger \mathbf{D}_{\mathbf{R}^\dagger}^{-\frac{1}{2}}$, respectively, where $\mathbf{R}^\dagger$ is the Moore-Penrose pseudo-inverse of $\mathbf{R}$.

Correlation hubs are screened by applying thresholds to the sample (partial) correlation matrix. A variable Z_i is declared a hub screening discovery at degree level $\delta \in \{1, 2, \ldots\}$ and threshold level $\rho \in [0, 1]$ if

$$|\{j : j \neq i, |\psi_{ij}| \geq \rho\}| \geq \delta,$$

where $\boldsymbol{\Psi} = \mathbf{R}$ for correlation screening and $\boldsymbol{\Psi} = \mathbf{P}$ for partial correlation screening. We denote by $N_{\delta,\rho} \in \{0, \ldots, p\}$ the total number of hub screening discoveries at levels δ, ρ.

Correlation hub screening can also be interpreted in terms of the *(partial) correlation graph* $\mathscr{G}_\rho(\boldsymbol{\Psi})$, depicted in Fig. 1 and defined as follows. The vertices of $\mathscr{G}_\rho(\boldsymbol{\Psi})$ are $v_1, \cdots, v_p$ which correspond to $Z_1, \cdots, Z_p$, respectively. For $1 \leq i, j \leq p$, v_i and v_j are connected by an edge in $\mathscr{G}_\rho(\boldsymbol{\Psi})$ if the magnitude of the sample (partial)

correlation coefficient between Z_i and Z_j is at least ρ. A vertex of $\mathscr{G}_\rho(\boldsymbol{\Psi})$ is called a δ-hub if its degree, the number of incident edges, is at least δ. Then, the number of discoveries $N_{\delta,\rho}$ defined earlier is the number of δ-hubs in the graph $\mathscr{G}_\rho(\boldsymbol{\Psi})$.

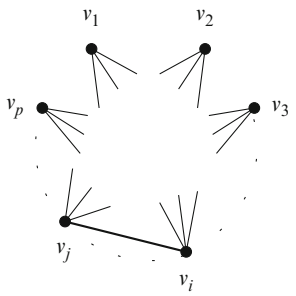


Fig. 1 Complex-valued (partial) correlation hub screening thresholds the sample correlation or partial correlation matrix, denoted generically by the matrix $\boldsymbol{\Psi}$, to find variables Z_i that are highly correlated with other variables. This is equivalent to finding hubs in a graph $\mathscr{G}_\rho(\boldsymbol{\Psi})$ with p vertices $v_1, \cdots, v_p$. For $1 \le i, j \le p$, v_i is connected to v_j in $\mathscr{G}_\rho(\boldsymbol{\Psi})$ if $|\psi_{ij}| \ge \rho$.

U-score Representation of Correlation Matrices

Our theory for complex-valued correlation screening is based on the U-score representation of the sample correlation and partial correlation matrices. Similarly to the real case [9], it can be shown that there exists an $(m-1) \times p$ complex-valued matrix $\mathbb{U}_{\mathbf{R}}$ with unit-norm columns $\mathbf{u}_{\mathbf{R}}^{(i)} \in \mathbb{C}^{m-1}$ such that the following representation holds:

$$\mathbf{R} = \mathbb{U}_{\mathbf{R}}^H \mathbb{U}_{\mathbf{R}}. \tag{13}$$

Similar to Lemma 1 in [9] it is straightforward to show that

$$\mathbf{R}^\dagger = \mathbb{U}_{\mathbf{R}}^H (\mathbb{U}_{\mathbf{R}} \mathbb{U}_{\mathbf{R}}^H)^{-2} \mathbb{U}_{\mathbf{R}}.$$

Hence, by defining $\mathbb{U}_{\mathbf{P}} = (\mathbb{U}_{\mathbf{R}} \mathbb{U}_{\mathbf{R}}^H)^{-1} \mathbb{U}_{\mathbf{R}} \mathbf{D}_{\mathbb{U}_{\mathbf{R}}^H (\mathbb{U}_{\mathbf{R}} \mathbb{U}_{\mathbf{R}}^H)^{-2} \mathbb{U}_{\mathbf{R}}}^{-\frac{1}{2}}$ we have the representation

$$\mathbf{P} = \mathbb{U}_{\mathbf{P}}^H \mathbb{U}_{\mathbf{P}}, \tag{14}$$

where the $(m-1) \times p$ matrix $\mathbb{U}_{\mathbf{P}}$ has unit-norm columns $\mathbf{u}_{\mathbf{P}}^{(i)} \in \mathbb{C}^{m-1}$.

Properties of U-scores

The U-score factorizations in (13) and (14) show that sample (partial) correlation matrices can be represented in terms of unit vectors in $\mathbb{C}^{m-1}$. This subsection presents definitions and properties related to U-scores that will be used in Section "Number of Hub Discoveries in the High-Dimensional Limit".

We denote the unit spheres in $\mathbb{R}^{m-1}$ and $\mathbb{C}^{m-1}$ as S_{m-1} and T_{m-1}, respectively. The surface areas of S_{m-1} and T_{m-1} are denoted as a_{m-1} and b_{m-1}, respectively. Define the interleaving function $h : \mathbb{R}^{2m-2} \to \mathbb{C}^{m-1}$ as below:

$$h([x_1, x_2, \cdots, x_{2m-2}]^T) = \\ [x_1 + x_2\sqrt{-1}, x_3 + x_4\sqrt{-1}, \cdots, x_{2m-3} + x_{2m-2}\sqrt{-1}]^T.$$

Note that $h(.)$ is a one-to-one and onto function and it maps S_{2m-2} to T_{m-1}.

For a fixed vector $\mathbf{u} \in T_{m-1}$ and a threshold $0 \le \rho \le 1$ define the spherical cap in T_{m-1}:

$$A_\rho(\mathbf{u}) = \{\mathbf{y} : \mathbf{y} \in T_{m-1}, |\mathbf{y}^H \mathbf{u}| \ge \rho\}.$$

Also define P_0 as the probability that a random point $\mathbf{Y}$ that is uniformly distributed on T_{m-1} falls into $A_\rho(\mathbf{u})$. Below we give a simple expression for P_0 as a function of ρ and m.

Lemma 2. *Let* $\mathbf{Y}$ *be an* $(m-1)$*-dimensional complex-valued random vector that is uniformly distributed over* T_{m-1}*. We have* $P_0 = \mathbb{P}\left(\mathbf{Y} \in A_\rho(\mathbf{u})\right) = (1-\rho^2)^{m-2}$.

Proof. Without loss of generality we assume $\mathbf{u} = [1, 0, \cdots, 0]^T$. We have

$$P_0 = \mathbb{P}(|Y_1| \ge \rho) = \mathbb{P}(\mathrm{Re}(Y_1)^2 + \mathrm{Im}(Y_1)^2 \ge \rho^2).$$

Since $\mathbf{Y}$ is uniform on T_{m-1}, we can write $\mathbf{Y} = \mathbf{X}/\|\mathbf{X}\|_2$, in which $\mathbf{X}$ is complex-valued random vector whose entries are i.i.d. complex-valued Gaussian variables with mean 0 and variance 1. Thus,

$$\begin{aligned} P_0 &= \mathbb{P}\left(\left(\mathrm{Re}(X_1)^2 + \mathrm{Im}(X_1^2)\right)/\|\mathbf{X}\|_2^2 \ge \rho^2\right) \\ &= \mathbb{P}\left((1-\rho^2)\left(\mathrm{Re}(X_1)^2 + \mathrm{Im}(X_1)^2\right) \ge \rho^2 \sum_{k=2}^{m-1} \mathrm{Re}(X_k)^2 + \mathrm{Im}(X_k)^2\right). \end{aligned}$$

Define $V_1 = \mathrm{Re}(X_1)^2 + \mathrm{Im}(X_1)^2$ and $V_2 = \sum_{k=2}^{m-1} \mathrm{Re}(X_k)^2 + \mathrm{Im}(X_k)^2$. V_1 and V_2 are independent and have chi-squared distributions with 2 and $2(m-2)$ degrees of freedom, respectively [20]. Therefore,

$$\begin{aligned} P_0 &= \int_0^\infty \int_{\rho^2 v_2/(1-\rho^2)}^\infty \chi^2_2(v_1)\chi^2_{2(m-2)}(v_2)dv_1dv_2 \\ &= \int_0^\infty \chi^2_{2(m-2)}(v_2) \int_{\rho^2 v_2/(1-\rho^2)}^\infty \frac{1}{2}e^{-v_1/2}dv_1dv_2 \end{aligned}$$

$$= \int_0^\infty \frac{1}{2^{m-2}\Gamma(m-2)} v_2^{m-3} e^{-v_2/2} e^{-\frac{\rho^2}{2(1-\rho^2)}v_2} dv_2$$
$$= \frac{1}{\Gamma(m-2)}(1-\rho^2)^{m-2} \int_0^\infty x^{m-3} e^{-x} dx$$
$$= \frac{1}{\Gamma(m-2)}(1-\rho^2)^{m-2}\Gamma(m-2) = (1-\rho^2)^{m-2},$$

in which we have made a change of variable $x = \frac{v_2}{2(1-\rho^2)}$. □

Under the assumption that the joint pdf of $\mathbf{Z}$ exists, the p columns of the U-score matrix have joint pdf $f_{\mathbf{U}_1,\ldots,\mathbf{U}_p}(\mathbf{u}_1,\ldots,\mathbf{u}_p)$ on $T_{m-1}^p = \times_{i=1}^p T_{m-1}$. The following $(\delta+1)$-fold average of the joint pdf will play a significant role in Section "Number of Hub Discoveries in the High-Dimensional Limit". This $(\delta+1)$-fold average is defined as

$$\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*\delta+1}}}(\mathbf{u}_1,\ldots,\mathbf{u}_{\delta+1}) = \frac{1}{(2\pi)^{\delta+1} p\binom{p-1}{\delta}} \times$$
$$\sum_{1\le i_1<\cdots<i_\delta\le p, i_{\delta+1}\notin\{i_1,\cdots,i_\delta\}} \int_0^{2\pi}\int_0^{2\pi}\cdots\int_0^{2\pi}$$
$$f_{\mathbf{U}_{i_1},\ldots,\mathbf{U}_{i_\delta},\mathbf{U}_{i_{\delta+1}}}(e^{\sqrt{-1}\theta_1}\mathbf{u}_1,\ldots,e^{\sqrt{-1}\theta_\delta}\mathbf{u}_\delta, e^{\sqrt{-1}\theta}\mathbf{u}_{\delta+1})\, d\theta_1\cdots d\theta_\delta\, d\theta.$$

Also for a joint pdf $f_{\mathbf{U}_1,\ldots,\mathbf{U}_{\delta+1}}(\mathbf{u}_1,\ldots,\mathbf{u}_{\delta+1})$ on $T_{m-1}^{\delta+1}$ define

$$J(f_{\mathbf{U}_1,\ldots,\mathbf{U}_{\delta+1}}) = a_{2m-2}^{\delta} \int_{S_{2m-2}} f_{\mathbf{U}_1,\ldots,\mathbf{U}_{\delta+1}}(h(\mathbf{u}),\ldots,h(\mathbf{u}))d\mathbf{u}.$$

Note that $J(f_{\mathbf{U}_1,\ldots,\mathbf{U}_{\delta+1}})$ is proportional to the integral of $f_{\mathbf{U}_1,\ldots,\mathbf{U}_{\delta+1}}$ over the manifold $\mathbf{u}_1 = \ldots = \mathbf{u}_{\delta+1}$. The quantity $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}})$ is key in determining the asymptotic average number of hubs in a complex-valued correlation network. This will be described in more detail in Sec. "Number of Hub Discoveries in the High-Dimensional Limit".

Let $\mathbf{i} = (i_0, i_1, \ldots, i_\delta)$ be a set of distinct indices, i.e., $1 \le i_0 \le p, 1 \le i_1 < \ldots < i_\delta \le p$, and $i_1, \ldots, i_\delta \ne i_0$. For a U-score matrix $\mathbb{U}$ define the dependency coefficient between the columns $\mathbf{U}_\mathbf{i} = \{\mathbf{U}_{i_0}, \mathbf{U}_{i_1}, \ldots, \mathbf{U}_{i_\delta}\}$ and their complementary k-NN (k-nearest neighbor) set $A_k(\mathbf{i})$ defined in (29) and Fig. 2 as

$$\Delta_{p,m,k,\delta}(\mathbf{i}) = \left\| (f_{\mathbf{U}_\mathbf{i}|\mathbf{U}_{A_k(\mathbf{i})}} - f_{\mathbf{U}_\mathbf{i}})/f_{\mathbf{U}_\mathbf{i}} \right\|_\infty,$$

where $\|\cdot\|_\infty$ denotes the supremum norm. The average of these coefficients is defined as

$$\|\Delta_{p,m,k,\delta}\|_1 = \frac{1}{p\binom{p-1}{\delta}} \sum_{i_0=1}^{p} \sum_{\substack{i_1,\ldots,i_\delta\ne i_0 \\ 1\le i_1<\ldots<i_\delta\le p}} \Delta_{p,m,k,\delta}(\mathbf{i}). \tag{15}$$

Number of Hub Discoveries in the High-Dimensional Limit

We now present the main theoretical result on complex-valued correlation screening. The following theorem gives asymptotic expressions for the mean number of δ-hubs and the probability of discovery of at least one δ-hub in the graph $\mathscr{G}_\rho(\mathbf{\Psi})$. It also gives bounds on the rates of convergence to these approximations as the dimension p increases and $\rho \to 1$. We use $\mathbb{U} = [\mathbf{U}_1, \cdots, \mathbf{U}_p]$ as a generic notation for the U-score representation of the sample (partial) correlation matrix. The asymptotic expression for the mean $\mathbb{E}[N_{\delta,\rho}]$ is denoted by Λ and is given by

$$\Lambda = p\binom{p-1}{\delta} P_0^\delta J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}). \tag{16}$$

Define $\eta_{p,\delta}$ as

$$\eta_{p,\delta} = p^{1/\delta}(p-1)P_0 = p^{1/\delta}(p-1)(1-\rho^2)^{(m-2)}, \tag{17}$$

where the last equation is due to Lemma 2. The parameter k below represents an upper bound on the true hub degree, i.e., the number of nonzero entries in any row of the population covariance matrix $\mathbf{\Sigma}$. Also let $\varphi(\delta)$ be the function that takes values $\varphi(\delta) = 2$ for $\delta = 1$ and $\varphi(\delta) = 1$ for $\delta > 1$.

Theorem 2. *Let $\mathbb{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_p]$ be a $(m-1) \times p$ random matrix with $\mathbf{U}_i \in T_{m-1}$ where $m > 2$. Let $\delta \geq 1$ be a fixed integer. Assume the joint pdf of any subset of the $\mathbf{U}_i$s is bounded and differentiable. Then, with Λ defined in (16),*

$$\left|\mathbb{E}[N_{\delta,\rho}] - \Lambda\right| \leq O\left(\eta_{p,\delta}^\delta \max\left\{\eta_{p,\delta} p^{-1/\delta}, (1-\rho)^{1/2}\right\}\right). \tag{18}$$

Furthermore, let $N^*_{\delta,\rho}$ be a Poisson distributed random variable with rate $\mathbb{E}[N^*_{\delta,\rho}] = \Lambda/\varphi(\delta)$. If $(p-1)P_0 \leq 1$, then

$$\left|\mathbb{P}(N_{\delta,\rho} > 0) - \mathbb{P}(N^*_{\delta,\rho} > 0)\right| \leq$$
$$\begin{cases} O\left(\eta_{p,\delta}^\delta \max\left\{\eta_{p,\delta}^\delta (k/p)^{\delta+1}, Q_{p,k,\delta}, \|\Delta_{p,m,k,\delta}\|_1, p^{-1/\delta}, (1-\rho)^{1/2}\right\}\right), & \delta > 1 \\ O\left(\eta_{p,1} \max\left\{\eta_{p,1}(k/p)^2, \|\Delta_{p,m,k,1}\|_1, p^{-1}, (1-\rho)^{1/2}\right\}\right), & \delta = 1 \end{cases}, \tag{19}$$

with $Q_{p,k,\delta} = \eta_{p,\delta}\left(k/p^{1/\delta}\right)^{\delta+1}$ and $\|\Delta_{p,m,k,\delta}\|_1$ defined in (15).

Proof. The proof is similar to the proof of proposition 1 in [9]. First we prove (18). Let $\phi_i = I(d_i \geq \delta)$ be the indicator of the event that $d_i \geq \delta$, in which d_i represents the degree of the vertex v_i in the graph $\mathscr{G}_\rho(\mathbf{\Psi})$. We have $N_{\delta,\rho} = \sum_{i=1}^p \phi_i$. With ϕ_{ij} being the indicator of the presence of an edge in $\mathscr{G}_\rho(\mathbf{\Psi})$ between vertices v_i and v_j we have the relation

$$\phi_i = \sum_{l=\delta}^{p-1} \sum_{\mathbf{k}\in\check{\mathscr{C}}_i(p-1,l)} \prod_{j=1}^{l} \phi_{ik_j} \prod_{q=l+1}^{p-1} (1-\phi_{ik_q}) \tag{20}$$

where we have defined the index vector $\mathbf{k} = (k_1,\ldots,k_{p-1})$ and the set

$$\check{\mathscr{C}}_i(p-1,l) =$$

$$\{\mathbf{k} : k_1 < \ldots < k_l, k_{l+1} < \ldots < k_{p-1}\, k_j \in \{1,\ldots,p\} - \{i\}, k_j \neq k_{j'}\}.$$

The inner summation in (20) simply sums over the set of distinct indices not equal to i that index all $\binom{p-1}{l}$ different types of products of the form: $\prod_{j=1}^{l} \phi_{ik_j} \prod_{q=l+1}^{p-1}(1-\phi_{ik_q})$. Subtracting $\sum_{\mathbf{k}\in\check{\mathscr{C}}_i(p-1,\delta)} \prod_{j=1}^{\delta} \phi_{ik_j}$ from both sides of (20)

$$\begin{aligned}
&\phi_i - \sum_{\mathbf{k}\in\check{\mathscr{C}}_i(p-1,\delta)} \prod_{j=1}^{\delta} \phi_{ik_j} \\
&\quad = \sum_{l=\delta+1}^{p-1} \sum_{\mathbf{k}\in\check{\mathscr{C}}_i(p-1,l)} \prod_{j=1}^{l} \phi_{ik_j} \prod_{q=l+1}^{p-1} (1-\phi_{ik_q}) \\
&\qquad + \sum_{\mathbf{k}\in\check{\mathscr{C}}_i(p-1,l)} \sum_{q=\delta+1}^{p-1} (-1)^{q-\delta} \\
&\qquad \sum_{k'_{\delta+1}<\ldots<k'_q, \{k'_{\delta+1},\ldots,k'_q\}\subset\{k_{\delta+1},\ldots,k_{p-1}\}} \prod_{j=1}^{l} \phi_{ik_j} \prod_{s=\delta+1}^{q} \phi_{ik'_s}
\end{aligned} \tag{21}$$

in which we have used the expansion

$$\prod_{q=\delta+1}^{p-1} (1-\phi_{ik_q}) = 1 + \sum_{q=\delta+1}^{p-1} (-1)^{q-\delta} \sum_{k'_{\delta+1}<\ldots<k'_q, \{k'_{\delta+1},\ldots,k'_q\}\subset\{k_{\delta+1},\ldots,k_{p-1}\}} \prod_{s=\delta+1}^{q} \phi_{ik'_s}.$$

The following simple asymptotic representation will be useful in the sequel. For any $i_1,\ldots,i_k \in \{1,\ldots,p\}$, $i_1 \neq \cdots \neq i_k \neq i$, $k \in \{1,\ldots,p-1\}$,

$$\begin{aligned}
\mathbb{E}\left[\prod_{j=1}^{k} \phi_{ii_j}\right] &= \int_{S_{2m-2}} \int_{h^{-1}(A_\rho(\boldsymbol{v}))} \cdots \int_{h^{-1}(A_\rho(\boldsymbol{v}))} \\
&\quad f_{\mathbf{U}_{i_1},\ldots,\mathbf{U}_{i_k},\mathbf{U}_i}(h(\mathbf{v}_1),\cdots,h(\mathbf{v}_k),h(\mathbf{v}))\, d\mathbf{v}_1\cdots d\mathbf{v}_k\, d\mathbf{v} \\
&\leq P_0^k a_{2m-2}^k M_{k|1},
\end{aligned} \tag{22}$$

where $P_0, A_\rho(\mathbf{u})$, and the function $h(.)$ are defined in Sec. "Properties of U-scores". Moreover,

$$M_{k|1} = \max_{i_1\neq\cdots\neq i_{k+1}} \left\| f_{\boldsymbol{U}_{i_1},\ldots,\boldsymbol{U}_{i_k}|\boldsymbol{U}_{i_{k+1}}} \right\|_\infty.$$

The following simple generalization of (22) to arbitrary product indices ϕ_{ij} will also be needed

$$\mathbb{E}\left[\prod_{l=1}^{q}\phi_{i_l j_l}\right] \le P_0^q a_{2m-2}^q M_{|Q|}, \tag{23}$$

where $Q = \text{unique}(\{i_l, j_l\}_{l=1}^q)$ is the set of unique indices among the distinct pairs $\{(i_l, j_l)\}_{l=1}^q$ and $M_{|Q|}$ is a bound on the joint pdf of $\mathbf{U}_Q$.

Define the random variable

$$\theta_i = \binom{p-1}{\delta}^{-1} \sum_{\mathbf{k}\in\breve{\mathscr{C}}_i(p-1,\delta)} \prod_{j=1}^{\delta}\phi_{ik_j}.$$

We show below that for sufficiently large p

$$\left|\mathbb{E}[\phi_i] - \binom{p-1}{\delta}\mathbb{E}[\theta_i]\right| \le \gamma_{p,\delta}((p-1)P_0)^{\delta+1}, \tag{24}$$

where $\gamma_{p,\delta} = \max_{\delta+1\le l<p}\{a_{2m-2}^l M_{l|1}\}\left(e - \sum_{l=0}^{\delta}\frac{1}{l!}\right)\left(1+(\delta!)^{-1}\right)$ and $M_{l|1}$ is a least upper bound on any l-dimensional joint pdf of the variables $\{\mathbf{U}_i\}_{j\ne i}^p$ conditioned on $\mathbf{U}_i$.

To show inequality (24) take expectations of (21) and apply the bound (22) to obtain

$$\begin{aligned}&\left|\mathbb{E}[\phi_i] - \binom{p-1}{\delta}\mathbb{E}[\theta_i]\right| \le \\ &\left|\sum_{l=\delta+1}^{p-1}\binom{p-1}{l}P_0^l a_{2m-2}^l M_{l|1} + \binom{p-1}{\delta}\sum_{l=1}^{p-1-\delta}\binom{p-1-\delta}{l}P_0^{\delta+l}a_{2m-2}^{\delta+l}M_{\delta+l|1}\right| \\ &\le A(1+(\delta!)^{-1}),\end{aligned} \tag{25}$$

where

$$A = \sum_{l=\delta+1}^{p-1}\binom{p-1}{l}((p-1)P_0)^l a_{2m-2}^l M_{l|1}.$$

Line (25) follows from the identity $\binom{p-1-\delta}{l}\binom{p-1}{\delta} = \binom{p-1}{l+\delta}\binom{l+\delta}{l}$ and a change of index in the second summation on the previous line. Since $(p-1)P_0 < 1$

$$\begin{aligned}|A| &\le \max_{\delta+1\le l<p}\{a_{2m-2}^l M_{l|1}\}\sum_{l=\delta+1}^{p-1}\binom{p-1}{l}((p-1)P_0)^l \\ &\le \max_{\delta+1\le l<p}\{a_{2m-2}^l M_{l|1}\}\left(e - \sum_{l=0}^{\delta}\frac{1}{l!}\right)((p-1)P_0)^{\delta+1}.\end{aligned}$$

Application of the mean value theorem to the integral representation (22) yields

$$\left|\mathbb{E}[\theta_i] - P_0^{\delta} J(\overline{f_{\mathbf{U}_{*1-i},\ldots,\mathbf{U}_{*\delta-i},\mathbf{U}_i}})\right| \leq \widetilde{\gamma}_{p,\delta}((p-1)P_0)^{\delta} r, \tag{26}$$

where

$$\overline{f_{\mathbf{U}_{*1-i},\ldots,\mathbf{U}_{*\delta-i},\mathbf{U}_i}}(\mathbf{u}_1, \ \ldots, \mathbf{u}_{\delta+1}) = \frac{1}{(2\pi)^{\delta}\binom{p-1}{\delta}} \sum_{\substack{1\leq i_1<\cdots<i_\delta\leq p \\ i\notin\{i_1,\cdots,i_\delta\}}} \int_0^{2\pi}\cdots\int_0^{2\pi} f_{\mathbf{U}_{i_1},\ldots,\mathbf{U}_{i_\delta},\mathbf{U}_i}(e^{\sqrt{-1}\theta_1}\mathbf{u}_1, \ \ldots, e^{\sqrt{-1}\theta_\delta}\mathbf{u}_\delta, \mathbf{u}_{\delta+1})\, d\theta_1\cdots d\theta_\delta,$$

$r = \sqrt{2(1-\rho)}$, $\widetilde{\gamma}_{p,\delta} = 2a_{2m-2}^{\delta+1}\dot{M}_{\delta+1|1}/\delta!$, and $\dot{M}_{\delta+1|1}$ is a bound on the norm of the gradient

$$\nabla_{\mathbf{u}_{i_1},\ldots,\mathbf{u}_{i_\delta}} \overline{f_{\mathbf{U}_{*1-i},\ldots,\mathbf{U}_{*\delta-i}|\mathbf{U}_i}(\mathbf{u}_{i_1},\ldots,\mathbf{u}_{i_\delta}|\mathbf{u}_i)}.$$

Combining (24)–(26) and the relation $r = O((1-\rho)^{1/2})$,

$$\left|\mathbb{E}[\phi_i] - \binom{p-1}{\delta} P_0^{\delta} J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}})\right| \leq O\left(((p-1)P_0)^{\delta} \max\left\{(p-1)P_0, (1-\rho)^{1/2}\right\}\right).$$

Summing over i and recalling definitions (16) and (17) of Λ and $\eta_{p,\delta}$,

$$\begin{aligned}\left|\mathbb{E}[N_{\delta,\rho}] - \Lambda\right| &\leq O\left(p((p-1)P_0)^{\delta} \max\left\{(p-1)P_0, (1-\rho)^{1/2}\right\}\right) \\ &= O\left(\eta_{p,\delta}^{\delta} \max\left\{\eta_{p,\delta} p^{-1/\delta}, (1-\rho)^{1/2}\right\}\right).\end{aligned}$$

This establishes the bound (18).

Next we prove bound (19) by using the Chen-Stein method [21]. Define

$$\widetilde{N}_{\delta,\rho} = \frac{1}{\varphi(\delta)} \sum_{i_0=1}^{p} \sum_{1\leq i_1<\ldots<i_\delta\leq p} \prod_{j=1}^{\delta} \phi_{i_0 i_j}, \tag{27}$$

where the second sum is over the indices $1 \leq i_1 < \ldots < i_\delta \leq p$ such that $i_j \neq i_0, 1 \leq j \leq \delta$. For $\mathbf{i} \overset{\text{def}}{=} (i_0, i_1, \ldots, i_\delta)$ define the index set $B_{\mathbf{i}} = B_{i_0,i_1,\ldots,i_\delta} = \{(j_0, j_1, \ldots, j_\delta) : j_l \in \mathscr{N}_k(i_l) \cup \{i_l\}, l = 0, \ldots, \delta\} \cap \mathscr{C}^{<}$ where $\mathscr{C}^{<} = \{(j_0, \ldots, j_\delta) : 1 \leq j_0 \leq p, 1 \leq j_1 < \cdots < j_\delta \leq p, j_l \neq j_0, 1 \leq l \leq \delta\}$. These index the distinct sets of points $\mathbf{U}_{\mathbf{i}} = \{\mathbf{U}_{i_0}, \mathbf{U}_{i_1}, \ldots, \mathbf{U}_{i_\delta}\}$ and their respective k-NNs. Note that $|B_{\mathbf{i}}| \leq k^{\delta+1}$. Identifying $\widetilde{N}_{\delta,\rho} = \sum_{\mathbf{i}\in\mathscr{C}^{<}} \prod_{l=1}^{\delta} \phi_{i_0 i_l}$ and $N^*_{\delta,\rho}$, a Poisson distributed random variable with rate $\mathbb{E}[\widetilde{N}_{\delta,\rho}]$, the Chen-Stein bound [21, Theorem 1] is

$$2\max_{A}|\mathbb{P}(\widetilde{N}_{\delta,\rho}\in A)-\mathbb{P}(N^{*}_{\delta,\rho}\in A)|\le b_1+b_2+b_3, \tag{28}$$

where

$$b_1=\sum_{\mathbf{i}\in\mathscr{C}^<}\sum_{\mathbf{j}\in B_{\mathbf{i}}}\mathbb{E}\left[\prod_{l=1}^{\delta}\phi_{i_0i_l}\right]\mathbb{E}\left[\prod_{q=1}^{\delta}\phi_{j_0j_q}\right],$$

$$b_2=\sum_{\mathbf{i}\in\mathscr{C}^<}\sum_{\mathbf{j}\in B_{\mathbf{i}-\{\mathbf{i}\}}}\mathbb{E}\left[\prod_{l=1}^{\delta}\phi_{i_0i_l}\prod_{q=1}^{\delta}\phi_{j_0j_q}\right],$$

and, for $p_{\mathbf{i}}=\mathbb{E}[\prod_{l=1}^{\delta}\phi_{i_0i_l}]$,

$$b_3=\sum_{\mathbf{i}\in\mathscr{C}^<}\mathbb{E}\left[\mathbb{E}\left[\prod_{l=1}^{\delta}\phi_{i_0i_l}-p_{\mathbf{i}}\middle|\phi_{\mathbf{j}}:\mathbf{j}\notin B_{\mathbf{i}}\right]\right].$$

Over the range of indices in the sum of b_1 $\mathbb{E}[\prod_{l=1}^{\delta}\phi_{ii_l}]$ is of order $O(P_0^{\delta})$, by (23), and therefore

$$b_1\le O\left(p^{\delta+1}k^{\delta+1}P_0^{2\delta}\right)=O\left(\eta_{p,\delta}^{2\delta}(k/p)^{\delta+1}\right),$$

which follows from definition (17). More care is needed to bound b_2 due to the repetition of characteristic functions ϕ_{ij}. Since $\mathbf{i}\neq\mathbf{j}$, $\mathbb{E}[\prod_{l=1}^{\delta}\phi_{i_0i_l}\prod_{q=1}^{\delta}\phi_{j_0j_q}]$ is a multiplication of at least $\delta+1$ different characteristic functions, hence by (23),

$$\mathbb{E}[\prod_{l=1}^{\delta}\phi_{i_0i_l}\prod_{q=1}^{\delta}\phi_{j_0j_q}]=O\left(P_0^{\delta+1}\right).$$

Therefore, we conclude that

$$b_2\le O\left(p^{\delta+1}k^{\delta+1}P_0^{\delta+1}\right).$$

Next we bound the term b_3 in (28). The set

$$A_k(\mathbf{i})=B_{\mathbf{i}}^{c}-\{\mathbf{i}\} \tag{29}$$

indexes the complementary k-NN of $\mathbf{U}_{\mathbf{i}}$ (see Fig. 2) so that, using the representation (23),

$$\begin{aligned}b_3&=\sum_{\mathbf{i}\in\mathscr{C}^<}\mathbb{E}\left[\mathbb{E}\left[\prod_{l=1}^{\delta}\phi_{i_0i_l}-p_{\mathbf{i}}\middle|\mathbf{U}_{A_k(\mathbf{i})}\right]\right]\\&=\sum_{\mathbf{i}\in\mathscr{C}^<}\int_{S_{2m-2}^{|A_k(\mathbf{i})|}}d\boldsymbol{u}_{A_k(\mathbf{i})}\left(\prod_{l=1}^{\delta}\int_{S_{2m-2}}d\boldsymbol{u}_{i_0}\int_{A(r,\boldsymbol{u}_{i_0})}d\boldsymbol{u}_{i_l}\right)\\&\quad\left(\frac{f_{\boldsymbol{U}_{\mathbf{i}}|\boldsymbol{U}_{A_k}}(\boldsymbol{u}_{\mathbf{i}}|\boldsymbol{u}_{A_k(\mathbf{i})})-f_{\boldsymbol{U}_{\mathbf{i}}}(\boldsymbol{u}_{\mathbf{i}})}{f_{\boldsymbol{U}_{\mathbf{i}}}(\boldsymbol{u}_{\mathbf{i}})}\right)f_{\boldsymbol{U}_{\mathbf{i}}}(\boldsymbol{u}_{\mathbf{i}})f_{\boldsymbol{U}_{A_k(\mathbf{i})}}(\boldsymbol{u}_{A_k(\mathbf{i})})\end{aligned}$$

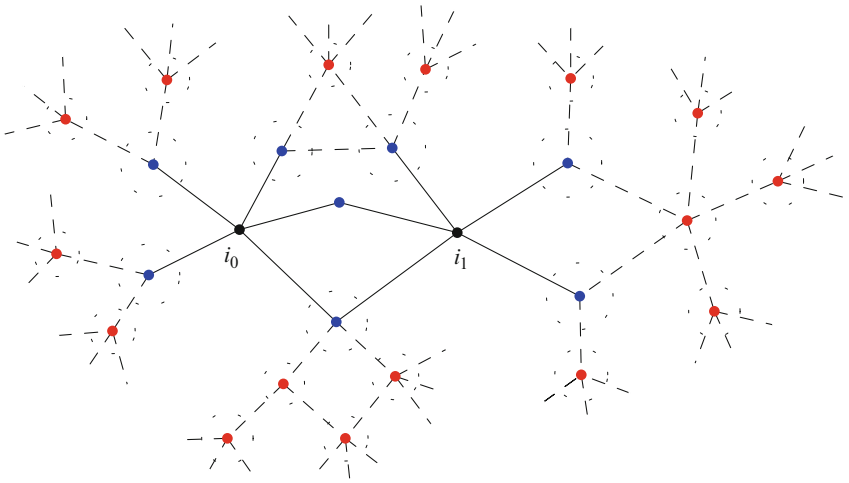


Fig. 2 The complementary k-NN set $A_k(\mathbf{i})$ illustrated for $\delta = 1$ and $k = 5$. Here we have $\mathbf{i} = (i_0, i_1)$. The vertices i_0, i_1, and their k-NNs are depicted in black and blue, respectively. The complement of the union of $\{i_0, i_1\}$ and its k-NNs is the complementary k-NN set $A_k(\mathbf{i})$ and is depicted in red.

$$\leq O\left(p^{\delta+1}P_0^{\delta}\|\Delta_{p,m,k,\delta}\|_1\right) = O\left(\eta_{p,\delta}^{\delta}\|\Delta_{p,m,k,\delta}\|_1\right).$$

Note that by definition of $\widetilde{N}_{\delta,\rho}$ we have $\widetilde{N}_{\delta,\rho} > 0$ if and only if $N_{\delta,\rho} > 0$. This yields

$$\begin{aligned}
&\left|\mathbb{P}(N_{\delta,\rho} > 0) - (1 - \exp(-\Lambda))\right| \leq \left|\mathbb{P}(\widetilde{N}_{\delta,\rho} > 0) - \mathbb{P}(N_{\delta,\rho} > 0)\right| + \\
&\left|\mathbb{P}(\widetilde{N}_{\delta,\rho} > 0) - \left(1 - \exp(-\mathbb{E}[\widetilde{N}_{\delta,\rho}])\right)\right| + \left|\exp(-\mathbb{E}[\widetilde{N}_{\delta,\rho}]) - \exp(-\Lambda)\right| \\
&\leq b_1 + b_2 + b_3 + O\left(\left|\mathbb{E}[\widetilde{N}_{\delta,\rho}] - \Lambda\right|\right). \qquad (30)
\end{aligned}$$

Combining the above inequalities on b_1, b_2, and b_3 yields the first three terms in the argument of the "max" on the right side of (19).

It remains to bound the term $|\mathbb{E}[\widetilde{N}_{\delta,\rho}] - \Lambda|$. Application of the mean value theorem to the multiple integral (23) gives

$$\left|\mathbb{E}\left[\prod_{l=1}^{\delta}\phi_{ii_l}\right] - P_0^{\delta}J\left(f_{\boldsymbol{U}_{i_1},\ldots,\boldsymbol{U}_{i_\delta},\boldsymbol{U}_i}\right)\right| \leq O\left(P_0^{\delta}r\right).$$

Applying relation (27) yields

$$\left|\mathbb{E}[\widetilde{N}_{\delta,\rho}] - p\binom{p-1}{\delta}P_0^{\delta}J\left(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}\right)\right| \leq O\left(p^{\delta+1}P_0^{\delta}r\right) = O\left(\eta_{p,\delta}^{\delta}r\right).$$

Combine this with (30) to obtain bound (19). This completes the proof of Theorem 2. □

An immediate consequence of Theorem 2 is the following result, similar to Proposition 2 in [9], which provides asymptotic expressions for the mean number of δ-hubs and the probability of the event $N_{\delta,\rho} > 0$ as p goes to ∞ and ρ converges to 1 at a prescribed rate.

Corollary 2. *Let $\rho_p \in [0,1]$ be a sequence converging to one as $p \to \infty$ such that $\eta_{p,\delta} = p^{1/\delta}(p-1)(1-\rho_p^2)^{(m-2)} \to e_{m,\delta} \in (0,\infty)$. Then,*

$$\lim_{p\to\infty} \mathbb{E}[N_{\delta,\rho_p}] = \Lambda_\infty = e_{m,\delta}^{\delta}/\delta! \lim_{p\to\infty} J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}). \tag{31}$$

Assume that $k = o(p^{1/\delta})$ and that for the weak dependency coefficient $\|\Delta_{p,m,k,\delta}\|_1$, defined via (15), we have $\lim_{p\to\infty} \|\Delta_{p,m,k,\delta}\|_1 = 0$. Then,

$$\mathbb{P}(N_{\delta,\rho_p} > 0) \to 1 - \exp(-\Lambda_\infty/\varphi(\delta)). \tag{32}$$

Corollary 2 shows that in the limit $p \to \infty$, the number of detected hubs depends on the true population correlations only through the quantity $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}})$. In some cases $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}})$ can be evaluated explicitly. Similar to the argument in [9], it can be shown that if the population covariance matrix $\boldsymbol{\Sigma}$ is sparse in the sense that its nonzero off-diagonal entries can be arranged into a $k \times k$ submatrix by reordering rows and columns, then

$$J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}) = 1 + O(k/p).$$

Hence, if $k = o(p)$ as $p \to \infty$, the quantity $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}})$ converges to 1. If $\boldsymbol{\Sigma}$ is diagonal, then $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}) = 1$ exactly. In such cases, the quantity Λ_∞ in Corollary 2 does not depend on the unknown underlying distribution of the U-scores. As a result, the expected number of δ-hubs in $\mathscr{G}_\rho(\boldsymbol{\Psi})$ and the probability of discovery of at least one δ-hub do not depend on the underlying distribution. We will see in Sec. "Application to Spectral Screening of Multivariate Gaussian Time Series" that this result is useful in assigning statistical significance levels to vertices of the graph $\mathscr{G}_\rho(\boldsymbol{\Psi})$.

Phase Transitions and Critical Threshold

It can be seen from Theorem 2 and Corollary 2 that the number of δ-hub discoveries exhibits a phase transition in the high-dimensional regime where the number of variables p can be very large relative to the number of samples m. Specifically, assume that the population covariance matrix $\boldsymbol{\Sigma}$ is block sparse as in Section "Number of Hub Discoveries in the High-Dimensional Limit". Then, as the correlation threshold ρ is reduced, the number of δ-hub discoveries abruptly increases to the maximum, p. Conversely as ρ increases, the number of discoveries quickly approaches

zero. Similarly, the family-wise error rate (i.e., the probability of discovering at least one δ-hub in a graph with no true hubs) exhibits a phase transition as a function of ρ. Figure 3 shows the family-wise error rate obtained via expression (32) for $\delta = 1$ and $p = 1000$, as a function of ρ and the number of samples m. It is seen that for a fixed value of m there is a sharp transition in the family-wise error rate as a function of ρ.

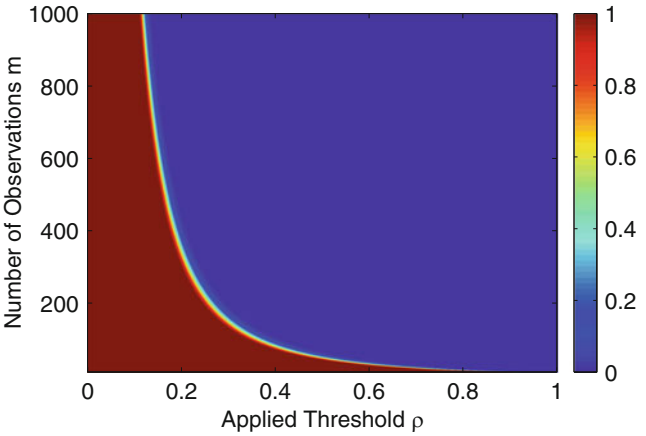


Fig. 3 Family-wise error rate as a function of correlation threshold ρ and number of samples m for $p = 1000, \delta = 1$. The phase transition phenomenon is clearly observable in the plot.

The phase transition phenomenon motivates the definition of a critical threshold $\rho_{c,\delta}$ as the threshold ρ satisfying the following slope condition:

$$\partial \mathbb{E}[N_{\delta,\rho}]/\partial \rho = -p.$$

Using (16) the solution of the above equation can be approximated via the expression below:

$$\rho_{c,\delta} = \sqrt{1 - (c_{m,\delta}(p-1))^{-2\delta/(\delta(2m-3)-2)}}, \tag{33}$$

where $c_{m,\delta} = b_{m-1}\delta J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}})$. The screening threshold ρ should be chosen greater than $\rho_{c,\delta}$ to prevent excessively large numbers of false positives. Note that the critical threshold $\rho_{c,\delta}$ also does not depend on the underlying distribution of the U-scores when the covariance matrix $\mathbf{\Sigma}$ is block sparse.

Expression (33) is similar to the expression obtained in [9] for the critical threshold in real-valued correlation screening. However, in the complex-valued case the coefficient $c_{m,\delta}$ and the exponent of the term $c_{m,\delta}(p-1)$ are different from the real case. This generally results in smaller values of $\rho_{c,\delta}$ for fixed m and δ.

Figure 4 shows the value of $\rho_{c,\delta}$ obtained via (33) as a function of m for different values of δ and p. The critical threshold decreases as either the sample size

m increases, the number of variables p decreases, or the vertex degree δ increases. Note that even for ten billion (10^{10}) dimensions (upper triplet of curves in the figure) only a relatively small number of samples are necessary for complex-valued correlation screening to be useful. For example, with $m = 200$ one can reliably discover connected vertices ($\delta = 1$ in the figure) having correlation greater than $\rho_{c,\delta} = 0.5$.

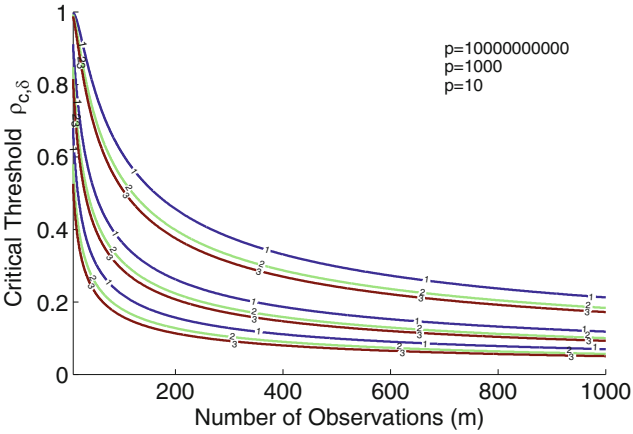


Fig. 4 The critical threshold $\rho_{c,\delta}$ as a function of the sample size m for $\delta = 1, 2, 3$ (curve labels) and $p = 10, 1000, 10^{10}$ (bottom to top triplets of curves). The figure shows that the critical threshold decreases as either m or δ increases. When the number of samples m is small the critical threshold is close to 1 in which case reliable hub discovery is impossible. However, a relatively small increment in m is sufficient to reduce the critical threshold significantly. For example, for $p = 10^{10}$, only $m = 200$ samples are enough to bring $\rho_{c,1}$ down to 0.5.

Application to Spectral Screening of Multivariate Gaussian Time Series

In this section, the complex-valued correlation hub screening method of Section "Complex-Valued Correlation Hub Screening" is applied to stationary multivariate Gaussian time series. Assume that the time series $X^{(1)}, \cdots, X^{(p)}$ defined in Section "Spectral Representation of Multivariate Time Series" satisfy the conditions of Corollary 1. Assume also that a total of $N = n \times m$ time samples of $X^{(1)}, \cdots, X^{(p)}$ are available. We divide the N samples into m parts of n consecutive samples and we take the n-point DFT of each part. Therefore, for each time series, at each frequency $f_i = (i-1)/n$, $1 \leq i \leq n$, m samples are available. This allows us to construct a (partial) correlation graph corresponding to each frequency. We denote the (partial) correlation graph corresponding to frequency f_i and correlation threshold ρ_i as $\mathscr{G}_{f_i,\rho_i}$. $\mathscr{G}_{f_i,\rho_i}$ has p vertices $v_1, v_2, \cdots, v_p$ corresponding to time series $X^{(1)}, X^{(2)}, \cdots, X^{(p)}$,

respectively. Vertices v_k and v_l are connected if the magnitude of the sample (partial) correlation between the DFTs of $X^{(k)}$ and $X^{(l)}$ at frequency f_i (i.e. the sample (partial) correlation between $Y^{(k)}(i-1)$ and $Y^{(l)}(i-1)$) is at least ρ_i.

Consider a single frequency f_i and the null hypothesis, $\mathscr{H}_0$, that the correlations among the time series $X^{(1)}, X^{(2)}, \cdots, X^{(p)}$ at frequency f_i are block sparse in the sense of Section "Number of Hub Discoveries in the High-Dimensional Limit". As discussed in Sec. "Number of Hub Discoveries in the High-Dimensional Limit", under $\mathscr{H}_0$ the expected number of δ-hubs and the probability of discovery of at least one δ-hub in graph $\mathscr{G}_{f_i,\rho_i}$ are not functions of the unknown underlying distribution of the data. Therefore, the results of Corollary 2 may be used to quantify the statistical significance of declaring vertices of $\mathscr{G}_{f_i,\rho_i}$ to be δ-hubs. The statistical significance is represented by the p-value, defined in general as the probability of having a test statistic at least as extreme as the value actually observed assuming that the null hypothesis $\mathscr{H}_0$ is true. In the case of correlation hub screening, the p-value $pv_\delta(j)$ assigned to vertex v_j for being a δ-hub is the maximal probability that v_j maintains degree δ given the observed sample correlations, assuming that the block-sparse hypothesis $\mathscr{H}_0$ is true. The detailed procedure for assigning p-values is similar to the procedure in [9] for real-valued correlation screening and is illustrated in Fig. 5. Equation (33) helps in choosing the initial threshold ρ^*.

- Initialization:
  1. Choose a degree threshold $\delta \geq 1$.
  2. Choose an initial threshold $\rho^* > \rho_{c,\delta}$.
  3. Calculate the degree d_j of each vertex of graph $\mathscr{G}_{\rho^*}(\blacksquare)$.
  4. Select a value of $\delta \in \{1, \cdots, \max_{1 \leq j \leq p} d_j\}$.
- For each $j = 1, \cdots, p$ find $\rho_j(\delta)$ as the δth greatest element of the jth row of the sample (partial) correlation matrix.
- Approximate the p-value corresponding to vertex v_j as $pv_\delta(j) \approx 1 - \exp(-\mathbb{E}[N_{\delta,\rho_j(\delta)}]/\varphi(\delta))$, where $\mathbb{E}[N_{\delta,\rho_j(\delta)}]$ is approximated by the limiting expression (31) using $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}) = 1$.
- Screen variables by thresholding the p-values $pv_\delta(j)$ at desired significance level.

Fig. 5 Procedure for assigning p-values to the vertices of $\mathscr{G}_{\rho^*}(\boldsymbol{\Psi})$.

Given Corollary 1, for $i \neq j$ the correlation graphs $\mathscr{G}_{f_i,\rho_i}$ and $\mathscr{G}_{f_j,\rho_j}$ and their associated inferences are approximately independent. Thus, we can solve multiple inference problems by first performing correlation hub screening on each graph as discussed above and then aggregating the inferences at each frequency in a straightforward manner. Examples of aggregation procedures are described below.

Disjunctive Hubs

One task that can be easily performed is finding the p-value for a given time series to be a hub in at least one of the graphs $\mathscr{G}_{f_1,\rho_1},\cdots,\mathscr{G}_{f_n,\rho_n}$. More specifically, for each $j=1,\ldots,p$ denote the p-values for vertex v_j being a δ-hub in $\mathscr{G}_{f_1,\rho_1},\cdots,\mathscr{G}_{f_n,\rho_n}$ by $pv_{f_1,\rho_1,\delta}(j),\cdots,pv_{f_n,\rho_n,\delta}(j)$, respectively. These p-values are obtained using the method of Fig. 5. Then, $pv_\delta(j)$, the p-value for the vertex v_j being a δ-hub in at least one of the frequency graphs $\mathscr{G}_{f_1,\rho_1},\cdots,\mathscr{G}_{f_n,\rho_n}$, can be approximated as

$$\mathbb{P}(\exists i: d_{j,f_i}\geq\delta\,|\mathscr{H}_0)\approx\hat{pv}_\delta(j)=1-\prod_{i=1}^{n}(1-pv_{f_i,\rho_i,\delta}(j)),$$

in which d_{j,f_i} is the degree of v_j in the graph $\mathscr{G}_{f_i,\rho_i}$.

Conjunctive Hubs

Another property of interest is the existence of a hub at all frequencies for a particular time series. In this case we have

$$\mathbb{P}(\forall i: d_{j,f_i}\geq\delta\,|\mathscr{H}_0)\approx\check{pv}_\delta(j)=\prod_{i=1}^{n}pv_{f_i,\rho_i,\delta}(j).$$

General Persistent Hubs

The general case is the event that at least K frequencies have hubs of degree at least δ at vertex v_j. For this general case we have

$$\mathbb{P}(\exists i_1,\ldots,i_K: d_{j,f_{i_1}}\geq\delta,\ldots d_{j,f_{i_K}}\geq\delta\,|\mathscr{H}_0)=$$
$$\sum_{k'=K}^{n}\sum_{\substack{i_1<\ldots<i_{k'},i_{k'+1}<\ldots<i_n\\ \{i_1,\ldots,i_n\}=\{1,\ldots,n\}}}\prod_{l=1}^{k'}pv_{f_{i_l},\rho_{i_l},\delta}(j)\prod_{l'=k'+1}^{n}\left(1-pv_{f_{i_{l'}},\rho_{i_{l'}},\delta}(j)\right).$$

Experimental Results

Phase Transition Phenomenon and Mean Number of Hubs

We first performed numerical simulations to confirm Theorem 2 and Corollary 2 for complex-valued correlation screening. Samples were generated from p uncorrelated

complex Gaussian random variables. Figure 6 shows the number of discovered 1-hubs for $p = 1000$ and several sample sizes m. The plots from left to right correspond to $m = 2000, 1000, 500, 100, 50, 20, 10, 6$ and 4, respectively. The phase transition phenomenon is clearly observed in the plot. Table 1 shows the predicted value obtained from formula (33) for the critical threshold. As can be seen in Fig. 6, the empirical phase transition thresholds approximately match the predicted values of Table 1. Moreover, to confirm the accuracy of equation (31) in Corollary 2, we list the number of hubs for $m = 100$ in Table 2. The left column shows the empirical average number of hubs of degree at least $\delta = 1, 2, 3, 4$ in a network of i.i.d. complex Gaussian random variables. The numbers in this column are obtained by averaging 1000 independent experiments. The right column shows the predicted value of $\mathbb{E}[N_{\delta,\rho}]$ obtained via formula (31) with $J(\overline{f_{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*(\delta+1)}}}) = 1$ for the i.i.d. case. As we see the empirical and predicted values are close to each other.

Table 1 The value of critical threshold $\rho_{c,\delta}$ obtained from formula (33) for $p = 1000$ complex variables and $\delta = 1$. The predicted $\rho_{c,\delta}$ approximates the phase transition thresholds in Fig. 6.

m	2000	1000	500	100	50	20	10	6	4
$\rho_{c,\delta}$	0.05	0.07	0.10	0.24	0.35	0.56	0.78	0.94	0.99

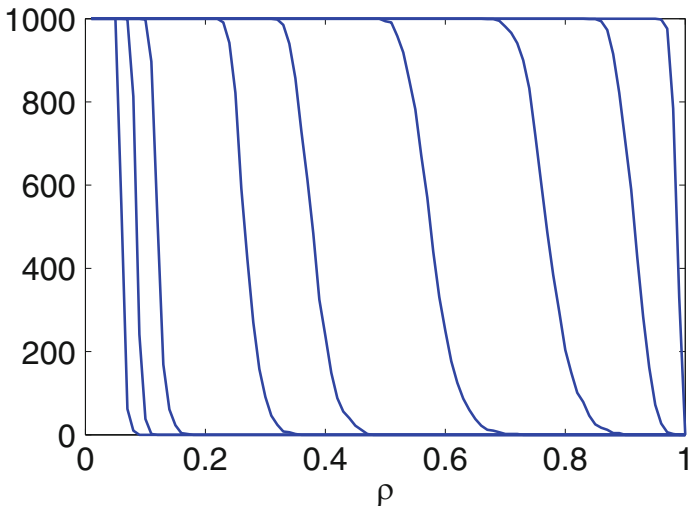


Fig. 6 Phase transition phenomenon: the number of 1-hubs in the sample correlation graph corresponding to uncorrelated complex Gaussian variables as a function of correlation threshold ρ. Here, $p = 1000$ and the plots from left to right correspond to $m = 2000, 1000, 500, 100, 50, 20, 10, 6$, and 4, respectively.

Table 2 Empirical average number of discovered hubs vs. predicted average number of discovered hubs in an uncorrelated complex Gaussian network. Here $p = 1000$, $m = 100$, $\rho = 0.28$. The empirical values are obtained by performing 1000 independent experiments.

degree threshold	empirical ($\mathbb{E}[N_{\delta,\rho}]$)	predicted ($\mathbb{E}[N_{\delta,\rho}]$)
$d_i \geq \delta = 1$	284	335
$d_i \geq \delta = 2$	45	56
$d_i \geq \delta = 3$	5	6
$d_i \geq \delta = 4$	0	0

Asymptotic Independence of Spectral Components for AR(1) Model

To illustrate the asymptotic independence property and convergence rate of Theorem 1, we considered the simple case of an AR(1) process,

$$X(k) = \varphi_1 X(k-1) + \varepsilon(k), \quad k \geq 1, \tag{34}$$

in which $X(0) = 0, \varphi_1 = 0.9$ and $\varepsilon(.)$ is a stationary Gaussian process with no temporal correlation and standard deviation 1. We performed Monte Carlo simulations to compute the correlation between spectral components at different frequencies for window sizes $n = 10, 20, \ldots, 250$. More specifically, we set $k = 1$ and $l = 2$ and empirically estimated $|\mathrm{cor}\,(Y(k), Y(l))\,|$ using 50000 Monte Carlo trials for each value of window size n. Figure 7 shows the result of this experiment. It is observable that the magnitude of $\mathrm{cor}\,(Y(k), Y(l))$ is bounded above by the function $10/n$. This observation is consistent with Theorem 1.

Spectral Correlation Screening of a Band-Pass Multivariate Time Series

Next we analyzed the performance of the proposed complex-valued correlation screening framework on a synthetic data set for which the expected results are known.

We synthesized a multivariate stationary Gaussian time series using the following procedure. Here we set $p = 1000, N = 12000$, and $m = n = 100$. The discrepancy between N and the product mn is explained below. Let $X(k), 0 \leq k \leq N-1$ be a sequence of i.i.d. zero-mean Gaussian random variables (i.e. white Gaussian noise) with standard deviation of 1. The p time series $X^{(1)}(k), \ldots, X^{(p)}(k), 0 \leq k \leq N-1$ are obtained from $X(k)$ by band-pass filtering and adding independent white Gaussian noise. Specifically,

$$X^{(i)}(k) = h_i(k) \star X(k) + N_i(k), \quad 1 \leq i \leq p, 0 \leq k \leq N-1,$$

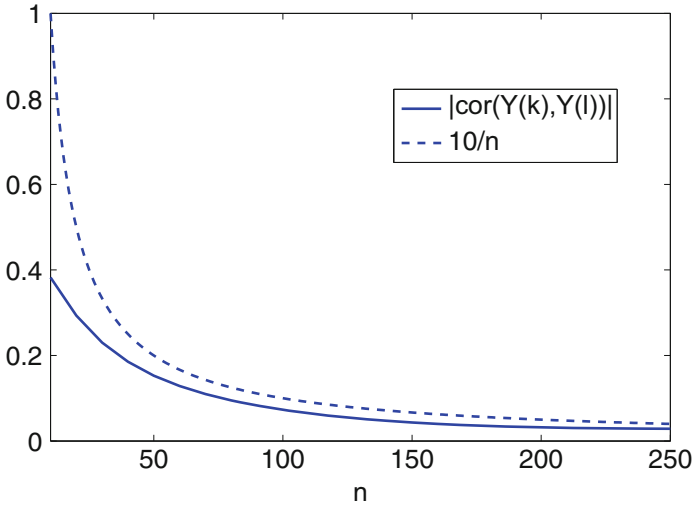


Fig. 7 Correlation coefficient $|\mathrm{cor}(Y(1),Y(2))|$ as a function of window size n, empirically estimated using 50000 Monte Carlo trials. Here $Y(.)$ is the DFT of the AR(1) process (34). The magnitude of the correlation for $n = 10, 20, \ldots, 250$ is bounded above by the function $10/n$. This observation is consistent with the convergence rate in Theorem 1.

in which $\star$ represents the convolution operator, $h_i(.)$ is the impulse response of the ith band-pass filter, and $N_i(.)$ is an independent white Gaussian noise series whose standard deviation is 0.1. Since stable filtering of a stationary series results in another stationary series, the obtained series $X^{(1)}(k), \ldots, X^{(p)}(k)$ are stationary and Gaussian. For $i = 10l, 1 \leq l \leq 50$, $h_i(k)$ is the impulse response of a band-pass filter with pass band $f \in [(4l-1)/400, 4l/400]$. We approximate the ideal band-pass filters with finite impulse response (FIR) Chebyshev filters [16]. Also for $i = 500 + 10l, 1 \leq l \leq 50$ we set $h_i(k) = h_{i-500}(k)$. For all of the other values of i (i.e., $i \neq 10l$) we set $h_i(k) = 0, 0 \leq k \leq N-1$.

Figure 8 shows the signal part of the time series (i.e., $h_i(k) \star X(k)$) for $i = 100, 200, 300, 400$. It is seen that the first 2000 samples of the signals reflect the transient response of the filters. These 2000 samples are not included for the purpose of correlation screening. Hence, the actual number of time samples considered is $mn = 10000$. Figure 9 shows the magnitude of the DFTs of the signals, $Y^{(i)}(k)$, for $i = 50, 100, \ldots, 500$. The band-pass structure of the signals is clearly observable in the figure.

We first constructed a correlation matrix for the time series $X^{(1)}(k), \ldots, X^{(p)}(k)$ from their simultaneous time samples. Figure 10 illustrates the structure of the thresholded sample correlation matrix and the corresponding correlation graph. Note that this is a real-valued correlation screening problem in the time domain. The correlation threshold used here is $\rho = 0.2$ which is well above the critical threshold $\rho_{c,1} = 0.028$ obtained via formula (10) in [9] for $p = 1000$ and $N = 10000$.

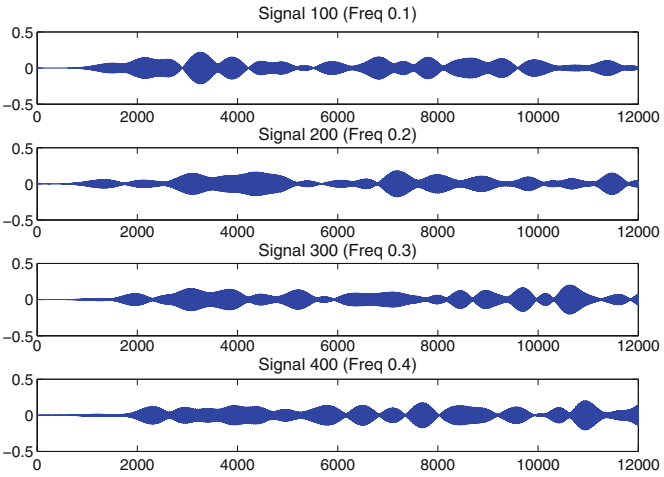


Fig. 8 Signal part of the band-pass time series $X^{(i)}(k)$ (i.e., $h_i(k) \star X(k)$) for $i = 100, 200, 300, 400$.

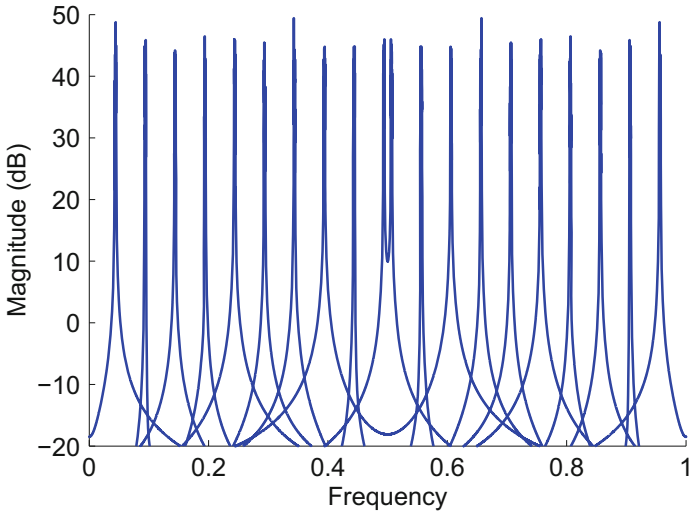


Fig. 9 DFT magnitude of the band-pass signals $h_i(k) \star X(k)$ (i.e., $20\log_{10}(|Y^{(i)}(.)|)$) as a function of frequency for $i = 50, 100, \ldots, 500$.

To examine the spectral structure of the correlations in Fig. 10, we then performed complex-valued correlation screening on the spectra of the time series $X^{(1)}(k), \ldots, X^{(p)}(k)$. Figure 11 shows the constructed correlation graphs $\mathscr{G}_{f,\rho}$ for $f = [0.1, 0.2, 0.3, 0.4]$ and correlation threshold $\rho = 0.9$, which corresponds to a $\delta = 1$ false positive rate $\mathbb{P}(N_{\delta,\rho} > 0) \approx 10^{-65}$ (using $\delta = 1$ in the asymptotic

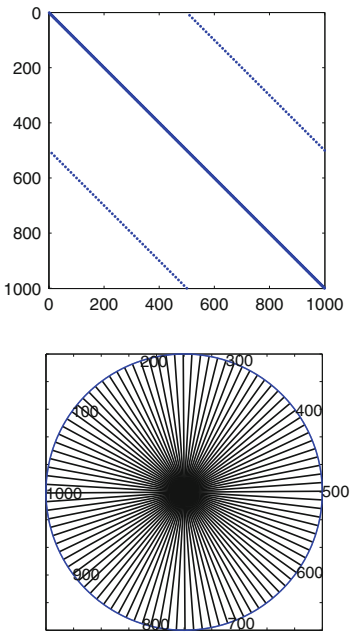


Fig. 10 (Left) The structure of the thresholded sample correlation matrix in the time domain. (Right) The correlation graph corresponding to the thresholded sample correlation matrix in the time domain.

relation (32) with $\Lambda_\infty = e^{\delta}_{m,\delta}/\delta!$ as specified by (31)). Note that the value of the correlation threshold is set to be higher than the critical threshold $\rho_c = 0.24$. It can be observed that performing complex-valued spectral correlation screening at each frequency correctly discovers the correlations between the time series which are active around that frequency. As an example, for $f = 0.2$ the discovered hubs (for $\delta = 1$) are the time series $X^{(i)}(k)$ for $i \in \{200, 700\}$. These time series are the ones that are active at frequency $f = 0.2$. Under the null hypothesis of diagonal covariance matrices, the p-values for the discovered hubs are of order 10^{-65} or smaller. These results show that complex-valued spectral correlation screening is able to resolve the sources of correlation between time series in the spectral domain.

Conclusion

This chapter presented a spectral method for correlation analysis of stationary multivariate Gaussian time series with a focus on identifying correlation hubs. The asymptotic independence of spectral components at different frequencies allows the problem to be decomposed into independent problems at each frequency, thus improving computational and statistical efficiency for high-dimensional time series.

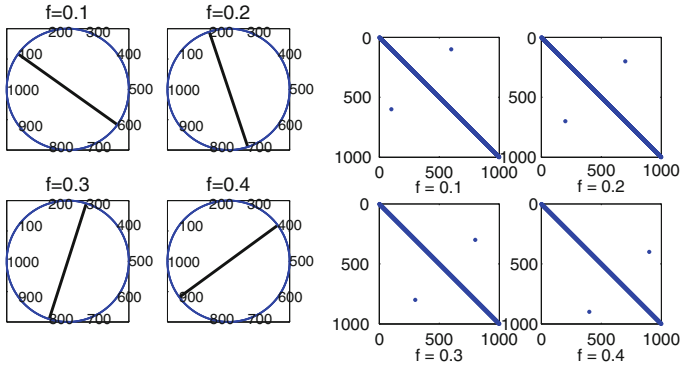


Fig. 11 Spectral correlation graphs $\mathscr{G}_{f,\rho}$ for $f = [0.1, 0.2, 0.3, 0.4]$ and correlation threshold $\rho = 0.9$, which corresponds to a false positive probability of 10^{-65}. The data used here are a set of synthetic time series obtained by band-pass filtering of a Gaussian white noise series with the band-pass filters shown in Fig. 9. As can be seen, complex correlation screening is able to extract the correlations at specific frequencies. This is not directly feasible in the time domain analysis.

The method of complex-valued correlation screening is then applied to detect hub variables at each frequency. Using a characterization of the number of hubs discovered by the method, thresholds for hub screening can be selected to avoid an excessive number of false positives or negatives, and the statistical significance of hub discoveries can be quantified. The theory specifically considers the high-dimensional case where the number of samples at each frequency can be significantly smaller than the number of time series. Experimental results validated the theory and illustrated the applicability of complex-valued correlation screening to the spectral domain.

Acknowledgment

This work was partially supported by AFOSR grant FA9550-13-1-0043.

References

1. M.C. Vuran, Ö.B. Akan, I.F. Akyildiz, Spatio-temporal correlation: theory and applications for wireless sensor networks. Comput. Netw. **45**(3), 245–259 (2004)
2. R. Paffenroth, P. du Toit, R. Nong, L. Scharf, A.P. Jayasumana, V. Bandara, Space-time signal processing for distributed pattern detection in sensor networks. IEEE J. Sel. Top. Sign. Process. **7**(1), 38–49 (2013)
3. K.J. Friston, J.T. Ashburner, S.J. Kiebel, T.E. Nichols, W.D. Penny, *Statistical Parametric Mapping: The Analysis of Functional Brain Images* (Academic, Boston, 2011)

4. P. Zhang, Y. Huang, S. Shekhar, V. Kumar, Correlation analysis of spatial time series datasets: a filter-and-refine approach, in *Advances in Knowledge Discovery and Data Mining*, ed. by K.-Y. Whang, J. Jeon, K. Shim, J. Srivastava (Springer, Berlin, 2003), pp. 532–544
5. R.S. Tsay, *Analysis of Financial Time Series*, vol. 543 (Wiley, Hoboken, 2005)
6. M. Stanley, S. Gervais-Ducouret, J. Adams, Intelligent sensor hub benefits for wireless sensor networks, in *Sensors Applications Symposium (SAS), 2012 IEEE*. IEEE (2012), pp. 1–6
7. Y. Li, M.T. Thai, W. Wu, *Wireless Sensor Networks and Applications* (Springer, Berlin, 2008)
8. E. Bullmore, O. Sporns, Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. **10**(3), 186–198 (2009)
9. A. Hero, B. Rajaratnam, Hub discovery in partial correlation graphs. IEEE Trans. Inf. Theory **58**(9), 6064–6078 (2012)
10. X. Chen, M. Xu, W.B. Wu et al., Covariance and precision matrix estimation for high-dimensional time series. Ann. Stat. **41**(6), 2994–3021 (2013)
11. A. Hero, B. Rajaratnam, Large-scale correlation screening. J. Am. Stat. Assoc. **106**(496), 1540–1552 (2011)
12. H. Firouzi, B. Rajaratnam, A. Hero, Predictive correlation screening: application to two-stage predictor design in high dimension, in *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, (2013)
13. R. Durrett, *Probability: Theory and Examples*, vol. 3 (Cambridge University Press, Cambridge, 2010)
14. U. Grenander, G. Szegő, *Toeplitz Forms and Their Applications* (University of California Press, Berkeley, 1958)
15. R.M. Gray, Toeplitz and circulant matrices: a review. Found. Trends Commun. Inf. Theory **2**(3), 155–239 (2006). doi:10.1561/0100000006. http://www.dx.doi.org/10.1561/0100000006
16. A.V. Oppenheim, R.W. Schafer, J.R. Buck et al., *Discrete-Time Signal Processing*, vol. 2 (Prentice-Hall, Englewood Cliffs, 1989)
17. J.B. Conway, *A Course in Functional Analysis*, vol. 96 (Springer, Berlin, 1990)
18. J.D. Hamilton, *Time Series Analysis*, vol. 2 (Princeton University Press, Princeton, 1994)
19. A.C. Micheas, D.K. Dey, K.V. Mardia, Complex elliptical distributions with application to shape analysis. J. Stat. Plann. Inference **136**(9), 2961–2982 (2006)
20. M.K. Simon, *Probability Distributions Involving Gaussian Random Variables: A Handbook for Engineers and Scientists* (Springer, Berlin/New York, 2007)
21. R. Arratia, L. Goldstein, L. Gordon, Poisson approximation and the Chen-Stein method. Stat. Sci. **5**(4), 403–424 (1990)

A Spectral Analysis Approach for Experimental Designs

R.A. Bailey, Persi Diaconis, Daniel N. Rockmore, and Chris Rowley

Abstract In this paper we show how the approach of spectral analysis generalizes the standard ANOVA-based techniques for studying data from designed experiments. Several examples are worked out in detail, including a thorough analysis of Calvin's famous ice cream data.

Key words: analysis of variance, block design, designed experiment, diallel experiment, irreducible subspace, orthogonal decomposition, permutation representation

Introduction

Designed experiments are used in many fields of enquiry. Government scientists may wish to compare the effects of different insecticides (including no insecticide) on colonies of bumble bees close to the fields where the insecticides are sprayed.

R.A. Bailey
School of Mathematics and Statistics, University of St Andrews, St Andrews, Fife, KY16 9SS, UK

School of Mathematical Sciences, Queen Mary University of London (Emerita), Mile End Road, London E1 4NS, UK
e-mail: rab@mcs.st-and.ac.uk

P. Diaconis
Department of Statistics, Sequoia Hall, 390 Serra Mall, Stanford University, Stanford, CA 94305-4065, USA
e-mail: diaconis@math.stanford.edu

D.N. Rockmore (✉)
Department of Mathematics and Neukom Institute for Computational Science, Dartmouth College, Hanover, NH 03755-3551, USA
e-mail: rockmore@cs.dartmouth.edu

C. Rowley
Faculty of Mathematics, Computing and Technology, Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, Buckinghamshire, MK7 6AA, UK

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_14

Pharmaceutical companies experiment with new drugs, in various doses, to cure certain diseases, or alleviate symptoms. Psychologists may run a trial to see which of three teaching methods is most effective in helping autistic children to understand emotions in other people. In the manufacturing industry, there are frequent experiments to investigate how the process may be improved by changing the raw materials, changing their quantities, or altering parts of the process. (See [36] for examples and a classical statistical approach.)

For all of these, the items being compared, such as insecticides, drugs, teaching methods, or raw materials, are called *treatments*. There may be a single treatment factor, or treatments may consist of all combinations of two or more factors: for example, five varieties of cow-peas with three different methods of cultivation, giving fifteen treatments altogether. One of the treatments may also be "untreated".

In order to compare the treatments, they have to be applied to something or somebody: for example, a treatment may be applied to a field, a whole farm, an ill person for a certain amount of time, a child, a group of children, one part of the factory for a month, and so on. Measurements are made on these, either on each whole item, or on smaller units, such as each child in the class. We call these *observational units*.

We now formalize these ideas. In a designed experiment, a finite set Γ of treatments is applied to a finite set Ω of observational units. A measurement y_ω is made on each unit ω in Ω, thus giving a data vector y in the vector space $\mathbb{R}^\Omega$ of all real functions on Ω. Twin problems are how to design the experiment and how to analyze the data that it produces.

There is a long history of group theory being used to develop and design the combinatorial structures used in such experiments (see, for example, [1, 26]). As proposed by Diaconis in [23] and the many references therein, *spectral analysis* (that is, Fourier analysis for the symmetry group of choice) is a non-model-based approach to analyzing data that may be carried out in the presence a natural symmetry group. Spectral analysis seeks to approximate the data vector as a sum of its projections into orthogonal, symmetry-invariant, and naturally interpretable subspaces of $\mathbb{R}^\Omega$, where orthogonality is with respect to the standard inner product. The paper [24] contains several spectral analysis examples as well as other examples of ways in which group theory enters statistical analysis.

In this paper we explore some new ideas relating to the spectral analysis of data from a designed experiment. We connect it to classical theory and show how, in a number of cases, this approach provides new information. Necessarily, this analysis depends on the representation theory of the associated symmetry group, and the attendant calculations depend on certain representation (Fourier) theoretic computations and algorithms. All of this is developed as we go along; the book [23] contains all the necessary representation-theoretic background.

In Section "Treatments" we look at families of subspaces of $\mathbb{R}^\Gamma$ that may be used to model expectation, introducing several examples. In Section "Treatment Permutations" we introduce a group of permutations of Γ, and compare the decomposition into irreducible subspaces with that obtained from the modeling approach.

Section "Observational Units" takes a similar approach to $\mathbb{R}^\Omega$, where structure on Ω might suggest models for the expectation of a random vector Y on Ω or for its variance-covariance matrix $\mathrm{Var}(Y)$. Again, these may be linked to a group of permutations of Ω.

The following sections combine these decompositions of $\mathbb{R}^\Gamma$ and $\mathbb{R}^\Omega$ when an allocation of treatments to observational units effectively makes $\mathbb{R}^\Gamma$ a subspace of $\mathbb{R}^\Omega$. The overall philosophy of analysis of variance is summarized in Section "Analysis of Variance". In the most straightforward case, treated in Section "The Orthogonal Case", there is geometric orthogonality between all subspaces. Otherwise, as explained in Section "The General Non-Orthogonal Case", more complicated algebra is needed. In Section "Incomplete-Block Designs" we restrict attention to incomplete-block designs, which avoid some of the complications of the general case while still showing interesting behavior.

Section "Ice Cream Data" introduces the subgroup of both previous groups that preserves structure on Γ and Ω as well as preserving the embedding of $\mathbb{R}^\Gamma$ in $\mathbb{R}^\Omega$. An example discussed in depth shows the utility of the approach from spectral analysis.

Finally, Sections "Strong symmetries of orthogonal designs" and "Strong symmetries of incomplete-block designs" consider this subgroup in the contexts of orthogonal designs and incomplete-block designs. This subgroup is usually smaller than the previous ones, so its decomposition may have more subspaces. What meaning can we attach to them?

Treatments

In this section we ignore the observational units, and pretend temporarily that there is only one observation on each treatment. Different structure on the set of treatments can lead to different plausible models for the response. Under suitable conditions, a family of models leads to an orthogonal decomposition of the vector space of real functions on the treatments.

Let Γ be the finite set of treatments. A *linear model* is typically a subspace V of $\mathbb{R}^\Gamma$: if a measurement is made on each treatment, then we expect the resulting vector of measurements to lie in V or close to V. In fact, linear models are frequently presented in notation that is shorthand for saying that a whole (finite) family $\mathscr{F}$ of subspaces is being considered. It is usual to assume that this family is closed under intersection and under vector-space summation.

Two subspaces V_1 and V_2 are defined in [54] to be *geometrically orthogonal* to each other if $V_1 \cap (V_1 \cap V_2)^\perp$ is orthogonal to $V_2 \cap (V_1 \cap V_2)^\perp$ (here $^\perp$ denotes orthogonal complement). If $\mathscr{F}$ is closed under $\cap$ and $+$ and every pair of subspaces in $\mathscr{F}$ is geometrically orthogonal, then there is a collection of pairwise orthogonal subspaces $\{W_j : j \in J\}$ such that $\sum_{V \in \mathscr{F}} V = \bigoplus_{j \in J} W_j$ and every subspace in $\mathscr{F}$ is a direct sum of some of the spaces W_j. This is called *orthogonal treatment structure* in [5, 6]. Not all sums of W_j spaces need occur in $\mathscr{F}$.

Denote by V_0 the 1-dimensional subspace consisting of constant vectors. It is usually assumed that V_0 belongs to $\mathscr{F}$.

Example 1. Suppose that Γ consists of all combinations of the n levels of treatment factor C with the m levels of treatment factor D. For example, factor C might be

three different non-zero quantities of aspirin and factor D might give two different times of day for taking the aspirin.

One obvious model subspace is V_C (of dimension n), which consists of all vectors which are constant on each level of C. This model is appropriate when the factor D has no effect. The m-dimensional subspace V_D is defined similarly. Then $V_C \cap V_D = V_0$ and V_C is geometrically orthogonal to V_D. The subspace $V_C + V_D$ is called the *additive model*. If this is appropriate, then the difference between two given levels of C does not depend on the level of D. Finally, the whole nm-dimensional space $\mathbb{R}^\Gamma$ is the *full model*, allowing unrelated measurements on all treatments.

Figure 1 shows the Hasse diagram for this family of subspaces. The dimension of each is shown beside the corresponding dot.

Put $W_0 = V_0$, $W_C = V_C \cap V_0^\perp$, $W_D = V_D \cap V_0^\perp$, and $W_{CD} = (V_C + V_D)^\perp$. These spaces are called the *grand mean*, the *main effect* of C, the *main effect* of D, and the C-by-D *interaction*, respectively. Strictly speaking, it is the orthogonal projection of the vector of measurements onto each W-subspace that has this name. Now $\dim(W_0) = 1$, $\dim(W_C) = n-1$, $\dim(W_D) = m-1$ and $\dim(W_{CD}) = (n-1)(m-1)$. These subspaces are mutually orthogonal. Furthermore, $V_0 = W_0$, $V_C = W_0 \oplus W_C$, $V_D = W_0 \oplus W_D$ and $\mathbb{R}^\Gamma = W_0 \oplus W_C \oplus W_D \oplus W_{CD}$.

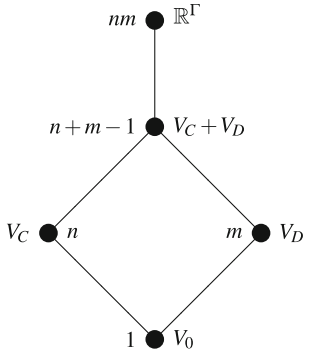


Fig. 1 Hasse diagram of subspaces in Example 1

Example 2. The half-diallel. Let Γ consist of all unordered pairs from a set of size n. This occurs in so-called *half-diallel* experiments in plant breeding, where $\{i, j\}$ denotes the cross between parental lines i and j (see [28]). Another example with $n = 5$ is an experiment to compare all fruit-juices made from equal quantities of two of orange, grapefruit, mango, pineapple, and passionfruit, to find the effect on the drinker's blood-pressure.

The family of model subspaces consists of V_0, V_1 and $\mathbb{R}^\Gamma$, where V_1 consists of all functions f of the form $f(\{i, j\}) = \psi_i + \psi_j$. Figure 2 gives the Hasse diagram. Since each pair of subspaces is related by inclusion, geometric orthogonality is assured,

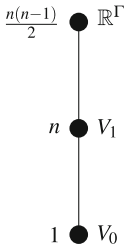


Fig. 2 Hasse diagram of subspaces in Example 2

and the whole space is decomposed as $W_0 \oplus W_1 \oplus W_2$, where $W_0 = V_0$, $W_1 = V_1 \cap V_0^\perp$ and $W_2 = V_1^\perp$. The last two are called *general combining ability* and *specific combining ability*, respectively. Now $\dim(W_0) = 1$, $\dim(W_1) = n-1$ and $\dim(W_2) = n(n-3)/2$.

Example 3. Now let Γ consist of all ordered pairs of distinct elements from a set of size n. This occurs in breeding if the gender of the parents is relevant. It occurs in a modification of the fruit-juice example if the treatments consist of all instructions like "drink orange juice at breakfast and mango juice at lunch".

One model subspace is the space V_S of *symmetric* functions f, for which $f((i,j)) = f((j,i))$ for all i and j. Another is the space V_A of *antisymmetric* functions f, for which $f((i,j)) = -f((j,i))$ for all i and j. A further obvious space V_P consists of parental effects: f is in V_P if there are constants α_i and β_j such that $f((i,j)) = \alpha_i + \beta_j$ for all i and j, which is similar to the additive model in Example 1.

There are (at least) four interesting subspaces of V_P: $f \in V_P \cap V_S$ if $\alpha_i = \beta_i$ for all i; $f \in V_P \cap V_A$ if $\alpha_i = -\beta_i$ for all i; $f \in V_1$ if $\beta_j = 0$ for all j; and $f \in V_2$ if $\alpha_i = 0$ for all i. Now the four subspaces $V_P \cap V_S \cap V_0^\perp$, $V_P \cap V_A$, $V_1 \cap V_0^\perp$ and $V_2 \cap V_0^\perp$ all have dimension $n-1$; the sum of any two is $V_P \cap V_0^\perp$; and no pair is geometrically orthogonal except for the first two.

On the other hand, $V_P^\perp$ is the orthogonal direct sum of W_S and W_A, where $W_S = V_S \cap V_P^\perp$, which has dimension $n(n-3)/2$ and is analogous to W_2 in Example 2, and $W_A = V_A \cap V_P^\perp$, which has dimension $(n-1)(n-2)/2$. See Figure 3.

Now the lack of a canonical orthogonal decomposition of $V_P \cap V_0^\perp$ can lead to difficulties in model choice.

Example 4. Suppose that we wish to compare six makes of strawberry ice cream. Sixty people take part in the experiment, so that each tastes four makes and rates one of these, giving it a score out of 100. Note that each make is tasted in the presence of all possible triples of other makes. Thus Γ consists of the 60 pairs $(i, \{j,k,l\})$ where i is rated in the presence of j, k, and l. One model subspace V consists of functions f for which $f((i,\{j,k,l\})) = \gamma_i + \theta_{ij} + \theta_{ik} + \theta_{il}$, where $\theta_{ij}, \theta_{ik}, \theta_{il}$ each account for the effect of tasting i in the presence of j, k and l. Note that V has dimension 30 because the subspace V_1 with $\theta_{ij} = \alpha_i$ for all i and j is the same as the subspace with $\theta_{ij} = 0$

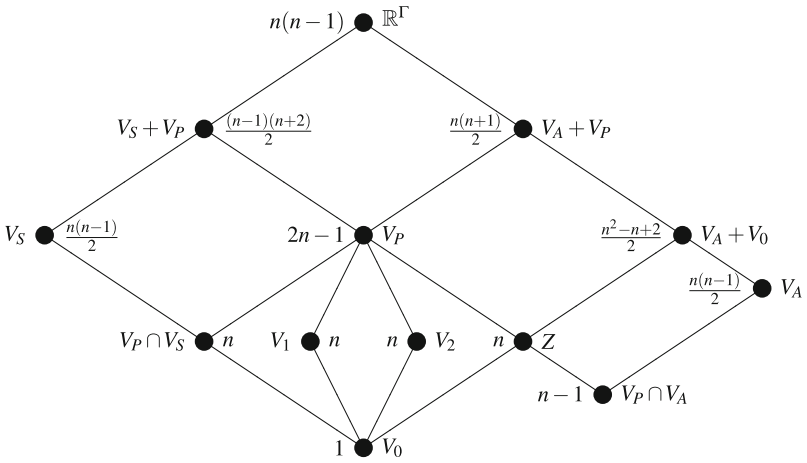


Fig. 3 Hasse diagram of subspaces in Example 3: Z is $(V_P \cap V_A) + V_0$

for all i and j. If we consider only subspaces of V, we appear to obtain the same family of models as in Example 3; however, their interpretation as subspaces of $\mathbb{R}^\Gamma$ is different.

Figure 4 represents the elements of Γ in Examples 3 and 4 when $n = 6$. Elements of Γ are identified by the labels of the rows and columns with "$\times$" marks those cells for pairings that do not occur as elements. The real numbers in the other cells thus represent a function in $\mathbb{R}^\Gamma$: in both cases it is the function in $V_P \cap V_A$ with $\alpha_1 = -\beta_1 = 1$ and $\alpha_i = \beta_i = 0$ if $i \neq 1$. In Example 3, this function is orthogonal to $V_S \cap V_P$; in Example 4, it is not. In Example 3, there seems to be no reason to prefer the orthogonal decomposition $(V_1 \cap V_0^\perp) \oplus (V_P \cap V_1^\perp)$ over $(V_2 \cap V_0^\perp) \oplus (V_P \cap V_2^\perp)$, whereas in Example 4 the explicit inclusion of γ_i in the formula seems to favor the former. In Example 4 there is a function h in $V_1 \cap V_0^\perp$ with $\gamma_1 = -10/3$ and $\gamma_i = 2/3$ if $i \neq 1$: when this is added to the function f shown in Figure 4(**b**), the result is constant on each column.

Treatment Permutations

Often, the set Γ of treatments has some combinatorial symmetry which is preserved by a group G_1 of permutations of Γ. In this section we see how this may be used to derive an orthogonal decomposition of $\mathbb{R}^\Gamma$. We write the elements of G_1 on the right of their arguments, so that composition is done from left to right.

The permutation representation of G_1 associated with its action on Γ is an isomorphism ρ_1 from G_1 to group of permutation matrices in $\mathbb{R}^{\Gamma\times\Gamma}$, whose rows and columns are indexed by Γ: for g in G_1, the (α,β)-entry of $\rho_1(g)$ is 1 if $\alpha g = \beta$ and is 0 otherwise. These matrices act on $\mathbb{R}^\Gamma$, which can be decomposed as a direct sum

	1	2	3	4	5	6
1	×	1	1	1	1	1
2	−1	×	0	0	0	0
3	−1	0	×	0	0	0
4	−1	0	0	×	0	0
5	−1	0	0	0	×	0
6	−1	0	0	0	0	×

a

	1 2 3 4	1 2 3 5	1 2 3 6	1 2 4 5	1 2 4 6	1 2 5 6	1 3 4 5	1 3 4 6	1 3 5 6	1 4 5 6	2 3 4 5	2 3 4 6	2 3 5 6	2 4 5 6	3 4 5 6
1	3	3	3	3	3	3	3	3	3	3	×	×	×	×	×
2	−1	−1	−1	−1	−1	−1	×	×	×	×	0	0	0	0	×
3	−1	−1	−1	×	×	×	−1	−1	−1	×	0	0	0	×	0
4	−1	×	×	−1	−1	×	−1	−1	×	−1	0	0	×	0	0
5	×	−1	×	−1	×	−1	−1	×	−1	−1	0	×	0	0	0
6	×	×	−1	×	−1	−1	×	−1	−1	−1	×	0	0	0	0

b

Fig. 4 A function in $\mathbb{R}^{\Gamma}$: (**a**) Example 3 (**b**) Example 4

of subspaces which are invariant under G_1 and irreducible under G_1. The centralizer algebra $\mathscr{C}(G_1)$ of G_1 is the set of matrices in $\mathbb{R}^{\Gamma\times\Gamma}$ which commute with $\rho_1(g)$ for all g in G_1.

Recall that, in general, a representation of degree N of a group G over a field F is just a homomorphism from G to the general linear group $\mathrm{GL}(N,F)$, where N is a non-negative integer. Two such representations are *equivalent* if and only if they differ by a change of basis. Thus, the trace function is an invariant of any equivalence class of representations. A representation is *irreducible* if and only if it is not equivalent to a direct sum of (non-trivial) representations. Up to equivalence, there are only a finite number of irreducible representations of a given group over a given field. Following community standards, we refer to both the collection of matrices and the underlying vector space with associated group action as "the representation".

Since every representation over $\mathbb{R}$ can also be considered to be a representation over $\mathbb{C}$, we sometimes need to write words like "real-irreducible" or "complex-irreducible" to make clear which field we are talking about. General theory is usually expressed over the complex numbers, giving decompositions of $\mathbb{C}^{\Gamma}$. However, only $\mathbb{R}^{\Gamma}$ occurs for actual data so we use real representations. See Section "Strong symmetries of incomplete-block designs" for the modification from $\mathbb{C}^{\Gamma}$ to $\mathbb{R}^{\Gamma}$.

Let π_1 be the permutation character of G_1, so that $\pi_1(g) = \mathrm{trace}(\rho_1(g))$ for g in G_1. Let $\{\chi_i : i \in \mathscr{I}\}$ be the real-irreducible characters of G_1. Then there are non-negative integers m_i for i in $\mathscr{I}$ such that $\pi_1 = \sum_i m_i\chi_i$. If $m_i > 0$, then there is a corresponding *homogeneous* subspace U_i of $\mathbb{R}^{\Gamma}$: it has dimension $m_i \deg(\chi_i)$; it is G_1-invariant; it is orthogonal to U_j if $i \neq j$. If $m_i = 1$, then U_i is G_1-irreducible; otherwise, it can be decomposed as a direct sum of m_i irreducible subspaces, all admitting isomorphic actions of G_1, in infinitely many ways.

If $m_i \in \{0,1\}$ for all i in $\mathscr{I}$, then π_1 is said to be *real-multiplicity-free*. Then $\mathbb{R}^{\Gamma}$ has a unique decomposition as a direct sum of orthogonal G_1-irreducible subspaces. We may be able to use this decomposition to explain the data vector. Otherwise, there is a choice of such decompositions.

What light does this approach from representation theory throw on the previous examples? In Example 1, we may consider Γ to be an $n \times m$ rectangle. Then we may take G_1 to be $\mathrm{Sym}(n) \times \mathrm{Sym}(m)$ (where $\mathrm{Sym}(n)$ denotes the symmetric group on n letters) in its product action, so that $(\alpha,\beta)(g,h) = (\alpha g, \beta h)$. Its permutation character is real-multiplicity-free, and the corresponding irreducible subspaces are precisely the subspaces W_0, W_C, W_D, and W_{CD} used by statisticians.

In Example 2, it is natural to take G_1 to be $\mathrm{Sym}(n)$ in its action on unordered pairs, so that $\{\alpha,\beta\}g = \{\alpha g, \beta g\}$. In the notation of [35], the vector space supporting the permutation representation is denoted $M^{n-2,2}$, which has a unique G_1-irreducible decomposition as $S^n \oplus S^{n-1,1} \oplus S^{n-2,2}$. In the notation of Example 2, S^n is the representation on W_0, $S^{n-1,1}$ is the representation on W_1, and $S^{n-2,2}$ is the representation on W_2.

In Example 3, we take G_1 to be $\mathrm{Sym}(n)$ in its action on ordered pairs, so that $(\alpha,\beta)g = (\alpha g, \beta g)$. The resulting permutation representation is denoted $M^{n-2,1,1}$. The decomposition (see [35]) of this representation is below.

	$M^{n-2,1,1}$	$=$	S^n	$\oplus$	$2S^{n-1,1}$	$\oplus$	$S^{n-2,2}$	$\oplus$	$S^{n-2,1,1}$
dim	$n(n-1)$		1		$2 \times (n-1)$		$\frac{n(n-3)}{2}$		$\frac{(n-1)(n-2)}{2}$

Here S^n, $S^{n-2,2}$, and $S^{n-2,1,1}$ are the representations on the subspaces W_0, W_S, and W_A, respectively. The notation $2S^{n-1,1}$ denotes a representation which is the direct sum of two representations isomorphic to $S^{n-1,1}$. Note that it is such a sum in infinitely many ways, and there is no canonical decomposition of the corresponding homogenous subspace of dimension $2(n-1)$, which is precisely $V_P \cap V_0^\perp$. Thus the group theory reinforces the previous discussion.

In [57], Yates proposed decomposing $V_P \cap V_0^\perp$ as $(V_P \cap V_S \cap V_0^\perp) \oplus (V_P \cap V_A)$. Using the above representation theory in [32], James was able to show that this was, in some sense, an arbitrary choice. Fortini gave a different decomposition in [27].

In Example 4, let G_1 be $\mathrm{Sym}(6)$ in its action on pairs (i,K) where K is a 4-subset of a 6-set and $i \in K$. Such a pair is equivalent to the partition with parts $\{i\}, K \setminus \{i\}$ and the complement of K, so the resulting permutation representation is $M^{3,2,1}$, whose decomposition (see [35]) is below.

	$M^{3,2,1}$	$=$	S^6	$\oplus$	$2S^{5,1}$	$\oplus$	$2S^{4,2}$	$\oplus$	$S^{4,1,1}$	$\oplus$	$S^{3,3}$	$\oplus$	$S^{3,2,1}$
dim	60		1		2×5		2×9		10		5		16.

(1)

Of course, S^6 is the representation on V_0. We have already seen that $V_P \cap V_0^\perp$ is the homogeneous subspace for $S^{5,1}$, that W_S affords one copy of $S^{4,2}$ and W_A affords $S^{4,1,1}$. The action of $\mathrm{Sym}(6)$ on 4-subsets (the column indices in Fig. 4(**b**)) is $M^{4,2}$, whose decomposition is $S^6 \oplus S^{5,1} \oplus S^{4,2}$. Thus the 15-dimensional subspace V_B of functions which are constant on each 4-subset includes a 5-dimensional subspace affording $S^{5,1}$, which must therefore be contained in $V_P \cap V_0^\perp$. Indeed, the end of Section "Treatments" gives a non-zero function in $V_P \cap V_0^\perp$ which is constant on each 4-subset.

Therefore $V_B \cap V_P^{\perp}$ is a 9-dimensional subspace affording $S^{4,2}$. How is this related to W_S? Figure 5 displays the function in W_S defined by $\theta_{ij} = 1$ if $\{i,j\} = \{1,2\}$ or $\{3,4\}$, $\theta_{ij} = -1$ if $\{i,j\} = \{1,3\}$ or $\{2,4\}$, and $\theta_{ij} = 0$ otherwise. This is in neither V_B nor $V_B^{\perp}$.

We return to this example in detail in Section "Ice Cream Data".

	12/34	12/35	12/36	12/45	12/46	12/56	13/45	13/46	13/56	14/56	23/45	23/46	23/56	24/56	34/56
1	0	0	0	1	1	1	−1	−1	−1	0	×	×	×	×	×
2	0	1	1	0	0	1	×	×	×	×	−1	−1	0	−1	×
3	0	−1	−1	×	×	×	0	0	−1	×	1	1	0	×	1
4	0	×	×	−1	−1	×	1	1	×	0	0	0	×	−1	1
5	×	0	×	0	×	0	0	×	0	0	0	×	0	0	0
6	×	×	0	×	0	0	×	0	0	0	×	0	0	0	0

Fig. 5 A function in W_S which is neither constant on columns nor orthogonal to columns

Observational Units

Now we temporarily ignore the treatments, and think about the observational units to be used in the experiment. Again, there is a corresponding real vector space, and we seek meaningful orthogonal decompositions of this. Such a decomposition may be defined by inherent factors or by a group of symmetries.

In a designed experiment, there is a finite set Ω of observational units to which treatments are applied; later, some response is measured on each unit. Even before the treatments are applied, inherent features of Ω may suggest something about the pattern of the response.

Example 5. An experiment comparing different methods of soil preparation for a single cereal variety might use k fields on each of b farms, with a single method on each field. Then Ω consists of the bk fields. Let Y be the hypothetical random vector of responses. If some farms produce consistently better results than others, then, in the absence of treatment differences, the expected value of the response should just depend on the farm. On the other hand, differences between farms may change from season to season, but then it is plausible that fields within a farm are more alike than fields in different farms. In this case, the grouping of fields within farms affects the covariance (matrix) $\mathrm{Var}(Y)$.

More generally, if Ω consists of bk observational units grouped into b blocks of size k, let V_B be the subspace of $\mathbb{R}^{\Omega}$ consisting of vectors which are constant on each block. The first approach assumes that $E(Y) \in V_B$. In this case, blocks are said to have *fixed effects*. It is usual to assume that $\mathrm{Var}(Y) = \sigma^2 I$ in this case. The second approach assumes that $E(Y) \in V_0$ and that the covariance has the form $\mathrm{Var}(Y) = \sigma^2 I + k\sigma_B^2 Q_B$, where Q_B is the matrix of orthogonal projection onto V_B. Blocks then

are said to have *random effects*. In this case, the eigenspaces of $\mathrm{Var}(Y)$ are V_B and $V_B^\perp$. Both approaches make it natural to consider the decomposition $V_0 \oplus W_B \oplus V_B^\perp$ of $\mathbb{R}^\Omega$, where $W_B = V_B \cap V_0^\perp$.

Let G_2 be a group of permutations of Ω that preserve structure on Ω, such as the partition into blocks, before treatments are allocated. Let π_2 be the permutation character of G_2. If π_2 is real-multiplicity-free, then there is a unique decomposition of $\mathbb{R}^\Omega$ as a sum of G_2-irreducible subspaces, which may be pertinent for data analysis. Even without this uniqueness, a decomposition into G_2-invariant subspaces may give insight into the data.

When Ω consists of b blocks of size k, we may take G_2 to be the wreath product $\mathrm{Sym}(k)\mathrm{wr}\mathrm{Sym}(b)$ in its imprimitive action. This has a subgroup $\mathrm{Sym}(k)$ for each block, permuting just the units within it, and a subgroup $\mathrm{Sym}(b)$ permuting the set of whole blocks. The action is real-multiplicity-free, with irreducibles V_0, W_B and $V_B^\perp$.

Example 6. Another common structure for Ω is a rectangle with r rows and c columns. This may be an actual physical rectangle on the ground, or an abstract one where, for example, rows represent time-periods and columns represent people. As in Example 1, $\mathbb{R}^\Omega$ has a natural decomposition as $W_0 \oplus W_R \oplus W_C \oplus W_{RC}$. If rows and columns have fixed effects, then $E(Y) \in W_0 \oplus W_R \oplus W_C$. If they have random effects, then $\mathrm{Var}(Y) = \sigma^2 I + c\sigma_R^2 Q_R + r\sigma_C^2 Q_C$, whose eigenspaces are W_0, W_R, W_C and W_{RC}. These four subspaces are the irreducibles of $\mathrm{Sym}(r) \times \mathrm{Sym}(c)$ in its product action.

The key part of an experimental design is the function $\tau\colon \Omega \to \Gamma$, which allocates treatment $\tau(\omega)$ to observational unit ω. This allocation is normally *randomized* before treatments are applied: a permutation g is chosen at random from a suitable group G_2 of permutations of Ω, and τ is replaced by τ^g, where $\tau^g(\omega) = \tau(\omega g^{-1})$. It is argued in [1, 4] that this justifies the assumption that $\mathrm{Var}(Y) \in \mathscr{C}(G_2)$. Since $\mathrm{Var}(Y)$ is symmetric, its eigenspaces form a G_2-invariant decomposition of $\mathbb{R}^\Omega$. If π_2 is real-multiplicity-free, these subspaces are known even if the corresponding eigenvalues are not.

Hannan [29] and Speed [52, 53] considered random variables Y such that $\mathrm{Var}(Y) \in \mathscr{C}(G_2)$ without the complication of Γ and τ. See those papers for more examples.

The three most common inherent structures in designed experiments are the two that we have discussed and the unstructured one, in which G_2 consists of all permutations of the units in Ω. The operations of nesting (units within blocks) and crossing (rows and columns) can be iterated, to give simple orthogonal block structures: see [43]. Their automorphism groups are generalized wreath products of symmetric groups, for all of which the relevant action is real-multiplicity-free: see [8]. The remaining examples in this paper use only these three common structures.

Analysis of Variance

Analysis of variance is often the statistician's first step towards analyzing the data vector. See [50] for an extensive study. This section gives a quick summary of the method, in language more familiar to algebraists.

Classical analysis of variance (ANOVA) depends on a given orthogonal decomposition of $\mathbb{R}^\Omega$ into subspaces W_j for j in some set J. Denote by P_j the linear operator of orthogonal projection onto W_j. Then the data vector y is the sum $\sum_{j\in J} P_j y$, and $P_j y$ is orthogonal to $P_i y$ if $i \neq j$. The *sum of squares* SS_j for W_j is defined to be $\|P_j y\|^2$, which is just $y^\top P_j y$. Since $I = \sum_{j\in J} P_j$, these sums of squares add to give $y^\top y$, the total sum of squares.

Put $d_j = \dim(W_j)$. The *mean square* MS_j for W_j is defined to be SS_j/d_j. If the data are purely random, in the sense that the responses are mutually independent random variables with the same expectation and variance, then all mean squares except MS_0 have the same expectation, where $W_0 = V_0$. More precisely, if Y is a random vector on Ω and $\mathrm{Var}(Y) = \sigma^2 I$, then $E(\mathrm{MS}_j) = \|E(P_j Y)\|^2/d_j + \sigma^2$. In general, if W_j is contained in an eigenspace of $\mathrm{Var}(Y)$ with eigenvalue ξ_i, then $E(\mathrm{MS}_j) = \|E(P_j Y)\|^2/d_j + \xi_i$.

Thus one approach to analyzing data is to calculate the mean squares and pick out those subspaces W_j whose mean squares are particularly large relative to the others: something interesting must be happening there. Part of the statistical theory of hypothesis testing is quantifying "particularly large". We do not go into details here.

The treatment allocation τ gives a subspace V_Γ of $\mathbb{R}^\Omega$ whose elements are constant on each treatment. Thus V_Γ is isomorphic to $\mathbb{R}^\Gamma$. To avoid complications, we assume that every treatment occurs on the same number of observational units. This ensures that any orthogonal decomposition of $\mathbb{R}^\Gamma$ into subspaces remains orthogonal when $\mathbb{R}^\Gamma$ is embedded into $\mathbb{R}^\Omega$ as V_Γ.

Here we assume that $\mathbb{R}^\Omega$ has a given G_2-invariant orthogonal decomposition and $\mathbb{R}^\Gamma$ has a given G_1-invariant orthogonal decomposition. When $\mathbb{R}^\Gamma$ is embedded in $\mathbb{R}^\Omega$ as V_Γ we need another orthogonal decomposition of $\mathbb{R}^\Omega$ which is related to the two previous ones and can be used for ANOVA. The next three sections show how to obtain such a decomposition in some cases, while indicating that it is difficult in general.

The Orthogonal Case

In this section, we start to combine the initial orthogonal decompositions of $\mathbb{R}^\Gamma$ and $\mathbb{R}^\Omega$. Of course, this depends on the way that the design map τ has embedded $\mathbb{R}^\Gamma$ in $\mathbb{R}^\Omega$ as V_Γ. The combination is relatively straightforward under a condition on τ known as orthogonality.

Given orthogonal decompositions of $\mathbb{R}^\Omega$ and $\mathbb{R}^\Gamma$, a design map τ is said to be *orthogonal* if every subspace in the given decomposition of V_Γ is geometrically orthogonal to every subspace in the given decomposition of $\mathbb{R}^\Omega$. Then the non-zero intersections of these subspaces give a canonical orthogonal decomposition of $\mathbb{R}^\Omega$ that refines both of the previous two.

Example 7. When $G_2 = \mathrm{Sym}(\Omega)$, the initial decomposition of $\mathbb{R}^\Omega$ is $V_0 \oplus V_0^\perp$. Such a design is called *completely randomized*. We always assume that V_0 is in the decomposition of $\mathbb{R}^\Gamma$, so now the combined decomposition simply adjoins $V_\Gamma^\perp$ to the decomposition of V_Γ. The subspace $V_\Gamma^\perp$ is known as *residual* in ANOVA.

Example 8. For a so-called *complete-block design*, there are b blocks of size k, where $k = |\Gamma|$, and every treatment occurs once in each block. If $G_1 = \mathrm{Sym}(k)$ and $G_2 = \mathrm{Sym}(k)\mathrm{wr}\mathrm{Sym}(b)$, then the initial decompositions are $V_0 + W_T$ (for $\mathbb{R}^\Gamma$) and $V_0 \oplus W_B \oplus V_B^\perp$ (for $\mathbb{R}^\Omega$). The combined decomposition is $V_0 \oplus W_B \oplus W_T \oplus (V_B + V_T)^\perp$, whose subspaces are usually called *grand mean*, *blocks*, *treatments*, and *residual*, respectively. If G_1 gives a finer decomposition of $\mathbb{R}^\Gamma$, then this gives a finer decomposition of W_T without affecting W_B or $(V_B + V_T)^\perp$.

Example 9. A third popular orthogonal design is the Latin square. Here Ω is an $n \times n$ rectangle, where $n = |\Gamma|$. The initial decomposition of $\mathbb{R}^\Omega$ is $V_0 \oplus W_R \oplus W_C \oplus (V_R + V_C)^\perp$. When treatments are applied in a Latin square, V_Γ is orthogonal to $W_R \oplus W_C$, so $(V_R + V_C)^\perp$ is decomposed as $W_T \oplus (V_R + V_C + V_T)^\perp$, with the second part being called *residual*. Again, any finer initial decomposition of $\mathbb{R}^\Gamma$ simply gives a decomposition of W_T.

Example 10. The simplest case in which more than one subspace in the initial decomposition of $\mathbb{R}^\Omega$ is split up is the so-called *split-plot design*. The treatments are as in Example 1. For Ω, there are rn blocks, each containing m observational units, so that $\mathbb{R}^\Omega = V_0 \oplus W_B \oplus V_B^\perp$. Each level of C is applied to r whole blocks, and each level of D is applied to one observational unit per block. Thus $V_C < V_B$, while W_D and W_{CD} are both orthogonal to V_B. The combined decomposition is

$$\begin{array}{ccccccccc} V_0 & \oplus & W_C & \oplus & (W_B \cap W_C^\perp) & \oplus & W_D & \oplus & W_{CD} & \oplus & (V_B^\perp \cap V_\Gamma^\perp) \\ 1 & & n-1 & & n(r-1) & & m-1 & & (n-1)(m-1) & & n(m-1)(r-1) \end{array},$$

where the dimension is shown underneath each subspace. The third subspace is called *block residual*. Under the assumption that treatments affect expectation and blocks have random effects, MS_C is compared with the mean square for block residual while MS_D and MS_{CD} are both compared with the mean square for $V_B^\perp \cap V_\Gamma^\perp$.

In general, if W is a subspace in the initial decomposition of $\mathbb{R}^\Omega$ and $W \cap V_\Gamma^\perp$ is non-zero, then $W \cap V_\Gamma^\perp$ is called a *residual subspace*.

The General Non-Orthogonal Case

This section gives a quick overview of some difficulties that can occur when the design is not orthogonal. We show that the worst of these are avoided if the structure on Ω is a partition into blocks of equal size. The development follows [30, 44].

When there is not geometric orthogonality between the original decompositions of $\mathbb{R}^\Omega$ and V_Γ, it is normal to start with the decomposition of $\mathbb{R}^\Omega$ and then try to refine it. Let U be one of the subspaces in the original decomposition of $\mathbb{R}^\Omega$. Given an orthogonal decomposition $\bigoplus_{j\in J} W_j$ of V_Γ, one obvious step is to project each W_j onto U. There are two difficulties. The first is that, even if W_i is orthogonal to W_j, their projections onto U may no longer be orthogonal to each other.

The second difficulty needs more explanation. Suppose that $P_j y = v$. Denote by Q the (matrix of) orthogonal projection onto U. If ϕ is the angle between v and Qv, then the sum of squares for $Q(W_j)$ is $(\cos^2\phi)\,\|v\|^2$ plus other non-negative pieces. Even if there are no other contributions to this sum of squares, we need to know $\cos^2\phi$ in order to make a judgement about the size of v. It is therefore helpful if all vectors in W_j make the same angle with U.

Fortunately, both difficulties are solved if each space W_j is an eigenspace of PQP, where $P = \sum_j P_j$. If this eigenspace decomposition of V_Γ and the original decomposition of V_Γ have a common refinement, then it is used. If not, there are disagreements about how to proceed: see [47].

Now consider two different subspaces U_1 and U_2 in the original decomposition of $\mathbb{R}^\Omega$, with projectors Q_1 and Q_2. If $V_0 \oplus U_1 \oplus U_2 = \mathbb{R}^\Omega$ and $W_j \perp V_0$, then $P_j(Q_1 + Q_2)P_j = P_j$, and so W_j is an eigenspace of $P_jQ_1P_j$ if and only if it is an eigenspace of $P_jQ_2P_j$. However, if $V_0 \oplus U_1 \oplus U_2$ is not the whole of $\mathbb{R}^\Omega$, then PQ_1P may not commute with PQ_2P, in which case these matrices do not have common eigenspaces.

Given an original decomposition of $\mathbb{R}^\Omega$ into subspaces $U_1, \ldots, U_s$ with projectors $Q_1, \ldots, Q_s$, Houtman and Speed [30] defined the design to have *general balance* if the matrices $PQ_1P, \ldots, PQ_sP$ commute with each other, where P is the projector onto V_Γ. This implies that, if the structure on Ω is defined simply by a partition into blocks of equal size, then all designs are generally balanced.

For the rest of this paper, we restrict attention to block designs or orthogonal designs, so that general balance is assured. Even with this restriction, there are still plenty of complications.

Incomplete-Block Designs

Apart from split-plot designs like those in Example 10, block designs in which the blocks are too small to hold all the treatments are not orthogonal. In this section we give the algebraic approach to studying these, starting with the seminal work of James [31], which was extended by Mann [39] to linear models which may not arise from experimental designs.

When there are b blocks of size k, the initial decomposition of $\mathbb{R}^\Omega$ is $V_0 \oplus W_B \oplus V_B^\perp$, with corresponding projectors Q_0, Q_B and $I - Q_0 - Q_B$. Here $Q_0 = (bk)^{-1}J$, where J is the all-1 matrix, and $Q_0 + Q_B = k^{-1}J_B$, where the (ω_1, ω_2)-entry of J_B is 1 if ω_1 and ω_2 are in the same block and is 0 otherwise. Thus the linear operator $Q_0 + Q_B$ simply replaces each entry of y by the average value on each block, so it is similar to a Radon transform [10, 37].

Suppose that there are t treatments each occurring r times, so that $tr = bk$. In an *incomplete-block design*, $k < t$ and no treatment occurs more than once in any block. Such a design is *balanced* if there is a constant λ such that each pair of treatments occur together in exactly λ blocks.

Put $W_T = V_\Gamma \cap V_0^\perp$. The orthogonal projector P_T onto W_T is $r^{-1}J_T - Q_0$, where the (ω_1, ω_2)-entry of J_T is 1 if $\tau(\omega_1) = \tau(\omega_2)$ and is 0 otherwise. Thus $Q_0 + P_T$ is another averaging operator. Let $W_{T|B}$ be the orthogonal projection of W_T onto $V_B^\perp$, which is $(W_T + V_B) \cap V_B^\perp$. The classical analysis of data from an incomplete-block design uses the decomposition $V_0 \oplus W_B \oplus W_{T|B} \oplus (V_B + V_\Gamma)^\perp$. The second and third subspaces are called *blocks ignoring treatments* and *treatments eliminating blocks*, respectively.

James seems to have been one of the first to have studied ANOVA for balanced incomplete-block designs from the point of view of the algebraic properties of the idempotents which yield the sum of squares decomposition. In [31] he defined the "relationship algebra of an experimental design" as the complex algebra $\mathscr{A}$ generated by the matrices I, J, J_B, and J_T. He showed that $P_T Q_B P_T = (1-e)P_T$, where $e = t(k-1)/(t-1)k$, which is called the *efficiency factor* of the design. As James and Wilkinson later showed in [34], every vector in W_T has angle ϕ with W_B, where $\cos^2 \phi = 1 - e$.

James proposed refining the classical ANOVA by decomposing W_B into $Q_B(W_T)$ and its orthogonal complement. The latter is zero if and only if $b = t$; the former has dimension $t - 1$ (this gives an easy proof of Fisher's Inequality, which states that $b \geq t$ for a balanced incomplete-block design). He showed that the algebra $\mathscr{A}$ has dimension six if $b = t$ and dimension seven otherwise. Since the generating matrices are symmetric, the algebra $\mathscr{A}$ is semisimple and thus may be decomposed as a direct sum of matrix algebras, in this case two or three one-dimensional algebras and one 2×2 matrix algebra (which is a four-dimensional algebra). The one-dimensional algebras correspond to the subspaces V_0, $(V_B + V_\Gamma)^\perp$ and, if it is non-zero, $W_B \cap V_\Gamma^\perp$. The projector onto their orthogonal complement can be written as a sum of two idempotents in the algebra in infinitely many ways.

James closed [31] by remarking that "For certain designs, the relationship algebra is the commutator algebra of the representation of a group expressing the symmetry of the experimental design". We take up the relevance of this remark in Section "Strong symmetries of incomplete-block designs".

This work was extended to arbitrary incomplete-block designs in [34]. If $P_T Q_B P_T$ has rank less than $t - 1$, then $V_\Gamma \cap V_B^\perp$ is non-zero. The vectors in this subspace are said to have *canonical efficiency factor* 1. Let the distinct non-zero eigenvalues of $P_T Q_B P_T$ be $\mu_1, \ldots, \mu_s$, where μ_i has multiplicity d_i. Then there are corresponding d_i-dimensional subspaces U_i of W_B and V_i of W_T such that every vector in V_i makes

angle ϕ_i with U_i, and vice versa, where $\cos^2 \phi_i = \mu_i$, and $1 - \mu_i$ is the canonical efficiency factor for vectors in V_i. If $i \neq j$, then $U_i \perp V_j$, $U_i \perp U_j$ and $V_i \perp V_j$.

Now $\mathbb{R}^\Omega$ has an orthogonal decomposition as

$$V_0 \oplus (V_B \cap V_\Gamma^\perp) \oplus (V_\Gamma \cap V_B^\perp) \oplus (V_B + V_\Gamma)^\perp \oplus (U_1 + V_1) \oplus \cdots \oplus (U_s + V_s), \qquad (2)$$

where the second or third subspace may be zero. The algebra $\mathscr{A}$ is the sum of scalar algebras on the first four subspaces, plus one 2×2 matrix algebra on each of $U_1 + V_1$, $\ldots$, $U_s + V_s$.

When $s = 1$ but the design is not balanced then there are just two canonical efficiency factors, one of which is 1. Such designs are called *partial geometric designs* in [11, 12], *C-designs* in [49], and $1\frac{1}{2}$*-designs* in [45]. They appear to be rather useful (see [16]). One way of obtaining them is to exchange the roles of blocks and treatments (thus forming the *dual* design) in a balanced design with $b > t$. Other examples include transversal designs and the lattice designs of Yates [56]. Some classes of such designs have been shown to be optimal (see [9, 22]).

Ice Cream Data

So far, we have assumed that the measurement y_ω on ω is influenced by ω itself and its inherent relation to the rest of Ω, as well as the treatment $\tau(\omega)$ and its relation to the rest of Γ. This paradigm does not cover experiments where y_ω might also be influenced by the treatments on observational units which are, in some sense, near to ω. For example, tall varieties of sunflower will shade their shorter neighbors, or the taste of one ice cream may affect the score given by the taster to another ice cream.

In such circumstances, it seems appropriate to consider the group G of *strong* symmetries of the design. This is the subgroup of the group G_2 of permutations of Ω which preserve the partition of Ω defined by the inverse images under τ (in particular, they stabilize V_Γ) and whose induced permutations on Γ are in G_1.

Example 4 is, in fact, silly. It wastes resources, because 240 items are tasted but only 60 are rated. In the actual experiment reported by Calvin in [17], only 15 people took part, each tasting four items and scoring each one. The six treatments were six quantities of vanilla added to the basic ice cream. There was one taster for each subset of size four, and they were asked to give each tasted item an integer score between 0 and 5 inclusive. It is not clear whether the tasters were told that their four treatments were all different (e.g., four tasters gave the same score to two of their items).

In this case, $t = |\Gamma| = 6$ and G_1 is $\mathrm{Sym}(6)$ in the natural action. Also, $|\Omega| = 60$. Calvin considered the tasters as blocks, so that G_2 is $\mathrm{Sym}(4)\mathrm{wr}\mathrm{Sym}(15)$ in its imprimitive action. The design was a balanced incomplete-block design with fifteen blocks of size four. Thus G is $\mathrm{Sym}(6)$ in the action described at the end of Section "Treatment Permutations".

Let $B(\omega)$ be the block containing ω. Calvin proposed the model that $E(Y) = f + d$, where $d(\omega) = \delta_{B(\omega)}$ and $f(\omega) = \gamma_i + \theta_{ij} + \theta_{ik} + \theta_{il}$ if $\tau(\omega) = i$ and the other treatments in $B(\omega)$ are j, k, and l. Furthermore, $\theta_{ij} = -\theta_{ji}$ for all i and j. Thus $d \in V_B$; the γ-parameters give a vector in V_1; and the θ-parameters give a vector in V_A, in the notation used in Example 4.

As we saw in Sections "Treatments" and "Treatment Permutations", there is a 5-dimensional subspace $\widetilde{V}_B$ of V_B such that the sum of any two of $\widetilde{V}_B$, $V_1 \cap V_0^\perp$ and $V_P \cap V_A$ is the same 10-dimensional subspace, the homogeneous subspace for $S^{5,1}$. Thus the δ-, γ- and θ-parameters are not all identifiable. Calvin got around this problem by restricting the θ-parameters to give a vector in the 10-dimensional subspace W_A, which is the homogeneous subspace for $S^{4,1,1}$.

Calvin gave the ANOVA in Table 1. As is common for statisticians, he omitted the line for V_0, he wrote "d.f." (degrees of freedom) for "dimension", and he called the residual line "Error". We have added the column for subspaces to clarify what he meant by "Source of variation".

Table 1 Expanded version of the ANOVA table given by Calvin in [17]

Subspace	Source of variation	d.f.	S.S.	M.S.
V_0	Grand mean	1	390.15	390.15
W_B	Blocks (unadjusted)	14	18.10	1.29
$(V_1 + V_B) \cap V_B^\perp$	Treatments (adjusted)	5	71.17	14.23
W_A	Correlations (adjusted)	10	27.17	2.72
$(V_A + V_B)^\perp$	Error	30	40.41	1.35
V	Total	60	547.00	—

In (1) we gave the decomposition of V into homogeneous subspaces. The approach outlined in [48] (via the discrete Radon transform [10]) gives the sum of squares (S.S.) for each of these as follows.

$$\begin{array}{c|l} & M^{3,2,1} = S^6 \oplus 2S^{5,1} \oplus 2S^{4,2} \oplus S^{4,1,1} \oplus S^{3,3} \oplus S^{3,2,1} \\ \hline \text{S.S.} & 547 = 390.15 + 77.667 + 23.40 + 27.167 + 8.083 + 20.533 \end{array} \tag{3}$$

Let U_1 and U_2 be the reducible homogeneous subspaces of dimensions 10 and 18, respectively. We now explore different ways of decomposing these two subspaces into orthogonal irreducibles.

Each block can be labelled by the pair of treatments which are not present in it. Thus the fifteen blocks have structure similar to that in Example 2. The method given by Yates in [57] decomposes the sum of squares for blocks into 6.50 for $\widetilde{V}_B$ and 11.60 for $W_B \cap \widetilde{V}_B^\perp$. These two subspaces are $U_1 \cap V_B$ and $U_2 \cap V_B$, respectively. We have already seen that $\widetilde{V}_B^\perp \cap U_1 = V_P \cap V_A = (V_1 + V_B) \cap V_B^\perp$, whose sum of squares is given in Table 1 as 71.17 (to two decimal places). Then $6.50 + 71.17 = 77.67$, which gives the sum of squares for U_1, as confirmed in (3). The sum of squares for U_2 is 23.40, so the sum of squares for $U_2 \cap V_B^\perp$ is $23.40 - 11.60 = 11.80$.

A statistician expects there to be differences between the blocks. It is therefore standard to begin with the decomposition $V_0 \oplus W_B \oplus V_B^\perp$ and then refine that. Refining it into irreducibles gives the ANOVA in Table 2.

Table 2 ANOVA table obtained by refining the original decomposition (defined by blocks) into group-irreducibles

Original	Group refinement	d.f.	S.S.	M.S.
V_0	V_0	1	390.15	390.15
W_B	$U_1 \cap V_B$	5	6.50	1.30
	$U_2 \cap V_B$	9	11.60	1.29
$V_B^\perp$	$U_1 \cap V_B^\perp = V_P \cap V_A$	5	71.17	14.23
	W_A	10	27.17	2.72
	$U_2 \cap V_B^\perp$	9	11.80	1.31
	$S^{3,3}$	5	8.08	1.62
	$S^{3,2,1}$	16	20.53	1.28
V	Total	60	547.00	—

It is not unusual for the mean square for V_0 to be much larger than the rest. The interesting question for data analysis is: which other mean squares are significantly larger than the rest?

One notable feature of Table 2 is that the four smallest mean squares are all approximately equal (about 1.3). In particular, the two subspaces of W_B are among these, so it appears that there are no differences between blocks.

In fact, given the way that the experiment was carried out, this is not surprising. Each taster had to give four integer scores in the range $[0,5]$, and most of them thought that they should not give the same score twice. It was therefore almost impossible for one taster to give consistently higher scores than another.

If blocks are not important, what other natural subspaces of U_1 and U_2 should we look at? To Calvin, the next most obvious subspace of U_1 was W_T, where $W_T = V_1 \cap V_0^\perp$ and $f \in V_1$ if there are constants $\gamma_1, \ldots, \gamma_6$ such that

$$f((i,\{j,k,l\})) = \gamma_i. \tag{4}$$

The constant γ_i is estimated by the mean of the responses for treatment i, and this gives the sum of squares for W_T as 63.35.

However, Calvin also proposed that treatments should affect each other asymmetrically: the taster has a fixed, short scale, so if one treatment's score goes up then another comes down. Another submodel of his model is

$$f((i,\{j,k,l\})) = \mu + 3\alpha_i - \alpha_j - \alpha_k - \alpha_l, \tag{5}$$

where μ is an overall constant. This corresponds to the subspace $V_0 \oplus (V_P \cap V_A)$. Since $V_P \cap V_A = U_1 \cap V_B^\perp$, we have already seen that the sum of squares for this is 71.17. Thus model (5) fits the data better than model (4).

In Table 3, the tasters' scores (the data) are shown at the top of each box. Below that are the fitted values for model (5), which can be easily calculated from Calvin's

results. They fit the data rather well. The third row gives the fitted values for the more general asymmetric model

$$f((i,\{j,k,l\})) = \mu + \theta_{ij} + \theta_{ik} + \theta_{il} \qquad \text{where } \theta_{rs} = -\theta_{sr} \text{ for all } r \text{ and } s: \tag{6}$$

this corresponds to the subspace $V_0 \oplus (V_P \cap V_A) \oplus W_A = V_0 \oplus V_A$.

Table 3 Ice cream analysis: data (top row); fitted values in model (5) (second row); fitted values in model (6) (third row); fitted values in model (7) (bottom row)

	1 2 3 4	1 2 3 5	1 2 3 6	1 2 4 5	1 2 4 6	1 2 5 6	1 3 4 5	1 3 4 6	1 3 5 6	1 4 5 6	2 3 4 5	2 3 4 6	2 3 5 6	2 4 5 6	3 4 5 6
1	2	0	0	1	0	0	2	0	0	2	×	×	×	×	×
	1.02	0.99	0.86	0.75	0.62	0.60	0.65	0.52	0.49	0.25					
	1.49	0.70	-0.05	0.78	0.01	-0.76	2.09	1.32	0.55	0.61					
	2.42	2.33	2.17	3.25	3.08	3.00	2.50	2.33	2.25	3.17					
2	4	4	3	3	2	1	×	×	×	×	0	1	2	1	×
	2.46	2.44	2.31	2.20	2.06	2.04					1.73	1.60	1.58	1.33	
	3.82	3.05	3.28	2.11	2.34	1.57					1.34	1.57	0.80	-0.14	
	2.75	3.33	2.83	3.25	2.75	3.33					3.17	2.67	3.25	3.17	
3	0	3	4	×	×	×	3	0	2	×	2	1	3	×	4
	2.88	2.85	2.72				2.51	2.38	2.36		2.15	2.02	1.99		1.65
	1.90	2.47	1.86				2.84	2.24	2.80		2.42	1.82	2.38		2.76
	0.92	2.50	2.17				1.33	1.00	2.58		2.00	1.67	3.25		2.08
4	3	×	×	4	4	×	1	3	×	4	3	2	×	4	3
	3.85			3.58	3.45		3.48	3.35		3.08	3.12	2.99		2.72	2.62
	2.99			3.28	3.17		2.92	2.82		3.11	3.51	3.40		3.70	3.34
	1.75			2.58	3.08		1.50	2.00		2.83	1.17	1.67		2.50	1.42
5	×	5	×	5	×	4	4	×	3	1	1	×	5	5	1
		3.91		3.67		3.52	3.56		3.41	3.17	3.20		3.05	2.81	2.70
		3.97		4.03		4.65	2.34		2.97	3.03	2.92		3.55	3.61	1.92
		3.83		3.08		3.00	3.00		2.92	2.17	3.33		3.25	2.50	2.42
6	×	×	5	×	5	5	×	4	3	4	×	4	4	2	2
			4.31		4.06	4.04		3.96	3.94	3.70		3.60	3.58	3.33	3.23
			5.11		4.67	4.73		3.82	3.88	3.45		3.40	3.47	3.03	2.17
			2.83		2.92	2.33		3.00	2.42	2.50		3.00	2.42	2.50	2.58

The asymmetric effect is sometimes known as *competition*x. It occurs when neighboring treatments compete for finite resources, be they food or tasters' good opinions. In some situations, a symmetric effect is more natural: that is, $\theta_{ij} = \theta_{ji}$ for all i and j. In many wildlife habitats, there is synergy between organisms filling different niches, to each others' mutual benefit, so that θ_{ij} is positive. On the other hand, antagonism gives a symmetric effect with negative θ_{ij}.

The symmetric model is

$$f((i,\{j,k,l\})) = \mu + \theta_{ij} + \theta_{ik} + \theta_{il} \qquad \text{where } \theta_{rs} = \theta_{sr} \text{ for all } r \text{ and } s. \tag{7}$$

The corresponding subspace V_S is similar to the whole space in Example 2, and decomposes as $V_0 \oplus (V_P \cap V_S) \oplus W_S$. A slight modification of Yates's method gives the corresponding sums of squares as 390.15, 7.64, and 16.54, respectively. The fitted vector in V_S is shown in the last row of Table 3. It is clearly not as good a fit to the data as either of the two rows above.

We now have three natural ways of decomposing U_1 as a pair of orthogonal irreducible subspaces. Table 4 shows the corresponding sums of squares and mean squares. Starting with blocks or with the asymmetric treatment model (5) gives (**a**); starting with direct effects of treatments, which is model (4), gives (**b**); and starting with the symmetric treatment model (7) gives (**c**). Of these, the only one where the larger mean square corresponds to a meaningful subspace and the other mean square is about 1.3 is the first. Thus consideration of U_1 suggests that we should include the subspace $V_P \cap V_A$ in the explanation for the data but that the rest of U_1 is just random noise.

Of course, there is a fourth decomposition, into the G-irreducible subspace containing the projection of y onto U_1 and its orthogonal complement, with mean squares 15.53 and 0, respectively. This is the most extreme decomposition of U_1 into orthogonal irreducibles, but we cannot consider it seriously for data analysis. In the first place, this decomposition is not known before the data are obtained. In the second place, the zero mean square is just *too* small: when four others are around 1.3 then anything much smaller is suspicious.

Table 4 Three natural ways of decomposing the subspace U_1

subspace	S.S.	M.S.
$U_1 \cap V_B$	6.50	1.30
$V_P \cap V_A$	71.17	14.23
U_1	77.67	

(**a**) blocks or asymmetric model

subspace	S.S.	M.S.
W_T	63.35	12.67
$U_1 \cap W_T^{\perp}$	14.32	2.86
U_1	77.67	

(**b**) direct treatment effects

subspace	S.S.	M.S.
$V_P \cap V_S$	7.64	1.53
$U_1 \cap V_S^{\perp}$	70.03	14.01
U_1	77.67	

(**c**) symmetric model

For decomposing U_2, we have the two possibilities shown in Table 5 (ignoring the extra one defined by the data). Starting with blocks gives (**a**), while starting with the symmetric treatment model gives (**b**). In the first, both mean squares are about 1.3, which is consistent with random noise. In the second, the larger mean square is 1.84. This is less than twice 1.3, so is unlikely to indicate anything meaningful. Moreover, it corresponds to the subspace W_S. Figure 3 shows that any natural treatment subspace containing W_S must contain the whole of V_S, in particular $V_P \cap V_S$, whose contribution to the data we have already decided is just random noise. These considerations suggest that no part of U_2 is anything more than random noise.

Table 5 Two natural ways of decomposing the subspace U_2

subspace	S.S.	M.S.
$U_2 \cap V_B$	11.60	1.29
$U_2 \cap V_B^\perp$	11.80	1.31
U_2	23.40	

(**a**) blocks

subspace	S.S.	M.S.
W_S	16.54	1.84
$U_2 \cap W_S^\perp$	6.86	0.76
U_2	23.40	

(**b**) symmetric model

There are three remaining subspaces in Table 2. Of these, $S^{3,2,1}$ has the smallest mean square, while the mean square for $S^{3,3}$ is 1.62, less than that for W_S, which we have already decided to ignore. That leaves just W_A, with a mean square of 2.72. As Figure 3 shows, including W_A in the treatment subspace that already includes V_0 and $V_P \cap V_A$ gives the rather natural subspace $V_0 + V_A$. This corresponds to model (6).

The conclusion from the spectral analysis is that model (6) explains the data well. That is, the different quantities of vanilla compete with each other for the tasters' scores, but there is no evidence of any direct effect of quantities or any differences between tasters. These conclusions differ from those in [17], because Calvin assumed that the most important effects would be the differences between tasters and the differences between the direct effects of the quantities of vanilla.

Note that the computational aspects of this are an instance of computing projections of a data vector onto the isotypic components of a representation of the symmetric group (see, e.g., [25]). The general computational problem of isotypic projection for arbitrary groups is considered in [41] as well as [40].

Strong symmetries of orthogonal designs

Even without the complication of the effects of neighboring treatments, we can define the group G of strong symmetries of the design. Do its irreducible subspaces help us to analyze the data? In this section we revisit Examples 7–10 and consider their strong symmetries.

The simplest orthogonal case is the completely randomized design in Example 7, where Ω is unstructured and each treatment is applied to r observational units, for some integer r. Then $G = \mathrm{Sym}(r)\mathrm{wr}G_1$, and [19] shows that every decomposition of $\mathbb{R}^\Omega$ into orthogonal G-irreducible subspaces has the form $\left(\bigoplus_{j\in J} W_j\right) \oplus V_\Gamma^\perp$, where $\bigoplus_{j\in J} W_j$ is an orthogonal decomposition of V_Γ into G_1-irreducible subspaces. Here $V_\Gamma^\perp$ is the subspace which is classically called "Error" or "residual" in the ANOVA, and so the approach using strong symmetries gives nothing new.

For a complete-block design in b blocks of size k, as in Example 8, we have $|\Gamma| = k$, $G_2 = \mathrm{Sym}(k)\mathrm{wr}\mathrm{Sym}(b)$, and $G = G_1 \times \mathrm{Sym}(b)$ in its product action. Let U_0 and U_1 be the irreducibles of $\mathrm{Sym}(b)$ in its natural action, of dimensions 1 and $b-1$, respectively. If $\bigoplus_{j\in J} W_j$ is an orthogonal G_1-irreducible decomposition of $\mathbb{R}^\Gamma$, then the subspaces in an orthogonal G-irreducible decomposition of $\mathbb{R}^\Omega$ are $U_i \otimes W_j$ for i in $\{0,1\}$ and j in J. Here $U_0 \otimes W_0 = V_0$, and $U_1 \otimes W_0 = W_B$, which is the subspace for differences between blocks. The subspaces $U_0 \otimes W_j$, for j in $J \setminus \{0\}$, give the decomposition of $V_\Gamma \cap V_0^\perp$ specified by G_1. If $|J| = 2$, then the only remaining subspace is $U_1 \otimes W_1$, which is $(V_B + V_\Gamma)^\perp$, the unique residual subspace. This case was discussed, from the point of view of strong symmetries, in [29, 38]. However, if $|J| \geq 3$, then $(V_B + V_\Gamma)^\perp$ is not G-irreducible.

For example, suppose that $k = mn$ and that Γ is as in Example 1. The approach of Section "The Orthogonal Case" gives the ANOVA in Table 6(**a**), while consideration of strong symmetries gives the decomposition in Table 6(**b**). Which should be used?

There is disagreement among statisticians about how to answer this question. The approach described by Nelder in [43, 44] is that in Sections "The Orthogonal Case" and "The General Non-Orthogonal Case": start with the decomposition of $\mathbb{R}^\Omega$ determined by G_2 and refine it using the decomposition of V_Γ. If there are simply fifteen treatments, then $(V_B + V_\Gamma)^\perp$ is used as the residual subspace: why should this be decomposed if the fifteen treatments are all combinations of five varieties of cow-peas with three methods of cultivation? This gives the decomposition in Table 6(**a**). A popular alternative approach is to start with a list of *factors* (that is, partitions of Ω with named parts) and close it under infima, where the infimum of two partitions is their coarsest common refinement. This gives the decomposition in Table 6(**b**). These two approaches are contrasted in [13].

It is not uncommon for the initial structure on Ω to be defined by a family $\mathscr{P}_2$ of partitions of Ω (such as the partitions into blocks, rows or columns) and the structure on Γ to be defined by a family $\mathscr{P}_1$ of partitions of Γ (such as those defined by factors C and D in Example 1). Each partition defines the subspace of vectors which are constant on each of its parts. Two partitions are said to be *orthogonal* to each other if their corresponding subspaces are geometrically orthogonal. The design function τ enables us to consider partitions of Γ to be partitions of Ω which refine the partition defined by the inverse images of τ. If $\mathscr{P}_1 \cup \mathscr{P}_2$ contains the two trivial partitions of Ω, is closed under suprema, and has the property that each pair of partitions is orthogonal, then it defines an orthogonal decomposition $\mathscr{W}$ of $\mathbb{R}^\Omega$ [54]. Now G is the group of permutations of Ω which preserve every partition in $\mathscr{P}_1 \cup \mathscr{P}_2$. In order for the subspaces in $\mathscr{W}$ to be G-irreducible, it is necessary that each partition has all its parts of the same size (otherwise, G cannot be transitive on Ω) and that $\mathscr{P}_1 \cup \mathscr{P}_2$ be closed under infima. It is arguable that the problem with the preceding example of a factorial design in complete blocks is the lack of closure under infima.

Table 6 Two different decompositions for a factorial design in complete blocks

Subspace	Source of variation	d.f.
$U_0 \otimes W_0 = V_0$	Grand mean	1
$U_1 \otimes W_0$	Blocks	$b-1$
$U_0 \otimes W_C$	Main effect of C	$n-1$
$U_0 \otimes W_D$	Main effect of D	$m-1$
$U_0 \otimes W_{CD}$	C-by-D interaction	$(n-1)(m-1)$
$(V_B + V_\Gamma)^\perp$	Residual	$(b-1)(mn-1)$
V	Total	bmn

(**a**) Method of Section "The Orthogonal Case"

Subspace	Source of variation	d.f.
$U_0 \otimes W_0 = V_0$	Grand mean	1
$U_1 \otimes W_0$	Blocks	$b-1$
$U_0 \otimes W_C$	Main effect of C	$n-1$
$U_1 \otimes W_C$	Residual for main effect of C	$(b-1)(n-1)$
$U_0 \otimes W_D$	Main effect of D	$m-1$
$U_1 \otimes W_D$	Residual for main effect of D	$(b-1)(m-1)$
$U_0 \otimes W_{CD}$	C-by-D interaction	$(n-1)(m-1)$
$U_1 \otimes W_{CD}$	Residual for C-by-D-interaction	$(b-1)(n-1)(m-1)$
V	Total	bmn

(**b**) Irreducible subspaces of group of strong symmetries

If, in addition to satisfying the preceding properties, the lattice of partitions in $\mathscr{P}_1 \cup \mathscr{P}_2$ is distributive, then G is a generalized wreath product and its irreducible subspaces are precisely those in $\mathscr{W}$ [8]. In this case, the strong symmetries give the same decomposition as that in Section "The Orthogonal Case". The split-plot design in Example 10 is a case in point.

To show that lack of closure under infima is not the whole explanation, we conclude this section by considering the Latin-square design in Example 9, and suppose that $G_1 = \mathrm{Sym}(n)$. The partitions of Ω into rows, columns, and letters, together with the two trivial partitions, satisfy all the aforementioned conditions, except that the lattice is not distributive. Now G is the subgroup of $\mathrm{Sym}(n) \times \mathrm{Sym}(n)$ which preserves the partition into letters. If the Latin square is not the Cayley table of a group, then G may not even be transitive on Ω: indeed, it may be trivial. Even when it is such a Cayley table, the results in [2] show that there may be surprisingly many G-irreducibles in a decomposition of $\mathbb{R}^\Omega$. However, neither of the common approaches to ANOVA described above uses any finer decomposition than the one in Section "The Orthogonal Case".

Thus considerations of symmetries, partitions, combinatorial conditions, or models may lead to different analyses. The Latin square seems to be a relatively straightforward design, yet subtle differences in assumptions have led to arguments over the correct data analysis ever since Neyman [46].

Strong symmetries of incomplete-block designs

In this section we return to the incomplete-block designs of Section "Incomplete-Block Designs", and use the notation introduced there. Thus Ω consists of b blocks of size k, and Γ consists of t treatments, where $t > k$. We assume no structure on Γ. The group G of strong symmetries consists of all permutations of Ω which preserve the partition into blocks and the partition into treatments.

James argued in [32] that ANOVA should use a decomposition of $\mathbb{R}^\Omega$ into orthogonal G-irreducible subspaces. Here we compare this approach with that of Section "Incomplete-Block Designs".

Let ρ be the permutation representation of this action of G, with permutation character π. If $g \in G$ then $\rho(g)$ fixes the subspaces V_0, W_B and W_T. Therefore $\rho(g)$ commutes with Q_B and P_T as well as with I and J, and so $\mathscr{A} \subseteq \mathscr{C}(G)$. Hence each of the summands in (2) is G-invariant, while U_i is G-isomorphic to V_i for $i = 1, \ldots, s$.

For simplicity, write $V_B \cap V_\Gamma^\perp$ as W_{B-T}, $V_\Gamma \cap V_B^\perp$ as W_{T-B} and $(V_B + V_\Gamma)^\perp$ as W. Assume that $k \geq 2$ and $r \geq 2$, so that V_0 and W are both non-zero. Let δ be the number of subspaces among W_{B-T} and W_{T-B} that are non-zero, so that $\delta \in \{0, 1, 2\}$.

A block design is said to be *resolvable* if there is a partition of Ω into replicates, coarser than the partition into blocks, such that each treatment occurs once in each replicate. For a resolvable design, define W_R analogously to W_B. Then $\dim(W_R) = r - 1$ and $W_R \leq W_{B-T}$: hence the latter cannot be zero and so $\delta \geq 1$.

Returning to the general case, recall that the *rank* p of G is defined to be the number of orbits of G in its induced action on $\Omega \times \Omega$ (see [55]). If G is transitive on Ω, then p is equal to the number of orbits on Ω of the stabilizer in G of any element of Ω. Less obviously, p is also equal to the sum of the squares of the multiplicities of complex-irreducible characters in π.

As in Section "Treatment Permutations", there are non-negative integers m_i such that $\pi = \sum_{i \in \mathscr{I}} m_i \chi_i$, where $\{\chi_i : i \in \mathscr{I}\}$ is the set of real-irreducible characters of G. The relation to complex-irreducibles is explained in [4, 7, 18, 42, 51], as follows. The set $\mathscr{I}$ is the disjoint union of $\mathscr{I}_1$, $\mathscr{I}_2$ and $\mathscr{I}_3$. If $\chi \in \mathscr{I}_1$, then χ is also complex-irreducible, of real type; if $\chi \in \mathscr{I}_2$, then $\chi = 2\eta$, where η is a complex-irreducible of quaternionic type; if $\chi \in \mathscr{I}_3$, then $\chi = \zeta + \bar{\zeta}$, where ζ is complex-irreducible of complex type. Therefore

$$p = \sum_{i \in \mathscr{I}_1} m_i^2 + 4 \sum_{i \in \mathscr{I}_2} m_i^2 + 2 \sum_{i \in \mathscr{I}_3} m_i^2. \tag{8}$$

For an incomplete-block design, comparison of (8) with (2) shows that

$$p \geq 2 + \delta + 4s,$$

with equality if and only if each of V_0, W_{B-T}, W_{T-B}, W, U_1, …, U_s is G-irreducible, admitting a complex-irreducible character of real type, and there are no G-isomorphisms among this list of subspaces. In particular, if $p \leq 9$, then $s = 1$ and U_1 is G-irreducible.

Furthermore, if $p = 6$, then $s = 1$, $\delta = 0$, the design is a balanced incomplete-block design with $b = t$, and W is G-irreducible. If $p = 7$, then $s = 1$. In this case either $\delta = 1, t \neq b$, the design or its dual is a balanced incomplete-block design, and W is G-irreducible; or $\delta = 0$, the design is a balanced incomplete-block design with $b = t$, and W is the sum of two G-irreducible subspaces. If $p = 8$, then $s = 1$, then either $\delta = 2$, the design and its dual are both partial geometric designs, and W is G-irreducible; or $\delta = 1, t \neq b$, the design or its dual is a balanced incomplete-block design and one of W, W_{B-T}, W_{T-B} is the sum of two G-irreducible subspaces; or $\delta = 0$, the design is a balanced incomplete-block design with $b = t$, and W is the sum of three G-irreducible subspaces.

We now specialize these results to several well-known families of incomplete-block designs.

Example 11. Let q be a prime power. Then there is a Desarguesian projective plane Π of order q. Its points and lines can be used as the treatments and blocks in an incomplete-block design with $t = b = q^2 + q + 1$, $r = k = q + 1$ and $\delta = 0$. The group of strong symmetries is $\mathrm{P\Gamma L}(3,q)$, which is transitive on sets of four points in general position. This can be used to show that G has rank 6: the details are in [4, 14, 15, 20, 21, 32]. Hence a G-irreducible decomposition of $\mathbb{R}^\Omega$, with dimensions, is

	V_0	$\oplus$	W_B	$\oplus$	W_T	$\oplus$	W
dim	1		q^2+q		q^2+q		q^3,

where only the middle two subspaces are non-orthogonal to each other. In this case, the G-decomposition is the same as that used in classical ANOVA.

Example 12. Consider the affine plane Δ obtained from the projective plane Π in Example 11 by deleting one line and all points on it. This gives a resolvable balanced incomplete-block design with $t = q^2$, $b = q(q+1)$, $r = q+1$, $k = q$, and $\delta = 1$. This design is also known as a *balanced square lattice design*. Its group G of strong symmetries is the stabilizer in $\mathrm{P\Gamma L}(3,q)$ of the omitted line. This is transitive on the units in Ω, which may be identified with the flags (x, λ) where x is a point of Δ incident with the line λ of Δ. The stabilizer in G of (x, λ) has the following orbits on Ω:

$$\{(x,\lambda)\}$$
$$\{(x,\mu) : x \in \mu \neq \lambda\}$$
$$\{(z,\lambda) : x \neq z \in \lambda\}$$
$$\{(z,\mu) : z \in \mu, z \in \lambda, x \notin \mu\}$$
$$\{(z,\mu) : z \in \mu, z \notin \lambda, x \in \mu\}$$
$$\{(z,\mu) : z \in \mu, \mu \| \lambda\}$$
$$\{(z,\mu) : z \in \mu, z \notin \lambda, \mu \not\| \lambda, x \notin \mu\}.$$

Thus the rank of G is 7, and so a G-irreducible decomposition of $\mathbb{R}^\Omega$, with dimensions, is

	V_0	$\oplus$	W_R	$\oplus$	$(W_B \cap W_R^\perp)$	$\oplus$	W_T	$\oplus$	W
dim	1		q		q^2-1		q^2-1		$(q-1)^2(q+1)$

All pairs of subspaces are orthogonal, apart from the two with dimension q^2-1. Moreover, $W_R = W_{B-T}$. This decomposition was obtained by Burton and Chakravarti in [15].

For a resolvable block design, the classical ANOVA normally splits W_B into W_R and $W_B \cap W_R^\perp$. These subspaces are called *replicates* and *blocks within replicates*, respectively. So this is another example where the G-decomposition is the same as that used in classical ANOVA.

Example 13. The simple square lattice design introduced by Yates in [56] is resolvable with $r = 2$. The treatments are identified with an abstract $n \times n$ array, so that $t = n^2$, where $n > 1$. In the first replicate, the rows of the array are blocks; in the second replicate, the columns of the array are blocks. Thus $k = n$ and $b = 2n$.

Now $\dim(W_R) = 1$ and so $W_B \cap W_{B-T}^\perp$ has dimension at most $b-2$, which is $2(n-1)$. Therefore $\dim(W_{T-B}) = t - 1 - \dim(W_B \cap W_{B-T}^\perp) \geq n^2 - 1 - 2(n-1) = (n-1)^2 > 0$. Hence W_{B-T} and W_{T-B} are both non-zero and so $\delta = 2$. Both the design and its dual are partial geometric designs.

Now the group G of strong symmetries is $\mathrm{Sym}(n)\mathrm{wr}\mathrm{Sym}(2)$ in its product action. It is generated by all permutations of the set of rows ($\mathrm{Sym}(n)$), all permutations of the set of columns ($\mathrm{Sym}(n)$), and the interchange of rows and columns ($\mathrm{Sym}(2)$). It is transitive on flags. If x is the treatment in row λ and column μ, then the stabilizer in G of the flag (x, λ) has the following orbits on Ω:

$$\begin{aligned}
&\{(x,\lambda)\}\\
&\{(x,\mu)\}\\
&\{(z,\lambda) : x \neq z \in \lambda\}\\
&\{(z,\mu) : x \neq z \in \mu\}\\
&\{(z,\nu) : z \in \lambda, z \in \nu \neq \mu, \nu \text{ is a column}\}\\
&\{(z,\nu) : z \in \mu, z \in \nu \neq \lambda, \nu \text{ is a row}\}\\
&\{(z,\nu) : z \notin \lambda, z \in \nu \neq \mu, \nu \text{ is a column}\}\\
&\{(z,\nu) : z \notin \mu, z \in \nu \neq \lambda, \nu \text{ is a row}\}.
\end{aligned}$$

Hence the rank of G is 8, and the subspaces V_0, W_{B-T}, W_{T-B}, U_1, V_1 and W are all G-irreducible. Since W_R is G-invariant and $W_R \leq W_{B-T}$, we must have $W_R = W_{B-T}$. Hence the decomposition, with dimensions, is

	V_0	$\oplus$	W_R	$\oplus$	U_1	$\oplus$	V_1	$\oplus$	W_{T-B}	$\oplus$	W
dim	1		1		$2(n-1)$		$2(n-1)$		$(n-1)^2$		$(n-1)^2$

Only the pair U_1 and V_1 are non-orthogonal. In spite of the equality of their dimensions, the subspaces W_{T-B} and W are not G-isomorphic, because otherwise the rank would be greater.

Example 14. Projective spaces of higher dimension are also considered in [4, 15]. Here we consider dimension 3. Let q be a prime power and let Θ be the projective space of dimension 3 over the field with q elements. Take the treatments and blocks to be the points and planes of Θ, so that $t = b = q^3+q^2+q+1$ and $r = k = q^2+q+1$. The design and its dual are both balanced, and so $\delta = 0$. Now $G = \mathrm{P\Gamma L}(4,q)$, which is transitive on ordered sets of five points in general position, so the stabilizer in G of a flag (x,Ψ) has the following orbits on Ω:

$$\begin{aligned}
&\{(x,\Psi)\}\\
&\{(x,\Phi) : x \in \Phi \neq \Psi\}\\
&\{(z,\Psi) : x \neq z \in \Psi\}\\
&\{(z,\Phi) : x \neq z \in \Phi \neq \Psi, z \in \Psi, x \in \Phi\}\\
&\{(z,\Phi) : z \in \Phi, z \in \Psi, x \notin \Phi\}\\
&\{(z,\Phi) : z \in \Phi, z \notin \Psi, x \in \Phi\}\\
&\{(z,\Phi) : z \in \Phi, z \notin \Psi, x \notin \Phi\}.
\end{aligned}$$

Thus G has rank 7, and so U_1 and V_1 are G-irreducible while W is the sum of two G-irreducibles.

The space Θ contains $(q^2+1)(q^2+q+1)$ lines, each incident with $(q+1)^2$ flags. Let V_L be the subspace of $\mathbb{R}^\Omega$ spanned by the characteristic vectors of the lines. Then V_L is G-invariant. Analysis of the permutation characters of G on lines, flags, and points shows that V_L is the sum of three G-irreducible subspaces: one is V_0; one is G-isomorphic to both W_B and W_T; the third is orthogonal to $V_B + V_T$ and has dimension q^4+q^2. Hence the G-irreducible subspaces of W are $V_L \cap W$ and $W \cap V_L^\perp$. The decomposition, with dimensions, is

	V_0	$\oplus$	W_{B-T}	$\oplus$	W_{T-B}	$\oplus$	$(V_L \cap W)$	$\oplus$	$(W \cap V_L^\perp)$
dim	1		q^3+q^2+q		q^3+q^2+q		q^4+q^2		$q^5+q^4+q^3$.

In this case, the group of strong symmetries decomposes the classical residual subspace into two parts.

Example 15. A *triple square lattice* is made from a simple one by adding an extra replicate. A Latin square is superimposed on the $n \times n$ array. A block in the third replicate contains all treatments in a given letter of the square.

Now G is the group of all permutations of the square array which preserve the set of three partitions into rows, columns, and letters, in its action on the $3n^2$ flags. Depending on the Latin square, G may not be transitive, and finding a meaningful G-irreducible decomposition may be as difficult as for the case of a Latin-square design discussed at the end of Section "Strong symmetries of orthogonal designs".

These examples show that the decomposition defined by the group of strong symmetries may be the same as the classical one, may give further decomposition of the residual subspace, or may prove intractable. James [33] and Bailey [3] have both suggested that using the group of *weak* symmetries of Γ may give a meaningful decomposition of V_Γ. This group consists of those permutations of Γ whose

permutation matrices commute with $X^\top Q_B X$, where X is the $\Omega \times \Gamma$ incidence matrix whose (ω, i)-entry is equal to 1 if $\tau(\omega) = i$ and is equal to 0 otherwise. However, this approach does not get the extra residual subspace in Example 14, nor does it make Example 15 tractable.

References

1. R.A. Bailey, A unified approach to design of experiments. J. R. Stat. Soc. Ser. A **144**, 214–223 (1981)
2. R.A. Bailey, Latin squares with highly transitive automorphism groups. J. Aust. Math. Soc. Ser. A **33**, 18–22 (1982)
3. R.A. Bailey, Automorphism groups of block structures with and without treatments, in *Coding Theory and Design Theory, Part II, Design Theory*, ed. by D. Ray-Chaudhuri (Springer, New York, 1990), pp. 24–41
4. R.A. Bailey, Strata for randomized experiments (with discussion). J. R. Stat. Soc. Ser. B **53**, 27–78 (1991)
5. R.A. Bailey, *Design of Comparative Experiments* (Cambridge University Press, Cambridge, 2008)
6. R.A. Bailey, Structures defined by factors, in *Handbook on Design of Experiments*, ed. by D. Bingham, A.M. Dean, M. Morris, J. Stufken (Chapman and Hall, Boca Raton, 2014)
7. R.A. Bailey, C.A. Rowley, General balance and treatment permutations. Linear Algebra Appl. **127**, 183–225 (1990)
8. R.A. Bailey, C.E. Praeger, C.A. Rowley, T.P. Speed, Generalized wreath products of permutation groups. Proc. Lond. Math. Soc. **47**, 69–82 (1983)
9. R.A. Bailey, H. Monod, J.P. Morgan, Construction and optimality of affine-resolvable designs. Biometrika **82**, 187–200 (1995)
10. E. Bolker, The finite Radon transform. Contemp. Math. **63**, 27–50 (1987)
11. R.C. Bose, W.G. Bridges, M.S. Shrikhande, A characterization of partial geometric designs. Discret. Math. **16**, 1–7 (1976)
12. R.C. Bose, S.S. Shrikande, N.M. Singhi, Edge regular multigraphs and partial geometric designs with an application to the embedding of quasi-residual designs, in *Teorie Combinatoire (Rome, 1973)* Atti dei Convegni Lincei, vol. 17 (Accademia Nazionale dei Lincei, Roma, 1976), pp. 49–81
13. C.J. Brien, B.D. Harch, R.L. Correll, R.A. Bailey, Multiphase experiments with at least one later laboratory phase. I. Orthogonal designs. J. Agric. Biol. Environ. Stat. **16**, 422–450 (2011)
14. C.T. Burton, Automorphism groups of balanced incomplete block designs and their use in statistical model construction and analysis. PhD Thesis, University of North Carolina at Chapel Hill (1980)
15. C.T. Burton, I.M. Chakravarti, On the commutant algebras corresponding to the permutation representations of the full collineation groups of PG(k,s) and EG(k,s), $s = p^r$, $k \geq 2$. J. Math. Anal. Appl. **89**, 489–514 (1982)
16. T. Caliński, On some desirable patterns in block designs. Biometrics **27**, 275–292 (1971)
17. L.D. Calvin, Doubly balanced incomplete block designs for experiments in which the treatment effects are correlated. Biometrics **10**, 61–88 (1954)
18. P.J. Cameron, Bounding the rank of certain permutation groups. Math. Z. **124**, 243–352 (1972)
19. T. Ceccherini–Silberstein, F. Scarabotti, F. Tolli, *Representation Theory and Harmonic Analysis of Wreath Products of Finite Groups*. London Mathematical Society Lecture Notes Series, vol. 410 (Cambridge University Press, Cambridge, 2014)
20. I.M. Chakravarti, C.T. Burton, On the algebras of symmetries (groups of collineations) of designs from finite Desarguesian planes with applications in statistics. J. Math. Anal. Appl. **89**, 515–529 (1982)

21. I.M. Chakravarti, C.T. Burton, Symmetries (groups of automorphisms) of Desarguesian finite projective and affine planes and their role in statistical model construction, in *Statistics and Probability: Essays in Honor of C. R. Rao*, ed. by G. Kallianpur, P.R. Krishnaiah, J.K. Ghosh (North-Holland, Amsterdam, 1982), pp. 169–178
22. C.S. Cheng, R.A. Bailey, Optimality of some two-associate-class partially balanced incomplete-block designs. Ann. Stat. **19**, 1667–1671 (1991)
23. P. Diaconis, *Group Representations in Probability and Statistics.* IMS Lecture Notes—Monograph Series, vol. 11 (Institute of Mathematical Statistics, Hayward, 1988)
24. P. Diaconis, Applications of group representations to statistical problems, in *Proceedings of the ICM, Kyoto, Japan, 1990* (Springer, Tokyo, 1991), pp. 1037–1048
25. P. Diaconis, D. Rockmore, Efficient computation of isotypic projections for the symmetric group, in *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 11, ed. by L. Finkelstein, W. Kantor (American Mathematical Society, Providence, 1993), pp. 87–104
26. R.A. Fisher, The theory of confounding in factorial experiments in relation to the theory of groups. Ann. Eugen. **11**, 341–353 (1942)
27. P. Fortini, Representations of groups and the analysis of variance. PhD Thesis, Harvard University (1977)
28. B. Griffing, Concept of general and specific combining ability in relation to diallel crossing system. Aust. J. Biol. Sci. **9**, 463–493 (1956)
29. E.J. Hannan, Group representations and applied probability. J. Appl. Probab. **2**, 1–68 (1965)
30. A.M. Houtman, T.P. Speed, Balance in designed experiments with orthogonal block structure. Ann. Stat. **11**, 1069–1085 (1983)
31. A.T. James, The relationship algebra of an experimental design. Ann. Math. Stat. **28**, 993–1002 (1957)
32. A.T. James, Analysis of variance determined by symmetry and combinatorial properties of zonal polynomials, in *Statistics and Probability: Essays in Honor of C. R. Rao*, ed. by G. Kallianpur, P.R. Krishnaiah, J.K. Ghosh (North-Holland, Amsterdam, 1982), pp. 329–341
33. A.T. James, Contribution to the discussion of 'Strata for randomized experiments' by R. A. Bailey. J. R. Stat. Soc. Ser. B **53**, 71 (1991)
34. A.T. James, G.N. Wilkinson, Factorization of the residual operator and canonical decomposition of nonorthogonal factors in the analysis of variance. Biometrika **58**, 279–294 (1971)
35. G. James, A. Kerber, *The Representation Theory of the Symmetric Group*. Encyclopedia of Mathematics and Its Applications, vol. 16 (Addison–Wesley, Reading, 1981)
36. P.M.W. John, *Statistical Design and Analysis of Experiments* (Macmillan, New York, 1971)
37. J. Kung, The Radon transform of a combinatorial geometry. J. Comb. Theory Ser. A **26**, 97–102 (1979)
38. W. Ledermann, Representation theory and statistics, in *Séminaire Dubreil-Pisot (Algèbre et Théorie des Nombres)*, vol. 15, University of Paris (1967), pp. 15.01–15.08
39. H.B. Mann, The algebra of a linear hypothesis. Ann. Math. Stat. **31**, 1–15 (1960)
40. D.K. Maslen, D.N. Rockmore, Separation of variables and the computation of Fourier transforms on finite groups, I. J. Am. Math. Soc. **10**(1), 169–214 (1997)
41. D.K. Maslen, M.E. Orrison, D.N. Rockmore, Computing isotypic projections with the Lanczos iteration. SIAM J. Matrix Anal. Appl. **25**, 784–803 (2003)
42. A.D. McLaren, On group representations and invariant stochastic processes. Proc. Camb. Philos. Soc. **59**, 431–450 (1963)
43. J.A. Nelder, The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. Proc. R. Soc. Ser. A **283**, 147–162 (1965)
44. J.A. Nelder, The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance. Proc. R. Soc. Ser. A **283**, 163–178 (1965)
45. A. Neumaier, $t\frac{1}{2}$-Designs. J. Comb. Theory Ser. A **28**, 226–248 (1980)
46. J. Neyman, Statistical problems in agricultural experimentation. J. R. Stat. Soc. Suppl. **2**, 107–154 (1935)

47. S.C. Pearce, *The Agricultural Field Experiment* (Wiley, Chichester, 1983)
48. D.N. Rockmore, Some applications of generalized FFTs, in *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 28, ed. by L. Finkelstein, W. Kantor (American Mathematical Society, Providence, 1997), pp. 329–370
49. G.M. Saha, On Caliński's patterns in block designs. Sankhyā Ser. B **38**, 383–392 (1976)
50. H. Scheffé, *The Analysis of Variance* (Wiley, New York, 1959)
51. J.-P. Serre, *Linear Representations of Finite Groups* (Springer, New York, 1977)
52. T.P. Speed, ANOVA models with random effects: an approach via symmetry, in *Essays in Time Series and Allied Processes: Papers in Honour of E. J. Hannan*, ed. by J. Gani, M.B. Priestly (Applied Probability Trust, Sheffield, 1986), pp. 355–368
53. T.P. Speed, What is an analysis of variance? (with discussion). Ann. Stat. **15**, 885–910 (1987)
54. T. Tjur, Analysis of variance models in orthogonal designs (with discussion). Int. Stat. Rev. **52**, 33–81 (1984)
55. H. Wielandt, *Finite Permutation Groups* (Academic Press, New York, 1964)
56. F. Yates, A new method of arranging variety trials involving a large number of varieties. J. Agric. Sci. **26**, 424–455 (1936)
57. F. Yates, Analysis of data from all possible reciprocal crosses between a set of parental lines. Heredity **1**, 287–301 (1947)

The Synchrosqueezing transform for instantaneous spectral analysis

Gaurav Thakur

Abstract The Synchrosqueezing transform is a time-frequency analysis method that can decompose complex signals into time-varying oscillatory components. It is a form of time-frequency reassignment that is both sparse and invertible, allowing for the recovery of the signal. This article presents an overview of the theory and stability properties of Synchrosqueezing, as well as applications of the technique to topics in cardiology, climate science, and economics.

Key words: Synchrosqueezing, Time-frequency reassignment, Instantaneous frequency, Sparse signal representations

Introduction

The Synchrosqueezing transform is a time-frequency analysis method that can characterize signals with time-varying oscillatory properties. It is designed to analyze and decompose signals of the form

$$f(t) = \sum_{k=1}^{K} A_k(t) e^{2\pi i \phi_k(t)}, \tag{1}$$

where A_k and ϕ_k are time-varying amplitude and phase functions, respectively. The goal is to recover the *instantaneous frequencies* (IFs) $\{\phi_k'\}_{1\le k\le K}$ and the oscillatory components $\{A_k e^{2\pi i \phi_k}\}_{1\le k\le K}$. Signals of the form (1) arise in numerous scientific

G. Thakur (✉)
MITRE Corporation, 7515 Colshire Drive, McLean, VA 22102, USA
e-mail: gthakur@alumni.princeton.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_15

and engineering applications[1] but are not well represented in a traditional Fourier basis, where the individual elements of the basis fail to capture localized oscillations in the components $\{A_k e^{2\pi i \phi_k}\}$. Standard time-frequency methods such as the short-time Fourier transform (STFT) and the continuous wavelet transform (CWT) are often used to analyze such signals, but do not take advantage of any sparsity of the form (1) in the signal and incur a tradeoff in time-frequency resolution [5, 8]. Synchrosqueezing is a variant of time-frequency reassignment (TFR), a class of techniques that apply a nonlinear post-processing mapping to a conventional STFT or CWT plot. The mapping is designed to "push" the energy in an STFT closer to its most prominent frequencies, resulting in a sparse and concentrated time-frequency representation of the signal [2, 9]. However, traditional TFR methods result in a loss of information from the underlying transform and cannot be used to recover the original signal, and also often involve heuristics that are difficult to justify rigorously.

Synchrosqueezing combines the localization and sparsity properties of TFR with the invertibility of a traditional time-frequency transform, and is robust to a variety of disturbances in the signal. The main concepts behind Synchrosqueezing were originally introduced in the mid-1990s for audio signal analysis [6], but it has received much closer attention in recent years, with an extensive mathematical theory developed in [7] and [14]. Unlike traditional TFR, Synchrosqueezing performs the post-processing mapping only in the frequency direction and does so in a manner that preserves the total energy of the signal f, allowing for the decomposition of the signal into the components $\{A_k e^{2\pi i \phi_k}\}$. This article provides a concise survey of the Synchrosqueezing methodology and its associated theory, and also discusses real-world applications in several different domains where the technique has provided new insights.

The Synchrosqueezing process

The Synchrosqueezing transform was originally developed in [7] and [6] in terms of the CWT. We choose a (complex) mother wavelet ψ such that the Fourier transform $\hat{\psi}$ has strictly positive support and satisfies the standard admissibility condition $\int_0^\infty z^{-1} \hat{\psi}(z) dz < \infty$ [5]. The CWT $W_\psi f(a,t)$ at the scale a and time shift t is then given by

$$W_\psi f(a,t) = a^{-1/2} \int_{-\infty}^{\infty} f(u) \overline{\psi\left(\frac{u-t}{a}\right)} du. \tag{2}$$

[1] Such signals are often called "nonstationary" in these domains, although this terminology is not related to its meaning for random processes.

We then take the *phase transform* $\omega f(a,b)$, defined as the derivative of the complex phase of $W_\psi f$,

$$\omega f(a,t) = \frac{\frac{\partial}{\partial t} W_\psi f(a,t)}{2\pi i W_\psi f(a,t)}. \tag{3}$$

Intuitively, this nonlinear operator can be thought of as removing the influence of ψ from the CWT and "encoding" the localized frequency information we want. The key step is to consider the *CWT Synchrosqueezing transform*,

$$S_\varepsilon^{\delta,M} f(t,\eta) = \int_{\{(a,t):a\in[M^{-1},M],|W_\psi f(a,t)|>\varepsilon\}} a^{-3/2} W_\psi f(a,t) \frac{1}{\delta} h\left(\frac{\eta - \omega f(a,t)}{\delta}\right) da \tag{4}$$

for a test function $h \in C_0^\infty$, a sufficiently large parameter M, and sufficiently small $\delta > 0$ and $\varepsilon > 0$. The motivation for (4) is that it is a smoothed out approximation to

$$Sf(t,\eta) = \int_{\{(a,t):\eta=\omega f(a,t)\}} a^{-3/2} W_\psi f(a,t) da,$$

or in other words, a partial inversion of the CWT that is only taken over the level curves of the phase transform ωf and ignores the rest of the time-scale plane (a,t). This localization process allows us to recover the components $A_k e^{2\pi i \phi_k}$ more accurately than inverting the CWT over the entire time-scale plane. Alternatively, the mapping $W_\psi f(a,t) \to Sf(t,\eta)$ can be thought of as a reassignment operation that squeezes energy from the scales a into IFs η centered on the level curves of ωf, but leaves the total energy in $W_\psi f(a,t)$ at each time t unchanged. For appropriate signals f, the energy in the Synchrosqueezing transform $S_\varepsilon^{\delta,M} f(b,\eta)$ is concentrated precisely around the IF curves $\{\phi_k'(t)\}$. Finally, once $S_\varepsilon^{\delta,M} f$ is computed, we can recover each of the components by completing the inversion of the CWT and integrating over small bands around each IF curve,

$$R_{k,\varepsilon}^{\delta,M} f(t) = \frac{1}{\int_0^\infty \frac{\hat{\psi}(z)}{z} dz} \int_{|\eta-\phi_k'(t)|<\varepsilon} S_\varepsilon^{\delta,M} f(t,\eta) d\eta. \tag{5}$$

Under certain conditions, it can be shown that $R_{k,\varepsilon}^{\delta,M} f(t) \approx A_k(t) e^{2\pi i \phi_k(t)}$. In practice, an additional, intermediate step is needed to identify the integration bands in (5), which is typically accomplished by a ridge extraction method that determines the maxima in the time-frequency plot $|S_\varepsilon^{\delta,M} f(t,\eta)|$. A discretized formulation of the steps (2)–(5) and related computational details can be found in [15].

The main concepts behind Synchrosqueezing can also be applied to other underlying time-frequency representations. The paper [14] develops a parallel approach based on the short-time Fourier transform (STFT), which is shown to have some

advantages.[2] The *STFT Synchrosqueezing* process is similar to the above development, but instead of (2) is based on the *modified STFT* for an appropriate window function G,

$$V_G f(t,z) = \int_{-\infty}^{\infty} f(u) G(u-t) e^{-2\pi i z(u-t)} du. \tag{6}$$

This is simply the standard STFT with an additional modulation factor $e^{2\pi i z t}$, and can be thought of as a filter bank taken by sliding the window G over different frequency bands. The phase transform (3) and Synchrosqueezing transform (4), respectively, become

$$\widetilde{\omega} f(z,t) = \frac{\frac{\partial}{\partial t} V_G f(t,z)}{2\pi i V_G f(t,z)},$$

$$\widetilde{S}_{\varepsilon}^{\delta,M} f(t,\eta) = \int_{\{(t,z): z\in[M^{-1},M], |V_G f(t,\eta)|>\varepsilon\}} V_G f(t,z) \frac{1}{\delta} h\left(\frac{\eta - \widetilde{\omega} f(t,z)}{\delta}\right) dz, \tag{7}$$

and the components can be recovered by fully inverting (7) as before by taking

$$\widetilde{R}_{k,\varepsilon}^{\delta,M} f(t) = \frac{1}{\int_{-\infty}^{\infty} |G(z)|^2 dz} \int_{|\eta - \phi_k'| < \varepsilon} \widetilde{S}_{\varepsilon}^{\delta,M} f(t,\eta) d\eta. \tag{8}$$

A simple example of the time-frequency plots $|S_{\varepsilon}^{\delta,M}(t,\eta)|$ and $|\widetilde{S}_{\varepsilon}^{\delta,M} f(t,\eta)|$ is shown in Figure 1. While the traditional STFT and CWT plots are blurry, reflecting the fact that they are not sparse representations of the signal, the Synchrosqueezing transforms have a much more concentrated profile and distinct IF curves in the time-frequency plane. Several additional examples can be found in [15], comparing CWT Synchrosqueezing with TFR methods and other techniques. An open source MATLAB toolbox implementing both forms of Synchrosqueezing is available [3] and has facilitated the use of the technique across different disciplines.

We briefly describe several extensions of these concepts that have been developed. The paper [16] considers a variant of the signal model in (1), where the mode $A_k(t)e^{2\pi i \phi_k(t)}$ is replaced by a more general form $A_k(t)s(\phi_k(t))$ for a given "shape function" s chosen to fit a particular application at hand. This turns out to be a natural model for the analysis of electrocardiogram signals, in which the sharp spikes (see Figure 2) are not well represented by standard Fourier harmonics. In [12], another generalization is presented based on replacing (6) with a "generalized Fourier transform", i.e. an oscillatory integral of the form $\int_{-\infty}^{\infty} f(u) g(u-t) e^{-2\pi i z \theta(u)} du$ where θ is a nonlinear phase function incorporating some prior knowledge of the signal's structure. The paper [19] develops another approach based on wave packet transforms, which encompasses some aspects of both the CWT and STFT formulations.

[2] We present a slightly different formulation of the transform than [14] that is more comparable with the approach in [7].

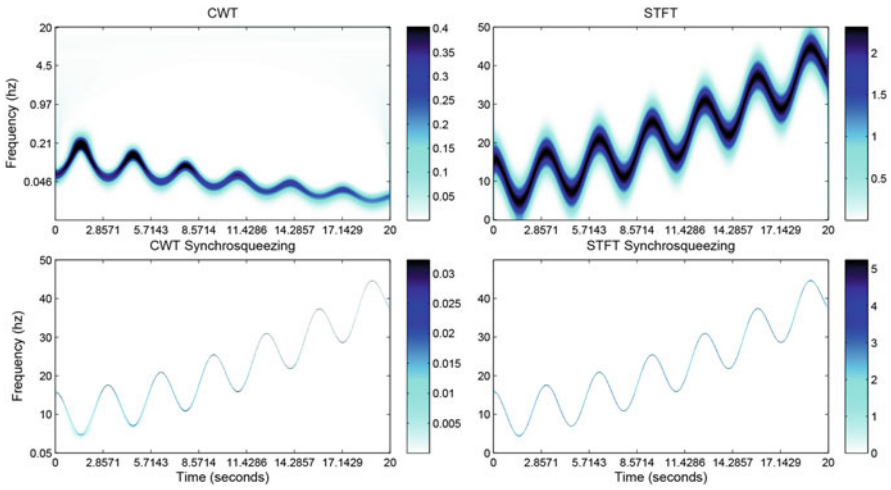


Fig. 1 Time-frequency plots of the signal $f(t) = \cos(2\pi(0.1t^{2.6} + 3\sin(2t) + 10t))$ under different transforms.

Theory

Synchrosqueezing has a fairly comprehensive mathematical theory developed for it, providing performance guarantees on selected classes of signals. As of 2014, most of the published theory in [7] and [15] covers the CWT version (4), but analogous results can be shown for the STFT formulation (7) from [14] using similar techniques. We review the results for the CWT case here, which are based on a sparsity model for the signal (1) in the frequency domain.

Definition 1. For given parameters $\varepsilon, d > 0$, we define the class $\mathscr{A}_{\varepsilon,d} = \{f : f(t) = \sum_{k=1}^{K} A_k(t) e^{2\pi i \phi_k(t)}\}$, where

$$A_k \in L^\infty \cap C^1, \quad \phi_k \in C^2, \quad \phi_k', \phi_k'' \in L^\infty, \quad A_k(t) > 0, \quad \phi_k'(t) > 0$$
$$\forall t \quad \left|A_k'(t)\right| \leq \varepsilon \left|\phi_k'(t)\right|, \quad \left|\phi_k''(t)\right| \leq \varepsilon \left|\phi_k'(t)\right|, \quad \text{and}$$

$$\frac{\phi_k'(t) - \phi_{k-1}'(t)}{\phi_k'(t) + \phi_{k-1}'(t)} \geq d. \tag{9}$$

The key condition here is (9), which says that higher frequency IFs are spaced exponentially further apart than lower frequency IFs. Under this signal model, the following result can be obtained [7].

Theorem 1. *Let $f \in \mathscr{A}_{\varepsilon,d}$ for some $\varepsilon, d > 0$, $h \in C_0^\infty$ with $\|h\|_{L^1} = 1$, and $\psi \in C^1$ with $\hat{\psi}$ supported in $[1-\Delta, 1+\Delta]$ for some $\Delta < \frac{d}{1+d}$. Let M be sufficiently large and*

define $\widetilde{\varepsilon} = \varepsilon^{1/3}$ *and the "scale band"* $Z_k = \{(a,b) : |a\phi_k'(t) - 1| < \Delta\}$. *If* $(a,t) \in Z_k$ *and* $|W_\psi f(a,t)| > \widetilde{\varepsilon}$, *then* $|\omega f(a,t) - \phi_k'(t)| \le \widetilde{\varepsilon}$. *Conversely, if* $(a,t) \notin Z_k$ *for any k, then* $|W_\psi f(a,t)| \le \widetilde{\varepsilon}$. *Furthermore, for some constant* C_1,

$$\left| \lim_{\delta \to 0} R^{\delta,M}_{k,\widetilde{\varepsilon}} f(t) - A_k(t) e^{2\pi i \phi_k(t)} \right| \le C_1 \widetilde{\varepsilon}.$$

This result says that the energy in the Synchrosqueezing time-frequency plane is concentrated around the IF curves $\{\phi_k'(t)\}$, and the inverted components f_k approximate the actual oscillatory components $\{A_k e^{2\pi i \phi_k}\}$. Additional results of this type were proved in [15], describing the robustness of the Synchrosqueezing transform under unstructured perturbations (e.g., quantization error) as well as white noise. We slightly paraphrase these theorems for clarity.

Theorem 2. *Let* f, ε, d, h, ψ, *and* Δ *be given as in Theorem 1. Let* $g = f + E$ *for some error* E *with* $\|E\|_{L^\infty}$ *sufficiently small. There are positive constants* M, C_2, C_3, *and* C_4 *such that the following holds. Let* $a \in [\frac{1}{M}, M]$. *If* $(a,t) \in Z_k$ *and* $|W_\psi g(a,t)| > C_2\widetilde{\varepsilon}$, *then* $|\omega g(a,t) - \phi_k'(t)| \le C_3\widetilde{\varepsilon}$. *Conversely, if* $(a,t) \notin Z_k$ *for any k, then* $|W_\psi g(a,t)| \le C_2\widetilde{\varepsilon}$. *Furthermore,*

$$\left| \lim_{\delta \to 0} R^{\delta,M}_{k,C_2\widetilde{\varepsilon}} g(t) - A_k(t) e^{2\pi i \phi_k(t)} \right| \le C_4 \widetilde{\varepsilon}.$$

Theorem 3. *Let* f, ε, d, h, ψ, *and* Δ *be given as in Theorem 1, with* ψ *also satisfying* $|\langle \psi, \psi' \rangle| < \|\psi\|_{L^2} \|\psi'\|_{L^2}$. *Let* $g = f + N$, *where* N *is Gaussian white noise with power* ε^{2+p} *for some* $p > 0$. *There are positive constants* M, E_1, E_2, C_2', C_3', *and* C_4' *such that the following holds. Let* $a \in [\frac{1}{M}, M]$. *If* $(a,t) \in Z_k$ *and* $|W_\psi g(a,t)| > C_2'\widetilde{\varepsilon}$, *then with probability* $1 - e^{-E_1\varepsilon^{-p}}$, $|\omega g(a,t) - \phi_k'(t)| \le C_3'\widetilde{\varepsilon}$. *Conversely, if* $(a,t) \notin Z_k$ *for any k, then with probability* $1 - e^{-E_2\varepsilon^{-p}}$, $|W_\psi g(a,t)| \le C_2'\widetilde{\varepsilon}$. *Furthermore, with probability* $1 - e^{-E_1\varepsilon^{-p}}$,

$$\left| \lim_{\delta \to 0} R^{\delta,M}_{k,C_2\widetilde{\varepsilon}} g(t) - A_k(t) e^{2\pi i \phi_k(t)} \right| \le C_4' \widetilde{\varepsilon}.$$

For STFT Synchrosqueezing, a result similar to Theorem 1 was proved in [14], although presented in slightly different terms there. The main distinction with the STFT approach is that the theory is developed for a different function class $\mathscr{B}_{\varepsilon,d}$, defined in the same way as $\mathscr{A}_{\varepsilon,d}$ in Definition 1 but with (9) replaced by the weaker separation requirement that $\inf_t \phi_k'(t) - \sup_t \phi_{k-1}'(t) > d$. The linear frequency scale of the modified STFT effectively allows the IF curves $\{\phi_k'\}$ to be spaced much closer to each other than the logarithmic scale of the CWT. In practical terms, STFT Synchrosqueezing is well suited for decomposing signals with multiple components that have closely packed IFs, especially at higher frequencies, while CWT Synchrosqueezing is more appropriate for studying low frequency, trend-like components in a signal.

We finally mention that the above results have mostly been formulated in a deterministic setting, where the signal of interest f is assumed to lie in the class $\mathscr{A}_{\varepsilon,d}$ but without any particular mechanism that generated it. The paper [4] develops extensions of these ideas to a stochastic model of the form $Y(t) = f(t) + T(t) + X(t)$, where f is essentially of the type $\mathscr{A}_{\varepsilon,d}$, T is a slowly varying trend, and X is an autoregressive moving average (ARMA) process with a time-dependent variance. The authors use CWT Synchrosqueezing to extract the components f, T, and X from an observed signal Y, and prove several results on confidence bounds and other aspects of the decomposition.

Applications

Due to its wide applicability, the Synchrosqueezing transform has been used to address problems in many diverse disciplines. The technique was first applied to topics in cardiology, specifically the analysis of electrocardiogram (ECG) signals [14, 17, 18]. The sharp spikes in an ECG signal are called the *R peaks* (see Figure 2) and encode important information about a patient's heart rate, respiration, and many other physiological properties. The analysis of respiration, or breathing characteristics, is important in many clinical applications such as testing for sleep apnea. However, recording the respiration directly requires hooking up a breathing apparatus (ventilator) to the patient and is often impractical to perform over a long period of time. A patient's respiration influences the ECG measurement and can be modeled as a low frequency envelope fitting over the R peaks, with the ECG signal's IF closely following the unobserved respiration signal's IF. The R peaks are not spaced uniformly but can be used to form an impulse train $\sum_k f(t_k)\delta(\cdot - t_k)$, where $\{t_k\}$ are the locations of the R peaks. Applying the STFT Synchrosqueezing transform to this impulse train provides an IF that accurately reflects short-range frequency variations in the respiration signal (Figure 2), and can be used for diagnosing irregularities in the patient's breathing.

Synchrosqueezing has also been used for the analysis of long-term trends in the global climate. The paper [15] studies sediment cores extracted from the ocean floor, in which the relative concentrations of the oxygen isotopes $\delta^{18}O$ and $\delta^{16}O$ indicate changes in the sea level, ice volume, and deep ocean temperature. These are caused by long-term fluctuations in the Earth's eccentricity and other rotational properties over time, known as *Milankovitch cycles*, which influence the amount of solar radiation received at the top of the atmosphere. The CWT Synchrosqueezing transform is used to analyze the $\delta^{18}O$ levels in several composite stacks of cores over the last 2.5 million years (Figure 3). It is able to distinguish the different Milankovitch cycles more accurately than the regular CWT, commonly used in this field, and identify when certain components faded away or became more prominent. The invertibility of the transform also allows one to extract the oscillatory components corresponding to each of the Milankovitch cycles, and better characterize some sudden changes in the climate between 0.5 and 1 million years ago.

Another application of Synchrosqueezing can be found in economics. The paper [10] studies the stability of the US financial system by considering time-frequency decompositions of equity indices, Treasury yields, foreign exchange rates, and several other macroeconomic time series. Each time series is thought of as the output of a dynamical system that produces slowly time-varying frequencies of the form (1), but which are interspersed by abrupt frequency transitions (*structural breaks*) that indicate the starting or stopping of new underlying dynamics. Among other events, the stock market crash in 1987 is contrasted with the global recession in 2008. It is shown that the former had a minimal impact on the dominant, low frequency components despite being prominent in the original data, while the latter was both preceded and followed by a variety of new dynamics, which left the economy in a permanently altered state (Figure 4). The authors also discuss a measure of instability in a time series called the "density index", taking the L^1 norm of the IFs at each point in time as a measure of how spread out or concentrated the frequencies are. A sharp jump in the density index corresponds to a structural break, which is shown to coincide with some of the major financial stress events over the last 25 years and which may provide "early warning" signs of future economic crises.

We briefly mention several other applications of Synchrosqueezing that have appeared in the literature. In [13], it is used to detect and analyze faults in a mechanical gearbox. The Synchrosqueezing plot of the gearbox's vibration signal reveals extra sideband components surrounding a central IF curve, which indicate the presence of a chipped gear in the transmission. In geophysics, [11] discusses the use of Synchrosqueezing to separate out resonant frequencies in data from micro-seismic experiments, which are used to study deformations in injection wells for oil extraction. Finally, [1] develops an automated trading strategy based on Synchrosqueezing, using the technique to model the relationship between correlated asset pairs such as the stocks of competing firms. The rise in one asset's price often precedes a fall in the other one, and a strategy based on identifying the prices' IFs is shown to describe short-range oscillations and outperform some standard approaches used in the industry.

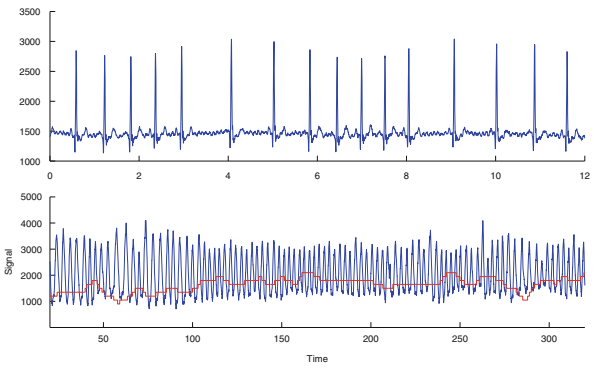


Fig. 2 Top: 10 second portion of ECG signal. Bottom: True respiration signal (blue) and the IF computed from the ECG signal's R peaks (red) using STFT Synchrosqueezing.

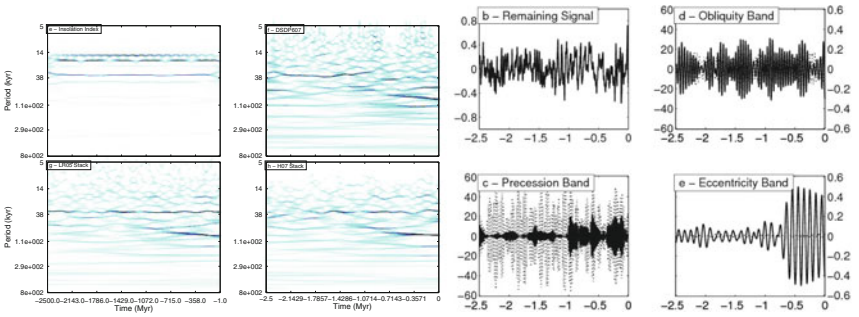


Fig. 3 Left: CWT Synchrosqueezing plots of the insolation index, a single core (DSPD07) and stacks of such cores (LR05 and H07). Right: Reconstructed oscillatory components, corresponding to the obliquity, precession, and eccentricity cycles.

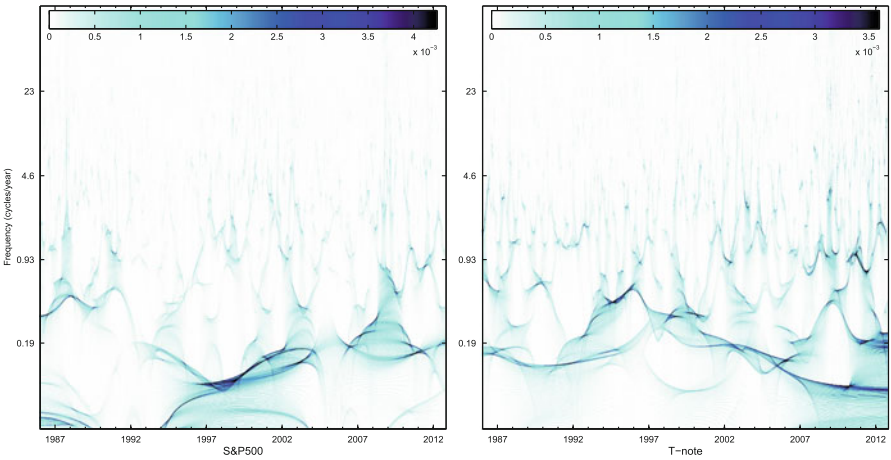


Fig. 4 CWT Synchrosqueezing plots of the S&P 500 price and the 10-year US Treasury yield.

References

1. A. Ahrabian, C.C. Took, D. Mandic, Algorithmic trading using phase synchronization. IEEE J. Sel. Top. Signal Process. **99**, 399–404 (2012)
2. F. Auger, P. Flandrin, Y.-T. Lin, S. McLaughlin, S. Meignen, T. Oberlin, H.-T. Wu, Time-frequency reassignment and synchrosqueezing. IEEE Signal Process. Mag. **30**, 32–41 (2013)
3. E. Brevdo, G. Thakur, H.-T. Wu, The synchrosqueezing toolbox (2013). https://www.web.math.princeton.edu/~ebrevdo/synsq/
4. Y.-C. Chen, M.-Y. Cheng, H.-T. Wu, Nonparametric and adaptive modeling of dynamic periodicity and trend with heteroscedastic and dependent errors. J. R. Stat. Soc. Ser. B **76**(3), 651–682 (2014)
5. I. Daubechies, *Ten Lectures on Wavelets* (Society for Industrial and Applied Mathematics, Philadelphia, 1992)

6. I. Daubechies, S. Maes, A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models, in *Wavelets in Medicine and Biology* ed. by A. Aldroubi, M. Unser (CRC Press, Boca Raton, 1996), pp. 527–546
7. I. Daubechies, J. Lu, H.-T. Wu, Synchrosqueezed wavelet transforms: an empirical mode decomposition-like tool. Appl. Comput. Harmon. Anal. **30**(2), 243–261 (2011)
8. P. Flandrin, *Time-Frequency/Time-Scale Analysis*. *Wavelet Analysis and Its Applications*, vol. 10 (Academic, San Diego, CA, 1999)
9. P. Flandrin, F. Auger, E. Chassande-Mottin, Time-frequency reassignment: from principles to algorithms, in *Applications in Time-Frequency Signal Processing*, ed. by A. Papandreou-Suppappola (CRC, Boca Raton, 2003)
10. S.K. Guharay, G.S. Thakur, F.J. Goodman, S.L. Rosen, D. Houser, Analysis of non-stationary dynamics in the financial system. Econ. Lett. **121**, 454–457 (2013)
11. R.H. Herrera, J.-B. Tary, M. van der Baan, Time-frequency representation of microseismic signals using the Synchrosqueezing transform. GeoConvention (2013). http://www.cspg.org/cspg/Conferences/Geoconvention/2013_Abstract_Archives.aspx
12. C. Li, M. Liang, A generalized synchrosqueezing transform for enhancing signal time-frequency separation. Signal Process. **92**, 2264–2274 (2012)
13. C. Li, M. Liang, Time-frequency analysis for gearbox fault diagnosis using a generalized synchrosqueezing transform. Mech. Syst. Signal Process. **26**, 205–217 (2012)
14. G. Thakur, H.-T. Wu, Synchrosqueezing-based recovery of instantaneous frequency from nonuniform samples. SIAM J. Math. Anal. **43**(5), 2078–2095 (2011)
15. G. Thakur, E. Brevdo, N.-S. Fuckar, H.-T. Wu, The Synchrosqueezing algorithm for time-varying spectral analysis: robustness properties and new paleoclimate applications. Signal Process. **93**, 1079–1094 (2013)
16. H.-T. Wu, Instantaneous frequency and wave shape functions (I). Appl. Comput. Harmon. Anal. **35**, 181–199 (2013)
17. H.-T. Wu, Y.-H. Chan, Y.-T. Lin, Y.-H. Yeh, Using synchrosqueezing transform to discover breathing dynamics from ECG signals. Appl. Comput. Harmon. Anal. **36**(2), 354–359 (2014)
18. H.-T. Wu, S.-S. Hseu, M.-Y. Bien, Y.R. Kou, I. Daubechies, Evaluating the physiological dynamics via Synchrosqueezing: Prediction of the Ventilator Weaning. IEEE Trans. Biomed. Eng. **61**(3), 736–744 (2014)
19. H. Yang, Synchrosqueezed wave packet transforms and diffeomorphism based spectral analysis for 1D general mode decompositions. arXiv:1311.4655 (2013)

Supervised non-negative matrix factorization for audio source separation

Pablo Sprechmann, Alex M. Bronstein, and Guillermo Sapiro

Abstract Source separation is a widely studied problem in signal processing. Despite the permanent progress reported in the literature it is still considered a significant challenge. This chapter first reviews the use of non-negative matrix factorization (NMF) algorithms for solving source separation problems, and proposes a new way for the supervised training in NMF. Matrix factorization methods have received a lot of attention in recent year in the audio processing community, producing particularly good results in source separation. Traditionally, NMF algorithms consist of two separate stages: a training stage, in which a generative model is learned; and a testing stage in which the pre-learned model is used in a high level task such as enhancement, separation, or classification. As an alternative, we propose a task-supervised NMF method for the adaptation of the basis spectra learned in the first stage to enhance the performance on the specific task used in the second stage. We cast this problem as a bilevel optimization program efficiently solved via stochastic gradient descent. The proposed approach is general enough to handle sparsity priors of the activations, and allow non-Euclidean data terms such as β-divergences. The framework is evaluated on speech enhancement.

Key words: Supervised learning, Task-specific learning, Bilevel optimization, NMF, Speech enhancement, Source separation

P. Sprechmann
New York University, New York, NY, USA
e-mail: pablo.sprechmann@nyu.edu

A.M. Bronstein
Tel Aviv University, Tel Aviv, Israel

Duke University, Durham, NC, USA
e-mail: bron@eng.tau.ac.il

G. Sapiro (✉)
Duke University, Durham, NC, USA
e-mail: guillermo.sapiro@duke.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_16

Introduction

The problem of isolating or enhancing an audio signal recorded in a noisy environment has been widely studied in the signal processing community [1, 2]. It becomes particularly challenging in the presence of non-stationary background noise, which is a very common situation in many applications encountered, e.g., in mobile telephony. In this chapter we address the problem of monaural source separation by applying matrix factorization algorithms on a transformed domain given by time-frequency representations of the signals.

The decomposition of time-frequency representations, such as the power or magnitude spectrogram, in terms of elementary atoms of a dictionary, has become a popular tool in audio processing. While many matrix factorization approaches have been used, models imposing non-negativity in their parameters have been proven to be significantly more effective for modeling complex audio mixtures. The non-negativity constraint ensures a parts-based decomposition [3], in which the elementary atoms can be thought as constructive building blocks of the input signal corresponding to interpretable spectral patterns of recurrent events. Non-negative matrix factorization (NMF) [3] and its probabilistic counterpart, the probabilistic latent component analysis (PLCA) [4], are the first instances of a great variety of approaches proposed over the last few years, see [5] for a recent review.

NMF can be applied with different levels of supervision [6, 7]. In this work we are interested in the supervised use of NMF, in which it is assumed that one has access to example audio signals at a training stage. In this setting, NMF is used to take advantage of the available data by pre-computing dictionaries that accurately represent the input signals. NMF has been successfully used in a great variety of audio processing problems ranging from music information retrieval to speech processing. In most approaches, the trained dictionaries are used to facilitate a high level task, such as speech separation [8–12], robust automatic speech recognition [13, 14], source identification [15], and bandwidth extension [16, 17], among many others. In the great majority of these approaches the dictionaries are pre-trained independently as a separate initial step not adapted to the subsequent (and ultimate) high level task. Initial works have recently shown the benefit of incorporating the actual objective of source separation into the training of the model, for example in NMF [18, 19] and deep (and recurrent) neural network based separation [20, 21]. It is worth mentioning that, in the context of classification, NMF has been also trained optimized in a discriminate way [19, 22, 23].

In this chapter we discuss in detail a supervised dictionary learning scheme that can be tailored for different specific high level tasks [18]. Following recent ideas proposed in the context of sparse coding [24], our training scheme is formulated as a bilevel optimization problem, which can be efficiently solved using standard stochastic optimization techniques. We use speech denoising as an example illustrating the power of the proposed framework. However, this technique is general and can be used for various audio applications involving NMF. We also show that these ideas can be employed in general regularized versions of NMF.

This chapter is organized as follows. In Section "Source separation via NMF" we begin by briefly summarizing NMF (and several of its commonly used extensions) in the context of audio source separation. We present the proposed supervised NMF framework in Section "Supervised NMF" and describe how to solve the associated optimization problem in Section "Optimization". Experimental results are presented in Section "Experimental results". In Section "Discussion" we conclude the chapter and discuss future lines of work.

Source separation via NMF

We consider the setting in which we observe a temporal signal $x(t)$ that is the sum of two speech signals $x_i(t)$, with $i = 1, 2$,

$$x(t) = x_1(t) + x_2(t), \tag{1}$$

and we aim at finding estimates $\widehat{x_i}(t)$. Let us define $\mathrm{x} \in \mathbb{R}^N$, a sampled version of the input signal satisfying, $x[n] = x(\frac{n}{f_s})$ with $n = 1, \ldots, N$, where f_s is the sampling rate.

NMF-based source separation techniques typically operate in two stages. First, the signal is represented in a feature space given by a non-linear analysis operator, typically defined (in the case of audio signals) as the magnitude of a time-frequency representation such as the Short-Time Fourier Transform (STFT). Other alternatives have also been explored [19, 25]. Then, a synthesis operator, given by the NMF, is applied to produce an unmixing in the feature space. The separation is obtained by inverting these representations. Performing the separation in the non-linear representation is key to the success of the algorithm. The magnitude of the STFT is in general sparse (simplifying the separation process) and invariant to variations in the phase (local translations), thus freeing the NMF model from learning this irrelevant variability. This comes at the expense of inverting the unmixed estimates in the feature space, which is a well-known problem usually referred to as the phase recovery problem [26].

Let us denote by $\mathrm{V} = \boldsymbol{\Phi}(\mathrm{x}) \in \mathbb{R}^{m \times n}$ a time frequency representation of x, comprising m frequency bins and n (usually overlapping) temporal frames. When the feature extractor $\boldsymbol{\Phi}$ is able to produce sparse representations of the sources (such as in the STFT), the following approximation holds,

$$\boldsymbol{\Phi}(\mathrm{x}) \approx \boldsymbol{\Phi}(\mathrm{x}_1) + \boldsymbol{\Phi}(\mathrm{x}_2),$$

for sufficiently distinct signals. The sum is approximate due to the non-linear effects of the phase. In such a setting, NMF attempts to find the non-negative activations $\mathrm{H}_i \in \mathbb{R}^{q \times n}$, $i = 1, 2$, best representing the different components in two non-negative dictionaries $\mathrm{W}_i \in \mathbb{R}^{m \times q}$. This task is achieved through the solution of the minimization problem

$$\min_{\mathrm{H}_i \geq 0} D(\mathrm{V}|\sum_{i=1,2} \mathrm{W}_i\mathrm{H}_i) + \lambda \sum_{i=1,2} \psi(\mathrm{H}_i)\ . \tag{2}$$

The first term in the optimization objective is a divergence measuring the dissimilarity between the input data V and combination of the estimated channels. Typically, this data fitting term is assumed to be separable,

$$D(\mathrm{A}|\mathrm{B}) = \sum_{i,j} D(a_{ij}|b_{ij}).$$

Significant attention has been devoted in the literature to the case in which the scalar divergence D in the right-hand side belongs to the family of the β-divergences [27],

$$D_\beta(a|b) = \begin{cases} \frac{a}{b} - \log\frac{a}{b} - 1 & : \beta = 0, \\ a\log a/b + (a-b) & : \beta = 1, \\ \frac{1}{\beta(\beta-1)}(a^\beta + (\beta-1)b^\beta - \beta a b^{\beta-1}) & : \text{otherwise.} \end{cases}$$

This family includes the three most widely used cost functions in NMF: the squared Euclidean distance ($\beta = 2$), the Kullback-Leibler divergence ($\beta = 1$), and the Itakura-Saito divergence ($\beta = 0$). For $\beta \geq 1$, the divergence is convex. The case of $\beta = 0$ is attractive despite the lack of convexity, due to the scale-invariance of the Itakura-Saito divergence, which makes the NMF procedure insensitive to volume changes [28].

The second term in the minimization objective is included to promote some desired structure of the activations. This is done using a designed regularization function ψ, whose relative importance is controlled by the parameters λ.

Once the optimal activations are solved for, the spectral envelopes of each source are estimated as $\mathrm{W}_i\mathrm{H}_i$. Since these estimated spectrum envelopes contain no phase information, a subsequent phase recovery stage is necessary. When the non-linearity is imposed as the magnitude of an invertible transform, $\mathscr{F}$, such as the STFT, a simple filtering strategy can be used [12]. In this case we have $\boldsymbol{\Phi}(\mathrm{x}) = |\mathscr{F}\{\mathrm{x}\}|$, where $\mathscr{F}\{\mathrm{x}\} \in \mathbb{C}^{m\times n}$ is a complex matrix. This strategy resembles Wiener filtering and has demonstrated very good results in practice. The recovered spectral envelopes are used to build soft masks to filter the input mixture signal,

$$\hat{\mathrm{x}}_i = \mathscr{F}^{-1}\{\mathrm{M}_i \circ \mathscr{F}\{\mathrm{x}\}\}, \quad \text{with} \quad \mathrm{M}_i = \frac{(\mathrm{W}_i\mathrm{H}_i^*)^p}{\sum_{j=1,2}(\mathrm{W}_j\mathrm{H}_j^*)^p}, \tag{3}$$

where H_i^* are the optimal activations obtained after solving (2), where multiplication denoted $\circ$, division, and exponentials are element-wise operations. The parameter p defines the smoothness of the mask. Note that when p goes to infinity, the mask becomes binary, choosing for each bin the larger of the two signals.

In this section we assumed that the dictionaries for each source were available beforehand for performing the demixing. This corresponds to a supervised version of NMF, in which the dictionaries for each source are trained independently from available training data. Specifically, this is achieved by solving

$$\min_{\mathrm{H}_i, \mathrm{W}_i \geq 0} D(\mathrm{V}_i | \mathrm{W}_i \mathrm{H}_i) + \lambda \, \psi(\mathrm{H}_i) \tag{4}$$

on a training set V_i of feature representations of the unmixed signals for each source.

As mentioned above, the underlying assumption is that the signals forming the mixture, and consequently the learned dictionaries, are sufficiently distinct to be unambiguously decomposed into $\mathrm{V} \approx \sum_{i=1,2} \mathrm{W}_i \mathrm{H}_i$. However, this assumption is often violated in practice, for which we would want to have the dictionaries W_i as incoherent as possible. In other words, the independently trained dictionaries do not ensure that the solutions $\mathrm{W}_1 \mathrm{H}_1$ and $\mathrm{W}_2 \mathrm{H}_2$ obtained from (2) will resemble the original components of the mixture.

Case study

The method proposed in this paper, described in Section "Supervised NMF", can be applied to a large family of approaches following the supervised NMF paradigm. In this paper, we opted to use a sparsity-regularized version of NMF as a case study. In this case, the regularizer ψ in (2) is given by the columns-wise ℓ_1 norm,

$$\psi(\mathrm{H}) = \lambda \|\mathrm{H}\|_1 + \frac{\mu}{2} \|\mathrm{H}\|_2^2. \tag{5}$$

For technical reasons, that will be clear in Section "Optimization", we also include an ℓ_2 regularizer on the activations.

Supervised NMF

As was discussed in the previous section, the optimization problem (5) is merely a proxy to the desired estimation problem. Standard dictionary learning applied independently to each source does not guarantee that its solutions will produce the best estimate of the unmixed sources even on mixtures created from the training data. Ideally, we would like to train dictionaries that explicitly maximize the performance directly on the source separation problem. In this section we describe a way of better posing this problem in the context of NMF.

Given a mixed input signal, x, the method described in Section "Source separation via NMF" defines an estimator of the signal components $\hat{\mathrm{x}}_i(\mathrm{W}_1, \mathrm{W}_2, \mathrm{x})$, where we made explicit their dependence on the dictionaries and the input signal. Ideally we would like to train the signal dictionaries to minimize the expected estimation risk of the estimation, for example, in terms of the mean squared error (MSE),

$$\{\mathrm{W}_i\}_{i=1,2} = \operatorname*{argmin}_{\mathrm{W}_i \geq 0} \sum_{i=1,2} \mathbb{E}_{\mathrm{x}_1, \mathrm{x}_2} \left\{ \|\mathrm{x}_i - \hat{\mathrm{x}}_i(\mathrm{W}_1, \mathrm{W}_2, \mathrm{x}_1 + \mathrm{x}_2)\|^2 \right\}.$$

Assuming that the signals are independent, we can write this expression as

$$\{\mathrm{W}_i\}_{i=1,2} = \underset{\mathrm{W}_i \geq 0}{\operatorname{argmin}} \int\int \sum_{i=1,2} \|\mathrm{x}_i - \hat{\mathrm{x}}_i(\mathrm{W}_1, \mathrm{W}_2, \mathrm{x}_1 + \mathrm{x}_2)\|^2 dP(\mathrm{x}_1) dP(\mathrm{x}_2),$$

where P are the distributions of each source. In practice, these distributions are latent; a common strategy to overcome this problem is to approximate the expected risk by computing the empirical risk over a finite set of training examples sampled from the source distributions. In what follows, we denote by $\mathscr{X}_i$ the available sets of training signals for each source. Then, the empirical risk is given by

$$\{\mathrm{W}_i\}_{i=1,2} = \underset{\mathrm{W}_i \geq 0}{\operatorname{argmin}} \frac{1}{|\mathscr{X}|} \sum_k \sum_{i=1,2} \|\mathrm{x}_i^k - \hat{\mathrm{x}}_i^k(\mathrm{W}_1, \mathrm{W}_2, \mathrm{x}^k)\|_2^2; \tag{6}$$

where the first sum (with the index k) goes over the elements in the product set, $\mathscr{X} = \mathscr{X}_1 \times \mathscr{X}_2$, containing all possible pairs of training signals. We used $\mathrm{x}^k = \mathrm{x}_1^k + \mathrm{x}_2^k$ to simplify the notation. While the empirical risk measures the performance of the estimators over the training set, the expected risk measures the expected performance over new data samples following the same distribution, that is, the generalization capabilities of the model. We can expect a good generalization when sufficient representative training data are available in advance.

When the feature space is given by an invertible transformation, the MSE in (6) can be computed in the (complex) transformed domain. From Parseval's theorem it follows that (6) is equivalent to

$$\{\mathrm{W}_i\}_{i=1,2} = \underset{\mathrm{W}_i \geq 0}{\operatorname{argmin}} \frac{1}{|\mathscr{X}|} \sum_k \sum_{i=1,2} \|\mathscr{F}\{\mathrm{x}_i^k\} - \mathrm{M}_i(\mathrm{W}_1, \mathrm{W}_2, \mathrm{x}^k) \mathscr{F}\{\mathrm{x}^k\}\|^2. \tag{7}$$

Note that the transformed representations $\mathscr{F}\{\mathrm{x}_i^k\}$ of the signals are complex.

As it was discussed in Section "Source separation via NMF", the standard setting for supervised NMF estimates the signal dictionaries independently solving (4) for each source. This approximation is pragmatic rather than principled, since the empirical loss given in (6) (or (7)) is difficult to compute. While the estimators $\hat{\mathrm{x}}_i$ (or the masks M_i) are functions of the dictionaries and the mixture signal, they cannot be computed in closed form as they depend on the solution of the optimization problem (2). Such optimization problems are referred to as *bilevel*. In the following section we describe how to solve the bilevel NMF dictionary learning problem when the divergence used in (2) is a convex $\beta-$divergence with appropriate regularization.

Finally, we note that another difficulty posed by the proposed training regime (common to any discriminative approach to source separation [19, 20]) is that the estimation of the dictionaries needs to be computed over the product set rather than each training set independently. This naturally increases the computational load of the training stage, however, it might not be a serious limitation as this can be done in an offline manner without affecting the computational load at testing time.

Optimization

As in any empirical risk minimization task, both formulations (6) and (7) are written as the average over a training set of a given cost function. We are going to adopt the formulation in the frequency domain, given in (7), since it has the additional advantage that can be easily separable on a frame-wise manner.

For now, we will assume that the regularizer in (2) is frame-wise separable, and defer the discussion of the more general case to Section "Implementation details". In this way, the cost function of the NMF problem also becomes frame-wise separable. In order to alleviate the notation, we are going to write the minimization of the empirical risk over a collection of frames rather than the actual audio signals. With this notation, the training data are composed by the set $\mathscr{X}_f$ containing pairs of frames of the form $(\mathrm{f}_1^j, \mathrm{f}_2^j)$, being $\mathrm{f}_i^j \in \mathbb{C}^m$ the $j-$th frame in the collection, corresponding to one column of the time frequency representation, $\mathscr{F}\{\mathrm{x}_i^k\}$, of some signal, x_i^k, in the original training set of signals $\mathscr{X}_i$. Now we denote the mixture as $\mathrm{f}^j = \mathrm{f}_1^j + \mathrm{f}_2^j$. Let us define the loss function

$$\ell(\mathrm{f}_1, \mathrm{f}_2, \mathrm{W}_1, \mathrm{W}_2, \mathrm{h}_1^*, \mathrm{h}_2^*) = \sum_{i=1,2} \|\mathrm{f}_i - \mathrm{M}_i(\mathrm{W}_1, \mathrm{W}_2, \mathrm{f}, \mathrm{h}_1^*, \mathrm{h}_2^*)\,\mathrm{f}\|^2, \tag{8}$$

where we made explicit the dependency of ℓ and the masks on the optimal activations h_1^* and h_2^*. These optimal activations are themselves functions of the input mixture and the dictionaries, $\mathrm{h}_i^* = \mathrm{h}_i^*(\mathrm{f}, \mathrm{W}_1, \mathrm{W}_2)$, and are obtained by solving the frame-wise version of (2) given by

$$\{\mathrm{h}_i^*\}_{i=1,2} = \underset{\mathrm{h}_i \geq 0}{\operatorname{argmin}}\, D_\beta(\mathrm{v} | \sum_{i=1,2} \mathrm{W}_i \mathrm{h}_i) + \sum_{i=1,2} \lambda\, \psi(\mathrm{h}_i)\ , \tag{9}$$

where, following the previous notation, $\mathrm{v} = \boldsymbol{\Phi}(\mathrm{f})$, and we explicitly wrote a ridge regression term controlled by the non-negative parameter μ. This is included to guarantee that (9) is strictly convex and has a unique solution. The supervised NMF problem can be stated as the optimization program given by

$$\{\mathrm{W}_i\}_{i=1,2} = \underset{\mathrm{W}_i \geq 0}{\operatorname{argmin}} \frac{1}{|\mathscr{X}_f|} \sum_j \ell(\mathrm{f}_1^j, \mathrm{f}_2^j, \mathrm{W}_1, \mathrm{W}_2, \mathrm{h}_1^*, \mathrm{h}_2^*). \tag{10}$$

This optimization problem is referred to as bilevel, with (10) and (9) being the high and low level problems, respectively. It is important to notice that while (10) depends on knowing the ground truth demixing, (9) only depends on the mixture signal, hence matching exactly the situation encountered at testing. As NMF itself, this bilevel optimization problem is non-convex. Hence, we aim at finding a good local minimizer. In what follows, we describe the general optimization algorithm used for this purpose.

Stochastic gradient descent

Problem (9) has a unique solution when $\beta \geq 1$ and $\mu > 0$, due to the strict convexity of the objective. In this situation, a local minimizer of (10) can be found via (projected) stochastic gradient descent (SGD) [29]. SGD is a gradient descent optimization algorithm for minimizing an objective function expressed as a sum or average of some training data of an almost-everywhere differentiable function. At each iteration, the gradient of the objective function is approximated using a randomly picked sub-sample.

At iteration j we randomly draw a sample pair from the training set of frames $\mathscr{X}_f$ and sum them together to obtain a mixture sample in the feature space, $\mathrm{v}^j = \boldsymbol{\Phi}(\mathrm{f}^j)$. Then the combined dictionary at iteration $j+1$, $\mathrm{W}^{j+1} = [\mathrm{W}_1^{j+1}, \mathrm{W}_2^{j+1}]$, is obtained by

$$\mathrm{W}^{j+1} \leftarrow \mathscr{P}(\mathrm{W}^j - \eta_j \nabla_{\mathrm{W}} \ell(\mathrm{f}_1^j, \mathrm{f}_2^j, \mathrm{W}_1^j, \mathrm{W}_2^j, \mathrm{h}_1^{*j}, \mathrm{h}_2^{*j}),$$

where $0 \leq \eta_i \leq \eta$ is a decreasing sequence of step-sizes, and $\mathscr{P}$ is a projection operator making the argument matrix be non-negative with column having the norm smaller or equal than one. Note that the learning requires the gradient $\nabla_{\mathrm{W}} \ell$, which in turn relies (via the chain rule) on the gradients of $\nabla_{\mathrm{M}_i} \ell$, $\nabla_{\mathrm{h}_i^*} \mathrm{M}_i$, and $\nabla_{\mathrm{W}} \mathrm{h}_i^*(\mathrm{v}, \mathrm{W})$. As in the context of dictionary learning for sparse coding [24], even though the h_i^* are obtained by solving a non-smooth optimization problem, they are almost everywhere differentiable, and one can compute their gradient with respect to W in a closed form. In the next section, we summarize the derivation of the gradients $\nabla_{\mathrm{W}} \ell$.

Following [24], we use a step size of the form $\eta_i = \eta \min(1, i_0/i)$ in all our experiments, which means that a fixed step size is used during the first i_0 iterations, after which it decays according to the $1/i$ annealing strategy. We set in all our experiments i_0 to be half of the total number of iterations. However, other standard tools commonly used in SGD optimization, such as momentum, could also be used. A common heuristic used in practice for accelerating the convergence speed of SGD algorithms consists in randomly drawing several samples (a mini batch) at each iteration instead of a single one. A natural initialization of the speech and noise dictionaries is the individual training via the solution of (4), as in standard supervised NMF denoising.

Gradient computation

Let us denote by ρ the objective function in (9),

$$\rho(\mathrm{W}, \mathrm{h}) = D_\beta(\mathrm{v}|\mathrm{Wh}) + \sum_{i=1,2} \lambda \psi(\mathrm{h}_i) + \mu ||\mathrm{h}_i||_2^2,$$

where, for simplicity, we define the vector $\mathrm{h} = [\mathrm{h}_1; \mathrm{h}_2]$ (using Matlab-like notation), containing the column-concatenated activations for each source, such that the product of h with the row-concatenated matrix $\mathrm{W} = [\mathrm{W}_1, \mathrm{W}_2]$ is well defined. Let us denote by Λ the active set of the solution of (9), that is, the indices of the non-zero coefficients of h^*. We use the sub-index Λ to indicate the sub-vector restricted to the active set, e.g., h^*_Λ. The first-order optimality conditions of (9) require the derivatives with respect to h_Λ to be zero,

$$\mathrm{h}^* \geq 0, \quad \nabla_{\mathrm{h}} \rho(\mathrm{W}, \mathrm{h}^*) \geq 0, \quad \mathrm{h}^* \circ \nabla_{\mathrm{h}} \rho(\mathrm{W}, \mathrm{h}^*) = 0, \tag{11}$$

where $\circ$ denotes element-wise multiplication (Hadamard product). For each coefficient in the active set of any stationary point of (9), the partial derivative of ρ with respect to that coefficient needs to be zero. Hence, if we look only at the active set, we have

$$[\nabla_{\mathrm{h}} \rho(\mathrm{W}, \mathrm{h}^*)]_\Lambda = \mathrm{W}_\Lambda^{\mathrm{T}} \Phi + \lambda \nabla_{\mathrm{h}} \sum_{i=1,2} \psi(\mathrm{h}_i^*)_\Lambda + \mu \, \mathrm{h}^*_\Lambda = 0, \tag{12}$$

where W_Λ is the matrix retaining only the columns of the dictionary associated with the active set, and $\Phi = (\mathrm{W}_\Lambda \mathrm{h}^*_\Lambda)^{\beta-2} \circ (\mathrm{W}_\Lambda \mathrm{h}^*_\Lambda - \mathrm{v})$. When ψ is the ℓ_1 norm as in the case of study described in Section "Case study", the derivative of the regularization term, $\nabla_{\mathrm{h}} \psi(\mathrm{h}_i) = \mathrm{p}$, is equal to a constant vector that assumes the value of one on the coefficients of Λ and zero otherwise.

For a given coordinate, say indexed by r, the conditions given in (11) imply three cases, either only one of $[\mathrm{h}^*]_r$ or $[\nabla_{\mathrm{h}} \rho(\mathrm{W}, \mathrm{h}^*)]_r$ are zero or both are. As it was shown in the sparse coding context [24], a key observation is that, almost surely, the set of active constraints in the solution of (9) remains constant on a local neighborhood of v and W. That is, for small changes in the dictionary, the active set Λ remains constant. The only points in which h^* is non-differentiable are points where the active set changes.

Hence, we know that only the gradient $\nabla_{\mathrm{W}_\Lambda} \mathrm{h}^*$ will be non-zero, that is, changes in the columns of W that do not affect the coefficients in Λ do not affect the cost function. Since we cannot write h^* in closed form as a function of W, we need to perform implicit differentiation. Taking the derivative in (12) with respect to W_Λ we obtain

$$d\mathrm{W}_\Lambda^{\mathrm{T}} \phi + \mathrm{W}_\Lambda^{\mathrm{T}} \Phi (d\mathrm{W}_\Lambda \mathrm{h}^*_\Lambda + \mathrm{W}_\Lambda \, d\mathrm{h}^*_\Lambda) + \mu \, d\mathrm{h}^*_\Lambda = 0, \tag{13}$$

where we used d to denote the differentials, and

$$\Phi = \mathrm{diag}\Big((\mathrm{W}_\Lambda \mathrm{h}^*_\Lambda)^{\beta-2} + (\beta - 2)(\mathrm{W}_\Lambda \mathrm{h}^*_\Lambda)^{\beta-3} \circ (\mathrm{W}_\Lambda \mathrm{h}^*_\Lambda - \mathrm{v}) \Big). \tag{14}$$

We can obtain an expression for $d\mathrm{h}^*_\Lambda$ from (13) as

$$d\mathrm{h}^*_\Lambda = \mathrm{Q} (d\mathrm{W}_\Lambda^{\mathrm{T}} \phi + \mathrm{W}_\Lambda^{\mathrm{T}} \Phi d\mathrm{W}_\Lambda \mathrm{h}^*_\Lambda), \tag{15}$$

where $\mathrm{Q} = (\mathrm{W}_\Lambda^{\mathrm{T}} \Phi \mathrm{W}_\Lambda + \mu \mathrm{I})^{-1}$. Note that the size of the matrix being inverted is given by the sparsity level of the representation. Now we can proceed to compute

the gradient of the loss function with respect to the dictionary. Invoking the chain rule, we have

$$\nabla_{\mathrm{W}}\ell = \mathrm{trace}(\nabla_{\mathrm{h}^*}\ell^{\mathrm{T}} d\mathrm{h}^*) + \nabla_{\mathrm{W}}\hat{\ell}, \tag{16}$$

where $\nabla_{\mathrm{W}}\hat{\ell}$ represents the gradient of ℓ with respect to W assuming h^* fixed. To compute the gradient $\nabla_{\mathrm{h}^*}\ell$ one has to also use the chain rule considering the definition of the masks given in (3). Combining (15) and (16) and using the properties of the trace function, it follows that

$$\nabla_{\mathrm{W}}\ell = \phi\,\xi^{\mathrm{T}} + \Phi\mathrm{W}_{\Lambda}\,\xi\,\mathrm{h}_{\Lambda}^{*\,\mathrm{T}} + \nabla_{\mathrm{W}}\hat{\ell}, \tag{17}$$

where $\xi = \mathrm{Q}\nabla_{\mathrm{h}^*}\ell$.

Implementation details

There are a few important implementation that need to be considered in practice. First, the $\beta-$divergences are not differentiable at zero when $\beta \leq 2$. A common way to solve this problem is to consider a translated version of the divergence instead, which is obtained by adding a small constant in the second argument,

$$\widetilde{D}_{\beta}(a|b) = D_{\beta}(a|b+\delta)$$

where $\delta > 0$ is a small constant. In our experiments we used $\delta = 0.001$. It is worth mentioning that this is common practice even in every setting of NMF in order to avoid instabilities produced by extremely large values.

During the iterations of the SGD algorithm, the estimation of the gradient of the cost function on the current sample (or mini-batch) requires the computation of the optimal activations h^* by solving (9). The precision with which these activations are computed is very important for obtaining meaningful gradients. In that sense, it is preferable to use algorithms with fast converge rates, for example the least angle regression (LARS) in the case of $\beta = 2$ [30], or the alternating method of multipliers (ADMM) [31] in the case of $\beta \leq 2$. While running multiplicative algorithms for a small number of iterations produces satisfactory results when running NMF for separation, their slow convergence rate makes them extremely inefficient in this case, requiring a very large number of iterations for computing meaningful gradients.

Experimental results

Data sets. We evaluated the separation performance of the proposed methods on a subset of the GRID dataset [32]. Three randomly chosen sets of distinct clips each were used for training (500 clips), validation (10 clips), and testing (50 clips).

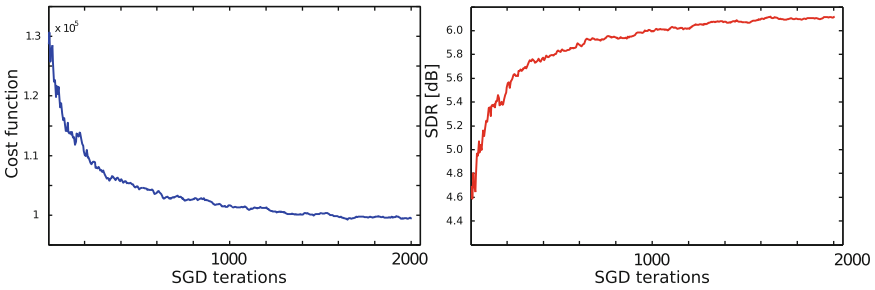


Fig. 1 Evolution of the average high level cost function (left) and the average SDR (in dB) on the validation set (mixed at $SNR = 0dB$) with the SGD iterations for task-specific NMF with $\beta = 1$.

The clips were resampled to 8 KHz. For the noise signals we used the AURORA corpus [33], which contains six categories of noise recorded from different real environments (street, restaurant, car, exhibition, train, and airport). Three sets of distinct clips each were used for training (15 clips), validation (3 clips), and testing (15 clips).

Evaluation measures. As the evaluation criteria, we used the *source-to-distortion ratio* (SDR), *source-to-interference ratio* (SIR), and *source-to-artifact ratio* (SAR) from the BSS-EVAL metrics [34]. We also computed the standard *signal-to-noise ratio* (SNR). When dealing with several frames, we computed a global score (GSDR, GSIR, GSAR, and GSNR) by averaging the metrics over all test clips from the same speaker and noise weighted by the clip duration.

The goal of this experiment was to apply the proposed approach in the context of audio denoising. Here the noise is considered as a source and modeled explicitly. We used dictionaries of size 60 and 10 atoms for representing the speech and the noise, respectively. These values were obtained using cross-validation. We used different values of the parameter λ for the signal and the noise, namely $\lambda_s = 0.1$ for speech and $\lambda_n = 0$ for the noise (the latter means that no sparsity was promoted in the representation of the noise) and $\mu = 0.001$. As an example, we used $\beta = 1$ and $\beta = 0$, and $\alpha = 0$ in the high level cost (10). For the SGD algorithm we used $\eta = 0.1$ and minibatch of size 50. These were obtained by trying several values of during a small number of iterations, keeping those producing the lowest error on a small validation set. All training signals were mixed at 5 dB.

Results. Figure 1 shows the evolution of the high level cost (10) and the SDR on the validation set with the SGD iterations. The algorithm converges to a dictionary that achieves about 2 dB better SDR on the validation set, this behavior is also verified on the test set. Tables 1 and 2 show results for the proposed approach on the test setting. We compare the performance of standard supervised sparse-NMF (referred simply as NMF) against the performance of the same model trained in the proposed task-specific manner (referred as TS-NMF) on denoising two with different SNR levels. Observe that the task-specific supervision leads to improvements in performance, maintaining (at 5dB SNR) the improvements observed on the

Table 1 Average performance (in dB) for NMF and proposed supervised NMF methods measured in terms of SDR, SIR, SAR, and SNR. Speech and noise were mixed at $5dB$ of SNR. The standard deviation of each result is shown in brackets.

	SDR	SIR	SAR	SNR
NMF $\beta = 1$	7.5 [1.5]	13.7 [0.9]	8.9 [1.7]	8.2 [1.3]
TS-NMF $\beta = 1$	9.5 [1.4]	15.2 [0.7]	11.0 [1.7]	10.0 [1.2]
TS-NMF $\beta = 0$	8.6 [1.3]	14.1 [1.2]	10.3 [1.5]	9.1 [1.1]

Table 2 See description of Table 1. In this case, speech and noise were mixed at $0dB$ of SNR.

	SDR	SIR	SAR	SNR
NMF $\beta = 1$	4.5 [1.1]	9.3 [0.9]	6.8 [1.2]	5.8 [0.8]
TS-NMF $\beta = 1$	6.3 [1.0]	11.9 [0.7]	8.0 [1.1]	7.2 [0.8]
TS-NMF $\beta = 0$	5.2 [1.2]	12.0 [1.7]	6.6 [1.2]	6.3 [0.9]

validation set. Interestingly, the method also works when using $\beta = 0$ (Itakura-Saito), even if the developments in Section "Optimization" are technically not valid in this case, since the divergence is not convex. While the non-convexity of the problem implies that there might be multiple minimums, we initialize the pursuit algorithm always with the exact same initial condition (all zeros). Intuitively, one can expect that a small perturbation on the dictionary will the local minims of the solution change slightly and consequently the algorithm will still converge to the same (perturbed) minimum.

Discussion

In this chapter we reviewed the use of NMF for solving source separation problems. We discussed different ways of solving the supervised training of the NMF model and proposed an algorithm for the task-supervised training of NMF models following the ideas introduced in [24] in the context of sparse coding. Unlike standard supervised NMF, the proposed approach matches the optimization objective used at the train and testing stages. In this way, the dictionaries can be trained to optimize the performance of the specific task. We cast this problem as bilevel optimization that can be efficiently solved via stochastic gradient descent. The proposed approach allows non-Euclidean data terms such as β-divergences. A simple case study of sparse-NMF with task specific supervision demonstrates promising results.

Acknowledgments

Work partially supported by ONR, NSF, NGA, AFOSR, BSF, ARO, and ERC.

References

1. P.C. Loizou, *Speech Enhancement: Theory and Practice*, vol. 30 (CRC, Boca Raton, 2007)
2. E. Hänsler, G. Schmidt, *Speech and Audio Processing in Adverse Environments* (Springer, Berlin, 2008)
3. D.D. Lee, H.S. Seung, Learning parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
4. P. Smaragdis, B. Raj, M. Shashanka, A probabilistic latent variable model for acoustic modeling, in *NIPS*, vol. 148, 2006
5. P. Smaragdis, C. Fevotte, G Mysore, N. Mohammadiha, M. Hoffman, Static and dynamic source separation using nonnegative factorizations: a unified view. IEEE Signal Process. Mag. **31**(3), 66–75 (2014)
6. P. Smaragdis, B. Raj, M. Shashanka, Supervised and semi-supervised separation of sounds from single-channel mixtures, in *Independent Component Analysis and Signal Separation*, ed. by M.K. Davies, C.J. James, S.A. Abdallah, M.D. Plumbley (Springer, Berlin, 2007), pp. 414–421
7. N. Mohammadiha, P. Smaragdis, A. Leijon, Supervised and unsupervised speech enhancement using nonnegative matrix factorization. IEEE Trans. Audio Speech Lang. Process. **21**(10), 2140–2151 (2013)
8. M.N. Schmidt, R.K. Olsson, Single-channel speech separation using sparse non-negative matrix factorization, in *INTERSPEECH*, Sept 2006
9. M.V.S. Shashanka, B. Raj, P. Smaragdis, Sparse overcomplete decomposition for single channel speaker separation, in *ICASSP*, 2007
10. C. Joder, F. Weninger, F. Eyben, D. Virette, B. Schuller, Real-time speech separation by semi-supervised nonnegative matrix factorization, in *LVA/ICA* (2012), pp. 322–329
11. Z. Duan, G.J. Mysore, P. Smaragdis, Online plca for real-time semi-supervised source separation, in *LVA/ICA* (2012), pp. 34–41
12. M.N. Schmidt, J. Larsen, F.-T. Hsiao, Wind noise reduction using non-negative sparse coding, in *MLSP*, pp. 431–436, Aug 2007
13. J.F. Gemmeke, T. Virtanen, A. Hurmalainen, Exemplar-based sparse representations for noise robust automatic speech recognition. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2067–2080 (2011)
14. F. Weninger, M. Wöllmer, J.T. Geiger, B. Schuller, J.F. Gemmeke, A. Hurmalainen, T. Virtanen, G. Rigoll, Non-negative matrix factorization for highly noise-robust asr: to enhance or to recognize? in *ICASSP* (2012), pp. 4681–4684
15. P. Sprechmann, I. Ramirez, P. Cancela, G. Sapiro, Collaborative sources identification in mixed signals via hierarchical sparse modeling, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011* . IEEE (2011), pp. 5816–5819
16. D. Bansal, B. Raj, P. Smaragdis, Bandwidth expansion of narrowband speech using non-negative matrix factorization, in *INTERSPEECH* (2005), pp. 1505–1508
17. J. Han, G.J. Mysore, B. Pardo, Audio imputation using the non-negative hidden markov model, in *LVA/ICA* (2012), pp. 347–355
18. P. Sprechmann, A.M. Bronstein, G. Sapiro, Supervised non-euclidean sparse NMF via bilevel optimization with applications to speech enhancement, in *HSCMA*. IEEE (2014), pp. 11–15
19. F. Weninger, J. Le Roux, J.R. Hershey, S. Watanabe, Discriminative NMF and its application to single-channel source separation, in *Proceedings of ISCA Interspeech*, 2014
20. P.-S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, Deep learning for monaural speech separation, in *ICASSP* (2014), pp. 1562–1566
21. F. Weninger, J. Le Roux, J.R. Hershey, B. Schuller, Discriminatively trained recurrent neural networks for single-channel speech separation, in *Proceedings of IEEE GlobalSIP 2014 Symposium on Machine Learning Applications in Speech Processing*, 2014
22. N. Boulanger-Lewandowski, Y. Bengio, P. Vincent, Discriminative non-negative matrix factorization for multiple pitch estimation, in *ISMIR*. Citeseer (2012), pp. 205–210

23. T. Ben Yakar, P. Sprechmann, R. Litman, A.M. Bronstein, G. Sapiro, Bilevel sparse models for polyphonic music transcription, in *ISMIR* (2013), pp. 65–70
24. J. Mairal, F. Bach, J. Ponce, Task-driven dictionary learning. IEEE Trans. Pattern Anal. Mach. Intel. **34**(4), 791–804 (2012)
25. J. Bruna, P. Sprechmann, Y. Le Cun, Source separation with scattering non-negative matrix factorization, in *ICASSP*, 2015
26. R.W. Gerchberg, W. Owen Saxton, A practical algorithm for the determination of the phase from image and diffraction plane pictures. Optik **35**, 237–246 (1972)
27. C. Févotte, J. Idier, Algorithms for nonnegative matrix factorization with the β-divergence. Neural Comput. **23**(9), 2421–2456 (2011)
28. C. Févotte, N. Bertin, J.-L. Durrieu, Nonnegative matrix factorization with the itakura-saito divergence. With application to music analysis. Neural Comput. **21**(3), 793–830 (2009)
29. B. Colson, P. Marcotte, G. Savard, An overview of bilevel optimization. Ann. Oper. Res. **153**(1), 235–256 (2007)
30. B. Efron, T. Hastie, I. Johnstone, R. Tibshirani et al., Least angle regression. Ann. Stat. **32**(2), 407–499 (2004)
31. D.L. Sun, C. Févotte, Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014
32. M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. **120**, 2421 (2006)
33. D. Pearce, H.-G. Hirsch, The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions, in *INTERSPEECH* (2000), pp. 29–32
34. E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)

Index

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7

Applied and Numerical Harmonic Analysis (71 volumes)

A. Saichev and W.A. Woyczyñski: *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)
O. Bratteli, P. Jorgensen, and B. Treadway: *Wavelets Through a Looking Glass* (ISBN 978-0-8176-4280-80

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7

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. Woyczyski: *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. Begué, 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. Begué, J. Benedetto, W. Czaja, and K.A Okoudjou: *Excursions in Harmonic Analysis, Volume 4* (ISBN 978-3-319-20187-0)

For an up-to-date list of ANHA titles, please visit **http://www.springer.com/series/4968**