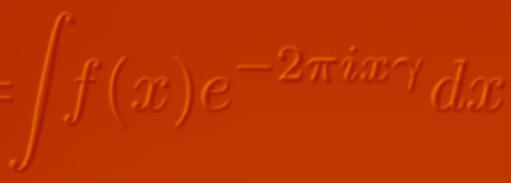
Applied and Numerical Harmonic Analysis

Holger Boche
Giuseppe Caire
Robert Calderbank
Maximilian März
Gitta Kutyniok
Rudolf Mathar
Editors



Compressed Sensing and its Applications

Second International MATHEON Conference 2015

Birkhäuser

Applied and Numerical Harmonic Analysis

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

Holger Boche • Giuseppe Caire • Robert Calderbank
Maximilian März • Gitta Kutyniok • Rudolf Mathar
Editors

Compressed Sensing and its Applications

Second International MATHEON Conference 2015

Birkhäuser

Editors
Holger Boche
Fakultät für Elektrotechnik und
Informationstechnik
Technische Universität München
Munich, Bavaria, Germany

Robert Calderbank
Department of Electrical
& Computer Engineering
Duke University
Durham, North Carolina, USA

Gitta Kutyniok
Institut für Mathematik
Technische Universität Berlin
Berlin, Germany

Giuseppe Caire
Institut für Telekommunikationssysteme
Technische Universität Berlin
Berlin, Germany

Maximilian März
Institut für Mathematik
Technische Universität Berlin
Berlin, Germany

Rudolf Mathar
Lehrstuhl und Institute für Statistik
RWTH Aachen
Aachen, Germany

ISSN 2296-5009 ISSN 2296-5017 (electronic)
Applied and Numerical Harmonic Analysis
ISBN 978-3-319-69801-4 ISBN 978-3-319-69802-1 (eBook)
https://doi.org/10.1007/978-3-319-69802-1

Library of Congress Control Number: 2017960841

Mathematics Subject Classification (2010): 94A12, 94A20, 68U10, 90C25, 15B52

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

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, e.g., by the mechanical synthesizers in tidal analysis.

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

Nowadays, some of the theory of PDEs has given way to the study of Fourier integral operators. Problems in antenna theory are studied in terms of unimodular trigonometric polynomials. Applications of Fourier analysis abound in signal processing, whether with the fast Fourier transform (FFT), or filter design, or the adaptive modeling inherent in time-frequency-scale methods such as wavelet theory.

The coherent states of mathematical physics are translated and modulated Fourier transforms, and these are used, in conjunction with the uncertainty principle, for dealing with signal reconstruction in communications theory. We are back to the raison d'être of the *ANHA* series!

University of Maryland
College Park, MD, USA

John J. Benedetto
Series Editor

Preface

The key challenge of compressed sensing arises from the task of recovering signals from a small collection of measurements, exploiting the fact that high-dimensional signals are typically governed by intrinsic low-complexity structures, for instance, being sparse in an orthonormal basis. While the reconstruction from such compressed, typically randomly selected measurements is well studied from a theoretical perspective, there also exist numerous efficient recovery algorithms exhibiting excellent practical performance and thereby making compressed sensing relevant to many different applications. In fact, from an early stage on, the field has greatly benefited from the interaction between mathematics, engineering, computer science, and physics, leading to new theoretical insights as well as significant improvements of real-world applications.

From the point of view of applied mathematics, the field makes use of tools from applied harmonic analysis, approximation theory, linear algebra, convex optimization, and probability theory, while the applications encompass many areas such as image processing, sensor networks, radar technology, quantum computing, or statistical learning, to name just a very few. Nowadays, it is fair to say that more than 10 years after its emergence, the field of compressed sensing has reached a mature state, where many of the underlying mathematical foundations are quite well understood. Therefore, some of the techniques and results are now being transferred to other related areas, leading to a broader conception of compressed sensing and opening up new possibilities for applications.

In December 2015, the editors of this volume organized the *Second International MATHEON Conference on Compressed Sensing and its Applications* at the Technische Universität Berlin. This conference was supported by the research center for *Mathematics for Key Technologies* (MATHEON), as well as the *German Research Foundation (DFG)*. It was attended by more than 150 participants from 18 different countries, and as in the first workshop of this series in 2013, experts in a variety of different research areas were present. This diverse background of participants led to a very fruitful exchange of ideas and to stimulating discussions.

This book is the second volume in the *Applied and Numerical Harmonic Analysis* book series on *Compressed Sensing and its Applications,* presenting state-of-the-art

monographs on various topics in compressed sensing and related fields. It is aimed at a broad readership, reaching from graduate students to senior researchers in applied mathematics, engineering, and computer science.

This volume features contributions by two of the plenary speakers (chapters "On the Global-Local Dichotomy in Sparsity Modeling" and "Fourier Phase Retrieval: Uniqueness and Algorithms"), namely, Michael Elad (Technion—Israel Institute of Technology) and Yonina C. Eldar (Technion—Israel Institute of Technology), and by ten invited speakers (chapters "Compressed Sensing Approaches for Polynomial Approximation of High-Dimensional Functions," "Multisection in the Stochastic Block Model Using Semidefinite Programming," "Recovering Signals with Unknown Sparsity in Multiple Dictionaries," "Compressive Classification and the Rare Eclipse Problem," "Weak Phase Retrieval," "Cubatures on Grassmannians: Moments, Dimension Reduction, and Related Topics," "A Randomized Tensor Train Singular Value Decomposition," "Versatile and Scalable Cosparse Methods for Physics-Driven Inverse Problems," "Total Variation Minimization in Compressed Sensing," "Compressed Sensing in Hilbert Spaces"), namely, Ben Adcock (Simon Fraser University), Alfonso S. Bandeira (Massachusetts Institute of Technology), Peter G. Casazza (University of Missouri, Columbia), Mike E. Davies (University of Edinburgh), Martin Ehler (University of Vienna), Rémi Gribonval (INRIA Rennes), Felix Krahmer (Technische Universität München), Dustin G. Mixon (Air Force Institute of Technology), Reinhold Schneider (Technische Universität Berlin), and Philip Schniter (The Ohio State University, Columbus).

In the following, we will give a brief outline of the content of each chapter. For an introduction and a self-contained overview on compressed sensing and its major achievements, we refer the reader to chapter "A Survey of Compressed Sensing" of the first volume of this book series (*Boche, H., Calderbank, R., Kutyniok, G., and Vybiral, J. (eds.), Compressed Sensing and its Applications: MATHEON Workshop 2013. Birkhäuser Boston, 2015).*

Two of the chapters focus on phase retrieval: chapter "Fourier Phase Retrieval: Uniqueness and Algorithms" contains a detailed overview on Fourier phase retrieval and practical algorithms, whereas chapter "Weak Phase Retrieval" introduces a weaker formulation of the classical phase retrieval. Another key topic is the question how sparsity-promoting transformations are used in compressed sensing. In this realm, chapter "On the Global-Local Dichotomy in Sparsity Modeling" analyzes the gap between local and global sparsity in dictionaries, chapter "Versatile and Scalable Cosparse Methods for Physics-Driven Inverse Problems" focuses on the use of the analysis formulation in physics-driven inverse problems, chapter "Total Variation Minimization in Compressed Sensing" gives an overview over total variation minimization in compressed sensing, and chapter "Recovering Signals with Unknown Sparsity in Multiple Dictionaries" uses iterative reweighting for recovering signals with unknown sparsity in multiple dictionaries. Several chapters focus entirely on mathematical aspects, such as chapter "Compressed Sensing Approaches for Polynomial Approximation of High-Dimensional Functions" which exploits compressed sensing for approximating functions with polynomials. In chapter "Compressed Sensing in Hilbert Spaces," compressed sensing is considered in the abstract

framework of Hilbert spaces, and chapter "Compressive Classification and the Rare Eclipse Problem" deals with random projections of convex sets. The other chapters study new frontiers in related areas, such as detecting community-like structures in graphs via the stochastic block model (chapter "Multisection in the Stochastic Block Model Using Semidefinite Programming"), cubatures on Grassmannians and their connection to the recovery of sparse probability measures (chapter "Cubatures on Grassmannians: Moments, Dimension Reduction, and Related Topics"), and an examination of randomized tensor train singular value decompositions (chapter "A Randomized Tensor Train Singular Value Decomposition").

We would like to thank the following current and former members of the research group "Applied Functional Analysis" at the Technische Universität Berlin without whom this conference would not have been possible: Axel Flinth, Martin Genzel, Mijail Guillemard, Anja Hedrich, Sandra Keiper, Anton Kolleck, Maximilian Leitheiser, Jackie Ma, Philipp Petersen, Friedrich Philipp, Mones Raslan, Martin Schäfer, and Yizhi Sun.

München, Germany — Holger Boche
Berlin, Germany — Giuseppe Caire
Durham, USA — Robert Calderbank
Berlin, Germany — Gitta Kutyniok
Berlin, Germany — Maximilian März
Aachen, Germany — Rudolf Mathar
July 2017

Contents

On the Global-Local Dichotomy in Sparsity Modeling

Dmitry Batenkov, Yaniv Romano, and Michael Elad

Abstract The traditional sparse modeling approach, when applied to inverse problems with large data such as images, essentially assumes a sparse model for small overlapping data patches and processes these patches as if they were independent from each other. While producing state-of-the-art results, this methodology is suboptimal, as it does not attempt to model the entire global signal in any meaningful way—a nontrivial task by itself.

In this paper we propose a way to bridge this theoretical gap by constructing a global model from the bottom-up. Given local sparsity assumptions in a dictionary, we show that the global signal representation must satisfy a constrained underdetermined system of linear equations, which forces the patches to agree on the overlaps. Furthermore, we show that the corresponding global pursuit can be solved via local operations. We investigate conditions for unique and stable recovery and provide numerical evidence corroborating the theory.

Keywords Sparse representations · Inverse problems · Convolutional sparse coding

D. Batenkov (✉)
Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
e-mail: batenkov@mit.edu

Y. Romano
Department of Electrical Engineering, Technion - Israel Institute of Technology, 32000 Haifa, Israel
e-mail: yromano@tx.technion.ac.il

M. Elad
Department of Computer Science, Technion - Israel Institute of Technology, 32000 Haifa, Israel
e-mail: elad@cs.technion.ac.il

H. Boche et al. (eds.), *Compressed Sensing and its Applications*,
Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-69802-1_1

1 Introduction

1.1 The Need for a New Local-Global Sparsity Theory

The sparse representation model [17] provides a powerful approach to various inverse problems in image and signal processing such as denoising [18, 37], deblurring [14, 57], and super-resolution [47, 56], to name a few [38]. This model assumes that a signal can be represented as a sparse linear combination of a few columns (called atoms) taken from a matrix termed dictionary. Given a signal, the sparse recovery of its representation over a dictionary is called sparse coding or pursuit (such as the orthogonal matching pursuit, OMP, or basis pursuit, BP). Due to computational and theoretical aspects, when treating high-dimensional data, various existing sparsity-inspired methods utilize local patched-based representations rather than the global ones, i.e., they divide a signal into small overlapping blocks (patches), reconstruct these patches using standard sparse recovery techniques, and subsequently average the overlapping regions [11, 17]. While this approach leads to highly efficient algorithms producing state-of-the-art results, the global signal prior remains essentially unexploited, potentially resulting in suboptimal recovery.

As an attempt to tackle this flaw, methods based on the notion of *structured sparsity* [19, 29, 30, 32, 55] started to appear; for example, in [14, 37, 47] the observation that a patch may have similar neighbors in its surroundings (often termed the self-similarity property) is injected to the pursuit, leading to improved local estimations. Another possibility to consider the dependencies between patches is to exploit the multi-scale nature of the signals [36, 40, 53]. A different direction is suggested by the expected patch log likelihood (EPLL) method [40, 52, 60], which encourages the patches of the final estimate (i.e., after the application of the averaging step) to comply with the local prior. Also, a related work [45, 46] suggests promoting the local estimations to agree on their shared content (the overlap) as a way to achieve a coherent reconstruction of the signal.

Recently, an alternative to the traditional patch-based prior was suggested in the form of the convolutional, or shift-invariant, sparse coding (CSC) model [10, 25, 27, 28, 49, 54]. Rather than dividing the image into local patches and processing each of these independently, this approach imposes a specific structure on the global dictionary—a concatenation of banded circulant matrices—and applies a global pursuit. A thorough theoretical analysis of this model was proposed very recently in [41, 42], providing a clear understanding of its success.

The empirical success of the above algorithms indicates the great potential of reducing the inherent gap that exists between the independent local processing of patches and the global nature of the signal at hand. However, a key and highly desirable part is still missing—a theory which would suggest how to modify the basic sparse model to take into account the mutual dependencies between the patches, what approximation methods to use, and how to efficiently design and learn the corresponding structured dictionary.

1.2 Content and Organization of the Paper

In this paper we propose a systematic investigation of the signals which are implicitly defined by local sparsity assumptions. A major theme in what follows is that the presence of patch overlaps reduces the number of degrees of freedom, which, in turn, has theoretical and practical implications. In particular, this allows more accurate estimates for uniqueness and stability of local sparse representations, as well as better bounds on performance of existing sparse approximation algorithms. Moreover, the global point of view allows for development of new pursuit algorithms, which consist of local operation on one hand, while also taking into account the patch overlaps on the other hand. Some aspects of the offered theory are still incomplete, and several exciting research directions emerge as well.

The paper is organized as follows. In Section 2 we develop the basic framework for signals which are patch-sparse, building the global model from the "bottom-up," and discuss some theoretical properties of the resulting model. In Section 3 we consider the questions of reconstructing the representation vector and of denoising a signal in this new framework. We describe "globalized" greedy pursuit algorithms [43] for these tasks, where the patch disagreements play a major role. We show that the frequently used local patch averaging (LPA) approach is in fact suboptimal in this case. In Section 4 and Appendix E: Generative Models for Patch-Sparse Signals, we describe several instances/classes of the local-global model in some detail, exemplifying the preceding definitions and results. The examples include piecewise constant signals, signature-type (periodic) signals, and more general bottom-up models. In Section 5 we present results of some numerical experiments, where in particular we show that one of the new globalized pursuits, inspired by the ADMM algorithm [9, 23, 24, 33], turns out to have superior performance in all the cases considered. We conclude the paper in Section 6 by discussing possible research directions.

2 Local-Global Sparsity

We start with the local sparsity assumptions for every patch and subsequently provide two complimentary characterizations of the resulting global signal space. On one hand, we show that the signals of interest admit a global "sparse-like" representation with a dictionary of convolutional type and with additional linear constraints on the representation vector. On the other hand, the signal space is in fact a union of linear subspaces, where each subspace is a kernel of a certain linear map. To complement and connect these points of view, in Appendix E: Generative Models for Patch-Sparse Signals, we show that the original local dictionary must carry a combinatorial structure, and based on this structure, we develop a generative model for patch-sparse signals. Concluding this section, we provide some theoretical analysis of the properties of the resulting model, in particular uniqueness and

stability of representation. For this task, we define certain measures of the dictionary, similar to the classical spark, coherence function, and the restricted isometry property, which take the additional dictionary structure into account. In general, this additional structure implies possibly better uniqueness as well as stability to perturbations; however, it is an open question to show they are provably better in certain cases.

2.1 Preliminaries

Let $[m]$ denote the set $\{1, 2, \ldots, m\}$. If D is an $n \times m$ matrix and $S \subset [m]$ is an index set, then D_S denotes the submatrix of D consisting of the columns indexed by S.

Definition 1 (Spark of a Matrix). Given a dictionary $D \in \mathbb{R}^{n\times m}$, the *spark* of D is defined as the minimal number of columns which are linearly dependent:

$$\sigma(D) := \min\{j : \exists S \subset [m],\ |S| = j,\ \operatorname{rank} D_S < j\}. \tag{1}$$

Clearly $\sigma(D) \leqslant n + 1$.

Definition 2. Given a vector $\alpha \in \mathbb{R}^m$, the ℓ_0 pseudo-norm is the number of nonzero elements in α:

$$\|\alpha\|_0 := \#\{j : \alpha_j \neq 0\}.$$

Definition 3. Let $D \in \mathbb{R}^{n\times m}$ be a dictionary with normalized atoms. The μ_1 coherence function (Tropp's Babel function) is defined as

$$\mu_1(s) := \max_{i\in[m]} \max_{S\subset[m]\setminus\{i\},\ |S|=s} \sum_{j\in S} \left|\langle d_i, d_j\rangle\right|.$$

Definition 4. Given a dictionary D as above, the restricted isometry constant of order k is the smallest number δ_k such that

$$(1 - \delta_k)\|\alpha\|_2^2 \leqslant \|D\alpha\|_2^2 \leqslant (1 + \delta_k)\|\alpha\|_2^2$$

for every $\alpha \in \mathbb{R}^m$ with $\|\alpha\|_0 \leqslant k$.

For any matrix M, we denote by $\mathscr{R}(M)$ the column space (range) of M.

2.2 Globalized Local Model

In what follows we treat one-dimensional signals $x \in \mathbb{R}^N$ of length N, divided into $P = N$ overlapping patches of equal size n (so that the original signal is thought

to be periodically extended). The other natural choice is $P = N - n + 1$, but for simplicity of derivations, we consider only the periodic case.

Let $R_1 := \left[I_{n\times n}\ \mathbf{0}\ \mathbf{0}\ \ldots\ \mathbf{0}\right] \in \mathbb{R}^{n\times N}$, and for each $i = 2, \ldots, P$, we define $R_i \in \mathbb{R}^{n\times N}$ to be the circular column shift of R_1 by $n \cdot (i-1)$ entries, i.e., this operator extracts the i-th patch from the signal in a circular fashion.

Definition 5. Given local dictionary $D \in \mathbb{R}^{n\times m}$, sparsity level $s < n$, signal length N, and the number of overlapping patches P, the *globalized local sparse* model is the set

$$\mathscr{M} = \mathscr{M}(D, s, P, N) := \left\{x \in \mathbb{R}^N,\ R_i x = D\alpha_i,\ \|\alpha_i\|_0 \leqslant s\ \forall i = 1, \ldots, P\right\}. \tag{2}$$

This model suggests that each patch, $R_i x$ is assumed to have an s-sparse representation α_i, and this way we have characterized the global x by describing the local nature of its patches.

Next we derive a "global" characterization of $\mathscr{M}$. Starting with the equations

$$R_i x = D\alpha_i, \qquad i = 1, \ldots, P,$$

and using the equality $I_{N\times N} = \frac{1}{n}\sum_{i=1}^{P} R_i^T R_i$, we have a representation

$$x = \frac{1}{n}\sum_{i=1}^{P} R_i^T R_i x = \sum_{i=1}^{P}\left(\frac{1}{n}R_i^T D\right)\alpha_i.$$

Let the global "convolutional" dictionary D_G be defined as the horizontal concatenation of the (vertically) shifted versions of $\frac{1}{n}D$, i.e., (see Figure 1 on page 5)

$$D_G := \left[\left(\frac{1}{n}R_i^T D\right)\right]_{i=1,\ldots P} \in \mathbb{R}^{N\times mP}. \tag{3}$$

Let $\Gamma \in \mathbb{R}^{mP}$ denote the concatenation of the local sparse codes, i.e.,

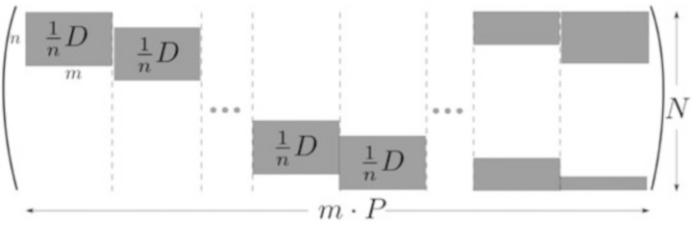


Fig. 1 The global dictionary D_G. After permuting the columns, the matrix becomes a union of circulant Toeplitz matrices, hence the term "convolutional".

$$\Gamma := \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_P \end{bmatrix}.$$

Given a vector Γ as above, we will denote by $\tilde{R}_i$ the operator of extracting its i-th portion,[1], i.e., $\tilde{R}_i \Gamma \equiv \alpha_i$.

Summarizing the above developments, we have the global convolutional representation for our signal as follows:

$$x = D_G \Gamma. \tag{4}$$

Next, applying R_i to both sides of (4) and using (2), we obtain

$$D\alpha_i = R_i x = R_i D_G \Gamma. \tag{5}$$

Let $\Omega_i := R_i D_G$ denote the i-th stripe from the global convolutional dictionary D_G. Thus (5) can be rewritten as

$$\underbrace{\left[\mathbf{0} \ldots \mathbf{0}\ D\ \mathbf{0} \ldots \mathbf{0}\right]}_{:=Q_i} \Gamma = \Omega_i \Gamma, \tag{6}$$

or $(Q_i - \Omega_i)\,\Gamma = 0$. Since this is true for all $i = 1, \ldots, P$, we have shown that the vector Γ satisfies

$$\underbrace{\begin{bmatrix} Q_1 - \Omega_1 \\ \vdots \\ Q_P - \Omega_P \end{bmatrix}}_{:=M \in \mathbb{R}^{nP \times mP}} \Gamma = 0.$$

Thus, the condition that the patches $R_i x$ agree on the overlaps is equivalent to the global representation vector Γ residing in the null-space of the matrix M.

An easy computation provides the dimension of this null-space (see proof in Appendix A: Proof of Lemma 1), or in other words the overall number of degrees of freedom of admissible Γ.

[1]Notice that while R_i extracts the i-th patch from the signal x, the operator $\tilde{R}_i$ extracts the representation α_i of $R_i x$ from Γ.

Lemma 1. *For any frame $D \in \mathbb{R}^{n\times m}$ (i.e., a full rank dictionary), we have*

$$\dim\ker M = N\,(m-n+1)\,.$$

Note that in particular for $m = n$, we have $\dim\ker M = N$, and since in this case D is invertible, we have $R_i x = D\alpha_i$ where $\alpha_i = D^{-1}R_i x$, so that every signal admits a unique representation $x = D_G\Gamma$ with $\Gamma = \left(D^{-1}R_1 x, \ldots, D^{-1}R_P x\right)^T$.

As we shall demonstrate now, the equation $M\Gamma = 0$ represents the requirement that the local sparse codes $\{\alpha_i\}$ are not independent but rather should be such that the corresponding patches $D\alpha_i$ agree on the overlaps.

Definition 6. Define the "extract from top/bottom" operators $S_T \in \mathbb{R}^{(n-1)\times n}$ and $S_B \in \mathbb{R}^{(n-1)\times n}$:

$$S_{T(op)} = \left[I_{n-1}\ \mathbf{0}\right], \quad S_{B(ottom)} = \left[\mathbf{0}\ I_{n-1}\right].$$

The following result is proved in Appendix B: Proof of Lemma 2.

Lemma 2. *Let $\Gamma = [\alpha_1, \ldots, \alpha_P]^T$. Under the above definitions, the following are equivalent:*

1. *$M\Gamma = 0$;*
2. *For each $i = 1, \ldots, P$, we have $S_B D\alpha_i = S_T D\alpha_{i+1}$.*

Definition 7. Given $\Gamma = [\alpha_1, \ldots, \alpha_P]^T \in \mathbb{R}^{mP}$, the $\|\cdot\|_{0,\infty}$ pseudo-norm is defined by

$$\|\Gamma\|_{0,\infty} := \max_{i=1,\ldots,P} \|\alpha_i\|_0.$$

Thus, every signal complying with the patch-sparse model, with sparsity s for each patch, admits the following representation.

Theorem 1. *Given D, s, P, and N, the globalized local sparse model (2) is equivalent to*

$$\begin{aligned}\mathcal{M} &= \left\{x \in \mathbb{R}^N :\ x = D_G\Gamma,\ M\Gamma = 0,\ \|\Gamma\|_{0,\infty} \leqslant s\right\} \\ &= \left\{x \in \mathbb{R}^N :\ x = D_G\Gamma,\ M_*\Gamma = 0,\ \|\Gamma\|_{0,\infty} \leqslant s\right\},\end{aligned} \tag{7}$$

where the matrix $M_* \in \mathbb{R}^{(n-1)P\times mP}$ is defined as

$$M_* := \begin{bmatrix} S_B D & -S_T D & & \\ & S_B D & -S_T D & \\ & & \ddots & \ddots \end{bmatrix}.$$

Proof. If $x \in \mathscr{M}$ (according to (2)), then by the above construction x belongs to the set defined by the RHS of (7) (let's call it $\mathscr{M}^*$ for the purposes of this proof only). In the other direction, assume that $x \in \mathscr{M}^*$. Now $R_i x = R_i D_G \Gamma = \Omega_i \Gamma$, and since $M\Gamma = 0$, we have $R_i x = Q_i \Gamma = D\tilde{R}_i \Gamma$. Denote $\alpha_i := \tilde{R}_i \Gamma$, and so we have that $R_i x = D\alpha_i$ with $\|\alpha_i\|_0 \leqslant s$, i.e., $x \in \mathscr{M}$ by definition. The second part follows from Lemma 2. □

We say that α_i is a *minimal* representation of x_i if $x_i = D\alpha_i$ such that the matrix $D_{\operatorname{supp}\alpha_i}$ has full rank—and therefore the atoms participating in the representation are linearly independent.[2]

Definition 8. Given a signal $x \in \mathscr{M}$, let us denote by $\rho(x)$ the set of all locally sparse *and minimal* representations of x:

$$\rho(x) := \left\{ \Gamma \in \mathbb{R}^{mP} : \ \|\Gamma\|_{0,\infty} \leqslant s, \ x = D_G \Gamma, \ M\Gamma = 0, \ D_{\operatorname{supp}\tilde{R}_i\Gamma} \text{ is full rank} \right\}.$$

Let us now go back to the definition (2). Consider a signal $x \in \mathscr{M}$, and let $\Gamma \in \rho(x)$. Denote $S_i := \operatorname{supp}\tilde{R}_i\Gamma$. Then we have $R_i x \in \mathscr{R}(D_{S_i})$, and therefore we can write $R_i x = P_{S_i} R_i x$, where P_{S_i} is the orthogonal projection operator onto $\mathscr{R}(D_{S_i})$. In fact, since D_{S_i} is full rank, we have $P_{s_i} = D_{S_i} D_{S_i}^{\dagger}$ where $D_{S_i}^{\dagger} = \left(D_{S_i}^T D_{S_i}\right)^{-1} D_{S_i}^T$ is the Moore-Penrose pseudoinverse of D_{S_i}.

Definition 9. Given a support sequence $\mathscr{S} = (S_1, \ldots, S_P)$, define the matrix $A_{\mathscr{S}}$ as follows:

$$A_{\mathscr{S}} := \begin{bmatrix} (I_n - P_{S_1}) R_1 \\ (I_n - P_{S_2}) R_2 \\ \vdots \\ (I_n - P_{S_P}) R_P \end{bmatrix} \in \mathbb{R}^{nP \times N}.$$

The map $A_{\mathscr{S}}$ measures the local patch discrepancies, i.e., how "far" is each local patch from the range of a particular subset of the columns of D.

Definition 10. Given a model $\mathscr{M}$, denote by $\Sigma_{\mathscr{M}}$ the set of all valid supports, i.e.,

$$\Sigma_{\mathscr{M}} := \{(S_1, \ldots, S_P) : \ \exists x \in \mathscr{M}, \ \Gamma amma \in \rho(x) \ \text{ s.t. } \forall i = 1, \ldots, P : \\ S_i = \operatorname{supp}\tilde{R}_i \Gamma \}.$$

With this notation in place, it is immediate to see that the global signal model is a union of subspaces.

Theorem 2. *The global model is equivalent to the union of subspaces*

$$\mathscr{M} = \bigcup_{\mathscr{S} \in \Sigma_{\mathscr{M}}} \ker A_{\mathscr{S}}.$$

[2]Notice that α_i might be a minimal representation but not a unique one with minimal sparsity. For discussion of uniqueness, see Subsection 2.3.

Remark 1. Contrary to the well-known union of subspaces model [7, 35], the subspaces $\{\ker A_{\mathscr{S}}\}$ do not have in general a sparse joint basis, and therefore our model is distinctly different from the well-known block-sparsity model [19, 20].

An important question of interest is to estimate $\dim \ker A_{\mathscr{S}}$ for a given $\mathscr{S} \in \Sigma_{\mathscr{M}}$. One possible solution is to investigate the "global" structure of the corresponding signals (as is done in Subsection 4.1 and Subsection 4.2), while another option is to utilize information about "local connections" (Appendix E: Generative Models for Patch-Sparse Signals).

2.3 Uniqueness and Stability

Given a signal $x \in \mathscr{M}$, it has a globalized representation $\Gamma \in \rho(x)$ according to Theorem 1. When is such a representation unique, and under what conditions can it be recovered when the signal is corrupted with noise?

In other words, we study the problem

$$\min \|\Gamma\|_{0,\infty} \qquad \text{s.t. } D_G\Gamma = D_G\Gamma_0,\ M\Gamma = 0 \qquad (P_{0,\infty})$$

and its noisy version

$$\min \|\Gamma\|_{0,\infty} \qquad \text{s.t. } \|D_G\Gamma - D_G\Gamma_0\| \leqslant \varepsilon,\ M\Gamma = 0 \qquad \left(P^{\varepsilon}_{0,\infty}\right).$$

For this task, we define certain measures of the dictionary, similar to the classical spark, coherence function, and the restricted isometry property, which take the additional dictionary structure into account. In general, the additional structure implies *possibly* better uniqueness as well as stability to perturbations; however, it is an open question to show they are *provably* better in certain cases.

The key observation is that the global model $\mathscr{M}$ imposes a constraint on the allowed local supports.

Definition 11. Denote the set of allowed local supports by

$$\mathscr{T} := \{T :\ \exists\,(S_1, \ldots, T, \ldots, S_P) \in \Sigma_{\mathscr{M}}\}.$$

Recall the definition of the spark (1). Clearly $\sigma(D)$ can be equivalently rewritten as

$$\sigma(D) = \min\{j :\ \exists S_1, S_2 \subset [m],\ |S_1 \cup S_2| = j,\ \operatorname{rank} D_{S_1 \cup S_2} < j\}. \tag{8}$$

Definition 12. The *globalized spark* $\sigma^*(D)$ is

$$\sigma^*(D) := \min\{j :\ \exists S_1, S_2 \in \mathscr{T},\ |S_1 \cup S_2| = j,\ \operatorname{rank} D_{S_1 \cup S_2} < j\}. \tag{9}$$

The following proposition is immediate by comparing (8) with (9).

Proposition 1. $\sigma^*(D) \geqslant \sigma(D)$.
The globalized spark provides a uniqueness result in the spirit of [15].

Theorem 3 (Uniqueness). *Let $x \in \mathscr{M}(D, s, N, P)$. If there exists $\Gamma \in \rho(x)$ for which $\|\Gamma\|_{0,\infty} < \frac{1}{2}\sigma^*(D)$ (i.e., it is a sufficiently sparse solution of $P_{0,\infty}$), then it is the unique solution (and so $\rho(x) = \{\Gamma\}$).*

Proof. Suppose that there exists $\Gamma_0 \in \rho(x)$ which is different from Γ. Put $\Gamma_1 := \Gamma - \Gamma_0$, then $\|\Gamma_1\|_{0,\infty} < \sigma^*(D)$, while $D_G\Gamma_1 = 0$ and $M\Gamma_1 = 0$. Denote $\beta_j := \tilde{R}_j\Gamma_1$. By assumption, there exists an index i for which $\beta_i \neq 0$, but we must have $D\beta_j = 0$ for every j, and therefore $D_{\operatorname{supp}\beta_i}$ must be rank-deficient—contradicting the fact that $\|\beta_i\| < \sigma^*(D)$. □

In classical sparsity, we have the bound

$$\sigma(D) \geqslant \min\{s : \ \mu_1(s-1) \geqslant 1\}, \tag{10}$$

where μ_1 is given by Definition 3. In a similar fashion, the globalized spark σ^* can be bounded by an appropriate analog of "coherence"—however, computing this new coherence appears to be in general intractable.

Definition 13. Given the model $\mathscr{M}$, we define the following globalized coherence function

$$\mu_1^*(s) := \max_{S \in \mathscr{T} \cup \mathscr{T}, |S| = s} \max_{j \in S} \sum_{k \in S \setminus \{j\}} \left|\langle d_j, d_k \rangle\right|,$$

where $\mathscr{T} \cup \mathscr{T} := \{S_1 \cup S_2 : \ S_1, S_2 \in \mathscr{T}\}$.

Theorem 4. *The globalized spark σ^* can be bounded by the globalized coherence as follows*[3]*:*

$$\sigma^*(D) \geqslant \min\{s : \ \mu_1^*(s) \geqslant 1\}.$$

Proof. Following closely the corresponding proof in [15], assume by contradiction that

$$\sigma^*(D) < \min\{s : \ \mu_1^*(s) \geqslant 1\}.$$

Let $S^* \in \mathscr{T} \cup \mathscr{T}$ with $|S^*| = \sigma^*(D)$ for which D_{s^*} is rank-deficient. Then the restricted Gram matrix $G := D_{S^*}^T D_{S^*}$ must be singular. On the other hand, $\mu_1^*(|S^*|) < 1$, and so in particular

$$\max_{j \in S^*} \sum_{k \in S^* \setminus \{j\}} \left|\langle d_j, d_k \rangle\right| < 1.$$

[3] In general $\min\{s : \ \mu_1^*(s-1) \geqslant 1\} \neq \max\{s : \ \mu_1^*(s) < 1\}$ because the function μ_1^* need not be monotonic.

But that means that G is diagonally dominant and therefore $\det G \neq 0$, a contradiction. □

We see that $\mu_1^*(s+1) \leqslant \mu_1(s)$ since the outer maximization is done on a smaller set. Therefore, in general the bound of Theorem 4 appears to be sharper than (10).

A notion of globalized RIP can also be defined as follows.

Definition 14. The globalized RIP constant of order k associated to the model $\mathscr{M}$ is the smallest number $\delta_{k,\mathscr{M}}$ such that

$$(1-\delta_{k,\mathscr{M}})\|\alpha\|_2^2 \leqslant \|D\alpha\|_2^2 \leqslant (1+\delta_{k,\mathscr{M}})\|\alpha\|_2^2$$

for every $\alpha \in \mathbb{R}^m$ with $\operatorname{supp}\alpha \in \mathscr{T}$.
Immediately one can see the following (recall Definition 4).

Proposition 2. *The globalized RIP constant is upper bounded by the standard RIP constant:*

$$\delta_{k,\mathscr{M}} \leqslant \delta_k.$$

Definition 15. The generalized RIP constant of order k associated to signals of length N is the smallest number $\delta_k^{(N)}$ such that

$$\left(1-\delta_k^{(N)}\right)\|\Gamma\|_2^2 \leqslant \|D_G\Gamma\|_2^2 \leqslant \left(1+\delta_k^{(N)}\right)\|\Gamma\|_2^2$$

for every $\Gamma \in \mathbb{R}^{mN}$ satisfying $M\Gamma = 0$, $\|\Gamma\|_{0,\infty} \leqslant k$.

Proposition 3. *We have*

$$\delta_k^{(N)} \leqslant \frac{\delta_{k,\mathscr{M}} + (n-1)}{n} \leqslant \frac{\delta_k + (n-1)}{n}.$$

Proof. Obviously it is enough to show only the leftmost inequality. If $\Gamma = (\alpha_i)_{i=1}^N$ and $\|\Gamma\|_{0,\infty} \leqslant k$, this gives $\|\alpha_i\|_0 \leqslant k$ for all $i = 1,\ldots,P$. Further, setting $x := D_G\Gamma$ we clearly have $\Gamma \in \rho(x)$ and so $\operatorname{supp}\Gamma \in \Sigma_{\mathscr{M}}$. Thus $\operatorname{supp}\alpha_i \in \mathscr{T}$, and therefore

$$(1-\delta_{k,\mathscr{M}})\|\alpha_i\|_2^2 \leqslant \|D\alpha_i\|_2^2 \leqslant (1+\delta_{k,\mathscr{M}})\|\alpha_i\|_2^2.$$

By Corollary 3 we know that for every Γ satisfying $M\Gamma = 0$, we have

$$\|D_G\Gamma\|_2^2 = \frac{1}{n}\sum_{i=1}^N \|D\alpha_i\|_2^2.$$

Now for the lower bound,

$$\|D_G\Gamma\|_2^2 \geqslant \frac{1-\delta_{k,\mathscr{M}}}{n}\sum_{i=1}^{N}\|\alpha_i\|_2^2 = \left(1-1+\frac{1-\delta_{k,\mathscr{M}}}{n}\right)\|\Gamma\|_2^2$$
$$= \left(1-\frac{\delta_{k,\mathscr{M}}+(n-1)}{n}\right)\|\Gamma\|_2^2.$$

For the upper bound,

$$\|D_G\Gamma\|_2^2 \leqslant \frac{1+\delta_{k,\mathscr{M}}}{n}\sum_{i=1}^{N}\|\alpha_i\|_2^2 < \left(1+\frac{\delta_{k,\mathscr{M}}+1}{n}\right)\|\Gamma\|_2^2$$
$$\leqslant \left(1+\frac{\delta_{k,\mathscr{M}}+(n-1)}{n}\right)\|\Gamma\|_2^2.$$

□

Theorem 5 (Uniqueness and Stability of $P_{0,\infty}$ via RIP). *Suppose that $\delta_{2s}^{(N)} < 1$, and suppose further that $x = D_G\Gamma_0$ with $\|\Gamma_0\|_{0,\infty} = s$ and $\|D_G\Gamma_0 - x\|_2 \leqslant \varepsilon$. Then every solution $\hat{\Gamma}$ of the noise-constrained $P_{0,\infty}^{\varepsilon}$ problem*

$$\hat{\Gamma} \leftarrow \arg\min_{\Gamma}\|\Gamma\|_{0,\infty} \text{ s.t. } \|D_G\Gamma - x\| \leqslant \varepsilon,\ M\Gamma = 0$$

satisfies

$$\|\hat{\Gamma} - \Gamma_0\|_2^2 \leqslant \frac{4\varepsilon^2}{1-\delta_{2s}^{(N)}}.$$

In particular, Γ_0 is the unique solution of the noiseless $P_{0,\infty}$ problem.

Proof. Immediate using the definition of the globalized RIP:

$$\|\hat{\Gamma} - \Gamma_0\|_2^2 < \frac{1}{1-\delta_{2s}^{(N)}}\|D_G\left(\hat{\Gamma}-\Gamma_0\right)\|_2^2 \leqslant \frac{1}{1-\delta_{2s}^{(N)}}\left(\|D_G\hat{\Gamma} - x\|_2 + \|D_G\Gamma_0 - x\|_2\right)^2$$
$$\leqslant \frac{4\varepsilon^2}{1-\delta_{2s}^{(N)}}.$$

□

3 Pursuit Algorithms

In this section we consider the problem of efficient projection onto the model $\mathscr{M}$. First we treat the "oracle" setting, i.e., when the supports of the local patches (and therefore of the global vector Γ) are known. We show that the local patch averaging

(LPA) method is not a good projector; however, repeated application of it does achieve the desired result.

For the non-oracle setting, we consider "local" and "globalized" pursuits. The former type does not use any dependencies between the patches, and tries to reconstruct the supports α_i completely locally, using standard methods such as OMP—and as we demonstrate, it can be guaranteed to succeed in more cases than the standard analysis would imply. However a possibly better alternative exists, namely, a "globalized" approach with the patch disagreements as a major driving force.

3.1 Global (Oracle) Projection, Local Patch Averaging (LPA) and the Local-Global Gap

Here we briefly consider the question of efficient projection onto the subspace $\ker A_{\mathscr{S}}$, given $\mathscr{S}$.

As customary in the literature [12], the projector onto $\ker A_{\mathscr{S}}$ can be called *an oracle.* In effect, we would like to compute

$$x_G(y, \mathscr{S}) := \arg\min_x \|y - x\|_2^2 \qquad s.t.\ A_{\mathscr{S}} x = 0, \tag{11}$$

given $y \in \mathbb{R}^N$.

To make things concrete, let us assume the standard Gaussian noise model:

$$y = x + \mathscr{N}\left(0, \sigma^2 I\right), \tag{12}$$

and let the mean squared error (MSE) of an estimator $f(y)$ of x be defined as usual, i.e., $MSE(f) := \mathbb{E}\|f(y) - x\|_2^2$. The following is well-known.

Proposition 4. *In the Gaussian noise model* (12)*, the performance of the oracle estimator* (11) *is*

$$MSE(x_G) = (\dim \ker A_{\mathscr{S}})\, \sigma^2.$$

Let us now turn to the local patch averaging (LPA) method. This approach suggests denoising an input signal by (i) breaking it into overlapping patches, (ii) denoising each patch independently, followed by (iii) averaging the local reconstructions to form the global signal estimate. The local denoising step is done by solving pursuit problems, estimating the local supports S_i, while the averaging step is the solution to the minimization problem:

$$\hat{x} = \arg\min_x \sum_{i=1}^{P} \|R_i x - P_{S_i} R_i y\|_2^2,$$

where y is the noisy signal. This has a closed-form solution:

$$\hat{x}_{LPA} = \left(\sum_i R_i^T R_i\right)^{-1} \left(\sum_i R_i^T P_{S_i} R_i\right) y = \underbrace{\left(\frac{1}{n}\sum_i R_i^T P_{S_i} R_i\right)}_{:=M_A} y. \tag{13}$$

Again, the following fact is well-established.

Proposition 5. *In the Gaussian noise model* (12)*, the performance of the averaging estimator* (13) *is*

$$MSE\left(\hat{x}_{LPA}\right) = \sigma^2 \sum_{i=1}^{N} \lambda_i,$$

where $\{\lambda_1, \ldots, \lambda_N\}$ *are the eigenvalues of* $M_A M_A^T$.

Thus, there exists a *local-global gap* in the oracle setting, illustrated in Figure 2 on page 14. In Subsection 4.1 we estimate this gap for a specific case of piecewise constant signals.

The following result is proved in Appendix C: Proof of Theorem 6.

Theorem 6. *For any* $\mathscr{S}$*, we have*

$$\lim_{k\to\infty} M_A^k = P_{\ker A_{\mathscr{S}}},$$

where $P_{\ker A_{\mathscr{S}}}$ *is the orthogonal projector onto* $\ker A_{\mathscr{S}}$*. Therefore for any* y*, iterations of* (13) *starting at* y *converge to* $x_G(y)$ *with a linear rate.*

From the proof it is evident that the rate of convergence depends on the eigenvalues of M_A (which turn out to be related to the singular values of $A_{\mathscr{S}}$). Analyzing these

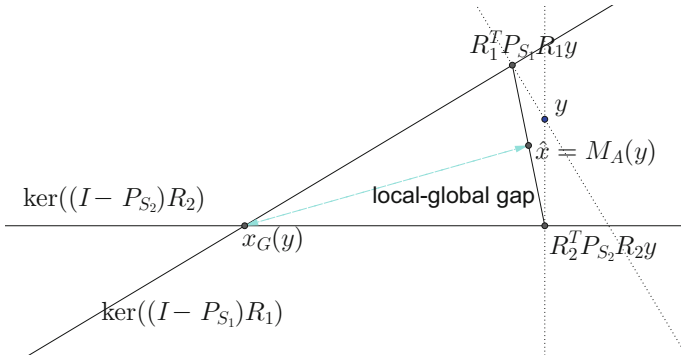


Fig. 2 The local-global gap, oracle setting. Illustration for the case $P = 2$. In details, the noisy signal y can be either projected onto $\ker A_{\mathscr{S}}$ (the point $x_G(y)$) or by applying the LPA (the point $\hat{x} = M_A(y)$). The difference between those two is the local-global gap, which can be significant.

eigenvalues (and therefore the convergence rate) appears to be a difficult problem for general $\mathcal{M}$ and $\mathcal{S}$. In Theorem 8 we show one example where we consider the related problem of estimating the sum $\sum_{i=1}^{N} \lambda_i$ appearing in Proposition 5, in the case of the piecewise constant model (providing estimates for the local-global gap as well).

To conclude, we have shown that *the iterated LPA algorithm provides an efficient method for computing the global oracle projection* x_G.

3.2 Local Pursuit Guarantees

Now we turn to the question of projection onto the model $\mathcal{M}$ when the support of Γ is not known.

Here we show that running OMP [13, 43] on each patch extracted from the signal in fact succeeds in more cases than can be predicted by the classical unconstrained sparse model for each patch. We use the modified coherence function (which is unfortunately intractable to compute):

$$\eta_1^*(s) := \max_{S \in \mathcal{T}} \left(\max_{j \in S} \sum_{k \in S \setminus \{j\}} \left| \langle d_k, d_j \rangle \right| + \max_{j \notin S} \sum_{k \in S} \left| \langle d_k, d_j \rangle \right| \right).$$

The proof of the following theorem is very similar to proving the guarantee for the standard OMP via the Babel function (Definition 3); see e.g., [22, Theorem 5.14]—and therefore we do not reproduce it here.

Theorem 7. *If* $\eta_1^*(s) < 1$*, then running OMP on each patch extracted from any* $x \in \mathcal{M}$ *will recover its true support.*

Since the modified coherence function takes the allowed local supports into consideration, one can readily conclude that

$$\eta_1^*(s) \leqslant \mu_1(s) + \mu_1(s-1),$$

and therefore Theorem 7 gives in general a possibly better guarantee than the one based on μ_1.

3.3 Globalized Pursuits

We now turn to consider several pursuit algorithms, aiming at solving the $P_{0,\infty}/P_{0,\infty}^{\varepsilon}$ problems, in the globalized model. The main question is how to project the patch supports onto the nonconvex set $\Sigma_{\mathcal{M}}$.

The core idea is to relax the constraint $M_*\Gamma = 0,\ \|\Gamma\|_{0,\infty} \leqslant s$ and allow for some patch disagreements, so that the term $\|M_*\Gamma_k\|$ is not exactly zero. Intuitive explanation is as follows: the disagreement term "drives" the pursuit, and the probability of success is higher because we only need to "jump-start" it with the first patch, and then by strengthening the weight of the penalty related to this constraint, the supports will "align" themselves correctly. Justifying this intuition, at least in some cases, is a future research goal.

3.3.1 Q-OMP

Given $\beta > 0$, we define

$$Q_\beta := \begin{bmatrix} D_G \\ \beta M_* \end{bmatrix}.$$

The main idea of the Q-OMP algorithm is to substitute the matrix Q_β as a proxy for the constraint $M_*\Gamma = 0$, by plugging it as a dictionary to the OMP algorithm. Then, given the obtained support $\mathscr{S}$, as a way to ensure that this constraint is met, one can construct the matrix $A_{\mathscr{S}}$ and project the signal onto the subspace $\ker A_{\mathscr{S}}$ (in Subsection 3.1 we show how such a projection can be done efficiently). The Q-OMP algorithm is detailed in Algorithm 1. Let us reemphasize the point that various values of β correspond to different weightings of the model constraint $M_*\Gamma = 0$ and this might possibly become useful when considering relaxed models (see Section 6).

Algorithm 1 The Q-OMP algorithm—a globalized pursuit

Given: noisy signal y, dictionary D, local sparsity s, parameter $\beta > 0$

1. Construct the matrix Q_β.
2. Run the OMP algorithm on the vector $Y := \begin{bmatrix} y \\ \mathbf{0} \end{bmatrix}$, with the dictionary Q_β and sparsity sN. Obtain the global support vector $\hat{\Gamma}$ with $\operatorname{supp}\hat{\Gamma} = \hat{\mathscr{S}}$.
3. Construct the matrix $A_{\hat{\mathscr{S}}}$ and project y onto $\ker A_{\hat{\mathscr{S}}}$.

3.3.2 ADMM-Inspired Approach

In what follows we extend the above idea and develop an ADMM-inspired pursuit [9, 23, 24, 33].

We start with the following global objective:

$$\hat{x} \leftarrow \arg\min_x \|y - x\|_2^2 \quad \text{s.t. } x = D_G\Gamma, M_*\Gamma = 0,\ \|\Gamma\|_{0,\infty} \leqslant K.$$

Clearly, it is equivalent to $\hat{x} = D_G \hat{\Gamma}$, where

$$\hat{\Gamma} \leftarrow \arg\min_{\Gamma} \|y - D_G \Gamma\|_2^2 \quad \text{s.t. } M_* \Gamma = 0,\ \|\Gamma\|_{0,\infty} \leqslant K. \tag{14}$$

Applying Corollary 3, we have the following result.

Proposition 6. *The following problem is equivalent to* (14)*:*

$$\begin{aligned} \hat{\Gamma} &\leftarrow \arg\min_{\{\alpha_i\}} \sum_{i=1}^{P} \|R_i y - D\alpha_i\|_2^2 \\ \text{s.t. } & S_B D\alpha_i = S_T D\alpha_{i+1} \text{ and } \|\alpha_i\|_0 < K \text{ for } i = 1, \ldots, P. \end{aligned} \tag{15}$$

We propose to approximate solution of the nonconvex problem (15) as follows. Define new variables z_i (which we would like to be equal to α_i eventually), and rewrite the problem by introducing the following variable splitting (here Z is the concatenation of all the z_i's):

$$\left\{\hat{\Gamma}, \hat{Z}\right\} \leftarrow \arg\min_{\Gamma, Z} \sum_{i=1}^{P} \|R_i y - D\alpha_i\|_2^2 \quad \text{s.t. } S_B D\alpha_i = S_T D z_{i+1},\ \alpha_i = z_i,\ \|\alpha_i\|_0 \leqslant K.$$

The constraints can be written in concise form

$$\underbrace{\begin{bmatrix} I \\ S_B D \end{bmatrix}}_{:=A} \alpha_i = \underbrace{\begin{bmatrix} I & 0 \\ 0 & S_T D \end{bmatrix}}_{:=B} \begin{pmatrix} z_i \\ z_{i+1} \end{pmatrix},$$

and so globally we would have the following structure (for $N = 3$)

$$\underbrace{\begin{bmatrix} A & & \\ & A & \\ & & A \end{bmatrix}}_{:=\tilde{A}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \underbrace{\begin{bmatrix} I & & & \\ & S_T D & & \\ & & I & \\ & & & S_T D \\ & & & I \\ S_T D & & & \end{bmatrix}}_{:=\tilde{B}} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$$

Our ADMM-inspired method is defined in Algorithm 2.

Algorithm 2 The ADMM-inspired pursuit for $P^{\varepsilon}_{0,\infty}$.

Given: noisy signal y, dictionary D, local sparsity s, parameter $\rho > 0$. The augmented Lagrangian is

$$L_\rho\left(\{\alpha_i\},\{z_i\},\{u_i\}\right) = \sum_{i=1}^{P} \|R_i y - D\alpha_i\|_2^2 + \frac{\rho}{2}\sum_{i=1}^{P} \|A\alpha_i - B\begin{pmatrix} z_i \\ z_{i+1}\end{pmatrix} + u_i\|_2^2.$$

1. Repeat until convergence:

   a. Minimization wrt $\{\alpha_i\}$ is a batch-OMP:

$$\alpha_i^{k+1} \leftarrow \arg\min_{\alpha_i} \|R_i y - D\alpha_i\|_2^2 + \frac{\rho}{2}\|A\alpha_i - B\begin{pmatrix} z_i^k \\ z_{i+1}^k\end{pmatrix} + u_i^k\|_2^2, \quad s.t. \|\alpha_i\|_0 \leqslant K$$

$$\alpha_i^{k+1} \leftarrow OMP\left(\tilde{D} = \begin{bmatrix} D \\ \sqrt{\frac{\rho}{2}}A\end{bmatrix},\ \tilde{y}_i^k = \begin{pmatrix} R_i y \\ \sqrt{\frac{\rho}{2}}\left(B\begin{pmatrix} z_i^k \\ z_{i+1}^k\end{pmatrix} - u_i^k\right)\end{pmatrix}, K\right).$$

   b. Minimization wrt z is a least squares problem with a sparse matrix, which can be implemented efficiently:

$$Z^{k+1} \leftarrow \arg\min_{Z} \|\tilde{A}\Gamma^{k+1} + U^k - \tilde{B}Z\|_2^2$$

   c. Dual update:

$$U^{k+1} \leftarrow \tilde{A}\Gamma^{k+1} - \tilde{B}Z + U^k.$$

2. Compute $\hat{y} := D_G\hat{\Gamma}$.

4 Examples

We now turn to present several classes of signals that belong to the proposed globalized model, where each of these is obtained by imposing a special structure on the local dictionary. Then, we demonstrate how one can sample from $\mathscr{M}$ and generate such signals. Additional examples are given in Appendix E: Generative Models for Patch-Sparse Signals.

4.1 Piecewise Constant (PWC) Signals

The (unnormalized) Heaviside $n \times n$ dictionary H_n is the upper triangular matrix with 1's in the upper part (see Figure 3 on page 19). Formally, each local atom d_i of length n is expressed as a step function, given by $d_i^T = [\mathbf{1_i}\ ,\ \mathbf{0_{n-i}}]^T$, $1 \leq i \leq n$, where $\mathbf{1_i}$ is a vector of ones of length i. Similarly, $\mathbf{0}_{n-i}$ is a zero vector of length $n-i$. The following property is verified by noticing that H_n^{-1} is the discrete difference operator.

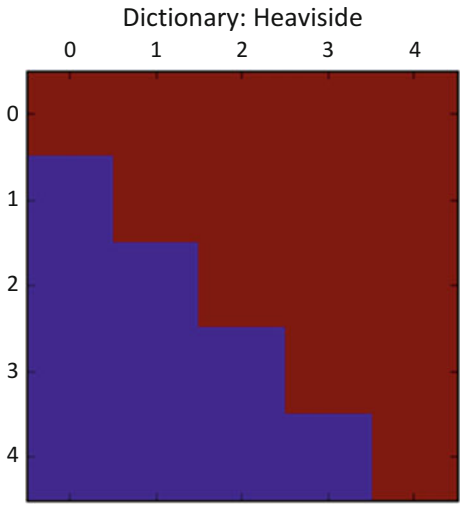


Fig. 3 Heaviside dictionary H_4. Red is 1, blue is 0.

Proposition 7. *If a patch $x_i \in \mathbb{R}^n$ has $L-1$ steps, then its (unique) representation in the Heaviside dictionary H_n has at most L nonzeros.*

Corollary 1. *Let $x \in \mathbb{R}^N$ be a piecewise constant signal with at most $L-1$ steps per each segment of length n (in the periodic sense). Then*

$$x \in \mathscr{M}\left(H_n, L, N, P = N\right).$$

Remark 2. The model $\mathscr{M}\left(H_n, L, N, P = N\right)$ contains also some signals having exactly L steps in a particular patch, but those patches must have their last segment with zero height.
As an example, one might synthesize signals with sparsity $\|\Gamma\|_{0,\infty} \leq 2$ according to the following scheme:

1. Draw at random the support of $\Gamma amma$ with the requirement that the distance between the jumps within the signal will be at least the length of a patch (this allows at most two nonzeros per patch, one for the step and the second for the bias/DC).
2. Multiply each step by a random number.

The global subspace $A_{\mathscr{S}}$ and the corresponding global oracle denoiser x_G (11) in the PWC model can be explicitly described.

Proposition 8. *Let $x \in \mathbb{R}^N$ consist of s constant segments with lengths ℓ_r, $r = 1, \ldots, s$, and let Γ be the (unique) global representation of x in $\mathscr{M}$ (i.e., $\rho(x) = \{\Gamma\}$). Denote $B := \operatorname{diag}\left(B_r\right)_{r=1}^{s}$, where $B_r = \frac{1}{\ell_r}\mathbf{1}_{\ell_r \times \ell_r}$ Then*

1. *We have*

$$\ker A_{\operatorname{supp}\Gamma} = \ker\left(I_N - B\right), \tag{16}$$

and therefore $\dim\ker A_{\operatorname{supp}\Gamma} = s$ *and* $MSE\left(\hat{x}_G\right) = s\sigma^2$ *under the Gaussian noise model* (12).

2. *Furthermore, the global oracle estimator* x_G *is given by*

$$x_G\left(y, \operatorname{supp}\Gamma\right) = By, \tag{17}$$

i.e., the global oracle is the averaging operator within the constant segments of the signal.

Proof. Every signal $y \in \ker A_{\operatorname{supp}\Gamma}$ has the same "local jump pattern" as x, and therefore it also has the same *global* jump pattern. That is, every such y consists of s constant segments with lengths ℓ_r. It is an easy observation that such signals satisfy $y = By$, which proves (16). It is easy to see that $\dim\ker\left(I_{\ell_r} - B_r\right) = 1$, and therefore

$$\dim\ker\left(I_N - \operatorname{diag}\left(B_r\right)_{r=1}^{s}\right) = s.$$

The proof of 1) is finished by invoking Proposition 4.

To prove (17), notice that by the previous discussion the null-space of $A_{\operatorname{supp}\Gamma}$ is spanned by the orthogonal set $e_r = \frac{1}{\sqrt{\ell_r}}\left[0,\ldots,0,\underbrace{1,1,\ldots,1}_{\ell_r},0,\ldots,0\right]^T$, $r = 1,\ldots,s$. Let $K = [e_1,\ldots,e_s]$, then $x_G = KK^{\dagger} = KK^T$. It can be easily verified by direct computation that $KK^T = B$. □

It turns out that the LPA performance (and the local-global gap) can be accurately described by the following result. We provide an outline of proof in Appendix D: Proof of Theorem 8.

Theorem 8. *Let* $x \in \mathbb{R}^N$ *consist of* s *constant segments with lengths* ℓ_r, $r = 1,\ldots,s$, *and assume the Gaussian noise model* (12). *Then*

1. *There exists a function* $R\left(n,\alpha\right) : \mathbb{N}\times\mathbb{N} \to \mathbb{R}^+$, *with* $R\left(n,\alpha\right) > 1$, *such that*

$$MSE\left(\hat{x}_{LPA}\right) = \sigma^2 \sum_{r=1}^{s} R\left(n,\ell_r\right).$$

2. *The function* $R\left(n,\alpha\right)$ *satisfies:*

   a. $R\left(n,\alpha\right) = 1 + \frac{\alpha\left(2\alpha H_\alpha^{(2)} - 3\alpha + 2\right) - 1}{n^2}$ *if* $n \geqslant \alpha$, *where* $H_\alpha^{(2)} = \sum_{k=1}^{\alpha}\frac{1}{k^2}$;
   b. $R\left(n,\alpha\right) = \frac{11}{18} + \frac{2\alpha}{3n} + \frac{6\alpha-11}{18n^2}$ *if* $n \leqslant \frac{\alpha}{2}$.

Corollary 2. *The function $R(n,\alpha)$ is monotonically increasing in α (with n fixed) and monotonically decreasing in n (with α fixed). Furthermore,*

1. $\lim_{n\to\infty} R(n,n) = \frac{\pi^2}{3} - 2 \approx 1.29$;
2. $\lim_{n\to\infty} R(n,2n) = \frac{35}{18} \approx 1.94$.

Thus, for reasonable choices of the patch size, the local-global gap is roughly a constant multiple of the number of segments, reflecting the global complexity of the signal.

For numerical examples of reconstructing the PWC signals using our local-global framework, see Subsection 5.2.

4.2 Signature-Type Dictionaries

Another type of signals that comply with our model are those represented via a signature dictionary, which has been shown to be effective for image restoration [3]. This dictionary is constructed from a small signal, $x \in \mathbb{R}^m$, such that its every patch (in varying location, extracted in a cyclic fashion), $R_i x \in \mathbb{R}^n$, is a possible atom in the representation, namely, $d_i = R_i x$. As such, every consecutive pair of atoms $(i, i+1)$ is essentially a pair of overlapping patches that satisfy $S_B d_i = S_T d_{i+1}$ (before normalization). The complete algorithm is presented for convenience in Algorithm 3.

Algorithm 3 Constructing the signature dictionary

1. Choose the base signal $x \in \mathbb{R}^m$.
2. Compute $D(x) = [R_1 x, R_2 x, \ldots, R_m x]$, where R_i extracts the i-th patch of size n in a cyclic fashion.
3. Normalization: $\tilde{D}(x) = [d_1, \ldots, d_m]$, where $d_i = \frac{R_i x}{\|R_i x\|_2}$.

Given D as above, one can generate signals $y \in \mathbb{R}^N$, where N is an integer multiple of m, with s nonzeros per patch, by the easy procedure outlined below.

1. Init: Construct a base signal $b \in \mathbb{R}^N$ by replicating $x \in \mathbb{R}^m$ N/m times (note that b is therefore periodic). Set $y = 0$.
2. Repeat for $j = 1, \ldots, s$:
   a. Shift: Circularly shift the base signal by t_j positions, denoted by $\text{shift}(b, t_j)$, for some $t_j = 0, 1, \ldots, m-1$ (drawn at random).
   b. Aggregate: $y = y + \omega_j \cdot \text{shift}(b, t_j)$, where ω is an arbitrary random scalar.

Notice that a signal constructed in this way must be periodic, as it is easily seen that

$$\ker A_{\mathscr{S}} = \text{span}\left\{\text{shift}\,(b, t_i)\right\}_{i=1}^{s},$$

while the support sequence $\mathscr{S}$ is

$$\mathscr{S} = ([t_1, t_2, \ldots, t_s], [t_1, t_2, \ldots, t_s] + 1, \ldots, [t_1, t_2, \ldots, t_s] + N) \pmod{m}.$$

Assuming that there are no additional relations between the single atoms of D except those from the above construction, all $\mathscr{S} \in \Sigma_{\mathscr{M}}$ are easily seen to be of the above form.

In Figure 4 on page 23, we give an example of a signature-type dictionary D for $(n, m) = (6, 10)$ and a signal x with $N = P = 30$ together with its corresponding sparse representation Γ.

Remark 3. It might seem that every $n \times m$ Hankel matrix such as the one shown in Figure 4 on page 23 produces a signature-type dictionary with a nonempty signal space $\mathscr{M}$. However this is not the case, because such a dictionary will usually fail to generate signals of length larger than $n + m - 1$.

4.2.1 Multi-Signature Dictionaries

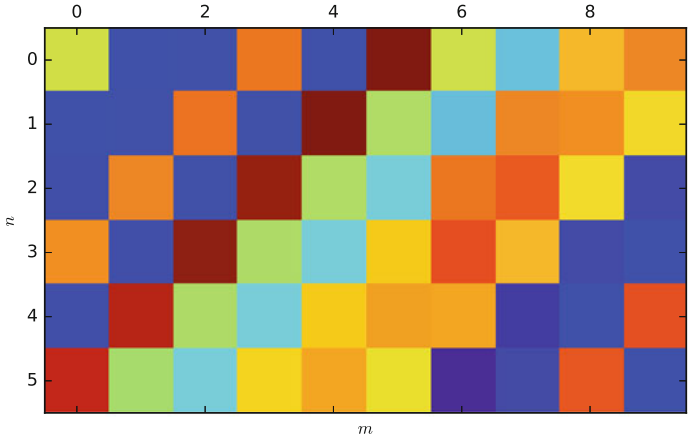
One can generalize the construction of Subsection 4.2 and consider k-tuples of initial base signals $x_i, \ldots, x_k$, instead of a single x. The desired dictionary D will consist of corresponding k-tuples of atoms, which are constructed from those base signals. In order to avoid ending up with the same structure as the case $k = 1$, we also require a "mixing" of the atoms. The complete procedure is outlined in Algorithm 4.

Algorithm 4 Constructing the multi-signature dictionary

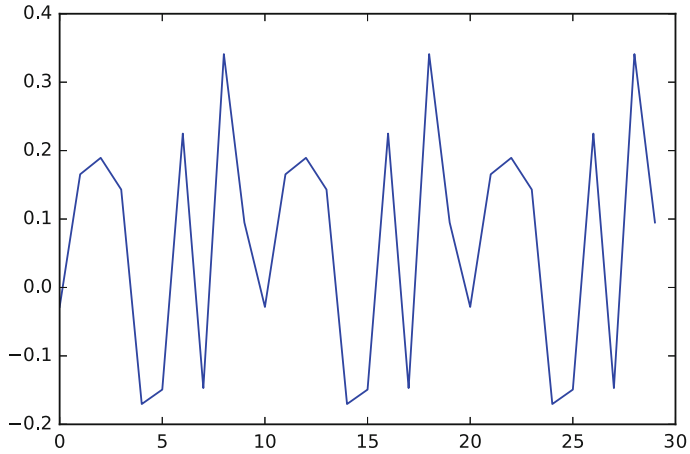
1. Input: n, m, k such that k divides m. Put $r := \frac{m}{k}$.
2. Select a signal basis matrix $X \in \mathbb{R}^{r \times k}$ and r nonsingular transfer matrices $M_i \in \mathbb{R}^{k \times k}$, $i = 1, \ldots, r$.
3. Repeat for $i = 1, \ldots, r$:
   a. Let $Y_i = \left[y_{i,1}, \ldots, y_{i,k}\right] \in \mathbb{R}^{n \times k}$, where each $y_{i,j}$ is the i-th patch (of length n) of the signal x_j.
   b. Put the k-tuple $\left[d_{i,1}, \ldots, d_{i,k}\right] = Y_i \times M_i$ as the next k atoms in D.

In order to generate a signal of length N from $\mathscr{M}$, one can follow these steps (again we assume that m divides N):

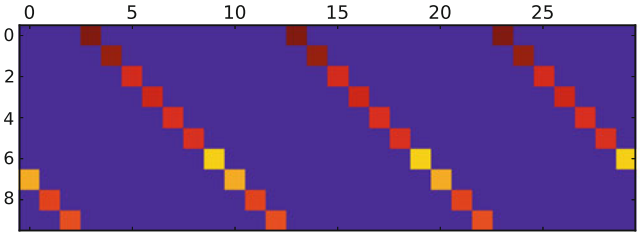
1. Create a base signal matrix $X^G \in \mathbb{R}^{N \times k}$ by stacking $k\frac{N}{m}$ copies of the original basis matrix X. Set $y = 0$.
2. Repeat for $j = 1, \ldots, s$:
   a. Select a base signal $b_j \in \mathscr{R}\left(X^G\right)$ and shift it (in a circular fashion) by some $t_j = 0, 1, \ldots, R - 1$.
   b. Aggregate: $y = y + \text{shift}(b_j, t_j)$ (note that here we do not need to multiply by a random scalar).



(a) The dictionary matrix D



(b) The signal $x \in \ker A_{\mathscr{S}}$ for $\mathscr{S}$ generated by $t_1 = 6$ and $s = 1$, with $P = N = 30$.



(c) The coefficient matrix Γ corresponding to the signal x in (c)

Fig. 4 An example of the signature dictionary with $n = 6$, $m = 10$. See Remark 3.

This procedure will produce a signal y of local sparsity $k \cdot s$. The corresponding support sequence can be written as

$$\mathscr{S} = (s_1, s_2, \ldots, s_N),$$

where $s_i = s_1 + i \pmod m$ and

$$s_1 = [(t_1, 1), (t_1, 2), \ldots, (t_1, k), \ldots, (t_s, 1), (t_s, 2), \ldots, (t_s, k)].$$

Here (t_j, i) denotes the atom $d_{t_j,i}$ in the notation of Algorithm 4. The corresponding signal space is

$$\ker A_{\mathscr{S}} = \operatorname{span}\left\{\operatorname{shift}\left(X^G, t_j\right)\right\}_{j=1}^{s},$$

and it is of dimension $k \cdot s$.

An example of a multi-signature dictionary and corresponding signals may be seen in Figure 5 on page 25.

4.3 Convolutional Dictionaries

An important class of signals is the *sparse convolution model,* where each signal $x \in \mathbb{R}^N$ can be written as a linear combination of shifted "waveforms" $\boldsymbol{d_i} \in \mathbb{R}^n$, each $\boldsymbol{d_i}$ being a column in the local dictionary $D' \in \mathbb{R}^{n \times m}$. More conveniently, any such x can be represented as a circular convolution of $\boldsymbol{d_i}$ with a (sparse) "feature map" $\boldsymbol{\psi_i} \in \mathbb{R}^N$:

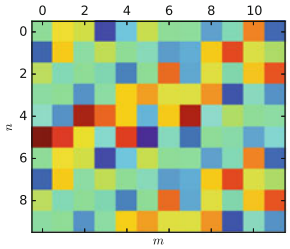
$$x = \sum_{i=1}^{m} \boldsymbol{d_i} *_N \boldsymbol{\psi_i}. \tag{18}$$

Such signals arise in various applications, such as audio classification [6, 26, 50], neural coding [16, 44], and mid-level image representation and denoising [31, 58, 59].

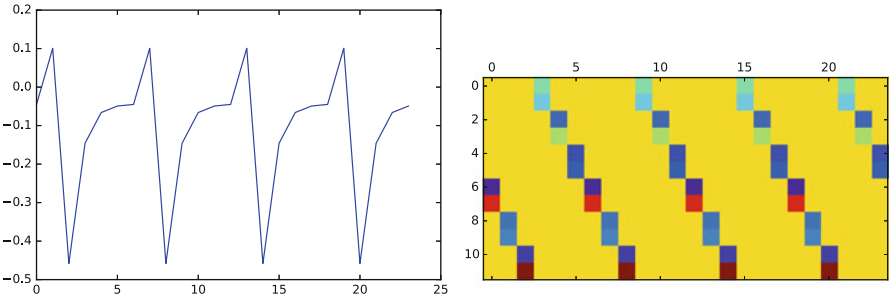
Formally, the convolutional class can be recast into the patch-sparse model of this paper as follows. First, we can rewrite (18) as

$$x = \underbrace{\left[\boldsymbol{C}_1 \; \boldsymbol{C}_2 \; \ldots \; \boldsymbol{C}_m\right]}_{:=\boldsymbol{E}} \boldsymbol{\Psi},$$

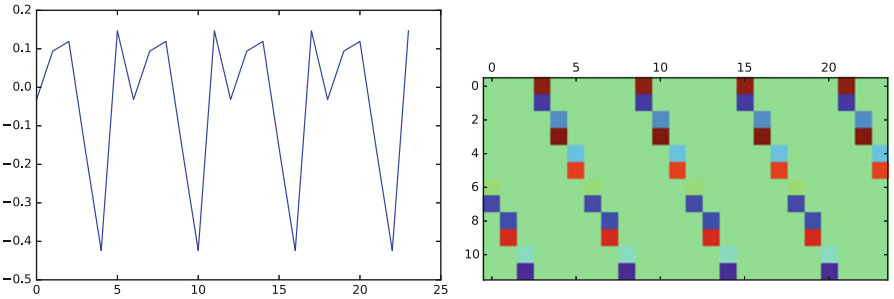
where each $\boldsymbol{C}_i \in \mathbb{R}^{N \times N}$ is a banded circulant matrix with its first column being equal to $\boldsymbol{d_i}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{Nm}$ is the concatenation of the $\boldsymbol{\psi_i}$'s. It is easy to see that by permuting the columns of $\boldsymbol{E}$, one obtains precisely the global convolutional dictionary nD_G based on the local dictionary D' (recall (3)). Therefore we obtain



(a) The dictionary D



(b) The first signal and its sparse representation in $\ker A_{\mathscr{S}}$ with $N = 24$, $s = 1$ and $t_1 = 5$.



(c) The second signal and its sparse representation in $\ker A_{\mathscr{S}}$.

Fig. 5 Example of multi-signature dictionary with $n = 10$, $m = 12$, *and* $k = 2$.

$$x = \underbrace{D_G\left(D'\right)}_{:=D'_G}\boldsymbol{\Gamma'}. \tag{19}$$

While it is tempting to conclude from comparing (19) and (4) that the convolutional model is equivalent to the patch-sparse model, an essential ingredient is missing, namely, the requirement of equality on overlaps, $M\boldsymbol{\Gamma'} = 0$. Indeed, nothing in the definition of the convolutional model restricts the representation $\boldsymbol{\Psi}$ (and therefore $\boldsymbol{\Gamma'}$); therefore, in principle the number of degrees of freedom remains Nm, as compared to $N(m-n+1)$ from Proposition 15.

To fix this, following [42], we apply R_i to (19) and obtain $R_i x = R_i D'_G \boldsymbol{\Gamma'}$. The "stripe" $\Omega'_i = R_i D'_G$ has only $(2n-1)\,m$ nonzero consecutive columns, and in fact the nonzero portion of Ω'_i is equal for all i. This implies that every x_i has a representation $x_i = \Theta \boldsymbol{\gamma_i}$ in the "pseudo-local" dictionary

$$\Theta\left(D'\right) := \left[Z_B^{(n-1)} D' \ \ldots \ D' \ \ldots \ Z_T^{(n-1)} D'\right] \in \mathbb{R}^{n \times (2n-1)m},$$

where the operators $Z_B^{(k)}$ and $Z_T^{(k)}$ are given by Definition 6 in Appendix B: Proof of Lemma 2. If we now assume that our convolutional signals satisfy

$$\|\boldsymbol{\gamma_i}\|_0 \leqslant s \qquad \forall i,$$

then we have shown that they belong to $\mathscr{M}\left(\Theta\left(D'\right), s, P, N\right)$ and thus can be formally treated by the framework we have developed.

It turns out that this direct approach is quite naive, as the dictionary $\Theta\left(D'\right)$ is extremely ill-equipped for sparse reconstruction (e.g., it has repeated atoms, and therefore $\mu\left(\Theta\left(D'\right)\right) = 1$). To tackle this problem, a convolutional sparse coding framework was recently developed in [42], where the explicit dependencies between the sparse representation vectors $\boldsymbol{\gamma_i}$ (and therefore the special structure of the corresponding constraint $M\left(D'\right)\boldsymbol{\Gamma'} = 0$) were exploited quite extensively, resulting in efficient recovery algorithms and nontrivial theoretical guarantees. We refer the reader to [42] for further details and examples.

5 Numerical Experiments

In this section, we test the effectiveness of the globalized model for recovering the signals from Section 4, both in the noiseless and noisy cases. For the PWC, we show a real-world example. These results are also compared to several other approaches such as the LPA, total variation denoising (for the PWC), and a global pursuit based on OMP.

5.1 Signature-Type Signals

In this section we investigate the performance of the pursuit algorithms on signals complying with the signature dictionary model elaborated in Subsection 4.2, constructed from one or two base signals ($k = 1, 2$), and allowing for varying values of s. We compare the results to both LPA and a global pursuit, which uses the dictionary explicitly constructed from the signature model. In detail, the global dictionary D^* is an $N \times (km)$ matrix consisting of the base signal matrix X^G and all its shifts, i.e. (recall the definitions in Subsection 4.2.1)

$$D^* = \left[\text{shift}\left(X^G, i\right)_{i=0}^{m-1}\right].$$

Given that, the global OMP algorithm is defined to run for $k \cdot s$ steps on D^*.

5.1.1 Constructing the Dictionary

In the context of the LPA algorithm, the condition for its success in recovering the representation is a function of the mutual coherence of the local dictionary—the smaller this measure, the larger the number of nonzeros that are guaranteed to be recovered. Leveraging this, we aim at constructing $D \in \mathbb{R}^{n\times m}$ of a signature type that has a small coherence. This can be cast as an optimization problem

$$D = \widetilde{D}\left(x_0\right), \; x_0 = \arg\min_{x\in\mathbb{R}^m} \mu\left(\tilde{D}\left(x\right)\right),$$

where $\tilde{D}\left(x\right)$ is computed by Algorithm 3 (or Algorithm 4) and μ is the (normalized) coherence function.

In our experiments, we choose $(n, m) = (15, 20)$ for $k = 1$ and $(n, m) = (10, 20)$ for $k = 2$. We minimize the above loss function via gradient descent, resulting in $\mu(\tilde{D}(x)) = 0.20$ for $k = 1$ and $\mu = 0.26$ for $k = 2$. We used the TensorFlow open source package [1]. As a comparison, the coherence of a random signature dictionary is about 0.5.

5.1.2 Noiseless Case

In this setting, we test the ability of the globalized OMP (Subsection 3.3.1) to perfectly recover the sparse representation of clean signature-type signals. Figure 6 compares the proposed algorithm (for different choices of $\beta \in \{0.25, 0.5, 1, 2, 5\}$) with the LPA by providing their probability of success in recovering the true sparse vectors, averaged over 10^3 randomly generated signals of length $N = 100$. For brevity we show only the results for $k = 1$ here.

From a theoretical perspective, since $\mu(D) = 0.20$, the LPA algorithm is guaranteed to recover the representation when $\|\Gamma\|_{0,\infty} \leq 3$, as indeed it does. Comparing the LPA approach to the globalized OMP, one can observe that for $\beta \geq 1$ the latter consistently outperforms the former, having a perfect recovery for $\|\Gamma\|_{0,\infty} \leq 4$. Another interesting insight of this experiment is the effect of β on the performance; roughly speaking, a relatively large value of this parameter results in a better success rate than the very small ones, thereby emphasizing importance of the constraint $M_*\Gamma = 0$. On the other hand, β should not be too large since the importance of the signal is reduced compared to the constraint, which might lead to deterioration in the success rate (see the curve that corresponds to $\beta = 5$ in Figure 6).

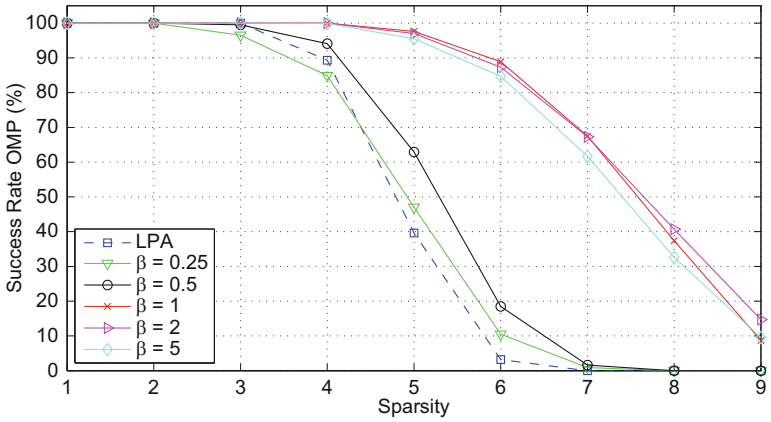
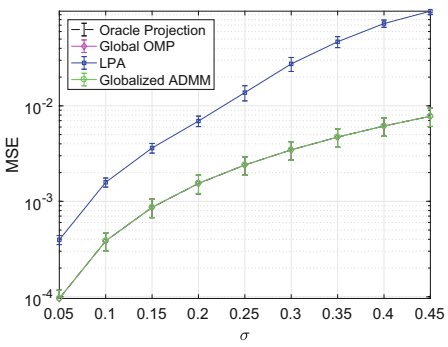


Fig. 6 Probability of the success (%) of the globalized OMP (for various values of β) and the LPA algorithms to perfectly recover the sparse representations of test signals from the signature dictionary model, averaged over 10^3 realizations, as a function of sparsity per patch.

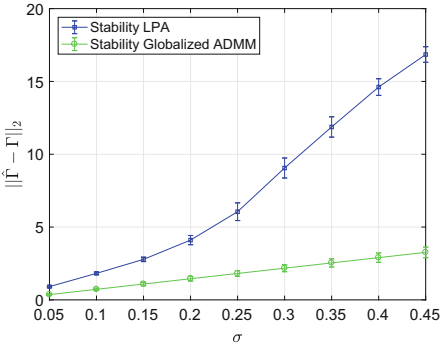
5.1.3 Noisy Case

In what follows, the stability of the proposed globalized ADMM-inspired pursuit is tested and compared to the traditional LPA algorithm, as well as to the global OMP. In addition to the above, we provide the restoration performance of the oracle estimator, serving as an indication for the best possible denoising that can be achieved. In this case, the oracle projection matrix A_S is constructed according to the ground-truth support S.

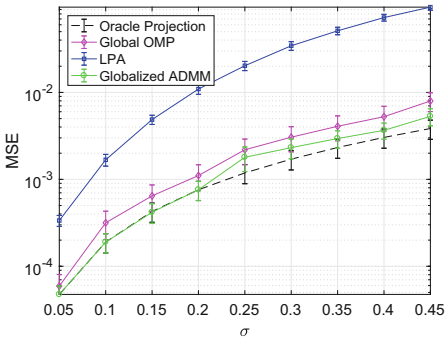
We generate ten random signature-type signals, where each of these is corrupted by white additive Gaussian noise with standard deviation σ, ranging from 0.05 up to 0.5. The global number of nonzeros is injected to the global OMP, and the information regarding the local sparsity is utilized both by the LPA algorithm as well as by our ADMM-inspired pursuit (which is based on local sparse recovery operations). Following Figure 7 parts (a, c), which plot the mean squared error (MSE) of the estimation as a function of the noise level, the ADMM-inspired pursuit achieves the best denoising performance, having similar results to the oracle estimator for all noise levels and sparsity factors. The source of superiority of the ADMM-inspired pursuit might be its inherent ability to obtain an estimation that perfectly fits to the globalized model. The second best algorithm is the global OMP; using complete global information about the signal space, this fact is to be expected. The LPA algorithm is the least accurate; it shows that for our signals the assumption of patch independence severely degrades performance. This sheds light on the difficulty of finding the true supports, the nontrivial solution of this problem, and the great advantage of the proposed globalized model.



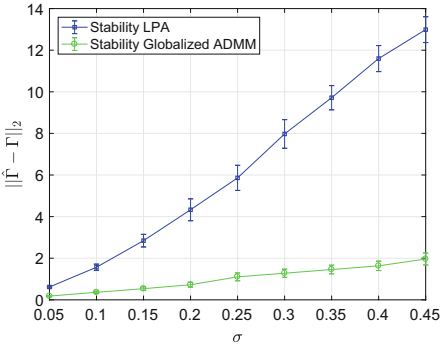
(a) $k = 1, s = 3$, denoising. The globalized ADMM-inspired (green curve) and oracle projection (black curve) coincide with Global OMP (magenta curve).



(b) $k = 1, s = 3$, stability



(c) $k = 2, s = 1$, denoising



(d) $k = 2, s = 1$, stability

Fig. 7 (a, c) Denoising performance of the global OMP, ADMM-inspired pursuit, and LPA algorithm for signals from the signature model with $(a)\, k = 1, s = 5$ and $(c)\, k = 2, s = 1$. The performance of the oracle estimator is provided as well, demonstrating the best possible restoration that can be achieved. (b, d) Stability of the ADMM-inspired pursuit and LPA algorithm for $(b)\, k = 1, s = 5$ and $(d)\, k = 2, s = 1$. For (a, b) the signal size was $N = 100$, while for (c, d) it was $N = 80$.

Similar conclusion holds for the stable recovery of the sparse representations. Per each pursuit algorithm, Figure 7 parts (b, d) illustrate the ℓ_2 distance between the original sparse vector Γ and its estimation $\hat{\Gamma}$, averaged over the different noise realizations. As can be seen, the ADMM-inspired pursuit achieves the most stable recovery, outperforming the LPA algorithm especially in the challenging cases of high noise levels and/or large sparsity factors.

5.2 *Denoising PWC Signals*

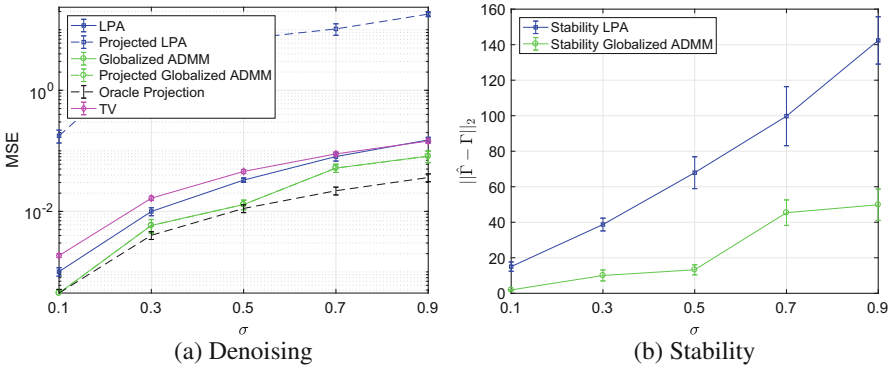
5.2.1 Synthetic Data

In this scenario, we test the ability of the globalized ADMM-inspired pursuit to restore corrupted PWC signals and compare these to the outcome of the LPA algorithm.

In addition, we run the total variation (TV) denoising [48] on the signals, which is known to perform well on PWC. We chose the regularization parameter in the TV by running an exhaustive search over a wide range of values per input signal and picked the one that minimizes the MSE between the estimated and the true signal. Notice that this results in the best possible denoising performance that can be obtained by the TV.

The projected versions of both ADMM-inspired pursuit and LPA are provided along with the one of the oracle estimator. Following the description in Section 4.1, we generate a signal of length $N = 200$, composed of patches of size $n = m = 20$ with a local sparsity of at most 2 nonzeros in the $\ell_{0,\infty}$ sense. These signals are then contaminated by a white additive Gaussian noise with σ in the range of 0.1 to 0.9.

The restoration performance (in terms of MSE) of the abovementioned algorithms and their stability are illustrated in Figure 8, where the results are averaged over 10 noise realizations. As can be seen, the globalized approach significantly outperforms the LPA algorithm for all noise levels. Furthermore, when $\sigma \leq 0.5$, the ADMM-inspired pursuit performs similarly to the oracle estimator. One can also notice that the ADMM-inspired pursuit and its projected version result in the very same estimation, i.e., this algorithm forces the signal to conform with the patch-sparse model globally. On the other hand, following the visual illustration given in Figure 9, the projected version of the LPA algorithm has only two nonzero segments,



(a) Denoising (b) Stability

Fig. 8 (*a*) Denoising performance and (*b*) stability for various noise levels, tested for signals from the piecewise constant model with $\|\Gamma\|_{0,\infty} \leq 2$.

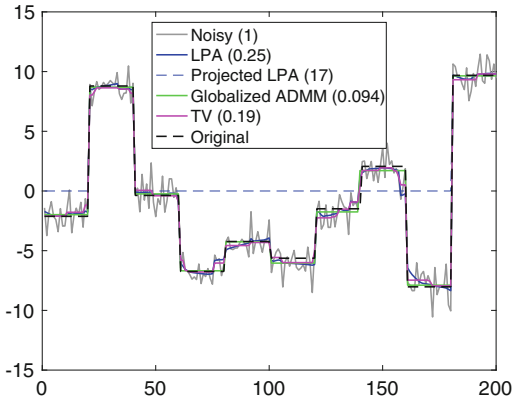


Fig. 9 Denoising of a PWC signal contaminated with additive Gaussian noise ($\sigma = 1.1$) via several pursuit algorithms: input noisy signal (MSE = 1.0), LPA algorithm (MSE = 0.25), projected LPA (MSE = 17), ADMM-inspired pursuit (MSE = 0.094), and TV (MSE=0.19). Projected ADMM is identical to the ADMM-inspired pursuit.

which are the consequence of almost complete disagreement in the support (local inconsistency). This is also reflected in Figure 8a, illustrating that even for a very small noise level ($\sigma = 0.1$), the projected version of the LPA algorithm has a very large estimation error (MSE ≈ 0.18) compared to the one of the ADMM-inspired pursuit (MSE ≈ 0.0004), indicating that the former fails in obtaining a consistent representation of the signal. The TV method is unable to take into account the local information, resulting in reconstruction of lesser quality than both the ADMM-inspired and the LPA.

5.2.2 Real-World Data

Here we apply the globalized ADMM for the PWC model on a real-world DNA copy number data from [51]. The data (see also [34]) come from a single experiment on 15 fibroblast cell lines with each array containing over 2000 (mapped) BACs (bacterial artificial chromosomes) spotted in triplicate. The results (see Figure 10) appear to be reasonably significant.

6 Discussion

In this work we have presented an extension of the classical theory of sparse representations to signals which are locally sparse, together with novel pursuit algorithms. We envision several promising research directions which might emerge from this work.

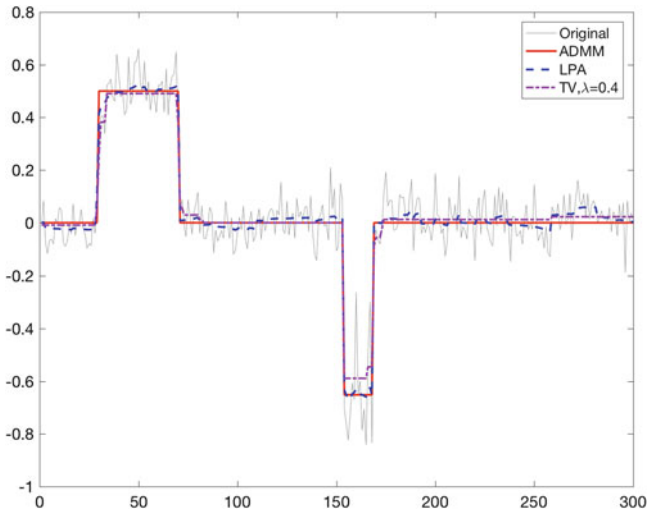


Fig. 10 Applying the PWC reconstruction to a single fibroblast cell line, as described in [51]. The value of λ in TV was chosen empirically based on visual quality. For the ADMM, we chose $n = 40$ and $k = 2$. The ordinate is the normalized average of the log base 2 test over reference ratio of the cell line.

6.1 Relation to Other Models

Viewed globally, the resulting signal model can be considered a sort of "structured sparse" model; however, in contrast to other such constructions ([29, 30, 32, 55] and others), our model incorporates both structure in the representation coefficients and a structured dictionary.

The recently developed framework of convolutional sparse coding (CSC) [41, 42] bears some similarities to our work, in that it, too, has a convolutional representation of the signal via a dictionary identical in structure to D_G. However, the underlying local sparsity assumptions are drastically different in the two models, resulting in very different guarantees and algorithms. That said, we believe that it would be important to provide precise connections between the results, possibly leading to their deeper understanding. First steps in this direction are outlined in Subsection 4.3.

6.2 Further Extensions

The decomposition of the global signal $x \in \mathbb{R}^N$ into its patches,

$$x \mapsto (R_i x)_{i=1}^P , \tag{20}$$

is a special case of a more general decomposition, namely,

$$x \mapsto (w_i \mathscr{P}_i x)_{i=1}^{P}, \tag{21}$$

where $\mathscr{P}_i$ is the (orthogonal) projection onto a subspace W_i of $\mathbb{R}^N$ and w_i are some weights. This observation naturally places our theory, at least partially, into the framework of *fusion frames*, a topic which is generating much interest recently in the applied harmonic analysis community [21, Chapter 13]. In fusion frame theory, which is motivated by applications such as distributed sensor networks, the starting point is precisely the decomposition (21). Instead of the reconstruction formula $x = \sum_i \frac{1}{n} R_i^T R_i x$, in fusion frame theory we have

$$x = \sum_i w_i^2 S_{\mathscr{W}}^{-1} (\mathscr{P}_i x),$$

where $S_{\mathscr{W}}$ is the associated *fusion frame operator*. The natural extension of our work to this setting would seek to enforce some sparsity of the projections $\mathscr{P}_i x$. Perhaps the most immediate variant of (20) in this respect would be to drop the periodicity requirement, resulting in a slightly modified R_i operators near the endpoints of the signal. We would like to mention some recent works which investigate different notions of fusion frame sparsity [2, 4, 8].

Another intriguing possible extension of our work is to relax the complete overlap requirement between patches and consider an "approximate patch sparsity" model, where the patch disagreement vector $M\Gamma$ is not zero but "small." In some sense, one can imagine a full "spectrum" of such models, ranging from a complete agreement (this work) to an arbitrary disagreement (such as in the CSC framework mentioned above).

6.3 Learning Models from Data

The last point above brings us to the question of how to obtain "good" models, reflecting the structure of the signals at hand (such as speech/images, etc.). We hope that one might use the ideas presented here in order to create novel learning algorithms. In this regard, the main difficulty is how to parametrize the space of allowed models in an efficient way. While we presented some initial ideas in Appendix E: Generative Models for Patch-Sparse Signals, in the most general case (incorporating the approximate sparsity direction above), the problem remains widely open.

Acknowledgments The research leading to these results has received funding from the European Research Council under European Union's Seventh Framework Programme, ERC Grant agreement no. 320649. The authors would also like to thank Jeremias Sulam, Vardan Papyan, Raja Giryes, and Gitta Kutinyok for inspiring discussions.

Appendix A: Proof of Lemma 1

Proof. Denote $Z := \ker M$ and consider the linear map $A : Z \to \mathbb{R}^N$ given by the restriction of the "averaging map" $D_G : \mathbb{R}^{mP} \to \mathbb{R}^N$ to Z.

1. Let us see first that $im\,(A) = \mathbb{R}^N$. Indeed, for every $x \in \mathbb{R}^N$, consider its patches $x_i = R_i x$. Since D is full rank, there exist $\{\alpha_i\}$ for which $D\alpha_i = x_i$. Then setting $\Gamma := (\alpha_1, \ldots, \alpha_P)$, we have both $D_G\Gamma = x$ and $M\Gamma = 0$ (by construction, see Section 2), i.e., $\Gamma \in Z$ and the claim follows.
2. Define

$$J := \ker D \times \ker D \times \ldots \ker D \subset \mathbb{R}^{mP}.$$

   We claim that $J = \ker A$.

   a. In one direction, let $\Gamma = (\alpha_1, \ldots, \alpha_P) \in \ker A$, i.e., $M\Gamma = 0$ and $D_G\Gamma = 0$. Immediately we see that $\frac{1}{n} D\alpha_i = 0$ for all i, and therefore $\alpha_i \in \ker D$ for all i, thus $\Gamma \in J$.
   b. In the other direction, let $\Gamma = (\alpha_1, \ldots, \alpha_P) \in J$, i.e., $D\alpha_i = 0$. Then the local representations agree, i.e., $M\Gamma = 0$, thus $\Gamma \in Z$. Furthermore, $D_G\Gamma = 0$ and therefore $\Gamma \in \ker A$.

3. By the fundamental theorem of linear algebra, we conclude

$$\begin{aligned} \dim Z &= \dim im\,(A) + \dim \ker A = N + \dim J \\ &= N + (m-n)\,N = N\,(m-n+1)\,. \end{aligned}$$

□

Appendix B: Proof of Lemma 2

We start with an easy observation.

Proposition 9. *For any vector $\rho \in \mathbb{R}^N$, we have*

$$\|\rho\|_2^2 = \frac{1}{n} \sum_{j=1}^{N} \|R_j \rho\|_2^2.$$

Proof. Since

$$\|\rho\|_2^2 = \sum_{j=1}^{N} \rho_j^2 = \frac{1}{n} \sum_{j=1}^{N} n\rho_j^2 = \frac{1}{n} \sum_{j=1}^{N} \sum_{k=1}^{n} \rho_j^2,$$

we can rearrange the sum and get

$$\|\rho\|_2^2 = \frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{N}\rho_j^2 = \frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{N}\rho_{(j+k) \mod N}^2 = \frac{1}{n}\sum_{j=1}^{N}\sum_{k=1}^{n}\rho_{(j+k) \mod N}^2$$
$$= \frac{1}{n}\sum_{j=1}^{N}\|R_j\rho\|_2^2.$$

□

Corollary 3. *Given $M\Gamma = 0$, we have*

$$\|y - D_G\Gamma\|_2^2 = \frac{1}{n}\sum_{j=1}^{N}\|R_j y - D\alpha_j\|_2^2.$$

Proof. Using Proposition 9, we get

$$\|y - D_G\Gamma\|_2^2 = \frac{1}{n}\sum_{j=1}^{N}\|R_j y - R_j D_G\Gamma\|_2^2 = \frac{1}{n}\sum_{j=1}^{N}\|R_j y - \Omega_j\Gamma\|_2^2.$$

Now since $M\Gamma = 0$, then by definition of M, we have $\Omega_j\Gamma = D\alpha_j$ (see (6)), and this completes the proof. □

Recall Definition 6. Multiplying the corresponding matrices gives

Proposition 10. *We have the following equality for all $i = 1, \ldots P$:*

$$S_B R_i = S_T R_{i+1}. \tag{22}$$

To facilitate the proof, we introduce extension of Definition 6 to multiple shifts as follows.

Definition 16. Let n be fixed. For $k = 0, \ldots, n-1$ let

1. $S_T^{(k)} := \begin{bmatrix} I_{n-k} & \mathbf{0} \end{bmatrix}$ and $S_B^{(k)} := \begin{bmatrix} \mathbf{0} & I_{n-k} \end{bmatrix}$ denote the operators extracting the top (resp. bottom) $n-k$ entries from a vector of length n; the matrices have dimension $(n-k) \times n$.
2. $Z_B^{(k)} := \begin{bmatrix} S_B^{(k)} \\ \mathbf{0}_{k\times n} \end{bmatrix}$ and $Z_T^{(k)} := \begin{bmatrix} \mathbf{0}_{k\times n} \\ S_T^{(k)} \end{bmatrix}$.
3. $W_B^{(k)} := \begin{bmatrix} \mathbf{0}_{k\times n} \\ S_B^{(k)} \end{bmatrix}$ and $W_T^{(k)} := \begin{bmatrix} S_T^{(k)} \\ \mathbf{0}_{k\times n} \end{bmatrix}$.

Note that $S_B = S_B^{(1)}$ and $S_T = S_T^{(1)}$. We have several useful consequences of the above definitions. The proofs are carried out via elementary matrix identities and are left to the reader.

Proposition 11. *For any $n \in \mathbb{N}$, the following hold:*

1. $Z_T^{(k)} = \left(Z_T^{(1)}\right)^k$ *and* $Z_B^{(k)} = \left(Z_B^{(1)}\right)^k$ *for* $k = 0, \ldots, n-1$;
2. $W_T^{(k)} W_T^{(k)} = W_T^{(k)}$ *and* $W_B^{(k)} W_B^{(k)} = W_B^{(k)}$ *for* $k = 0, \ldots, n-1$;
3. $W_T^{(k)} W_B^{(j)} = W_B^{(j)} W_T^{(k)}$ *for* $j, k = 0, \ldots, n-1$;
4. $Z_B^{(k)} = Z_B^{(k)} W_B^{(k)}$ *and* $Z_T^{(k)} = Z_T^{(k)} W_T^{(k)}$ *for* $k = 0, \ldots, n-1$;
5. $W_B^{(k)} = Z_T^{(1)} W_B^{(k-1)} Z_B^{(1)}$ *and* $W_T^{(k)} = Z_B^{(1)} W_T^{(k-1)} Z_T$ *for* $k = 1, \ldots, n-1$;
6. $Z_B^{(k)} Z_T^{(k)} = W_T^{(k)}$ *and* $Z_T^{(k)} Z_B^{(k)} = W_B^{(k)}$ *for* $k = 0, \ldots, n-1$;
7. $(n-1)\, I_{n\times n} = \sum_{k=1}^{n-1} \left(W_B^{(k)} + W_T^{(k)}\right)$.

Proposition 12. *If the vectors $u_1, \ldots, u_N \in \mathbb{R}^n$ satisfy pairwise*

$$S_B u_i = S_T u_{i+1},$$

then they also satisfy for each $k = 0, \ldots, n-1$ the following:

$$W_B^{(k)} u_i = Z_T^{(k)} u_{i+k}, \tag{23}$$

$$Z_B^{(k)} u_i = W_T^{(k)} u_{i+k}. \tag{24}$$

Proof. It is easy to see that the condition $S_B u_i = S_T u_{i+1}$ directly implies

$$Z_B^{(1)} u_i = W_T^{(1)} u_{i+1}, \quad W_B^{(1)} u_i = Z_T^{(1)} u_{i+1} \quad \forall i. \tag{25}$$

Let us first prove (23) by induction on k. The base case $k = 1$ is precisely (25). Assuming validity for $k-1$ and $\forall i$, we have

$$\begin{aligned}
W_B^{(k)} u_i =& Z_T^{(1)} W_B^{(k-1)} Z_B^{(1)} u_i && \text{(by Proposition 11, item 5)}\\
=& Z_T^{(1)} W_B^{(k-1)} W_T^{(1)} u_{i+1} && \text{(by (25))}\\
=& Z_T^{(1)} W_T^{(1)} W_B^{(k-1)} u_{i+1} && \text{(by Proposition 11, item 3)}\\
=& Z_T^{(1)} W_T^{(1)} Z_T^{(k-1)} u_{i+k} && \text{(by the induction hypothesis)}\\
=& Z_T^{(1)} Z_T^{(k-1)} u_{i+k} && \text{(by Proposition 11, item 4)}\\
=& Z_T^{(k)} u_{i+k}. && \text{(by Proposition 11, item 1)}
\end{aligned}$$

To prove (24) we proceed as follows:

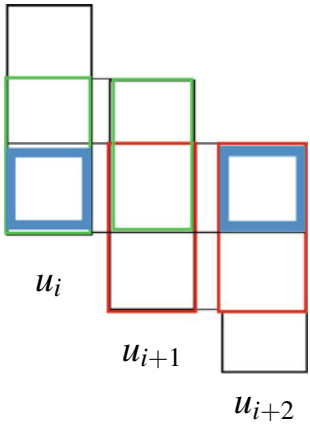


Fig. 11 Illustration to the proof of Proposition 12. The green pair is equal, as well as the red pair. It follows that the blue elements are equal as well.

$$\begin{aligned} Z_B^{(k)} u_i &= Z_B^{(k)} W_B^{(k)} u_i && \text{(by Proposition 11, item 4)} \\ &= Z_B^{(k)} Z_T^{(k)} u_{i+k} && \text{(by (23) which is already proved)} \\ &= W_T^{(k)} u_{i+k}. && \text{(by Proposition 11, item 6)} \end{aligned}$$

This finishes the proof of Proposition 12. □

Example 1. To help the reader understand the claim of Proposition 12, consider the case $k = 2$, and take some three vectors u_i, u_{i+1}, u_{i+2}. We have $S_B u_i = S_T u_{i+1}$ and also $S_B u_{i+1} = S_T u_{i+2}$. Then clearly $S_B^{(2)} u_i = S_T^{(2)} u_{i+2}$ (see Figure 11 on page 37) and therefore $W_B^{(2)} u_i = Z_T^{(2)} u_{i+2}$.

Let us now present the proof of Lemma 2.

Proof. We show equivalence in two directions.

- (1) $\Longrightarrow$ (2): Let $M\Gamma = 0$. Define $x := D_G \Gamma$, and then further denote $x_i := R_i x$. Then on the one hand:

$$\begin{aligned} x_i &= R_i D_G \Gamma \\ &= \Omega_i \Gamma && \text{(definition of } \Omega_i) \\ &= D\alpha_i. && (M\Gamma = 0) \end{aligned}$$

  On the other hand, because of (22) we have $S_B R_i x = S_T R_{i+1} x$, and by combining the two, we conclude that $S_B D\alpha_i = S_T D\alpha_{i+1}$.
- (2) $\Longrightarrow$ (1): In the other direction, suppose that $S_B D\alpha_i = S_T D\alpha_{i+1}$. Denote $u_i := D\alpha_i$. Now consider the product $\Omega_i \Gamma$ where $\Omega_i = R_i D_G$. One can easily be convinced that in fact

$$\Omega_i \Gamma = \frac{1}{n}\left(\sum_{k=1}^{n-1}\left(Z_B^{(k)} u_{i-k} + Z_T^{(k)} u_{i+k}\right) + u_i\right).$$

Therefore

$$\begin{aligned}(\Omega_i - Q_i)\,\Gamma &= \frac{1}{n}\left(u_i + \sum_{k=1}^{n-1}\left(Z_B^{(k)} u_{i-k} + Z_T^{(k)} u_{i+k}\right)\right) - u_i \\ &= \frac{1}{n}\left(\sum_{k=1}^{n-1}\left(W_T^{(k)} u_i + W_B^k u_i\right) - (n-1)\,u_i\right) \qquad \text{(by Proposition 12)} \\ &= 0. \qquad \text{(by Proposition 11, item 7)}\end{aligned}$$

Since this holds for all i, we have shown that $M\Gamma = 0$.

□

Appendix C: Proof of Theorem 6

Recall that $M_A = \frac{1}{n}\sum_i R_i^T P_{s_i} R_i$. We first show that M_A is a contraction.

Proposition 13. $\|M_A\|_2 \leqslant 1$.

Proof. Closely following a similar proof in [45], divide the index set $\{1, \ldots, N\}$ into n groups representing *non-overlapping* patches: for $i = 1, \ldots, n$ let

$$K(i) := \left\{i, i+n, \ldots, i + \left(\left\lfloor \frac{N}{n} \right\rfloor - 1\right) n\right\} \mod N.$$

Now

$$\begin{aligned}\|M_A x\|_2 &= \frac{1}{n}\left\|\sum_{i=1}^{N} R_i^T P_{s_i} R_i x\right\|_2 \\ &= \frac{1}{n}\left\|\sum_{i=1}^{n}\sum_{j\in K(i)} R_j^T P_{s_j} R_j x\right\|_2 \\ &\leqslant \frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{j\in K(i)} R_j^T P_j R_j x\right\|_2.\end{aligned}$$

By construction, $R_j R_k^T = \mathbf{0}_{n\times n}$ for $j, k \in K(i)$ and $j \neq k$. Therefore for all $i = 1, \ldots, n$ we have

$$\left\| \sum_{j \in K(i)} R_j^T P_{s_j} R_j x \right\|_2^2 = \sum_{j \in K(i)} \left\| R_j^T P_{s_j} R_j x \right\|_2^2 \leqslant \sum_{j \in K(i)} \left\| R_j x \right\|_2^2 \leqslant \|x\|_2^2 .$$

Substituting in back into the preceding inequality finally gives

$$\|M_A x\|_2 \leqslant \frac{1}{n} \sum_{i=1}^{n} \|x\|_2 = \|x\|_2 .$$

□

Now let us move on to prove Theorem 6.

Proof. Define

$$\hat{P}_i := (I - P_{s_i}) R_i.$$

It is easy to see that

$$\sum_i \hat{P}_i^T \hat{P}_i = A_{\mathscr{S}}^T A_{\mathscr{S}}.$$

Let the SVD of $A_{\mathscr{S}}$ be

$$A_{\mathscr{S}} = U \Sigma V^T.$$

Now

$$V \Sigma^2 V^T = A_{\mathscr{S}}^T A_{\mathscr{S}} = \sum_i \hat{P}_i^T \hat{P}_i = \sum_i R_i^T R_i - \underbrace{\sum_i R_i^T P_{s_i} R_i}_{:=T} = nI - T.$$

Therefore $T = nI - V \Sigma V^T$, and

$$M_A = \frac{1}{n} T = I - \frac{1}{n} V \Sigma^2 V^T = V \left(I - \frac{\Sigma^2}{n} \right) V^T.$$

This shows that the eigenvalues of M_A are $\tau_i = 1 - \frac{\sigma_i^2}{n}$ where $\{\sigma_i\}$ are the singular values of $A_{\mathscr{S}}$. Thus we obtain

$$M_A^k = V \operatorname{diag} \left\{ \tau_i^k \right\} V^T.$$

If $\sigma_i = 0$ then $\tau_i = 1$, and in any case, by Proposition 13, we have $|\tau_i| \leqslant 1$. Let the columns of the matrix W consist of the singular vectors of $A_{\mathscr{S}}$ corresponding to $\sigma_i = 0$ (and so span $W = \mathscr{N}(A_{\mathscr{S}})$), then

$$\lim_{k\to\infty} M_A^k = WW^T.$$

Thus, as $k \to \infty$, M_A^k tends to the orthogonal projector onto $\mathscr{N}(A_{\mathscr{S}})$. The convergence is evidently linear, the constant being dependent upon $\{\tau_i\}$. □

Appendix D: Proof of Theorem 8

Recall that the signal consists of s constant segments of corresponding lengths $\ell_1, \ldots, \ell_s$. We would like to compute the MSE for every pixel within every such segment of length $\alpha := \ell_r$. For each patch, the oracle provides the locations of the jump points within the patch.

Let us calculate the MSE for pixel with index 0 inside a constant (**nonzero**) segment $[-k, \alpha - k - 1]$ with value v (Figure 12 on page 41 might be useful). The oracle estimator has the explicit formula

$$\hat{x}_A^{r,k} = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{b_j - a_j + 1} \sum_{i=a_j}^{b_j} (v + z_i), \tag{26}$$

where $j = 1, \ldots, n$ corresponds to the index of the overlapping patch containing the pixel, intersecting the constant segment on $\left[a_j, b_j\right]$, so that

$$\begin{aligned} a_j &= -\min(k, n-j), \\ b_j &= \min(\alpha - k - 1, j - 1). \end{aligned}$$

Now, the oracle error for the pixel is

$$\begin{aligned} \hat{x}_A^{r,k} - v &= \frac{1}{n}\sum_{j=1}^{n} \frac{1}{b_j - a_j + 1} \sum_{i=a_j}^{b_j} z_i \\ &= \sum_{i=-k}^{\alpha-k-1} c_{i,\alpha,n,k} z_i, \end{aligned}$$

where the coefficients $c_{i,\alpha,n,k}$ are some *positive* rational numbers depending only on i, α, n and k. It is easy to check by rearranging the above expression that

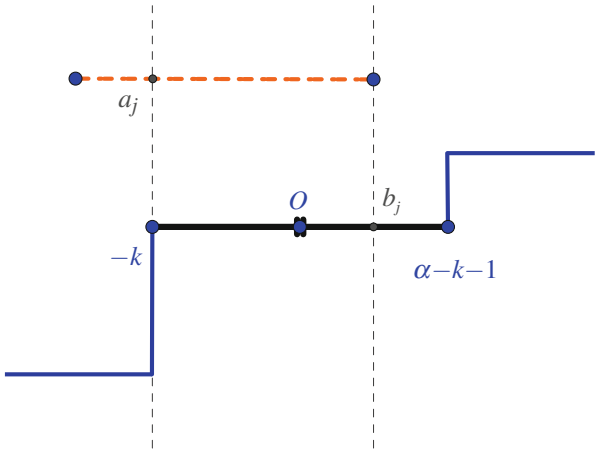


Fig. 12 The oracle estimator for the pixel O in the segment (black). The orange line is patch number $j = 1, \ldots, n$, and the relevant pixels are between a_j and b_j. The signal itself is shown to extend beyond the segment (blue line).

$$\sum_{i=-k}^{\alpha-k-1} c_{i,\alpha,n,k} = 1, \tag{27}$$

and furthermore, denoting $d_i := c_{i,\alpha,n,k}$ for fixed α, n, k, we also have that

$$d_{-k} < d_{-k+1} < \ldots d_0 > d_1 > \ldots d_{\alpha-k-1}. \tag{28}$$

Example 2. $n = 4,\ \alpha = 3$

- For $k = 1$:

$$\begin{aligned}\hat{x}_A^{r,k} - v &= \frac{1}{4}\left(\frac{1}{2}+\frac{1}{3}+\frac{1}{3}+\frac{1}{2}\right)z_0 + \frac{1}{4}\left(\frac{1}{2}+\frac{1}{3}+\frac{1}{3}\right)z_{-1} + \frac{1}{4}\left(\frac{1}{3}+\frac{1}{3}+\frac{1}{2}\right)z_1 \\ &= \underbrace{\frac{7}{24}}_{d_{-1}} z_{-1} + \underbrace{\frac{5}{12}}_{d_0} z_0 + \underbrace{\frac{7}{24}}_{d_1} z_1\end{aligned}$$

- For $k = 2$:

$$\begin{aligned}\hat{x}_A^{r,k} - v &= \frac{1}{4}\left(\frac{1}{3}+\frac{1}{3}+\frac{1}{2}+1\right)z_0 + \frac{1}{4}\left(\frac{1}{3}+\frac{1}{3}+\frac{1}{2}\right)z_{-1} + \frac{1}{4}\left(\frac{1}{3}+\frac{1}{3}\right)z_{-2} \\ &= \frac{13}{24}z_0 + \frac{7}{24}z_{-1} + \frac{1}{6}z_{-2}\end{aligned}$$

Now consider the optimization problem

$$\min_{c\in\mathbb{R}^\alpha} c^T c \quad \text{s.t } \mathbf{1}^T c = 1.$$

It can be easily verified that it has the optimal value $\frac{1}{\alpha}$, attained at $c^* = \alpha\mathbf{1}$. From this, (27) and (28), it follows that

$$\sum_{i=-k}^{\alpha-k-1} c_{i,\alpha,n,k}^2 > \frac{1}{\alpha}.$$

Since the z_i are i.i.d., we have

$$\mathbb{E}\left(\hat{x}_A^{r,k} - v\right)^2 = \sigma^2 \sum_{i=-k}^{\alpha-k-1} c_{i,\alpha,n,k}^2,$$

while for the entire nonzero segment of length $\alpha = \ell_r$

$$E_r := \mathbb{E}\left(\sum_{k=0}^{\alpha-1}\left(\hat{x}_A^{r,k} - v\right)^2\right) = \sum_{k=0}^{\alpha-1}\mathbb{E}\left(\hat{x}_A^{r,k} - v\right)^2 = \sigma^2 \sum_{k=0}^{\alpha-1}\sum_{i=-k}^{\alpha-k-1} c_{i,\alpha,n,k}^2.$$

Defining

$$R\left(n,\alpha\right) := \sum_{k=0}^{\alpha-1}\sum_{i=-k}^{\alpha-k-1} c_{i,\alpha,n,k}^2,$$

we obtain that $R\left(n,\alpha\right) > 1$ and furthermore

$$\mathbb{E}\left\|\hat{x}_A - x\right\|^2 = \sum_{r=1}^{s} E_r = \sigma^2 \sum_{r=1}^{s} R\left(n,\ell_r\right) > s\sigma^2.$$

This proves item (1) of Theorem 8. For showing the explicit formulas for $R\left(n,\alpha\right)$ in item (2), we have used automatic symbolic simplification software MAPLE [39].

By construction (26), it is not difficult to see that if $n \geqslant \alpha$ then

$$R(n,\alpha) = \frac{1}{n^2}\sum_{k=0}^{\alpha-1}\Big(\sum_{j=0}^{k}\left(2H_{\alpha-1} - H_k + \frac{n-\alpha+1}{\alpha} - H_{\alpha-1-j}\right)^2$$
$$+ \sum_{j=k+1}^{\alpha-1}\left(2H_{\alpha-1} - H_{\alpha-k-1} + \frac{n-\alpha+1}{\alpha} - H_j\right)^2\Big),$$

where $H_k := \sum_{i=1}^{k}\frac{1}{i}$ is the k-th harmonic number. This simplifies to

$$R(n,\alpha) = 1 + \frac{\alpha(2\alpha H_\alpha^{(2)} + 2 - 3\alpha) - 1}{n^2},$$

where $H_k^{(2)} = \sum_{i=1}^{k} \frac{1}{i^2}$ is the k-th harmonic number of the second kind.

On the other hand, for $n \leqslant \frac{\alpha}{2}$ we have

$$R(n,\alpha) = \sum_{k=0}^{n-2} c_{n,k}^{(1)} + \sum_{k=n-1}^{\alpha-n} c_{n,k}^{(2)} + \sum_{k=\alpha-n+1}^{\alpha-1} c_{n,\alpha-1-k}^{(1)},$$

where

$$c_{n,k}^{(1)} = \frac{1}{n^2}\left(\sum_{j=k}^{n-1}\big(H_{n-1} - H_j + \frac{k+1}{n}\big)^2 + \sum_{i=n-k}^{n-1}\big(\frac{n-i}{n}\big)^2 + \sum_{i=0}^{k-1}\big(H_{n-1} - H_k + \frac{k-i}{n}\big)^2\right)$$

and

$$c_{n,k}^{(2)} = \frac{1}{n^2}\left(\sum_{j=k-n+1}^{k}\left(\frac{j-k+n}{n}\right)^2 + \sum_{j=k+1}^{k+n-1}\left(\frac{k+n-j}{n}\right)^2\right).$$

Automatic symbolic simplification of the above gives

$$R\,(n,\alpha) = \frac{11}{18} + \frac{2\alpha}{3n} - \frac{5}{18n^2} + \frac{\alpha-1}{3n^3}.$$

Appendix E: Generative Models for Patch-Sparse Signals

In this section we propose a general framework aimed at generating signals from the patch-sparse model. Our approach is to construct a graph-based model for the dictionary and subsequently use this model to generate dictionaries and signals which turn out to be much richer than those considered in Section 4.

Local Support Dependencies

We start by highlighting the importance of the local connections (recall Lemma 2) between the neighboring patches of the signal and therefore between the corresponding subspaces containing those patches. This in turn allows to characterize $\Sigma_{\mathscr{M}}$ as the set of all "realizable" paths in a certain dependency graph derived from the dictionary D. This point of view allows to describe the model $\mathscr{M}$ using only the intrinsic properties of the dictionary, in contrast to Theorem 2.

Proposition 14. *Let* $0 \neq x \in \mathscr{M}$ *and* $\Gamma amma \in \rho\,(x)$ *with* $\operatorname{supp}\Gamma = (S_1, \ldots, S_P)$. *Then for* $i = 1, \ldots, P$

$$\operatorname{rank}\left[S_B D_{S_i} \;\; -S_T D_{S_{i+1}}\right] < |S_i| + |S_{i+1}| \leqslant 2s, \tag{29}$$

where by convention $\operatorname{rank} \emptyset = -\infty$.

Proof. $x \in \mathscr{M}$ implies by Lemma 2 that for every $i = 1, \dots P$

$$[S_B D \quad -S_T D] \begin{bmatrix} \alpha_i \\ \alpha_{i+1} \end{bmatrix} = 0.$$

But

$$[S_B D \quad -S_T D] \begin{bmatrix} \alpha_i \\ \alpha_{i+1} \end{bmatrix} = \left[S_B D_{S_i} \quad -S_T D_{S_{i+1}}\right] \begin{bmatrix} \alpha_i | S_i \\ \alpha_{i+1} | S_{i+1} \end{bmatrix} = 0,$$

and therefore the matrix $\left[S_B D_{S_i} \;\; -S_T D_{S_{i+1}}\right]$ must be rank-deficient. Note in particular that the conclusion still holds if one (or both) of the $\{s_i, s_{i+1}\}$ is empty. □

The preceding result suggests a way to describe all the supports in $\Sigma_{\mathscr{M}}$.

Definition 17. Given a dictionary D, we define an abstract directed graph $\mathscr{G}_{D,s} = (V, E)$, with the vertex set

$$V = \{(i_1, \dots, i_k) \subset \{1, \dots, m\} : \quad \operatorname{rank} D_{i_1, \dots, i_k} = k < n\},$$

and the edge set

$$E = \left\{ (S_1, S_2) \in V \times V : \quad \operatorname{rank} [S_B D_{S_1} \quad -S_T D_{S_2}] < \min \{n - 1, |S_1| + |S_2|\} \right\}.$$

In particular, $\emptyset \in V$ and $(\emptyset, \emptyset) \in E$ with $\operatorname{rank} [\emptyset] := -\infty$.

Remark 4. It might be impossible to compute $\mathscr{G}_{D,s}$ in practice. However we set this issue aside for now and only explore the theoretical ramifications of its properties.

Definition 18. The set of all directed paths of length P in $\mathscr{G}_{D,s}$, not including the self-loop $\underbrace{(\emptyset, \emptyset, \dots \emptyset)}_{\times P}$, is denoted by $\mathscr{C}_{\mathscr{G}}(P)$.

Definition 19. A path $\mathscr{S} \in \mathscr{C}_{\mathscr{G}}(P)$ is called *realizable* if $\dim \ker A_{\mathscr{S}} > 0$. The set of all realizable paths in $\mathscr{C}_{\mathscr{G}}(P)$ is denoted by $\mathscr{R}_{\mathscr{G}}(P)$.
Thus we have the following result.

Theorem 9. *Suppose* $0 \neq x \in \mathscr{M}$. *Then*

1. *Every representation* $\Gamma = (\alpha_i)_{i=1}^{P} \in \rho(x)$ *satisfies* $\operatorname{supp} \Gamma \in \mathscr{C}_{\mathscr{G}}(P)$, *and therefore*

$$\Sigma_{\mathscr{M}} \subseteq \mathscr{R}_{\mathscr{G}}(P). \tag{30}$$

2. The model $\mathscr{M}$ can be characterized "intrinsically" by the dictionary as follows:

$$\mathscr{M} = \bigcup_{\mathscr{S} \in \mathscr{R}_{\mathscr{G}}(P)} \ker A_{\mathscr{S}}. \tag{31}$$

Proof. Let $\operatorname{supp} \Gamma = (S_1, \ldots, S_P)$ with $S_i = \operatorname{supp} \alpha_i$ if $\alpha_i \neq \mathbf{0}$, and $S_i = \emptyset$ if $\alpha_i = \mathbf{0}$. Then by Proposition 14, we must have that

$$\operatorname{rank} \left[S_B D_{S_i} \; - S_T D_{S_{i+1}} \right] < |S_i| + |S_{i+1}| \leqslant 2s.$$

Furthermore, since $\Gamma \in \rho(x)$ we must have that D_{s_i} is full rank for each $i = 1, \ldots, P$. Thus $(S_i, S_{i+1}) \in \mathscr{G}_{D,s}$, and so $\operatorname{supp} \Gamma \in \mathscr{R}_{\mathscr{G}}(P)$. Since by assumption $\operatorname{supp} \Gamma \in \Sigma_{\mathscr{M}}$, this proves (30).

To show (31), notice that if $\operatorname{supp} \Gamma amma \in \mathscr{R}_{\mathscr{G}}(P)$, then for every $x \in \ker A_{\operatorname{supp} \Gamma}$, we have $R_i x = P_{S_i} R_i x$, i.e., $R_i x = D\alpha_i$ for some α_i with $\operatorname{supp} \alpha_i \subseteq S_i$. Clearly in this case $|\operatorname{supp} \alpha_i| \leqslant s$ and therefore $x \in \mathscr{M}$. The other direction of (31) follows immediately from the definitions. □

Definition 20. The dictionary D is called "(s, P)-good" if

$$|\mathscr{R}_{\mathscr{G}}(P)| > 0.$$

Theorem 10. *The set of "(s, P)-good" dictionaries has measure zero in the space of all $n \times m$ matrices.*

Proof. Every low-rank condition defines a finite number of algebraic equations on the entries of D (given by the vanishing of all the $2s \times 2s$ minors of $\left[S_B D_{S_i} \; S_T D_{S_j} \right]$). Since the number of possible graphs is finite (given fixed n, m and s), the resulting solution set is a finite union of semi-algebraic sets of low dimension and hence has measure zero. □

Constructing "Good" Dictionaries

The above considerations suggest that the good dictionaries are hard to come by; here we provide an example of an explicit construction.

We start by defining an abstract graph $\mathscr{G}$ with some desirable properties, and subsequently look for a nontrivial realization D of the graph, so that in addition $\mathscr{R}_{\mathscr{G}} \neq \emptyset$.

In this context, we would want $\mathscr{G}$ to contain *sufficiently many different long cycles*, which would correspond to long signals and a rich resulting model $\mathscr{M}$. In contrast with the models from Subsection 4.2 (where all the graphs consist of a single cycle), one therefore should allow for some branching mechanism. An example of a possible $\mathscr{G}$ is given in Figure 13 on page 46. Notice that due to the

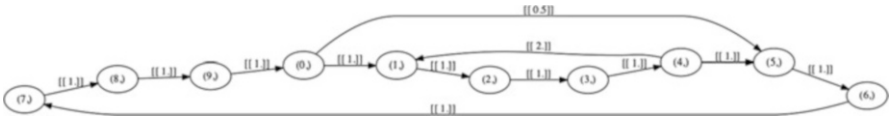


Fig. 13 A possible dependency graph $\mathscr{G}$ with $m = 10$. In this example, $|\mathscr{C}_{\mathscr{G}}(70)| = 37614$.

structure of $\mathscr{G}$, there are many possible paths in $\mathscr{C}_{\mathscr{G}}(P)$. In fact, a direct search algorithm yields $|\mathscr{C}_{\mathscr{G}}(70)| = 37614$.

Every edge in $\mathscr{G}$ corresponds to a conditions of the form (29) imposed on the entries of D. As discussed in Theorem 10, this in turn translates to a set of algebraic equations. So the natural idea would be to write out the large system of such equations and look for a solution over the field $\mathbb{R}$ by well-known algorithms in numerical algebraic geometry [5]. However, this approach is highly impractical because these algorithms have (single or double) exponential running time. We consequently propose a simplified, more direct approach to the problem.

In detail, we replace the low-rank conditions (29) with more explicit and restrictive ones below.

Assumptions(*) For each $(S_i, S_j) \in \mathscr{G}$ we have $|S_i| = |S_j| = k$. We require that $\operatorname{span} S_B D_{S_i} = \operatorname{span} S_T D_{S_j} = \Lambda_{i,j}$ with $\dim \Lambda_{i,j} = k$. Thus there exists a non-singular transfer matrix $C_{i,j} \in \mathbb{R}^{k\times k}$ such that

$$S_B D_{S_i} = C_{i,j} S_T D_{S_j}. \tag{32}$$

In other words, every column in $S_B D_{S_i}$ must be a specific linear combination of the columns in $S_T D_{S_j}$. This is much more restrictive than the low-rank condition, but on the other hand, given the matrix $C_{i,j}$, it defines a set of linear constraints on D. To summarize, the final algorithm is presented in Algorithm 5. In general, nothing guarantees that for a particular choice of $\mathscr{G}$ and the transfer matrices, there is a nontrivial solution D; however, in practice we do find such solutions. For example, taking the graph from Figure 13 on page 46 and augmenting it with the matrices $C_{i,j}$ (scalars in this case), we obtain a solution over $\mathbb{R}^6$ which is shown in Figure 14 on page 47. Notice that while the resulting dictionary has a Hankel-type structure similar to what we have seen previously, the additional dependencies between the atoms produce a rich signal space structure, as we shall demonstrate in the following section.

Algorithm 5 Finding a realization D of the graph $\mathscr{G}$

1. Input: a graph $\mathscr{G}$ satisfying the **Assumptions(*)** above, and the dimension n of the realization space $\mathbb{R}^n$.
2. Augment the edges of $\mathscr{G}$ with arbitrary nonsingular transfer matrices $C_{i,j}$.
3. Construct the system of linear equations given by (32).
4. Find a nonzero D solving the system above over $\mathbb{R}^n$.

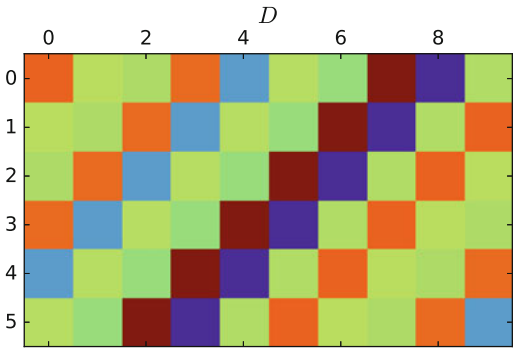


Fig. 14 A realization $D \in \mathbb{R}^{6\times 10}$ of $\mathscr{G}$ from Figure 13 on page 46.

Generating Signals

Now suppose the graph $\mathscr{G}$ is known (or can be easily constructed). Then this gives a simple procedure to generate signals from $\mathscr{M}$, presented in Algorithm 6.

Algorithm 6 Constructing a signal from $\mathscr{M}$ via $\mathscr{G}$

1. Construct a path $\mathscr{S} \in \mathscr{C}_{\mathscr{G}}(P)$.
2. Construct the matrix $A_{\mathscr{S}}$.
3. Find a nonzero vector in $\ker A_{\mathscr{S}}$.

Let us demonstrate this on the example in Figure 13 on page 46 and Figure 14 on page 47. Not all paths in $\mathscr{C}_{\mathscr{G}}$ are realizable, but it turns out that in this example we have $|\mathscr{R}_{\mathscr{G}}(70)| = 17160$. Three different signals and their supports $\mathscr{S}$ are shown in Figure 15 on page 48. As can be seen from these examples, the resulting model $\mathscr{M}$ is indeed much richer than the signature-type construction from Subsection 4.2.

An interesting question arises: given $\mathscr{S} \in \mathscr{C}_{\mathscr{G}}(P)$, can we say something about $\dim\ker A_{\mathscr{S}}$? In particular, when is it strictly positive (i.e., when $\mathscr{S} \in \mathscr{R}_{\mathscr{G}}(P)$?) While in general the question seems to be difficult, in some special cases this number can be estimated using only the properties of the local connections (S_i, S_{i+1}), by essentially counting the additional "degrees of freedom" when moving from patch i to patch $i+1$. To this effect, we prove two results.

Proposition 15. *For every $\mathscr{S} \in \mathscr{R}_{\mathscr{G}}(P)$, we have*

$$\dim\ker A_{\mathscr{S}} = \dim\ker M_*^{(\mathscr{S})}.$$

Proof. Notice that

$$\ker A_{\mathscr{S}} = \left\{ D_G^{(\mathscr{S})} \Gamma_{\mathscr{S}},\ M_*^{(\mathscr{S})} \Gamma_{\mathscr{S}} = 0 \right\} = im\left(D_G^{(\mathscr{S})}|_{\ker M_*^{(\mathscr{S})}} \right),$$

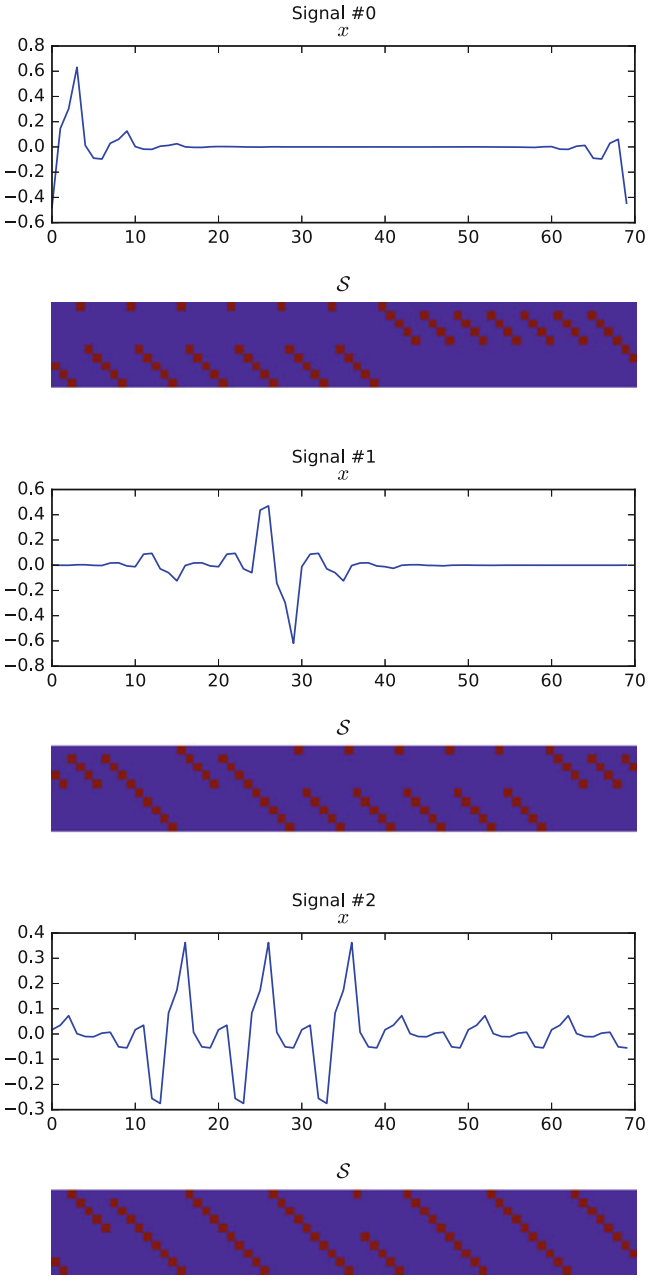


Fig. 15 Examples of signals from $\mathscr{M}$ and the corresponding supports $\mathscr{S}$.

and therefore $\dim\ker A_{\mathscr{S}} \leqslant \dim\ker M_*^{(\mathscr{S})}$. Furthermore, the map $D_G^{(\mathscr{S})}|_{\ker M_*^{(\mathscr{S})}}$ is injective, because if $D_G^{(\mathscr{S})}\Gamma_{\mathscr{S}} = 0$ and $M_*^{(\mathscr{S})}\Gamma_{\mathscr{S}} = 0$, we must have that $D_{S_i}\alpha_i|_{S_i} = 0$ and, since D_{s_i} has full rank, also $\alpha_i = 0$. The conclusion follows. □

Proposition 16. *Assume that the model satisfies* ***Assumptions(*)*** *above. Then for every* $\mathscr{S} \in \mathscr{R}_{\mathscr{G}}(P)$

$$\dim\ker A_{\mathscr{S}} \leqslant k.$$

Proof. The idea is to construct a spanning set for $\ker M_*^{(\mathscr{S})}$ and invoke Proposition 15. Let us relabel the nodes along $\mathscr{S}$ to be $1, 2, \ldots, P$. Starting from an arbitrary α_1 with support $|S_1| = k$, we use (32) to obtain, for $i = 1, 2, \ldots, P-1$, a formula for the next portion of the global representation vector $\Gamma amma$

$$\alpha_{i+1} = C_{i,i+1}^{-1}\alpha_i. \tag{33}$$

This gives a set Δ consisting of overall k linearly independent vectors $\Gamma amma_i$ with $\operatorname{supp}\Gamma_i = \mathscr{S}$. It may happen that equation (33) is not satisfied for $i = P$. However, every Γ with $\operatorname{supp}\Gamma = \mathscr{S}$ and $M_*^{(\mathscr{S})}\Gamma amma_{\mathscr{S}} = 0$ must belong to $\operatorname{span}\Delta$, and therefore

$$\dim\ker M_*^{(\mathscr{S})} \leqslant \dim\operatorname{span}\Delta = k.$$

□

We believe that Proposition 16 can be extended to more general graphs, not necessarily satisfying **Assumptions(*)**. In particular, the following estimate appears to hold for a general model $\mathscr{M}$ and $\mathscr{S} \in \mathscr{R}_{\mathscr{G}}(P)$:

$$\dim\ker A_{\mathscr{S}} \leqslant |S_1| + \sum_i \left(|S_{i+1}| - \operatorname{rank}\left[S_B D_{S_i}\ S_T D_{S_{i+1}}\right]\right).$$

We leave the rigorous proof of this result to a future work.

Further Remarks

While the model presented in this section is the hardest to analyze theoretically, even in the restricted case of **Assumptions(*)** (when does a nontrivial realization of a given $\mathscr{G}$ exist? How does the answer depend on n? When $\mathscr{R}_{\mathscr{G}}(P) \neq \emptyset$? etc?), we hope that this construction will be most useful in applications such as denoising of natural signals.

References

1. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: large-scale machine learning on heterogeneous systems (2015). http://tensorflow.org/. Software available from tensorflow.org
2. R. Aceska, J.L. Bouchot, S. Li, Local sparsity and recovery of fusion frames structured signals. preprint (2015). http://www.mathc.rwth-aachen.de/~bouchot/files/pubs/FusionCSfinal.pdf
3. M. Aharon, M. Elad, Sparse and redundant modeling of image content using an image-signature-dictionary. SIAM J. Imag. Sci. **1**(3), 228–247 (2008)
4. U. Ayaz, S. Dirksen, H. Rauhut, Uniform recovery of fusion frame structured sparse signals. Appl. Comput. Harmon. Anal. **41**(2), 341–361 (2016). https://doi.org/10.1016/j.acha.2016.03.006. http://www.sciencedirect.com/science/article/pii/S1063520316000294
5. S. Basu, R. Pollack, M.F. Roy, *Algorithms in Real Algebraic Geometry*. Algorithms and Computation in Mathematics, 2nd edn., vol. 10 (Springer, Berlin, 2006)
6. T. Blumensath, M. Davies, Sparse and shift-invariant representations of music. IEEE Trans. Audio Speech Lang. Process. **14**(1), 50–57 (2006). http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1561263
7. T. Blumensath, M.E. Davies, Sampling theorems for signals from the union of finite-dimensional linear subspaces. IEEE Trans. Inf. Theory **55**(4), 1872–1882 (2009). http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4802322
8. P. Boufounos, G. Kutyniok, H. Rauhut, Sparse recovery from combined fusion frame measurements. IEEE Trans. Inf. Theory **57**(6), 3864–3876 (2011). https://doi.org/10.1109/TIT.2011.2143890
9. S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011). http://dx.doi.org/10.1561/2200000016
10. H. Bristow, A. Eriksson, S. Lucey, Fast convolutional sparse coding. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 391–398
11. A.M. Bruckstein, D.L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**(1), 34–81 (2009). http://epubs.siam.org/doi/abs/10.1137/060657704
12. E.J. Candes, Modern statistical estimation via oracle inequalities. Acta Numer. **15**, 257–325 (2006). http://journals.cambridge.org/abstract_S0962492906230010
13. S. Chen, S.A. Billings, W. Luo, Orthogonal least squares methods and their application to non-linear system identification. Int. J. Control. **50**(5), 1873–1896 (1989)
14. W. Dong, L. Zhang, G. Shi, X. Li, Nonlocally centralized sparse representation for image restoration. IEEE Trans. Image Process. **22**(4), 1620–1630 (2013)
15. D.L. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. Proc. Natl. Acad. Sci. **100**(5), 2197–2202 (2003). doi:10.1073/pnas.0437847100. http://www.pnas.org/content/100/5/2197
16. C. Ekanadham, D. Tranchina, E.P. Simoncelli, A unified framework and method for automatic neural spike identification. J. Neurosci. Methods **222**, 47–55 (2014). doi:10.1016/j.jneumeth.2013.10.001. http://www.sciencedirect.com/science/article/pii/S0165027013003415
17. M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing* (Springer, New York, 2010)

18. M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans. Image Process. **15**(12), 3736–3745 (2006)
19. Y.C. Eldar, M. Mishali, Block sparsity and sampling over a union of subspaces, in *2009 16th International Conference on Digital Signal Processing* (IEEE, New York, 2009), pp. 1–8. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5201211
20. Y.C. Eldar, M. Mishali, Robust recovery of signals from a structured union of subspaces. IEEE Trans. Inf. Theory **55**(11), 5302–5316 (2009)
21. Finite Frames - Theory and Applications. http://www.springer.com/birkhauser/mathematics/book/978-0-8176-8372-6
22. S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Springer, New York, 2013). http://link.springer.com/content/pdf/10.1007/978-0-8176-4948-7.pdf
23. D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**(1), 17–40 (1976)
24. R. Glowinski, On alternating direction methods of multipliers: a historical perspective, in *Modeling, Simulation and Optimization for Science and Technology* (Springer, Dordrecht, 2014), pp. 59–82
25. R. Grosse, R. Raina, H. Kwong, A.Y. Ng, Shift-invariance sparse coding for audio classification (2012). arXiv preprint arXiv: 1206.5241
26. R. Grosse, R. Raina, H. Kwong, A.Y. Ng, Shift-invariance sparse coding for audio classification. arXiv: 1206.5241 [cs, stat] (2012). http://arxiv.org/abs/1206.5241. arXiv: 1206.5241
27. S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, L. Zhang, Convolutional sparse coding for image super-resolution, in *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 1823–1831
28. F. Heide, W. Heidrich, G. Wetzstein, Fast and flexible convolutional sparse coding, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, New York, 2015), pp. 5135–5143
29. J. Huang, T. Zhang, D. Metaxas, Learning with structured sparsity. J. Mach. Learn. Res. **12**, 3371–3412 (2011)
30. J. Huang, T. Zhang, et al., The benefit of group sparsity. Ann. Stat. **38**(4), 1978–2004 (2010)
31. K. Kavukcuoglu, P. Sermanet, Y.l. Boureau, K. Gregor, M. Mathieu, Y.L. Cun, Learning convolutional feature hierarchies for visual recognition, in *Advances in Neural Information Processing Systems*, ed. by J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, A. Culotta, vol. 23 (Curran Associates, Red Hook, 2010), pp. 1090–1098. http://papers.nips.cc/paper/4133-learning-convolutional-feature-hierarchies-for-visual-recognition.pdf
32. A. Kyrillidis, L. Baldassarre, M.E. Halabi, Q. Tran-Dinh, V. Cevher, Structured sparsity: discrete and convex approaches, in *Compressed Sensing and Its Applications*. Applied and Numerical Harmonic Analysis, ed. by H. Boche, R. Calderbank, G. Kutyniok, J. Vybíral (Springer, Cham, 2015), pp. 341–387. http://link.springer.com/chapter/10.1007/978-3-319-16042-9_12. https://doi.org/10.1007/978-3-319-16042-9_12
33. P.L. Lions, B. Mercier, Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**(6), 964–979 (1979)
34. M.A. Little, N.S. Jones, Generalized methods and solvers for noise removal from piecewise constant signals. II. New methods. Proc. R. Soc. Lond. A: Math. Phys. Eng. Sci. rspa20100674 (2011). doi:https://doi.org/10.1098/rspa.2010.0674. http://rspa.royalsocietypublishing.org/content/early/2011/06/07/rspa.2010.0674
35. Y.M. Lu, M.N. Do, A theory for sampling signals from a union of subspaces. IEEE Trans. Signal Process. **56**, 2334–2345 (2007)
36. J. Mairal, G. Sapiro, M. Elad, Learning multiscale sparse representations for image and video restoration. Multiscale Model. Simul. **7**(1), 214–241 (2008)
37. J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Non-local sparse models for image restoration. in *2009 IEEE 12th International Conference on Computer Vision* (IEEE, New York, 2009), pp. 2272–2279

38. J. Mairal, F. Bach, J. Ponce, Sparse modeling for image and vision processing. Found. Trends Comput. Graph. Vis. **8**(2–3), 85–283 (2014). https://doi.org/10.1561/0600000058. http://www.nowpublishers.com/article/Details/CGV-058
39. Maplesoft, a division of Waterloo Maple Inc. http://www.maplesoft.com
40. V. Papyan, M. Elad, Multi-scale patch-based image restoration. IEEE Trans. Image Process. **25**(1), 249–261 (2016). https://doi.org/10.1109/TIP.2015.2499698
41. V. Papyan, Y. Romano, M. Elad, Convolutional neural networks analyzed via convolutional sparse coding. J. Mach. Learn. Res. **18**(83), 1–52 (2017)
42. V. Papyan, J. Sulam, M. Elad, Working locally thinking globally: theoretical guarantees for convolutional sparse coding. IEEE Trans. Signal Process. **65**(21), 5687–5701 (2017)
43. Y.C. Pati, R. Rezaiifar, P. Krishnaprasad, Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition, in *Asilomar Conference on Signals, Systems and Computers* (IEEE, New York, 1993), pp. 40–44
44. R. Quiroga, Spike sorting. Scholarpedia **2**(12), 3583 (2007). https://doi.org/10.4249/scholarpedia.3583
45. Y. Romano, M. Elad, Boosting of image denoising algorithms. SIAM J. Imag. Sci. **8**(2), 1187–1219 (2015)
46. Y. Romano, M. Elad, Patch-disagreement as away to improve K-SVD denoising, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York, 2015), pp. 1280–1284
47. Y. Romano, M. Protter, M. Elad, Single image interpolation via adaptive nonlocal sparsity-based modeling. IEEE Trans. Image Process. **23**(7), 3085–3098 (2014)
48. L.I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms. Physica D **60**(1), 259–268 (1992). http://www.sciencedirect.com/science/article/pii/016727899290242F
49. C. Rusu, B. Dumitrescu, S. Tsaftaris, Explicit shift-invariant dictionary learning. IEEE Signal Process. Lett. **21**, 6–9 (2014). http://www.schur.pub.ro/Idei2011/Articole/SPL_2014_shifts.pdf
50. E. Smith, M.S. Lewicki, Efficient coding of time-relative structure using spikes. Neural Comput. **17**(1), 19–45 (2005). http://dl.acm.org/citation.cfm?id=1119614
51. A.M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A.K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J.P. Yue, J.W. Gray, A.N. Jain, D. Pinkel, D.G. Albertson, Assembly of microarrays for genome-wide measurement of DNA copy number. Nat. Genet. **29**(3), 263–264 (2001). https://doi.org/10.1038/ng754. https://www.nature.com/ng/journal/v29/n3/full/ng754.html
52. J. Sulam, M. Elad, Expected patch log likelihood with a sparse prior, in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition* (Springer, New York, 2015), pp. 99–111
53. J. Sulam, B. Ophir, M. Elad, Image denoising through multi-scale learnt dictionaries, in *2014 IEEE International Conference on Image Processing (ICIP)* (IEEE, New York, 2014), pp. 808–812
54. J.J. Thiagarajan, K.N. Ramamurthy, A. Spanias, Shift-invariant sparse representation of images using learned dictionaries, in *IEEE Workshop on Machine Learning for Signal Processing, 2008, MLSP 2008* (2008), pp. 145–150 https://doi.org/10.1109/MLSP.2008.4685470
55. J.A. Tropp, A.C. Gilbert, M.J. Strauss, Algorithms for simultaneous sparse approximation. Part i: greedy pursuit. Signal Process. **86**(3), 572–588 (2006)
56. J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**(11), 2861–2873 (2010)
57. G. Yu, G. Sapiro, S. Mallat, Solving inverse problems with piecewise linear estimators: From gaussian mixture models to structured sparsity. IEEE Trans. Image Process. **21**(5), 2481–2499 (2012). https://doi.org/10.1109/TIP.2011.2176743
58. M.D. Zeiler, D. Krishnan, G.W. Taylor, R. Fergus, Deconvolutional networks, in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE, New York, 2010), pp. 2528–2535. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5539957

59. M. Zeiler, G. Taylor, R. Fergus, Adaptive deconvolutional networks for mid and high level feature learning, in *2011 IEEE International Conference on Computer Vision (ICCV)*, pp. 2018–2025 (2011). doi:10.1109/ICCV.2011.6126474
60. D. Zoran, Y. Weiss, From learning models of natural image patches to whole image restoration, in *2011 IEEE International Conference on Computer Vision (ICCV)* (IEEE, New York, 2011), pp. 479–486. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6126278

Fourier Phase Retrieval: Uniqueness and Algorithms

Tamir Bendory, Robert Beinert, and Yonina C. Eldar

Abstract The problem of recovering a signal from its phaseless Fourier transform measurements, called Fourier phase retrieval, arises in many applications in engineering and science. Fourier phase retrieval poses fundamental theoretical and algorithmic challenges. In general, there is no unique mapping between a one-dimensional signal and its Fourier magnitude, and therefore the problem is ill-posed. Additionally, while almost all multidimensional signals are uniquely mapped to their Fourier magnitude, the performance of existing algorithms is generally not well-understood. In this chapter we survey methods to guarantee uniqueness in Fourier phase retrieval. We then present different algorithmic approaches to retrieve the signal in practice. We conclude by outlining some of the main open questions in this field.

T. Bendory (✉)
The Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ, USA
e-mail: tamir.bendory@princeton.edu

R. Beinert
Institute of Mathematics and Scientific Computing, University of Graz, Heinrichstraße 36, 8010 Graz, Austria

The Institute of Mathematics and Scientific Computing is a member of NAWI Graz (http://www.nawigraz.at/). The author is supported by the Austrian Science Fund (FWF) within the project P 28858.
e-mail: robert.beinert@uni-graz.at

Y.C. Eldar
The Andrew and Erna Viterbi Faculty of Electrical Engineering, Technion - Israel Institute of Technology, Haifa, Israel

The author is supported by the European Union's Horizon 2020 research and innovation program under grant agreement no. 646804-ERC-COG-BNYQ, and from the Israel Science Foundation under Grant no. 335/14.
e-mail: yonina@ee.technion.ac.il

H. Boche et al. (eds.), *Compressed Sensing and its Applications*,
Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-69802-1_2

Keywords Phase retrieval · Ptychography · Ultra-short pulse characterization · Uniqueness guarantees · Masked fourier phase retrieval · Semidefinite programming · Non-convex optimization · Alternating projections · Finitely supported and sparse signals

1 Introduction

The task of recovering a signal from its Fourier transform magnitude, called *Fourier phase retrieval*, arises in many areas in engineering and science. The problem has a rich history, tracing back to 1952 [112]. Important examples for Fourier phase retrieval naturally appear in many optical settings since optical sensors, such as a charge-coupled device (CCD) and the human eye, are insensitive to phase information of the light wave. A typical example is coherent diffraction imaging (CDI) which is used in a variety of imaging techniques [26, 35, 39, 94, 107, 111]. In CDI, an object is illuminated with a coherent electromagnetic wave, and the far-field intensity diffraction pattern is measured. This pattern is proportional to the object's Fourier transform, and therefore the measured data is proportional to its Fourier magnitude. Phase retrieval also played a key role in the development of the DNA double helix model [57]. This discovery awarded Watson, Crick, and Wilkins the Nobel Prize in Physiology or Medicine in 1962 [1]. Additional examples for applications in which Fourier phase retrieval appears are X-ray crystallography, speech recognition, blind channel estimation, astronomy, computational biology, alignment, and blind deconvolution [2, 11, 23, 56, 64, 102, 119, 126, 133, 138].

Fourier phase retrieval has been a long-standing problem since it raises difficult challenges. In general, there is no unique mapping between a one-dimensional signal and its Fourier magnitude, and therefore the problem is ill-posed. Additionally, while almost all multidimensional signals are uniquely mapped to their Fourier magnitude, the performance and stability of existing algorithms are generally not well-understood. In particular, it is not clear when given methods recover the true underlying signal. To simplify the mathematical analysis, in recent years attention has been devoted to a family of related problems, frequently called *generalized phase retrieval*. This refers to the setting in which the measurements are the phaseless inner products of the signal with known vectors. Particularly, the majority of works studied inner products with random vectors. Based on probabilistic considerations, a variety of convex and non-convex algorithms were suggested, equipped with stability guarantees from near-optimal number of measurements; see [4, 5, 34, 37, 46, 60, 125, 132, 134] to name a few works along these lines.

Here, we focus on the original Fourier phase retrieval problem and study it in detail. We begin by considering the ambiguities of Fourier phase retrieval [13, 16, 17]. We show that while in general a one-dimensional signal cannot be determined from its Fourier magnitude, there are several exceptional cases, such as minimum phase [68] and sparse signals [73, 103]. For general signals, one can guarantee uniqueness by taking multiple measurements, each one with a different mask. This setup is called masked phase retrieval [34, 72] and has several interesting

special cases, such as the short-time Fourier transform (STFT) phase retrieval [20, 47, 75] and vectorial phase retrieval [83, 104–106]. For all aforementioned setups, we present algorithms and discuss their properties. We also study the closely related frequency-resolved optical gating (FROG) methods [24, 25] and multidimensional Fourier phase retrieval [80].

The outline of this chapter is as follows. In Section 2 we formulate the Fourier phase retrieval problem. We also introduce several of its variants, such as masked Fourier phase retrieval and STFT phase retrieval. In Section 3 we discuss the fundamental problem of uniqueness, namely, conditions under which there is a unique mapping between a signal and its phaseless measurements. Section 4 is devoted to different algorithmic approaches to recover a signal from its phaseless measurements. Section 5 concludes the chapter and outlines some open questions. We hope that highlighting the gaps in the theory of phase retrieval will motivate more research on these issues.

2 Problem Formulation

In this section, we formulate the Fourier phase retrieval problem and introduce notation.

Let $x \in \mathbb{C}^N$ be the underlying signal we wish to recover. In Fourier phase retrieval, the measurements are given by

$$y[k] = \left| \sum_{n=0}^{N-1} x[n] e^{-2\pi jkn/\tilde{N}} \right|^2, \quad k = 0, \ldots, K-1. \tag{1}$$

Unless otherwise mentioned, we consider the over-sampled Fourier transform, i.e., $\tilde{N} = K = 2N - 1$, since in this case the acquired data is equivalent to the autocorrelation of x as explained in Section 3.1. We refer to this case as the *classical phase retrieval problem*. As will be discussed in the next section, in general the classical phase retrieval problem is ill-posed. Nevertheless, some special structures may impose uniqueness. Two important examples are sparse signals obeying a nonperiodic support [73, 103] and minimum phase signals [68]; see Section 3.2.

For general signals, a popular method to guarantee a unique mapping between the signal and its phaseless Fourier measurements is by utilizing several masks to introduce redundancy in the acquired data. In this case, the measurements are given by

$$y[m,k] = \left| \sum_{n=0}^{N-1} x[n] d_m[n] e^{-2\pi jkn/\tilde{N}} \right|^2, \quad k = 0, \ldots, K-1, \quad m = 0, \ldots, M-1, \tag{2}$$

where d_m are M known masks. In matrix notation, this model can be written as

$$y[m,k] = \left| f_k^* D_m x \right|^2, \quad k = 0, \ldots, K-1, \quad m = 0, \ldots, M-1, \tag{3}$$

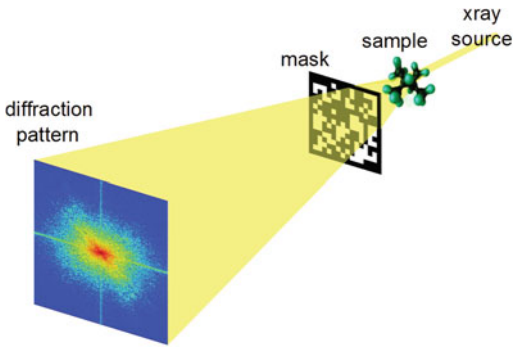


Fig. 1 An illustration of a typical masked phase retrieval setup (courtesy of [34]).

where f_k^* is the kth row of the DFT matrix $F \in \mathbb{C}^{K\times N}$ and $D_m \in \mathbb{C}^{N\times N}$ is a diagonal matrix that contains the entries of the mth mask. Classical phase retrieval is a special case in which $M = 1$ and $D_0 = I_N$ are the identity matrix.

There are several experimental techniques to generate masked Fourier measurements in optical setups [34]. One method is to insert a mask or a phase plate after the object [86]. Another possibility is to modulate the illuminating beam by an optical grating [87]. A third alternative is oblique illumination by illuminating beams hitting the object at specified angles [51]. An illustration of a masked phase retrieval setup is shown in Figure 1.

An interesting special case of masked phase retrieval is signal reconstruction from phaseless STFT measurements. Here, all masks are translations of a reference mask, i.e., $d_m[n] = d[mL-n]$, where L is a parameter that determines the overlapping factor between adjacent windows. Explicitly, the STFT phase retrieval problem takes on the form

$$y[m,k] = \left|\sum_{n=0}^{N-1} x[n]d[mL-n]e^{-2\pi jkn/\tilde{N}}\right|^2, \qquad (4)$$
$$k = 0,\dots,K-1, \quad m = 0,\dots,\lceil N/L\rceil - 1.$$

The reference mask d is referred to as *STFT window*. We denote the length of the STFT window by W, namely, $d[n] = 0$ for $n = W,\dots,N-1$ for some $W \le N$.

The problem of recovering a signal from its STFT magnitude arises in several applications in optics and speech processing. Particularly, it serves as the model of a popular variant of an ultrashort laser pulse measurement technique called frequency-resolved optical gating (FROG) which is introduced in Section 3.5 (the variant is referred to as X-FROG) [24, 25]. Another application is ptychography in which a moving probe (pinhole) is used to sense multiple diffraction measurements [89, 92, 108]. An illustration of a conventional ptychography setup is given in Figure 2. A closely related problem is Fourier ptychography [138].

The next section is devoted to the question of uniqueness, namely, under what conditions on the signal x and the masks d_m there exists a unique mapping between

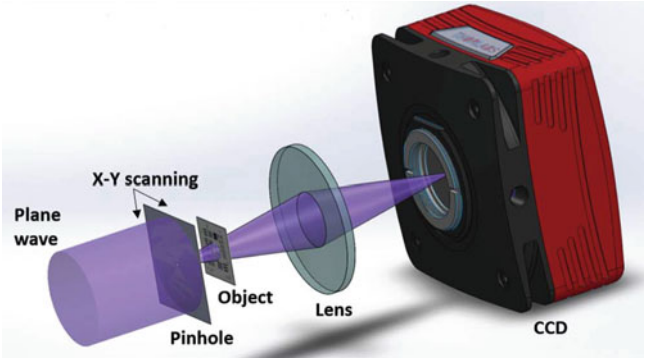


Fig. 2 An illustration of a conventional ptychography setup (courtesy of [120]).

the signal and the phaseless measurements. In Section 4 we survey different algorithmic approaches to recover the underlying signal x from the acquired data.

3 Uniqueness Guarantees

The aim of this section is to survey several approaches to ensure uniqueness of the discrete phase retrieval problem. We begin our study in Section 3.1 by considering the ambiguities arising in classical phase retrieval and provide a complete characterization of the solution set. Although the problem is highly ambiguous in general, uniqueness can be ensured if additional information about the signal is available. In Section 3.2, we first consider uniqueness guarantees based on the knowledge of the absolute value or phase of some signal entries. Next, we study sparse and minimum phase signals, which are uniquely defined by their Fourier magnitude and can be recovered by stable algorithms. In Sections 3.3 and 3.4, we show that for general signals, the ambiguities may be avoided by measuring the Fourier magnitudes of the interaction of the true signal with multiple deterministic masks or with several shifts of a fixed window. In Section 3.5 we study uniqueness guarantees for the closely related FROG methods. Finally, in Section 3.6 we survey the multidimensional phase retrieval problems and their properties that differ significantly from the one-dimensional setting.

3.1 Trivial and Non-Trivial Ambiguities

Considering the measurement model (1) of the classical phase retrieval problem, we immediately observe that the true signal $x \in \mathbb{C}^N$ cannot be recovered uniquely. For instance, the rotation (multiplication with a unimodular factor), the translation, or the conjugate reflection do not modify the Fourier magnitudes. Without further

a priori constraints, the unknown signal x is hence only determined up to these so-called *trivial ambiguities*, which are of minor interest. Besides the trivial ambiguities, the classical phase retrieval problem usually has a series of further non-trivial solutions, which can strongly differ from the true signal. For instance, the two non-trivially different signals

$$x_1 = (1, 0, -2, 0, -2)^T \qquad \text{and} \qquad x_2 = ((1-\sqrt{3}), 0, 1, 0, (1+\sqrt{3}))^T$$

yield the same Fourier magnitudes $y[k]$ in (1); see [119].

To characterize the occurring non-trivial ambiguities, one can exploit the close relation between the given Fourier magnitudes $y[k]$ with $k = 0, \ldots, 2N-2$ in (1) and the autocorrelation signal

$$a[n] = \sum_{m=0}^{N-1} \overline{x[m]}\, x[m+n], \qquad n = -N+1, \ldots, N-1,$$

with $x[n] = 0$ for $n < 0$ and $n \geq N$, [16, 29]. For this purpose, we consider the product of the polynomial $X(z) = \sum_{n=0}^{N-1} x[n]\, z^n$ and the reversed polynomial $\tilde{X}(z) = z^{N-1}\overline{X}(z^{-1})$, where $\overline{X}$ denotes the polynomial with conjugate coefficients. Note that $X(z^{-1})$ coincides with the usual z-transform of the signal $x \in \mathbb{C}^N$. Assuming that $x[0] \neq 0$ and $x[N-1] \neq 0$, we have

$$X(z)\,\tilde{X}(z) = z^{N-1} \sum_{n=0}^{N-1} x[n]\, z^n \cdot \sum_{m=0}^{N-1} \overline{x[m]}\; z^{-m} = \sum_{n=0}^{2N-2} a[n-N+1]\, z^n =: A(z),$$

where $A(z)$ is the autocorrelation polynomial of degree $2N-2$.

Since the Fourier magnitude (1) can be written as

$$y[k] = e^{2\pi jk(N-1)/\tilde{N}}\, X\big(e^{-2\pi jk/\tilde{N}}\big)\, \tilde{X}\big(e^{-2\pi jk/\tilde{N}}\big) = e^{2\pi jk(N-1)/\tilde{N}}\, A\big(e^{-2\pi jk/\tilde{N}}\big),$$

the autocorrelation polynomial $A(z)$ is completely determined by the $2N-1$ samples $y[k]$. The classical phase retrieval problem is thus equivalent to the recovery of $X(z)$ from

$$A(z) = X(z)\,\tilde{X}(z).$$

Comparing the roots of $X(z)$ and $\tilde{X}(z)$, we observe that the roots of the autocorrelation polynomial $A(z)$ occur in reflected pairs $(\gamma_j, \overline{\gamma}_j^{-1})$ with respect to the unit circle. The main problem in the recovery of $X(z)$ is now to decide whether γ_j or $\overline{\gamma}_j^{-1}$ is a root of $X(z)$. On the basis of this observation, all ambiguities—trivial and non-trivial—are characterized in the following way.

Theorem 1 ([16]). *Let $x \in \mathbb{C}^N$ be a complex-valued signal with $x[0] \neq 0$ and $x[N-1] \neq 0$ and with the Fourier magnitudes $y[k]$, $k = 0, \ldots, 2N-2$, in (1). Then*

the polynomial $X'(z) = \sum_{n=0}^{N-1} x'[n]\, z^n$ *of each signal* $x' \in \mathbb{C}^N$ *with* $y'[k] = y[k]$ *can be written as*

$$X'(z) = e^{j\alpha} \sqrt{|a[N-1] \prod_{i=1}^{N-1} |\beta_i|^{-1}} \cdot \prod_{i=1}^{N-1} (z - \beta_i),$$

where $\alpha \in [-\pi, \pi)$ *and where* β_i *is chosen from the reflected zero pairs* $(\gamma_i, \overline{\gamma}_i^{-1})$ *of the autocorrelation polynomial* $A(z)$*. Moreover, up to* 2^{N-2} *of these solutions may be non-trivially different.*

Since the support length N of the true signal x is directly encoded in the degree of the autocorrelation polynomial, all signals x' with $y'[k] = y[k]$ in Theorem 1 have the same length, and the trivial shift ambiguity does not occur. The multiplication by $e^{j\alpha}$ is related to the trivial rotation ambiguity. The trivial conjugate reflection ambiguity is also covered by Theorem 1, since this corresponds to the reflection of all zeros β_i at the unit circle and to an appropriate rotation of the whole signal. Hence, at least two of the 2^{N-1} possible zero sets $\{\beta_1, \ldots, \beta_{N-1}\}$ always correspond to the same non-trivial solution, which implies that the number of non-trivial solutions of the classical phase retrieval problem is bounded by 2^{N-2}.

The actual number of non-trivial ambiguities for a specific phase retrieval problem, however, strongly depends on the zeros of the true solution. If L denotes the number of zero pairs $(\gamma_\ell, \overline{\gamma}_\ell^{-1})$ of the autocorrelation polynomial $A(z)$ not lying on the unit circle, and m_ℓ the multiplicities of these zeros, then the different zero sets $\{\beta_1, \ldots, \beta_{N-1}\}$ in Theorem 1 can consist of s_ℓ roots γ_ℓ and $(m_\ell - s_\ell)$ roots $\overline{\gamma}_\ell^{-1}$, where s_ℓ is an integer between 0 and m_ℓ. Due to the trivial conjugation and reflection ambiguity, the corresponding phase retrieval problem has exactly

$$\left\lceil \frac{1}{2} \prod_{\ell=1}^{L} (m_\ell + 1) \right\rceil$$

non-trivial solutions [13, 53]. If, for instance, all zero pairs $(\gamma_\ell, \overline{\gamma}_\ell^{-1})$ are unimodular, then the problem is even uniquely solvable.

3.2 Ensuring Uniqueness in Classical Phase Retrieval

To overcome the non-trivial ambiguities, and to ensure uniqueness in the phase retrieval problem, one can rely on suitable a priori conditions or further information about the true signal. For instance, if the sought signal represents an intensity or a probability distribution, then it has to be real-valued and nonnegative. Unfortunately, this natural constraint does not guarantee uniqueness [14]. More appropriate priors like minimum phase or sparsity ensure uniqueness for almost every or, even, for every possible signal. Additional information about some entries of the true signal like the magnitude or the phase also guarantee uniqueness in certain settings.

3.2.1 Information About Some Entries of the True Signal

One approach to overcome the non-trivial ambiguities is to use additional information about some entries of the otherwise unknown signal x. For instance, in wave-front sensing and laser optics [114], besides the Fourier intensity, the absolute values $|x[0]|, \ldots, |x[N-1]|$ of the sought signal x are available. Interestingly, already one absolute value $|x[N-1-\ell]|$ within the support of the true signal x almost always ensures uniqueness.

Theorem 2 ([17]). *Let ℓ be an arbitrary integer between 0 and $N-1$. Then almost every complex-valued signal $x \in \mathbb{C}^N$ with support length N can be uniquely recovered from $y[k]$, $k = 0, \ldots, 2N-2$, in (1) and $|x[N-1-\ell]|$ up to rotations if $\ell \neq (N-1)/2$. In the case $\ell = (N-1)/2$, the reconstruction is almost surely unique up to rotations and conjugate reflections.*

The uniqueness guarantee in Theorem 2 cannot be improved by the knowledge of further or, even, all absolute values $|x[0]|, \ldots, |x[N-1]|$ of the true signal. More precisely, one can explicitly construct signals that are not uniquely defined by their Fourier magnitudes $y[k]$ and all temporal magnitudes $|x[n]|$ for every possible signal length [17]. In order to recover a signal from its Fourier magnitudes and all temporal magnitudes numerically, several multilevel Gauss-Newton methods have been proposed in [81, 82, 114]. Under certain conditions, the convergence of these algorithms to the true solution is guaranteed, and they allow signal reconstruction from noise-free as well as from noisy data.

The main idea behind Theorem 2 exploits $|x[N-1-\ell]|$ to show that the zero sets $\{\beta_1, \ldots, \beta_{N-1}\}$ of signals that cannot be recovered uniquely (up to trivial ambiguities) form an algebraic variety of lesser dimension. This approach can be transferred to further kinds of information about some entries of x. For instance, the knowledge of at least two phases of the true signal also guarantees uniqueness almost surely.

Theorem 3 ([17]). *Let ℓ_1 and ℓ_2 be different integers in $\{0, \ldots, N-1\}$. Then almost every complex-valued signal $x \in \mathbb{C}^N$ with support length N can be uniquely recovered from $y[k]$, $k = 0, \ldots, 2N-2$, in (1), $\arg x[N-1-\ell_1]$, and $\arg x[N-1-\ell_2]$ whenever $\ell_1 + \ell_2 \neq N-1$. In the case $\ell_1 + \ell_2 = N-1$, the recovery is only unique up to conjugate reflection except for $\ell_1 = 0$ and $\ell_2 = N-1$, where the set of non-trivial ambiguities is not reduced at all.*

As a consequence of Theorems 2 and 3, the classical phase retrieval problem is almost always uniquely solvable if at least one entry of the true signal x is known. Unfortunately, there is no algorithm that knows how to exploit the given entries to recover the complete signal in a stable and efficient manner.

Corollary 1. *Let ℓ be an arbitrary integer between 0 and $N-1$. Then almost every complex-valued signal $x \in \mathbb{C}^N$ with support length N can be uniquely recovered from $y[k]$, $k = 0, \ldots, 2N-2$, in (1) and $x[N-1-\ell]$ if $\ell \neq (N-1)/2$. In the case $\ell = (N-1)/2$, the reconstruction is almost surely unique up to conjugate reflection.*

Corollary 1 is a generalization of [137], where the recovery of real-valued signals $x \in \mathbb{R}^N$ from their Fourier magnitude $y[k]$ and one of their end points $x[0]$ or $x[N-1]$ is studied. In contrast to Theorems 2 and 3, the classical phase retrieval problem becomes unique if enough entries of the true signal are known beforehand.

Theorem 4 ([16, 96]). *Each complex-valued signal $x \in \mathbb{C}^N$ with signal length N is uniquely determined by $y[k]$, $k = 0, \dots, 2N-2$, in (1) and the $\lceil N/2 \rceil$ left end points $x[0], \dots, x[\lceil N/2 \rceil - 1]$.*

3.2.2 Sparse Signals

In the last section, the true signal x could be any arbitrary vector in $\mathbb{C}^N$. In the following, we consider the classical phase retrieval problem under the assumption that the unknown signal is sparse, namely, that only a small number of entries are non-zero. Sparse signals have been studied thoroughly in the last two decades; see, for instance, [32, 43, 45]. Phase retrieval problems of sparse signals arise in crystallography [78, 103] and astronomy [29, 103], for example. In many cases, the signal is sparse under an unknown transform. In the context of phase retrieval, a recent paper suggests a new technique to learn, directly from the phaseless data, the sparsifying transformation and the sparse representation of the signals simultaneously [127].

The union of all k-sparse signals in $\mathbb{C}^N$, which have at most k non-zero entries, is here denoted by $\mathscr{S}_k^N$. Since $\mathscr{S}_k^N$ with $k < N$ is a k-dimensional submanifold of $\mathbb{C}^N$ and hence itself a Lebesgue null set, Theorem 2 and Corollary 1 cannot be employed to guarantee uniqueness of the sparse phase retrieval problem. Further, if the k non-zero entries lie at equispaced positions within the true signal x, i.e., the support is of the form $\{n_0 + Lm : m = 0, \dots, k-1\}$ for some positive integers n_0 and L, this specific phase retrieval problem is equivalent to the recovery of a k-dimensional vector from its Fourier intensity [73]. Due to the non-trivial ambiguities, which are characterized by Theorem 1, the assumed sparsity cannot always avoid non-trivial ambiguities.

In general, the knowledge that the true signal is sparse has a beneficial effect on the uniqueness of phase retrieval. Under the restriction that the unknown signal x belongs to the class $\mathscr{T}_k^N$ of all k-sparse signals in $\mathbb{C}^N$ without equispaced support, which is again a k-dimensional submanifold, the uniqueness is ensured for almost all signals.

Theorem 5 ([73]). *Almost all signals $x \in \mathscr{T}_k^N$ can be uniquely recovered from their Fourier magnitudes $y[k]$, $k = 0, \dots, 2N-2$, in (1) up to rotations.*

Although Theorem 5 gives a theoretical uniqueness guarantee, it is generally a non-trivial task to decide whether a sparse signal is uniquely defined by its Fourier intensity. However, if the true signal does not possess any collisions, uniqueness is always given [103]. In this context, a sparse signal x has a *collision* if there exist

four indices i_1, i_2, i_3, i_4 within the support of x so that $i_1 - i_2 = i_3 - i_4$. A sparse signal without collisions is called *collision-free*. For instance, the signal

$$x = (0, 0, 1, 0, -2, 0, 1, 0, 0, 3, 0, 0)^T \in \mathbb{R}^{12}$$

is not collision-free since the index difference $6 - 4 = 2$ is equal to $4 - 2 = 2$.

Theorem 6 ([103]). *Assume that the signal $x \in \mathscr{S}_k^N$ with $k < N$ has no collisions.*

- *If $k \neq 6$, then x can be uniquely recovered from $y[k]$, $k = 0, \ldots, 2N-2$, in (1) up to trivial ambiguities;*
- *If $k = 6$ and not all non-zero entries $x[n]$ have the same value, then x can be uniquely recovered from $y[k]$, $k = 0, \ldots, 2N-2$, in (1) up to trivial ambiguities;*
- *If $k = 6$ and all non-zero entries $x[n]$ have the same value, then x can be uniquely recovered from $y[k]$, $k = 0, \ldots, 2N-2$, in (1) almost surely up to trivial ambiguities.*

The uniqueness guarantees in Theorem 6 remain valid for k-sparse continuous-time signals, which are composed of k pulses at arbitrary positions. More precisely, the continuous-time signal f is here given by $f(t) = \sum_{i=0}^{k-1} c_i \, \delta(t - t_i)$, where δ is the Dirac delta function, $c_i \in \mathbb{C}$ and $t_i \in \mathbb{R}$. In this setting, the uniqueness can be guaranteed by $\mathscr{O}(k^2)$ samples of the Fourier magnitude [18].

In Section 4.4, we discuss different algorithms to recover sparse signals $x \in \mathbb{C}^N$ that work well in practice.

3.2.3 Minimum Phase Signals

Based on the observation that each non-trivial solution of the classical phase retrieval problem is uniquely characterized by the zero set $\{\beta_1, \ldots, \beta_{N-1}\}$ in Theorem 1, one of the simplest ideas to enforce uniqueness is to restrict these zeros in an appropriate manner. Under the assumption that the true signal x is a minimum phase signal, which means that all zeros β_i chosen from the reflected zero pairs $(\gamma_i, \overline{\gamma}_i^{-1})$ of the autocorrelation polynomial $A(z)$ lie inside the unit circle, the corresponding phase retrieval problem is uniquely solvable [67, 68].

Although the minimum phase constraint guarantees uniqueness, the question arises on how to ensure that an unknown signal is minimum phase. Fortunately, each complex-valued signal x may be augmented to a minimum phase signal.

Theorem 7 ([68]). *For every $x \in \mathbb{C}^N$, the augmented signal*

$$x_{\min} = (\delta, x[0], \ldots, x[N-1])^T,$$

with $|\delta| \geq \|x\|_1$ is a minimum phase signal.

Consequently, if the Fourier intensity of the augmented signal $x_{\min}$ is available, then the true signal x can always be uniquely recovered up to trivial ambiguities.

Moreover, the minimum phase solution x can be computed (up to rotations) from the Fourier magnitude y as in (1) by a number of efficient algorithms [68]. Due to the trivial conjugate reflection ambiguity, this approach can be applied to maximum-phase signals whose zeros lie outside the unit circle.

The minimum phase solution of a given phase retrieval problem may be determined in a stable manner using an approach by Kolmogorov [68]. For simplicity, we restrict ourselves to the real case $x \in \mathbb{R}^N$ with $x[N-1] > 0$. The main idea is to determine the logarithm of the reversed polynomial $\tilde{X}(z) = z^{N-1} \sum_{n=0}^{N-1} \bar{x}[n]\, z^{-n}$ from the given data $y[k]$. Under the assumption that all roots of x strictly lie inside the unit circle, the analytic function $\log \tilde{X}(z)$ may be written as

$$\log \tilde{X}(z) = \sum_{n=0}^{\infty} \alpha_n z^n, \qquad (\alpha_n \in \mathbb{R})$$

where the unit circle $|z| = 1$ is contained in the region of convergence. Substituting $z = e^{-j\omega}$ with $\omega \in \mathbb{R}$, we have

$$\Re\big[\log \tilde{X}(e^{-j\omega})\big] = \sum_{n=0}^{\infty} \alpha_n \cos \omega n \quad \text{and} \quad \Im\big[\log \tilde{X}(e^{-j\omega})\big] = -\sum_{n=0}^{\infty} \alpha_n \sin \omega n,$$

where $\Re[\cdot]$ and $\Im[\cdot]$ denote the real and imaginary parts, respectively. Since the real and imaginary parts are a Hilbert transform pair, $\Im\big[\log \tilde{X}(e^{-j\omega})\big]$ is completely defined by $\Re\big[\log \tilde{X}(e^{-j\omega})\big]$. Because of the identity $|\tilde{X}(e^{-j\omega})|^2 = |A(e^{-j\omega})|$, the real part may be computed from the autocorrelation polynomial $A(z)$ by

$$\Re\big[\log \tilde{X}(e^{-j\omega})\big] = \tfrac{1}{2} \log |A(e^{-j\omega})|.$$

Finally, the autocorrelation polynomial $A(z)$ is completely determined by the Fourier magnitudes $y[k]$, $k = 0, \ldots, 2N-2$, leading to the recovery of the true minimum phase signal x. Based on this idea, one can construct numerical algorithms that guarantee stable signal recovery under the presence of noise [68].

3.3 Phase Retrieval with Deterministic Masks

A further possibility to obtain additional information about the underlying signal x is to measure its Fourier magnitude with respect to different masks as described in (2) and (3). Assuming that the masks are constructed randomly, one can show that the corresponding phase retrieval problem has a unique solution up to rotations almost surely or, at least, with high probability. Depending on the random model, the number of employed masks to recover a one-dimensional signal $x \in \mathbb{C}^N$ varies from $O(\log N)$ over $O((\log N)^2)$ to $O((\log N)^4)$; see [6, 63], and [33], respectively. Moreover, in the multidimensional case, two independent masks are sufficient to

guarantee uniqueness of almost every signal up to rotations [50]. As the following results show, in the deterministic setup, already a very small number of specifically constructed masks ensure uniqueness for most signals.

Theorem 8 ([72]). *Almost all complex-valued signals $x \in \mathbb{C}^N$ can be uniquely recovered from $y[m,k]$, $m = 1, 2$, and $k = 0, \dots, 2N-2$, as in (2) up to rotations if the two masks d_1, $d_2 \in \mathbb{C}^N$ satisfy*

- *$d_1[n] \neq 0$ or $d_2[n] \neq 0$ for each $0 \leq n \leq N-1$,*
- *$d_1[n]d_2[n] \neq 0$ for some $0 \leq n \leq N-1$.*

For some masks d_1 and d_2, one can overcome the "almost all" in Theorem 8 and obtain uniqueness of the corresponding phase retrieval problem.

Theorem 9 ([72]). *If the diagonal matrices D_1, D_2 correspond to the two masks*

$$d_1[n] = 1 \quad (0 \leq n \leq N-1) \qquad \text{and} \qquad d_2[n] = \begin{cases} 0 & n = 0 \\ 1 & 1 \leq n, \leq N-1, \end{cases} \tag{5}$$

then every complex-valued signal $x \in \mathbb{C}^N$ with $x[0] \neq 0$ can be uniquely recovered from $y[m,k]$, $m = 1, 2$ and $k = 0, \dots, 2N-2$, up to rotations.

A different approach to exploit deterministic masks in order to overcome the ambiguity in phase retrieval is discussed in [71] and can be proven by using the characterization in Theorem 1. More explicitly, here the two masks

$$d_1[n] = \begin{cases} 1, & 0 \leq n \leq L-1, \\ 0, & L \leq n \leq N-1, \end{cases} \qquad \text{and} \qquad d_2[n] = \begin{cases} 0, & 0 \leq n \leq L-1, \\ 1, & L \leq n \leq N-1, \end{cases} \tag{6}$$

for some L between 1 and $N-2$ are used. Pictorially, the mask d_1 blocks the right-hand side of the underlying signal x and d_2 the left-hand side.

For the signals x, D_1x, and D_2x, where D_i is the diagonal matrix with respect to the mask d_i, we define the polynomials X, X_1, and X_2 by

$$X(z) = \sum_{n=0}^{N-1} x[n]\, z^n, \quad X_1(z) = \sum_{n=0}^{L-1} x[n]\, z^n \quad \text{and} \quad X_2(z) = \sum_{n=1}^{N-L-1} x[n+L]\, z^n.$$

Different from the autocorrelation functions of D_1x and D_2x, which are simply given by $A_1(z) = X_1(z)\tilde{X}_1(z)$ and $A_2(z) = X_2(z)\tilde{X}_2(z)$, the autocorrelation function $A(z)$ of the true signal x can be written as

$$\begin{aligned} A(z) &= \big(X_1(z) + z^L X_2(z)\big)\big(z^{N-L-1}\tilde{X}_1(z) + \tilde{X}_2(z)\big) \\ &= z^{N-L-1}A_1(z) + X_1(z)\tilde{X}_2(z) + z^{N-1}\tilde{X}_1(z)X_2(z) + z^L A_2(z), \end{aligned}$$

since $X(z) = X_1(z) + z^L X_2(z)$. Due to the fact that $X_1(z)\tilde{X}_2(z)$ and $z^{N-1}\tilde{X}_1(z)X_2(z)$ have no common monomials with the same degree, one can determine the polynomials

$$X_1(z)\tilde{X}_1(z), \quad X_1(z)\tilde{X}_2(z), \quad \tilde{X}_1(z)X_2(z), \quad \text{and} \quad X_2(z)\tilde{X}_2(z), \tag{7}$$

from the autocorrelation functions $A(z)$, $A_1(z)$, and $A_2(z)$.

As mentioned before, the reversed polynomials $\tilde{X}_i(z)$ correspond to the reflected zero set of $X_i(z)$ with respect to the unit circle. Hence, assuming that the zeros of $D_1 x$ and $D_2 x$ are pairwise different, one can determine both zero sets by comparing the roots of the four polynomials (7), which yields the following result.

Theorem 10 ([71]). *Let $x \in \mathbb{C}^N$, and assume that the zeros ξ_i and η_ℓ of*

$$X_1(z) = x[L-1]\prod_{i=1}^{L-1}(z-\xi_i), \quad \textit{and} \quad X_2(z) = x[N-1]\prod_{\ell=1}^{N-L-1}(z-\eta_\ell),$$

are pairwise different. Then the signal x can be uniquely recovered up to rotations from the Fourier magnitudes $y[m,k]$, $m = 0,1,2$ and $k = 0,\dots,2N-2$, with the masks $d_0 \equiv 1$ and d_1, d_2 in (6).

The phase retrieval problem in Theorem 10 is equivalent to the recovery of $x_1 = D_1 x$ and $x_2 = D_2 x$ with support $\{0,\dots,L-1\}$ and $\{L,\dots,N-1\}$ from the Fourier magnitudes of x_1, x_2, and $x_1 + x_2$. More generally, the recovery of two arbitrary signals $x_1, x_2 \in \mathbb{C}^N$ from their Fourier magnitudes and the Fourier magnitude of the interference $x_1 + x_2$ has been studied in [16, 79]. Theorem 10 is a specific instance of the uniqueness guarantee given in [16]. Furthermore, these problems are closely related to the vectorial phase retrieval problem introduced in [84, 104, 105], where the Fourier magnitudes of a second interference $x_1 + jx_2$ are employed.

A further example for phase retrieval with deterministic masks is considered in [34], where the three masks are defined by

$$d_0[n] = 1, \qquad d_1[n] = 1 + e^{2\pi j s n/N}, \quad \text{and} \quad d_2[n] = 1 + e^{2\pi j(sn/N - 1/4)}, \tag{8}$$

for a nonnegative integer s. The masks d_1 and d_2 here interfere the unknown signal x with a modulated version of the unknown signal itself, which yields the Fourier magnitudes $|\hat{x}[k] + \hat{x}[k-s]|^2$ and $|\hat{x}[k] - j\,\hat{x}[k-s]|^2$. Together with the Fourier magnitudes $|\hat{x}[k]|^2$, for almost every signal, the relative phases $\phi[k-s] - \phi[k]$ of the Fourier transform $\hat{x}[k] = |x[k]|\, e^{j\phi[k]}$ can be determined. If s is relatively prime with N, then the Fourier transform $\hat{x}$ and thus the true signal x are recovered up to rotations.

Theorem 11 ([34]). *Let $x \in \mathbb{C}^N$ be a signal with non-vanishing DFT. Then x is uniquely recovered from $y[m,k]$ with $K = \tilde{N} = N$ and the masks in (8) up to rotations if and only if the nonnegative integer s is relatively prime with N.*

The masks in (8) as well as the uniqueness guarantee in Theorem 11 can be generalized to multidimensional phase retrieval [34]. If $\tilde{N}$ is replaced by $2N-1$, every signal $x \in \mathbb{C}^N$ is uniquely recovered up to rotation from its Fourier magnitudes $y[m,k]$ in (2) with masks $d_0[n] = 1$ and $d_i[n] = 1 + e^{j\alpha_i} e^{2\pi jsn/N}$, $i = 1, 2$, where $\alpha_i \in [-\pi, \pi)$ and where s can be nearly every real number [15]. Several further examples of deterministic masks which allow a unique recovery are detailed in [15, 34, 71, 72] and references therein. In Section 4.2, we consider semidefinite relaxation algorithms which stably recover the unknown signal from its masked Fourier magnitudes (2) under noise.

3.4 Phase Retrieval from STFT Measurements

We next consider uniqueness guarantees for the recovery of an unknown signal from the magnitude of its STFT as defined in (4). This problem can be interpreted as a sequence of classical phase retrieval problems, where some entries of the underlying signals have to coincide. Obviously, the STFT phase retrieval problem cannot be solved uniquely if the parameter L is greater than or equal to the window length W, since the classical problems are then independent from each other.

Under the assumption that the known window d does not vanish, i.e., $d[n] \neq 0$ for $n = 0, \ldots, W-1$, some of the first uniqueness guarantees were established in [96].

Theorem 12 ([96]). *Let d be a non-vanishing window of length $W > 2$, and let L be an integer in $\{1, \ldots, \lfloor W/2 \rfloor\}$. If the signal $x \in \mathbb{C}^N$ with support length N has at most $W - 2L$ consecutive zeros between any two non-zero entries, and if the first L entries of x are known, then x can be uniquely recovered from $y[m,k]$ with $K = 2W - 1$ in (4).*

The main idea behind Theorem 12 is that the corresponding classical phase retrieval problems are solved sequentially. For instance, the case $m = 1$ is equivalent to recovering a signal in $\mathbb{C}^{L+1}$ from its Fourier intensity and the first L entries. The uniqueness of this phase retrieval problem is guaranteed by Theorem 4. Since the true signal x has at most $W - 2L$ consecutive zeros, the remaining subproblems can also be reduced to the setting considered in Theorem 4.

Knowledge of the first L entries of x in Theorem 12 is a strong restriction in practice. Under the a priori constraint that the unknown signal is non-vanishing everywhere, the first L entries are not needed to ensure uniqueness.

Theorem 13 ([75]). *Let d be a non-vanishing window of length W satisfying $L < W \leq N/2$. Then almost all non-vanishing signals can be uniquely recovered up to rotations from their STFT magnitudes $y[m,k]$ in (4) with $2W \leq K \leq N$ and $\tilde{N} = N$.*

For some classes of STFT windows, the uniqueness is guaranteed for all non-vanishing signals [20, 47]. Both references use a slightly different definition of

the STFT, where the STFT window in (4) is periodically extended over the support $\{0, \dots, N-1\}$, i.e., the indices of the window d are considered as modulo the signal length N.

Theorem 14 ([47]). *Let d be a periodic window with support length $W \geq 2$ and $2W - 1 \leq N$, and assume that the length-N DFT of $|d[n]|^2$ is non-vanishing. If N and $W - 1$ are co-prime, then every non-vanishing signal $x \in \mathbb{C}^N$ can be uniquely recovered from its STFT magnitudes $y[m, k]$ in (4) with $L = 1$ and $K = \tilde{N} = N$ up to rotations.*

Theorem 15 ([20]). *Let d be a periodic window of length W, and assume that the length-N DFT of $|d[n]|^2$ and $d[n]d[n-1]$ are non-vanishing. Then every non-vanishing signal $x \in \mathbb{C}^N$ can be uniquely recovered from its STFT magnitudes $y[m, k]$ in (4) with $L = 1$ and $K = \tilde{N} = N$ up to rotations.*

If we abandon the constraint that the underlying signal is non-vanishing, then the behavior of the STFT phase retrieval problem changes dramatically, and the recovery of the unknown signal becomes much more challenging. For example, if the unknown signal x possesses more than $W - 1$ consecutive zero entries, then the signal can be divided in two parts, whose STFTs are completely independent. An explicit non-trivial ambiguity for this specific setting is constructed in [47]. Depending on the window length, there are thus some natural limitations on how far uniqueness can be ensured for sparse signals.

Theorem 16 ([75]). *Let d be a non-vanishing window of length W satisfying $L < W \leq N/2$. Then almost all sparse signals with less than $\min\{W - L, L\}$ consecutive zeros can be uniquely recovered up to rotations from their STFT magnitudes $y[m, k]$ in (4) with $2W \leq K \leq N$ and $\tilde{N} = N$.*

In [27], the STFT is interpreted as measurements with respect to a Gabor frame. Under certain conditions on the generator of the frame, every signal $x \in \mathbb{C}^N$ is uniquely recovered up to rotations. Further, the true signal x is given as a closed-form solution. For the STFT model in (4), this implies the following uniqueness guarantee.

Theorem 17 ([27]). *Let d be a periodic window of length W, and assume that the length-N DFT of $d[n]d[n-m]$ is non-vanishing for $m = 0, \dots, N-1$. Then every signal $x \in \mathbb{C}$ can be uniquely recovered from its STFT magnitudes $y[m, k]$ in (4) with $L = 1$ and $K = \tilde{N} = N$ up to rotations.*

The main difference between Theorem 17 and the uniqueness results before is that the unknown signal $x \in \mathbb{C}^N$ can have arbitrarily many consecutive zeros. On the other hand, the STFT window must have a length of at least $N/2$ in order to ensure that $d[n]d[n-m]$ is not the zero vector. Thus, the theorem is only relevant for long windows. A similar result was derived in [20], followed by a stable recovery algorithm; see Section 4.3.

3.5 FROG Methods

An important optical application for phase retrieval is ultrashort laser pulse characterization [128, 130]. One way to overcome the non-uniqueness of Fourier phase retrieval in this application is by employing a measurement technique called X-FROG (see also Section 2). In X-FROG, a reference window is used to gate the sought signal, resulting in the STFT phase retrieval model (4). However, in practice it is quite hard to generate and measure such a reference window. Therefore, in order to generate redundancy in ultrashort laser pulse measurements, it is common to correlate the signal with itself. This method is called frequency-resolved optical gating (FROG).

FROG is probably the most commonly used approach for full characterization of ultrashort optical pulses due to its simplicity and good experimental performance. A FROG apparatus produces a 2D intensity diagram of an input pulse by interacting the pulse with delayed versions of itself in a nonlinear-optical medium, usually using a second harmonic generation (SHG) crystal [42]. This 2D signal is called a FROG trace and is a quartic function of the unknown signal. An illustration of the FROG setup is presented in Figure 3. Here we focus on SHG FROG, but other types of nonlinearities exist for FROG measurements. A generalization of FROG, in which two different unknown pulses gate each other in a nonlinear medium, is called blind FROG. This method can be used to characterize simultaneously two signals [136]. In this case, the measured data is referred to as a blind FROG trace and is quadratic in both signals. We refer to the problems of recovering a signal from its blind FROG trace and FROG trace as *bivariate phase retrieval* and *quartic phase retrieval*, respectively.

In bivariate phase retrieval, we acquire, for each delay step m, the power spectrum of

$$x_m[n] = x_1[n]\, x_2[n + mL],$$

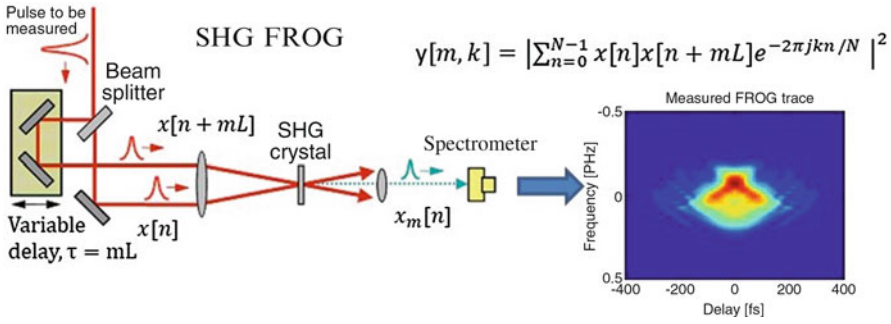


Fig. 3 Illustration of the SHG FROG technique (courtesy of [24]).

where L, as in the STFT phase retrieval setup, determines the overlap factor between adjacent sections. The acquired data is given by

$$\begin{aligned} y[m,k] &= \left| \sum_{n=0}^{N-1} x_m[n]\, e^{-2\pi jkn/N} \right|^2 \\ &= \left| \sum_{n=0}^{N-1} x_1[n]\, x_2[n+mL] e^{-2\pi jkn/N} \right|^2 . \end{aligned} \tag{9}$$

Quartic phase retrieval is the special case in which $x_1 = x_2$.

The trivial ambiguities of bivariate phase retrieval are described in the following proposition.

Proposition 1 ([24]). *Let $x_1, x_2 \in \mathbb{C}^N$, and let $x_m[n] := x_1[n]x_2[n+mL]$ for some fixed L. Then, the following signals have the same phaseless bivariate measurements $y[m,k]$ as the pair x_1, x_2:*

1. *multiplication by global phases $x_1 e^{j\psi_1}, x_2 e^{j\psi_2}$ for some $\psi_1, \psi_2 \in \mathbb{R}$,*
2. *the shifted signal $\tilde{x}_m[n] = x_m[n-n_0]$ for some $n_0 \in \mathbb{Z}$,*
3. *the conjugated and reflected signal $\grave{x}_m[n] = \overline{x_m[-n]}$,*
4. *modulation, $x_1[n]e^{-2\pi jk_0 n/N}$, $x_2[n]e^{2\pi jk_0 n/N}$ for some $k_0 \in \mathbb{Z}$.*

The fundamental question of uniqueness for FROG methods has been analyzed first in [113] for the continuous setup. The analysis of the discrete setup appears in [24].

Theorem 18 ([24]). *Let $L = 1$, and let $\hat{x}_1$ and $\hat{x}_2$ be the Fourier transforms of x_1 and x_2, respectively. Assume that $\hat{x}_1$ has at least $\lceil (N-1)/2 \rceil$ consecutive zeros (e.g., band-limited signal). Then, almost all signals are determined uniquely, up to trivial ambiguities, from the measurements $y[m,k]$ in (9) and the knowledge of $|\hat{x}_1|$ and $|\hat{x}_2|$. By trivial ambiguities we mean that x_1 and x_2 are determined up to global phase, time shift, and conjugate reflection.*

For the FROG setup, i.e., $x_1 = x_2$, this result has been recently extended for $L > 1$; see [25].

Several heuristic techniques have been proposed to estimate an underlying signal from its FROG trace. These algorithms are based on a variety of methods, such as alternating projections, gradient descent, and iterative PCA [77, 122, 129].

3.6 Multidimensional Phase Retrieval

In a wide range of real-world applications like crystallography or electron microscopy, the natural objects of interest correspond to two- or three-dimensional signals. More generally, the r-dimensional phase retrieval problem consists of the recovery of an unknown r-dimensional signal $x \in \mathbb{C}^{N_1 \times \cdots \times N_r}$ from its Fourier magnitudes

$$y[k] = \left| \sum_{n \in \mathbb{Z}_N} x[n] e^{-(2\pi)^r j k \cdot n / \tilde{N}_1 \dots \tilde{N}_r} \right|^2, \tag{10}$$

$$k \in \{0, \dots, K_1 - 1\} \times \cdots \times \{0, \dots, K_r - 1\},$$

with $n = (n_1, \dots, n_r)^T$ and $\mathbb{Z}_N = \{0, \dots, N_1 - 1\} \times \cdots \times \{0, \dots, N_r - 1\}$. Unless otherwise mentioned, we assume $\tilde{N}_i = K_i = 2N_i - 1$.

Clearly, rotations, transitions, or conjugate reflections of the true signal lead to trivial ambiguities. Besides these similarities, the ambiguities of the multidimensional phase retrieval problem are very different from those of its one-dimensional counterpart. More precisely, non-trivial ambiguities occur only in very rare cases, and almost every signal is uniquely defined by its Fourier magnitude up to trivial ambiguities.

Similarly to Section 3.1, the non-trivial ambiguities can be characterized by exploiting the autocorrelation. Here the related polynomial

$$X(z) = \sum_{n \in \mathbb{Z}_n} x[n] \, z^n = \prod_{i=1}^{I} X_i(z),$$

with $z^n = z^{n_1} \cdots z^{n_r}$ is uniquely factorized (up to multiplicative constants) into irreducible factors $X_i(z)$, which means that the X_i cannot be represented as a product of multivariate polynomials of lesser degree. The main difference with the one-dimensional setup is that most multivariate polynomials consist of only one irreducible factor X_i. Denoting the multivariate reversed polynomial by

$$\tilde{X}_i(z) = z^M \, \overline{X}_i(z^{-1}),$$

with $z^M = z^{M_1} \cdots z^{M_r}$, where M_ℓ is the degree of X_i with respect to the variable z_ℓ, the non-trivial ambiguities in the multidimensional setting are characterized as follows.

Theorem 19 ([65]). *Let $x \in \mathbb{C}^{N_1 \times \cdots \times N_r}$ be the complex-valued signal related to the polynomial $X(z) = \prod_{i=1}^{I} X_i(z)$, where $X_i(z)$ are non-trivial irreducible polynomials. Then the polynomial $X'(z) = \sum_{n \in \mathbb{Z}_N} x'[n] \, z^n$ of each signal $x' \in \mathbb{C}^{N_1 \times \cdots \times N_r}$ with Fourier magnitudes $y'[k] = y[k]$ in (10) can be written as*

$$X'(z) = \prod_{i \in J} X_i(z) \cdot \prod_{i \notin J} \tilde{X}_i(z),$$

for some index set $J \subset \{1, \dots, I\}$.

Thus, the phase retrieval problem is uniquely solvable up to trivial ambiguities if the algebraic polynomial $X(z)$ of the true signal x is irreducible or if all but one factor $X_i(z)$ are invariant under reversion [65]. In contrast to the one-dimensional

case, where the polynomial $X(z)$ may always be factorized into linear factors with respect to the zeros β_i, cf. Theorem 1, most multivariate polynomials cannot be factorized as mentioned above.

Theorem 20 ([66]). *The subset of the r-variate polynomials $X(z_1, \ldots, z_r)$ with $r > 1$ of degree $M_\ell > 1$ in z_ℓ which are reducible over the complex numbers corresponds to a set of measure zero.*

Consequently, the multidimensional phase retrieval problem has a completely different behavior than its one-dimensional counterpart.

Corollary 2. *Almost every r-dimensional signal with $r > 1$ is uniquely defined by its Fourier magnitudes $y[k]$ in (10) up to trivial ambiguities.*

Investigating the close connection between the one-dimensional and two-dimensional problem formulations, the different uniqueness properties have been studied in [80]. Particularly, one can show that the two-dimensional phase retrieval problem corresponds to a one-dimensional problem with additional constraints, which almost always guarantee uniqueness. Despite these uniqueness guarantees, there are no systematic methods to estimate an r-dimensional signal from its Fourier magnitude [9, 80]. The most popular techniques are based on alternating projection algorithms as discussed in Section 4.1.

4 Phase Retrieval Algorithms

The previous section presented conditions under which there exists a unique mapping between a signal and its Fourier magnitude (up to trivial ambiguities). Yet, the existence of a unique mapping does not imply that we can actually estimate the signal in a stable fashion. The goal of this section is to present different algorithmic approaches for the inverse problem of recovering a signal from its phaseless Fourier measurements. In the absence of noise, this task can be formulated as a feasibility problem over a non-convex set

$$\begin{aligned} \text{find}_{z\in\mathbb{C}^N} \quad \text{subject to} \quad & y[m,k] = \left|f_k^* D_m z\right|^2, \\ & k = 0, \ldots, K-1, \quad m = 0, \ldots, M-1. \end{aligned} \tag{11}$$

Recall that (11) covers the classical and STFT phase retrieval problems as special cases.

From the algorithmic point of view, it is often more convenient to formulate the problem as a minimization problem. Two common approaches are to minimize the intensity-based loss function

$$\min_{z\in\mathbb{C}^N} \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} \left(y[m,k] - \left|f_k^* D_m z\right|^2 \right)^2, \tag{12}$$

or the amplitude-based loss function (see, for instance, [54, 131, 134])

$$\min_{z\in\mathbb{C}^N} \sum_{k=0}^{K-1}\sum_{m=0}^{M-1} \left(\sqrt{y[m,k]} - \left|f_k^* D_m z\right|\right)^2. \tag{13}$$

The chief difficulty arises from the non-convexity of these loss functions. For example, if x is a real signal, then (12) is a sum of MK quartic polynomials. Hence, there is no reason to believe that a gradient algorithm will converge to the target signal from an arbitrary initialization. To demonstrate this behavior, we consider an STFT phase retrieval setup for which a unique solution is guaranteed (see Theorem 15). We attempt to recover the signal by employing two methods: a gradient descent algorithm that minimizes (12) and the classical Griffin-Lim algorithm (see Section 4.1 and Algorithm 2). Both techniques were initialized from 100 different random vectors. As can be seen in Figure 4, even for long windows, the algorithms do not always converge to the global minimum. Furthermore, the success rate decreases with the window's length. In what follows, we present different systematic approaches to recover a signal from its phaseless Fourier measurements and discuss their advantages and shortcomings.

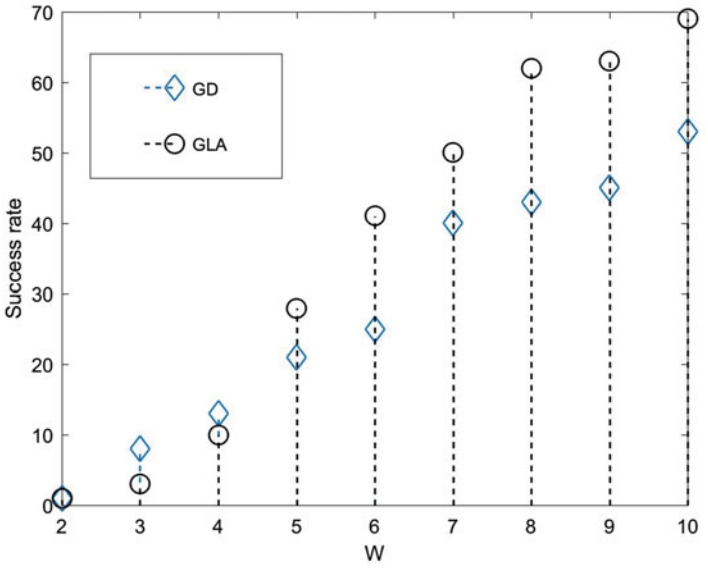


Fig. 4 This figure examines the empirical success rate of a gradient algorithm (GD) that minimizes (4) and the Griffin-Lim algorithm (GLA) as presented in Algorithm 2 for the STFT phase retrieval problem with a rectangular window. For each value of W, 100 experiments were conducted with $N = 23$ and $L = 1$ in a noise-free environment. Note that a unique solution is guaranteed according to Theorem 15. The underlying signals and the initializations were drawn from an i.i.d. normal distribution. A success was declared when the objective function is below 10^{-4}.

The rest of this section is organized as follows. We begin in Section 4.1 by introducing the classical algorithms which are based on alternating projections. Then, we proceed in Section 4.2 with convex programs based on semidefinite programming (SDP) relaxations. SDPs have gained popularity in the recent years as they provide good numerical performance and theoretical guarantees. We present SDP-based algorithms for masked phase retrieval, STFT phase retrieval, and minimum phase and sparse signals. In Section 4.3, we survey additional non-convex algorithms with special focus on STFT phase retrieval. Section 4.4 presents several algorithms specialized for the case of phase retrieval of sparse signals.

4.1 Alternating Projection Algorithms

In their seminal work [58], Gerchberg and Saxton considered the problem of recovering a signal from its Fourier and temporal magnitude. They proposed an intuitive solution which iterates between two basic steps. The algorithm begins with an arbitrary initial guess. Then, at each iteration, it imposes the known Fourier magnitude $|\hat{x}|$ and temporal magnitude $|x|$ consecutively. This process proceeds until a stopping criterion is attained.

The basic concept of the Gerchberg-Saxton algorithm was extended by Fienup in 1982 to a variety of phase retrieval settings [54, 55]. Fienup suggested to replace the temporal magnitude constraint by other alternative constraints in the time domain. Examples for such constraints are the knowledge of the signal's support or few entries of the signal, nonnegativity, or a known subspace in which the signal lies. Recently, it was also suggested to incorporate sparsity priors [95]. These algorithms have the desired property of *error reduction*. Let $\hat{x}^{(\ell)}$ be the Fourier transform of the estimation in the ℓth iteration. Then, it can be shown that the quantity $E_\ell := \sum_k ||\hat{x}[k]| - |\hat{x}^{(\ell)}[k]||^2$ is monotonically non-increasing. This class of methods is best understood as *alternating projection* algorithms [48, 91, 99]. Namely, each iteration consists of two consecutive projections onto sets defined by the spectral and temporal constraints. As the first step projects onto a non-convex set (and in some cases, the temporal projection is non-convex as well), the iterations may not converge to the target signal. The method is summarized in Algorithm 1, where we use the definition

$$\operatorname{sign}(z[n]) := \begin{cases} \frac{z[n]}{|z[n]|}, & z[n] \neq 0, \\ 0, & z[n] \neq 0. \end{cases}$$

Over the years, many variants of the basic alternating projection scheme have been suggested. A popular algorithm used for CDI applications is the *hybrid input-output* (HIO), which consists of an additional correction step in the time domain [54]. Specifically, the last stage of each iteration is of the form

Algorithm 1 General scheme of alternating projection algorithms

Input: Spectral magnitude $|\hat{x}|$ and additional temporal constraint on x
Output: x_{est} - estimation of x
Initialization: random input vector $x^{(0)}$, $\ell = 0$
while halting criterion false **do**:

- $\ell \leftarrow \ell + 1$
- Compute the Fourier transform of current estimation $\hat{x}^{(\ell)}$
- Keep phase, update spectral magnitude $\hat{z}^{(\ell)} = |\hat{x}| \operatorname{sign}(\hat{x}^{(\ell)})$
- Compute $z^{(\ell)}$, the inverse Fourier transform of $\hat{z}^{(\ell)}$
- Impose temporal constraints on $z^{(\ell)}$ to obtain $x^{(\ell)}$

end while
Return: $x_{est} \leftarrow x^{(\ell)}$

$$x^{(\ell)}[n] = \begin{cases} z^{(\ell)}[n], & n \notin \gamma, \\ x^{(\ell-1)}[n] - \beta z^{(\ell)}[n], & n \in \gamma, \end{cases}$$

where γ is the set of indices for which $z^{(\ell)}$ violates the temporal constraint (e.g., support constraint, nonnegativity) and β is a small parameter. While there is no proof that the HIO converges, it tends to avoid local minima in the absence of noise. Additionally, it is known to be sensitive to the prior knowledge accuracy [119]. For additional related algorithms, we refer the interested reader to [10, 38, 49, 88, 93, 109].

Griffin and Lim proposed a modification of Algorithm 1 specialized for STFT phase retrieval [62]. In this approach, the last step at each iteration harnesses the knowledge of the STFT window to update the signal estimation. The Griffin-Lim heuristic is summarized in Algorithm 2.

Algorithm 2 Griffin-Lim algorithm

Input: STFT magnitude $|\hat{x}_d[m,k]|$
Output: x_{est} - estimation of x
Initialization: random input vector $x^{(0)}$, $\ell = 0$
while halting criterion false **do**:

- $\ell \leftarrow \ell + 1$
- Compute the STFT of current estimation $\hat{x}_d^{(\ell)}$
- Keep phase, update STFT magnitudes $\hat{z}^{(\ell)} = |\hat{x}_d| \operatorname{sign}(\hat{x}_d^{(\ell)})$
- For each fixed m, compute $z_m^{(\ell)}$, the inverse Fourier transform of $\hat{z}^{(\ell)}$
- Update signal estimate $x^{(\ell)}[n] = \frac{\sum_m z_m^{(\ell)}[n]\, d[mL-n]}{\sum_m |d[mL-n]|^2}$

end while
Return: $x_{est} \leftarrow x^{(\ell)}$

4.2 Semidefinite Relaxation Algorithms

In recent years, algorithms based on convex relaxation techniques have attracted considerable attention [34, 132]. These methods are based on the insight that while the feasibility problem (11) is quadratic with respect to x, it is linear in the matrix xx^*. This leads to a natural convex SDP relaxation that can be solved in polynomial time using standard solvers like CVX [61]. In many cases, these relaxations achieve excellent numerical performance followed by theoretical guarantees. However, the SDP relaxation optimizes over N^2 variables, and therefore its computational complexity is quite high.

SDP relaxation techniques begin by reformulating the measurement model (3) as a linear function of the Hermitian rank-one matrix $X := xx^*$:

$$y[m,k] = (f_k^* D_m x)^* (f_k^* D_m x) = x^* D_m^* f_k f_k^* D_m x = \operatorname{trace}(D_m^* f_k f_k^* D_m X).$$

Consequently, the problem of recovering x from y can be posed as the feasibility problem of finding a rank-one Hermitian matrix which is consistent with the measurements:

$$\begin{aligned} \text{find} \quad X \in \mathscr{H}^N \quad \text{subject to} \quad & X \succeq 0, \quad \operatorname{rank}(X) = 1, \\ & y[m,k] = \operatorname{trace}(D_m^* f_k f_k^* D_m X), \\ & k = 0, \ldots, K-1, \quad m = 0, \ldots, M-1, \end{aligned} \tag{14}$$

where $\mathscr{H}^N$ is the set of all $N \times N$ Hermitian matrices. If there exists a matrix X satisfying all the constraints of (14), then it determines x up to global phase. The feasibility problem (14) is non-convex due to the rank constraint. A convex relaxation may be obtained by omitting the rank constraint leading to the SDP [34, 59, 115, 132]:

$$\begin{aligned} \text{find} \quad X \in \mathscr{H}^N \quad \text{subject to} \quad & X \succeq 0, \\ & y[m,k] = \operatorname{trace}(D_m^* f_k f_k^* D_m X), \\ & k = 0, \ldots, K-1, \quad m = 0, \ldots, M-1. \end{aligned} \tag{15}$$

If the solution of (15) happens to be of rank one, then it determines x up to global phase. In practice, it is useful to promote a low-rank solution by minimizing an objective function over the constraints of (15). A typical choice is the trace function, which is the convex hull of the rank function for Hermitian matrices. The resulting SDP relaxation algorithm is summarized in Algorithm 3.

The SDP relaxation for the classical phase retrieval problem (i.e., $M = 1$ and $D_0 = I_N$) was investigated in [68]. It was shown that SDP relaxation achieves the optimal cost function value of (12). However, recall that in general the classical phase retrieval problem does not admit a unique solution. Minimum phase signals are an exception as explained in Section 3.2.3. Let a be the autocorrelation sequence

Algorithm 3 SDP relaxation for phase retrieval with masks

Input: Fourier magnitudes $y[m,k]$ as given in (2) and the masks D_m, $m = 0,\dots,M-1$
Output: x_{est} - estimation of x
Solve:

$$\min_{x\in\mathscr{H}^N} \operatorname{trace}(X) \quad \text{subject to} \quad \begin{aligned} & X \succeq 0, \\ & y[m,k] = \operatorname{trace}(D_m^* f_k f_k^* D_m X), \\ & k = 0,\dots,K-1, \quad m = 0,\dots,M-1. \end{aligned}$$

Return : x_{est} - the best rank-one approximation of the SDP's solution.

of the estimated signal from Algorithm 3. If x is the minimum phase, then it can be recovered by the following program:

$$\max_{X\in\mathscr{H}^N} \quad X[0,0] \quad \text{subject to} \quad \begin{aligned} & X \succeq 0, \quad \operatorname{trace}(\Theta_k X) = a[k], \\ & k = 0,\dots,N-1, \end{aligned} \tag{16}$$

where Θ_k is a Toeplitz matrix with ones in the kth diagonal and zero otherwise. The solution of (16), X_{MP}, is guaranteed to be rank one so that $X_{MP} = xx^*$. See Section 3.2.3 for a different algorithm to recover minimum phase signals.

An SDP relaxation for deterministic masks was investigated in [72], where the authors consider two types of masks. Here, we consider the two masks, d_1 and d_2, given in (5). Let D_1 and D_2 be the diagonal matrices associated with d_1 and d_2, respectively, and assume that each measurement is contaminated by bounded noise ε. Then, it was suggested to estimate the signal by solving the following convex program

$$\min_{X\in\mathscr{H}^N} \quad \operatorname{trace}(X) \quad \text{subject to} \quad \begin{aligned} & X \succeq 0, \\ & |y[m,k] - \operatorname{trace}(D_m^* f_k f_k^* D_m X)| \le \varepsilon, \\ & k = 0,\dots,2N-1, \quad m = 0,1. \end{aligned} \tag{17}$$

This program achieves stable recovery in the sense that the recovery error is proportional to the noise level and reduces to zero in the noise-free case. Note, however, that in the presence of noise, the solution is not likely to be rank one.

Theorem 21 ([72]). *Consider a signal $x \in \mathbb{C}^N$ satisfying $\|x\|_1 \le \beta$ and $|x[0]| \ge \gamma > 0$. Suppose that the measurements are taken with the diagonal matrices D_1 and D_2 (masks) associated with d_1 and d_2 given in (5). Then, the solution $\widetilde{X}$ of the convex program (17) obeys*

$$\|\widetilde{X} - xx^*\|_2 \le C(\beta,\gamma)\varepsilon$$

for some numerical constant $C(\beta,\gamma)$.

Phase retrieval from STFT measurements using SDP was considered in [75]. Here, SDP relaxation in the noiseless case takes on the form

$$\begin{aligned} \min_{X\in\mathscr{H}^N} \quad \mathrm{trace}(X) \quad \text{subject to} \quad & X \succeq 0, \\ & y[m,k] = \mathrm{trace}(D_m^* f_k f_k^* D_m X), \\ & k = 0,\dots,K-1, \quad m = 0,\dots,M-1, \end{aligned} \tag{18}$$

where $M = \lceil N/L \rceil$ is the number of STFT windows and $\tilde{N} = N$ (see (4)). In [75], it was proven that (18) recovers the signal exactly under the following conditions.

Theorem 22 ([75]). *The convex program* (18) *has a unique feasible matrix* $X = xx^*$ *for almost all non-vanishing signals* x *if:*

- $d[n] \neq 0$ *for* $n = 0,\dots,W-1$,
- $2L \leq W \leq N/2$,
- $4L \leq K \leq N$,
- $x[n]$ *is known a priori for* $0 \leq n \leq \left\lfloor \frac{L}{2} \right\rfloor$,

where W *is the window's length.*

Note that for $L = 1$, no prior knowledge on the entries of x is required.

An interesting implication of this theorem is that recovery remains exact even if we merely have access to the low frequencies of the data. This property is called *super-resolution* and will be discussed in more detail in the context of sparse signals. Numerically, the performance of (18) is better than Theorem 22 suggests. Specifically, it seems that for $W \geq 2L$, (18) recovers xx^* exactly without any prior knowledge on the entries of x, as demonstrated in Figure 5 (a similar example is given in [75]). Additionally, the program is stable in the presence of noise.

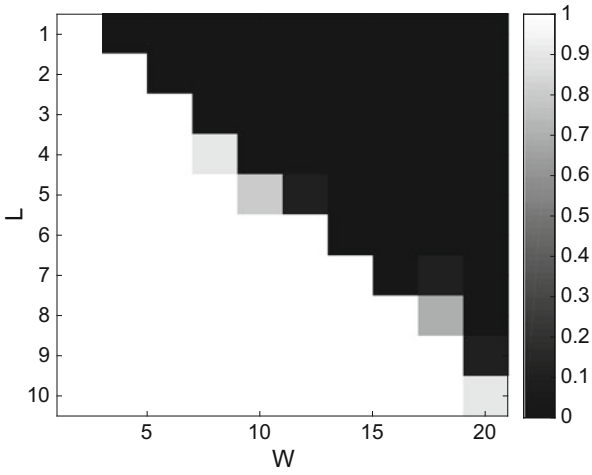


Fig. 5 The empirical success rate of the SDP relaxation for STFT phase retrieval (18) with a rectangular window of length W, i.e., $d[n] = 1$ for $n = 0,\dots,W-1$. For each pair (W, L), 100 complex signals of length $N = 40$ were drawn from an i.i.d. normal distribution. The figure presents the empirical success rate. An experiment was declared as a success if the recovery error is below 10^{-4}.

4.3 Additional Non-Convex Algorithms

In this section, we present additional non-convex algorithms for phase retrieval with special focus on STFT phase retrieval. A naive way to estimate the signal from its phaseless measurements is by minimizing the non-convex loss functions (12) or (13) by employing a gradient descent scheme. However, as demonstrated in Figure 4, this algorithm is likely to converge to a local minimum due to the non-convexity of the loss functions. Hence, the key is to introduce an efficient method to initialize the non-convex algorithm sufficiently close to the global minimum.

A recent paper [20] suggests an initialization technique for STFT phase retrieval, which we now describe. Consider the one-dimensional Fourier transform of the data with respect to the frequency variable (see (4)), given by

$$\tilde{y}[m,\ell] = \sum_{n=0}^{N-1} x[n]\overline{x[n+\ell]}d[mL-n]\overline{d[mL-n-\ell]},$$

where $\tilde{N} = N$ and both the signal and the window are assumed to be periodic. For fixed ℓ, we obtain the linear system of equations

$$\tilde{y}_\ell = G_\ell x_\ell, \tag{19}$$

where $\tilde{y}_\ell = \{\tilde{y}[m,\ell]\}_{m=0}^{\lceil N/L\rceil-1}$, $x_\ell \in \mathbb{C}^N$ is the ℓth diagonal of the matrix xx^* and $G_\ell \in \mathbb{C}^{\lceil N/L\rceil\times N}$ is the matrix with (m,n)th entry given by $d[mL-n]\overline{d[mL-n-\ell]}$. For $L = 1$, G_ℓ is a circulant matrix. Clearly, recovering x_ℓ for all ℓ is equivalent to recovering xx^*. Hence, the ability to estimate x depends on the properties of the window which determines G_ℓ. To make this statement precise, we use the following definition.

Definition 1. A window d is called an *admissible window of length* W if for all $\ell = -(W-1),\dots,(W-1)$ the associated circulant matrices G_ℓ in (19) are invertible. An example of an admissible window is a rectangular window with $W \le N/2$ and N a prime number. If the STFT window is sufficiently long and admissible, then the STFT phase retrieval has a closed-form solution. This solution can be obtained by the principal eigenvector of a matrix, constructed as the solution of a least-squares problem according to (19). This algorithm is summarized in Algorithm 4.

Theorem 23 ([20]). *Let $L = 1$, and suppose that d is an admissible window of length $W \ge \lceil \frac{N+1}{2} \rceil$ (see Definition 1). Then, Algorithm 4 recovers any complex signal up to global phase.*

In many cases, the window is shorter than $\lceil \frac{N+1}{2} \rceil$. However, the same technique can be applied to initialize a refinement process, such as a gradient method or the Griffin-Lim algorithm (GLA). In this case, the distance between the initial vector (the output of Algorithm 4) and the target signal can be estimated as follows.

Algorithm 4 Least-squares algorithm for STFT phase retrieval with $L = 1$

Input: The STFT magnitude $y[m,k]$ as given in (4) with $\widetilde{N} = N$
Output: x_{est} - estimation of x

1. Compute $\tilde{y}[m,\ell]$, the one-dimensional DFT of $y[m,k]$ with respect to the second variable.
2. Construct a matrix X_0 such that

$$\operatorname{diag}(X_0,\ell) = \begin{cases} G_\ell^\dagger \tilde{y}_\ell, & \ell = -(W-1),\cdots,(W-1), \\ 0, & \text{otherwise}, \end{cases}$$

where $G_\ell \in \mathbb{R}^{N\times N}$ and $\tilde{y}_\ell$ are given in (19).

Return:

$$x_{est} = \sqrt{\sum_{n\in P}(G_0^\dagger y_0)[n]}\, x_p,$$

where $P := \{n : (G_0^\dagger y_0)[n] > 0\}$ and x_p is the principle (unit-norm) eigenvector of X_0.

Theorem 24 ([20]). *Suppose that $L = 1$, $\|x\|_2 = 1$, d is an admissible window of length $W \geq 2$ and that $\|x\|_\infty \leq \sqrt{B/N}$ for some $0 < B \leq N/(2N-4W+2)$. Then, the output x_0 of Algorithm 4 satisfies*

$$\min_{\phi\in[0,2\pi)} \|x - x_0 e^{j\phi}\|_2^2 \leq 1 - \sqrt{1 - 2B\,\frac{N-2W+1}{N}}.$$

For $L > 1$, it is harder to obtain a reliable estimation of the diagonals of xx^*. Nevertheless, a simple heuristic is proposed in [20] based on the smoothing properties of typical STFT windows. Figure 6 shows experiments corroborating the effectiveness of this initialization approach for $L > 1$.

We have seen that under some conditions, the STFT phaseless measurements provide partial information on the matrix xx^*. In some cases, the main diagonal of xx^*, or equivalently the temporal magnitude of x, is also measured. Therefore, if the signal is non-vanishing, then all entries of the matrix xx^* can be normalized to have unit modulus. This in turn implies that the STFT phase retrieval problem is equivalent to estimating the missing entries of a rank-one matrix with unit modulus entries (i.e., phases). This problem is known as *phase synchronization*. In recent years, several algorithms for phase synchronization were suggested and analyzed, among them are eigenvector-based methods, SDP relaxations, projected power methods, and approximate message passing algorithms [7, 8, 28, 36, 100, 123]. Recent papers [69, 70] adopted this approach and suggested spectral and greedy algorithms for STFT phase retrieval. These methods are accompanied by stability guarantees and can be modified for phase retrieval using masks. The main shortcoming of this approach is that it relies on a good estimation of the temporal magnitudes which may not always be available.

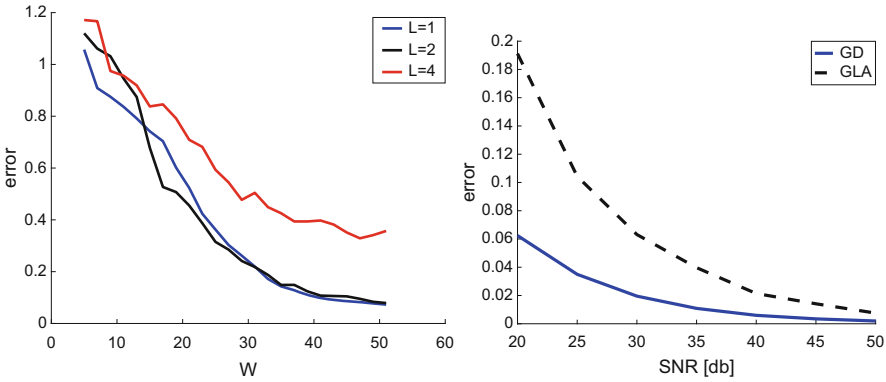


Fig. 6 **(left)** Average error (over 50 experiments) of the initialization method of Algorithm 4 as a function of W and L. The experiments were conducted on a signal of length $N = 101$ with a Gaussian window $d[n] = e^{-n^2/2\sigma^2}$. The window length was set to $W = 3\sigma$. **(right)** Average normalized recovery error (over 20 experiments) of the gradient descent (GD) and Griffin-Lim algorithm (GLA) in the presence of normal i.i.d. noise. Both algorithms were initialized by Algorithm 4. The experiments were conducted on signals of length $N = 53$ with a rectangular window of length $W = 19$ and $L = 2$.

Another interesting approach has been recently proposed in [101]. This paper suggests a multistage algorithm based on spectral clustering and angular synchronization. It is shown that the algorithm achieves stable estimation (and exactness in the noise-free setting) with only $O(N \log N)$ phaseless STFT measurements. Nevertheless, the algorithm builds upon random STFT windows of length N, while most applications use shorter windows.

4.4 Algorithms for Sparse Signals

In this section, we assume that the signal is sparse with a sparsity level defined as

$$s = \{\#n : x[n] \neq 0\}.$$

In this case, the basic phase retrieval problem (12) can be modified to the constrained least-squares problem

$$\min_{z\in\mathbb{C}^N} \sum_{k=0}^{K-1}\sum_{n=0}^{M-1} \left(y[m,k] - \left|f_k^* D_m z\right|^2\right)^2 \quad \text{subject to} \quad \|z\|_0 \leq s, \tag{20}$$

where we use $\|\cdot\|_0$ as the standard ℓ_0 pseudo-norm counting the non-zero entries of a signal.

Many phase retrieval algorithms for sparse signals are modifications of known algorithms for the non-sparse case. For instance, gradient algorithms were modified to take into account the sparsity structure. The underlying idea of these algorithms is to add a thresholding step at each iteration. Theoretical analysis of these algorithms for phase retrieval with random sensing vectors is considered in [30, 135]. A similar modification for the HIO algorithm was proposed in [95]. Modifications of SDP relaxation methods for phase retrieval with random sensing vectors were considered in [85, 97, 98]. Here, the core idea is to incorporate a sparse-promoting regularizer in the objective function. However, this technique cannot be adapted directly to Fourier phase retrieval because of the trivial ambiguities of translation and conjugate reflection; see a detailed explanation in [74]. To overcome this barrier, a two-stage sparse-phase retrieval (TSPR) algorithm was proposed in [73]. The first stage of the algorithm involves estimating the support of the signal directly from the support of its autocorrelation. This problem is equivalent to the *turnpike problem* of estimating a set of integers from their pairwise distances [124]. Once the support is known, the second stage involves solving an SDP to estimate the missing amplitudes. It was proven that TSPR recovers signals exactly in the noiseless case as long as the sparsity level is less than $O(N^{1/2})$. In the noisy setting, recovery is robust for sparsity level lower than $O(N^{1/4})$. A different SDP-based approach was suggested in [115]. This method proposes to promote a sparse solution by the log-det heuristic [52] and an $\ell_1 - \ell_2$ constraint on the matrix xx^*.

An alternative class of algorithms that has been proven to be highly effective for sparse signals is the class of greedy algorithms; see, for instance, [12, 90]. For phase retrieval tasks, a greedy optimization algorithm called GESPAR (GrEedy Sparse PhAse Retrieval) is proposed in [118]. The algorithm was applied for a variety of optical tasks, such as CDI and phase retrieval through waveguide arrays [116, 117, 121]. GESPAR is a local search algorithm, based on iteratively updating the signal support. Specifically, two elements are being updated at each iteration by swapping. Then, a non-convex objective function that takes the support into account is minimized by a damped Gauss-Newton method. The swap is carried out between the support element which corresponds to the smallest entry (in absolute value) and the off-support element with maximal gradient value of the objective function. A modification of GESPAR for STFT phase retrieval was presented in [47]. A schematic outline of GESPAR is given in Algorithm 5; for more details, see [118].

In practice, many optical measurement processes blur the fine details of the acquired data and act as low-pass filters. In these cases, one aims at estimating the signal from its low-resolution Fourier magnitudes. This problem combines two classical problems: phase retrieval and super-resolution. In recent years, super-resolution for sparse signals has been investigated thoroughly [3, 19, 21, 22, 31, 44]. In Theorem 22, we have seen that the SDP (18) can recover a signal from its low-resolution STFT magnitude. The problem of recovering a signal from its low-resolution phaseless measurements using masks was considered in [76, 110]. It was

Algorithm 5 A schematic outline of GESPAR algorithm; for details see [118]

Input: Fourier magnitude y as in (2) and sparsity level s
Output: x_{est} - estimation of x
Initialization:

- Generate a random support set $S^{(0)}$ of size s
- Employ a damped Gauss-Newton method with support $S^{(0)}$ and obtain an initial estimation $x^{(0)}$
- Set $\ell = 0$

while halting criterion false **do**:

- $\ell \leftarrow \ell + 1$
- Update support by swapping two entries, one in $S^{(\ell-1)}$ and one in the complementary set
- Minimize a non-convex objective with the given support $S^{(\ell)}$ using the damped Gauss-Newton method to obtain $x^{(\ell)}$

end while
Return: $x_{est} \leftarrow x^{(\ell)}$

proven that exact recovery may be obtained by only few[1] carefully designed masks if the underlying signal is sparse and its support is not clustered (this requirement is also known as the separation condition). An extension to the continuous setup was suggested in [41]. A combinatorial algorithm for recovering a signal from its low-resolution Fourier magnitude was suggested in [40]. The algorithm recovers an s-sparse signal exactly from $2s^2 - 2s + 2$ low-pass magnitudes. Nevertheless, this algorithm is unstable in the presence of noise due to error propagation.

5 Conclusion

In this chapter, we studied the problem of Fourier phase retrieval. We focused on the question of uniqueness, presented the main algorithmic approaches and discussed their properties. To conclude the chapter, we outline several fundamental gaps in the theory of Fourier phase retrieval.

Although many different methods have been proposed and analyzed in the last decade for Fourier phase retrieval, alternating projection algorithms maintained their popularity. Nevertheless, the theoretical understanding of these algorithms is limited. Another fundamental open question regards multidimensional phase retrieval. While almost all multidimensional signals are determined uniquely by their Fourier magnitude, there is no method that provably recovers the signal.

In many applications in optics, the measurement process acts as a low-pass filter. Hence, a practical algorithm should recover the missing phases (phase retrieval) and resolve the fine details of the data (super-resolution). In this chapter, we

[1] Specifically, several combinations of masks are suggested. Each combination consists of three to five deterministic masks.

surveyed several works dealing with the combined problem. Nonetheless, the current approaches are based on inefficient SDP programs [41, 75, 76, 110] or lack theoretical analysis [20, 115]. Additionally, even if all frequencies are available, it is still not clear what is the maximal sparsity that enables efficient and stable recovery of a signal from its Fourier magnitude.

In ultrashort laser pulse characterization, it is common to use the FROG methods that were introduced in Section 3.5. It is interesting to understand the minimal number of measurements which can guarantee uniqueness for FROG-type methods. Additionally, a variety of algorithms are applied to estimate signals from FROG-type measurements; a theoretical understanding of these algorithms is required.

References

1. https://www.nobelprize.org/nobel_prizes/medicine/laureates/1962/
2. E. Abbe, T. Bendory, W. Leeb, J. Pereira, N. Sharon, A. Singer, Multireference alignment is easier with an aperiodic translation distribution. Preprint (2017). arXiv:1710.02793
3. J.-M. Azais, Y. De Castro, F. Gamboa, Spike detection from inaccurate samplings. Appl. Comput. Harmon. Anal. **38**(2), 177–195 (2015)
4. R. Balan, P. Casazza, D. Edidin, On signal reconstruction without phase. Appl. Comput. Harmon. Anal. **20**(3), 345–356 (2006)
5. A.S. Bandeira, J. Cahill, D.G. Mixon, A.A. Nelson, Saving phase: injectivity and stability for phase retrieval. Appl. Comput. Harmon. Anal. **37**(1), 106–125 (2014)
6. A.S. Bandeira, Y. Chen, D.G. Mixon, Phase retrieval from power spectra of masked signals. Inf. Interference J. IMA **3**(2), 83–102 (2014)
7. A.S. Bandeira, Y. Chen, A. Singer, Non-unique games over compact groups and orientation estimation in cryo-EM (2015). arXiv preprint arXiv: 1505.03840
8. A.S. Bandeira, N. Boumal, A. Singer, Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. Math. Program. **163**(1), 145–167 (2017)
9. H.H. Bauschke, P.L. Combettes, D.R. Luke, Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization. J. Opt. Soc. Am. A Opt. Image Sci. Vis. **19**(7), 1334–1345 (2002)
10. H.H. Bauschke, P.L. Combettes, D.R. Luke, Hybrid projection–reflection method for phase retrieval. J. Opt. Soc. Am. A Opt. Image Sci. Vis. **20**(6), 1025–1034 (2003)
11. B. Baykal, Blind channel estimation via combining autocorrelation and blind phase estimation. IEEE Trans. Circuits Syst. Regul. Pap. **51**(6), 1125–1131 (2004)
12. A. Beck, Y.C. Eldar, Sparsity constrained nonlinear optimization: optimality conditions and algorithms. SIAM J. Optim. **23**(3), 1480–1509 (2013)
13. R. Beinert, Ambiguities in one-dimensional phase retrieval from Fourier magnitudes. Ph.D. thesis (Georg-August-Universität, Göttingen, 2015). http://hdl.handle.net/11858/00-1735-0000-0028-8691-3
14. R. Beinert, Non-negativity constraints in the one-dimensional discrete-time phase retrieval problem. Inf. Inference A J. IMA (2016). https://doi.org/10.1093/imaiai/iaw018
15. R. Beinert, One-dimensional phase retrieval with additional interference measurements. Results Math. (2016). https://doi.org/10.1007/s00025-016-0633-9
16. R. Beinert, G. Plonka, Ambiguities in one-dimensional discrete phase retrieval from Fourier magnitudes. J. Fourier Anal. Appl. **21**(6), 1169–1198 (2015)
17. R. Beinert, G. Plonka, Enforcing uniqueness in one-dimensional phase retrieval by additional signal information in time domain. Appl. Comput. Harmon. Anal. (2017). https://doi.org/10.1016/j.acha.2016.12.002

18. R. Beinert, G. Plonka, Sparse phase retrieval of one-dimensional signals by Prony's method. Front. Appl. Math. Stat. (Mathematics of Computation and Data Science) **3**(5), 1–10 (2017)
19. T. Bendory, Robust recovery of positive stream of pulses. IEEE Trans. Signal Process. **65**(8), 2114–2122 (2017)
20. T. Bendory, Y.C. Eldar, N. Boumal, Non-convex phase retrieval from STFT measurements. IEEE Trans. Inf. Theory **PP**(99) (2017). https://doi.org/doi:10.1109/TIT.2017.2745623
21. T. Bendory, S. Dekel, A. Feuer, Super-resolution on the sphere using convex optimization. IEEE Trans. Signal Process. **63**(9), 2253–2262 (2015)
22. T. Bendory, S. Dekel, A. Feuer, Robust recovery of stream of pulses using convex optimization. J. Math. Anal. Appl. **442**(2), 511–536 (2016)
23. T. Bendory, N. Boumal, C. Ma, Z. Zhao, A. Singer, Bispectrum inversion with application to multireference alignment. IEEE Trans. Signal Process. **PP**(99) (2017). https://doi.org/doi:10.1109/TSP.2017.2775591
24. T. Bendory, P. Sidorenko, Y.C. Eldar, On the uniqueness of FROG methods. IEEE Signal Process Lett. **24**(5), 722–726 (2017)
25. T. Bendory, D. Edidin, Y.C. Eldar, On signal reconstruction from FROG measurements. Preprint (2017). arXiv:1706.08494
26. J. Bertolotti, E.G. van Putten, C. Blum, A. Lagendijk, W.L. Vos, A.P. Mosk, Non-invasive imaging through opaque scattering layers. Nature **491**(7423), 232–234 (2012)
27. I. Bojarovska, A. Flinth, Phase retrieval from Gabor measurements. J. Fourier Anal. Appl. **22**(3), 542–567 (2016)
28. N. Boumal, Nonconvex phase synchronization. SIAM J. Optim. **26**(4), 2355–2377 (2016)
29. Y.M. Bruck, L.G. Sodin, On the ambiguity of the image reconstruction problem. Opt. Commun. **30**(3), 304–308 (1979)
30. T.T. Cai, X. Li, Z. Ma, et al., Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. Ann. Stat. **44**(5), 2221–2251 (2016)
31. E.J. Candès, C. Fernandez-Granda, Towards a mathematical theory of super-resolution. Commun. Pure Appl. Math. **67**(6), 906–956 (2014)
32. E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
33. E.J. Candès, X. Li, M. Soltanolkotabi, Phase retrieval from coded diffraction patterns. Appl. Comput. Harmon. Anal. **39**(2), 277–299 (2015)
34. E.J. Candès, Y.C. Eldar, T. Strohmer, V. Voroninski, Phase retrieval via matrix completion. SIAM Rev. **57**(2), 225–251 (2015)
35. H.N. Chapman, A. Barty, M.J. Bogan, S. Boutet, M. Frank, S.P. Hau-Riege, S. Marchesini, B.W. Woods, S. Bajt, W.H. Benner, et al., Femtosecond diffractive imaging with a soft-X-ray free-electron laser. Nat. Phys. **2**(12), 839–843 (2006)
36. Y. Chen, E. Candès, The projected power method: an efficient algorithm for joint alignment from pairwise differences (2016). arXiv preprint arXiv: 1609.05820
37. Y. Chen, E.J. Candés, Solving random quadratic systems of equations is nearly as easy as solving linear systems. Commun. Pure Appl. Math. **70**(5), 822–883 (2017)
38. C.-C. Chen, J. Miao, C.W. Wang, T.K. Lee, Application of optimization technique to noncrystalline X-ray diffraction microscopy: guided hybrid input-output method. Phys. Rev. B **76**(6), 064113 (2007)
39. B. Chen, R.A. Dilanian, S. Teichmann, B. Abbey, A.G. Peele, G.J. Williams, P. Hannaford, L.V. Dao, H.M. Quiney, K.A. Nugent, Multiple wavelength diffractive imaging. Phys. Rev. A **79**(2), 023809 (2009)
40. Y. Chen, Y.C. Eldar, A.J. Goldsmith, An algorithm for exact super-resolution and phase retrieval, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York, 2014), pp. 754–758
41. M. Cho, C. Thrampoulidis, W. Xu, B. Hassibi, Phaseless super-resolution in the continuous domain (2016). arXiv preprint arXiv: 1609.08522

42. K.W. DeLong, R. Trebino, J. Hunter, W.E. White, Frequency-resolved optical gating with the use of second-harmonic generation. J. Opt. Soc. Am. B **11**(11), 2206–2215 (1994)
43. D.L. Donoho, Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
44. V. Duval, G. Peyré, Exact support recovery for sparse spikes deconvolution. Found. Comput. Math. **15**(5), 1315–1355 (2015)
45. Y.C. Eldar, *Sampling Theory: Beyond Bandlimited Systems* (Cambridge University Press, Cambridge, 2015)
46. Y.C. Eldar, S. Mendelson, Phase retrieval: stability and recovery guarantees. Appl. Comput. Harmon. Anal. **36**(3), 473–494 (2014)
47. Y.C. Eldar, P. Sidorenko, D.G. Mixon, S. Barel, O. Cohen, Sparse phase retrieval from short-time Fourier measurements. IEEE Signal Process Lett. **22**(5), 638–642 (2015)
48. V. Elser, Phase retrieval by iterated projections. J. Opt. Soc. Am. A **20**(1), 40–55 (2003)
49. V. Elser, Solution of the crystallographic phase problem by iterated projections. Acta Crystallogr. A: Found. Crystallogr. **59**(3), 201–209 (2003)
50. A. Fannjiang, Absolute uniqueness of phase retrieval with random illumination. Inverse Prob. **28**(7), 20 (2012)
51. A. Faridian, D. Hopp, G. Pedrini, U. Eigenthaler, M. Hirscher, W. Osten, Nanoscale imaging using deep ultraviolet digital holographic microscopy. Opt. Express **18**(13), 14159–14164 (2010)
52. M. Fazel, H. Hindi, S.P. Boyd, Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices, in *American Control Conference, 2003. Proceedings of the 2003*, vol. 3 (IEEE, New York, 2003), pp. 2156–2162
53. L. Fejér, Über trigonometrische Polynome. J. Reine Angew. Math. **146**(2), 53–82 (1916)
54. J.R. Fienup, Phase retrieval algorithms: a comparison. Appl. Opt. **21**(15), 2758–2769 (1982)
55. J.R. Fienup, Reconstruction of a complex-valued object from the modulus of its fourier transform using a support constraint. J. Opt. Soc. Am. A **4**(1), 118–123 (1987)
56. J.R. Fienup, C. Dainty, Phase retrieval and image reconstruction for astronomy, in *Image Recovery: Theory and Application* (Elsevier, Amsterdam, 1987), pp. 231–275
57. L. Garwin, T. Lincoln, *A Century of Nature: Twenty-One Discoveries that Changed Science and the World* (University of Chicago Press, Chicago, 2010)
58. R.W. Gerchberg, W.O. Saxton, A practical algorithm for the determination of phase from image and diffraction plane pictures. Optik **35**, 237 (1972)
59. M.X. Goemans, D.P. Williamson, Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. J. ACM (JACM) **42**(6), 1115–1145 (1995)
60. T. Goldstein, C. Studer, Phasemax: convex phase retrieval via basis pursuit (2016). arXiv preprint arXiv: 1610.07531
61. M. Grant, S. Boyd, Y. Ye, CVX: Matlab software for disciplined convex programming (2008)
62. D.W. Griffin, J.S. Lim, Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoust. Speech Signal Process. **32**(2), 236–243 (1984)
63. D. Gross, F. Krahmer, R. Kueng, Improved recovery guarantees for phase retrieval form coded diffraction patterns. Appl. Comput. Harmon. Anal. **42**(1), 37–64 (2017)
64. R.W. Harrison, Phase problem in crystallography. J. Opt. Soc. Am. A **10**(5), 1046–1055 (1993)
65. M.H. Hayes, The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform. IEEE Trans. Acoust. Speech Signal Process. (ASSP) **30**(2), 140–154 (1982)
66. M.H. Hayes, J.H. McClellan, Reducible polynomials in more than one variable. Proc. IEEE (Proc. Lett.) **70**(2), 197–198 (1982)
67. M.H. Hayes, J.S. Lim, A.V. Oppenheim, Signal reconstruction from phase or magnitude. IEEE Trans. Acoust. Speech Signal Process. (ASSP) **28**(6), 672–680 (1980)
68. K. Huang, Y.C. Eldar, N.D. Sidiropoulos, Phase retrieval from 1D Fourier measurements: convexity, uniqueness, and algorithms. IEEE Trans. Signal Process. **64**(23), 6105–6117 (2016)

69. M.A. Iwen, B. Preskitt, R. Saab, A. Viswanathan, Phase retrieval from local measurements: improved robustness via eigenvector-based angular synchronization (2016). arXiv preprint arXiv: 1612.01182
70. M.A. Iwen, A. Viswanathan, Y. Wang, Fast phase retrieval from local correlation measurements. SIAM J. Imag. Sci. **9**(4), 1655–1688 (2016)
71. K. Jaganathan, B. Hassibi, Reconstruction of signals from their autocorrelation and cross-correlation vectors, with applications to phase retrieval and blind channel estimation (2016). arXiv: 1610.02620v1
72. K. Jaganathan, Y. Eldar, B. Hassibi, Phase retrieval with masks using convex optimization, in *2015 IEEE International Symposium on Information Theory (ISIT)* (IEEE, New York, 2015), pp. 1655–1659
73. K. Jaganathan, S. Oymak, B. Hassibi, Sparse phase retrieval: uniqueness guarantees and recovery algorithms. IEEE Trans. Signal Process. **65**(9), 2402–2410 (2017)
74. K. Jaganathan, Y.C. Eldar, B. Hassibi, Phase retrieval: an overview of recent developments, in *Optical Compressive Imaging*, ed. by A. Stern (CRC Press, Boca Raton, 2016)
75. K. Jaganathan, Y.C. Eldar, B. Hassibi, STFT phase retrieval: uniqueness guarantees and recovery algorithms. IEEE J. Sel. Top. Sign. Proces. **10**(4), 770–781 (2016)
76. K. Jaganathan, J. Saunderson, M. Fazei, Y.C. Eldar, B. Hassibi, Phaseless super-resolution using masks, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York, 2016), pp. 4039–4043
77. D.J. Kane, Principal components generalized projections: a review [invited]. J. Opt. Soc. Am. B **25**(6), A120–A132 (2008)
78. W. Kim, M.H. Hayes, The phase retrieval problem in X-ray crystallography, in *Proceedings: ICASSP 91: 1991 International Conference on Acoustics, Speech and Signal Processing*, May 14–17, 1991, vol. 3 (IEEE Signal Processing Society, Piscataway, 1991), pp. 1765–1768
79. W. Kim, M.H. Hayes , Phase retrieval using a window function. IEEE Trans. Signal Process. **41**(3), 1409–1412 (1993)
80. D. Kogan, Y.C. Eldar, D. Oron, On the 2D phase retrieval problem. IEEE Trans. Signal Process. **65**(4), 1058–1067 (2016)
81. D. Langemann, M. Tasche, Phase reconstruction by a multilevel iteratively regularized Gauss-Newton method. Inverse Prob. **24**(3), 035006(26) (2008)
82. D. Langemann, M. Tasche, Multilevel phase reconstruction for a rapidly decreasing interpolating function. Results Math. **53**(3–4), 333–340 (2009)
83. B. Leshem, R. Xu, Y. Dallal, J. Miao, B. Nadler, D. Oron, N. Dudovich, O. Raz, Direct single-shot phase retrieval from the diffraction pattern of separated objects. Nat. Commun. **7**, 10820 (2016)
84. B. Leshem, O. Raz, A. Jaffe, B. Nadler, The discrete sign problem: uniqueness, recovery algorithms and phase retrieval applications. Appl. Comput. Harmon. Anal. (2017). https://doi.org/10.1016/j.acha.2016.12.003, http://www.sciencedirect.com/science/article/pii/S1063520316300987?via%3Dihub
85. X. Li, V. Voroninski, Sparse signal recovery from quadratic measurements via convex programming. SIAM J. Math. Anal. **45**(5), 3019–3033 (2013)
86. Y.J. Liu, B. Chen, E.R. Li, J.Y. Wang, A. Marcelli, S.W. Wilkins, H. Ming, Y.C. Tian, K.A. Nugent, P.P. Zhu, et al., Phase retrieval in X-ray imaging based on using structured illumination. Phys. Rev. A **78**(2), 023817 (2008)
87. E.G. Loewen, E. Popov, *Diffraction Gratings and Applications* (CRC Press, Boca Raton, 1997)
88. D.R. Luke, Relaxed averaged alternating reflections for diffraction imaging. Inverse Prob. **21**(1), 37 (2004)
89. A.M. Maiden, M.J. Humphry, F. Zhang, J.M. Rodenburg, Superresolution imaging via ptychography. J. Opt. Soc. Am. A **28**(4), 604–612 (2011)
90. S.G. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process. **41**(12), 3397–3415 (1993)

91. S. Marchesini, Invited article: a unified evaluation of iterative projection algorithms for phase retrieval. Rev. Sci. Instrum. **78**(1), 011301 (2007)
92. S. Marchesini, Y.-C. Tu, H.-T. Wu, Alternating projection, ptychographic imaging and phase synchronization. Appl. Comput. Harmon. Anal. **41**(3), 815–851 (2015). https://doi.org/10.1016/j.acha.2015.06.005, http://www.sciencedirect.com/science/article/pii/S1063520315000913?via%3Dihub
93. A.V. Martin, F. Wang, N.D. Loh, T. Ekeberg, F.R.N.C. Maia, M. Hantke, G. van der Schot, C.Y. Hampton, R.G. Sierra, A. Aquila, et al., Noise-robust coherent diffractive imaging with a single diffraction pattern. Opt. Express **20**(15), 16650–16661 (2012)
94. J. Miao, P. Charalambous, J. Kirz, D. Sayre, Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. Nature **400**(6742), 342–344 (1999)
95. S. Mukherjee, C.S. Seelamantula, An iterative algorithm for phase retrieval with sparsity constraints: application to frequency domain optical coherence tomography, in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York, 2012), pp. 553–556
96. S. Nawab, T. Quatieri, J. Lim, Signal reconstruction from short-time Fourier transform magnitude. IEEE Trans. Acoust. Speech Signal Process. **31**(4), 986–998 (1983)
97. H. Ohlsson, A.Y. Yang, R. Dong, S.S. Sastry, Compressive phase retrieval from squared output measurements via semidefinite programming (2011), p. 1111. arXiv preprint arXiv
98. S. Oymak, A. Jalali, M. Fazel, Y.C. Eldar, B. Hassibi, Simultaneously structured models with application to sparse and low-rank matrices. IEEE Trans. Inf. Theory **61**(5), 2886–2908 (2015)
99. E. Pauwels, A. Beck, Y.C. Eldar, S. Sabach, On Fienup methods for regularized phase retrieval. Preprint (2017). arXiv:1702.08339
100. A. Perry, A.S. Wein, A.S. Bandeira, A. Moitra, Message-passing algorithms for synchronization problems over compact groups (2016). arXiv preprint arXiv: 1610.04583
101. G.E. Pfander, P. Salanevich, Robust phase retrieval algorithm for time-frequency structured measurements (2016). arXiv preprint arXiv: 1611.02540
102. L. Rabiner, B.-H. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Upper Saddle River, 1993)
103. J. Ranieri, A. Chebira, Y.M. Lu, M. Vetterli, Phase retrieval for sparse signals: uniqueness conditions (2013). Preprint. arXiv: 1308.3058v2
104. O. Raz, O. Schwartz, D. Austin, A.S. Wyatt, A. Schiavi, O. Smirnova, B. Nadler, I.A. Walmsley, D. Oron, N. Dudovich, Vectorial phase retrieval for linear characterization of attosecond pulses. Phys. Rev. Lett. **107**(13), 133902 (2011)
105. O. Raz, N. Dudovich, B. Nadler, Vectorial phase retrieval of 1-D signals. IEEE Trans. Signal Process. **61**(7), 1632–1643 (2013)
106. O. Raz, B. Leshem, J. Miao, B. Nadler, D. Oron, N. Dudovich, Direct phase retrieval in double blind fourier holography. Opt. Express **22**(21), 24935–24950 (2014)
107. I.K. Robinson, I.A. Vartanyants, G.J. Williams, M.A. Pfeifer, J.A. Pitney, Reconstruction of the shapes of gold nanocrystals using coherent X-ray diffraction. Phys. Rev. Lett. **87**(19), 195505 (2001)
108. J.M. Rodenburg, Ptychography and related diffractive imaging methods. *Advances in Imaging and Electron Physics* vol. 150, (Elsevier, Amsterdam, 2008), pp. 87–184
109. J.A. Rodriguez, R. Xu, C.-C. Chen, Y. Zou, J. Miao, Oversampling smoothness: an effective algorithm for phase retrieval of noisy diffraction intensities. J. Appl. Crystallogr. **46**(2), 312–318 (2013)
110. F. Salehi, K. Jaganathan, B. Hassibi, Multiple illumination phaseless super-resolution (MIPS) with applications to phaseless DOA estimation and diffraction imaging (2017). arXiv preprint arXiv: 1701.03515
111. R.L. Sandberg, C. Song, P.W. Wachulak, D.A. Raymondson, A. Paul, B. Amirbekian, E. Lee, A.E. Sakdinawat, L.-O. Chan, M.C. Marconi, et al., High numerical aperture tabletop soft X-ray diffraction microscopy with 70-nm resolution. Proc. Natl. Acad. Sci. **105**(1), 24–27 (2008)

112. D. Sayre, Some implications of a theorem due to Shannon. Acta Crystallogr. **5**(6), 843–843 (1952)
113. B. Seifert, H. Stolz, M. Tasche, Nontrivial ambiguities for blind frequency-resolved optical gating and the problem of uniqueness. J. Opt. Soc. Am. B **21**(5), 1089–1097 (2004)
114. B. Seifert, H. Stolz, M. Donatelli, D. Langemann, M. Tasche, Multilevel Gauss-Newton methods for phase retrieval problems. J. Phys. A Math. Gen. **39**(16), 4191–4206 (2006)
115. Y. Shechtman, Y.C. Eldar, A. Szameit, M. Segev, Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing. Opt. Express **19**(16), 14807–14822 (2011)
116. Y. Shechtman, Y.C. Eldar, O. Cohen, M. Segev, Efficient coherent diffractive imaging for sparsely varying objects. Opt. Express **21**(5), 6327–6338 (2013)
117. Y. Shechtman, E. Small, Y. Lahini, M. Verbin, Y.C. Eldar, Y. Silberberg, M. Segev, Sparsity-based super-resolution and phase-retrieval in waveguide arrays. Opt. Express **21**(20), 24015–24024 (2013)
118. Y. Shechtman, A. Beck, Y.C. Eldar, Gespar: efficient phase retrieval of sparse signals. IEEE Trans. Signal Process. **62**(4), 928–938 (2014)
119. Y. Shechtman, Y.C. Eldar, O. Cohen, H.N. Chapman, J. Miao, M. Segev, Phase retrieval with application to optical imaging: a contemporary overview. IEEE Signal Process. Mag. **32**(3), 87–109 (2015)
120. P. Sidorenko, O. Cohen, Single-shot ptychography. Optica **3**(1), 9–14 (2016)
121. P. Sidorenko, A. Fleischer, Y. Shechtman, Y.C. Eldar, M. Segev, O. Cohen, Sparsity-based super-resolution coherent diffractive imaging of (practically) 1D images using extreme UV radiation, in *CLEO: QELS_Fundamental Science* (Optical Society of America, Washington, DC, 2013), p. QF1C–7
122. P. Sidorenko, O. Lahav, Z. Avnat, O. Cohen, Ptychographic reconstruction algorithm for frequency-resolved optical gating: super-resolution and supreme robustness. Optica **3**(12), 1320–1330 (2016)
123. A. Singer, Angular synchronization by eigenvectors and semidefinite programming. Appl. Comput. Harmon. Anal. **30**(1), 20–36 (2011)
124. S.S. Skiena, W.D. Smith, P. Lemke, Reconstructing sets from interpoint distances, in *Proceedings of the Sixth Annual Symposium on Computational Geometry* (ACM, New York, 1990), pp. 332–339
125. M. Soltanolkotabi, Structured signal recovery from quadratic measurements: breaking sample complexity barriers via nonconvex optimization (2017). arXiv preprint arXiv: 1702.06175
126. M. Stefik, Inferring DNA structures from segmentation data. Artif. Intell. **11**(1), 85–114 (1978)
127. A.M. Tillmann, Y.C. Eldar, J. Mairal, DOLPHIn–dictionary learning for phase retrieval. IEEE Trans. Signal Process. **64**(24), 6485–6500 (2016)
128. R. Trebino, *Frequency-Resolved Optical Gating: The Measurement of Ultrashort Laser Pulses* (Springer, New York, 2012)
129. R. Trebino, D.J. Kane, Using phase retrieval to measure the intensity and phase of ultrashort pulses: frequency-resolved optical gating. J. Opt. Soc. Am. A **10**(5), 1101–1111 (1993)
130. R. Trebino, K.W. DeLong, D.N. Fittinghoff, J.N. Sweetser, M.A. Krumbügel, B.A. Richman, D.J. Kane, Measuring ultrashort laser pulses in the time-frequency domain using frequency-resolved optical gating. Rev. Sci. Instrum. **68**(9), 3277–3295 (1997)
131. I. Waldspurger, Phase retrieval with random gaussian sensing vectors by alternating projections (2016). arXiv preprint arXiv: 1609.03088
132. I. Waldspurger, A. d'Aspremont, S. Mallat, Phase recovery, MaxCut and complex semidefinite programming. Math. Program. **149**(1–2), 47–81 (2015)
133. A. Walther, The question of phase retrieval in optics. J. Mod. Opt. **10**(1), 41–49 (1963)
134. G. Wang, G.B. Giannakis, Y.C. Eldar, Solving systems of random quadratic equations via truncated amplitude flow (2016). arXiv preprint arXiv: 1605.08285
135. G. Wang, L. Zhang, G.B. Giannakis, M. Akçakaya, J. Chen, Sparse phase retrieval via truncated amplitude flow (2016). arXiv preprint arXiv: 1611.07641

136. T.C. Wong, J. Ratner, V. Chauhan, J. Cohen, P.M. Vaughan, L. Xu, A. Consoli, R. Trebino, Simultaneously measuring two ultrashort laser pulses on a single-shot using double-blind frequency-resolved optical gating. J. Opt. Soc. Am. B **29**(6), 1237–1244 (2012)
137. L. Xu, P. Yan, T. Chang, Almost unique specification of discrete finite length signal: from its end point and Fourier transform magnitude, in *Proceedings : ICASSP 87 : IEEE International Conference on Acoustics, Speech, and Signal*, vol. 12 (IEEE, New York, 1987), pp. 2097–2100
138. L.-H. Yeh, J. Dong, J. Zhong, L. Tian, M. Chen, G. Tang, M. Soltanolkotabi, L. Waller, Experimental robustness of Fourier ptychography phase retrieval algorithms. Opt. Express **23**(26), 33214–33240 (2015)

Compressed Sensing Approaches for Polynomial Approximation of High-Dimensional Functions

Ben Adcock, Simone Brugiapaglia, and Clayton G. Webster

Abstract In recent years, the use of sparse recovery techniques in the approximation of high-dimensional functions has garnered increasing interest. In this work we present a survey of recent progress in this emerging topic. Our main focus is on the computation of polynomial approximations of high-dimensional functions on d-dimensional hypercubes. We show that smooth, multivariate functions possess expansions in orthogonal polynomial bases that are not only approximately sparse but possess a particular type of structured sparsity defined by so-called lower sets. This structure can be exploited via the use of weighted ℓ^1 minimization techniques, and, as we demonstrate, doing so leads to sample complexity estimates that are at most logarithmically dependent on the dimension d. Hence the curse of dimensionality – the bane of high-dimensional approximation – is mitigated to a significant extent. We also discuss several practical issues, including unknown noise (due to truncation or numerical error), and highlight a number of open problems and challenges.

Keywords High-dimensional approximation · Weighted ℓ^1 minimization · Orthogonal polynomials · Lower sets

1 Introduction

The approximation of high-dimensional functions is a fundamental difficulty in a large number of fields, including neutron, tomographic and magnetic resonance image reconstruction, uncertainty quantification (UQ), optimal control, and parameter identification for engineering and science applications. In addition, this problem

B. Adcock (✉) · S. Brugiapaglia
Simon Fraser University, Burnaby, BC, Canada
e-mail: ben_adcock@sfu.ca; simone_brugiapaglia@sfu.ca

C.G. Webster
University of Tennessee and Oak Ridge National Lab, Oak Ridge, TN, USA
e-mail: webstercg@math.utk.edu; webstercg@ornl.gov

H. Boche et al. (eds.), *Compressed Sensing and its Applications*,
Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-69802-1_3

naturally arises in computational solutions to kinetic plasma physics equations, the many-body Schrödinger equation, Dirac and Maxwell equations for molecular electronic structures and nuclear dynamic computations, options pricing equations in mathematical finance, Fokker-Planck and fluid dynamics equations for complex fluids, turbulent flow, quantum dynamics, molecular life sciences, and nonlocal mechanics. The subject of intensive research over the last half-century, high-dimensional approximation is made challenging by the *curse of dimensionality*, a phrase coined by Bellman [7]. Loosely speaking, this refers to the tendency of naïve approaches to exhibit exponential blow-up in complexity with increasing dimension. Progress is possible, however, by placing restrictions on the class of functions to be approximated, for example, smoothness, anisotropy, sparsity, and compressibility. Well-known algorithms such as sparse grids [14, 54, 55, 69], which are specifically designed to capture this behavior, can mitigate the curse of dimensionality to a substantial extent.

While successful, however, such approaches typically require strong *a priori* knowledge of the functions being approximated, e.g., the parameters of the anisotropic behavior, or costly adaptive implementations to estimate the anisotropy during the approximation process. The efficient approximation of high-dimensional functions in the absence of such knowledge remains a significant challenge.

In this chapter, we consider new methods for high-dimensional approximation based on the techniques of compressed sensing. Compressed sensing is an appealing approach for reconstructing signals from underdetermined systems, i.e., with far smaller number of measurements compared to the signal length [16, 31]. This approach has emerged in the last half a dozen years as an alternative to more classical approximation schemes for high-dimensional functions, with the aim being to overcome some of the limitations mentioned above. Under natural sparsity or compressibility assumptions, it enjoys a significant improvement in sample complexity over traditional methods such as discrete least squares, projection, and interpolation [37, 38]. Our intention in this chapter is to both present an overview of existing work in this area, focusing particularly on the mitigation of the curse of dimensionality, and to highlight existing open problems and challenges.

1.1 Compressed Sensing for High-Dimensional Approximation

Compressed sensing asserts that a vector $\boldsymbol{x} \in \mathbb{C}^n$ possessing a k-sparse representation in a fixed orthonormal basis can be recovered from a number of suitably chosen measurements m that are linear in k and logarithmic in the ambient dimension n. In practice, recovery can be achieved via a number of different approaches, including convex optimization (ℓ^1 minimization), greedy or thresholding algorithms.

Let $f : D \to \mathbb{C}$ be a function, where $D \subseteq \mathbb{R}^d$ is a domain in $d \gg 1$ dimensions. In order to apply compressed sensing techniques to approximate f, we must first address the following three questions:

(i) In which orthonormal system of functions $\{\phi_i\}_{i=1}^n$ does f have an approximately sparse representation?

(ii) Given suitable assumptions on f (e.g., smoothness) how fast does the best k-term approximation error decay?
(iii) Given such a system $\{\phi_i\}_{i=1}^n$, what are suitable measurements to take of f?

The concern of this chapter is the approximation of smooth functions, and as such we will use orthonormal bases consisting of multivariate orthogonal polynomials. In answer to (i) and (ii) in Section 2, we discuss why this choice leads to approximate sparse representations for functions with suitable smoothness and characterize the best k-term approximation error in terms of certain regularity conditions. As we note in Section 2.2, practical examples of such functions include parameter maps of many different types of parametric PDEs.

For sampling, we evaluate f at a set of points $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m \in D$. This approach is simple and particularly well-suited in practical problems. In UQ, for example, it is commonly referred to as a *nonintrusive* approach [46] or *stochastic collocation* [52]. More complicated measurement procedures – for instance, *intrusive* procedures such as inner products with respect to a set of basis functions – are often impractical or even infeasible, since, for example, they require computation of high-dimensional integrals. The results presented in Section 3 identify appropriate (random) choices of the sample points $\{\boldsymbol{z}_i\}_{i=1}^m$ and bound for the number of measurements m under which f can be stably and robustly recovered from the data $\{f(\boldsymbol{z}_i)\}_{i=1}^m$.

1.2 Structured Sparsity

The approximation of high-dimensional functions using polynomials differs from standard compressed sensing in several key ways. Standard compressed sensing exploits sparsity of the finite vector of coefficients $\boldsymbol{c} \in \mathbb{C}^n$ of a finite-dimensional signal $\boldsymbol{x} \in \mathbb{C}^n$. However, polynomial coefficients of smooth functions typically possess more detailed structure than just sparsity. Loosely speaking, coefficients corresponding to low polynomial orders tend to be larger than coefficients corresponding to higher orders. This raises several questions:

(iv) What is a reasonable *structured sparsity* model for polynomial coefficients of high-dimensional functions?
(v) How can such structured sparsity be exploited in the reconstruction procedure, and by how much does doing this reduce the number of measurements required?

In Section 2.3 it is shown that high-dimensional functions can be approximated with quasi-optimal rates of convergence by k-term polynomial expansions with coefficients lying in so-called *lower* sets of multi-indices. As we discuss, sparsity in lower sets is a type of structured sparsity, and in Section 3 we show how it can be exploited by replacing the classical ℓ^1 regularizer by a suitable weighted ℓ^1-norm. Growing weights penalize high-degree polynomial coefficients, and when chosen appropriately, they act to promote lower set structure. In Section 3.2 nonuniform recovery techniques are used to identify a suitable choice of weights. This choice of

weights is then adopted in Section 3.5 to establish quasi-optimal uniform recovery guarantees for compressed sensing of polynomial expansions using weighted ℓ^1 minimization.

The effect of this weighted procedure is a substantially improved recovery guarantee over the case of unweighted ℓ^1 minimization, specifically, a measurement condition that is only logarithmically dependent on the dimension d and polynomially dependent on the sparsity k. Hence the curse of dimensionality is almost completely avoided. As we note in Section 3.3, these polynomial rates of growth in k agree with the best known recovery guarantees for oracle least-squares estimators.

1.3 Dealing with Infinity

Another way in which the approximation of high-dimensional functions differs from standard compressed sensing is that functions typically have infinite (as opposed to finite) expansions in orthogonal polynomial bases. In order to apply compressed sensing techniques, this expansion must be truncated in a suitable way. This leads to the following questions:

(vi) What is a suitable truncation of the infinite expansion?
(vii) How does the corresponding truncation error affect the overall reconstruction?

In Section 3 a truncation strategy – corresponding to a hyperbolic cross index set – is proposed based on the lower set structure. The issue of truncation error (question (vii)) presents some technical issues, both theoretical and practical, since this error is usually unknown *a priori*. In Section 3.6 we discuss a means to overcome these issues via a slightly modified optimization problem. Besides doing so, another benefit of the approach developed therein is that it yields approximations to f that also interpolate at the sample points $\{z_i\}_{i=1}^m$, a desirable property for certain applications. Furthermore, the results given in Section 3.6 also address the robustness of the recovery to unknown errors in the measurements. This is a quite common phenomenon in applications, since function samples are often the result of (inexact) numerical computations.

1.4 Main Results

We now summarize our main results. In order to keep the presentation brief, in this chapter we limit ourselves to functions defined on the unit hypercube $D = (-1, 1)^d$ and consider expansions in orthonormal polynomial bases $\{\phi_{\boldsymbol{i}}\}_{\boldsymbol{i}\in\mathbb{N}_0^d}$ of Chebyshev or Legendre type. We note in passing, however, that many of our results apply immediately (or extend straightforwardly) to more general systems of functions. See Section 4 for some further discussion.

Let ν be the probability measure under which the basis $\{\phi_{\boldsymbol{i}}\}_{\boldsymbol{i}\in\mathbb{N}_0^d}$ is orthonormal. Our main result is as follows:

Theorem 1. *Let $k \in \mathbb{N}$, $0 < \varepsilon < 1$, $\{\phi_i\}_{i\in\mathbb{N}_0^d}$ be the orthonormal Chebyshev or Legendre basis on $D = (-1,1)^d$, $\Lambda = \Lambda_k^{\mathrm{HC}}$ be the hyperbolic cross of index k and define weights $\boldsymbol{u} = (u_i)_{i\in\mathbb{N}_0^d}$, where $u_i = \|\phi_i\|_{L^\infty}$. Suppose that*

$$m \gtrsim k^\gamma \left(\log^2(k)\min\{d + \log(k), \log(2d)\log(k)\} + \log(k)\log(\log(k)/\varepsilon)\right),$$

where $\gamma = \frac{\log(3)}{\log(2)}$ or $\gamma = 2$ for Chebyshev or Legendre polynomials, respectively, and draw $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m \in D$ independently according to ν. Then with probability at least $1-\varepsilon$ the following holds. For any $f \in L^2_\nu(D) \cap L^\infty(D)$ satisfying

$$\left\| f - \sum_{i\in\Lambda} c_i \phi_i \right\|_{L^\infty} \leq \eta, \tag{1}$$

for some known $\eta \geq 0$, it is possible to compute, via solving a $\ell^1_{\boldsymbol{u}}$ minimization problem of size $m \times n$ where $n = |\Lambda|$, an approximation $\tilde{f}$ from the samples $\boldsymbol{y} = (f(\boldsymbol{z}_j))_{j=1}^m$ that satisfies

$$\|f - \tilde{f}\|_{L^2_\nu} \lesssim \frac{\sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}}}{k^{\gamma/2}} + \eta, \qquad \left\|f - \tilde{f}\right\|_{L^\infty} \lesssim \sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + k^{\gamma/2}\eta. \tag{2}$$

Here $\boldsymbol{c}$ are the coefficients of f in the basis $\{\phi_i\}_{i\in\mathbb{N}_0^d}$ and $\sigma_{s,L}(\boldsymbol{c})_{1,\boldsymbol{u}}$ is the $\ell^1_{\boldsymbol{u}}$-norm error of the best approximation of $\boldsymbol{c}$ by a vector that is k-sparse and lower.

Note that the condition (1) is strong, since it assumes an *a priori* upper bound on the expansion error is available. Such a condition is unlikely to be met in practice. In Section 3.6 we discuss recovery results for general f without such *a priori* bounds.

1.5 Existing Literature

The first results on compressed sensing with orthogonal polynomials in the one-dimensional setting appeared in [60], based on earlier work in sparse trigonometric expansions [58]. This was extended to the higher-dimensional setting in [72]. Weighted ℓ^1 minimization was studied in [61], and recovery guarantees given in terms of so-called weighted sparsity. However, this does not lead straightforwardly to explicit measurement conditions for quasi-best k-term approximation. The works [1, 22] introduced new guarantees for weighted ℓ^1 minimization of nonuniform and uniform types, respectively, leading to optimal sample complexity estimates for recovering high-dimensional functions using k-term approximations in lower sets. Theorem 1 is based on results in [22]. Relevant approaches to compressed sensing in infinite dimensions have also been considered in [1, 2, 4, 11, 13, 66]

Applications of compressed sensing to UQ, specifically the computation of polynomial chaos expansions of parametric PDEs, can be found in [10, 32, 47, 56, 59, 73]

and references therein. Throughout this chapter we use random sampling from the orthogonality measure of the polynomial basis. We do this for its simplicity, and the theoretical optimality of the recovery guarantees in terms of the dimension d. Other strategies, which typically seek a smaller error or lower polynomial factor of k in the sample complexity, have been considered in [39, 41, 44, 52, 53, 64, 71]. Working toward a similar end, various approaches have also been considered to learn a better sparsity basis [43, 74] or to use additional gradient samples [57]. In this chapter, we focus on fixed bases of Chebyshev or Legendre polynomials in the unit cube. For results in $\mathbb{R}^d$ using Hermite polynomials, see [39, 41, 53].

In some scenarios, a suitable lower set may be known in advance or be computed via an adaptive search. In this case, least-squares methods may be suitable. A series of works have studied the sample complexity of such approaches in the context of high-dimensional polynomial approximation [19, 24, 28, 29, 40, 48–51, 53]. We review a number of these results in Section 3.3.

2 Sparse Polynomial Approximation of High-Dimensional Functions

2.1 Setup and Notation

We first require some notation. For the remainder of this chapter, $D = (-1,1)^d$ will be the d-dimensional unit cube. The vector $\boldsymbol{z} = (z_1, \dots, z_d)$ will denote the variable in D, and $\boldsymbol{i} = (i_1, \dots, i_d) \in \mathbb{N}_0^d$ will be a multi-index. Let $\nu^{(1)}, \dots, \nu^{(d)}$ be probability measures on the unit interval $(-1,1)$. We consider the tensor product probability measure ν on D given by $\nu = \nu^{(1)} \otimes \cdots \otimes \nu^{(d)}$. Let $\{\phi_i^{(k)}\}_{i=0}^{\infty}$ be an orthonormal polynomial basis of $L^2_{\nu^{(k)}}(-1,1)$ and define the corresponding tensor product orthonormal basis $\{\phi_{\boldsymbol{i}}\}_{\boldsymbol{i}\in\mathbb{N}_0^d}$ of $L^2_\nu(D)$ by

$$\phi_{\boldsymbol{i}} = \phi_{i_1}^{(1)} \otimes \cdots \otimes \phi_{i_d}^{(d)}, \qquad \boldsymbol{i} = (i_1, \dots, i_d) \in \mathbb{N}_0^d.$$

We let $\|\cdot\|_{L^2_\nu}$ and $\langle \cdot, \cdot \rangle_{L^2_\nu}$ denote the norm and inner product on $L^2_\nu(D)$, respectively.

Let $f \in L^2_\nu(D) \cap L^\infty(D)$ be the function to be approximated, and write

$$f = \sum_{\boldsymbol{i}\in\mathbb{N}_0^d} c_{\boldsymbol{i}} \phi_{\boldsymbol{i}}, \tag{3}$$

where $c_{\boldsymbol{i}} = \langle f, \phi_{\boldsymbol{i}} \rangle_{L^2_\nu}$ are the coefficients of f in the basis $\{\phi_{\boldsymbol{i}}\}_{\boldsymbol{i}\in\mathbb{N}_0^d}$. We define

$$\boldsymbol{c} = (c_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}_0^d} \in \ell^2(\mathbb{N}_0^d),$$

to be the infinite vector of coefficients in this basis.

Example 1. Our main example will be Chebyshev or Legendre polynomials. In one dimension, these are orthogonal polynomials with respect to the weight functions

$$\mathrm{d}\nu = \frac{1}{2}\,\mathrm{d}z \quad \text{(Legendre)}, \qquad \mathrm{d}\nu = \frac{1}{\pi\sqrt{1-z^2}}\,\mathrm{d}z \quad \text{(Chebyshev)},$$

respectively. For simplicity, we will consider only tensor products of the same types of polynomials in each coordinate. The corresponding tensor product measures on D are consequently defined as:

$$\mathrm{d}\nu = 2^{-d}\,\mathrm{d}\boldsymbol{z} \quad \text{(Legendre)}, \qquad \mathrm{d}\nu = \prod_{j=1}^{d} \frac{1}{\pi\sqrt{1-z_j^2}}\,\mathrm{d}\boldsymbol{z} \quad \text{(Chebyshev)}.$$

We note also that many of the results presented below extend to more general families of orthogonal polynomials, e.g., Jacobi polynomials (see Remark 5).

As discussed in Section 1.3, it is necessary to truncate the infinite expansion (3) to a finite one. Throughout, we let $\Lambda \subset \mathbb{N}_0^d$ be a subset of size $|\Lambda| = n$, and define the truncated expansion

$$f_\Lambda = \sum_{\boldsymbol{i}\in\Lambda} c_{\boldsymbol{i}}\phi_{\boldsymbol{i}}.$$

We write $\boldsymbol{c}_\Lambda$ for the finite vector of coefficients with multi-indices in Λ. Whenever necessary, we will assume an ordering $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_n$ of the multi-indices in Λ, so that

$$f_\Lambda = \sum_{j=1}^{n} c_{\boldsymbol{i}_j}\phi_{\boldsymbol{i}_j}, \qquad \boldsymbol{c}_\Lambda = (c_{\boldsymbol{i}_j})_{j=1}^n \in \mathbb{C}^n.$$

We will adopt the usual convention and view $\boldsymbol{c}_\Lambda$ interchangeably as a vector in $\mathbb{C}^n$ and as an element of $\ell^2(\mathbb{N}_0^d)$ whose entries corresponding to indices $\boldsymbol{i} \notin \Lambda$ are zero.

2.2 Regularity and Best k-Term Approximation

In the high-dimensional setting, we assume the regularity of f is such that the complex continuation of f, represented as the map $f : \mathbb{C}^d \to \mathbb{C}$, is a holomorphic function on $\mathbb{C}^d$. In addition, for $1 \le k \le n$, we let

$$\Sigma_k = \left\{\boldsymbol{c} \in \ell^2(\mathbb{N}_0^d) : |\mathrm{supp}(\boldsymbol{c})| \le k\right\},$$

be the set of k-sparse vectors, and

$$\sigma_k(\boldsymbol{c})_1 = \inf_{\boldsymbol{d}\in\Sigma_k} \|\boldsymbol{c}-\boldsymbol{d}\|_1,$$

be the error of the best k-term approximation of $\boldsymbol{c}$, measured in the ℓ^1 norm.

Recently, for smooth functions as described above, sparse recovery of the polynomial expansion (3) with the use of compressed sensing has shown tremendous promise. However, this approach requires a small uniform bound of the underlying basis, given by

$$\Theta = \sup_{i\in\Lambda} \|\phi_i\|_{L^\infty(D)},$$

as the sample complexity m required to recover the best k-term approximation (up to a multiplicative constant) scales with the following bound (see, e.g., [35])

$$m \gtrsim \Theta^2 k \times \log \text{ factors}. \tag{4}$$

This poses a challenge for many multivariate polynomial approximation strategies as Θ is prohibitively large in high dimensions. In particular, for d-dimensional problems, $\Theta = 2^{d/2}$ for Chebyshev systems and so-called preconditioned Legendre systems [60]. Moreover, when using the standard Legendre expansion, the theoretical number of samples can exceed the cardinality of the polynomial subspace, unless the subspace a priori excludes all terms of high total order (see, e.g., [41, 72]). Therefore, the advantages of sparse polynomial recovery methods, coming from reduced sample complexity, are eventually overcome by the curse of dimensionality, in that such techniques require at least as many samples as traditional sparse interpolation techniques in high dimensions [37, 54, 55]. Nevertheless, in the next section we describe a common characteristic of the polynomial space spanned by the best k-terms, that we will exploit to overcome the curse of dimensionality in the sample complexity bound (4). As such, our work also provides a fair comparison with existing numerical polynomial approaches in high dimensions [6, 18–20, 65].

2.3 Lower Sets and Structured Sparsity

In many engineering and science applications, the target functions, despite being high-dimensional, are smooth and often characterized by a rapidly decaying polynomial expansion, whose most important coefficients are of low order [21, 25, 27, 42]. In such situations, the quest for finding the approximation containing the largest k terms can be restricted to polynomial spaces associated with *lower (also known as downward closed or monotone)* sets. These are defined as follows:

Definition 1. A set $S \subseteq \mathbb{N}_0^d$ is lower if, whenever $\boldsymbol{i} = (i_1, \ldots, i_d) \in S$ and $\boldsymbol{i}' = (i'_1, \ldots, i'_d) \in \mathbb{N}_0^d$ satisfies $i'_j \leq i_j$ for all $j = 1, \ldots, d$, then $\boldsymbol{i}' \in S$.

The practicality of downward closed sets is mainly computational, and has been demonstrated in different approaches such as quasi-optimal strategies, Taylor expansion, interpolation methods, and discrete least squares (see [6, 18–22, 26, 27, 49, 51, 62, 65] and references therein). For instance, in the context of parametric PDEs, it was shown in [21] that for a large class of smooth differential operators, with a certain type of anisotropic dependence on z, the solution map $z \mapsto f(z)$ can be approximated by its best k-term expansions associated with index sets of cardinality k, resulting in algebraic rates $k^{-\alpha}$, $\alpha > 0$ in the uniform and/or mean average sense. The same rates are preserved with index sets that are lower. In addition, such lower sets of cardinality k also enable the equivalence property $\|\cdot\|_{L^2_\nu(D)} \leq \|\cdot\|_{L^\infty} \leq k^\gamma \|\cdot\|_{L^2_\nu(D)}$ in arbitrary dimensions d with, e.g., $\gamma = 2$ for the uniform measure and $\gamma = \frac{\log 3}{\log 2}$ for Chebyshev measure.

Rather than best k-term approximation, we now consider best k-term approximation in a lower set. Hence, we replace Σ_k with

$$\Sigma_{k,L} = \left\{ \boldsymbol{c} \in \ell^2(\mathbb{N}_0^d) : |\mathrm{supp}(\boldsymbol{c})| \leq k, \ \mathrm{supp}(\boldsymbol{c}) \text{is lower} \right\},$$

and $\sigma_k(\boldsymbol{c})_1$ with the quantity

$$\sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{w}} = \inf_{\boldsymbol{d} \in \Sigma_{k,L}} \|\boldsymbol{c} - \boldsymbol{d}\|_{1,\boldsymbol{w}}. \tag{5}$$

Here $\boldsymbol{w} = (w_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}_0^d}$ is a sequence of positive weights and $\|\boldsymbol{c}\|_{1,\boldsymbol{w}} = \sum_{\boldsymbol{i} \in \mathbb{N}_0^d} w_{\boldsymbol{i}} |c_{\boldsymbol{i}}|$ is the norm on $\ell^1_{\boldsymbol{w}}(\mathbb{N}_0^d)$.

Remark 1. Sparsity in lower sets is an example of a so-called *structured sparsity* model. Specifically, $\Sigma_{k,L}$ is the subset of Σ_k corresponding to the union of all k-dimensional subspaces defined by lower sets:

$$\Sigma_{k,L} \equiv \bigcup_{\substack{|S|=k \\ S \text{ lower}}} \{\boldsymbol{c} : \mathrm{supp}(\boldsymbol{c}) \subseteq S\} \subset \bigcup_{|S|=k} \{\boldsymbol{c} : \mathrm{supp}(\boldsymbol{c}) \subseteq S\} \equiv \Sigma_k.$$

Structured sparsity models have been studied extensively in compressed sensing (see, e.g., [5, 9, 30, 33, 66] and references therein). There are a variety of general approaches for exploiting such structure, including greedy and iterative methods (see, for example, [5]) and convex relaxations [66]. A difficulty with lower set sparsity is that projections onto $\Sigma_{k,L}$ cannot be easily computed [22], unlike the case of Σ_k. Therefore, in this chapter we shall opt for a different approach based on $\ell^1_{\boldsymbol{w}}$ minimization with suitably chosen weights $\boldsymbol{w}$. See Section 4 for some further discussion.

3 Compressed Sensing for Multivariate Polynomial Approximation

Having introduced tensor orthogonal polynomials as a good basis for obtaining (structured) sparse representation of smooth, multivariate functions, we now turn our attention to computing quasi-optimal approximations of such a function f from the measurements $\{f(z_i)\}_{i=1}^m$.

It is first necessary to choose the sampling points $z_1, \dots, z_m$. From now on, following an approach that has become standard in compressed sensing [35], we shall assume that these points are drawn randomly and independently according to the probability measure ν. We remark in passing that this may not be the best choice in practice. However, such an approach yields recovery guarantees with measurement conditions that are essentially independent of d, thus mitigating the curse of dimensionality. In Section 4 we briefly discuss other strategies for choosing these points which may convey some practical advantages.

3.1 Exploiting Lower Set-Structured Sparsity

Let $\boldsymbol{c} \in \ell^2(\mathbb{N}_0^d)$ be the infinite vector of coefficients of a function $f \in L^2_\nu(D)$. Suppose that $\Lambda \subset \mathbb{N}_0^d$, $|\Lambda| = n$ and notice that

$$\boldsymbol{y} = A\boldsymbol{c}_\Lambda + \boldsymbol{e}_\Lambda, \tag{6}$$

where $\boldsymbol{y} \in \mathbb{C}^m$ and $A \in \mathbb{C}^{m\times n}$ are the finite vector and matrix given by

$$\boldsymbol{y} = \frac{1}{\sqrt{m}} \left(f(z_j)\right)_{j=1}^m, \qquad A = \frac{1}{\sqrt{m}} \left(\phi_{i_k}(z_j)\right)_{j,k=1}^{m,n}, \tag{7}$$

respectively, and

$$\boldsymbol{e}_\Lambda = \frac{1}{\sqrt{m}} \left(f(z_j) - f_\Lambda(z_j)\right)_{j=1}^m = \frac{1}{\sqrt{m}} \left(\sum_{i\notin\Lambda} c_i \phi_i(z_j) \right)_{j=1}^m, \tag{8}$$

is the vector of remainder terms corresponding to the coefficients c_i with indices outside Λ. Our aim is to approximate $\boldsymbol{c}$ up to an error depending on $\sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{w}}$, i.e., its best k-term approximation in a lower set (see (5)). In order for this to be possible, it is necessary to choose Λ so that it contains all lower sets of cardinality k. A straightforward choice is to make Λ exactly equal to the union of all such sets, which transpires to be precisely the hyperbolic cross index set with index k. That is,

$$\bigcup_{\substack{|S|=k \\ S \text{ lower}}} S = \left\{ \boldsymbol{i} = (i_1, \dots, i_d) \in \mathbb{N}_0^d : \prod_{j=1}^{d} (i_j + 1) \leq k \right\} = \Lambda_k^{\mathrm{HC}}. \tag{9}$$

It is interesting to note that this union is a finite set, due to the lower set assumption. Had one not enforced this additional property, the union would be infinite and equal to the whole space $\mathbb{N}_0^d$.

We shall assume that $\Lambda = \Lambda_k^{\mathrm{HC}}$ from now on. For later results, it will be useful to know the cardinality of this set. While an exact formula in terms of k and d is unknown, there are a variety of different upper and lower bounds. In particular, we shall make use of the following result:

$$n = \left| \Lambda_k^{\mathrm{HC}} \right| \leq \min \left\{ 2k^3 4^d, \mathrm{e}^2 k^{2+\log_2(d)} \right\}. \tag{10}$$

See [17, Thm. 3.7] and [45, Thm. 4.9], respectively.

With this in hand, we now wish to obtain a solution $\hat{\boldsymbol{c}}_\Lambda$ of (6) which approximates $\boldsymbol{c}_\Lambda$, and therefore $\boldsymbol{c}$ due to the choice of Λ, up to an error determined by its best approximation in a lower set of size k. We shall do this by weighted ℓ^1 minimization. Let $\boldsymbol{w} = (w_{\boldsymbol{i}})_{\boldsymbol{i} \in \Lambda}$ be a vector of positive weights and consider the problem

$$\min_{\boldsymbol{d} \in \mathbb{C}^n} \|\boldsymbol{d}\|_{1,\boldsymbol{w}} \text{ s.t. } \|\boldsymbol{y} - A\boldsymbol{d}\|_2 \leq \eta, \tag{11}$$

where $\|\boldsymbol{d}\|_{1,\boldsymbol{w}} = \sum_{j=1}^{n} w_{i_j} |d_{i_j}|$ is the weighted ℓ^1-norm and $\eta \geq 0$ is a parameter that will be chosen later. Since the weights $\boldsymbol{w}$ are positive we shall without loss of generality assume that

$$w_{\boldsymbol{i}} \geq 1, \quad \forall \boldsymbol{i}.$$

Our choice of these weights is based on the desire to exploit the lower set structure. Indeed, since lower sets inherently penalize higher indices, it is reasonable (and will turn out to be the case) that appropriate choices of increasing weights will promote this type of structure.

For simplicity, we shall assume for the moment that η is chosen so that

$$\eta \geq \|\boldsymbol{e}_\Lambda\|_2. \tag{12}$$

In particular, this implies that the exact vector $\boldsymbol{c}_\Lambda$ is a feasible point of the problem (11). As was already mentioned in Section 1.4, this assumption is a strong one and is unreasonable for practical scenarios where good *a priori* estimates on the expansion error $f - f_\Lambda$ are hard to obtain. In Section 3.6 we address the removal of this condition.

3.2 Choosing the Optimization Weights: Nonuniform Recovery

Our first task is to determine a good choice of optimization weights. For this, techniques from nonuniform recovery[1] are particularly useful.

At this stage it is convenient to define the following. First, for a vector of weights $\boldsymbol{w}$ and a subset S we let

$$|S|_{\boldsymbol{w}} = \sum_{i \in S} w_i^2, \tag{13}$$

be the *weighted* cardinality of S. Second, for the orthonormal basis $\{\phi_i\}_{i \in \mathbb{N}_0^d}$ we define the *intrinsic* weights $\boldsymbol{u} = (u_i)_i$ as

$$u_i = \|\phi_i\|_{L^\infty}. \tag{14}$$

Note that $u_i = \|\phi_i\|_{L^\infty} \geq \|\phi_i\|_{L^2_\nu} = 1$ since ν is a probability measure. With this in hand, we now have the following result (see [1, Thm. 6.1]):

Theorem 2. *Let $\Lambda \subset \mathbb{N}_0^d$ with $|\Lambda| = n$, $0 < \varepsilon < \mathrm{e}^{-1}$, $\eta \geq 0$, $\boldsymbol{w} = (w_i)_{i \in \Lambda}$ be a set of weights, $\boldsymbol{c} \in \ell^2(\mathbb{N}_0^d)$ and $S \subseteq \Lambda$, $S \neq \emptyset$, be any fixed set. Draw $z_1, \ldots, z_m$ independently according to the measure ν, let A, $\boldsymbol{y}$ and $\boldsymbol{e}_\Lambda$ be as in (7) and (8), respectively, and suppose that η satisfies (12). Then, with probability at least $1 - \varepsilon$, any minimizer $\hat{\boldsymbol{c}}_\Lambda$ of (11) satisfies*

$$\|\boldsymbol{c} - \hat{\boldsymbol{c}}_\Lambda\|_2 \lesssim \lambda \sqrt{|S|_{\boldsymbol{w}}}\, (\eta + \|\boldsymbol{c} - \boldsymbol{c}_\Lambda\|_{1,\boldsymbol{u}}) + \|\boldsymbol{c} - \boldsymbol{c}_S\|_{1,\boldsymbol{w}}, \tag{15}$$

provided

$$m \gtrsim \left(|S|_{\boldsymbol{u}} + \max_{i \in \Lambda \setminus S} \{u_i^2 / w_i^2\} |S|_{\boldsymbol{w}} \right) L, \tag{16}$$

where $\lambda = 1 + \frac{\sqrt{\log(\varepsilon^{-1})}}{\log(2n\sqrt{|S|_{\boldsymbol{w}}})}$ and $L = \log(\varepsilon^{-1}) \log\left(2n\sqrt{|S|_{\boldsymbol{w}}}\right)$.

Suppose for simplicity that $\boldsymbol{c}$ were exactly sparse and let $S = \mathrm{supp}(\boldsymbol{c})$ and $\eta = 0$. Then this result asserts exact recovery of $\boldsymbol{c}$, provided the measurement condition (16) holds. Ignoring the log factor L, this condition is determined by

$$\mathscr{M}(S; \boldsymbol{u}, \boldsymbol{w}) = |S|_{\boldsymbol{u}} + \max_{i \in \Lambda \setminus S} \{u_i^2 / w_i^2\} |S|_{\boldsymbol{w}}. \tag{17}$$

[1] By nonuniform recovery, we mean results that guarantee recovery of a fixed vector $\boldsymbol{c}_\Lambda$ from a single realization of the random matrix A. Conversely, uniform recovery results consider recovery of all sparse (or structured sparse) vectors from a single realization of A. See, for example, [35] for further discussion.

The first term is the weighted cardinality of S with respect to the intrinsic weights $\boldsymbol{u}$ and is independent of the choice of optimization weights $\boldsymbol{w}$. The second term depends on these weights, but the possibly large size of $|S|_{\boldsymbol{w}}$ is compensated by the factor $\max_{\boldsymbol{i}\in\Lambda\setminus S}\{u_{\boldsymbol{i}}^2/w_{\boldsymbol{i}}^2\}$.

Seeking to minimize $\mathscr{M}(S;\boldsymbol{u},\boldsymbol{w})$, it is natural to choose the weights $\boldsymbol{w}$ so that the second term in (17) is equal to the first. This is easily achieved by the choice

$$w_{\boldsymbol{i}} = u_{\boldsymbol{i}}, \quad \forall \boldsymbol{i}, \tag{18}$$

with the resulting measurement condition being simply

$$m \gtrsim |S|_{\boldsymbol{u}} \log(\varepsilon^{-1}) \log(2n\sqrt{|S|_{\boldsymbol{u}}}). \tag{19}$$

From now on, we primarily consider the weights (18).

Remark 2. Theorem 2 is a nonuniform recovery guarantee for weighted ℓ^1 minimization. Its proof uses the well-known golfing scheme [36], following similar arguments to those given in [4, 15] for unweighted ℓ^1 minimization. Unlike the results in [4, 15], however, it gives a measurement condition in terms of a fixed set S, rather than the sparsity k (or weighted sparsity). In other words, no sparsity (or structured sparsity) model is required at this stage. Such an approach was first pursued in [8] in the context of block sampling in compressed sensing. See also [23].

3.3 Comparison with Oracle Estimators

As noted above, the condition (19) does not require S to be a lower set. In Section 3.4 we shall use this property in order to estimate $|S|_{\boldsymbol{u}}$ in terms of the sparsity k. First, however, it is informative to compare (19) to the measurement condition of an oracle estimator. Suppose that the set S were known. Then a standard estimator for $\boldsymbol{c}$ is the least-squares solution

$$\check{\boldsymbol{c}}_S = (A_S)^{\dagger}\boldsymbol{y}, \tag{20}$$

where $A_S \in \mathbb{C}^{m\times|S|}$ is the matrix formed from the columns of A with indices belonging to S and $\dagger$ denotes the pseudoinverse. Stable and robust recovery via this estimator follows if the matrix A_S is well-conditioned. For this, one has the following well-known result:

Proposition 1. *Let $0 < \delta, \varepsilon < 1$, $S \subset \mathbb{N}_0^d$, $|S| = k$ and suppose that m satisfies*

$$m \gtrsim \delta^{-2}|S|_{\boldsymbol{u}} \log(2k\varepsilon^{-1}). \tag{21}$$

Draw $z_1, \dots, z_m$ independently according to the measure ν and let A be as in (7). Then, with probability at least $1 - \varepsilon$, the matrix A_S satisfies

$$\|(A_S)^* A_S - I\|_2 \leq \delta,$$

where $I \in \mathbb{C}^{k \times k}$ is the identity matrix and $\|\cdot\|_2$ is the matrix 2-norm.

See, for example, [1, Lem. 8.2]. Besides the log factor, (21) is the same sufficient condition as (19). Thus the weighted ℓ^1 minimization estimator $\hat{\boldsymbol{c}}_\Lambda$ with weights $\boldsymbol{w} = \boldsymbol{u}$ requires roughly the same measurement condition as the oracle least-squares estimator. Of course, the former requires no *a priori* knowledge of S.

Remark 3. In fact, one may prove a slightly sharper estimate than (21) where $|S|_{\boldsymbol{u}}$ is replaced by the quantity

$$\sup_{z \in D} \sum_{i \in S} |\phi_i(z)|^2. \tag{22}$$

See, for example, [24]. Note that $\sum_{i \in S} |\phi_i(z)|^2$ is the so-called Christoffel function of the subspace spanned by the functions $\{\phi_i\}_{i \in S}$. However, (22) coincides with $|S|_{\boldsymbol{u}}$ whenever the polynomials ϕ_i achieve their absolute maxima at the same point in D. This is the case for any Jacobi polynomials whenever the parameters satisfy $\max\{\alpha, \beta\} \geq -1/2$ [63, Thm. 7.32.1]; in particular, Legendre and Chebyshev polynomials (see Example 1), and tensor products thereof.

3.4 Sample Complexity for Lower Sets

The measurement condition (19) determines the sample complexity in terms of the weighted cardinality $|S|_{\boldsymbol{u}}$ of the set S. When a structured sparsity model is applied to S – in particular, lower set sparsity – one may derive estimates for $|S|_{\boldsymbol{u}}$ in terms of just the cardinality $k = |S|$ and the dimension d.

Lemma 1. *Let $2 \leq k \leq 2^{d+1}$. If $\{\phi_i\}_{i \in \mathbb{N}_0^d}$ is the tensor Chebyshev basis then*

$$k^{\log(3)/\log(2)}/3 \leq \max\left\{|S|_{\boldsymbol{u}} : S \subset \mathbb{N}_0^d,\ |S| \leq k, S \text{ lower}\right\} \leq k^{\log(3)/\log(2)},$$

where $|S|_{\boldsymbol{u}}$ and $\boldsymbol{u}$ are as in (13) and (14), respectively. If $\{\phi_i\}_{i \in \mathbb{N}_0^d}$ is the tensor Legendre basis then

$$k^2/4 \leq \max\left\{|S|_{\boldsymbol{u}} : S \subset \mathbb{N}_0^d,\ |S| \leq k, S \text{ lower}\right\} \leq k^2.$$

Moreover, the upper estimates hold for all $k \geq 2$.

See [19, Lem. 3.7]. With this in hand, we now have the following result:

Theorem 3. *Consider the setup in Theorem 2 with $k \geq 2$, $\Lambda = \Lambda_k^{\mathrm{HC}}$ the hyperbolic cross (9), weights $\boldsymbol{w} = \boldsymbol{u}$, and $\{\phi_i\}_{i\in\mathbb{N}_0^d}$ the tensor Legendre or Chebyshev basis. Then any minimizer $\hat{\boldsymbol{c}}_\Lambda$ of (11) with weights $\boldsymbol{w} = \boldsymbol{u}$ satisfies*

$$\|\boldsymbol{c} - \hat{\boldsymbol{c}}_\Lambda\|_2 \lesssim \lambda k^{\gamma/2} \left(\eta + \|\boldsymbol{c} - \boldsymbol{c}_\Lambda\|_{1,\boldsymbol{u}}\right) + \sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}},$$

with probability at least $1 - \varepsilon$, provided

$$m \gtrsim k^\gamma \log(\varepsilon^{-1}) \min\left\{d + \log(k), \log(2d)\log(k)\right\},$$

where $\lambda = 1 + \frac{\sqrt{\log(\varepsilon^{-1})}}{\log(k)}$ and where $\gamma = \log(3)/\log(2)$ or $\gamma = 2$ in the Chebyshev or Legendre case, respectively.

Proof. Let $S \subset \mathbb{N}_0^d$, $|S| \leq k$ be a lower set such that $\|\boldsymbol{c} - \boldsymbol{c}_S\|_{1,\boldsymbol{u}} = \sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}}$. By Lemma 1 we have $|S|_{\boldsymbol{u}} \leq k^\gamma$. We now apply Theorem 2 with $\boldsymbol{w} = \boldsymbol{u}$, and use this result and the bound (10) for $n = |\Lambda_k^{\mathrm{HC}}|$. □

Remark 4. It is worth noting that the lower set assumption drastically reduces the sample complexity. Indeed, for the case of Chebyshev polynomials one has

$$\max\left\{|S|_{\boldsymbol{u}} : S \subset \mathbb{N}_0^d,\ |S| \leq k\right\} = 2^d k.$$

In other words, in the absence of the lower set condition, one can potentially suffer exponential blow-up with dimension d. Note that this result follows straightforwardly from the explicit expression for the weights $\boldsymbol{u}$ in this case: namely,

$$u_{\boldsymbol{i}} = 2^{\|\boldsymbol{i}\|_0/2}, \tag{23}$$

where $\|\boldsymbol{i}\|_0 = |\left\{j : i_j \neq 0\right\}|$ for $\boldsymbol{i} = (i_1, \ldots, i_d) \in \mathbb{N}_0^d$ (see, e.g., [1]). On the other hand, for Legendre polynomials the corresponding quantity is infinite, since in this case the weights

$$u_{\boldsymbol{i}} = \prod_{j=1}^{d} \sqrt{2i_j + 1}, \tag{24}$$

are unbounded. Moreover, even if S is constrained to lie in the hyperbolic cross $\Lambda = \Lambda_k^{\mathrm{HC}}$, one still has a worst-case estimate that is polynomially large in k [22].

Remark 5. We have considered only tensor Legendre and Chebyshev polynomial bases. However, Theorem 3 readily extends to other types of orthogonal polynomials. All that is required is an upper bound for

$$\max\left\{|S|_{\boldsymbol{u}} : S \subset \mathbb{N}_0^d,\ |S| \leq k, S \text{ lower}\right\},$$

in terms of the sparsity k. For example, suppose that $\{\phi_{\boldsymbol{i}}\}_{\boldsymbol{i}\in\mathbb{N}_0^d}$ is the tensor ultraspherical polynomial basis, corresponding to the measure

$$\mathrm{d}\nu = (c_\alpha)^d \prod_{j=1}^{d} (1-z_j^2)^\alpha \,\mathrm{d}\boldsymbol{z}, \qquad c_\alpha = \left(\int_{-1}^{1} (1-z^2)^\alpha \,\mathrm{d}z \right)^{-1}.$$

If the parameter α satisfies $2\alpha + 1 \in \mathbb{N}$ then [49, Thm. 8] gives that

$$\max\left\{ |S|_{\boldsymbol{u}} : S \subset \mathbb{N}_0^d,\ |S| \leq k, S \text{ lower} \right\} \leq k^{2\alpha+2}.$$

This result includes the Legendre case ($\alpha = 0$) given in Lemma 1, as well as the case of Chebyshev polynomials of the second kind ($\alpha = 1/2$). A similar result also holds for tensor Jacobi polynomials for parameters $\alpha, \beta \in \mathbb{N}_0$ (see [49, Thm. 9]).

3.5 Quasi-Optimal Approximation: Uniform Recovery

As is typical of a nonuniform recovery guarantee, the error bound in Theorem 3 has the limitation that it relates the ℓ^2-norm of the error with the best k-term, lower approximation error in the $\ell^1_{\boldsymbol{u}}$-norm. To obtain stronger estimates we now consider uniform recovery techniques.

We first require an extension of the standard restricted isometry property (RIP) to the case of sparsity in lower sets. To this end, for $k \in \mathbb{N}$ we now define the quantity

$$s(k) = \max\left\{ |S|_{\boldsymbol{u}} : S \subset \mathbb{N}_0^d,\ |S| \leq k, S \text{ lower} \right\}. \tag{25}$$

The following extension of the RIP was introduced in [22]:

Definition 2. A matrix $A \in \mathbb{C}^{m\times n}$ has the lower restricted isometry property (lower RIP) of order k if there exists as $0 < \delta < 1$ such that

$$(1-\delta)\,\|\boldsymbol{c}\|_2^2 \leq \|A\boldsymbol{c}\|_2^2 \leq (1+\delta)\,\|\boldsymbol{c}\|_2^2, \quad \forall \boldsymbol{c} \in \mathbb{C}^n,\ |\mathrm{supp}(\boldsymbol{c})|_{\boldsymbol{u}} \leq s(k).$$

If $\delta = \delta_{k,L}$ is the smallest constant such that this holds, then $\delta_{k,L}$ is the k^{th} lower restricted isometry constant (lower RIC) of A.

We shall use the lower RIP to establish stable and robust recovery. For this, we first note that the lower RIP implies a suitable version of the robust null space property (see [22, Prop. 4.4]):

Lemma 2. *Let $k \geq 2$ and $A \in \mathbb{C}^{m\times n}$ satisfy the lower RIP of order αk with constant*

$$\delta = \delta_{\alpha k,L} < 1/5,$$

where $\alpha = 2$ if the weights $\boldsymbol{u}$ arise from the tensor Legendre basis and $\alpha = 3$ if the weights arise from the tensor Chebyshev basis. Then for any $S \subseteq \Lambda_s^{\mathrm{HC}}$ with $|S|_{\boldsymbol{u}} \leq s(k)$ and any $\boldsymbol{d} \in \mathbb{C}^n$,

$$\|\boldsymbol{d}_S\|_2 \leq \frac{\rho}{\sqrt{s(k)}} \|\boldsymbol{d}_{S^c}\|_{1,\boldsymbol{u}} + \tau \|A\boldsymbol{d}\|_2,$$

where $\rho = \frac{4\delta}{1-\delta} < 1$ and $\tau = \frac{\sqrt{1+\delta}}{1-\delta}$.

With this in hand, we now establish conditions under which the lower RIP holds for matrices A defined in (7). The following result was shown in [22]:

Theorem 4. *Fix $0 < \varepsilon < 1$, $0 < \delta < 1/13$, let $\{\phi_i\}_{i \in \mathbb{N}_0^d}$ be as in Section 2.1 and $\boldsymbol{u}$ be as in (14) and suppose that*

$$m \gtrsim \frac{s(k)}{\delta^2} L,$$

where $s(k)$ is as in (25) and

$$L = \log\left(\frac{s(k)}{\delta^2}\right)\left(\frac{1}{\delta^4}\log\left(2\frac{s(k)}{\delta^2}\log\left(\frac{s(k)}{\delta^2}\right)\right)\log(2n) + \frac{1}{\delta}\log\left(\frac{1}{\gamma\delta}\log\left(\frac{k(s)}{\delta^2}\right)\right)\right).$$

Draw $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ independently according to ν and let $A \in \mathbb{C}^{m \times n}$ be as in (7). Then with probability at least $1 - \varepsilon$, the matrix A satisfies the lower RIP of order k with constant $\delta_{k,L} \leq 13\delta$.

Combining this with the previous lemma now gives the following uniform recovery guarantee:

Theorem 5. *Let $0 < \varepsilon < 1$, $k \geq 2$ and*

$$m \asymp k^{\gamma} L, \tag{26}$$

where $\gamma = \log(3)/\log(2)$ or $\gamma = 2$ in the tensor Chebyshev or tensor Legendre cases, respectively, and

$$L = \left(\log^2(k)\min\{d + \log(k), \log(2d)\log(k)\} + \log(k)\log(\log(k)/\varepsilon)\right). \tag{27}$$

Let $\Lambda = \Lambda_k^{\mathrm{HC}}$ be the hyperbolic cross index set, $\{\phi_i\}_{i \in \mathbb{N}_0^d}$ be the tensor Legendre or Chebyshev polynomial basis and draw $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ independently according to the corresponding measure ν. Then with probability at least $1 - \varepsilon$ the following holds. For any $f \in L^2(D) \cap L^{\infty}(D)$ the approximation

$$\tilde{f} = \sum_{i \in \Lambda} \hat{c}_i \phi_i,$$

where $\hat{\boldsymbol{c}}_\Lambda = (\hat{c}_i)_{i\in\Lambda}$ *is a solution of (11) with A,* $\boldsymbol{y}$ *and* η *given by (7) and (12), respectively, and weights* $\boldsymbol{w} = \boldsymbol{u}$*, satisfies*

$$\left\|f - \tilde{f}\right\|_{L^\infty} \leq \|\boldsymbol{c} - \hat{\boldsymbol{c}}_\Lambda\|_{1,\boldsymbol{u}} \lesssim \sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + k^{\gamma/2}\eta, \tag{28}$$

and

$$\left\|f - \tilde{f}\right\|_{L^2_v} = \|\boldsymbol{c} - \hat{\boldsymbol{c}}_\Lambda\|_2 \lesssim \frac{\sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}}}{k^{\gamma/2}} + \eta, \tag{29}$$

where $\boldsymbol{c} \in \ell^2(\mathbb{N}_0^d)$ *are the coefficients of f in the basis* $\{\phi_i\}_{i\in\mathbb{N}_0^d}$*.*

Proof. Let $\alpha = 2$ or $\alpha = 3$ in the Legendre or Chebyshev case, respectively. Condition (26), Lemma 1 and Theorem 4 imply that the matrix A satisfies the lower RIP of order αk with constant $\delta_{\alpha k,L} \leq 1/6 < 1/5$. Now let S be a lower set of cardinality $|S| = k$ such that

$$\|\boldsymbol{c} - \boldsymbol{c}_S\|_{1,\boldsymbol{u}} = \sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}}, \tag{30}$$

set $\boldsymbol{d} = \boldsymbol{c}_\Lambda - \hat{\boldsymbol{c}}_\Lambda$ and $T = \Lambda\backslash S$. Note that

$$\begin{aligned}\|\boldsymbol{d}_T\|_{1,\boldsymbol{u}} &\leq \|\boldsymbol{c}_T\|_{1,\boldsymbol{u}} + \|\hat{\boldsymbol{c}}_T\|_{1,\boldsymbol{u}} \\ &= 2\|\boldsymbol{c}_T\|_{1,\boldsymbol{u}} + \|\boldsymbol{c}_S\|_{1,\boldsymbol{u}} + \|\hat{\boldsymbol{c}}_T\|_{1,\boldsymbol{u}} - \|\boldsymbol{c}_\Lambda\|_{1,\boldsymbol{u}} \\ &\leq 2\|\boldsymbol{c}_T\|_{1,\boldsymbol{u}} + \|\boldsymbol{d}_S\|_{1,\boldsymbol{u}} + \|\hat{\boldsymbol{c}}_\Lambda\|_{1,\boldsymbol{u}} - \|\boldsymbol{c}_\Lambda\|_{1,\boldsymbol{u}} \leq 2\sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + \|\boldsymbol{d}_S\|_{1,\boldsymbol{u}},\end{aligned}$$

since $\hat{\boldsymbol{c}}_\Lambda$ is a solution of (11) and $\boldsymbol{c}_\Lambda$ is feasible for (11) due to the choice of η. By Lemma 2 we have

$$\|\boldsymbol{d}_T\|_{1,\boldsymbol{u}} \leq 2\sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + \sqrt{s(k)}\|\boldsymbol{d}_S\|_2 \leq 2\sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + \rho\|\boldsymbol{d}_T\|_{1,\boldsymbol{u}} + \tau\sqrt{s(k)}\|A\boldsymbol{d}\|_2,$$

where $\rho \leq 4/5$ and $\tau \leq \sqrt{42}/5$. Therefore

$$\|\boldsymbol{d}_T\|_{1,\boldsymbol{u}} \lesssim \sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + \sqrt{s(k)}\|A\boldsymbol{d}\|_2 \lesssim \sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + \sqrt{s(k)}\eta,$$

where in the second step we use the fact that $\boldsymbol{d} = \boldsymbol{c}_\Lambda - \hat{\boldsymbol{c}}_\Lambda$ is the difference of two vectors that are both feasible for (11). Using this bound and Lemma 2 again gives

$$\|\boldsymbol{d}\|_{1,\boldsymbol{u}} \lesssim \sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + \sqrt{s(k)}\eta, \tag{31}$$

and since $\boldsymbol{c} - \hat{\boldsymbol{c}}_\Lambda = \boldsymbol{d} + \boldsymbol{c} - \boldsymbol{c}_\Lambda$, we deduce that

$$\|\boldsymbol{c} - \hat{\boldsymbol{c}}_\Lambda\|_{1,\boldsymbol{u}} \leq \|\boldsymbol{d}\|_{1,\boldsymbol{u}} + \|\boldsymbol{c} - \boldsymbol{c}_\Lambda\|_{1,\boldsymbol{u}} \lesssim \sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + \sqrt{s(k)}\eta. \tag{32}$$

Due to the definition of the weights $\boldsymbol{u}$, we have $\|f-\tilde{f}\|_{L^\infty} \leq \|\boldsymbol{c}-\hat{\boldsymbol{c}}_\Lambda\|_{1,\boldsymbol{u}}$, and therefore, after noting that $s(k) \lesssim k^\gamma$ (see Lemma 1) we obtain the first estimate (28). For the second estimate let S be such that

$$\|\boldsymbol{c}-\boldsymbol{c}_S\|_2 = \min\{\|\boldsymbol{c}-\boldsymbol{d}\|_2 : |\mathrm{supp}(\boldsymbol{d})|_{\boldsymbol{u}} \leq s(k)\},$$

and set $T = S^c$. Let $\boldsymbol{d} = \boldsymbol{c} - \hat{\boldsymbol{c}}_\Lambda$ and write $\|\boldsymbol{d}\|_2 \leq \|\boldsymbol{d}_S\|_2 + \|\boldsymbol{d}_T\|_2$. Via a weighted Stechkin estimate [61, Thm. 3.2] we have $\|\boldsymbol{d}_T\|_2 \leq \frac{1}{\sqrt{s(k)-\|\boldsymbol{u}\|_\infty}}\|\boldsymbol{d}\|_{1,\boldsymbol{u}}$. For tensor Chebyshev and Legendre polynomials, one has $\|\boldsymbol{u}\|_\infty \leq \frac{3}{4}s(k)$ (see [22, Lem. 4.1]), and therefore $\|\boldsymbol{d}_T\|_2 \lesssim \frac{1}{\sqrt{s(k)}}\|\boldsymbol{d}\|_{1,\boldsymbol{u}}$. We now apply Lemma 2 to deduce that $\|\boldsymbol{d}\|_2 \lesssim \frac{1}{\sqrt{s(k)}}\|\boldsymbol{d}\|_{1,\boldsymbol{u}} + \eta$. Recall that $s(k) \gtrsim k^\gamma$ due to Lemma 1. Hence (32) now gives $\|\boldsymbol{d}\|_2 \lesssim \frac{\sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}}}{k^{\gamma/2}} + \eta$, as required. □

For the Legendre and Chebyshev cases, Theorem 5 proves recovery with quasi-optimal k-term rates of approximation subject to the same measurement condition (up to log factors) as the oracle least-squares estimator. In particular, the sample complexity is polynomial in k and at most logarithmic in the dimension d, thus mitigating the curse of dimensionality to a substantial extent. We remark in passing that this result can be extended to general Jacobi polynomials (recall Remark 5). Furthermore, the dependence on d can be removed altogether by considering special classes of lower sets, known as *anchored* sets [29].

3.6 Unknown Errors, Robustness, and Interpolation

A drawback of the main results so far (Theorems 3 and 5) is that they assume the *a priori* bound (12), i.e.

$$\frac{1}{m}\sum_{j=1}^m |f(\boldsymbol{z}_j) - f_\Lambda(\boldsymbol{z}_j)|^2 \leq \eta^2, \tag{33}$$

for some known η. Note that this is implied by the slightly stronger condition

$$\|f - f_\Lambda\|_{L^\infty} \leq \eta.$$

Such an η is required in order to formulate the optimization problem (11) to recover f. Moreover, in view of the error bounds in Theorems 3 and 5, one expects a poor estimation of η to yield a larger recovery error. Another drawback of the current approach is that the approximation $\tilde{f}$ does not interpolate f, a property which is sometimes desirable in applications.

We now consider the removal of the condition (12). This follows the work of [3, 12]. To this end, let $\eta \geq 0$ be arbitrary, i.e., (33) need not hold, and consider the minimization problem

$$\min_{\boldsymbol{d}\in\mathbb{C}^n} \|\boldsymbol{d}\|_{1,\boldsymbol{u}} \text{ s.t. } \|\boldsymbol{y} - A\boldsymbol{d}\|_2 \leq \eta. \tag{34}$$

If $\hat{\boldsymbol{c}}_\Lambda = (\hat{c}_i)_{i\in\Lambda}$ is a minimizer of this problem, we define, as before, the corresponding approximation

$$\tilde{f} = \sum_{i\in\Lambda} \hat{c}_i \phi_i.$$

Note that if $\eta = 0$ then $\tilde{f}$ exactly interpolates f at the sample points $\{\boldsymbol{z}_j\}_{j=1}^m$.

An immediate issue with the minimization problem (34) is that the truncated vector of coefficients $\boldsymbol{c}_\Lambda$ is not generally feasible. Indeed, $\boldsymbol{y} - A\boldsymbol{c}_\Lambda = \boldsymbol{e}_\Lambda$, where $\boldsymbol{e}_\Lambda$ is as in (8) and is generally nonzero. In fact, is not even guaranteed that the feasibility set of (34) is nonempty. However, this will of course be the case whenever A has full rank m. Under this assumption, one then has the following (see [3]):

Theorem 6. *Let ε, k, m, γ, Λ, $\{\phi_i\}_{i\in\mathbb{N}_0^d}$ and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ be as in Theorem 5. Then with probability at least $1-\varepsilon$ the following holds. For any $\eta \geq 0$ and $f \in L^2(D)\cap L^\infty(D)$ the approximation*

$$\tilde{f} = \sum_{i\in\Lambda} \hat{c}_i \phi_i,$$

where $\hat{\boldsymbol{c}}_\Lambda = (\hat{c}_i)_{i\in\Lambda}$ is a solution of (34) with A and $\boldsymbol{y}$ given by (7) satisfies

$$\left\|f - \tilde{f}\right\|_{L^\infty} \leq \|\boldsymbol{c} - \hat{\boldsymbol{c}}_\Lambda\|_{1,\boldsymbol{u}} \lesssim \sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + k^{\gamma/2}\left(\eta + \|\boldsymbol{e}_\Lambda\|_2 + T_{\boldsymbol{u}}(A,\Lambda,\boldsymbol{e}_\Lambda,\eta)\right) \tag{35}$$

and

$$\left\|f - \tilde{f}\right\|_{L^2_v} = \|\boldsymbol{c} - \hat{\boldsymbol{c}}_\Lambda\|_2 \lesssim \frac{\sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}}}{k^{\gamma/2}} + \eta + \|\boldsymbol{e}_\Lambda\|_2 + T_{\boldsymbol{u}}(A,\Lambda,\boldsymbol{e}_\Lambda,\eta), \tag{36}$$

where $\boldsymbol{c} \in \ell^2(\mathbb{N}_0^d)$ are the coefficients of f in the basis $\{\phi_i\}_{i\in\mathbb{N}_0^d}$, $\boldsymbol{e}_\Lambda$ is as in (8) and

$$T_{\boldsymbol{u}}(A,\Lambda,\boldsymbol{e}_\Lambda,\eta) = \min\left\{\frac{\|\boldsymbol{d}\|_{1,\boldsymbol{u}}}{k^{\gamma/2}} : \boldsymbol{d}\in\mathbb{C}^n,\ \|A\boldsymbol{d} - \boldsymbol{e}_\Lambda\|_2 \leq \eta\right\}. \tag{37}$$

Proof. We follow the steps of the proof of Theorem 5 with some adjustments to take into account the fact that $\boldsymbol{c}_\Lambda$ may not be feasible. First, let S be such that (30) holds and set $\boldsymbol{d} = \boldsymbol{c}_\Lambda - \hat{\boldsymbol{c}}_\Lambda$ and $T = \Lambda\backslash S$. Then, arguing in a similar way we see that

$$\begin{aligned}\|\boldsymbol{d}_T\|_{1,\boldsymbol{u}} &\leq 2\sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + \|\boldsymbol{d}_S\|_{1,\boldsymbol{u}} + \|\hat{\boldsymbol{c}}_\Lambda\|_{1,\boldsymbol{u}} - \|\boldsymbol{c}_\Lambda\|_{1,\boldsymbol{u}} \\ &\leq 2\sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + \|\boldsymbol{d}_S\|_{1,\boldsymbol{u}} + \|\boldsymbol{g} - \boldsymbol{c}_\Lambda\|_{1,\boldsymbol{u}},\end{aligned}$$

where $\boldsymbol{g}\in\mathbb{C}^n$ is any point in the feasible set of (34). By Lemma 2 we have

$$\|\boldsymbol{d}_T\|_{1,\boldsymbol{u}} \leq 2\sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + \rho\|\boldsymbol{d}_T\|_{1,\boldsymbol{u}} + \tau\sqrt{s(k)}\|A\boldsymbol{d}\|_2 + \|\boldsymbol{g}-\boldsymbol{c}_\Lambda\|_{1,\boldsymbol{u}}.$$

Notice that $\|A\boldsymbol{d}\|_2 = \|\boldsymbol{y}-\boldsymbol{e}_\Lambda - A\hat{\boldsymbol{c}}_\Lambda\|_2 \leq \|\boldsymbol{e}_\Lambda\|_2 + \eta$, and therefore

$$\|\boldsymbol{d}_T\|_{1,\boldsymbol{u}} \lesssim \sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + \sqrt{s(k)}\,(\|\boldsymbol{e}_\Lambda\|_2 + \eta) + \|\boldsymbol{g}-\boldsymbol{c}_\Lambda\|_{1,\boldsymbol{u}}.$$

Hence, by similar arguments, it follows that

$$\|\boldsymbol{c}-\boldsymbol{c}_\Lambda\|_{1,\boldsymbol{u}} \lesssim \sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}} + k^{\gamma/2}\,(\|\boldsymbol{e}_\Lambda\|_2 + \eta) + \|\boldsymbol{g}-\boldsymbol{c}_\Lambda\|_{1,\boldsymbol{u}}, \tag{38}$$

for any feasible point $\boldsymbol{g}$. After analogous arguments, we also deduce the following bound in the ℓ^2-norm:

$$\|\boldsymbol{c}-\boldsymbol{c}_\Lambda\|_2 \lesssim \frac{\sigma_{k,L}(\boldsymbol{c})_{1,\boldsymbol{u}}}{k^{\gamma/2}} + (\|\boldsymbol{e}_\Lambda\|_2 + \eta) + k^{-\gamma/2}\|\boldsymbol{g}-\boldsymbol{c}_\Lambda\|_{1,\boldsymbol{u}}. \tag{39}$$

To complete the proof, we consider the term $\|\boldsymbol{g}-\boldsymbol{c}_\Lambda\|_{1,\boldsymbol{u}}$. Write $\boldsymbol{g} = \boldsymbol{c}_\Lambda + \boldsymbol{g}'$ and notice that $\boldsymbol{g}$ is feasible if and only if $\boldsymbol{g}'$ satisfies $\|A\boldsymbol{g}' - \boldsymbol{e}_\Lambda\| \leq \eta$. Since $\boldsymbol{g}'$ is arbitrary we get the result. □

The two error bounds (35) and (36) in this theorem are analogous to (28) and (29) in Theorem 5. They remove the condition $\eta \geq \|\boldsymbol{e}_\Lambda\|_2$ at the expense of an additional term $T_{\boldsymbol{u}}(A, \Lambda, \boldsymbol{e}_\Lambda, \eta)$. We now provide a bound for this term (see [3]):

Theorem 7. *Consider the setup of Theorem 6, and let $T_{\boldsymbol{u}}(A, \Lambda, \boldsymbol{e}_\Lambda, \eta)$ be as in (37). If A has full rank, then*

$$T_{\boldsymbol{u}}(A, \Lambda, \boldsymbol{e}_\Lambda, \eta) \leq \frac{k^{\alpha/2}\sqrt{L}}{\sigma_{\min}\left(\sqrt{\frac{m}{n}}A^*\right)} \max\left\{\|\boldsymbol{e}_\Lambda\|_2 - \eta, 0\right\}, \tag{40}$$

where L is as in (27) and $\alpha = 1, 2$ in the Chebyshev or Legendre cases, respectively.

Proof. If $\eta \geq \|\boldsymbol{e}_\Lambda\|_2$ then the result holds trivially. Suppose now that $\eta < \|\boldsymbol{e}_\Lambda\|_2$. Since $\|\boldsymbol{e}_\Lambda\|_2 \neq 0$ in this case, we can define $\boldsymbol{d} = (1 - \eta/\|\boldsymbol{e}_\Lambda\|_2)A^\dagger\boldsymbol{e}_\Lambda$, where $A^\dagger$ denotes the pseudoinverse. Then $\boldsymbol{d}$ satisfies $\|A\boldsymbol{d} - \boldsymbol{e}_\Lambda\|_2 = \eta$, and therefore

$$k^{\gamma/2}T_{\boldsymbol{u}}(A, \Lambda, \boldsymbol{e}_\Lambda, \eta) \leq \|d\|_{1,\boldsymbol{u}} \leq \sqrt{|\Lambda|_{\boldsymbol{u}}}\|\boldsymbol{d}\|_2 \leq \frac{\sqrt{|\Lambda|_{\boldsymbol{u}}}}{\sigma_{\min}(A^*)}\,(\|\boldsymbol{e}_\Lambda\|_2 - \eta).$$

Equation (26) implies that $\sqrt{\frac{m}{k^\gamma}} \lesssim \sqrt{L}$, and hence

$$T_{\boldsymbol{u}}(A, \Lambda, \boldsymbol{e}_\Lambda, \eta) \lesssim \sqrt{\frac{|\Lambda|_{1,\boldsymbol{u}}}{n}}\frac{\sqrt{L}}{\sigma_{\min}\left(\sqrt{\frac{m}{n}}A^*\right)}\,(\|\boldsymbol{e}_\Lambda\|_2 - \eta). \tag{41}$$

It remains to estimate $|\Lambda|_{1,\boldsymbol{u}}$. For the Chebyshev case, we apply (23) to give

$$|\Lambda|_{1,\boldsymbol{u}} = \sum_{\boldsymbol{i}\in\Lambda} 2^{\|\boldsymbol{i}\|_0} \leq \sum_{\boldsymbol{i}\in\Lambda} \prod_{j=1}^{d} (i_j+1) \leq k \sum_{\boldsymbol{i}\in\Lambda} 1 = kn$$

where in the penultimate step we used the definition of the hyperbolic cross (9). For the Legendre case, we use (24) to get

$$|\Lambda|_{1,\boldsymbol{u}} = \sum_{\boldsymbol{i}\in\Lambda} \prod_{j=1}^{d} (2i_j+1) \leq \sum_{\boldsymbol{i}\in\Lambda} 2^{\|\boldsymbol{i}\|_0} \prod_{j=1}^{d} (i_j+1) \leq k^2 n.$$

This completes the proof. □

The error bound (40) suggests that the effect of removing the condition $\eta \geq \|\boldsymbol{e}_\Lambda\|_2$ is at most a small algebraic factor in k, a log factor and term depending on the minimal singular value of the scaled matrix $\sqrt{\frac{m}{n}}A^*$. We discuss this latter term further in below. Interestingly, this bound suggests that a good estimate of $\|\boldsymbol{e}_\Lambda\|_2$ (when available) can reduce this error term. Indeed, one has $T_{\boldsymbol{u}}(A,\Lambda,\boldsymbol{e}_\Lambda,\eta) \to 0$ linearly in $\|\boldsymbol{e}_\Lambda\|_2 - \eta \to 0^+$. Hence estimation procedures aiming to tune η – for example, cross validation (see Section 3.7) – are expected to yield reduced error over the case $\eta = 0$, for example.

It is beyond the scope of this chapter to provide theoretical bounds on the minimal singular value of the scaled matrix $\sqrt{\frac{m}{n}}A^*$. We refer to [12] for a more comprehensive treatment of such bounds. However, we note that it is reasonable to expect that $\sigma_{\min}(\sqrt{\frac{m}{n}}A^*) \approx 1$ under appropriate conditions on m and n. Indeed:

Lemma 3. *Let $B = \mathbb{E}\left(\frac{m}{n}AA^*\right)$, where A is the matrix of Theorem 6. Then the minimal eigenvalue of B is precisely $1 - 1/n$.*

Proof. We have $\mathbb{E}\left(\frac{m}{n}AA^*\right)_{j,l} = \mathbb{E}\left(\frac{1}{n}\sum_{\boldsymbol{i}\in\Lambda}\phi_{\boldsymbol{i}}(\boldsymbol{z}_j)\phi_{\boldsymbol{i}}(\boldsymbol{z}_l)\right)$. When $l = j$ this gives $\mathbb{E}\left(\frac{m}{n}AA^*\right)_{j,j} = 1$. Conversely, since $\{\phi_{\boldsymbol{i}}\}_{\boldsymbol{i}\in\mathbb{N}_0^d}$ are orthogonal polynomials one has $\int_D \phi_{\boldsymbol{i}}(\boldsymbol{z})\,\mathrm{d}\nu = \langle\phi_{\boldsymbol{i}},\phi_{\boldsymbol{0}}\rangle_{L^2_\nu} = \delta_{\boldsymbol{i},\boldsymbol{0}}$, and therefore for $l \neq j$ one has $\mathbb{E}\left(\frac{m}{n}AA^*\right)_{j,l} = \frac{1}{n}\sum_{\boldsymbol{i}\in\Lambda}\left(\int_D \phi_{\boldsymbol{i}}(\boldsymbol{z})\,\mathrm{d}\nu\right)^2 = \frac{1}{n}$. It is now a straightforward calculation to show that $\lambda_{\min}(B) = 1 - 1/n$. □

Remark 6. Although complete theoretical estimates $T_{\boldsymbol{u}}(A,\Lambda,\boldsymbol{e}_\Lambda,\eta)$ are outside the scope of this work, it is straightforward to derive a bound that can be computed. Indeed, it follows immediately from (41) that

$$T_{\boldsymbol{u}}(A,\Lambda,\boldsymbol{e}_\Lambda,\eta) \lesssim Q_{\boldsymbol{u}}(A)\sqrt{L}\max\{\|\boldsymbol{e}_\Lambda\|_2 - \eta, 0\},$$

where

$$Q_{\boldsymbol{u}}(A) = \sqrt{\frac{|\Lambda|_{1,\boldsymbol{u}}}{n}}\frac{1}{\sigma_{\min}\left(\sqrt{\frac{m}{n}}A^*\right)}. \tag{42}$$

Hence, up to the log factor, the expected robustness of (34) can be easily checked numerically. See Section 3.7 for some examples of this approach.

Remark 7. For pedagogical reasons, we have assumed the truncation of f to f_Λ is the only source of error $\boldsymbol{e}_\Lambda$ affecting the measurements $\boldsymbol{y}$ (recall (8)). There is no reason for this to be the case, and $\boldsymbol{e}_\Lambda$ may incorporate other errors without changing any of the above results. We note that concrete applications often give rise to other sources of unknown error. For example, in UQ, we usually aim at approximating a function of the form $f(\boldsymbol{z}) = q(u(\boldsymbol{z}))$, where $u(\boldsymbol{z})$ is the solution to a PDE depending on some random coefficients $\boldsymbol{z}$ and q is a quantity of interest (see, e.g., [32, 73]). In this case, each sample $f(\boldsymbol{z}_j)$ is typically subject to further sources of inaccuracy, such as the numerical error associated with the PDE solver employed to compute $u(\boldsymbol{z}_j)$ (e.g., a finite element method) and, possibly, the error committed evaluating q on $u(\boldsymbol{z}_j)$ (e.g., numerical integration).

Remark 8. Our analysis based on the estimation of the tail error (37) can be compared with the robustness analysis of basis pursuit based on the so-called *quotient property* [35]. However, this analysis is limited to the case of *basis pursuit*, corresponding to the optimization program (34) with $\boldsymbol{u} = \mathbf{1}$ (i.e., unweighted ℓ^1 norm) and $\eta = 0$. In the context of compressed sensing, random matrices that are known to fulfill the quotient property with high probability are gaussian, subgaussian, and Weibull matrices [34, 70]. For further details we refer to [12].

3.7 Numerical Results

We conclude this chapter with a series of numerical results. First, in Figures 1 and 2 we show the approximation of several functions via weighted ℓ^1 minimization. Weights of the form $w_i = (u_i)^\alpha$ are used for several different choices of α. In agreement with the discussion in Section 3.2, the choice $\alpha = 1$, i.e., $w_i = u_i$ generally gives among the smallest error. Moreover, while larger values of α sometime give a smaller error, this is not the case for all functions. Notice that in all cases unweighted ℓ^1 minimization gives a worse error than weighted ℓ^1 minimization. As is to be expected, the improvement offered by weighted ℓ^1 minimization in the Chebyshev case is less significant in moderate dimensions than for Legendre polynomials.

The results in Figures 1 and 2 were computed by solving weighted ℓ^1 minimization problems with η set arbitrarily to $\eta = 10^{-12}$ (we make this choice rather than $\eta = 0$ to avoid potential infeasibility issues in the solver). In particular, the condition (33) is not generally satisfied. Following Remark 6, we next assess the size of the constant $Q_{\boldsymbol{u}}(A)$ defined in (42). Table 1 shows the magnitude of this constant for the setups considered in Figures 1 and 2. Over all ranges of m considered, this constant is never more than 20 in magnitude. That is to say, the additional effect due to the unknown truncation error $\|\boldsymbol{e}_\Lambda\|_2$ is relatively small.

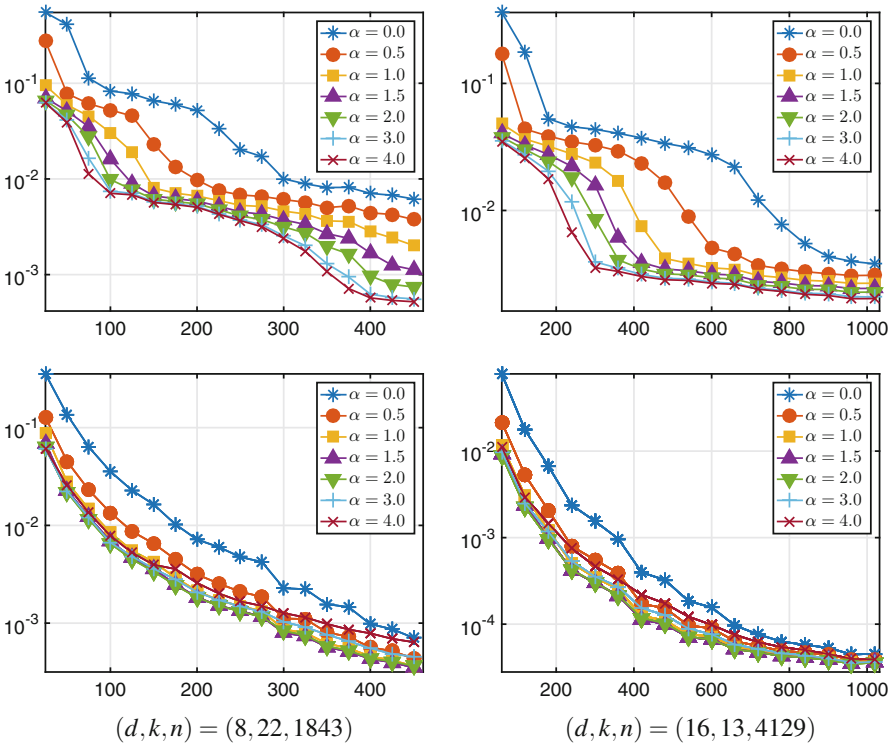


Fig. 1 The error $\|f - \widetilde{f}\|_{L^\infty}$ (averaged over 50 trials) against m for Legendre polynomials. Here $\widetilde{f} = \sum_{i \in \Lambda} \hat{c}_i \phi_i$, where $\hat{\boldsymbol{c}}_\Lambda$ is a solution of (11) with weights $w_i = (u_i)^\alpha$ and $\Lambda = \Lambda_k^{\mathrm{HC}}$ a hyperbolic cross index set. The functions used were $f(\boldsymbol{y}) = \prod_{k=d/2+1}^{d} \cos(16 y_k / 2^k) / \prod_{k=1}^{d/2} (1 - y_k/4^k)$ and $f(\boldsymbol{y}) = \exp\left(-\sum_{k=1}^{d} y_k/(2d)\right)$ (top and bottom, respectively). The weighted ℓ^1 minimization problem was solved using the SPGL1 package [67, 68] with a maximum of 100,000 iterations and $\eta = 10^{-12}$.

In view of Remark 7, in Figure 3 we assess the performance of weighted ℓ^1 minimization in the presence of external sources of error corrupting the measurements. In order to model this scenario, we consider the problem (11) where the vector of measurements is corrupted by additive noise

$$\boldsymbol{y} = \frac{1}{\sqrt{m}} (f(\boldsymbol{z}_j))_{j=1}^{m} + \boldsymbol{n}, \tag{43}$$

or, equivalently, by recalling (6),

$$\boldsymbol{y} = A \boldsymbol{c}_\Lambda + \boldsymbol{e}_\Lambda + \boldsymbol{n}. \tag{44}$$

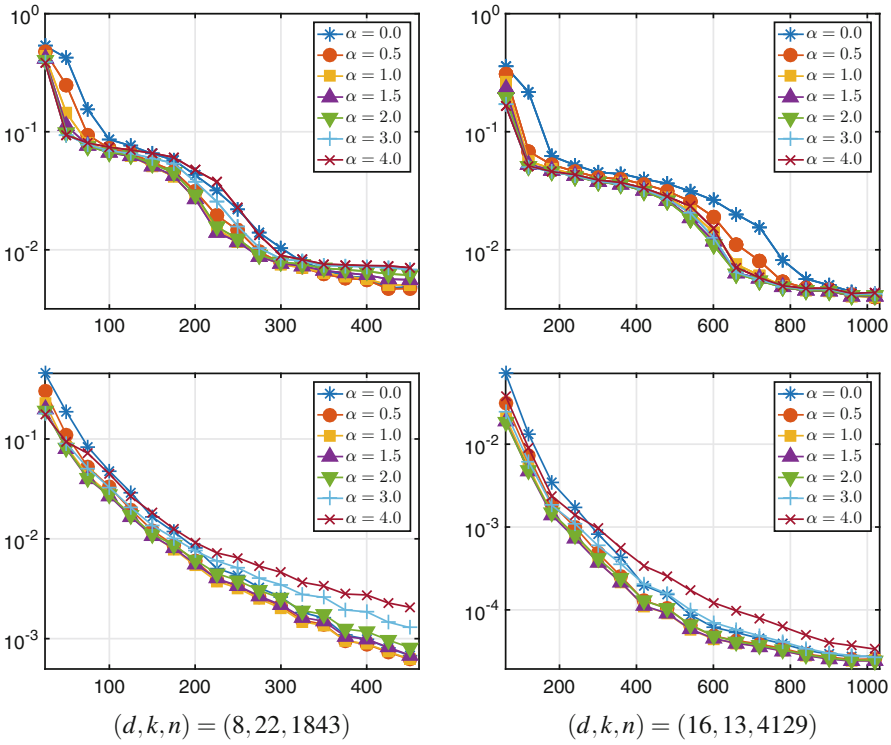


Fig. 2 The same as Figure 1 but with Chebyshev polynomials.

Table 1 The constant $Q_u(A)$ (averaged over 50 trials) for the setup considered in Figures 1 and 2.

	m	125	250	375	500	625	750	875	1000
$(d,k,n) = (8,22,1843)$	Chebyshev	2.65	3.07	3.53	3.95	4.46	5.03	5.78	6.82
	Legendre	6.45	7.97	8.99	10.5	12.1	13.7	15.8	18.6
	m	250	500	750	1000	1250	1500	1750	2000
$(d,k,n) = (16,13,4129)$	Chebyshev	2.64	2.93	3.30	3.63	3.99	4.41	4.95	5.62
	Legendre	5.64	6.20	6.85	7.60	8.32	8.99	10.1	11.1

We randomly generate the noise as $\boldsymbol{n} = 10^{-3}\boldsymbol{g}/\|\boldsymbol{g}\|_2$, where $\boldsymbol{g} \in \mathbb{R}^m$ is a standard random gaussian vector, so that $\|\boldsymbol{n}\|_2 = 10^{-3}$. Considering weights $\boldsymbol{w} = (u_i^\alpha)_{i\in\Lambda}$, with $\alpha = 0, 1$, we compare the error obtained when the parameter η in (11) is chosen according to each of the following three strategies:

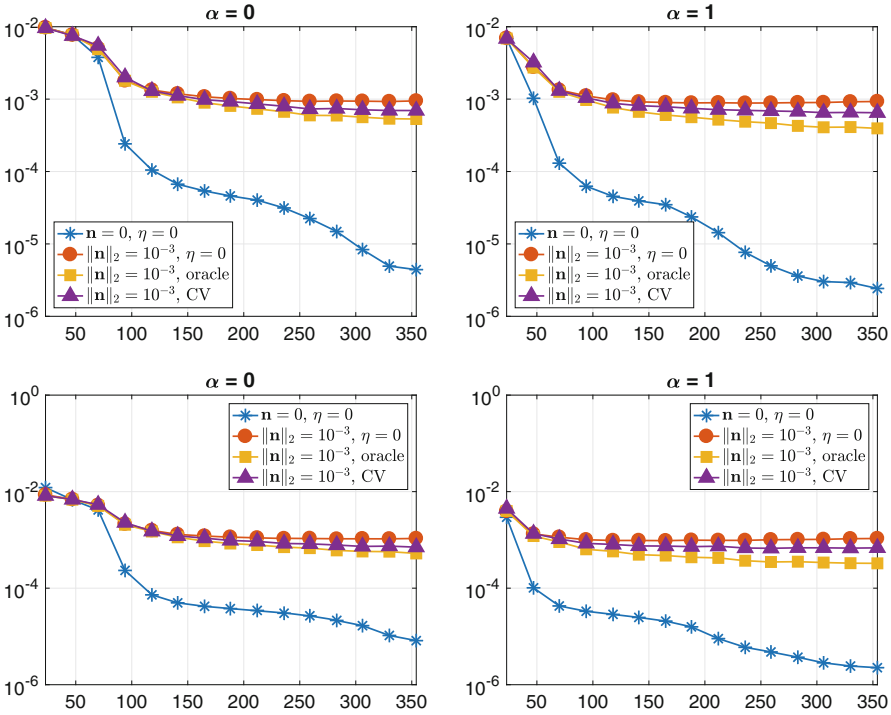


Fig. 3 The error $\|f - \tilde{f}\|_{L^2_\nu}$ against m. Here $\tilde{f} = \sum_{i\in\Lambda} \hat{c}_i \phi_i$, where $\hat{\boldsymbol{c}}_\Lambda$ is a solution of (11) with weights $\boldsymbol{w} = (u_i^\alpha)_{i\in\Lambda}$, with $\alpha = 0$ (left) and $\alpha = 1$ (right), and $\boldsymbol{y}$ defined as in (43). Regarding $\{\phi_i\}_{i\in\Lambda}$ and ν, the Chebyshev polynomials with the Chebyshev measure are employed in the top line and the Legendre polynomials with the uniform measure in the bottom line. We choose $d = 8$ and $\Lambda = \Lambda_{19}^{\mathrm{HC}}$ with $n = |\Lambda| = 1771$. For each value of m, we average the error over 50 trials considering three different strategies for the choice of η: namely, $\eta = 0$, estimation via oracle least squares, and cross validation (CV). The function approximated is $f(\boldsymbol{y}) = \exp\left(-\sum_{k=1}^d \cos(y_k)/(8d)\right)$.

1. $\eta = 0$, corresponding to enforcing the exact constraint $A\boldsymbol{d} = \boldsymbol{y}$ in (11);
2. $\eta = \eta_{oracle} = \|A\hat{\boldsymbol{c}}_{oracle} - \boldsymbol{y}\|_2$, where $f_{oracle} = \sum_{i\in\Lambda} (\hat{c}_{oracle})_i \phi_i$ is the oracle least-squares solution based on $10n$ random samples of f distributed according to ν;
3. η is estimate using a cross validation approach, as described in [32, Section 3.5], where the search of η is restricted to the values of the form $10^k \cdot \eta_{oracle}$, where k belongs to a uniform grid of 11 equispaced points on the interval $[-3, 3]$, $3/4$ of the samples are used as reconstruction samples and $1/4$ as validation samples.

The results are in accordance with the estimate (36). Indeed, as expected, for any value of α, the recovery error associated with $\boldsymbol{n} = \boldsymbol{0}$ and $\eta = 0$ is always lower than the recovery error associated with $\boldsymbol{n} \neq \boldsymbol{0}$ and any choice of η. This can be

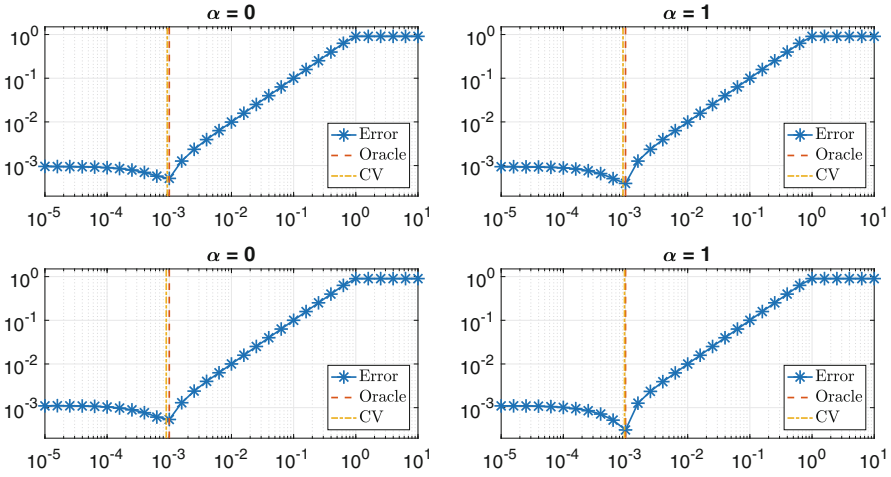


Fig. 4 Recovery error $\|f - \widetilde{f}\|_{L^2_\nu}$ (averaged over 50 trials) against η, in the same setting as in Figure 3. We use Chebyshev and Legendre polynomials in the top and bottom rows, respectively. We consider $\eta = 10^k$, with k belonging to a uniform grid of 31 points on the interval $[-5, 1]$. The vertical lines represent the estimated values of η (averaged over 50 trials) based on oracle least squares (red-dashed line) and cross validation (yellow dashed-dotted line). The weights are chosen as $\boldsymbol{w} = (\boldsymbol{u}_i^\alpha)_{i\in\Lambda}$, with $\alpha = 0$ (left) and $\alpha = 1$ (right).

explained by the fact that, in the right-hand side of (36), the terms $\sigma_{k,L}(\boldsymbol{c})/k^{\gamma/2}$ and $\|\boldsymbol{e}_\Lambda\|_2$ are dominated by $\eta + T_{\boldsymbol{u}}(A, \Lambda, \boldsymbol{e}_\Lambda, \eta)$ when $\boldsymbol{n} \neq \boldsymbol{0}$. Moreover, estimating η via oracle least squares (strategy 2) gives better results than cross validation (strategy 3), which in turn is better than the neutral choice $\eta = 0$ (strategy 1). Finally, we note that the discrepancy among the three strategies is accentuated as α gets larger.

In the next experiment we highlight the importance of the parameter η when solving (11) with measurements subject to external sources of error (recall Remark 7). We corrupt the measurements by adding random noise $\boldsymbol{n}$ with norm $\|\boldsymbol{n}\|_2 = 10^{-3}$, analogously to (43). Then, for different values of η from 10^{-5} to 10, we solve (11) with weights $\boldsymbol{w} = (\boldsymbol{u}_i^\alpha)_{i\in\Lambda}$ and $\alpha = 0, 1$. The resulting recovery errors with respect to the L^2_ν norm (averaged over 50 trials) are plotted as a function of η in Figure 4. For every value of α, the resulting curve is constant for the smallest and largest values of η. In between, the curve exhibits a global minimum, which corresponds to an optimal calibration of η. The values of η estimated via oracle least squares and cross validation are both able to approximate the global minimum on average. However, cross validation has a larger standard deviation compared to the former (see Table 2). This explains why the performance of cross validation is suboptimal in Figure 3. We also notice that the global minimum is more pronounced as α gets larger, in accordance to the observations in Figure 3.

Table 2 Mean ± standard deviation for the values of η estimated via oracle least squares and cross validation over 50 trials in Figure 4.

α	Chebyshev		Legendre	
	Oracle	Cross validation	Oracle	Cross validation
0	1.0e−03 ± 7.2e−09	9.3e−04 ± 3.8e−04	1.0e−03 ± 3.7e−09	9.0e−04 ± 4.0e−04
1	1.0e−03 ± 4.9e−09	9.1e−04 ± 4.0e−04	1.0e−03 ± 3.6e−09	9.7e−04 ± 3.6e−04

4 Conclusions and Challenges

The concern of this chapter has been the emerging topic of compressed sensing for high-dimensional approximation. As shown, smooth, multivariate functions are compressible in orthogonal polynomial bases. Moreover, their coefficients have a certain form of structured sparsity corresponding to so-called lower sets. The main result of this work is that such structure can be exploited via weighted ℓ^1-norm regularizers. Doing so leads to sample complexity estimates that are at most logarithmically dependent on the dimension d, thus mitigating the curse of dimensionality to a substantial extent.

As discussed in Section 1.5, this topic has garnered much interest over the last half a dozen years. Yet challenges remain. We conclude by highlighting a number of open problems in this area:

Unbounded domains We have considered only bounded hypercubes in this chapter. The case of unbounded domains presents additional issues. While Hermite polynomials (orthogonal on $\mathbb{R}$) have been considered in the case of unweighted ℓ^1 minimization in [39, 41, 53], the corresponding measurement conditions exhibit exponentially large factors in either the dimension d or degree k of the (total degree) index space used. It is unclear how to obtain dimension-independent measurement conditions in this setting, even for structured sparsity in lower sets.

Sampling strategies Throughout we have considered sampling i.i.d. according to the orthogonality measure of the basis functions. This is by no means the only choice, and various other sampling strategies have been considered in other works [39, 41, 44, 52, 53, 64, 71]. Empirically, several of these approaches are known to give some benefits. However, it is not known how to design sampling strategies which lead to better measurement conditions than those given in Theorem 5. A singular challenge is to design a sampling strategy for which m need only scale linearly with k. We note in passing that this has been achieved for the oracle least-squares estimator (recall Section 3.3) [28]. However, it is not clear how to extend this approach to a compressed sensing framework.

Alternatives to weighted ℓ^1 minimization. As discussed in Remark 1, lower set structure is a type of structured sparsity model. We have used weighted ℓ^1 minimization to promote such structure. Yet other approaches may convey benefits. Different, but related, types of structured sparsity have been exploited in the past

using greedy or iterative algorithms [5, 9, 30, 33], or by designing appropriate convex regularizers [66]. This remains an interesting problem for future work.

Recovering Hilbert-valued functions. We have focused on compressed sensing-based polynomial approximation of high-dimensional functions whose coefficients belong to the complex domain $\mathbb{C}$. However, an important problem in computational science, especially in the context of UQ and optimal control, involves the approximation of parametric PDEs. Current compressed sensing techniques proposed in literature [10, 22, 32, 47, 56, 59, 73] only approximate functionals of parameterized solutions, e.g., evaluation at a single spatial location, whereas a more robust approach should consider an ℓ_1-regularized problem involving Hilbert-valued signals, i.e., signals where each coordinate is a function in a Hilbert space, which can provide a direct, global reconstruction of the solutions in the entire physical domain. However, to achieve this goal new iterative minimization procedures as well as several theoretical concepts will need to be extended to the Hilbert space setting. The advantages of this approach over pointwise recovery with standard techniques will include: (i) for many parametric and stochastic model problems, global estimate of solutions in the physical domain is a quantity of interest; (ii) the recovery guarantees of this strategy can be derived from the decay of the polynomial coefficients in the relevant function space, which is well-known in the existing theory; and (iii) the global reconstruction only assumes *a priori* bounds of the tail expansion in energy norms, which are much more realistic than pointwise bounds.

Acknowledgements The first and second authors acknowledge the support of the Alfred P. Sloan Foundation and the Natural Sciences and Engineering Research Council of Canada through grant 611675. The second author acknowledges the Postdoctoral Training Center in Stochastics of the Pacific Institute for the Mathematical Sciences for the support. The third author acknowledges support by the US Defense Advanced Research Projects Agency, Defense Sciences Office under contract and award numbers HR0011619523 and 1868-A017-15; the US Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under contract number ERKJ259 and ERKJ314; and the Laboratory Directed Research and Development program at the Oak Ridge National Laboratory, which is operated by UT-Battelle, LLC., for the US Department of Energy under Contract DE-AC05-00OR22725.

References

1. B. Adcock, Infinite-dimensional compressed sensing and function interpolation. Found. Comput. Math., 1–41 (2017). https://doi.org/10.1007/s10208-017-9350-3
2. B. Adcock, Infinite-dimensional ℓ^1 minimization and function approximation from pointwise data. Constr. Approx. **45**(3), 345–390 (2017)
3. B. Adcock, A. Bao, S. Brugiapaglia, Correcting for unknown errors in sparse high-dimensional function approximation (2017). arXiv:1711.07622
4. B. Adcock, A.C. Hansen. Generalized sampling and infinite-dimensional compressed sensing. Found. Comput. Math. **16**(5), 1263–1323 (2016)

5. R.G. Baraniuk, V. Cevher, M.F. Duarte, C. Hedge, Model-based compressive sensing. IEEE Trans. Inform. Theory **56**(4), 1982–2001 (2010)
6. J. Beck, F. Nobile, L. Tamellini, R. Tempone, Convergence of quasi-optimal Stochastic Galerkin methods for a class of PDEs with random coefficients. Comput. Math. Appl. **67**(4), 732–751 (2014)
7. R.E. Bellman, *Adaptive Control Processes: A Guided Tour* (Princeton University Press, Princeton, 1961)
8. J. Bigot, C. Boyer, P. Weiss, An analysis of block sampling strategies in compressed sensing. IEEE Trans. Inform. Theory **64**(4), 2125–2139 (2016)
9. T. Blumensath, Sampling theorems for signals from the union of finite-dimensional linear subspaces. IEEE Trans. Inform. Theory **55**(4), 1872–1882 (2009)
10. J.-L. Bouchot, H. Rauhut, C. Schwab, Multi-level Compressed Sensing Petrov-Galerkin discretization of high-dimensional parametric PDEs (2017). arXiv:1701.01671
11. S. Brugiapaglia, COmpRessed SolvING: sparse approximation of PDEs based on compressed sensing, Ph.D. thesis, Politecnico di Milano, Milano, 2016
12. S. Brugiapaglia, B. Adcock, Robustness to unknown error in sparse regularization (2017). arXiv:1705.10299
13. S. Brugiapaglia, F. Nobile, S. Micheletti, S. Perotto, A theoretical study of compressed solving for advection-diffusion-reaction problems. Math. Comput. **87**(309), 1–38 (2018)
14. H.-J. Bungartz, M. Griebel, Sparse grids. Acta Numer. **13**, 147–269 (2004)
15. E.J. Candès, Y. Plan, A probabilistic and RIPless theory of compressed sensing. IEEE Trans. Inform. Theory **57**(11), 7235–7254 (2011)
16. E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inform. Theory **52**(1), 489–509 (2006)
17. A. Chernov, D. Dũng, New explicit-in-dimension estimates for the cardinality of high-dimensional hyperbolic crosses and approximation of functions having mixed smoothness. J. Compl. **32**, 92–121 (2016)
18. A. Chkifa, A. Cohen, R. DeVore, C. Schwab, Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs. Modél. Math. Anal. Numér. **47**(1), 253–280 (2013)
19. A. Chkifa, A. Cohen, G. Migliorati, F. Nobile, R. Tempone, Discrete least squares polynomial approximation with random evaluations – application to parametric and stochastic elliptic PDEs. ESAIM Math. Model. Numer. Anal. **49**(3), 815–837 (2015)
20. A. Chkifa, A. Cohen, C. Schwab, High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs. Found. Comput. Math. **14**(4), 601–633 (2014)
21. A. Chkifa, A. Cohen, C. Schwab, Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs. J. Math. Pures Appl. **103**, 400–428 (2015)
22. A. Chkifa, N. Dexter, H. Tran, C.G. Webster, Polynomial approximation via compressed sensing of high-dimensional functions on lower sets. Math. Comput. arXiv:1602.05823 (2016, to appear)
23. I.-Y. Chun, B. Adcock, Compressed sensing and parallel acquisition. IEEE Trans. Inform. Theory **63**(8), 4760–4882 (2017). arXiv:1601.06214
24. A. Cohen, M.A. Davenport, D. Leviatan, On the stability and accuracy of least squares approximations. Found. Comput. Math. **13**, 819–834 (2013)
25. A. Cohen, R. Devore Approximation of high-dimensional parametric PDEs. Acta Numer. **24**, 1–159 (2015)
26. A. Cohen, R.A. DeVore, C. Schwab, Convergence rates of best *N*-term Galerkin approximations for a class of elliptic sPDEs. Found. Comput. Math. **10**, 615–646 (2010)
27. A. Cohen, R.A. DeVore, C. Schwab, Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDEs. Anal. Appl. **9**(1), 11–47 (2011)
28. A. Cohen, G. Migliorati, Optimal weighted least-squares methods (2016). arXiv:1608.00512
29. A. Cohen, G. Migliorati, F. Nobile, Discrete least-squares approximations over optimized downward closed polynomial spaces in arbitrary dimension. Constr. Approx. **45**(3), 497–519 (2017)

30. M.A. Davenport, M.F. Duarte, Y.C. Eldar, G. Kutyniok, Introduction to compressed sensing, in *Compressed Sensing: Theory and Applications* (Cambridge University Press, Cambridge, 2011)
31. D.L. Donoho, Compressed sensing. IEEE Trans. Inform. Theory **52**(4), 1289–1306 (2006)
32. A. Doostan, H. Owhadi, A non-adapted sparse approximation of PDEs with stochastic inputs. J. Comput. Phys. **230**(8), 3015–3034 (2011)
33. M.F. Duarte, Y.C. Eldar, Structured compressed sensing: from theory to applications. IEEE Trans. Signal Process. **59**(9), 4053–4085 (2011)
34. S. Foucart, Stability and robustness of ℓ_1-minimizations with weibull matrices and redundant dictionaries. Linear Algebra Appl. **441**, 4–21 (2014)
35. S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Birkhauser, Basel, 2013)
36. D. Gross, Recovering low-rank matrices from few coefficients in any basis. IEEE Trans. Inform. Theory **57**(3), 1548–1566 (2011)
37. M. Gunzburger, C.G. Webster, G. Zhang, Stochastic finite element methods for partial differential equations with random input data. Acta Numer. **23**, 521–650 (2014)
38. M. Gunzburger, C.G. Webster, G. Zhang, Sparse collocation methods for stochastic interpolation and quadrature, in *Handbook of Uncertainty Quantification* (Springer, New York, 2016), pp. 1–46
39. L. Guo, A. Narayan, T. Zhou, Y. Chen, Stochastic collocation methods via L1 minimization using randomized quadratures. SIAM J. Sci. Comput. **39**(1), A333–A359 (2017). arXiv:1602.00995
40. J. Hampton, A. Doostan, Coherence motivated sampling and convergence analysis of least squares polynomial Chaos regression. Comput. Methods Appl. Mech. Eng. **290**, 73–97 (2015)
41. J. Hampton, A. Doostan, Compressive sampling of polynomial chaos expansions: convergence analysis and sampling strategies. J. Comput. Phys. **280**, 363–386 (2015)
42. V.H. Hoang, C. Schwab, Regularity and generalized polynomial chaos approximation of parametric and random 2nd order hyperbolic partial differential equations. Anal. Appl. **10**(3), 295–326 (2012)
43. J.D. Jakeman, M.S. Eldred, K. Sargsyan, Enhancing l_1-minimization estimates of polynomial chaos expansions using basis selection. J. Comput. Phys. **289**, 18–34 (2015). arXiv:1407.8093
44. J.D. Jakeman, A. Narayan, T. Zhou, A generalized sampling and preconditioning scheme for sparse approximation of polynomial chaos expansions. SIAM J. Sci. Comput. **39**(3), A1114–A1144 (2017). arXiv:1602.06879
45. T. Kühn, W. Sickel, T. Ullrich, Approximation of mixed order Sobolev functions on the d-torus: asymptotics, preasymptotics, and d-dependence. Constr. Approx. **42**(3), 353–398 (2015)
46. O.P. Le Maître, O.M. Knio, *Spectral Methods for Uncertainty Quantification* (Springer, New York, 2010)
47. L. Mathelin, K.A. Gallivan, A compressed sensing approach for partial differential equations with random input data. Commun. Comput. Phys. **12**(4), 919–954 (2012)
48. G. Migliorati, Polynomial approximation by means of the random discrete L^2 projection and application to inverse problems for PDEs with stochastic data, Ph.D. thesis, Politecnico di Milano, Milano, 2013
49. G. Migliorati, Multivariate Markov-type and Nikolskii-type inequalities for polynomials associated with downward closed multi-index sets. J. Approx. Theory **189**, 137–159 (2015)
50. G. Migliorati, F. Nobile, Analysis of discrete least squares on multivariate polynomial spaces with evaluations at low-discrepancy point sets. J. Complexity **31**(4), 517–542 (2015)
51. G. Migliorati, F. Nobile, E. von Schwerin, R. Tempone, Analysis of the discrete L^2 projection on polynomial spaces with random evaluations. Found. Comput. Math. **14**, 419–456 (2014)
52. A. Narayan, T. Zhou, Stochastic collocation on unstructured multivariate meshes. Commun. Comput. Phys. **18**(1), 1–36 (2015)
53. A. Narayan, J.D. Jakeman, T. Zhou, A Christoffel function weighted least squares algorithm for collocation approximations. Math. Comput. **86**(306), 1913–1947 (2014). arXiv:1412.4305

54. F. Nobile, R. Tempone, C.G. Webster, An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. **46**(5), 2411–2442 (2008)
55. F. Nobile, R. Tempone, C.G. Webster, A sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. **46**(5), 2309–2345 (2008)
56. J. Peng, J. Hampton, A. Doostan, A weighted ℓ_1-minimization approach for sparse polynomial chaos expansions. J. Comput. Phys. **267**, 92–111 (2014)
57. J. Peng, J. Hampton, A. Doostan, On polynomial chaos expansion via gradient-enhanced ℓ_1-minimization. J. Comput. Phys. **310**, 440–458 (2016)
58. H. Rauhut, Random sampling of sparse trigonometric polynomials. Appl. Comput. Harmon. Anal. **22**(1), 16–42 (2007)
59. H. Rauhut, C. Schwab, Compressive sensing Petrov-Galerkin approximation of high dimensional parametric operator equations. Math. Comput. **86**, 661–700 (2017)
60. H. Rauhut, R. Ward, Sparse Legendre expansions via ℓ_1-minimization. J. Approx. Theory **164**(5), 517–533 (2012)
61. H. Rauhut, R. Ward, Interpolation via weighted ℓ_1 minimization. Appl. Comput. Harmon. Anal. **40**(2), 321–351 (2016)
62. M.K. Stoyanov, C.G. Webster, A dynamically adaptive sparse grid method for quasi-optimal interpolation of multidimensional functions. Comput. Math. Appl. **71**(11), 2449–2465 (2016)
63. G. Szegö, *Orthogonal Polynomials* (American Mathematical Society, Providence, RI, 1975)
64. G. Tang, G. Iaccarino, Subsampled Gauss quadrature nodes for estimating polynomial chaos expansions. SIAM/ASA J. Uncertain. Quantif. **2**(1), 423–443 (2014)
65. H. Tran, C.G. Webster, G. Zhang, Analysis of quasi-optimal polynomial approximations for parameterized PDEs with deterministic and stochastic coefficients. Numer. Math. **137**(2), 451–493 (2017). arXiv:1508.01821
66. Y. Traonmilin, R. Gribonval, Stable recovery of low-dimensional cones in Hilbert spaces: one RIP to rule them all. Appl. Comput. Harm. Anal. (2017). https://doi.org/10.1016/j.acha.2016.08.004
67. E. van den Berg, M.P. Friedlander, SPGL1: a solver for large-scale sparse reconstruction (June 2007), http://www.cs.ubc.ca/labs/scl/spgl1
68. E. van den Berg, M.P. Friedlander, Probing the Pareto frontier for basis pursuit solutions. SIAM J. Sci. Comput. **31**(2), 890–912 (2008)
69. C.G. Webster, Sparse grid stochastic collocation techniques for the numerical solution of partial differential equations with random input data, Ph.D. thesis, Florida State University, Tallahassee, 2007
70. P. Wojtaszczyk, Stability and instance optimality for gaussian measurements in compressed sensing. Found. Comput. Math. **10**(1), 1–13 (2010)
71. Z. Xu, T. Zhou, On sparse interpolation and the design of deterministic interpolation points. SIAM J. Sci. Comput. **36**(4), 1752–1769 (2014)
72. L. Yan, L. Guo, D. Xiu, Stochastic collocation algorithms using ℓ_1-minimization. Int. J. Uncertain. Quantif. **2**(3), 279–293 (2012)
73. X. Yang, G.E. Karniadakis, Reweighted ℓ_1 minimization method for stochastic elliptic differential equations. J. Comput. Phys. **248**, 87–108 (2013)
74. X. Yang, H. Lei, N.A. Baker, G. Lin, Enhancing sparsity of Hermite polynomial expansions by iterative rotations. J. Comput. Phys. **307**, 94–109 (2016). arXiv:1506.04344

Multisection in the Stochastic Block Model Using Semidefinite Programming

Naman Agarwal, Afonso S. Bandeira, Konstantinos Koiliaris, and Alexandra Kolla

Abstract We consider the problem of identifying underlying community-like structures in graphs. Toward this end, we study the stochastic block model (SBM) on k-clusters: a random model on $n = km$ vertices, partitioned in k equal sized clusters, with edges sampled independently across clusters with probability q and within clusters with probability p, $p > q$. The goal is to recover the initial "hidden" partition of $[n]$. We study semidefinite programming (SDP)-based algorithms in this context. In the regime $p = \frac{\alpha \log(m)}{m}$ and $q = \frac{\beta \log(m)}{m}$, we show that a certain natural SDP-based algorithm solves the problem of *exact recovery* in the k-community SBM, with high probability, whenever $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{1}$, as long as $k = o(\log n)$. This threshold is known to be the information theoretically optimal. We also study the case when $k = \theta(\log(n))$. In this case however, we achieve recovery guarantees that no longer match the optimal condition $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{1}$, thus leaving achieving optimality for this range an open question.

Keywords Graph partitioning · Random models · Stochastic block model · Semidefinite programming · Dual certificate

N. Agarwal
Princeton University, Princeton, NJ, USA
e-mail: namana@cs.princeton.edu

A.S. Bandeira (✉)
Department of Mathematics, Courant Institute of Mathematical Sciences and Center for Data Science, New York University, New York, NY, USA
e-mail: bandeira@cims.nyu.edu

K. Koiliaris · A. Kolla
University of Illinois Urbana - Champaign, Urbana, IL, USA
e-mail: koiliar2@illinois.edu; akolla@illinois.edu

H. Boche et al. (eds.), *Compressed Sensing and its Applications*,
Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-69802-1_4

1 Introduction

Identifying underlying structure in graphs is a primitive question for scientists: can existing communities be located in a large graph? Is it possible to partition the vertices of a graph into strongly connected clusters? Several of these questions have been shown to be hard to answer, even approximately, so instead of looking for worst-case guarantees, attention has shifted toward average-case analyses. In order to study such questions, the usual approach is to consider a random [26] or a semi-random [19, 24] generative model of graphs and use it as a benchmark to test existing algorithms or to develop new ones. With respect to identifying underlying community structure, the stochastic block model (SBM) (or planted partition model) has, in recent times, been one of the most popular choices. Its growing popularity is largely due to the fact that its structure is simple to describe, but at the same time it has interesting and involved phase transition properties which have only recently been discovered [1, 2, 7, 14, 17, 21, 22, 27, 28, 30].

In this paper we consider the SBM on k-communities defined as follows. Let n be a multiple of m, $V = [n]$ be the set of vertices, and $P = \{P_i\}$ be a partition of them into k equal sized clusters each of size $m = \frac{n}{k}$. Construct a random graph G on V by adding an edge for any two vertices in the same cluster independently with probability p and any two vertices across distinct clusters independently with probability q where $p > q$. We will write $G \sim \mathcal{G}_{p,q,k}$ to denote that a graph G is generated from the above model. Given such a G, the goal is to recover (with high probability) the initial hidden partition P.

The SBM can be seen as an extension of the Erdős-Rényi random graph model [18] with the additional property of possessing a nontrivial underlying community structure (something which the Erdős-Rényi model lacks). This richer structure not only makes this model interesting to study theoretically but also renders it closer to real-world inputs, which tend to have a community structure. It is also worth noting that, as pointed out in [14], a slight generalization of the SBM encompasses several classical planted random graph problems including planted clique [4, 26], planted coloring [3], planted dense subgraph [5], and planted partition [11, 16, 19].

There are two natural problems that arise in the context of the SBM: *exact recovery*, where the aim is to recover the hidden partition completely, and *detection*, where the aim is to recover the partition better than what a random guess would achieve. In this paper we focus on exact recovery. Note that exact recovery necessarily requires the hidden clusters to be connected (since otherwise there would be no way to match the partitions in one component to another component), and it is easy to see that the threshold for connectivity occurs when $p = \Omega(\log(m)/m)$. Therefore, the right scale for the threshold behavior of the parameters p, q is $\Theta(\log(m)/m)$, which is what we consider in this paper.

In the case of two communities ($k = 2$), Abbe et al. [2] recently established a sharp phase transition phenomenon from information theoretic impossibility to computational feasibility of exact recovery. However, the existence of such a phenomenon in the case of $k > 2$ was left open until solved, for $k = O(1)$, in

independent parallel research [1, 22]. In this paper we resolve the above showing the existence of a sharp phase transition for $k = o(\log(n))$.

More precisely, in this work, we study a semidefinite programming (SDP)-based algorithm that, for $k = o(\log(n))$, recovers, for an optimal range of parameters, exactly the planted k-partition of $G \sim \mathscr{G}_{p,q,k}$ with high probability. The range of the parameters p, q is optimal in the following sense: it can be shown that this parameter range exhibits a sharp phase transition from information theoretic impossibility to computational feasibility through the SDP algorithm studied in this paper. An interesting aspect of our result is that, for $k = o(\log(n))$, the threshold is the same as for $k = 2$. This means that, even if an oracle reveals all of the cluster memberships except for two, the problem has essentially the same difficulty. We also consider the case when $k = \Theta(\log(n))$. Unfortunately, in this regime we can no longer guarantee exact recovery up to the proposed information theoretic threshold. Similar behavior was observed and reported by Chen et al. [14], and in our work we observe that the divergence between our information theoretic lower bound and our computational upper bound sets in at $k = \Theta(\log(n))$. This is formally summarized in the following theorems.

Theorem 1. *Given a graph $G \sim \mathscr{G}_{p,q,k}$ with $k = O(\log(m))$ hidden clusters each of size m and $p = \frac{\alpha \log(m)}{m}$ and $q = \frac{\beta \log(m)}{m}$, where $\alpha > \beta > 0$ are fixed constants, the semidefinite program (4), with probability $1 - n^{-\Omega(1)}$, recovers the clusters when:*

- *for $k = o(\log n)$, as long as*

$$\sqrt{\alpha} - \sqrt{\beta} > 1;$$

- *for $k = (\gamma + o(1)) \log(n)$ for a fixed γ, as long as*

$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{1 + c\sqrt{\beta\gamma}\left(1 + \log\left(\sqrt{\frac{\alpha}{\beta}}\right)\right)},$$

where c is a universal constant.

We complement the above theorem by showing the following lower bound which is a straightforward extension of the lower bound for $k = 2$ from [2].

Theorem 2. *Given a graph $G \sim \mathscr{G}_{p,q,k}$ with k hidden clusters each of size m where k is $o(m^{-\lambda})$ for any fixed $\lambda > 0$, if $p = \frac{\alpha \log(m)}{m}$ and $q = \frac{\beta \log(m)}{m}$, where $\alpha > \beta > 0$ are fixed constants, then it is information theoretically impossible to recover the clusters exactly with high probability if*

$$\sqrt{\alpha} - \sqrt{\beta} < 1\,.$$

Note that Theorem 2 establishes a sharp phase transition between computational feasibility and information theoretic impossibility when $k = o(\log(n))$. At $k \sim \log(n)$ we see that our lower and upper bounds diverge. We leave it as an open

problem to determine whether such divergence is necessary or a shortcoming of the SDP approach.

At the heart of our argument is the following theorem which establishes a sufficient condition for exact recovery with high probability.

Theorem 3. *Let $G \sim \mathcal{G}_{p,q,k}$, with probability $1 - n^{-\Omega(1)}$ over the choice of G, if the following condition is satisfied, the semidefinite program (4) recovers the hidden partition:*

$$\min_i \Delta(i) \geq \hat{c} \left(\sqrt{pn/k + qn} + q\sqrt{\frac{n}{k}\log(n)} + \sqrt{\log(n)} + \log(k) \right), \tag{1}$$

where $\hat{c}$ is a universal constant and $\Delta(i)$ is defined as the difference between the number of neighbors a vertex i has in its own cluster and the maximum number of neighbors it has in any other cluster (with respect to the hidden partition). In other words, with probability $1 - n^{-\Omega(1)}$, (1) implies exact recovery.

We are able to give sharp guarantees for the semidefinite programming algorithm based essentially on the behavior of inner and outer degrees of the vertices. This is achieved by constructing a candidate dual certificate and using bounds on the spectral norm of random matrices to show that the constructed candidate is indeed a valid one. The problem is then reduced to the easier task of understanding the typical values of such degrees. Remarkably, the conditions required for these quantities are very similar to the ones required for the problem to be information theoretically solvable (which essentially correspond to each node having larger in-degree than out-degree). This helps explain the optimality of our algorithm. The approach of reducing the validity of a dual certificate to conditions on an interpretable quantity appeared in [7] for a considerably simpler class of problems where the dual certificate construction is straightforward (which includes the stochastic block model for $k = 2$ but not $k > 2$). In contrast, in the current setting, the dual certificate construction is complex, rendering a different and considerably more involved analysis. Moreover, the estimates we need (both of spectral norms and of inner and outer degrees) do not fall under the class of the ones studied in [7].

We also show that our algorithm recovers the planted partitions exactly also in the presence of a monotone adversary, a semi-random model defined in [19].

1.1 Related Previous and Parallel Work

Graph partitioning problem has been studied over the years with various different objectives and guarantees. There has been significant recent literature concentration around the bipartition (bisection) and the general k-partition problems (multisection) in random and semi-random models [2, 14, 15, 17, 25, 27–32].

Some of the first results on partitioning random graphs were due to Bui et al. [12] who presented algorithms for finding bipartitions in dense graphs. Boppana [11] showed a spectral algorithm that for a large range of parameters recovers a planted bipartition in a graph. Feige and Kilian [19] present an SDP-based algorithm to solve

the problem of planted bipartition (along with the problems of finding independent sets and graph coloring). Independently, McSherry [26] gave a spectral algorithm that solved the problems of multisection, clique, and graph coloring.

More recently, a spate of results have established very interesting phase transition phenomena for SBMs, both for the case of *detection* and *exact* recovery. For the case of detection, where the aim is to recover partitions better than a random guess asymptotically, recent works of [25, 27, 28] established a striking sharp phase transition from information theoretic impossibility to computational feasibility for the case of $k = 2$. For the case of exact recovery, Abbe et al. [2], and independently [30], established the existence of a similar phase transition phenomenon albeit at a different parameter range. More recently the same phenomenon was shown to exist for a semidefinite programming relaxation, for $k = 2$ in [7, 21]. However, the works described above established phase transition for $k = 2$, and the case for larger k was left open. Our paper bridges the gap for larger k up to $o(\log(n))$ for the case of exact recovery. To put our work into context, the corresponding case of establishing such behavior for the problem of detection remains open. In fact, it is conjectured in [17, 27] that, for the detection problem, there exists a gap between the thresholds for computational feasibility and information theoretic impossibility for any k number of communities greater than 4. In this paper, we show that this is not the case for the exact recovery problem.

Chen et al. [14] also study the k-community SBM and provide convex programming-based algorithms and information theoretic lower bounds for exact recovery. Their results are similar to ours in the sense that they also conjecture a separation between information theoretic impossibility and computation feasibility as k grows. In comparison we focus strongly on the case of slightly superconstant k ($o(\log(n))$) and mildly growing k ($\Omega(\log(n))$) and show exact recovery to the optimal (even up to constants) threshold in the former case. Very recently in independent and parallel work, Abbe and Sandon [1] studied the problem of exact recovery for a fixed number of ($k > 2$) communities where the symmetry constraint (equality of cluster sizes and the probabilities of connection are same in different clusters) is removed. Our result, in contrast to theirs, is based on the integrality of a semidefinite relaxation, which has the added benefit of producing an explicit certificate for optimality (i.e., indeed when the solution is "integral," we know for sure that it is the optimal balanced k-partition). Abbe and Sandon [1] comment in their paper that their results can be extended for slightly superconstant k but leave it as future work. In another parallel and independent work, Hajek et al. [22] study semidefinite programming relaxations for exact recovery in SBMs and achieve similar results as ours. We remark that semidefinite program in consideration in [22] is the same as the semidefinite program (4) considered by us (up to an additive/multiplicative shift), and both works achieve the same optimality guarantee for $k = O(1)$. They also consider the problem of SBM with two unequal sized clusters and the binary censored block model. In contrast we show that the guarantees extend to the case even k is superconstant $o(\log(n))$ and provide sufficient guarantees for the case of $k = \theta(\log(n))$ pointing to a possible divergence between information theoretic possibility and computational feasibility at $k = \log(n)$ which we leave as an open question.

1.2 Preliminaries

In this section we describe the notation and definitions which we use through the rest of the paper.

Notation Throughout the rest of the paper we will be reserving capital letters such as X for matrices, and with $X[i,j]$ we will denote the corresponding entries. In particular, J will be used to denote the all-ones matrix and I the identity matrix. Let $A \bullet B$ be the element wise inner product of two matrices, i.e. $A \bullet B = Trace(A^T B)$. We note that the all the logarithms used in this paper are natural logarithms i.e. with the base e.

Let $G = (V, E)$ be a graph, n the number of vertices, and $A(G)$ its adjacency matrix. With $G \sim \mathscr{G}_{p,q,k}$ we denote a graph drawn from the stochastic block model distribution as described earlier with k denoting the number of hidden *clusters* each of size m. We denote the underlying hidden partition with $\{P_t\}$. Let $P(i)$ be the function that maps vertex i to the cluster containing i. To avoid confusion in the notation, note that with P_t we denote the t^{th} cluster and $P(i)$ denotes the cluster containing the vertex i. We now describe the definitions of a few quantities which will be useful in further discussion of our results as well as their proofs. Define $\delta_{i \to P_t}$ to be the "degree" of vertex i to cluster t. Formally

$$\delta_{i \to P_t} \triangleq \sum_{j \in P_t} A(G)[i,j] \,.$$

Similarly for any two clusters P_{t_1}, P_{t_2} define $\delta_{P_{t_1} \to P_{t_2}}$ as

$$\delta_{P_{t_1} \to P_{t_2}} \triangleq \sum_{i \in P_{t_1}} \sum_{j \in P_{t_2}} A(G)[i,j] \,.$$

Define the "in-degree" of a vertex i, denoted $\delta^{in}(i)$, to be the number of edges of going from the vertex to its own cluster:

$$\delta^{in}(i) \triangleq \delta_{i \to P(i)} \,,$$

also define $\delta^{out}_{\max}(i)$ to be the maximum "out-degree" of a vertex i to any other cluster:

$$\delta^{out}_{\max}(i) \triangleq \max_{P_t \neq P(i)} \delta_{i \to P_t} \,.$$

Finally, define

$$\Delta(i) \triangleq \delta^{in}(i) - \delta^{out}_{\max}(i) \,,$$

where $\Delta(i)$ will be the crucial parameter in our threshold. Remember that $\Delta(i)$ for $A(G)$ is a random variable and let $\Delta \triangleq \mathbb{E}[\Delta(i)]$ be its expectation (same for all i).

Organization The rest of this paper is structured as follows. In Section 2 we discuss the two SDP relaxations we consider in the paper. We state sufficient conditions for exact recovery for both of them as Theorems 4 and 3 and provide an intuitive explanation of why the condition (1) is sufficient for recovery up to the optimal threshold. We provide formal proofs of Theorems 1 and 2 in Sections 3.1 and 3.2, respectively. We provide the proof of Theorem 3 in Section 3.3. Further in Section 4 we show how our result can be extended to a semi-random model with a monotone adversary. We further provide an experimental evaluation of the SDPs in Section 5 followed by a discussion and connections with multireference alignment in Section 6.

2 SDP Relaxations and Main Results

In this section we present two candidate SDPs which we use to recover the hidden partition. The first SDP is inspired from the Max-k-Cut SDP introduced by Frieze and Jerrum [20] where we do not explicitly encode the fact that each cluster contains an equal number of vertices. In the second SDP, we encode the fact that each cluster has exactly m vertices explicitly. We state our main theorems which provide sufficient conditions for exact recovery in both SDPs. Indeed the latter SDP, being stronger, is the one we use to prove our main theorem, Theorem 1. Before describing the SDPs, let's first consider the maximum likelihood estimator (MLE) of the hidden partition. It is easy to see that the MLE corresponds to the following problem which we refer to as the multisection problem. Given a graph $G = (V, E)$, divide the set of vertices into k-clusters $\{P_t\}$ such that for all t_1, t_2, $|P_{t_1}| = |P_{t_2}|$ and the number of edges $(u, v) \in E$ such that $u \in P_{t_1}$ and $v \in P_{t_2}$ are minimized. (This problem has been studied under the name of Min-Balanced-k-partition [23].) In this section we consider two SDP relaxations for the multisection problem. Since SDPs can be solved in polynomial time, the relaxations provide polynomial time algorithms to recover the hidden partitions.

A natural relaxation to consider for the problem of multisection in the stochastic block model is the Min-k-Cut SDP relaxation studied by Frieze and Jerrum [20] (they actually study the Max-k-Cut problem, but we can analogously study the min cut version too.) The Min-k-Cut SDP formulates the problem as an instance of Min-k-Cut where one tries to separate the graph into k-partitions with the objective of minimizing the number of edges cut by the partition. Note that the k-Cut version does not have any explicit constraints for ensuring balancedness. However, studying Min-k-Cut through SDPs has a natural difficulty; the relaxation must explicitly contain a constraint that tells it to divide the graph into at least k-clusters. In the case of SBMs with the parameters $\alpha \frac{\log(n)}{n}$ and $\beta \frac{\log(n)}{n}$, one can try and overcome the above difficulty by making use of the fact that the generated graph is very sparse.

Thus, instead of looking directly at the min-k-Cut objective, we can consider the following objective: minimizing the difference between the number of edges cut and the number of non-edges cut. Indeed for sparse graphs, the second term in the difference is the dominant term, and hence the SDP has an incentive to produce more clusters. Note that the above objective can also be thought of as doing Min-k-Cut on the signed adjacency matrix $2A(G) - J$ (where J is the all-ones matrix). Following the above intuition, we consider the following SDP (2) which is inspired from the Max-k-Cut formulation of Feige and Jerrum [20]. In Section 6 we provide a reduction, to the k-Cut SDP we study in this paper, from a more general class of SDPs studied by Charikar et al. [13] for unique games and more recently by Bandeira et al. [10] in a more general setting:

$$\begin{aligned} \max\ & (2A(G) - J) \bullet Y \\ \text{s.t.}\ & Y_{ii} = 1 \quad (\forall\ i) \\ & Y_{ij} \geq -\frac{1}{k-1} \quad (\forall\ i,j) \\ & Y \succcurlyeq 0\ . \end{aligned} \tag{2}$$

To see that the above SDP is a relaxation of the multisection problem, note that for the hidden partition $\{P_t\}$, we can define a candidate solution Y^* as follows. $Y^*_{ij} = 1$ if i, j belong to the same cluster and $-\frac{1}{k-1}$ if i, j belong to different clusters. Note that although the objective does not directly minimize the number of edges cut, it is an additive/multiplicative shift of it. Given $G \sim \mathscr{G}_{p,q,k}$, define

$$\nu(i) \triangleq \delta_{in}(i) - \max_{i,j} \left(\delta_{i \to P(j)} + \delta_{j \to P(i)} - \frac{\delta_{P(j) \to P(i)}}{n/k} \right)$$

Theorem 4. *Let $G \sim \mathscr{G}_{p,q,k}$, with $p = \alpha \frac{\log(m)}{m}$ and $q = \beta \frac{\log(m)}{m}$ where α, β are constant. Consider the SDP given by (2). With probability $1 - n^{-\Omega(1)}$ over the choice of G, if the following condition is satisfied, then the SDP recovers the hidden partition:*

$$\min_i \nu(i) \geq \hat{c} \left(\sqrt{pn/k + qn} + \sqrt{\log(n)} \right)\ , \tag{3}$$

where $\hat{c}$ is a universal constant.

In other words with probability $1 - n^{-\Omega(1)}$, condition (3) implies exact recovery.

We provide a proof of the above theorem in Section 3.4, but we note the above condition is not an optimal one in terms of exact recovery, and we discuss this issue next. It is quite possible that the above SDP recovers the planted multisection all the way down to the threshold; however, we have not been able to establish this and leave it as an open question. Indeed to prove our results, we consider a stronger SDP with which we establish optimality. We have empirically tested the performance of both the SDPs and include the results in Section 5. We now take a closer look at the

above sufficient condition (3) and argue why the condition is not strong enough to achieve optimal results. It is not hard to see that

$$\mathbb{E}[\nu(i)] \sim p\frac{n}{k} - q\frac{n}{k} - O\left(\sqrt{q\frac{n}{k}\log(n)}\right)$$

Note that, in expectation, the maximization term in the definition of $\nu(i)$ has an extra $\log(n)$ term as the maximization runs through all i,j pairs. For the condition (3) to hold with at least a constant probability, we expect that it needs to be the case that

$$p\frac{n}{k} - q\frac{n}{k} - O\left(\sqrt{q\frac{n}{k}\log(n)}\right) \geq O\left(\sqrt{p\frac{n}{k} + q\frac{n}{k}k} + \sqrt{\log(n)}\right)$$

Substituting the parameter range that we are interested $p = \frac{\alpha\log(m)}{m}$ and $q = \frac{\beta\log(m)}{m}$, we require that

$$\alpha - \beta \geq O\left(\sqrt{\beta} + \sqrt{\frac{\beta k}{\log(n)}}\right)$$

Indeed from the above expression, it is clear that if $k << \log(n)$ the first term above dominates, and we cannot expect to get the tight results we hope for in Theorem 1. A closer look at the above calculation reveals that the major barrier toward achieving the optimal result is the additional $\log(n)$ factor due to the maximization over all i,j in the definition of $\nu(i)$. For instance, if one could replace the maximization term above with a term that takes the maximum per vertex over all clusters, one would pick up only a $\log(k)$ term (as there are only k-clusters) and hopefully achieve optimality.

In the context of the above discussion, we suggest the following SDP in which we explicitly add a per-row constraint bounding the number of vertices belonging to the same cluster as the vertex in contention:

$$\begin{aligned} &\max\ A(G) \bullet Y \\ &\text{s.t.}\ \ \sum_j Y_{ij} + \sum_j Y_{ji} = 2n/k \ \ (\forall\ i) \\ &\qquad Y_{ii} = 1 \ \ (\forall\ i) \\ &\qquad Y_{ij} \geq 0 \ \ (\forall\ i,j) \\ &\qquad Y \succcurlyeq 0\,. \end{aligned} \tag{4}$$

To see that the above SDP is a relaxation of the MLE discussed above, note that for any partition $P = \{P_i\}$, we can associate a canonical $n \times n$ matrix Y_P with it defined as

$$Y_P[i,j] = \begin{cases} 1 & \text{vertices } i \text{ and } j \text{ belong to the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

Note that Y_P satisfies the SDP constraints and the SDP maximizes the number of edges within the cluster which is equivalent to minimizing the number of edges across the clusters. The second constraint above, since Y is symmetric, says that the sum of the values along the row is n/k, which represents the number of vertices in a cluster. For the SDP above, we show the following theorem which is a restatement of Theorem 3.

Theorem 3. *Let $G \sim \mathcal{G}_{p,q,k}$. With probability $1 - n^{-\Omega(1)}$ over the choice of G, if the following condition is satisfied, then the SDP defined by (4) recovers the hidden partition:*

$$\min_i \Delta(i) \geq \hat{c}\left(\sqrt{pn/k + qn} + q\sqrt{\frac{n}{k}\log(n)} + \sqrt{\log(n)} + \log(k)\right), \tag{5}$$

In other words with probability $1 - n^{-\Omega(1)}$, condition (5) implies exact recovery.

We remark that the above statement is indeed true for all values of p, q. For the specific range that we are interested in, we show in Section 3.1 how condition (5) leads to the optimal threshold. The following is an intuitive explanation of why this is the case that condition in (5) for $k << \log(n)$ in Theorem 3 is optimal. As stated earlier the regime we consider is the case when $p = \frac{\alpha \log(m)}{m}$ and $q = \frac{\beta \log(m)}{m}$, where α and β are constants.

Note that for the MLE to succeed, the values of p and q should be such that $\min_i\{\delta^{in}(i) - \delta^{out}_{\max}(i)\} \geq 0$ w.h.p., since otherwise one expects there should be many such vertices i for which $\delta^{in}(i) - \delta_{i\to P_t} \leq 0$ for some $t \neq P(i)$ and in particular a pair t_1, t_2 such that there exists $i \in P_{t_1}, j \in P_{t_2}$ such that $\delta^{in}(i) - \delta_{i\to P_{t_2}} \leq 0$ as well as $\delta^{in}(j) - \delta_{i\to P_{t_1}} \leq 0$. This would imply that we can exchange the pairs i, j and get a better partition than the planted partition and therefore that the MLE itself does not recover the hidden partition. Recall that $\Delta(i) = \delta^{in} - \delta^{out}_{\max}(i)$. We now show that the deviation in $\Delta(i)$ required by Theorem 3 is $o(\mathbb{E}[\Delta(i)])$, and therefore informally one can expect, intuitively, that

$$\mathbb{P}(\min_i \Delta(i) \geq 0) \sim \mathbb{P}\left(\min_i \Delta(i) \geq o(\mathbb{E}[\Delta(i)])\right)$$

which implies that the SDP in Theorem 3 recovers the partition optimally. Indeed, the deviation required in Theorem 3 is $o(\mathbb{E}[\Delta(i)])$:

$$\begin{aligned}
&\frac{\left(\sqrt{pn/k + qn} + q\sqrt{n/k\log(n)} + \sqrt{\log(n)}\right)}{\mathbb{E}[\Delta(i)]} \\
&= \frac{O\left(\sqrt{\log(m)(\alpha + k\beta)}\right) + O(\sqrt{\log(n)})}{\Omega\left((\alpha - \beta)\log(m)\right)} \\
&= o(1)\,.
\end{aligned}$$

Above we assumed that $k = o(\log(n))$. In the next section, following from the intuition above, we prove Theorems 1 and 2 which imply that our SDP is optimal.

In Section 5 we present an experimental evaluation of the two SDPs considered in this section. The experiments corroborate Theorem 1 and also show that the SDP in (2) experimentally seems to have a similar recovery performance as the (stronger) SDP in (4); however we could only prove a suboptimal result about it. We leave the possible optimality of the SDP in (2) as an open question.

3 Proofs

In this section we collect all the proofs of the main theorems stated so far. We first prove Theorem 1 assuming Theorem 3 in Section 3.1. Further in Section 3.2, we prove Theorem 2. We then prove our main Theorem 3 in Section 3.3 regarding the SDP defined in (4). Finally we provide the proof of Theorem 4 regarding the SDP defined in (2) in Section 3.4.

3.1 Proof of Optimality: Theorem 1

Proof. We will use the condition of Theorem 3 and the following lemma to prove theorem 1.

Lemma 1. *Let* $p = \frac{\alpha \log(m)}{m}$ *and* $q = \frac{\beta \log(m)}{m}$. *Let* $k = \gamma \log(m)$ *(where* $\gamma = O(1)$*). Now we have that as long as*

$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{1 + c_1 \sqrt{\beta\gamma}\left(1 + \log\left(\sqrt{\frac{\alpha}{\beta}}\right)\right)} \tag{6}$$

then for sufficiently large n, we have that with probability at least $1 - n^{-\Omega(1)}$ $\forall i, t$

$$\delta^{in}(i) - \delta_{i \to P_t} > c_2\left(\sqrt{\beta\gamma}\log(n) + \sqrt{\alpha \log(n)}\right)$$

where $c_2 > 0$ *be any fixed number and* $c_1 > 0$ *in* (6) *is a constant depending on* c_2.

To complete the proof of Theorem 1, we first observe that for the given range of parameters $p = \frac{\alpha \log(m)}{m}$ and $q = \frac{\beta \log(m)}{m}$, condition (5) in Theorem 3 becomes

$$\hat{c}\left(\sqrt{pn/k + qn} + q\sqrt{\frac{n}{k}\log(n)} + \sqrt{\log(n)} + \log(k)\right)$$

$$\leq c_2\left(\sqrt{\beta k \log(m)} + \sqrt{\alpha \log(n)}\right)$$

However, Lemma 1 implies that with probability $1 - n^{-\Omega(1)}$ we have that if condition 6 is satisfied, then $\forall i, t$

$$\delta^{in}(i) - \delta_{i \to P_t} > c_2 \left(\sqrt{\beta\gamma} \log(n) + \sqrt{\alpha \log(n)} \right)$$

where $c_2 > 0$ depends on $\hat{c}$. Therefore, with probability $1 - n^{-\Omega(1)}$, the condition in (5) of Theorem 3 is satisfied which in turn implies the SDP in Theorem 3 recovers the clusters, which concludes the proof of Theorem 1. Note that setting $\gamma = o(1)$ we get the case $k = o(\log(n))$ and the above condition reduces to $\sqrt{\alpha} - \sqrt{\beta} > 1 + o_n(1)$. □

In the rest of the section, we prove Lemma 1. For the remainder of this section, we borrow the notation from Abbe et al. [2]. In [2, Definition 3, Section A.1], they define the following quantity $T(m, p, q, \delta)$ which we use:

Definition 1. Let m be a natural number, $p, q \in [0, 1]$, and $\delta \geq 0$, define

$$T(m, p, q, \delta) = \mathbb{P}\left[\sum_{i=1}^{m} (Z_i - W_i) \geq \delta \right],$$

where W_i are i.i.d Bernoulli(p) and Z_i are i.i.d. Bernoulli(q), independent of the W_i.

Let $Z = \sum_{i=1}^{m} Z_i$ and $W = \sum_{i=1}^{m} W_i$. The proof is similar to proof of [2, Lemma 8, Section A.1] with modifications.

Proof. (of Lemma 1) We will bound the probability of the bad event:

$$\delta^{in}(i) - \delta_{i \to P_t} \leq c_2 \left(\sqrt{\beta\gamma} \log(n) + \sqrt{\alpha \log(n)} \right).$$

Note that $\delta_{in}(i)$ is a binomial variable with parameter p and similarly $\delta_{i \to P_t}$ is a binomial variable with parameter q, and therefore, following the notation of [2], we have that the probability of this bad event is

$$T\left(m, p, q, -c_2 \left(\sqrt{\beta\gamma} \log(n) + \sqrt{\alpha \log(n)} \right)\right).$$

We show the following strengthening of their lemma.

Lemma 2. *Let W_i be a sequence of i.i.d Bernoulli$\left(\frac{\alpha \log(m)}{m}\right)$ random variables and Z_i an independent sequence of i.i.d Bernoulli$\left(\frac{\beta \log(m)}{m}\right)$ random variables; then the following bound holds for m sufficiently large:*

$$T\left(m, \frac{\alpha \log(m)}{m}, \frac{\beta \log(m)}{m}, -c_2\left(\sqrt{\beta\gamma}\log(n) + \sqrt{\alpha \log(n)}\right)\right) \leq$$
$$\exp\left(-\left(\alpha + \beta - 2\sqrt{\alpha\beta} - c_1\sqrt{\beta\gamma}\left(1 + \log\left(\sqrt{\frac{\alpha}{\beta}}\right)\right) + o(1)\right)\log(m)\right) \tag{7}$$

where $c_2 > 0$ is a fixed number and $c_1 > 0$ depends only on c_2.

Assuming the above lemma and taking a union bound over all clusters and vertices, we get the following sequence of equations which proves Theorem 1:

$$\mathbb{P}\left((\exists\, i, t)\ \delta^{in}(i) - \delta_{i \to P_t} \leq c_2\left(\sqrt{\beta\gamma}\log(n) + \sqrt{\alpha\log(n)}\right)\right)$$
$$\leq mk^2 \exp\left(-\left(\alpha + \beta - 2\sqrt{\alpha\beta} - c_1\sqrt{\beta\gamma}\left(1 + \log\left(\sqrt{\frac{\alpha}{\beta}}\right)\right) + o(1)\right)\log(m)\right)$$
$$\leq \exp\left(-\left(\alpha + \beta - 2\sqrt{\alpha\beta} - 1 - c_1\sqrt{\beta\gamma}\left(1 + \log\left(\sqrt{\frac{\alpha}{\beta}}\right)\right) + o(1)\right)\log(m)\right)$$
$$\leq m^{-\Omega(1)}$$
$$\leq n^{-\Omega(1)}$$

□

3.2 Proof of Optimality: Theorem 2

Proof. The theorem follows directly from the lower bound presented in [2]. They showed that [2, Theorem 1] when we sample $G \sim G_{p,q,2}$ with $p = \alpha' \frac{\log(n)}{n}$ and $q = \beta' \frac{\log(n)}{n}$, it is information theoretically impossible to correctly recover the clusters with high probability if

$$\sqrt{\alpha'} - \sqrt{\beta'} < \sqrt{2}$$

Now consider $G \sim G_{p,q,k}$ with $p = \alpha \frac{\log(m)}{m}$ and $q = \beta \frac{\log(m)}{m}$. Suppose that the algorithm was given the membership of vertices in all the clusters except two of them. A direct application of the above theorem yields that it is information theoretically impossible to correctly recover the two unrevealed clusters with high probability if

$$\sqrt{2\frac{\log(m)}{\log(n)}}(\sqrt{\alpha} - \sqrt{\beta}) < \sqrt{2}$$

which is equivalent to

$$\sqrt{\alpha} - \sqrt{\beta} < \frac{\log(n)}{\log(m)} = 1 + \frac{\log(k)}{\log(m)} = 1 + o_n(1)$$

which proves the bound. □

Proof of Lemma 2. The proof of Lemma 2 is a simple modification of the proof of [2, Lemma 8, Section A.1]. We mention the proof here for completeness.

Define $r = c_2\left(\sqrt{\beta\gamma}\log(n) + \sqrt{\alpha\log(n)}\right) \leq c_1\sqrt{\beta\gamma}\log(n)$ (for some fixed $c_1 > 0$ depending only on c_2), and let $Z = \sum Z_i$ and $W = \sum W_i$. We split T as follows:

$$T(m,p,q,-r) = \mathbb{P}\left(-r \leq Z - W \leq \log^2(m)\right) + \mathbb{P}\left(Z - W \geq \log^2(m)\right) .$$

Let's bound the second term first. A simple application of Bernstein's inequality (the calculations are shown in [2, Lemma 8, Section A.1]) shows that

$$\mathbb{P}\left(Z - W \geq \log^2(m)\right) \leq \exp\left(-\Omega(1)\frac{\log^2(m)}{\log(\log(m))}\right) .$$

We now bound the first term $\mathbb{P}\left(-r \leq Z - W \leq \log^2(m)\right)$. Define

$$\hat{r} = argmax_x \mathbb{P}(Z - W = -x)$$

Now it is easy to see that $\hat{r} = O(\log(m))$ (for $p = \alpha\frac{\log(m)}{m}$ and $q = \beta\frac{\log(m)}{m}$). Let $r_{max} = \max(r, \hat{r})$ and $r_{min} = \min(r, \hat{r})$.

$$\mathbb{P}\left(-r \leq Z - W \leq \log^2(m)\right) \leq (\log^2(m) + r_{max})\mathbb{P}(Z - W = -r_{min})$$

$$\leq (\log^2(m) + r_{max})\left(\sum_{k_2=r_{min}}^{\log^2(m)+r_{max}} \mathbb{P}(Z = k_2 - r)\mathbb{P}(W = k_2) + \sum_{k_2=\log^2(m)+r_{min}}^{m} \mathbb{P}(Z = k_2 - r)\mathbb{P}(W = k_2)\right)$$

$$\leq (\log^2(m) + r_{max})^2 \max_{k_2}\{\mathbb{P}(Z = k_2 - r_{min})\mathbb{P}(W = k_2)\} + (\log^2(m) + r_{max})\mathbb{P}(Z \geq \log^2(n))\mathbb{P}(W \geq \log^2(m))$$

The first inequality follows easily from considering both the cases $\hat{r} \geq r$ and $\hat{r} \leq r$. Similar probability estimates (using Bernstein) as before give that both

$$\mathbb{P}\left(Z \geq \log^2(m)\right), \mathbb{P}\left(W \geq \log^2(m)\right) \leq \exp\left(-\Omega(1)\frac{\log(m)}{\log(\log(m))}\right)$$

We now need to bound $\max_{k_2}\{\mathbb{P}(Z = k_2 - r)\mathbb{P}(W = k_2)\}$ for which we use Lemma 3 which is a modification of [2, Lemma 7, Section A.1]. Plugging the estimates from above and noting that $\max_{k_2}\{\mathbb{P}(Z = k_2 - r)\mathbb{P}(W = k_2)\} = T^*\left(m, p, q, \frac{r_{min}}{\log(m)}\right)$ (defined in Lemma 3), we get that

$$\mathbb{P}\left(-r \leq Z - W \leq \log^2(m)\right) \leq O(\log^4(n))T^*\left(m, p, q, \frac{r_{min}}{\log(m)}\right) + \log^2(n)\exp\left(-\Omega(1)\frac{\log(m)}{\log(\log(m))}\right)$$

Putting everything together, we get that

$$T(m, p, q, 0) \leq 2\log^4(n)T^*\left(m, p, q, \frac{r_{min}}{\log(m)}\right) + \log^2(n)\exp\left(-\Omega(1)\frac{\log(m)}{\log(\log(m))}\right) + \exp\left(-\Omega(1)\frac{\log(m)}{\log(\log(m))}\right)$$

Using Lemma 3 it follows from the above equation that

$$-\log(T(m, p, q, -r)) \geq -\Omega(\log(\log(m))) + g\left(\alpha, \beta, \frac{r_{min}}{\log(n)}\right)\log(m) - o(\log(m))$$

$$\geq \left(\alpha + \beta - 2\sqrt{\alpha\beta} - c_1\sqrt{\beta}\gamma\left(1 + \log\left(\sqrt{\frac{\alpha}{\beta}}\right)\right)\right)\log(m) - o(\log(m))$$

For the first inequality, we use Lemma 3 and set $\epsilon = \frac{r_{min}}{\log(n)}$. For the second inequality, we use the fact that $\epsilon \leq c_1\sqrt{\beta\gamma}$. □

Lemma 3. *Let $p = \frac{\alpha\log(m)}{m}$ and $q = \frac{\beta\log(m)}{m}$ and let W_i be a sequence of i.i.d Bernoulli-p random variables and Z_i an independent sequence of i.i.d Bernoulli-q random variables. Define*

$$V'(m, p, q, \tau, \epsilon) = \mathbb{P}\left(\sum Z_i = \tau\log(m)\right)\mathbb{P}\left(\sum W_i = (\tau + \epsilon)\log(m)\right)$$

$$= \binom{m}{\tau\log(m)} q^{\tau\log(m)}(1 - q)^{m - \tau\log(m)} \binom{m}{(\tau + \epsilon)\log(m)} p^{(\tau+\epsilon)\log(m)}(1 - p)^{m - (\tau+\epsilon)\log(m)},$$

where $\epsilon = O(1)$. *We also define the function*

$$g(\alpha,\beta,\epsilon) = (\alpha+\beta) - \epsilon\log(\alpha) - 2\sqrt{\left(\frac{\epsilon}{2}\right)^2 + \alpha\beta} + \frac{\epsilon}{2}\log\left(\alpha\beta\frac{\sqrt{(\frac{\epsilon}{2})^2+\alpha\beta}+\frac{\epsilon}{2}}{\sqrt{(\frac{\epsilon}{2})^2+\alpha\beta}-\frac{\epsilon}{2}}\right).$$

Then we have the following results for $T^*(m,p,q,\epsilon) = \max_{\tau>0} V'(m,p,q,\tau,\epsilon)$. *We have that, for* $m \in N$ *and* $\forall\tau > 0$

$$-\log(T^*(m,p,q,\epsilon)) \geq \log(m)g(\alpha,\beta,\epsilon) - o(\log(m)) \ .$$

Proof. The proof of the above lemma is computational and follows from carefully bounding the combinatorial coefficients. Note that

$$\log(V(m,p,q,\tau,\epsilon)) = \log\binom{m}{\tau\log(m)} + \log\binom{m}{(\tau+\epsilon)\log(m)} + \tau\log(m)\log(pq) + \epsilon\log(m)\log\left(\frac{p}{1-p} + (m-\tau\log(m))\log((1-p)(1-q))\right)$$

Substituting the values of p and q, we get

$$\begin{aligned}\log(V(m,p,q,\tau,\epsilon)) = \ & \log\binom{m}{\tau\log(m)} + \log\binom{m}{(\tau+\epsilon)\log(m)} \\ & + \tau\log(m)\left(\log(\alpha\beta) + 2\log\log(m) - 2\log(m)\right) \\ & + \epsilon\log(m)\left(\log(\alpha) + \log\log(m) - \log(m) + \alpha\frac{\log(m)}{m}\right) \\ & - \log(m)(\alpha+\beta) + o(\log(m))\end{aligned}$$

We now use the following easy inequality:

$$\log\binom{n}{k} \leq k\left(\log(ne) - \log(k)\right)$$

and now replacing this in the above equation gives us

$$-\log(V(m,p,q,\tau,\epsilon)) \geq \log(m)\left((\alpha+\beta) + (\tau+\epsilon)\log\left(\frac{\tau+\epsilon}{e}\right) + \tau\log\left(\frac{\tau}{\epsilon}\right) - \tau\log(\alpha\beta) - \epsilon\log(\alpha)\right) - o(\log(m) \qquad (8)$$

Now optimizing over τ proves the lemma. □

3.3 *Proof of Theorem 3*

In this section we prove our main theorem, Theorem 3, about the SDP defined by (4). We restate the SDP here:

$$\begin{aligned} \max\ & A(G) \bullet Y \\ \text{s.t.}\ & \sum_j Y_{ij} + \sum_j Y_{ji} = 2n/k \quad (\forall\ i) \\ & Y_{ii} = 1 \quad (\forall\ i) \\ & Y_{ij} \geq 0 \quad (\forall\ i,j) \\ & Y \succcurlyeq 0\,. \end{aligned} \tag{9}$$

Let Y^* be the matrix corresponding to the hidden partition $P^* = \{P_t\}$, i.e., $Y^*[i,j] = 1$ if i,j belong to the same cluster and 0 otherwise. Let $OPT(G)$ be the optimal value in the above SDP. We will show that Y^* is the unique solution to SDP (4) w.h.p as long as the conditions in Theorem 3 are satisfied. This would prove Theorem 3. Our proof will be based on a dual certificate. In that context consider the dual formulation of the above SDP which is the following:

$$\begin{aligned} & \textit{min}\ \text{Trace}(D) + (2n/k)\sum_i x_i \\ & \textit{s.t.}\ \ D + \sum_i x_i(R_i + C_i) - Z - A \succcurlyeq 0\,. \end{aligned} \tag{10}$$

where D is a diagonal matrix, x_i are scalars, Z is a nonnegative symmetric matrix (corresponding to the ≥ 0 constraints) with 0 in the diagonal entries, R_i is the matrix with 1 in every entry of row i and 0 otherwise, and $C_i = R_i^T$ is the matrix with 1 in every entry of column i and 0 otherwise, and we write A instead of $A(G)$ when there is no fear of confusion.

Let $DUAL(G)$ be the optimal value of the above dual program. We will first exhibit a valid dual solution $M^* = (D^*, \{x_i^*\}, Z^*)$ which, with high probability, has dual objective value δ such that $A \bullet Y^* = \delta$. But since $A \bullet Y^* \leq OPT(G) \leq DUAL(G)$ (by weak duality), we get that Y^* is an optimal solution to the above SDP. We will also show uniqueness via complementary slackness.

Before moving on further, it will be convenient to introduce the following definition which will be used in the proof later. We also encourage the reader to revisit the Notations section (Section 1.2) at this time as it would help with the reading of what follows.

Definition 2. Given a partition of n vertices $\{P_t\}_{t=1}^k$, we define the vectors $\{v_t\}$ to be the indicator vectors of the clusters. We further define the following subspaces, which are perpendicular to each other, and partition $\mathbb{R}^n$.

- $\mathbb{R}_k$: the subspace spanned by the vectors $\{v_t\}$, i.e., the subspace of vectors with equal values in each cluster,

- $\mathbb{R}_{n|k}$: the subspace perpendicular to $\mathbb{R}_k$, i.e., the subspace where the sum on each cluster is equal to 0.

At this point it is useful to look at what the complementary slackness condition implies. Since strong duality holds in the case of our SDP (easy to check that Slater's conditions are satisfied), we have that complementary slackness is zero which implies that

$$\text{Trace}(M^* Y^*) = \text{Trace}\left(M^* \sum v_t v_t^T\right) = 0\,.$$

for any optimal dual solution M^*. The above condition implies that for any such M^* (since M^* is PSD) it must be that the subspace $\mathbb{R}_k$ is an eigenspace with eigenvalue 0 which implies

$$(\forall i, t)\delta_{i\to P_t}(M^*) = 0\,. \tag{11}$$

Having established the conditions that must be satisfied by the optimal dual solution M^*, we describe our candidate dual solution

$$(D^*, \{x_i^*\}, Z^*)\,.$$

We begin by describing the choice of Z^*. If vertex i and j belong to the same cluster, then $Z^*[i,j] = 0$; otherwise

$$Z^*[i,j] = \left(\frac{\delta^{out}_{\max}(i)}{n/k} - \frac{\delta_{i\to P(j)}}{n/k}\right) + \left(\frac{\delta^{out}_{\max}(j)}{n/k} - \frac{\delta_{j\to P(i)}}{n/k}\right) + \left(\frac{\delta_{P(j)\to P(i)}}{(n/k)(n/k)} - \min_{t_1,t_2}\frac{\delta_{P_{t_1}\to P_{t_2}}}{(n/k)(n/k)}\right)\,.$$

It is easy to see that the matrix Z^* is symmetric by noting that exchanging j and i in the above expression leads to the same value. Also to see that each entry of Z^* is nonnegative, note that $Z^*[i,j]$ is the sum of nonnegative terms. Having defined Z^* as above, we choose x_i^* to be such that the condition given in equation (11) holds for the non-diagonal blocks, yielding:

$$x_i^* = \frac{\delta^{out}_{\max}(i)}{n/k} - \frac{1}{2}\min_{t_1,t_2}\frac{\delta_{P_{t_1}\to P_{t_2}}}{(n/k)(n/k)}\,.$$

And finally we define D^* to balance out the sum along the diagonal blocks from A as well as the x_i^*.

$$D^*[i,i] = \delta^{in}(i) - \delta^{out}_{\max}(i) - \sum_{j\in P(i)}\frac{\delta^{out}_{\max}(j)}{n/k} + \min_{t_1,t_2}\frac{\delta_{P_{t_1}\to P_{t_2}}}{n/k}\,.$$

Interestingly, this dual certificate construction seems to share some features with the one proposed by Awasthi et al. [6] for an SDP relaxation for k-means clustering. While we were not able to make a formal connection, it would be very interesting if the reason for the similarities was the existence of some type of canonical way of building certificates for clustering problems; we leave this for future investigations.

Now consider the objective for the dual program (10). It is easy to see that it is equal to

$$\mathrm{Trace}(D^*) + 2n/k \sum_i x_i^* = \sum_i \delta^{in}(i) = A(G) \bullet Y^* .$$

The following lemma implies that the abovementioned solution is a valid dual solution, proving that Y^* is an optimal solution to the above program (by weak duality).

Lemma 4. *The matrix $M^* = D^* + \sum_i x_i^*(R_i + C_i) - A - Z^*$ (as defined above) is such that with probability $1 - n^{-\Omega(1)}$, if the condition (5) is satisfied, then*

$$M^* \succeq 0 .$$

Proof. To prove this lemma, we first show that equation (11) is satisfied for M^*. This implies that the vectors $\{v_t\}$ which are indicator vectors for the clusters are an eigenvector with eigenvalue 0. Consider the value of $\delta_{i \to P_t}(M^*)$ when $P_t = P(i)$. In this case

$$\begin{aligned} \delta_{i \to P_t}(M^*) &= D^*[i,i] + \frac{n}{k} x_i^* + \sum_{i' \in P(i)} x_{i'}^* - \sum_{i' \in P(i)} A[i,i'] \\ &= 0 . \end{aligned}$$

where the last equality follows directly from the definitions of the dual certificate. Now consider the value of $\delta_{i \to P_t}(M^*)$ when $P_t \neq P(i)$. In this case

$$\begin{aligned} \delta_{i \to P_t}(M^*) &= \frac{n}{k} x_i^* + \sum_{j \in P_t} x_j^* - \sum_{j \in P_t} (Z[i,j] + A[i,j]) \\ &= \frac{n}{k} x_i^* + \sum_{j \in P_t} x_j^* - \sum_{j \in P_t} \left(\frac{\delta^{out}_{\max}(i)}{n/k} + \frac{\delta^{out}_{\max}(j)}{n/k} - \left(\frac{\delta_{i \to P(j)}}{n/k} + A[i,j] \right) + \right. \\ &\quad \left. \left(- \frac{\delta_{j \to P(i)}}{n/k} + \frac{\delta_{P(j) \to P(i)}}{(n/k)(n/k)} \right) - \min_{t_1, t_2} \frac{\delta_{P_{t_1} \to P_{t_2}}}{(n/k)(n/k)} \right) \\ &= \frac{n}{k} x_i^* + \sum_{j \in P_t} x_j^* - \sum_{j \in P_t} \left(\frac{\delta^{out}_{\max}(i)}{n/k} + \frac{\delta^{out}_{\max}(j)}{n/k} - \min_{t_1, t_2} \frac{\delta_{P_{t_1} \to P_{t_2}}}{(n/k)(n/k)} \right) \\ &= 0 . \end{aligned}$$

The third equality follows by noting that the terms in the parenthesis in the expression in the second line go to zero in summation. The fourth equality follows directly from the definitions.

The above implies that for all t, $M^* v_t = 0$. Therefore, we only need to show that M^* is PSD with high probability on the subspace $\mathbb{R}_{n|k}$ (which is perpendicular to $\mathbb{R}_k = span(\{v_k\})$). To that end, note that if a matrix W is such that for all i, $W[i, j_1] = W[i, j_2]$ when $P(j_1) = P(j_2)$, then for any $x \in \mathbb{R}_{n|k}$, $Wx = 0$, and similarly if for all j, $W[i_1, j] = W[i_2, j]$ when $P(i_1) = P(i_2)$, then for any $x \in \mathbb{R}_{n|k}$, $x^T W = 0$. Therefore, we have that $x^T Z^* x = x^T (R_i + C_i)x = 0$ and so $x^T M^* x = x^T D^* x - x^T A x$.

In order to finish the proof it is enough to show that for all $x \in \mathbb{R}_{n|k}$

$$x^T (D^* - A)x \geq 0 \,.$$

In order to prove the above equation and conclude the proof of Theorem 3, we use the following two lemmas.

Lemma 5. *Define $\lambda_{max}(A(G))$ to be the maximum over all $x \in \mathbb{R}_{n|k}$ of $x^T A(G)x$. With probability $1 - n^{-\Omega(1)}$ over the choice of G, $\lambda_{max}(A(G))$ is bounded by*

$$\lambda_{\max}(A(G)) \leq 3\sqrt{pn/k + qn} + c\sqrt{\log(n)} \,. \tag{12}$$

where c is a universal constant.

Proof. We use the following recent sharp concentration result [8, Corollary 3.12].

Theorem 5 (Bandeira et al. [8]). *Let X be an $n \times n$ symmetric matrix whose entries X_{ij} are independent centered random variables. Then there exists for any $0 < \epsilon \leq 1/2$ a universal constant $\tilde{c_\epsilon}$ such that for every $t \geq 0$*

$$\mathbb{P}\left(|X| \geq (1+\epsilon)2\sqrt{2}\tilde{\sigma} + t\right) \leq n e^{-t^2/\tilde{c_\epsilon}\sigma_*^2} \,,$$

where

$$\tilde{\sigma} = \max_i \sqrt{\sum_j \mathbb{E}[X_{ij}^2]}, \qquad \sigma_* = \max_{ij} \|X_{ij}\|_\infty \,.$$

We apply the above theorem to the matrix $A - \mathbb{E}[A]$. It is easy to see that the variance of any row $\tilde{\sigma}$ is upper bounded by

$$\tilde{\sigma} \leq \sqrt{p(1-p)n/k + q(1-q)n} \leq \sqrt{pn/k + qn} \,,$$

and $\sigma_* \leq 1$. Applying Theorem 5 with the above parameters $\tilde{\sigma} = \sqrt{pn/k + qn}$ and $\sigma_* = 1$, we get that with probability $1 - n^{-\Omega(1)}$

$$|A - \mathbb{E}[A]| \leq 3\sqrt{pn/k + qn} + c'\sqrt{\log(n)} \,.$$

where c' is a universal constant defined as $c' = 2\tilde{c_\epsilon}$ for $\epsilon = \frac{3}{2\sqrt{2}} - 1$ and $\tilde{c_\epsilon}$ defined by the statement of Theorem 5. Also note that $\mathbb{E}[A] + pI$ has the space $\mathbb{R}_{n|k}$ as an eigenspace with eigenvalue 0. Therefore, we have that for any unit vector $x \in \mathbb{R}_{n|k}$

$$\begin{aligned} |x^T A x| &\leq |A - \mathbb{E}[A]| + |x^T \mathbb{E}[A] x| \\ &\leq 3\sqrt{pn/k + qn} + c'\sqrt{\log(n)} + p \\ &\leq 3\sqrt{pn/k + qn} + c\sqrt{\log(n)}\,. \end{aligned}$$

where $c = c' + 1$. This proves Lemma 5. □

Lemma 6. *With probability* $1 - n^{-\Omega(1)}$*, we have that for all clusters* P_t

$$\sum_{j \in P_t} \frac{\delta^{out}_{\max}(j)}{n/k} \leq \frac{qn}{k} + 30\left(\sqrt{\frac{n\log(k)}{k}}q + \log(k) + \sqrt{\frac{n}{k}\log(n)} \cdot \max\left\{q, \sqrt{\frac{q\log(n)}{n/k}}, \frac{\log(n)}{n/k}\right\}\right), \tag{13}$$

and for all pairs of clusters P_{t_1} *and* P_{t_2}

$$\min_{t_1,t_2} \frac{\delta_{P_{t_1} \to P_{t_2}}}{n/k} \geq \frac{qn}{k} - 2\sqrt{q\log(n)}\,. \tag{14}$$

Proof. We prove Lemma 6 using the following.

Lemma 7. *For every vertex* i *we have that*

$$\mathbb{E}\left[\delta^{out}_{\max}(i)\right] \leq \frac{qn}{k} + 28\left(\sqrt{\frac{n\log(k)}{k}}q + \log(k)\right).$$

Proof. Consider $\delta^{out}_{\max}(i)$ for some i, this is defined to be the maximum of k random variables S_i with $S_i \sim \mathrm{Bin}(n/k, q)$ (the binomial distribution with parameters $n/k, q$) with variance $\frac{n}{k}\sigma^2$ where $\sigma^2 = q(1-q)$. Consider $\tilde{S}_i = S_i - \mathbb{E}[S_i]$. Let $\gamma = \sigma\sqrt{\frac{n\log(k)}{k}} + \log(k)$. From Corollary 1 we get that

$$\mathbb{P}\left(\tilde{S}_i \geq 4(t+1)\gamma\right) \leq \frac{1}{k^{t+1}}\,;$$

therefore by a union bound, we get that the

$$\mathbb{P}\left(\max_i \tilde{S}_i \geq 4(t+1)\gamma\right) \leq \frac{1}{k^t}\,.$$

Hence, we can bound the expectation by

$$\begin{aligned}
\mathbb{E}[\max_i \tilde{S}_i] &\leq 4\gamma + \sum_{t=1}^{\infty} 4(t+1)\gamma\, \mathbb{P}\left(\max_i \tilde{S}_i \geq 4t\gamma\right) \\
&\leq 4\gamma + \sum_{t=1}^{\infty} 4(t+1)\gamma \frac{1}{k^{t-1}} \\
&\leq 4\gamma + 4\gamma \left(\sum_{t=1}^{\infty} (t+1)\frac{1}{2^{t-1}}\right) \\
&\leq 4\gamma + 24\gamma \\
&\leq 28\left(\sigma\sqrt{\frac{n\log(k)}{k}} + \log(k)\right) .
\end{aligned}$$

It follows from the above that

$$\mathbb{E}[\delta^{out}_{\max}(i)] \leq \frac{n}{k}q + 28\left(\sqrt{\frac{n\log(k)}{k}}q + \log(k)\right) .$$

□

Using this, the proof of Lemma 6 is as follows. Note that by a direct application of the Chernoff bound described in Corollary 1 and with a union bound over all clusters and vertices, we get that with probability $1 - \frac{1}{n}$ for all vertices i and all clusters $P_t \neq P(i)$

$$\delta_{i\to P_t} \leq \frac{qn}{k} + 12\sqrt{\frac{qn}{k}\log(n)} + 12\log(n) .$$

Lets call the event that the above holds $\mathscr{E}$ and consider the sum

$$S(i) = \frac{\sum_{i' \in P(i)} \delta^{out}_{\max}(i)}{n/k} .$$

Let

$$\gamma = \frac{qn}{k} + 30\left(\sqrt{\frac{n\log(k)}{k}}q + \log(k) + \sqrt{\frac{n}{k}\log(n)} \cdot \max\left\{q, \sqrt{\frac{q\log(n)}{n/k}}, \frac{\log(n)}{n/k}\right\}\right) .$$

We have that

$$\begin{aligned}
\mathbb{P}\left(\exists i\, S(i) \geq \gamma\right) &= \mathbb{P}(\mathscr{E})\mathbb{P}\left(\exists i\, S(i) \geq \gamma \mid \mathscr{E}\right) + \mathbb{P}(\sim \mathscr{E})\mathbb{P}\left(\exists i\, S(i) \geq \gamma \mid \sim \mathscr{E}\right) \\
&\leq n^{-\Omega(1)} + \mathbb{P}\left(\exists i\, S(i) \geq \gamma \mid \sim \mathscr{E}\right) .
\end{aligned}$$

Now for a fixed i we will consider $\mathbb{P}(S(i) \geq \gamma \mid \sim \mathscr{E})$. Note that under the conditioning the individual entries in the sum above are still independent, and therefore the above is an average of independent random variables each of which is bounded by $\frac{qn}{k} + 12\sqrt{\frac{qn}{k}\log(n)} + 12\log(n)$ (by the conditioning). Also note that for any positive random variable X

$$\mathbb{E}[X \mid \sim \mathscr{E}] \leq \frac{\mathbb{E}[X]}{\mathbb{P}(\sim \mathscr{E})},$$

and since we have that $\mathbb{P}(\sim \mathscr{E}) \geq 1 - 1/n$, we get that

$$\mathbb{E}[S(i) \mid \sim \mathscr{E}] \leq \mathbb{E}[S(i)] + \frac{\mathbb{E}[S(i)]}{n-1}.$$

We now use Hoeffding's inequality 9 in the conditioned probability space (and remove the conditioning terms from the probability for ease of notation) to get that

$$\mathbb{P}(S(i) \geq \mathbb{E}[S(i)] + t) \leq \exp\left(-\frac{2\frac{n^2}{k^2}t^2}{\frac{n}{k}\left(\frac{qn}{k} + 12\sqrt{\frac{qn}{k}\log(n)} + 12\log(n)\right)^2}\right).$$

Now, if we choose

$$t = 25\sqrt{\frac{n}{k}\log(n)} \cdot \max\left\{q, \sqrt{\frac{q\log(n)}{n/k}}, \frac{\log(n)}{n/k}\right\},$$

and apply a union bound, we get that with

$$\mathbb{P}\left(\exists i\, S_i \geq \mathbb{E}[S_i] + t \mid \sim \mathscr{E}\right]) \leq n^{-\Omega(1)},$$

and now substituting the value of $\mathbb{E}[S(i) \mid \mathscr{E}]$ from before and being extremely liberal with the constants for n large enough, we have that

$$\mathbb{P}\left(\exists i\, S(i) \geq \frac{qn}{k} + 30\left(\sqrt{\frac{n\log(k)}{k}}q + \log(k) + \sqrt{\frac{n}{k}\log(n)} \cdot \max\left\{q, \sqrt{\frac{q\log(n)}{n/k}}, \frac{\log(n)}{n/k}\right\}\right)\right) \leq n^{-\Omega(1)}.$$

To show the second equation, note that for any pair of clusters t_1, t_2, $\delta_{P_{t_1} \to P_{t_2}}$ is a sum of $(n/k)^2$ independent random variables. Therefore, by a Chernoff bound from the second part of Theorem 7 and a union bound, we get that with probability $1 - n^{-\Omega(1)}$

$$\min_{t_1,t_2} \frac{\delta_{P_{t_1} \to P_{t_2}}}{n/k} \geq \frac{qn}{k} - 2\sqrt{q \log(n)} \,.$$

□

Using those two lemmas, we can now conclude the proof of Theorem 3 as follows:

We separate $D^* = D_1^* - D_2^*$, where D_1^*, D_2^* are diagonal matrices

$$D_1^*[i,i] = \delta^{in}(i) - \delta^{out}_{\max}(i)$$
$$D_2^*[i,i] = \sum_{j \in P(i)} \frac{\delta^{out}_{\max}(j)}{n/k} - \min_{t_1,t_2} \frac{\delta_{P_{t_1} \to P_{t_2}}}{n/k} \,.$$

Now for any $x \in \mathbb{R}_{n|k}$ let's consider $x^T(D^* - A)x$:

$$\begin{aligned}
x^T(D^* - A)x &\geq \min_i D_1^*[i,i] - \left(\max_i D_2^*[i,i] + \max_{x \in \mathbb{R}_{n|k}} x^T A x \right) \\
&\geq \min_i D_1^*[i,i] \\
&\quad - \left(30 \left(\sqrt{\frac{n \log(k)}{k}} q + \log(k) + \sqrt{\frac{n}{k} \log(n)} \cdot \max \left\{ q, \sqrt{\frac{q \log(n)}{n/k}}, \frac{\log(n)}{n/k} \right\} \right) \right. \\
&\quad \left. + 3\sqrt{pn/k + qn} + c\sqrt{\log(n)} \right) \\
&\geq \min_i D_1^*[i,i] - \hat{c} \left(\sqrt{pn/k + qn} + q\sqrt{\frac{n}{k} \log(n)} + \sqrt{\log(n)} + \log(k) \right) \\
&\geq 0 \,.
\end{aligned}$$

where $\hat{c}$ is a universal constant. The second inequality follows by direct substitutions from equations (12), (13), and (14); the third inequality follows from noting that n is large enough such that $\sqrt{qn} >> \sqrt{q}\log(n)$ and $\sqrt{\log(n)} \frac{\log(n)}{\sqrt{\frac{n}{k}}} << \sqrt{log(n)}$ and $\sqrt{qn \frac{\log(k)}{k}} \leq \sqrt{qn}$. The last inequality follows from condition 5 of Theorem 3. □

It is easy to show using complementary slackness that Y^* is indeed the unique optimal solution with high probability. For completeness we include the proof in the next section.

3.3.1 Proof of Uniqueness of the Solution

In this section we prove that Y^* is the unique optimal solution to the SDP considered in Section 3.3. To remind the reader, M^* was the candidate dual solution. For the rest of the section, we use the same notations we defined in Section 3.3. To show uniqueness, we make use of complementary slackness which implies that for any other optimal solution $\hat{Y}$ since with high probability $M^* = D^* - \sum_i x_i^* R_i - A(G) - Z^*$ is an optimal solution of the dual program, we have that

$$\hat{Y} \bullet M^* = 0\,.$$

But it is easy to see from the proof of Lemma 4 that we can make a stronger statement that the subspace $\mathbb{R}_k$ is the null space of M^* and on the perpendicular subspace $\mathbb{R}_{n|k}$ the lowest eigenvalue is strictly greater than 0. Combining this with the complementary slackness condition in particular implies that the span of the columns of $\hat{Y}$ is restricted to the span of $\mathbb{R}_k$. Hence, the conditions of the SDP (sum constraint, the diagonal value constraint, and the positivity constraint) force $\hat{Y} = Y^*$ if the column space of $\hat{Y}$ is the span of $\mathbb{R}_k$ which proves uniqueness.

3.4 Proof of Theorem 4

Proof. We extend the definitions of Section 1.2 to ease readability. We define the notion of relative degree $\bar{\delta}$ by defining it as the number of edges present minus the number of edges not present. In this light we define the following quantities extending the definitions from Section 1.2.

$\delta_{i\to P_t}$ to be the "degree" of vertex i to cluster t. Formally

$$\begin{aligned}
\bar{\delta}_{i\to P_t} &\triangleq 2\delta_{i\to P_t} - |P_t| \\
\bar{\delta}_{P_{t_1}\to P_{t_2}} &\triangleq 2\delta_{P_{t_1}\to P_{t_2}} - |P_{t_1}||P_{t_2}| \\
\bar{\delta}^{in}(i) &\triangleq 2\delta^{in}(i) - |P(i)|
\end{aligned}$$

We consider the following SDP in this section. Let J be the $n \times n$ matrix such that $J[i,j] = 1$ for all i,j.

$$\begin{aligned}
\max\ & (2 * A(G) - J) \bullet Y \\
\text{s.t.}\ & Y_{ii} = 1 \quad (\forall\ i) \\
& Y_{ij} \geq -\frac{1}{k-1} \quad (\forall\ i,j) \\
& Y \succcurlyeq 0\ .
\end{aligned} \tag{15}$$

The dual of the above SDP is as follows:

$$\begin{aligned} &\min \text{Trace}(D) + \frac{1}{k-1}\sum_{ij} Z[i,j] \\ &\text{s.t.} \quad D - Z - (2A(G) - J) \succcurlyeq 0\,. \end{aligned} \tag{16}$$

where Z is a symmetric entrywise nonnegative matrix with zeros in the diagonal and D is a diagonal matrix.

The optimal solution Y^* we have in mind is the matrix $Y^*_{ij} = 1$ if i,j belong to the same cluster and $-\frac{1}{k-1}$ if i,j belong to different clusters. Note that Y^* is PSD and is a valid solution of the primal. In this case it is easy to see that the value of the SDP is equal to

$$(2 * A(G) - J) \bullet Y^* = \sum_i \left(\bar{\delta}_{in}(i) - \frac{\sum_{t:P(i)\neq P_t} \bar{\delta}_{i\to P_t}}{k-1} \right)$$

We will exhibit a candidate dual solution D^*, Z^* such that

$$(2 * A(G) - J) \bullet Y^* = \text{Trace}(D) + \frac{1}{k-1}\sum_{ij} Z[i,j]$$

and with high probability $D^* - Z^* - (2A(G) - J) \succcurlyeq 0$ if condition (3) of the theorem is satisfied. Note that this implies through weak duality that Y^* is a solution of (2). The uniqueness of the solution can be proven exactly in the same way as in Section 3.3.1.

Before we define our candidate dual solution, we define the following quantity for ease of notation:

$$\begin{aligned} \bar{\delta}_{min} &\triangleq \min_{i,j} \left(-\bar{\delta}_{i\to P(j)} - \bar{\delta}_{j\to P(i)} + \frac{\bar{\delta}_{P(j)\to P(i)}}{(n/k)} \right) \\ &= \left(n/k - 2\max_{i,j} \left(\delta_{i\to P(j)} + \delta_{j\to P(i)} - \frac{\delta_{P(j)\to P(i)}}{(n/k)} \right) \right) \end{aligned} \tag{17}$$

We begin by describing the choice of Z^*. If vertex i and j belong to the same clusters, then $Z^*[i,j] = 0$; otherwise,

$$\begin{aligned} Z^*[i,j] &\triangleq \left(-\frac{\bar{\delta}_{i\to P(j)}}{n/k} - \frac{\bar{\delta}_{j\to P(i)}}{n/k} + \frac{\bar{\delta}_{P(j)\to P(i)}}{(n/k)(n/k)} - \frac{\bar{\delta}_{min}}{n/k} \right) \\ &= \left(1 - 2\left(\frac{\delta_{i\to P(j)}}{n/k} + \frac{\delta_{j\to P(i)}}{n/k} - \frac{\delta_{P(j)\to P(i)}}{(n/k)(n/k)} \right) - \frac{\bar{\delta}_{min}}{n/k} \right) \end{aligned}$$

Note that by definition (17), Z^* is a symmetric nonnegative matrix. We now define the diagonal matrix D^* as

$$D^*[i,i] \triangleq \bar{\delta}_{in}(i) + \bar{\delta}_{min} = 2\left(\delta_{in}(i) - \max_{i,j}\left(\delta_{i\to P(j)} + \delta_{j\to P(i)} - \frac{\delta_{P(j)\to P(i)}}{(n/k)}\right)\right)$$

A simple calculation now shows the first required property that

$$\text{Trace}(D) + \frac{1}{k-1}\sum_{ij} Z[i,j] = \sum_i \left(\bar{\delta}_{in}(i) - \frac{\sum_{t:P(i)\neq P_t} \bar{\delta}_{i\to P_t}}{k-1}\right) = (2*A(G) - J)\bullet Y^*$$

We now proceed to show that D^*, Z^* is a valid dual solution, i.e.,

$$M^* = D^* - Z^* - (2A - J) \succeq 0$$

To see this consider the following extension of the decomposition of the space $\mathbb{R}^n$ defined in Section 3.3.

Definition 3. Given a k-clustering of n vertices $\{P_t\}_{t=1}^k$, we define the vectors v_t to be the indicator vectors of the clusters. We further define the following subspaces, which are perpendicular to each other, and partition $\mathbb{R}^n$.

- $\nVdash$: the vectors with 1 in each coordinate
- $\mathbb{R}_{k-1}$: the $k-1$ dimensional subspace such that for every vector $v \in \mathbb{R}_{k-1}$, $v(i) = v(j)$ if $P(i) = P(j)$ and $< v, \nVdash >= 0$
- $\mathbb{R}_{n|k}$: the subspace perpendicular to $\mathbb{R}_{k-1} \cup \nVdash$, i.e., the subspace where the sum on each cluster is equal to 0.

The following are two easy observations that follow from simple calculations similar to the calculations shown in Section 3.3.

Observation 1. $(\forall\ \ v \in \mathbb{R}_{k-1})\ (D^* - Z^* - (2A - J))v = 0$

Observation 2. $(\forall\ \ v \in \mathbb{R}_{n|k})\ v^T Z^* v = 0$

We first focus on the subspace $\mathbb{R}_{n|k}$ and show that $\forall x \in \mathbb{R}_{n|k}$

$$x^T(D^* - Z^* - (2A - J)x = x^T(D^* - 2A)x \geq 0 \tag{18}$$

The proof of the above statement follows from the following set of inequalities:

$$\begin{aligned} x^T(D^* - 2A)x &\geq \min_i D^*[i,i] - 2\max_x x^T A(G)x \\ &\geq 2\min_i v(i) - 2\max_x x^T A(G)x \\ &\geq 2\left(\min_i v(i) - \hat{c}\left(\sqrt{pn/k + qn} + + \sqrt{\log(n)}\right)\right) \\ &\geq 0 \end{aligned}$$

where the second inequality above follows from substituting the values of $\bar{\delta}_{i\to P(t)}$ in terms of $\delta_{i\to P(t)}$ in the expression for $D^*[i,i]$ and using the definition of $\nu(i)$. The second inequality follows from Lemma 5, and the third inequality follows from the condition (3). Note that in condition (3) if we assume the constant to be $\hat{c}+1$ instead of $\hat{c}$, then we get a stronger property that the above quantity is in fact greater than $\sqrt{\log(n)}$ and not just positive. We use this below.

The above analysis shows that the matrix $M^* = D^* - Z^* - (2A - J)$ is PSD on the subspace $\mathbb{R}_{n|k}$. Let's now focus on a vector $y \in \mathbb{R}_{n|k} \oplus \not\Vdash$. Let $H^* = D^* - Z^* - 2A = M^* - J$. By appropriate scaling we can consider any $y = x + \delta\frac{\not\Vdash}{\sqrt{n}}$ (see footnote[1]) where $x \in \mathbb{R}_{n|k}$ is a unit vector and $\delta \geq 0$. In the analysis above, we explained that $x^T H^* x \geq \sqrt{\log(n)}\|x\|^2 = \sqrt{\log(n)}$. With these facts in place, consider $y^T M^* y$:

$$\begin{aligned} y^T M^* y &= x^T H^* x + \frac{\delta^2}{n}\not\Vdash^T J \not\Vdash + 2x^T H^* \frac{\delta}{\sqrt{n}}\not\Vdash \\ &\geq \sqrt{\log(n)} + \delta^2 n - 2\delta\|H^*\| \end{aligned}$$

where we use the fact that for unit vector x $\frac{x^T H^* \not\Vdash}{\sqrt{n}} \leq \|H^*\|$. Therefore, as long as we have that $4\|H^*\|^2 \leq 4n\sqrt{\log(n)}$, we have that $y^T M^* y \geq 0$ (as the expression is a quadratic in δ). Therefore, we need to control the spectral norm of H^*. We can show the above via very simple and fairly loose calculations:

$$\begin{aligned} \|H^*\| &\leq \|D^*\| + 2\|A\| + \|Z^*\| \\ &\leq \max D^*[i,i] + 2\delta_{max} + O(\delta_{max}) \\ &\leq O(\delta_{max}) \end{aligned}$$

where δ_{max} is the degree of the vertex with maximum degree in the graph G. The above equation follows with very loose approximations from the definitions. A simple Chernoff bound shows that with high probability, $\delta_{max} \leq pm + kqm + \sqrt{pm + kqm}\log(n) \leq O(k\log(n) + \log^{3/2}(n))$ where we have replaced p with $\alpha\frac{\log(m)}{m}$ and q with $\beta\frac{\log(m)}{m}$ which implies that $\|H^*\| \leq \sqrt{n}$ which completes the proof since we have shown that M^* is PSD. □

[1] Indeed by definition any vector $y \in \mathbb{R}_{n|k} \oplus \not\Vdash$ can be written as $x + \delta\frac{\not\Vdash}{\sqrt{n}}$ for some δ and $x \in \mathbb{R}_{n|k}$. For the purpose of proving positive definiteness, we can always divide by any positive number and can therefore consider $\frac{y}{\|x\|}$. Also note that we can consider y or $-y$ equivalently and hence can consider the case when $\delta > 0$.

4 Note About the Monotone Adversary

In this section, we extend our result to the following semi-random model considered in the paper of Feige and Kilian [19]. We first define a monotone adversary (we define it for the "homophilic" case). Given a graph G and a partition $P = \{P_i\}$, a *monotone* adversary is allowed to take any of the following two actions on the graph:

- Arbitrarily remove edges across clusters, i.e., (u, v) s.t. $P(u) \neq P(v)$.
- Arbitrarily add edges within clusters, i.e., (u, v) s.t. $P(u) = P(v)$.

Given a graph G let G_{adv} be the resulting graph after the adversary's actions. The adversary is monotone in the sense that the set of the optimal multisections in G_{adv} contains the set of the optimal multisections in G. Let $B(G)$ be the number of edges cut in the optimal multisection. We now consider the following semi-random model, where we first randomly pick a graph $G \sim \mathscr{G}_{p,q,k}$ and then the algorithm is given G_{adv} where the monotone adversary has acted on G. The following theorem shows that our algorithm is robust against such a monotone adversary.

Theorem 6. *Given a graph G_{adv} generated by a semi-random model described above, we have that with probability $1-o(1)$ the algorithm described in Section 3.3 recovers the original (hidden) partition. The probability is over the randomness in the production of $G \sim \mathscr{G}_{p,q,k}$ on which the adversary acts.*

Proof. We consider the SDP relaxation (4) as in the proof of Theorem 1. Let $Y^*(G)$ be the optimal solution of the SDP when we run it on the graph G. Now suppose $G \sim \mathscr{G}_{p,q,k}$. The proof of Theorem 1 shows that with high probability, $Y^*(G)$ is unique and it corresponds to the hidden partition. Suppose this event happens, we then show that for any graph G_{adv} generated by the monotone adversary after acting on G, $Y^*(G_{adv})$ is also unique and it is equal to $Y^*(G)$. This will prove Theorem 6.

Define $SDP_G(Y)$ to be the objective value (corresponding to the graph G) of a feasible matrix Y, i.e., $SDP_G(Y) = A(G) \bullet Y$. Note that since Y has only positive entries (since it is a feasible solution), we have that $A(G') \bullet Y \leq A(G) \bullet Y$, if G' is a subgraph of G. Also since $Y \succeq 0$ and its diagonal entries $Y_{ii} = 1$, we have that $|Y_{ij}| \leq 1$. Therefore, $A(G \cup e) \bullet Y \leq A(G) \bullet Y + 2$. Suppose the monotone adversary adds a total of r^+ edges and removes r^- edges. From the monotonicity of the adversary, it is easy to see that $A(G_{adv}) \bullet Y^*(G) = A(G) \bullet Y^*(G) + 2r^+$. However, for any other solution by the argument above, we have that $A(G_{adv}) \bullet Y \leq A(G) \bullet Y + 2r^+$. Also by our assumption, we have that $A(G) \bullet Y^*(G) < A(G) \bullet Y$ for any feasible $Y \neq Y^*(G)$. Putting it together we have that

$$A(G_{adv}) \bullet Y^*(G) = A(G) \bullet Y^*(G) + 2r^+ > A(G) \bullet Y + 2r^+ \geq A(G_{adv}) \bullet Y\,,$$

for any feasible $Y \neq Y^*(G)$, which proves the theorem. □

5 Experimental Evaluation

In this section we present some experimental results on the SDPs presented above. For both of the SDPs, we consider the case of $p = \alpha\frac{\log(m)}{m}$ and $q = \beta\frac{\log(m)}{m}$ with $k = 3$ and $m = 20$. We vary α and β, and for each pair of values, we take 10 independent instances, and the shade of gray in the square represents the fraction of instances for which the SDP was integral with lighter representing higher fractions of integrality. The red lines represent the curve we prove in our main Theorem 1, i.e., $\sqrt{\alpha} - \sqrt{\beta} > 1$.

Figure 1 corroborates our Theorem 1 as for SDP in (4) we observe that experimentally the performance almost exactly mimics what we prove. For the other (possibly) weaker SDP in (2), we see in Figure 2 that the performance is almost similar to the stronger SDP; however we were unable to prove it formally as discussed in Section 2. We leave this as an open question to show that SDP in (2) is integral all the way down to the information theoretic threshold (i.e., $\sqrt{\alpha}-\sqrt{\beta} > 1$). We observe from the experiments above that this indeed seems to be the case.

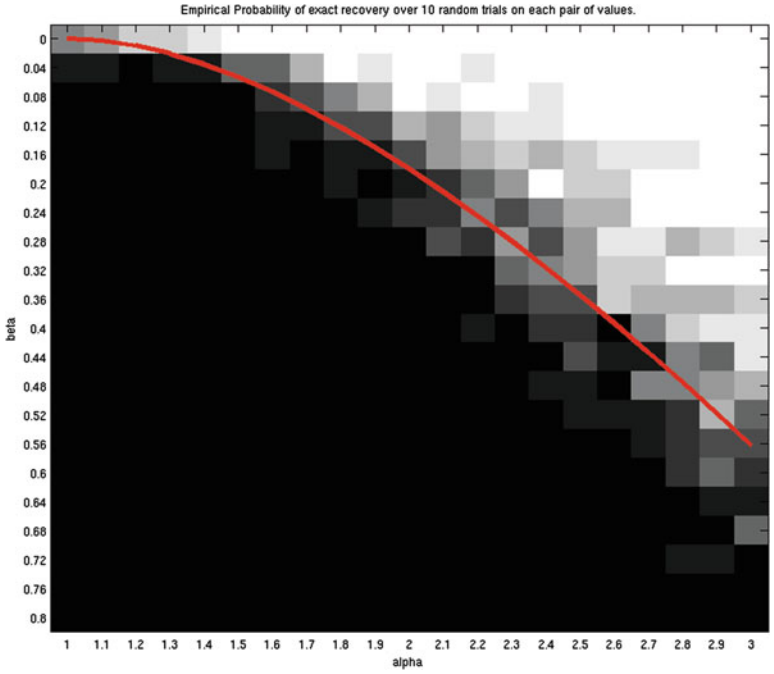


Fig. 1 Performance of SDP in (4). We consider the case of $p = \alpha\frac{\log(m)}{m}$ and $q = \beta\frac{\log(m)}{m}$ with $k = 3$ and $m = 20$. We vary α and β, and for each pair of values, we take 10 independent instances, and the shade of gray in the square represents the fraction of instances for which the SDP was integral with lighter representing higher fractions of integrality. The red line represents the curve we prove in our main Theorem 1, i.e., $\sqrt{\alpha} - \sqrt{\beta} > 1$

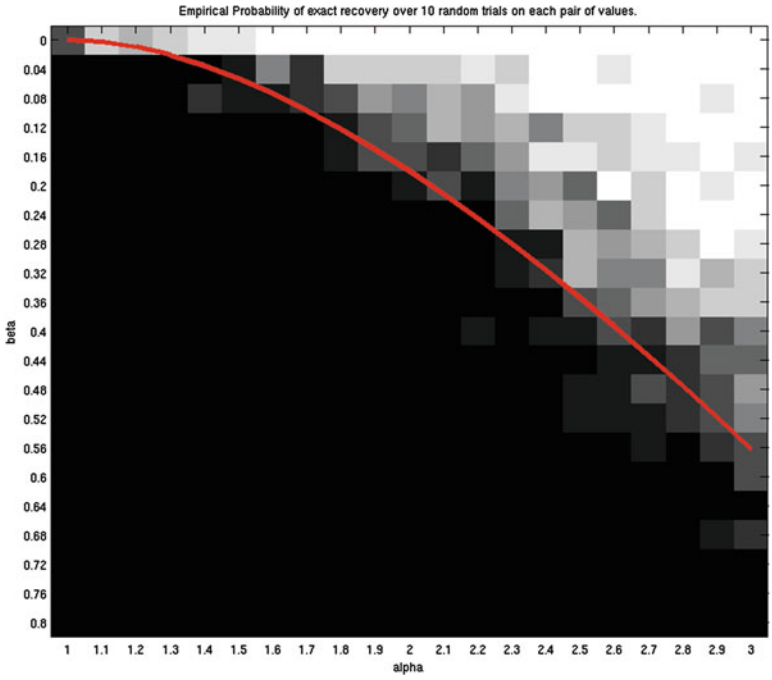


Fig. 2 Performance of SDP in (2). We consider the case of $p = \alpha \frac{\log(m)}{m}$ and $q = \beta \frac{\log(m)}{m}$ with $k = 3$ and $m = 20$. We vary α and β, and for each pair of values, we take 10 independent instances, and the shade of gray in the square represents the fraction of instances for which the SDP was integral with lighter representing higher fractions of integrality. The red line represents the curve we prove in Theorem 1, for the SDP (4), i.e., $\sqrt{\alpha} - \sqrt{\beta} > 1$

6 The Multireference Alignment SDP for Clustering

In this section we describe an interesting connection between the SDPs used for clustering and partitioning problems and others such as the ones used for the multireference signal alignment and the unique games problems.

For illustrative purposes we will consider a slightly different version of the balanced k-cut (multisection) problem described earlier. Instead of imposing that the graph is partitioned in equal sized clusters, we will consider the objective value to be maximized to be the difference between the number of *agreeing pairs* and *disagreeing pairs* where an agreeing pair is a pair of nodes connected by an edge that was picked to be in the same cluster or a pair of points not connected by an edge that is not in the same cluster, and disagreeing pairs are all the others. Note that, if the balanced partition constraint was enforced, this objective would be equivalent to the multisection one.

The multireference alignment problem in signal processing [9] consists of aligning n signals $y_1, \ldots, y_n$ with length k that are copies of a single signal but have been shifted and corrupted with white Gaussian noise. For $a \in [k]$, we set R_{l_i} to be

the $k \times k$ matrix that shifts the entries of vector by a coordinates. In this notation, the maximum likelihood estimator for the multireference alignment problem is given by the shifts $l_1, \ldots, l_n \in [k]$ that maximize

$$\sum_{i,j=1}^{n} \left\langle R_{l_i}^T y_i, R_{l_j}^T y_j \right\rangle = \sum_{i,j=1}^{n} \mathrm{Tr}\left[y_j y_i^T R_{l_i} R_{l_j}^T \right]. \tag{19}$$

A fruitful way of thinking about (19) is as a sum, over each pair i, j, of pairwise costs that depends on the choices of shifts for the variable in each pair. An example of a problem of this type is the celebrated unique games problem, and indeed the SDP approach developed in [9] for the multireference alignment problem is an adaptation of an SDP-based approximation algorithm for the unique games problems by Charikar et al. [13]. The objective in the alignment problem (19) has, however, an important property – the pairwise costs only depend on the relative choices of shifts. More precisely, both l_i and l_j being increased by the same amount have no effect on the pairwise cost relative to (i, j). In fact, there is a general framework for solving problems with this group invariance-type property, called nonunique games, when the group involved is compact [10]. The example above and SDP (2) that we will derive below are particular cases of this framework, but it is more enlightening to derive the SDP we will use for partitioning from the multireference alignment one.

To obtain an SDP for the partitioning problem, one can think of each node i as a signal y_i in $\mathbb{R}^k$ and think of a shift label as a cluster membership; the cost associated to the pair i, j should then, if the nodes are connected, $+1$ if the two signals are given the same shift, and -1 otherwise; if the nodes are not connected, it should be -1 if the two signals are given the same shift and $+1$ otherwise. This can be achieved by replacing $y_j y_i^T$ on the objective (19) by appropriate $k \times k$ matrices $C_{ij}^T = \frac{1}{k}\left(2I - \mathbf{1}\mathbf{1}^T\right)$ if i and j are connected and $C_{ij}^T = \frac{1}{k}\left(\mathbf{1}\mathbf{1}^T - 2I\right)$ if not. Our objective would then be

$$\sum_{a=1}^{k} \sum_{i,j \in \mathscr{C}_a} d_{ij} = -\sum_{i,j \in [n]} \mathrm{Tr}\left[C_{ij}^T R_{l_i} R_{l_j}^T \right],$$

where R_{l_i} is constrained to be a circulant permutation matrix (a shift operator).

The SDP relaxation proposed in [9] would then take the form

$$\begin{aligned} \max \ & \mathrm{Tr}(CX) \\ \text{s. t.} \ & X_{ii} = I_{k \times k} \\ & X_{ij}\mathbf{1} = \mathbf{1} \\ & X_{ij} \text{ is circulant} \\ & X \geq 0 \\ & X \succeq 0, \end{aligned} \tag{20}$$

In this section $X \geq 0$ for a matrix refers to entrywise ≥ 0.

It is clear, however, that (20) has many optimal solutions. Given an optimal selection of cluster labelings, any permutation of these labels will yield a solution with the same objective. For that reason we can adapt the SDP to consider the average of such solutions. This is achieved by restricting each block X_{ij} to be a linear combination of $I_{k\times k}$ and $\mathbf{1}\mathbf{1}^T$ (meaning that it is constant both on the diagonal and on the off-diagonal). Adding that constraint yields the following SDP.

$$\begin{aligned} \max\ & \mathrm{Tr}(CX) \\ \text{s. t. }\ & X_{ii} = I_{k\times k} \\ & X_{ij}\mathbf{1} = \mathbf{1} \\ & X_{ij} \text{ is circulant} \\ & \left(X_{ij}\right)_{aa} = \left(X_{ij}\right)_{11} \\ & \left(X_{ij}\right)_{ab} = \left(X_{ij}\right)_{12}, \ \forall_{a\neq b} \\ & X \succeq 0 \\ & X \succeq 0, \end{aligned} \tag{21}$$

Since the constraints in (21) imply

$$\left(X_{ij}\right)_{11} + (k-1)\left(X_{ij}\right)_{12} = 1,$$

(21) can be described completely in terms of the variables $\left(X_{ij}\right)_{11}$. For that reason we consider the matrix $Z \in \mathbb{R}^{n\times n}$ with entries $Z_{ij} = \left(X_{ij}\right)_{11}$. We can then rewrite (21) as

$$\begin{aligned} \max\ & \mathrm{Tr}\left(\tilde{C}Z\right) \\ \text{s. t. }\ & Z_{ii} = 1 \\ & Z \geq 0 \\ & Z^{(k)} \succeq 0, \end{aligned} \tag{22}$$

where $\tilde{C}_{ij} = kC_{ij}$ and $Z^{(k)}$ is the $nk \times nk$ matrix whose $n \times n$ diagonal blocks are equal to Z and whose $n \times n$ non-diagonal blocks are equal to $\frac{11^T - Z}{k-1}$. For example,

$$Z^{(2)} = \begin{bmatrix} Z & 11^T - Z \\ 11^T - Z & Z \end{bmatrix} \quad \text{and} \quad Z^{(3)} = \begin{bmatrix} Z & \frac{11^T - X}{2} & \frac{11^T - Z}{2} \\ \frac{11^T - Z}{2} & Z & \frac{11^T - Z}{2} \\ \frac{11^T - Z}{2} & \frac{11^T - Z}{2} & Z \end{bmatrix}.$$

The following lemma gives a simpler characterization for the intriguing $Z^{(k)} \succeq 0$ constraint.

Lemma 8. *Let Z be a symmetric matrix and $k \geq 2$ an integer. $Z^{(k)} \succeq 0$ if and only if $Z \succeq \frac{1}{k}\mathbf{1}\mathbf{1}^T$.*

Before proving Lemma 8, we note that it implies that we can succinctly rewrite (22) as

$$\begin{aligned} \max \ & \mathrm{Tr}\left(\tilde{C}Z\right) \\ \text{s. t. } & Z_{ii} = 1 \\ & Z \succeq 0 \\ & Z \succeq \tfrac{1}{k}\mathbf{1}\mathbf{1}^T. \end{aligned} \tag{23}$$

A simple change of variables $Y = \frac{k}{k-1}Z - \frac{1}{k-1}\mathbf{1}\mathbf{1}^T$ allows one to rewrite (23) as (for appropriate matrix C' and constant c')

$$\begin{aligned} \max \ & \mathrm{Tr}\left(C'Y\right) - c' \\ \text{s. t. } & Y_{ii} = 1 \\ & Y_{ij} \geq -\tfrac{1}{k-1} \\ & Y \succeq 0. \end{aligned} \tag{24}$$

Remarkably,(24) coincides with the classical semidefinite relaxation for the Max-k-Cut problem [20], which corresponds to (2) used in this paper.

Proof (of Lemma 8). Since, in this proof, we will be using $\mathbf{1}$ to refer to the all-ones vector in two different dimensions, we will include a subscript denoting the dimension of the all-ones vector.

The matrix $Z^{(k)}$ is block circulant, and so it can be block diagonalizable by a block DFT matrix, $F_{k\times k} \otimes I_{n\times n}$, where $F_{k\times k}$ is the $k \times k$ (normalized) DFT matrix and $\otimes$ is the Kronecker product. In other words,

$$\left(F_{k\times k} \otimes I_{n\times n}\right) Z^{(k)} \left(F_{k\times k} \otimes I_{n\times n}\right)^T$$

is block diagonal. Furthermore, note that

$$Z^{(k)} = \left(\mathbf{1}_k\mathbf{1}_k^T \otimes \frac{\mathbf{1}_n\mathbf{1}_n^T - Z}{k-\mathbf{1}}\right) - \left(I_{k\times k} \otimes \left[Z - \frac{\mathbf{1}_n\mathbf{1}_n^T - Z}{k-1}\right]\right).$$

Also, it is easy to check that

$$\left(F_{k\times k} \otimes I_{n\times n}\right)\left(I_{k\times k} \otimes \left[Z - \frac{\mathbf{1}_n\mathbf{1}_n^T - Z}{k-1}\right]\right)\left(F_{k\times k} \otimes I_{n\times n}\right)^T = I_{k\times k} \otimes \left[Z - \frac{\mathbf{1}_n\mathbf{1}_n^T - Z}{k-1}\right],$$

and

$$\left(F_{k\times k} \otimes I_{n\times n}\right)\left(\mathbf{1}_k\mathbf{1}_k^T \otimes \frac{\mathbf{1}_n\mathbf{1}_n^T - Z}{k-1}\right)\left(F_{k\times k} \otimes I_{n\times n}\right)^T = k\left(e_1e_1^T \otimes \frac{\mathbf{1}_n\mathbf{1}_n^T - Z}{k-1}\right),$$

This means that $(F_{k\times k} \otimes I_{n\times n})\, Z^{(k)}\, (F_{k\times k} \otimes I_{n\times n})^T$ is a block diagonal matrix with the first block equal to $\mathscr{A}$ and all other diagonal blocks equal to $\mathscr{B}$ where $\mathscr{A}$ and $\mathscr{B}$ are given by

$$\mathscr{A} = Z - \frac{\mathbf{1}_n\mathbf{1}_n^T - Z}{k-1} + k\frac{\mathbf{1}_n\mathbf{1}_n^T - Z}{k-1} = \mathbf{1}_n\mathbf{1}_n^T \text{ and } \mathscr{B} = Z - \frac{\mathbf{1}_n\mathbf{1}_n^T - Z}{k-1}.$$

Thus, the condition $Z^{(k)} \succeq 0$ is equivalent to $Z - \frac{\mathbf{1}_n\mathbf{1}_n^T - Z}{k-1} \succeq 0$ which can be rewritten as

$$Z - \frac{1}{k}\mathbf{1}_n\mathbf{1}_n^T \succeq 0.$$

□

Acknowledgements Most of the work presented in this paper was conducted while ASB was at Princeton University and partly conducted while ASB was at the Massachusetts Institute of Technology. ASB acknowledges support from AFOSR Grant No. FA9550-12-1-0317, NSF DMS-1317308, NSF DMS-1712730, and NSF DMS-1719545.

Appendix

Forms of Chernoff Bounds and Hoeffding Bounds Used in the Arguments

Theorem 7 (Chernoff). *Suppose $X_1 \ldots X_n$ be independent random variables taking values in $\{0, 1\}$. Let X denote their sum and let $\mu = \mathbb{E}[X]$ be its expectation. Then for any $\delta > 0$ it holds that*

$$\mathbb{P}\left(X > (1+\delta)\mu\right) < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}, \tag{25}$$

$$\mathbb{P}\left(X < (1-\delta)\mu\right) < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}. \tag{26}$$

A simplified form of the above bound is the following formula (for $\delta \leq 1$)

$$\mathbb{P}\left(X \geq (1+\delta)\mu\right) \leq e^{-\frac{\delta^2\mu}{3}},$$

$$\mathbb{P}\left(X \leq (1-\delta)\mu\right) \leq e^{-\frac{\delta^2\mu}{2}}.$$

Theorem 8 (Bernstein). *Suppose $X_1 \ldots X_n$ be independent random variables taking values in $[-M, M]$. Let X denote their sum and let $\mu = \mathbb{E}[X]$ be its expectation, then*

$$\mathbb{P}\left(|X-\mu| \geq t\right) \leq \exp\left(-\frac{1}{2}\frac{t^2}{\sum_i \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] + Mt/3}\right).$$

Corollary 1. *Suppose $X_1 \ldots X_n$ are i.i.d Bernoulli variables with parameter p. Let $\sigma = \sigma(X_i) = p(1-p)$; then we have that for any $r \geq 0$*

$$\mathbb{P}\left(X \geq \mu + \alpha\sigma\sqrt{n\log(r)} + \alpha\log(r)\right) \leq e^{-\frac{\alpha\log(r)}{4}}.$$

Proof. We have that $n\sigma^2 = np(1-p)$ and $M = 1$. We can now choose $t = \alpha\sigma\sqrt{n\log(r)} + \alpha\log(r)$. This implies that $\frac{n\sigma^2 + t/3}{t^2} \leq \frac{1}{\log(r)}\left(1/\alpha^2 + 1/3\alpha\right) \leq \frac{2}{\alpha\log(r)}$ which implies from Theorem 8 that $\mathbb{P}\left(X > \mu + \alpha\sigma\sqrt{n\log(r)} + \alpha\log(r)\right) \leq e^{-\frac{\alpha\log(r)}{4}}$. □

Theorem 9 (Hoeffding). *Let $X_1 \ldots X_n$ be independent random variables. Assume that the X_i are bounded in the interval $[a_i, b_i]$. Define the empirical mean of these variables as*

$$\bar{X} = \frac{\sum_i \bar{X}_i}{n},$$

then

$$\mathbb{P}\left(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t\right) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{27}$$

References

1. E. Abbe, C. Sandon, Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms (2015). Available online at arXiv:1503.00609 [math.PR]
2. E. Abbe, A.S. Bandeira, G. Hall, Exact recovery in the stochastic block model (2014). Available online at arXiv:1405.3267 [cs.SI]
3. N. Alon, N. Kahale, A spectral technique for coloring random 3-colorable graphs. SIAM J. Comput. **26**(6), 1733–1748 (1997)
4. N. Alon, M. Krivelevich, B. Sudakov, Finding a large hidden clique in a random graph, in *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, 25–27 January 1998, San Francisco, CA (1998), pp. 594–598
5. E. Arias-Castro, N. Verzelen, Community detection in random networks (2013). Available online at arXiv:1302.7099 [math.ST]
6. P. Awasthi, A.S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, R. Ward, Relax, no need to round: integrality of clustering formulations, in *6th Innovations in Theoretical Computer Science (ITCS 2015)* (2015)
7. A.S. Bandeira, Random Laplacian matrices and convex relaxations (2015). Available online at arXiv:1504.03987 [math.PR]

8. A.S. Bandeira, R.V. Handel, Sharp nonasymptotic bounds on the norm of random matrices with independent entries. Ann. Probab. **44**(4), 2479–2506 (2016)
9. A.S. Bandeira, M. Charikar, A. Singer, A. Zhu, Multireference alignment using semidefinite programming, in *5th Innovations in Theoretical Computer Science (ITCS 2014)* (2014)
10. A.S. Bandeira, Y. Chen, A. Singer, Non-unique games over compact groups and orientation estimation in cryo-em (2015). Available at arXiv:1505.03840 [cs.CV]
11. R.B. Boppana, Eigenvalues and graph bisection: an average-case analysis, in *Proceedings of the 28th Annual Symposium on Foundations of Computer Science, SFCS '87*, Washington, DC (IEEE Computer Society, Washington, 1987), pp. 280–285
12. T.N. Bui, S. Chaudhuri, F.T. Leighton, M. Sipser, Graph bisection algorithms with good average case behavior, in *25th Annual Symposium on Foundations of Computer Science*, 24–26 October 1984, West Palm Beach, FL (1984), pp. 181–192
13. M. Charikar, K. Makarychev, Y. Makarychev, Near-optimal algorithms for unique games, in *Proceedings of the Thirty-eighth Annual ACM Symposium on Theory of Computing, STOC '06*, New York, NY (ACM, New York, 2006), pp. 205–214
14. Y. Chen, J. Xu, Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices (2014). Available online at arXiv:1402.1267 [stat.ML]
15. P. Chin, A. Rao, V. Vu, Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery (2015). Available online at: arXiv:1501.05021
16. A. Condon, R.M. Karp, Algorithms for graph partitioning on the planted partition model. Random Struct. Algor. **18**(2), 116–140 (2001)
17. A. Decelle, F. Krzakala, C. Moore, L. Zdeborová, Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. Phys. Rev. E **84**, 066106 (2011)
18. P. Erdös, A. Renyi, On random graphs. I. Publ. Math. **6**, 290–297 (1959)
19. U. Feige, J. Kilian, Heuristics for semirandom graph problems. J. Comput. Syst. Sci. **63**(4), 639–671 (2001)
20. A.M. Frieze, M. Jerrum, Improved approximation algorithms for max k-cut and max bisection, in *Proceedings of the 4th International IPCO Conference on Integer Programming and Combinatorial Optimization* (Springer-Verlag, London, 1995), pp. 1–13
21. B. Hajek, Y. Wu, J. Xu, Achieving exact cluster recovery threshold via semidefinite programming (2014). Available online at arXiv:1412.6156 [stat.ML]
22. B. Hajek, Y. Wu, J. Xu, Achieving exact cluster recovery threshold via semidefinite programming: extensions (2015). Available online at arXiv:1502.07738 [stat.ML]
23. R. Krauthgamer, J. Naor, R. Schwartz, Partitioning graphs into balanced components, in *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '09* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2009), pp. 942–949
24. K. Makarychev, Y. Makarychev, A. Vijayaraghavan, Constant factor approximation for balanced cut in the PIE model, in *Symposium on Theory of Computing, STOC 2014*, New York, NY, May 31–June 03 (2014), pp. 41–49
25. L. Massoulié, Community detection thresholds and the weak Ramanujan property, in *Symposium on Theory of Computing, STOC 2014*, New York, NY, May 31–June 03 (2014), pp. 694–703
26. F. McSherry, Spectral partitioning of random graphs, in *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science, FOCS '01* (IEEE Computer SocietyWashington, DC, 2001), p. 529
27. E. Mossel, J. Neeman, A. Sly, Stochastic block models and reconstruction (2012). Available online at arXiv:1202.1499
28. E. Mossel, J. Neeman, A. Sly, A proof of the block model threshold conjecture (2013). Available online at arXiv:1311.4115
29. E. Mossel, J. Neeman, A. Sly, Belief propagation, robust reconstruction and optimal recovery of block models, in *Proceedings of The 27th Conference on Learning Theory, COLT 2014*, Barcelona, June 13–15 (2014), pp. 356–370

30. E. Mossel, J. Neeman, A. Sly, Consistency thresholds for binary symmetric block models (2014). Available online at arXiv: 1407.1591
31. V. Vu, A simple SVD algorithm for finding hidden partitions. Available online at arXiv: 1404.3918 (2014)
32. S.-Y. Yun, A. Proutiere, Accurate community detection in the stochastic block model via spectral algorithms (2014). Available online at arXiv: 1412.7335

Recovering Signals with Unknown Sparsity in Multiple Dictionaries

Rizwan Ahmad and Philip Schniter

Abstract Motivated by the observation that a given signal x may admit sparse representations in multiple dictionaries Ψ_d, but with varying levels of sparsity across dictionaries, we propose two new algorithms for signal reconstruction from noisy linear measurements. Our first algorithm extends the well-known basis pursuit denoising algorithm from the L1 regularizer $\|\Psi x\|_1$ to composite regularizers of the form $\sum_d \lambda_d \|\Psi_d x\|_1$ while self-adjusting the regularization weights λ_d. Our second algorithm extends the well-known iteratively reweighted L1 algorithm to the same family of composite regularizers. For each algorithm, we provide several interpretations: i) majorization-minimization (MM) applied to a non-convex log-sum-type penalty, ii) MM applied to an approximate ℓ_0-type penalty, iii) MM applied to Bayesian MAP inference under a particular hierarchical prior, and iv) variational expectation-maximization (VEM) under a particular prior with deterministic unknown parameters.A detailed numerical study suggests that, when compared to their non-composite counterparts, our composite algorithms yield significantly improvements in accuracy with only modest increases in computational complexity.

Keywords Composite regularization · Iterative reweighting algorithms · Majorization minimization · Sparse optimization · Variational inference · Bayesian inference

R. Ahmad
Department of Biomedical Engineering, The Ohio State University, Columbus, OH, USA
e-mail: ahmad.46@osu.edu

P. Schniter (✉)
Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA
e-mail: schniter.1@osu.edu

H. Boche et al. (eds.), *Compressed Sensing and its Applications*, Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-69802-1_5

1 Introduction

Consider the problem of recovering a signal or image $x \in \mathbb{C}^n$ from noisy linear measurements

$$y = \Phi x + e \in \mathbb{C}^m, \tag{1}$$

where the measurement operator $\Phi \in \mathbb{C}^{m \times n}$ is known and $e \in \mathbb{C}^m$ is additive noise. Such problems arise in imaging, communications, speech, radar, machine learning, and many other applications. We are particularly interested in the case where $m \ll n$, under which x cannot be uniquely determined from the measurements y, even in the absence of noise. This latter situation arises in many of the applications mentioned earlier, and it has recently been popularized under the framework of *compressive sensing* (CS) [12, 22, 27].

1.1 ℓ_2-Constrained Regularization

By incorporating (partial) prior knowledge about the signal and noise power, it may be possible to accurately recover x from $m \ll n$ measurements y. In this work, we consider signal recovery based on optimization problems of the form

$$\arg\min_x R(x) \;\; \text{s.t.} \;\; \|y - \Phi x\|_2 \leq \varepsilon, \tag{2}$$

where $\varepsilon \geq 0$ is a data-fidelity tolerance that reflects prior knowledge of the noise power and $R(x)$ is a penalty, or regularization, that reflects prior knowledge about the signal x [35]. We briefly summarize several common instances of $R(x)$ below.

1. If x is known to be *sparse* (i.e., contains sufficiently few non-zero coefficients) or approximately sparse, then one would ideally like to use the ℓ_0 penalty (i.e., counting "norm") $R(x) = \|x\|_0 \triangleq |\operatorname{supp}(x)|$. However, since this choice makes (2) NP-hard, it is rarely considered in practice.
2. The ℓ_1 penalty, $R(x) = \|x\|_1 = \sum_{j=1}^n |x_j|$, is a commonly used surrogate to ℓ_0 that makes (2) convex and thus solvable in polynomial time. Under this penalty, (2) is known as *basis pursuit denoising* [17] or as the *lasso* [44]. It is commonly used in *synthesis-based* CS [12, 22, 27].
3. Non-convex surrogates to the ℓ_0 penalty have also been proposed. Well-known varieties include the ℓ_p penalty $R(x) = \|x\|_p^p = \sum_{j=1}^n |x_j|^p$ with $p \in (0, 1)$, and the log-sum penalty $R(x) = \sum_{j=1}^n \log(\delta + |x_j|)$ with $\delta \geq 0$. Although (2) becomes difficult to solve exactly in a guaranteed manner, it can be approximated, leading to excellent empirical performance. Further details will be given below.
4. The choice $R(x) = \|\Psi x\|_1$, with known matrix $\Psi \in \mathbb{C}^{L \times n}$, is familiar from *analysis-based* CS [21]. Penalties of this form are appropriate when prior knowledge suggests that the transform coefficients Ψx are (approximately) sparse, as opposed to the signal x itself being sparse. In this case, (2) can be

solved by the *generalized lasso* [45]. When Ψ is a finite-difference operator, $\|\Psi x\|_1$ yields anisotropic *total variation* regularization [42].

5. Non-convex penalties can also be placed on the transform coefficients Ψx, leading to, e.g., $R(x) = \|\Psi x\|_p^p = \sum_{l=1}^{L} |\psi_l^T x|^p$ with $p \in (0,1)$ or $R(x) = \sum_{l=1}^{L} \log(\delta + |\psi_l^T x|)$ with $\delta \geq 0$, where ψ_l^T denotes the lth row of Ψ.

With a non-convex penalty $R(x)$, a popular approach to solving (2) is through *iteratively reweighted* $\ell_!$ (IRW-L1) [13, 46]. There, (2) with a fixed non-convex $R(x)$ is approximated by a sequence of convex problems

$$x^{(t)} = \arg\min_x R^{(t)}(x) \;\text{ s.t. }\; \|y - \Phi x\|_2^2 \leq \varepsilon \tag{3}$$

with $R^{(t)}(x) = \sum_{j=1}^{n} w_j^{(t)} |x_j|$ a weighted ℓ_1 norm, where the weights $w^{(t)}$ are computed from the previous estimate $x^{(t-1)}$. When $w_j^{(t)} = (\delta + |x_j^{(t-1)}|)^{-1}$ for a small constant $\delta \geq 0$, the IRW-L1 algorithm can be interpreted [13] as a majorization-minimization (MM) [29] approach to (2) under the *log-sum* penalty $R(x) = \sum_{j=1}^{n} \log(\delta + |x_j|)$, which can be considered as a non-convex surrogate to the ℓ_0 penalty. Various empirical and theoretical studies [13, 30, 46] of this latter case have shown performance surpassing that of the ℓ_1 penalty. Unconstrained formulations of IRW-L1 based on "$\arg\min_x R^{(t)}(x) + \gamma \|y - \Phi x\|_2^2$" have also been considered, such as in the seminal work [25]. Likewise, constrained and unconstrained versions of iteratively reweighted ℓ_2 were considered in [16, 19, 23, 25, 46]. See [35] for further discussion.

1.2 Sparsity-Inducing Composite Regularizers

In this work, we focus on sparsity-inducing *composite* regularizers of the form

$$R_1(x) \triangleq \sum_{d=1}^{D} \lambda_d \|\Psi_d x\|_1, \tag{4}$$

where each $\Psi_d \in \mathbb{C}^{L_d \times n}$ is a known analysis operator and $\lambda_d \geq 0$ is its regularization weight. Our goal is to recover the signal x from measurements (1) using a constrained optimization (2) under the composite regularizer (4). Doing so requires an optimization of the weights $\lambda \triangleq [\lambda_1, \ldots, \lambda_D]^T$ in (4). We are also interested in iteratively re-weighted extensions of this problem that, at iteration t, use composite regularizers of the form[1]

$$R^{(t)}(x) = \sum_{d=1}^{D} \lambda_d^{(t)} \|W_d^{(t)} \Psi_d x\|_1, \tag{5}$$

[1] Although (5) is over-parameterized, the form of (5) is convenient for algorithm development.

where $W_d^{(t)}$ are diagonal matrices. This latter approach requires the optimization of both $\lambda_d^{(t)}$ and $W_d^{(t)}$ for all d.

As a motivating example, suppose that $\{\Psi_d\}$ is a collection of orthonormal bases that includes, e.g., spikes, sines, and various wavelet bases. The signal x may be sparse in some of these bases, but not all. Thus, we would like to adjust each λ_d in (4) to appropriately weight the contribution from each basis. But it is not clear how to do this when x is unknown. As another example, suppose that x contains a (rasterized) sequence of images and that $\|\Psi_1 x\|_1$ measures temporal total-variation while $\|\Psi_2 x\|_1$ measures spatial total-variation. Intuitively, we would like to weight these two regularizations differently, depending on whether the image varies more in the temporal or spatial dimensions. But it is not clear how to do this when x is unknown.

1.3 Contributions

In this work, we propose novel iteratively reweighted approaches to sparse reconstruction based on composite regularizations of the form (4)–(5) with automatic tuning of the regularization weights λ and W_d. For each of our proposed algorithms, we will provide four interpretations:

1. MM applied to a non-convex log-sum-type penalty,
2. MM applied to an approximate ℓ_0-type penalty,
3. MM applied to Bayesian MAP inference based on Gamma and Jeffrey's hyper-priors [7, 24, 37], and
4. variational expectation maximization (VEM) [8, 36] applied to a Laplacian or generalized-Pareto prior with deterministic unknown parameters.

We show that the MM interpretation guarantees convergence in the sense of satisfying an asymptotic stationary point condition [34]. Moreover, we establish connections between our proposed approaches and existing IRW-L1 algorithms, and we provide novel VEM-based and Bayesian MAP interpretations of those existing algorithms.

Finally, through the detailed numerical study in Section 4, we establish that our proposed algorithms yield significant gains in recovery accuracy relative to existing methods with only modest increases in runtime. In particular, when $\{\Psi_d\}$ are chosen so that the sparsity of $\Psi_d x$ varies with d, this structure can be exploited for improved recovery. The more disparate the sparsity, the greater the improvement.

1.4 Related Work

As discussed above, the generalized lasso [45] is one of the most common approaches to L1-regularized analysis-CS [21], i.e., the optimization (2) under the

regularizer $R(x) = \|\Psi x\|_1$. The Co-L1 algorithm that we present in Section 2 can be interpreted as a generalization of this L1 method to *composite* regularizers of the form (4). Meanwhile, the iteratively reweighted extension of the generalized lasso, IRW-L1 [13], often yields significantly better reconstruction accuracy with a modest increase in complexity (e.g., [13, 14]). The Co-IRW-L1 algorithm that we present in Section 3 can then be interpreted as a generalization of this IRW-L1 method to *composite* regularizers of the form (5). The existing non-composite L1 and IRW-L1 approaches essentially place an identical weight $\lambda_d = 1$ on every term in (4)–(5) and thus make no attempt to leverage differences in the sparsity of the transform coefficients $\Psi_d x$ across the sub-dictionary index d. However, the numerical results that we present in Section 4 suggest that there can be significant advantages to optimizing λ_d, which is precisely what our methods do.

The problem of optimizing the weights λ_d of composite regularizers $R(x;\lambda) = \sum_d \lambda_d R_d(x)$ is a long-standing problem with a rich literature (see, e.g., the recent book [33]). However, the vast majority of that literature focuses on the Tikhonov case where $R_d(x)$ are quadratic (see, e.g., [11, 26, 28, 47]). One notable exception is [6], which assumes continuously differentiable $R_d(x)$ and thus does not cover our composite ℓ_1 prior (4). Another notable exception is [32], which assumes i) the availability of a noiseless training example of x to help tune the L1 regularization weights λ in (4), and ii) the trivial measurement matrix $\Phi = I$. In contrast, our proposed methods operate without any training and support generic measurement matrices Φ.

In the special case that each Ψ_d is composed of a subset of rows from the $n \times n$ identity matrix, the regularizers (4)–(5) can induce *group* sparsity in the recovery of x, in that certain sub-vectors $x_d \triangleq \Psi_d x$ of x are driven to zero, while others are not. The paper [40] develops an IRW-L1-based approach to group-sparse signal recovery for equal-sized non-overlapping groups that can be considered as a special case of the Co-L1 algorithm that we develop in Section 2. However, our approach is more general in that it handles possibly non-equal and/or overlapping groups, not to mention sparsity in a generic set of sub-dictionaries Ψ_d. Recently, Bayesian MAP group-sparse recovery was considered in [4]. However, the technique described there uses Gaussian scale mixtures or, equivalently, weighted-ℓ_2 regularizers $R(x;\lambda) = \sum_d \lambda_d \|x_d\|_2$, while our methods use weighted-ℓ_1 regularizers (4)–(5).

A recent work [2] considered the *unconstrained* version of the problem considered in this chapter, where the aim is to solve a non-convex optimization problem of the form

$$\arg\min_x R(x) + \gamma \|y - \Phi x\|_2, \tag{6}$$

for some $\gamma > 0$, through a sequence of convex problems

$$x^{(t)} = \arg\min_x \sum_{d=1}^{D} \lambda_d^{(t)} \|W_d^{(t)} \Psi_d x\|_1 + \gamma \|y - \Phi x\|_2^2, \tag{7}$$

where $\{\lambda_d^{(t)}, W_d^{(t)}\}_{d=1}^D$ are set using $x^{(t-1)}$. Although the unconstrained case bears some similarity to the *constrained* case considered in this chapter, each case leads to a distinct set of algorithms, interpretations, and analyses.

1.5 Notation

We use capital letters like Ψ for matrices, small letters like x for vectors, and $(\cdot)^T$ for transposition. We use $\|x\|_p \triangleq (\sum_j |x_j|^p)^{1/p}$ for the ℓ_p norm of vector x, with x_j representing the nth coefficient in x. When referring to the "mixed $\ell_{p,q}$ norm" of a matrix X, we mean $(\sum_d (\sum_l |x_{d,l}|^p)^{q/p})^{1/q}$ as in [31], where $x_{d,l}$ is the dth row and lth column of X. We adopt the index-set abbreviation $[D] \triangleq \{1, \ldots, D\}$ and use I to denote the identity matrix. We use $\nabla g(x)$ for the gradient of a functional $g(x)$ with respect to x and 1_A for the indicator function that returns the value 1 when A is true and 0 when A is false. We use $p(x; \lambda)$ for the pdf of random vector x under deterministic parameters λ and $p(x|\lambda)$ for the pdf of x conditioned on the random vector λ. We use $D_{\mathsf{KL}}(q\|p)$ to denote the Kullback-Leibler (KL) divergence of the pdf p from the pdf q, and we use $\mathbb{R}$ and $\mathbb{C}$ to denote the real and complex fields, respectively.

2 The Co-L1 Algorithm

We first propose the Composite-L1 (Co-L1) algorithm, which is summarized in Algorithm 1. There, L_d denotes the number of rows in Ψ_d.

Algorithm 1 The Co-L1 Algorithm

1: input: $\{\Psi_d\}_{d=1}^D, \Phi, y, \varepsilon \geq 0, \delta \geq 0$
2: initialization: $\lambda_d^{(1)} = 1 \; \forall d$
3: for $t = 1, 2, 3, \ldots$
4: $\quad x^{(t)} \leftarrow \arg\min_x \sum_{d=1}^D \lambda_d^{(t)} \|\Psi_d x\|_1$ s.t. $\|y - \Phi x\|_2 \leq \varepsilon$
5: $\quad \lambda_d^{(t+1)} \leftarrow \dfrac{L_d}{\delta + \|\Psi_d x^{(t)}\|_1}, \; d \in [D]$
6: end
7: output: $x^{(t)}$

The main computational step of Co-L1 is the constrained ℓ_1 minimization in line 4, which can be recognized as (2) under the composite regularizer R_1 from (4). This is a convex optimization problem that can be readily solved by existing techniques (e.g., Douglas-Rachford splitting [18], ADMM [1, 10], NESTA-UP [5], MFISTA via smoothing and decomposition [43], etc.), the specific choice of which is immaterial to this paper.

Note that Co-L1 requires the user to set a small regularization term $\delta \geq 0$ whose role is to prevent the denominator in line 5 from reaching zero. For typical choices of the analysis operators Ψ_d and ε, the vector $\Psi_d x^{(t)}$ will almost never be exactly zero, in which case it suffices to set $\delta = 0$. Also, Co-L1 requires the user to set the measurement fidelity constraint $\varepsilon \geq 0$. For additive white Gaussian noise (AWGN) of variance $\sigma^2 > 0$, the choice $\varepsilon = 0.8\sqrt{\sigma^2 m}$ works empirically well, and we used this setting for all numerical results in Section 4.

Co-L1's update of the weights λ, defined by line 5 of Algorithm 1, can be interpreted in various ways, as we detail below. For ease of explanation, we first consider the case where the signal x is real-valued and later discuss the complex-valued case in Section 2.6. As we will see, the steps in Algorithm 1 apply to both real- and complex-valued x.

Theorem 1 (Co-L1). *The Co-L1 algorithm in Algorithm 1 has the following interpretations:*

1. *MM applied to (2) under the log-sum penalty*

$$R_{ls}^{D}(x;\delta) \triangleq \sum_{d=1}^{D} L_d \log(\delta + \|\Psi_d x\|_1), \tag{8}$$

2. *as $\delta \to 0$, an approximate solution to the weighted $\ell_{1,0}$ [31] problem*

$$\arg\min_{x} \sum_{d=1}^{D} L_d \, 1_{\|\Psi_d x\|_1 > 0} \;\; s.t. \;\; \|y - \Phi x\|_2 \leq \varepsilon, \tag{9}$$

3. *for $\varepsilon = 0$, MM applied to Bayesian MAP estimation under a noiseless likelihood and the hierarchical prior*

$$p(x|\lambda) = \prod_{d=1}^{D} \left(\frac{\lambda_d}{2}\right)^{L_d} \exp\left(-\lambda_d \|\Psi_d x\|_1\right) \tag{10}$$

$$\lambda \sim i.i.d.\ \Gamma(0, \delta^{-1}) \tag{11}$$

   where $z_d \triangleq \Psi_d x \in \mathbb{R}^{L_d}$ is i.i.d. Laplacian given λ_d, and λ_d is Gamma distributed with scale parameter δ^{-1} and shape parameter zero, which becomes Jeffrey's non-informative hyperprior $p(\lambda_d) \propto 1_{\lambda_d > 0}/\lambda_d$ when $\delta = 0$.
4. *for $\varepsilon = 0$, variational EM under a noiseless likelihood and the prior*

$$p(x;\lambda) \propto \prod_{d=1}^{D} \left(\frac{\lambda_d}{2}\right)^{L_d} \exp\left(-\lambda_d(\|\Psi_d x\|_1 + \delta)\right), \tag{12}$$

   which, when $\delta = 0$, is i.i.d. Laplacian on $z_d = \Psi_d x \in \mathbb{R}^{L_d}$ with deterministic scale parameter $\lambda_d > 0$.

Proof. See Sections 2.1 to 2.5 below.

Importantly, the MM interpretation implies convergence (in the sense of an asymptotic stationary point condition) when $\delta > 0$, as detailed in Section 2.2.

2.1 Log-Sum MM Interpretation of Co-L1

Consider the optimization problem

$$\arg\min_{x} R_{\text{ls}}^{D}(x;\delta) \;\text{ s.t. }\; \|y-\Phi x\|_2 \le \varepsilon \tag{13}$$

with R_{ls}^{D} from (8). Inspired by [13, §2.3], we write (13) as

$$\arg\min_{x,u} \sum_{d=1}^{D} L_d \log\left(\delta + \sum_{l=1}^{L_d} u_{d,l}\right) \;\text{ s.t. }\; \begin{cases} \|y-\Phi x\|_2 \le \varepsilon \\ |\psi_{d,l}^T x| \le u_{d,l} \;\forall d,l, \end{cases} \tag{14}$$

where $\psi_{d,l}^T$ is the lth row of Ψ_d. Problem (14) is of the form

$$\arg\min_{v} g(v) \;\text{ s.t. }\; v \in C, \tag{15}$$

where $v = [u^T, x^T]^T$, C is a convex set,

$$g(v) = \sum_{d=1}^{D} L_d \log\left(\delta + \sum_{k\in K_d} v_k\right) \tag{16}$$

is a concave penalty, and the set $K_d \triangleq \{k : \sum_{d'=1}^{d-1} L_{d'} < k \le \sum_{d'=1}^{d} L_{d'}\}$ contains the indices k such that $v_k \in \{u_{d,l}\}_{l=1}^{L_d}$.

Majorization-minimization (MM) [29, 34] is a popular method to attack non-convex problems of this form. In particular, MM iterates the following two steps: (i) construct a surrogate $g(v; v^{(t)})$ that majorizes $g(v)$ at $v^{(t)}$, and (ii) update $v^{(t+1)} = \arg\min_{v\in C} g(v; v^{(t)})$. By "majorize," we mean that $g(v; v^{(t)}) \ge g(v)$ for all v with equality when $v = v^{(t)}$.

Due to the concavity of our g, we can construct a majorizing surrogate using the tangent of g at $v^{(t)}$. In particular, let ∇g denote the gradient of g w.r.t. v. Then

$$g(v; v^{(t)}) = g(v^{(t)}) + \nabla g(v^{(t)})^T [v - v^{(t)}] \tag{17}$$

majorizes $g(v)$ at $v^{(t)}$, and so the MM iterations become

$$v^{(t+1)} = \arg\min_{v\in C} \nabla g(v^{(t)})^T v \tag{18}$$

after neglecting the v-invariant terms. From (16), we see that

$$[\nabla g(v^{(t)})]_k = \begin{cases} \dfrac{L_{d(k)}}{\delta + \sum_{i \in K_{d(k)}} v_i^{(t)}} & \text{if } d(k) \neq 0 \\ 0 & \text{else,} \end{cases} \tag{19}$$

where $d(k)$ is the index $d \in [D]$ of the set K_d containing k, or 0 if no such set exists. Thus MM prescribes

$$v^{(t+1)} = \arg\min_{v \in C} \sum_{d=1}^{D} \sum_{k \in K_d} \frac{L_d v_k}{\delta + \sum_{i \in K_d} v_i^{(t)}}, \tag{20}$$

or equivalently

$$x^{(t+1)} = \arg\min_{x} \sum_{d=1}^{D} \frac{L_d \sum_{l=1}^{L_d} |\psi_{d,l}^T x|}{\delta + \sum_{l=1}^{L_d} |\psi_{d,l}^T x^{(t)}|} \text{ s.t. } \|y - \Phi x\|_2 \leq \varepsilon \tag{21}$$

$$= \arg\min_{x} \sum_{d=1}^{D} \lambda_d^{(t+1)} \|\Psi_d x\|_1 \text{ s.t. } \|y - \Phi x\|_2 \leq \varepsilon \tag{22}$$

for

$$\lambda_d^{(t+1)} = \frac{L_d}{\delta + \|\Psi_d x^{(t)}\|_1}, \tag{23}$$

which coincides with Algorithm 1. This establishes Part 1 of Theorem 1.

2.2 *Convergence of Co-L1*

The recent paper [34] studies the convergence of MM algorithms. In particular, it establishes that when the optimization objective $g(v)$ is differentiable in $v \in C$ with a Lipschitz continuous gradient, the MM sequence $\{v^{(t)}\}_{t \geq 1}$ satisfies an asymptotic stationary point (ASP) condition. Although it falls short of establishing convergence to a local minimum (which is very difficult for general non-convex optimization problems), the ASP condition is based on a classical necessary condition for a local minimum. In particular, using $\nabla g(v; d)$ to denote the directional derivative of g at v in the direction d, it is known [9] that $v_\star$ locally minimizes g over C only if $\nabla g(v_\star; v - v_\star) \geq 0$ for all $v \in C$. Thus, in [34], it is said that $\{v^{(t)}\}_{t \geq 1}$ satisfies an ASC condition if

$$\liminf_{t \to +\infty} \inf_{v \in C} \frac{\nabla g(v^{(t)}; v - v^{(t)})}{\|v - v^{(t)}\|_2} \geq 0. \tag{24}$$

In our case, g from (16) is indeed differentiable, with gradient given in (19). Moreover, [2, App. A] shows that the gradient is Lipschitz continuous when $\delta > 0$. Thus, the sequence of estimates produced by Algorithm 1 satisfies the ASP condition (24).

2.3 Approximate $\ell_{1,0}$ Interpretation of Co-L1

In the limit of $\delta \to 0$, the log-sum minimization

$$\arg\min_x \sum_{j=1}^{n} \log(\delta + |x_j|) \;\text{ s.t. }\; \|y - \Phi x\|_2 \le \varepsilon \tag{25}$$

is known [41] to be equivalent to ℓ_0 minimization

$$\arg\min_x \|x\|_0 \;\text{ s.t. }\; \|y - \Phi x\|_2 \le \varepsilon. \tag{26}$$

(See [2, App. B] for a proof.) This equivalence can be seen intuitively as follows. As $\delta \to 0$, the contribution to the regularization term $\sum_{j=1}^{n} \log(\delta + |x_j|)$ from each non-zero x_j remains finite, while that from each zero-valued x_j approaches $-\infty$. Since we are interested in minimizing the regularization term, we get a huge reward for each zero-valued x_j, or—equivalently—a huge penalty for each non-zero x_j.

To arrive at an ℓ_0 interpretation of the Co-L1 algorithm, we consider the corresponding optimization problem (13) in the limit that $\delta \to 0$. There we see that the regularization term $R_{\mathsf{ls}}^{D}(x;0)$ from (8) yields L_d huge rewards when $\|\Psi_d x\|_1 = 0$, or equivalently L_d huge penalties when $\|\Psi_d x\|_1 \neq 0$, for each $d \in [D]$. Thus, we can interpret Co-L1 as attempting to solve the optimization problem (9), which is a weighted version of the "$\ell_{p,q}$ mixed norm" problem from [31] for $p = 1$ and $q \to 0$. This establishes Part 2 of Theorem 1.

2.4 Bayesian MAP Interpretation of Co-L1

The MAP estimate [38] of x from y is

$$x_{\mathsf{MAP}} \triangleq \arg\max_x p(x|y) = \arg\min_x \big\{ -\log p(x|y) \big\} \tag{27}$$

$$= \arg\min_x \big\{ -\log p(x) - \log p(y|x) \big\}, \tag{28}$$

where (27) used the monotonicity of log and (28) used Bayes rule. In the case of a noiseless likelihood (e.g., AWGN with variance $\sigma^2 \to 0$), the second term in (28) is $+\infty$ unless $y = \Phi x$, and so

$$x_{\mathsf{MAP}} = \arg\min_x \big\{ -\log p(x) \big\} \;\text{ s.t. }\; y = \Phi x. \tag{29}$$

Recall that, with shape parameter κ and scale parameter θ, the Gamma pdf is $\Gamma(\lambda_d;\kappa,\theta) = 1_{\lambda_d>0}\lambda_d^{\kappa-1}\theta^{-\kappa}\exp(-\lambda_d/\theta)/\Gamma(\kappa)$ where $\Gamma(\kappa)$ is the Gamma function. Since $\Gamma(\lambda_d;\kappa,\theta) \propto 1_{\lambda_d>0}\lambda_d^{\kappa-1}\exp(-\lambda_d/\theta)$, we see that $\Gamma(\lambda_d;0,\infty) \propto 1_{\lambda_d>0}/\lambda_d$, which is Jeffrey's non-informative hyperprior [7, 24, 37] for the Laplace scale parameter λ_d. Then, according to (10)–(11), the prior equals

$$p(x) = \int_{\mathbb{R}^D} p(x|\lambda)p(\lambda)\mathrm{d}\lambda \tag{30}$$

$$\propto \prod_{d=1}^{D}\int_0^\infty \left(\frac{\lambda_d}{2}\right)^{L_d}\exp(-\lambda_d\|\Psi_d x\|_1)\frac{\exp(-\lambda_d\delta)}{\lambda_d}\mathrm{d}\lambda_d \tag{31}$$

$$= \prod_{d=1}^{D}\frac{(L_d-1)!}{\big(2(\|\Psi_d x\|_1+\delta)\big)^{L_d}}, \tag{32}$$

which implies that

$$-\log p(x) = \text{const} + \sum_{d=1}^{D} L_d \log\big(\|\Psi_d x\|_1 + \delta\big). \tag{33}$$

Thus (29), (33), and (8) imply

$$x_{\mathsf{MAP}} = \arg\min_x R_{\mathsf{ls}}^D(x;0) \ \text{ s.t. } \ y = \Phi x. \tag{34}$$

Finally, applying the MM algorithm to this optimization problem (as detailed in Section 2.1), we arrive at the $\varepsilon = 0$ version of Algorithm 1. This establishes Part 3 of Theorem 1.

2.5 *Variational EM Interpretation of Co-L1*

The variational expectation-maximization (VEM) algorithm [8, 36] is an iterative approach to maximum-likelihood (ML) estimation that generalizes the EM algorithm from [20]. We now provide a brief review of the VEM algorithm and describe how it can be applied to estimate λ in (12).

First, note that the log-likelihood can be written as

$$\log p(y;\lambda) = \int q(x)\log p(y;\lambda)\mathrm{d}x \tag{35}$$

$$= \int q(x)\log\left[\frac{p(x,y;\lambda)}{q(x)}\frac{q(x)}{p(x|y;\lambda)}\right]\mathrm{d}x \tag{36}$$

$$= \underbrace{\int q(x) \log \frac{p(x,y;\lambda)}{q(x)} \mathrm{d}x}_{\triangleq F\left(q(x);\lambda\right)} + \underbrace{\int q(x) \log \frac{q(x)}{p(x|y;\lambda)} \mathrm{d}x}_{\triangleq D_{\mathsf{KL}}\left(q(x) \middle\| p(x|y;\lambda)\right)}, \tag{37}$$

for an arbitrary pdf $q(x)$, where $D_{\mathsf{KL}}(q\|p)$ denotes the KL divergence of p from q. Because $D_{\mathsf{KL}}(q\|p) \geq 0$ for any q and p, we see that $F(q(x);\lambda)$ is a lower bound on $\log p(y;\lambda)$. The EM algorithm performs ML estimation by iterating

$$q^{(t)}(x) = \arg\min_{q} D_{\mathsf{KL}}\left(q(x) \middle\| p(x|y;\lambda^{(t)})\right) \tag{38}$$

$$\lambda^{(t+1)} = \arg\max_{\lambda} F(q^{(t)}(x);\lambda), \tag{39}$$

where the "E" step (38) tightens the lower bound and the "M" step (39) maximizes the lower bound.

The EM algorithm places no constraints on $q(x)$, in which case the solution to (38) is simply $q^{(t)}(x) = p(x|y;\lambda^{(t)})$, i.e., the posterior pdf of x under $\lambda = \lambda^{(t)}$. In many applications, however, this posterior is too difficult to compute and/or use in (39). To circumvent this problem, the VEM algorithm constrains $q(x)$ to some family of distributions Q that makes (38)–(39) tractable.

For our application of the VEM algorithm, we constrain to distributions of the form

$$q(x) \propto \lim_{\tau \to 0} \exp\left(\tfrac{1}{\tau} \log p(x|y;\lambda)\right), \tag{40}$$

which has the effect of concentrating the mass in $q(x)$ at its mode. Plugging this $q(x)$ and $p(x,y;\lambda) = p(y|x)p(x;\lambda)$ into (37), we see that the M step (39) reduces to

$$\lambda^{(t+1)} = \arg\max_{\lambda} \log p(x;\lambda)\big|_{x=x^{(t)}_{\mathsf{MAP}}} \tag{41}$$

$$\text{for } x^{(t)}_{\mathsf{MAP}} \triangleq \arg\max_{x} p(x|y;\lambda^{(t)}), \tag{42}$$

where (42) can be interpreted as the E step. For the particular $p(x;\lambda)$ in (12), we have that

$$\log p(x;\lambda) = \text{const} + \sum_{d=1}^{D} \big[L_d \log(\lambda_d) - \lambda_d(\|\Psi_d x\|_1 + \delta)\big], \tag{43}$$

and by zeroing the gradient w.r.t. λ, we find that (41) becomes

$$\lambda_d^{(t+1)} = \frac{L_d}{\left\|\Psi_d x^{(t)}_{\mathsf{MAP}}\right\|_1 + \delta}, \quad d \in [D]. \tag{44}$$

Meanwhile, from the noiseless MAP expression (29) and (43), we find that (42) becomes

$$x_{\mathsf{MAP}}^{(t)} = \arg\min_{x} \sum_{d=1}^{D} \lambda_d^{(t)} \|\Psi_d x\|_1 \;\text{ s.t. }\; y = \Phi x. \tag{45}$$

In conclusion, our VEM algorithm iterates the steps (44)–(45), which match the steps in Algorithm 1 for $\varepsilon = 0$. This establishes Part 4 of Theorem 1.

2.6 Co-L1 for Complex-Valued x

In Theorem 1 and Sections 2.1–2.5, real-valued x was assumed for ease of explanation. However, real-valuedness was employed only in defining the Laplacian pdfs (10) and (12). As we now show, the Co-L1 algorithm in Algorithm 1 can also be justified based on a complex-valued Laplacian pdf. For this, we focus on the VEM interpretation (recall Part 4 of Theorem 1), noting that a similar justification can be made based on the Bayesian MAP interpretation. In particular, we show that, for $\varepsilon = 0$, Algorithm 1 results from VEM inference under an noiseless likelihood and the signal prior

$$p(x;\lambda) \propto \prod_{d=1}^{D} \left(\frac{\lambda_d}{2\pi}\right)^{2L_d} \exp\left(-\lambda_d(\|\Psi_d x\|_1 + \delta)\right), \tag{46}$$

which, when $\delta = 0$, is i.i.d. Laplacian on $z_d = \Psi_d x \in \mathbb{C}^{L_d}$ with deterministic scale parameter $\lambda_d > 0$. To show this, we follow the steps in Section 2.5 up to the log-prior in (43), which now becomes

$$\log p(x;\lambda) = \text{const} + \sum_{d=1}^{D} \big[2L_d \log(\lambda_d) - \lambda_d(\|\Psi_d x\|_1 + \delta)\big]. \tag{47}$$

Zeroing the gradient w.r.t. λ, we find that the VEM update in (41) becomes

$$\lambda_d^{(t+1)} = \frac{2L_d}{\left\|\Psi_d x_{\mathsf{MAP}}^{(t)}\right\|_1 + \delta}, \quad d \in [D], \tag{48}$$

which differs from its real-valued counterpart (44) in a constant scaling of 2. However, this scaling does not affect $x_{\mathsf{MAP}}^{(t+1)}$ in (45) and thus does not affect the output $x^{(t)}$ of Algorithm 1 and thus can be ignored.

2.7 New Interpretations of the IRW-L1 Algorithm

The proposed Co-L1 algorithm is related to the analysis-CS formulation of the well-known IRW-L1 algorithm from [13]. For clarity, and for later use in Section 3, we summarize this latter algorithm in Algorithm 2 and note that the synthesis-CS formulation follows from the special case that $\Psi = I$.

Algorithm 2 The IRW-L1 Algorithm

1: input: $\Psi = [\psi_1, \ldots, \psi_L]^T, \Phi, y, \varepsilon \geq 0, \delta \geq 0$
2: initialization: $W^{(1)} = I$
3: for $t = 1, 2, 3, \ldots$
4: $\quad x^{(t)} \leftarrow \arg\min_x \|W^{(t)}\Psi x\|_1 \;\text{ s.t. }\; \|y - \Phi x\|_2 \leq \varepsilon$
5: $\quad W^{(t+1)} \leftarrow \operatorname{diag}\left\{\frac{1}{\delta + |\psi_1^T x^{(t)}|}, \cdots, \frac{1}{\delta + |\psi_L^T x^{(t)}|}\right\}$
6: end
7: output: $x^{(t)}$

Comparing Algorithm 2 to Algorithm 1, we see that IRW-L1 coincides with Co-L1 in the case that every sub-dictionary Ψ_d has dimension one, i.e., $L_d = 1\ \forall d$ and $D = L$, where $L \triangleq \sum_{d=1}^{D} L_d$ denotes the total number of analysis coefficients. Thus, the Co-L1 interpretations from Theorem 1 can be directly translated to IRW-L1 as follows.

Corollary 1 (IRW-L1). *The IRW-L1 algorithm from Algorithm 2 has the following interpretations:*

1. *MM applied to (2) under the log-sum penalty*

$$R_{ls}^{L}(x;\delta) = \sum_{l=1}^{L} \log(\delta + |\psi_l^T x|), \tag{49}$$

   recalling the definition of R_{ls}^{L} from (8),
2. *as $\delta \to 0$, an approximate solution to the ℓ_0 problem*

$$\arg\min_x \sum_{l=1}^{L} 1_{|\psi_l^T x|>0} \;\; s.t. \;\; \|y - \Phi x\|_2 \leq \varepsilon, \tag{50}$$

3. *for $\varepsilon = 0$, MM applied to Bayesian MAP estimation under a noiseless likelihood and the hierarchical prior*

$$p(x|\lambda) = \prod_{l=1}^{L} \frac{\lambda_l}{2} \exp\left(-\lambda_l |\psi_l^T x|\right) \tag{51}$$

$$\lambda \sim i.i.d.\ \Gamma(0, \delta^{-1}), \tag{52}$$

where $z_l = \psi_l^T x$ is Laplacian given λ_l, and λ_l is Gamma distributed with scale parameter δ^{-1} and shape parameter zero, which becomes Jeffrey's non-informative hyperprior $p(\lambda_l) \propto 1_{\lambda_l > 0}/\lambda_l$ when $\delta = 0$.

4. *for $\varepsilon = 0$, variational EM under a noiseless likelihood and the prior*

$$p(x;\lambda) \propto \prod_{l=1}^{L} \frac{\lambda_l}{2} \exp\left(-\lambda_l(|\psi_l^T x| + \delta)\right), \tag{53}$$

which, when $\delta = 0$, is independent Laplacian on $z = \Psi x \in \mathbb{R}^L$ under the positive deterministic scale parameters in λ.

While Part 1 and Part 2 of Corollary 1 were established for the synthesis-CS formulation of IRW-L1 in [13], we believe that Part 3 and Part 4 are novel interpretations of IRW-L1.

3 The Co-IRW-L1 Algorithm

We now propose the Co-IRW-L1-δ algorithm, which is summarized in Algorithm 3. Co-IRW-L1-δ can be thought of as a hybrid of the Co-L1 and IRW-L1 approaches from Algorithms 1 and 2, respectively. Like with Co-L1, the Co-IRW-L1-δ algorithm uses sub-dictionary dependent weights λ_d that are updated at each iteration t using a sparsity metric on $\Psi_d x^{(t)}$. But, like with IRW-L1, the Co-IRW-L1-δ algorithm also uses diagonal weight matrices $W_d^{(t)}$ that are updated at each iteration. As with both Co-L1 and IRW-L1, the computational burden of Co-IRW-L1-δ is dominated by the constrained ℓ_1 minimization problem in line 4 of Algorithm 3, which is readily solved by existing techniques like Douglas-Rachford splitting.

Algorithm 3 The Real-Valued Co-IRW-L1-δ Algorithm

1: input: $\{\Psi_d\}_{d=1}^D, \Phi, y, \varepsilon \geq 0, \delta_d > 0\ \forall d, \rho \geq 0,$
2: initialization: $\lambda_d^{(1)} = 1, W_d^{(1)} = I,\ \forall d \in [D]$
3: for $t = 1, 2, 3, \ldots$
4: $\quad x^{(t)} \leftarrow \arg\min_x \sum_{d=1}^{D} \lambda_d^{(t)} \| W_d^{(t)} \Psi_d x \|_1$ s.t. $\|y - \Phi x\|_2 \leq \varepsilon$
5: $\quad \lambda_d^{(t+1)} \leftarrow \left[\frac{1}{L_d} \sum_{l=1}^{L_d} \log\left(1 + \rho + \frac{|\psi_{d,l}^T x^{(t)}|}{\delta_d}\right) \right]^{-1} + 1,\ \forall d \in [D]$
6: $\quad W_d^{(t+1)} \leftarrow \text{diag}\left\{ \frac{1}{\delta_d(1+\rho) + |\psi_{d,1}^T x^{(t)}|}, \cdots, \frac{1}{\delta_d(1+\rho) + |\psi_{d,L_d}^T x^{(t)}|} \right\},\ \forall d \in [D]$
7: end
8: output: $x^{(t)}$

The Co-IRW-L1-δ algorithm can be interpreted in various ways, as we detail below. For clarity, we first consider fixed regularization parameters δ, and later, in Section 3.6, we describe how they can be adapted at each iteration, leading to the Co-IRW-L1 algorithm. Also, to simplify the development, we first consider the case where x is real-valued and later, in Section 3.7, discuss the complex-valued case.

Theorem 2 (Co-IRW-L1-δ). *The real-valued Co-IRW-L1-δ algorithm in Algorithm 3 has the following interpretations:*

1. *MM applied to (2) under the log-sum-log penalty*

$$R_{\text{lsl}}(x;\delta,\rho) \triangleq \sum_{d=1}^{D}\sum_{l=1}^{L_d} \log\left[\left(\delta_d(1+\rho)+|\psi_{d,l}^T x|\right)\sum_{i=1}^{L_d}\log\left(1+\rho+\frac{|\psi_{d,i}^T x|}{\delta_d}\right)\right], \tag{54}$$

2. *as $\rho \to 0$ and $\delta_d \to 0\ \forall d$, an approximate solution to the $\ell_0 + \ell_{0,0}$ problem*

$$\arg\min_x \|\Psi x\|_0 + \sum_{d=1}^{D} L_d\, 1_{\|\Psi_d x\|_0>0} \ \ s.t.\ \ \|y-\Phi x\|_2 \le \varepsilon, \tag{55}$$

3. *for $\varepsilon = 0$, MM applied to Bayesian MAP estimation under a noiseless likelihood and the hierarchical prior*

$$p(x|\lambda;\delta) = \prod_{d=1}^{D}\prod_{l=1}^{L_d}\frac{\lambda_d}{2\delta_d}\left(1+\rho+\frac{|\psi_{d,l}^T x|}{\delta_d}\right)^{-(\lambda_d+1)} \tag{56}$$

$$p(\lambda) = \prod_{d=1}^{D} p(\lambda_d),\ \ p(\lambda_d) \propto \begin{cases} \frac{1}{\lambda_d} & \lambda_d > 0 \\ 0 & else \end{cases}, \tag{57}$$

   where, when $\rho = 0$, the variables $z_d = \Psi_d x \in \mathbb{R}^{L_d}$ are i.i.d. generalized-Pareto [15] given λ_d, and $p(\lambda_d)$ is Jeffrey's non-informative hyperprior [7, 24, 37] for the random shape parameter λ_d.

4. *for $\varepsilon = 0$, variational EM under a noiseless likelihood and the prior*

$$p(x;\lambda,\delta) = \prod_{d=1}^{D}\prod_{l=1}^{L_d}\frac{\lambda_d-1}{2\delta_d}\left(1+\rho+\frac{|\psi_{d,l}^T x|}{\delta_d}\right)^{-\lambda_d}, \tag{58}$$

   where, when $\rho = 0$, the variables $z_d = \Psi_d x \in \mathbb{R}^{L_d}$ are i.i.d. generalized Pareto with deterministic shape parameter $\lambda_d > 1$ and scale parameter $\delta_d > 0$.

Proof. See Sections 3.1 to 3.5 below.

As with Co-L1, the MM interpretation implies convergence (in the sense of an asymptotic stationary point condition) when $\rho > 0$, as detailed in Section 3.2.

3.1 Log-Sum-Log MM Interpretation of Co-IRW-L1-δ

Consider the optimization problem

$$\arg\min_x R_{\mathsf{lsl}}(x;\delta,\rho) \;\text{ s.t. }\; \|y-\Phi x\|_2 \le \varepsilon, \tag{59}$$

with R_{lsl} defined in (54). We attack this optimization problem using the MM approach detailed in Section 2.1. The difference is that now the function g is defined as

$$g(v) = \sum_{d=1}^{D}\sum_{k\in K_d}\log\left[\left(\delta_d(1+\rho)+v_k\right)\sum_{i\in K_d}\log\left(1+\rho+\frac{v_i}{\delta_d}\right)\right] \tag{60}$$

$$= \sum_{d=1}^{D}\left[L_d\log\sum_{i\in K_d}\log\left(1+\rho+\frac{v_i}{\delta_d}\right)+\sum_{k\in K_d}\log\left(\delta_d(1+\rho)+v_k\right)\right], \tag{61}$$

which has a gradient of

$$[\nabla g(v^{(t)})]_k = \left(\frac{L_{d(k)}}{\sum\limits_{i\in K_{d(k)}}\log\left(1+\rho+\frac{v_i^{(t)}}{\delta_{d(k)}}\right)}+1\right)\frac{1}{\delta_{d(k)}(1+\rho)+v_k^{(t)}} \tag{62}$$

when $d(k)\neq 0$ and otherwise $[\nabla g(v^{(t)})]_k = 0$. Thus, recalling (18), MM prescribes

$$v^{(t+1)} = \arg\min_{v\in C}\sum_{d=1}^{D}\sum_{k\in K_d}\left(\frac{L_d}{\sum\limits_{i\in K_d}\log\left(1+\rho+\frac{v_i^{(t)}}{\delta_d}\right)}+1\right)\left(\frac{v_k}{\delta_d(1+\rho)+v_k^{(t)}}\right) \tag{63}$$

or equivalently

$$x^{(t+1)} = \arg\min_x\sum_{d=1}^{D}\sum_{l=1}^{L_d}\lambda_d^{(t+1)}\left(\frac{|\psi_{d,l}^T x|}{\delta_d(1+\rho)+|\psi_{d,l}^T x^{(t)}|}\right) \;\text{ s.t. }\; \|y-\Phi x\|_2\le\varepsilon \tag{64}$$

for

$$\lambda_d^{(t+1)} = \left[\frac{1}{L_d}\sum_{l=1}^{L_d}\log\left(1+\rho+\frac{|\psi_{d,l}^T x^{(t)}|}{\delta_d}\right)\right]^{-1}+1, \tag{65}$$

which coincides with Algorithm 3. This establishes Part 1 of Theorem 2.

3.2 Convergence of Co-IRW-L1-δ

The convergence of Co-IRW-L1-δ (in the sense of an asymptotic stationary point condition) for $\rho > 0$ can be shown using the same procedure as in Section 2.2. To do this, we only need to verify that the gradient ∇g in (62) is Lipschitz continuous when $\rho > 0$, which was done in [2, App. C].

3.3 Approximate $\ell_0 + \ell_{0,0}$ Interpretation of Co-IRW-L1-δ

Recalling the discussion in Section 2.3, we now consider the behavior of the $R_{\mathsf{lsl}}(x;\delta,\rho)$ regularizer in (54) as $\rho \to 0$ and $\delta_d \to 0\ \forall d$. For this, it helps to decouple (54) into two terms:

$$R_{\mathsf{lsl}}(x;\delta,\rho) \tag{66}$$
$$= \sum_{d=1}^{D}\sum_{l=1}^{L_d} \log\left(\delta_d(1+\rho) + |\psi_{d,l}^T x|\right) + \sum_{d=1}^{D}\sum_{l=1}^{L_d} \log\left[\sum_{i=1}^{L_d} \log\left(1+\rho+\frac{|\psi_{d,i}^T x|}{\delta_d}\right)\right].$$

As $\delta_d \to 0\ \forall d$, the first term in (66) contributes an infinite valued "reward" for each pair (d,l) such that $|\psi_{d,l}^T x| = 0$, or a finite valued cost otherwise. As for the second term, we see that $\lim_{\rho\to 0,\delta_d\to 0} \sum_{i=1}^{L_d} \log\left(1+|\psi_{d,i}^T x|/\delta_d+\rho\right) = 0$ if and only if $|\psi_{d,i}^T x| = 0\ \forall i \in [L_d]$, i.e., if and only if $\|\Psi_d x\|_0 = 0$. And when $\|\Psi_d x\|_0 = 0$, the second term in (66) contributes L_d infinite valued rewards. In summary, as $\rho \to 0$ and $\delta_d \to 0\ \forall d$, the first term in (66) behaves like $\|\Psi x\|_0$ and the second term like the weighted $\ell_{0,0}$ quasi-norm $\sum_{d=1}^{D} L_d 1_{\|\Psi_d x\|_0 > 0}$, as stated in (55). This establishes Part 2 of Theorem 2.

3.4 Bayesian MAP Interpretation of Co-IRW-L1-δ

To show that Co-IRW-L1-δ can be interpreted as Bayesian MAP estimation under the hierarchical prior (56)–(57), we first compute the prior $p(x)$. To start,

$$p(x) = \int_{\mathbb{R}^D} p(\lambda)p(x|\lambda)\mathrm{d}\lambda \tag{67}$$
$$\propto \prod_{d=1}^{D}\int_0^\infty \frac{1}{\lambda_d}\prod_{l=1}^{L_d}\frac{\lambda_d}{2\delta_d}\left(1+\rho+\frac{|\psi_{d,l}^T x|}{\delta_d}\right)^{-(\lambda_d+1)}\mathrm{d}\lambda_d. \tag{68}$$

Writing $(1+\rho+|\psi_{d,l}^T x|/\delta_d)^{-(\lambda_d+1)} = \exp(-(\lambda_d+1)Q_{d,l})$ for $Q_{d,l} \triangleq \log(1+\rho+|\psi_{d,l}^T x|/\delta_d)$, we get

$$p(x) \propto \prod_{d=1}^{D} \frac{1}{(2\delta_d)^{L_d}} \int_0^\infty \lambda_d^{L_d-1} e^{-(\lambda_d+1)\sum_{l=1}^{L_d} Q_{d,l}} \mathrm{d}\lambda_d. \tag{69}$$

Defining $Q_d \triangleq \sum_{l=1}^{L_d} Q_{d,l}$ and changing the variable of integration to $\tau_d \triangleq \lambda_d Q_d$, we find

$$p(x) \propto \prod_{d=1}^{D} \frac{e^{-Q_d}}{(2\delta_d Q_d)^{L_d}} \underbrace{\int_0^\infty \tau_d^{L_d-1} e^{-\tau_d} \mathrm{d}\tau_d}_{(L_d-1)!} \tag{70}$$

$$\propto \prod_{d=1}^{D} \left[\frac{1}{\delta_d \sum_{i=1}^{L_d} \log(1+\rho+\frac{|\psi_{d,i}^T x|}{\delta_d})} \right]^{L_d} \prod_{l=1}^{L_d} \frac{1}{1+\rho+\frac{|\psi_{d,l}^T x|}{\delta_d}} \tag{71}$$

$$= \prod_{d=1}^{D} \prod_{l=1}^{L_d} \left[\left(\delta_d(1+\rho)+|\psi_{d,l}^T x|\right) \sum_{i=1}^{L_d} \log\left(1+\rho+\frac{|\psi_{d,i}^T x|}{\delta_d}\right) \right]^{-1}, \tag{72}$$

which implies that

$$-\log p(x) = \text{const} + R_{\mathsf{lsl}}(x;\delta,\rho) \tag{73}$$

for $R_{\mathsf{lsl}}(x;\delta,\rho)$ defined in (54).

Plugging (73) into noiseless MAP expression (29), we have

$$x_{\mathsf{MAP}} = \arg\min_x R_{\mathsf{lsl}}(x;\delta,\rho) \ \text{ s.t. } \ y = \Phi x, \tag{74}$$

which is equivalent to the optimization problem in (59) when $\varepsilon = 0$. We showed in Section 3.1 that, by applying the MM algorithm to (59), we arrive at Algorithm 3. This establishes Part 3 of Theorem 2.

3.5 *Variational EM Interpretation of Co-IRW-L1-δ*

To justify the variational EM (VEM) interpretation of Co-IRW-L1-δ, we closely follow the approach used for Co-L1 in Section 2.5. The main difference is that now the prior takes the form of $p(x;\lambda,\delta)$ from (58). Thus, (43) becomes

$$\log p(x;\lambda,\delta) = \sum_{d=1}^{D}\sum_{l=1}^{L_d} \left[\log\left(\frac{\lambda_d-1}{\delta_d}\right) - \lambda_d \log\left(1+\rho+\frac{|\psi_{d,l}^T x|}{\delta_d}\right) \right] + \text{const} \tag{75}$$

and by zeroing the gradient w.r.t. λ, we see that the M step (44) becomes

$$\frac{1}{\lambda_d^{(t+1)} - 1} = \frac{1}{L_d} \log\left(1 + \rho + \frac{|\psi_{d,l}^T x_{\text{MAP}}^{(t)}|}{\delta_d}\right), \quad d \in [D], \tag{76}$$

where again $x_{\text{MAP}}^{(t)}$ denotes the MAP estimate of x under $\lambda = \lambda^{(t)}$. From (29) and (58), we see that

$$x_{\text{MAP}}^{(t)} = \arg\min_x \sum_{d=1}^{D} \lambda_d^{(t)} \sum_{l=1}^{L_d} \log\left(|\psi_{d,l}^T x| + \delta_d(1+\rho)\right) \text{ s.t. } y = \Phi x, \tag{77}$$

which (for $\rho = 0$) is a $\lambda^{(t)}$-weighted version of the IRW-L1 log-sum optimization problem (recall Part 1 of Corollary 1). To solve (77), we apply MM with inner iteration i. With a small modification of the MM derivation from Section 2.1, we obtain the two-step iteration

$$x_{\text{MAP}}^{(i)} = \arg\min_x \sum_{d=1}^{D} \lambda_d^{(t)} \|W_d^{(i)} \Psi_d x\|_1 \text{ s.t. } y = \Phi x \tag{78}$$

$$W_d^{(i+1)} = \text{diag}\left\{\frac{1}{\delta_d(1+\rho) + |\psi_{d,1}^T x^{(i)}|}, \cdots, \frac{1}{\delta_d(1+\rho) + |\psi_{d,L_d}^T x^{(i)}|}\right\}, \tag{79}$$

with $\lambda_d^{(t)}$ fixed at the value appearing in (77). Next, by using only a single MM iteration per VEM iteration, the MM index "i" can be equated with the VEM index "t," in which case the VEM algorithm becomes

$$x^{(t)} = \arg\min_x \sum_{d=1}^{D} \lambda_d^{(t)} \|W_d^{(t)} \Psi_d x\|_1 \text{ s.t. } y = \Phi x \tag{80}$$

$$W_d^{(t+1)} = \text{diag}\left\{\frac{1}{\delta_d(1+\rho) + |\psi_{d,1}^T x^{(t)}|}, \cdots, \frac{1}{\delta_d(1+\rho) + |\psi_{d,L_d}^T x^{(t)}|}\right\}, \forall d \tag{81}$$

$$\lambda_d^{(t+1)} = \left[\frac{1}{L_d} \log\left(1 + \rho + \frac{|\psi_{d,l}^T x^{(t)}|}{\delta_d}\right)\right]^{-1} + 1, \quad \forall d, \tag{82}$$

which matches the steps in Algorithm 3 under $\varepsilon = 0$. This establishes Part 4 of Theorem 2.

3.6 Co-IRW-L1

Until now, we have considered the Co-IRW-L1-δ parameters δ to be fixed and known. But it is not clear how to set these parameters in practice. Thus, in this section, we describe an extension of Co-IRW-L1-δ that adapts the δ vector at every iteration. The resulting procedure, which we will refer to as Co-IRW-L1, is summarized in Algorithm 4.

In the case of real-valued x, the expression for $\log p(x; \lambda, \delta)$ in line 6 of Algorithm 4 is given in (75) for $\lambda_d > 1$ and $\delta_d > 0$. Although there does not appear to be a closed-form solution to the joint maximization problem in line 6, it is over two real parameters and thus can be solved numerically without a significant computational burden.

Algorithm 4 The Co-IRW-L1 Algorithm

1: input: $\{\Psi_d\}_{d=1}^D, \Phi, y, \varepsilon \geq 0, \rho \geq 0$
2: if $x \in \mathbb{R}^n$, use $\Lambda = (1, \infty)$ and $\log p(x; \lambda, \delta)$ from (75);
if $x \in \mathbb{C}^n$, use $\Lambda = (2, \infty)$ and $\log p(x; \lambda, \delta)$ from (84).
3: initialization: $\lambda_d^{(1)} = 1, W_d^{(1)} = I, \ \forall d \in [D]$
4: for $t = 1, 2, 3, \ldots$
5: $x^{(t)} \leftarrow \arg\min_x \sum_{d=1}^D \lambda_d^{(t)} \|W_d^{(t)} \Psi_d x\|_1 \ \text{s.t.} \ \|y - \Phi x\|_2 \leq \varepsilon$
6: $(\lambda_d^{(t+1)}, \delta_d^{(t+1)}) \leftarrow \arg\max_{\lambda_d \in \Lambda, \delta_d > 0} \log p(x^{(t)}; \lambda, \delta), \ d \in [D]$
7: $W_d^{(t+1)} \leftarrow \text{diag}\left\{ \frac{1}{\delta_d^{(t+1)}(1+\rho) + |\psi_{d,1}^T x^{(t)}|}, \cdots, \frac{1}{\delta_d^{(t+1)}(1+\rho) + |\psi_{d,L_d}^T x^{(t)}|} \right\}, \ d \in [D]$
8: end
9: output: $x^{(t)}$

Algorithm 4 can be interpreted as a generalization of the VEM approach to Co-IRW-L1-δ that is summarized in Part 4 of Theorem 2 and detailed in Section 3.5. Whereas Co-IRW-L1-δ used VEM to estimate the λ parameters in the prior (58) for a fixed value of δ, Co-IRW-L1 uses VEM to *jointly* estimate (λ, δ) in (58). Thus, Co-IRW-L1 can be derived by repeating the steps in Section 3.5, except that now the maximization of $\log p(x; \lambda, \delta)$ in (75) is performed jointly over (λ, δ), as reflected by line 6 of Algorithm 4.

3.7 Co-IRW-L1 for Complex-Valued x

In Sections 3.1–3.6, the signal x was assumed to be real-valued. We now extend the previous results to the case of complex-valued x. For this, we focus on the Co-IRW-L1 algorithm, since Co-IRW-L1-δ follows as the special case where δ is fixed at a user-supplied value.

Recalling that Co-IRW-L1 was constructed by generalizing the VEM interpretation of Co-IRW-L1-δ, we reconsider this VEM interpretation for the case of complex-valued x. In particular, we assume an AWGN likelihood and the following complex-valued extension of the prior (58):

$$p(x;\lambda,\delta) \propto \prod_{d=1}^{D}\prod_{l=1}^{L_d} \frac{(\lambda_d-1)(\lambda_d-2)}{2\pi\delta_d^2}\left(1+\rho+\frac{|\psi_{d,l}^T x|}{\delta_d}\right)^{-\lambda_d}, \tag{83}$$

which is now i.i.d. generalized Pareto on $z_d = \Psi_d x \in \mathbb{C}^{L_d}$ with deterministic shape parameter $\lambda_d > 2$ and deterministic scale parameter $\delta_d > 0$. In this case, the log-prior (75) changes to

$$\log p(x;\lambda,\delta)$$
$$= \text{const} + \sum_{d=1}^{D}\sum_{l=1}^{L_d}\left[\log\left(\frac{(\lambda_d-1)(\lambda_d-2)}{\delta_d^2}\right) - \lambda_d \log\left(1+\rho+\frac{|\psi_{d,l}^T x|}{\delta_d}\right)\right], \tag{84}$$

which is then maximized over (λ,δ) in line 6 of Algorithm 4.

4 Numerical Results

We now present results from a numerical study into the performance of the proposed Co-L1 and Co-IRW-L1 methods, given as Algorithm 1 and Algorithm 4, respectively. Three experiments are discussed below, all of which focus on the problem of recovering an n-pixel image (or image sequence) x from m-sample noisy compressed measurements $y = \Phi x + e$, with $m \ll n$. In the first experiment, we recover synthetic 2D finite-difference signals; in the second experiment, we recover the Shepp-Logan phantom and the Cameraman image; and in the third experiment, we recover dynamic MRI sequences, also known as "cines."

As discussed in Section 1.4, Co-L1 can be considered as the composite extension of the standard L1-regularized L2-constrained approach to analysis CS, i.e., (2) under the non-composite L1 regularizer $R(x) = \|\Psi x\|_1$. Similarly, Co-IRW-L1 can be considered as the composite extension of the standard IRW approach to the same L1 problem. Thus, we compare our proposed composite methods against these two non-composite methods, referring to them simply as "L1" and "IRW-L1" in the sequel.

4.1 Experimental Setup

For the dynamic MRI experiment, we constructed Φ using randomly subsampled Fourier measurements at each time instant with a varying sampling pattern across time. More details are given in Section 4.4. For the other experiments, we used a "spread spectrum" operator [39] of the form $\Phi = DFC$, where $C \in \mathbb{R}^{n\times n}$ is diagonal matrix with i.i.d equiprobable ± 1 entries, $F \in \mathbb{C}^{n\times n}$ is the discrete Fourier transform (DFT), and $D \in \mathbb{R}^{m\times n}$ is a row-selection operator that selects m rows of $FC \in \mathbb{C}^{n\times n}$ uniformly at random.

In all cases, the noise e was zero-mean, white, and circular Gaussian (i.e., independent real and imaginary components of equal variance). Denoting the noise variance by σ^2, we define the measurement signal-to-noise ratio (SNR) as $\|y\|_2^2/(m\sigma^2)$ and the recovery SNR of signal estimate $\hat{x}$ as $\|x\|_2^2/\|x-\hat{x}\|_2^2$.

Note that, when x is real-valued, the measurements y will be complex-valued due to the construction of Φ. Thus, to allow the use of real-valued L1 solvers, we split each complex-valued element of y (and the corresponding rows of Φ and e) into real and imaginary components, resulting in a real-only model. However, to avoid possible redundancy issues caused by the conjugate symmetry of the noiseless Fourier measurements FCx, we ensured that D selected at most one sample from each complex-conjugate pair.

To implement the existing non-composite L1 and IRW-L1 methods, we used the Matlab codes linked[2] to the paper [14], which are based on Douglas-Rachford splitting [18]. All default settings were retained except that the maximum number of reweighting iterations was increased from 10 to 25, which resulted in improved recovery SNR. Then, to implement the weighted-ℓ_1 minimization step in Co-L1 and Co-IRW-L1, we used a similar Douglas-Rachford splitting technique. The maximum number of reweighting iterations for Co-L1 and Co-IRW-L1 was set at 25. For Co-L1, IRW-L1, and Co-IRW-L1, the t-indexed iterations in Algorithm 1, Algorithm 2, and Algorithm 4, respectively, were terminated when $\|x^{(t)} - x^{(t-1)}\|_2/\|x^{(t)}\|_2 < 1 \times 10^{-8}$. In all experiments we used $\varepsilon = 0.8\sqrt{\sigma^2 m}$ and $\delta = 0 = \rho$.

4.2 Synthetic 2D Finite-Difference Signals

Our first experiment aims to answer the following question. If we know that the sparsity of $\Psi_1 x$ differs from the sparsity of $\Psi_2 x$, then can we exploit this knowledge for signal recovery, even if we don't know *how* the sparsities are different? This is precisely the goal of composite regularizations like (4).

[2]Matlab codes for [14] are provided at `https://github.com/basp-group/sopt`.

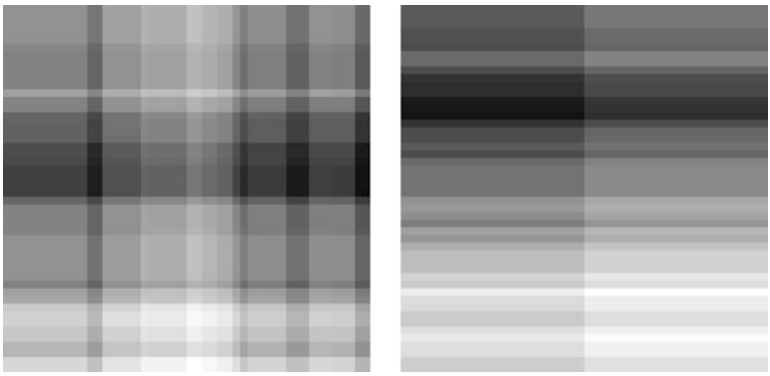


Fig. 1 Examples of the 2D finite-difference signal X used in the first experiment. On the left is a realization generated under a transition ratio of $\alpha = 14/14 = 1$, and on the right is a realization generated under $\alpha = 27/1 = 27$.

To investigate this question, we constructed 2D signals with finite-difference structure in both the vertical and horizontal domains. In particular, we constructed $X = x_1 1^T + 1 x_2^T$, where both $x_1 \in \mathbb{R}^{48}$ and $x_2 \in \mathbb{R}^{48}$ are finite-difference signals and $1 \in \mathbb{R}^{48}$ contains only ones. The locations of the transitions in x_1 and x_2 were selected uniformly at random and the amplitudes of the transitions were drawn i.i.d. zero-mean Gaussian. The total number of transitions in x_1 and x_2 was fixed at 28, but the ratio of the number of transitions in x_1 to the number in x_2, denoted by α, was varied from 1 to 27. The case $\alpha = 1$ corresponds to X having 14 vertical transitions and 14 horizontal transitions, while the case $\alpha = 27$ corresponds to X having 27 vertical transitions and a single horizontal transition. (See Figure 1 for examples.) Finally, the signal $x \in \mathbb{R}^n$ appearing in our model (1) was created by vectorizing X, yielding a total of $n = 48^2 = 2304$ pixels.

Given x, noisy observations $y = \Phi x + e$ were generated using the random "spread spectrum" measurement operator Φ described earlier at a sampling ratio of $m/n = 0.3$, with additive white Gaussian noise (AWGN) e scaled to achieve a measurement SNR of 40 dB. All recovery algorithms used vertical and horizontal finite-difference operators Ψ_1 and Ψ_2, respectively, with $\Psi = [\Psi_1^T, \Psi_2^T]^T$ in the non-composite case.

Figure 2 shows recovery SNR versus α for the non-composite L1 and IRW-L1 techniques and our proposed Co-L1 and Co-IRW-L1 techniques. Each SNR in the figure represents the median value from 45 trials, each using an independent realization of the triple (Φ, x, e). The figure shows that the recovery SNR of both L1 and IRW-L1 is roughly invariant to the transition ratio α, which makes sense because the overall sparsity of Ψx is fixed at 28 transitions by construction. In contrast, the recovery SNRs of Co-L1 and Co-IRW-L1 vary with α, with higher values of α yielding a more structured signal and thus higher recovery SNR when this structure is properly exploited.

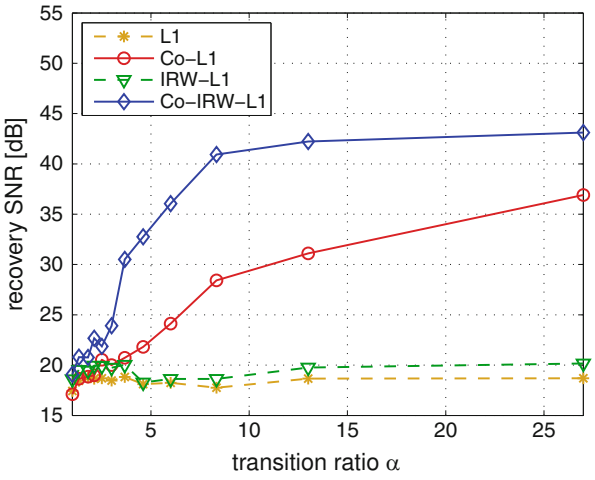


Fig. 2 Recovery SNR versus transition ratio α for the first experiment, which used 2D finite-difference signals, spread-spectrum measurements at $m/n = 0.3$, AWGN at 40 dB, and finite-difference operators for Ψ_d. Each recovery SNR represents the median value from 45 independent trials.

Fig. 3 Left: the Shepp-Logan phantom of size $n = 96 \times 96$. Right: the cropped Cameraman image of size $n = 96 \times 104$.

4.3 Shepp-Logan and Cameraman Recovery

For our second experiment, we investigate algorithm performance versus sampling ratio m/n when recovering the well-known Shepp-Logan phantom and Cameraman images. In particular, we used the $n = 96 \times 96$ Shepp-Logan phantom and the $n = 96 \times 104$ cropped Cameraman image shown in Figure 3, and we constructed compressed noisy measurements y using spread-spectrum Φ and AWGN e at a measurement SNR of 30 dB in the Shepp-Logan case and 40 dB in the Cameraman case.

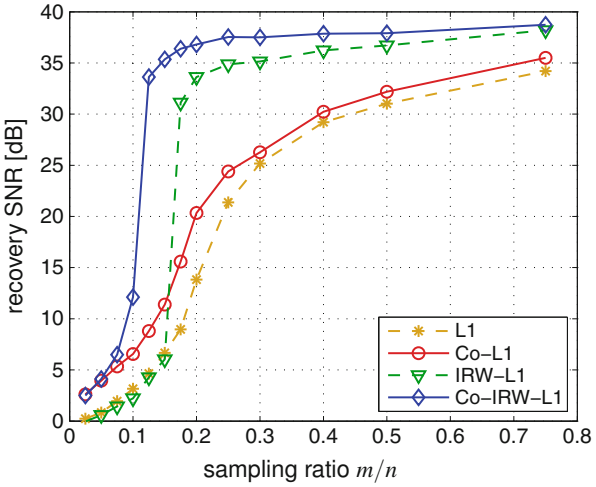


Fig. 4 Recovery SNR versus sampling ratio m/n for the Shepp-Logan phantom. Measurements were constructed using a spread-spectrum operator and AWGN at 30 dB SNR, and recovery used the UWT-db1 2D wavelet transform at two levels of decomposition. Each recovery SNR represents the median value from *seven* independent trials.

All algorithms used analysis operator $\Psi \in \mathbb{R}^{7n\times n}$ constructed from the undecimated Daubechies-1 2D wavelet transform (UWT-db1) with two levels of decomposition. However, the Co-L1 and Co-IRW-L1 algorithms treated each of the seven subbands of UWT-db1 as a separate sub-dictionary $\Psi_d \in \mathbb{R}^{n\times n}$ in their composite regularizers.

Figure 4 shows recovery SNR versus sampling ratio m/n for the Shepp-Logan phantom, while Figure 5 shows the same for the Cameraman image. Each recovery SNR represents the median value from seven independent realizations of (Φ, e). Both figures show that Co-L1 and Co-IRW-L1 outperform their non-composite counterparts, especially at low sampling ratios; the gap between Co-IRW-L1 and IRW-L1 closes at $m/n \geq 0.4$. Although not shown, similar results were observed with a level three decomposition of UWT-db1 and at higher (50 dB) and lower (25 dB) measurement SNRs.

4.4 Dynamic MRI

For our third experiment, we investigate a simplified version of the "dynamic MRI" (dMRI) problem. In dMRI, one attempts to recover a sequence of MRI images, known as an MRI cine, from highly under-sampled "k-t-domain" measurements $\{y_t\}_{t=1}^T$ constructed as

$$y_t = \Phi_t x_t + e_t, \tag{85}$$

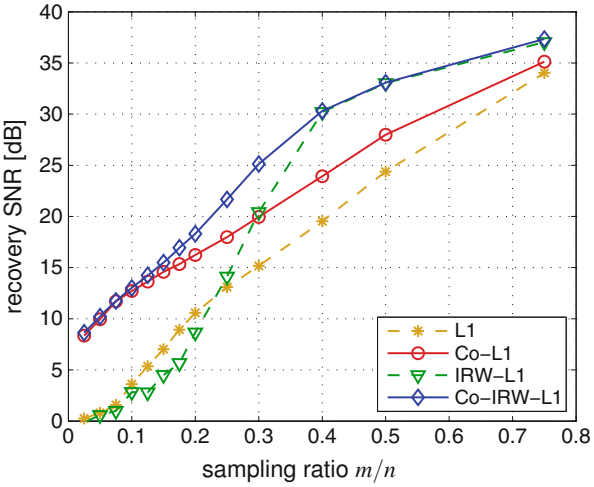


Fig. 5 Recovery SNR versus sampling ratio m/n for the cropped Cameraman image. Measurements were constructed using a spread-spectrum operator and AWGN at 40 dB SNR, and recovery used the UWT-db1 2D wavelet transform at two levels of decomposition. Each SNR value represents the median value from 7 independent trials.

where $x_t \in \mathbb{R}^{n_1 n_2}$ is a vectorized $(n_1 \times n_2)$-pixel image at time t, $\Phi_t \in \mathbb{R}^{m_1 \times n_1 n_2}$ is a subsampled Fourier operator at time t, and $e_t \in \mathbb{R}^m_1$ is AWGN. This real-valued Φ_t is constructed from the complex-valued $n_1 n_2 \times n_1 n_2$ 2D DFT matrix by randomly selecting $0.5 m_1$ rows and then splitting each of those rows into its real and imaginary components. Here, it is usually advantageous to vary the sampling pattern with time and to sample more densely at low frequencies, where most of the signal energy lies (e.g., [3]). Putting (85) into the form of our measurement model (1), we get

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} \Phi_1 & & \\ & \ddots & \\ & & \Phi_T \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_T \end{bmatrix}}_{x} + \underbrace{\begin{bmatrix} e_1 \\ \vdots \\ e_T \end{bmatrix}}_{e}, \tag{86}$$

with total measurement dimension $m = m_1 T$ and total signal dimension $n = n_1 n_2 T$.

As ground truth, we used a high-quality dMRI cardiac cine x of dimensions $n_1 = 144$, $n_2 = 85$, and $T = 48$. The left pane in Figure 6 shows a 144×85 image from this cine extracted at a single time t, while the middle pane shows a 144×48 spatio-temporal profile from this cine extracted at a single horizontal location. This middle pane shows that the temporal dimension is much more structured than the spatial dimension, suggesting that there may be an advantage to weighting the spatial and temporal dimensions differently in a composite regularizer.

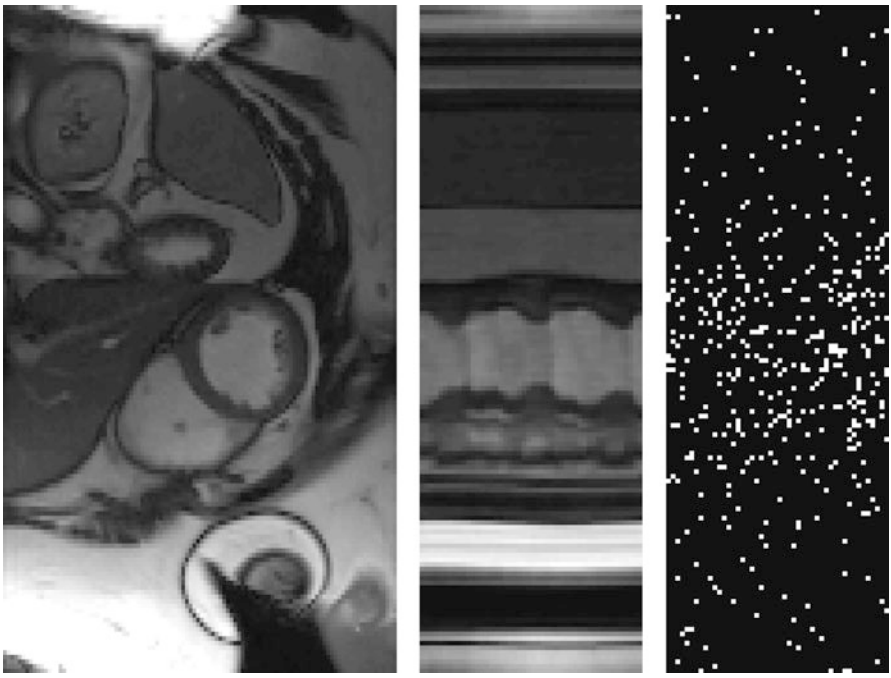


Fig. 6 Left: A 144×85 spatial slice from the $144 \times 85 \times 48$ dMRI dataset. Middle: The 144×48 spatio-temporal slice used for the dMRI experiment. Right: a realization of the variable-density k-space sampling pattern, versus time, at $m/n = 0.15$.

To test this hypothesis, we constructed an experiment where the goal was to recover the 144×48 spatio-temporal profile shown in the middle pane of Figure 6, as opposed to the full 3D cine, from subsampled k-t-domain measurements. For this purpose, we constructed measurements $\{y\}_{t=1}^{T}$ as described above, but with $n_2 = 1$ (and thus a 1D DFT), and used a variable density random sampling method. The right pane of Figure 6 shows a typical realization of the sampling pattern versus time. Finally, we selected the AWGN variance that yielded measurement SNR $=$ 30 dB.

For the non-composite L1 and IRW-L1 algorithms, we constructed the analysis operator $\Psi \in \mathbb{R}^{3n \times n}$ from a vertical concatenation of the db1-db3 Daubechies orthogonal 2D discrete wavelet bases, each with three levels of decomposition. For the Co-L1 and Co-IRW-L1 algorithms, we assigned each of the 30 sub-bands in Ψ to a separate sub-dictionary $\Psi_d \in \mathbb{R}^{L_d \times n}$. Note that the sub-dictionary size L_d decreases with the level in the decomposition. By weighting certain sub-dictionaries differently than others, the composite regularizers can exploit differences in spatial versus temporal structure.

Figure 7 shows recovery SNR versus sampling ratio m/n for the four algorithms under test. Each reported SNR represents the median SNR from seven independent

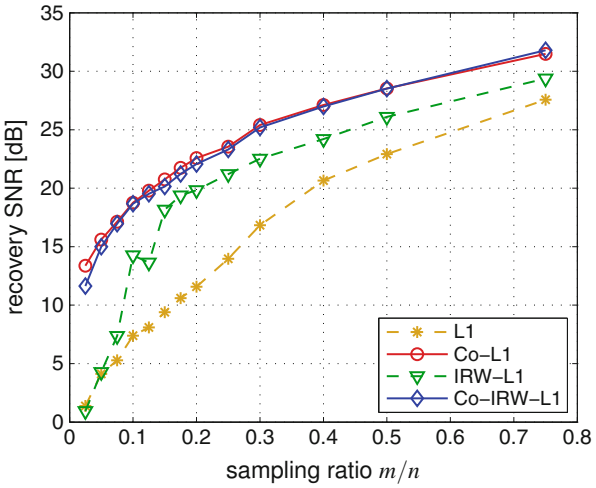


Fig. 7 Recovery SNR versus sampling ratio m/n for the dMRI experiment. Each SNR value represents the median value from 7 independent trials. Measurements were constructed using variable-density subsampled Fourier operator and AWGN at 30 dB measurement SNR, and recovery used a concatenation of db1–db3 orthogonal 2D wavelet bases at three levels of decomposition.

realizations of (Φ, e). The figure shows that Co-L1 and Co-IRW-L1 outperform their non-composite counterparts by $\geq$ 2 dB at all tested values of m/n, with larger gains at small m/n. Interestingly, Co-L1 and Co-IRW-L1 gave nearly identical recovery SNR in this experiment, which suggests that—for each d—the analysis coefficients *within* $\Psi_d x$ were of a similar magnitude. Although not shown here, we obtained similar results with other cine datasets and with an UWT-db1-based analysis operator.

For qualitative comparison, Figure 8 shows the spatio-temporal profile recovered by each of the four algorithms under test at $m/n = 0.15$ for a typical realization of (Φ, e). Compared to the ground-truth profile shown in the middle pane of Figure 6, the profiles recovered by L1 and IRW-L1 show visible artifacts that appear as vertical streaks. In contrast, the profiles recovered by Co-L1 and Co-IRW-L1 preserve most of the features present in the ground-truth profile.

4.5 Algorithm Runtime

Table 1 reports the average runtimes of the L1, Co-L1, IRW-L1, and Co-IRW-L1 algorithms for the experiments in Sections 4.3 and 4.4. There we see that the runtime of Co-L1 ranged between 1.5× to 3× that of L1, and the runtime of Co-IRW-L1 ranged between 1.5× to 3× the runtime of IRW-L1.

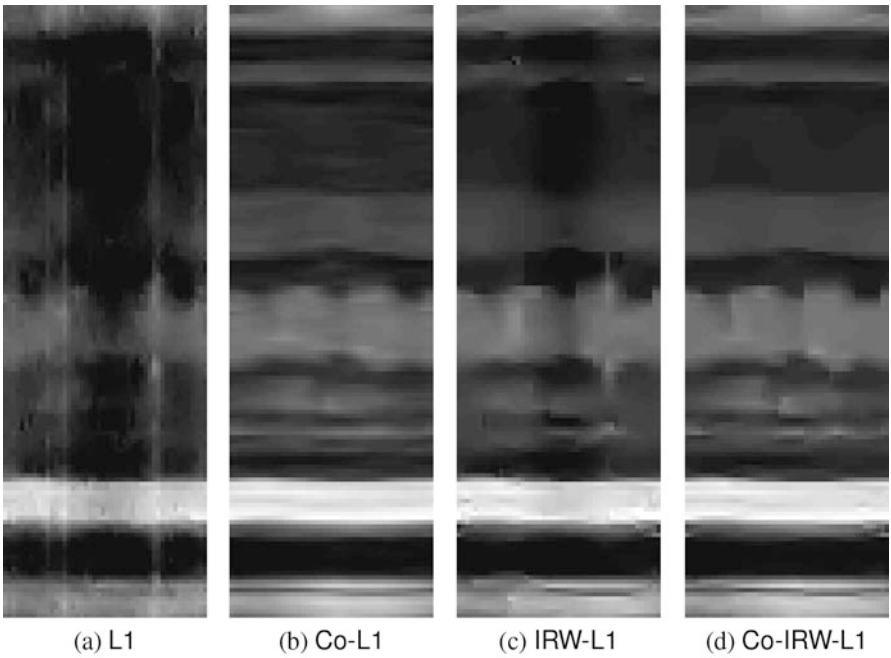


Fig. 8 Recovered dMRI spatio-temporal profiles at $m/n = 0.15$

Table 1 Computation times (in seconds) for the presented experimental studies. The times are averaged over trial runs and different sampling ratios.

	Shepp-Logan	Cameraman	MRI
L1	20.8	23.1	29.3
Co-L1	32.7	34.2	86.4
IRW-L1	45.9	48.4	54.1
Co-IRW-L1	72.1	96.4	131

5 Conclusions

Motivated by the observation that a given signal x admits sparse representations in multiple dictionaries Ψ_d but with varying levels of sparsity across dictionaries, we proposed two new algorithms for the reconstruction of (approximately) sparse signals from noisy linear measurements. Our first algorithm, Co-L1, extends the well-known lasso algorithm [17, 44, 45] from the L1 penalty $\|\Psi x\|_1$ to composite L1 penalties of the form (4) while self-adjusting the regularization weights λ_d. Our second algorithm, Co-IRW-L1, extends the well-known IRW-L1 algorithm [13, 14] to the same family of composite penalties while self-adjusting the regularization weights λ_d and the regularization parameters δ_d.

We provided several interpretations of both algorithms: i) majorization-minimization (MM) applied to a non-convex log-sum-type penalty, ii) MM applied to an approximate ℓ_0-type penalty, iii) MM applied to Bayesian MAP inference

under a particular hierarchical prior, and iv) variational expectation-maximization (VEM) under a particular prior with deterministic unknown parameters. Also, we leveraged the MM interpretation to establish convergence in the form of an asymptotic stationary point condition [34]. Furthermore, we noted that the Bayesian MAP and VEM viewpoints yield novel interpretations of the original IRW-L1 algorithm. Finally, we present a detailed numerical study that suggests that our proposed algorithms yield significantly improved recovery SNR when compared to their non-composite L1 and IRW-L1 counterparts with a modest (e.g., 1.5×-3×) increase in runtime.

Acknowledgements This work has been supported in part by NSF grants CCF-1218754 and CCF-1018368 and DARPA grant N66001-11-1-4090 and NIH grant R01HL135489. An early version of this work was presented at the 2015 ISMRM Annual Meeting and Exhibition.

References

1. M.V. Afonso, J.M. Bioucas-Dias, M.A.T. Figueiredo, Fast image recovery using variable splitting and constrained optimization. IEEE Trans. Image Process. **19**(9), 2345–2356 (2010)
2. R. Ahmad, P. Schniter, Iteratively reweighted ℓ_1 approaches to sparse composite regularization. IEEE Trans. Comput. Imaging **10**(2), 220–235 (2015)
3. R. Ahmad, H. Xue, S. Giri, Y. Ding, J. Craft, O.P. Simonetti, Variable density incoherent spatiotemporal acquisition (VISTA) for highly accelerated cardiac MRI. Magn. Reson. Med. 1266–1278 (2014). https://doi.org/10.1002/mrm.25507
4. S.D. Babacan, S. Nakajima, M.N. Do, Bayesian group-sparse modeling and variational inference. IEEE Trans. Signal Process. **62**(11), 2906–2921 (2014)
5. S. Becker, J. Bobin, E.J. Candès, NESTA: a fast and accurate first-order method for sparse recovery. SIAM J. Imaging Sci. **4**(1), 1–39 (2011)
6. M. Belge, M.E. Kilmer, E.L. Miller, Efficient determination of multiple regularization parameters in a generalized L-curve framework. Inverse Prob. **18**(4), 1161–1183 (2002)
7. J.O. Berger, *Statistical Decision Theory and Bayesian Analysis* (Springer, New York, 1985)
8. C.M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2007)
9. J.M. Borwein, A.S. Lewis, *Convex Analysis and Nonlinear Optimization* (Springer, New York, 2006)
10. S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2010)
11. C. Brezinski, M. Redivo-Zaglia, G. Rodriguez, S. Seatzu, Multi-parameter regularization techniques for ill-conditioned linear systems. Numer. Math. **94**(2), 203–228 (2003)
12. E.J. Candès, M.B. Wakin, An introduction to compressive sampling. IEEE Signal Process. Mag. **25**(2), 21–30 (2008)
13. E.J. Candès, M.B. Wakin, S. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization. J. Fourier Anal. Appl. **14**(5), 877–905 (2008)
14. R.E. Carrillo, J.D. McEwen, D. Van De Ville, J.P. Thiran, Y. Wiaux, Sparsity averaging for compressive imaging. IEEE Signal Process. Lett. **20**(6), 591–594 (2013)
15. V. Cevher, Learning with compressible priors, in *Proceedings of Neural Information Processing Systems Conference*, Vancouver, BC (2009), pp. 261–269
16. R. Chartrand, W. Yin, Iteratively reweighted algorithms for compressive sensing, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV (2008), pp. 3869–3872

17. S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**(1), 33–61 (1998)
18. P.L. Combettes, J.C. Pesquet, A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. IEEE J. Sel. Top. Sign. Proces. **1**(4), 6564–574 (2007)
19. I. Daubechies, R. DeVore, M. Fornasier, C.S. Güntürk, Iteratively reweighted least squares minimization for sparse recovery. Commun. Pure Appl. Math. **63**(1), 1–38 (2010)
20. 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)
21. M. Elad, P. Milanfar, R. Rubinstein, Analysis versus synthesis in signal priors. Inverse Prob. **23**, 947–968 (2007)
22. Y.C. Eldar, G. Kutyniok, *Compressed Sensing: Theory and Applications* (Cambridge University Press, New York, 2012)
23. M.A. Figueiredo, Adaptive sparseness for supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1150–1159 (2003)
24. M.A.T. Figueiredo, R.D. Nowak, Wavelet-based image estimation: an empirical Bayes approach using Jeffreys' noninformative prior. IEEE Trans. Image Process. **10**(9), 1322–1331 (2001)
25. M.A.T. Figueiredo, R.D. Nowak, Majorization-minimization algorithms for wavelet-based image restoration. IEEE Trans. Image Process. **16**(12), 2980–2991 (2007)
26. M. Fornasier, V. Naumova, S.V. Pereverzyev, Multi-parameter regularization techniques for ill-conditioned linear systems. SIAM J. Numer. Anal. **52**(4), 1770–1794 (2014)
27. S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Birkhäuser, New York, 2013)
28. S. Gazzola, P. Novati, Multi-parameter Arnoldi-Tikhonov methods. Electron. Trans. Numer. Anal. **40**, 452–475 (2013)
29. D.R. Hunter, K. Lange, A tutorial on MM algorithms. Am. Stat. **58**(1), 30–37 (2004)
30. M.A. Khajehnejad, M. Amin, W. Xu, A.S. Avestimehr, B. Hassibi, Improved sparse recovery thresholds with two-step reweighted ℓ_1 minimization, in *Proceedings of the IEEE International Symposium on Information Theory* (2010), pp. 1603–1607
31. M. Kowalski, Sparse regression using mixed norms. Appl. Comput. Harmon. Anal. **27**(2), 303–324 (2009)
32. K. Kunisch, T. Pock, A bilevel optimization approach for parameter learning in variational models. SIAM J. Imag. Sci. **6**(2), 938–983 (2013)
33. S. Lu, S.V. Pereverzev, *Regularization Theory for Ill-Posed Problems* (Walter de Gruyter, Berlin, 2013)
34. J. Mairal, Optimization with first-order surrogate functions, in *Proceeding International Conference on Machine Learning*, vol. 28 (2013), pp. 783–791
35. J. Mairal, F. Bach, J. Ponce, Sparse modeling for image and vision processing. Found. Trends Comput. Vis. **8**(2–3), 85–283 (2014)
36. R. Neal, G. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in *Learning in Graphical Models*, ed. by M.I. Jordan (MIT Press, Cambridge, MA, 1998), pp. 355–368
37. J.P. Oliveira, J.M. Bioucas-Dias, M.A.T. Figueiredo, Adaptive total variation image deblurring: a majorization-minimization approach. Signal Process. **89**(9), 1683–1693 (2009)
38. H.V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd edn. (Springer, New York, 1994)
39. G. Puy, P. Vandergheynst, R. Gribonval, Y. Wiaux, Universal and efficient compressed sensing by spread spectrum and application to realistic Fourier imaging techniques. EURASIP J. Appl. Signal Process. **2012**(6), 1–13 (2012)
40. A. Rakotomamonjy, Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms. Signal Process. **91**, 1505–1526 (2011)
41. B.D. Rao, K. Kreutz-Delgado, An affine scaling methodology for best basis selection. IEEE Trans. Signal Process. **47**, 187–200 (1999)

42. L.I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms. Physica D **60**, 259–268 (1992)
43. Z. Tan, Y. Eldar, A. Beck, A. Nehorai, Smoothing and decomposition for analysis sparse recovery. IEEE Trans. Signal Process. **62**(7), 1762–1774 (2014)
44. R. Tibshirani, Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B **58**(1), 267–288 (1996)
45. R.J. Tibshirani, Solution path of the generalized lasso. Ann. Stat. **39**(3), 1335–1371 (2011)
46. D. Wipf, S. Nagarajan, Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. IEEE J. Sel. Top. Sign. Process. **4**(2), 317–329 (2010)
47. P. Xu, Y. Fukuda, Y. Liu, Multiple parameter regularization: numerical solutions and applications to the determination of geopotential from precise satellite orbits. J. Geodesy **80**(1), 17–27 (2006)

Compressive Classification and the Rare Eclipse Problem

Afonso S. Bandeira, Dustin G. Mixon, and Benjamin Recht

Abstract This paper addresses the fundamental question of when convex sets remain disjoint after random projection. We provide an analysis using ideas from high-dimensional convex geometry. For ellipsoids, we provide a bound in terms of the distance between these ellipsoids and simple functions of their polynomial coefficients. As an application, this theorem provides bounds for compressive classification of convex sets. Rather than assuming that the data to be classified is sparse, our results show that the data can be acquired via very few measurements yet will remain linearly separable. We demonstrate the feasibility of this approach in the context of hyperspectral imaging.

Keywords Compressive classification · Gordon's theorem

A.S. Bandeira (✉)
Department of Mathematics, Courant Institute of Mathematical Sciences
and Center for Data Science, New York University, New York, NY, USA
e-mail: bandeira@cims.nyu.edu

D.G. Mixon
Department of Mathematics, The Ohio State University, Columbus, OH, USA

Department of Mathematics and Statistics, Air Force Institute of Technology,
Wright-Patterson AFB, Dayton, OH, USA
e-mail: mixon.23@osu.edu; dustin.mixon@afit.edu

B. Recht
Department of Electrical Engineering and Computer Science, University of California,
Berkeley, CA, USA

Department of Statistics, University of California, Berkeley, CA, USA
e-mail: brecht@eecs.berkeley.edu

H. Boche et al. (eds.), *Compressed Sensing and its Applications*,
Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-69802-1_6

1 Introduction

A decade of powerful results in compressed sensing and related fields have demonstrated that many signals that have low-dimensional latent structure can be recovered from very few compressive measurements. Building on this work, many researchers have shown that classification tasks can also be run on compressive measurements, provided that either the data or classifier is sparse in an appropriate basis [4, 8, 9, 11, 12, 20]. However, classification is a considerably simpler task than reconstruction, as there may be a large number of hyperplanes which successfully cleave the same data set. The question remains:

> Can we successfully classify data from even fewer compressive measurements than required for signal reconstruction?

Prior work on compressive classification has focused on preserving distances or inner products between data points. Indeed, since popular classifiers including the support vector machine and logistic regression only depend on dot products between data points, it makes sense that if dot products are preserved under a compressive measurement, then the resulting decision hyperplane should be close to the one computed on the uncompressed data.

In this paper, we take a different view of the compressive classification problem, and for some special cases, we are able to show that data can be classified with extremely few compressive measurements. Specifically, we assume that our data classes are circumscribed by disjoint convex bodies, and we seek to avoid intersection between distinct classes after projection. By studying the set of separating hyperplanes, we provide a general way to estimate the minimal dimension under which two bodies remain disjoint after random projection. In Section 3, we specialize these results to study ellipsoidal classes and give our main theoretical result—that k ellipsoids of sufficient pairwise separation remain separated after randomly projecting onto $O(\log k)$ dimensions. Here, the geometry of the ellipsoids plays an interesting and intuitive role in the notion of sufficient separation. Our results differ from prior work insofar as they can be applied to *full* dimensional data sets and are independent of the number of points in each class. We provide a comparison with principal component analysis in Section 4 by considering different toy examples of classes to illustrate strengths and weaknesses and then by applying both approaches to hyperspectral imaging data. We conclude in Section 5 with a discussion of future work.

2 Our Model and Related Work

In this section, we discuss our model for the classes as well as the underlying assumptions we apply throughout this paper. Consider an ensemble of classes $C_i \subseteq \mathbb{R}^N$ that we would like to classify. We assume that these classes are pairwise *linearly separable*, that is, for every pair i, j with $i \neq j$, there exists a hyperplane

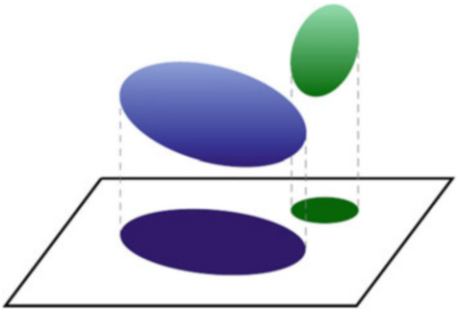


Fig. 1 Two sufficiently separated convex sets remain separated when projected onto a subspace. The rare eclipse problem asks for the smallest M such that this happens when projecting onto a random subspace of dimension M. Solving this problem for a given ensemble of classes enables dimensionality reduction in a way that ensures linear separability for classification

in $\mathbb{R}^N$ which separates C_i and C_j. Equivalently, we assume that the convex hulls $S_i := \text{hull}(C_i)$ are disjoint, and for simplicity, we assume these convex hulls are closed sets.

Linear separability is a particularly useful property in the context of classification, since to demonstrate nonmembership, it suffices to threshold an inner product with the vector normal to a separating hyperplane. Of course, in many applications, classes do not enjoy this (strong) property, but the property can be weakened to *near* linear separability, in which there exists a hyperplane that mostly distinguishes a pair of classes. One may also lift to a tensored version of the vector space and find linear separability there. Since linear separability is so useful, we use this property as the basis for our notion of distortion: We seek to project the classes $\{C_i\}_{i=1}^k$ in such a way that their images are linearly separable.

Our assumptions on the C_is and our notion of distortion both lead to a rather natural problem in convex geometry (see Figure 1 for an illustration):

Rare Eclipse Problem. Given a pair of disjoint closed convex sets $A, B \subseteq \mathbb{R}^N$ and $\eta > 0$, find the smallest M such that a random $M \times N$ projection P satisfies $PA \cap PB = \emptyset$ with probability $\geq 1 - \eta$.

At this point, we discuss some related work in the community. It appears that compressive classification was studied as early as 2006, when [12] considered a model in which each class is a point in Euclidean space. Interestingly, this bears some resemblance to the celebrated work in [14, 16], which used random projections to quickly approximate nearest neighbor search. The work in [8, 9] considered a more exotic family of classes, namely, low-dimensional manifolds—this is particularly applicable to the classification of images according to the primary object featured in each image. Along these lines of low-dimensional classes, there has since been some work in the case where classes are low-dimensional subspaces [18, 21] or unions thereof [3]. Specifically, [21] considers a Gaussian mixture model in which each Gaussian is supported on a different subspace. From

a slightly dual view, researchers have also shown that if the classifier is known to be sparse, then we can subsample the data itself, and the separating hyperplane can be determined from a number of examples roughly proportional to the sparsity of the hyperplane [4, 11, 20].

It is striking that, to date, all of the work in compressive classification has focused on classes of low dimension. This is perhaps an artifact of the mindset of compressed sensing, in which the projection preserves all information on coordinate planes of sufficiently small dimension. However, classification should not require nearly as much information as signal reconstruction does, and so we expect to be able to compressively classify into classes of full dimension; indeed, we allow two points in a common class to be mapped to the same compressive measurement, as this will not affect the classification. A Boolean version of this idea is studied in [1], which considers both random and optimality constructed projections. In the continuous setting, the closest existing work is that of Dasgupta [6, 7], which uses random projections to learn a mixture of Gaussians. In particular, Dasgupta shows that sufficiently separated Gaussians stay separated after random projection. In the next section, we prove a similar result about ellipsoids, but with a sharper notion of separation.

3 Theoretical Results

Given two disjoint closed convex bodies $A, B \subseteq \mathbb{R}^N$ and a projection dimension M, the rare eclipse problem asks whether a random $M \times N$ projection P of these bodies avoids collision, i.e., whether $PA \cap PB$ is typically empty. This can be recast as a condition on the $(N - M)$-dimensional null space of P:

$$PA \cap PB = \emptyset \qquad \Longleftrightarrow \qquad \text{Null}(P) \cap (A - B) = \emptyset,$$

where $A - B$ denotes the Minkowski difference of A and B. Of course, the null space of P is closed under scalar multiplication, and so avoiding $A - B$ is equivalent to avoiding the normalized versions of the members of $A - B$. Indeed, if we take S to denote the intersection between the unit sphere in $\mathbb{R}^N$ and the cone generated by $A - B$, then

$$PA \cap PB = \emptyset \qquad \Longleftrightarrow \qquad \text{Null}(P) \cap S = \emptyset.$$

Now suppose P is drawn so that its entries are iid $\mathcal{N}(0, 1)$. Then by rotational invariance, the distribution of its null space is uniform over the Grassmannian. As such, the rare eclipse problem reduces to a classical problem in convex geometry: Given a "mesh" (a closed subset of the unit sphere), how small must K be for a random K-dimensional subspace to "escape through the mesh," i.e., to avoid collision? It turns out that for this problem, the natural way to quantify the size of a mesh is according to its *Gaussian width*:

$$w(S) := \mathbb{E}_g\left[\sup_{z\in S}\langle z, g\rangle\right],$$

where g is a random vector with iid $\mathcal{N}(0,1)$ entries. Indeed, Gaussian width plays a crucial role in the following result, which is an improvement to the original (Corollary 3.4 in [10]); the proof is given in the appendix and follows the proof of Corollary 3.3 in [5] almost identically.

Gordon's Escape Through a Mesh Theorem. Take a closed subset S of the unit sphere in $\mathbb{R}^N$, and denote $\lambda_M := \mathbb{E}\|g\|_2$, where g is a random M-dimensional vector with iid $\mathcal{N}(0,1)$ entries. If $w(S) < \lambda_M$, then an $(N-M)$-dimensional subspace Y drawn uniformly from the Grassmannian satisfies

$$\Pr\Big(Y \cap S = \emptyset\Big) \geq 1 - \exp\left(-\frac{1}{2}\Big(\lambda_M - w(S)\Big)^2\right).$$

It is straightforward to verify that $\lambda_M \geq \sqrt{M-1}$, and so rearranging leads to the following corollary:

Corollary 1. *Take disjoint closed convex sets $A, B \subseteq \mathbb{R}^N$, and let $w_\cap$ denote the Gaussian width of the intersection between the unit sphere in $\mathbb{R}^N$ and the cone generated by the Minkowski difference $A - B$. Draw an $M \times N$ matrix P with iid $\mathcal{N}(0,1)$ entries. Then*

$$M > \Big(w_\cap + \sqrt{2\log(1/\eta)}\Big)^2 + 1 \Longrightarrow \Pr\big(PA \cap PB = \emptyset\big) \geq 1 - \eta.$$

Now that we have a sufficient condition on M, it is natural to wonder how tight this condition is. Recent work by Amelunxen et al. [2] shows that the Gordon's results are incredibly tight. Indeed, by an immediate application Theorem I and Proposition 10.1 in [2], we achieve the following characterization of a phase transition for the rare eclipse problem:

Corollary 2. *Take disjoint closed convex sets $A, B \subseteq \mathbb{R}^N$, and let $w_\cap$ denote the Gaussian width of the intersection between the unit sphere in $\mathbb{R}^N$ and the cone generated by the Minkowski difference $A - B$. Draw an $M \times N$ matrix P with iid $\mathcal{N}(0,1)$ entries. Then*

$$\begin{aligned} M \geq w_\cap^2 + \sqrt{16N\log(4/\eta)} + 1 &\Longrightarrow \Pr\big(PA \cap PB = \emptyset\big) \geq 1 - \eta, \\ M \leq w_\cap^2 - \sqrt{16N\log(4/\eta)} &\Longrightarrow \Pr\big(PA \cap PB = \emptyset\big) \leq \eta, \end{aligned}$$

Considering the second part of Corollary 2, the bound in Corollary 1 is essentially tight. Also, since Corollary 2 features an additional $\sqrt{N}$ factor in the error term of the phase transition, the bound in Corollary 1 is stronger than the first part of Corollary 2 when $w_\cap \ll \sqrt{N} - \sqrt{\log(1/\eta)}$, which corresponds to the regime where we can compress the most: $M \ll N$.

3.1 The Case of Two Balls

Corollaries 1 and 2 demonstrate the significance of Gaussian width to the rare eclipse problem. In this subsection, we observe these quantities to solve the rare eclipse problem in the special case where A and B are balls. Since each ball has its own parameters (namely, its center and radius), in this subsection, it is more convenient to write $A = S_1$ and $B = S_2$. The following lemma completely characterizes the difference cone $S_1 - S_2$:

Lemma 1. *For $i = 1, 2$, take balls $S_i := \{c_i + r_i x : x \in \mathscr{B}\}$, where $c_i \in \mathbb{R}^N$, $r_i > 0$ such that $r_1 + r_2 < \|c_1 - c_2\|$, and $\mathscr{B}$ denotes the ball centered at 0 of radius 1. Then the cone generated by the Minkowski difference $S_1 - S_2$ is the circular cone:*

$$\mathrm{Circ}(\alpha) := \{z : \langle z, c_1 - c_2\rangle \geq \|z\|\|c_1 - c_2\| \cos\alpha\},$$

where $\alpha \in (0, \pi/2)$ is the angle such that $\sin\alpha = (r_1 + r_2)/\|c_1 - c_2\|$.

In three dimensions, the fact that the difference cone is circular makes intuitive sense. The proof of Lemma 1 is routine and can be found in the appendix.

Considering the beginning on this section, it now suffices to bound the Gaussian width of the circular cone's intersection with the unit sphere $\mathbb{S}^{N-1}$. Luckily, this computation is already available as Proposition 4.3 in [2]:

$$\left(w(\mathrm{Circ}(\alpha)) \cap \mathbb{S}^{N-1}\right)^2 = N\sin^2\alpha + O(1).$$

See Figure 2 for an illustration of the corresponding phase transition. By Lemma 1 (and Corollaries 1 and 2), this means a random $M \times N$ projection will keep two balls from colliding provided

$$M \geq N\left(\frac{r_1 + r_2}{\|c_1 - c_2\|}\right)^2 + O(\sqrt{N}).$$

Note that there is a big payoff in the separation $\|c_1 - c_2\|$ between the balls. Indeed, doubling the separation decreases the required projection dimension by a factor of 4.

3.2 The Case of Two Ellipsoids

Now that we have solved the rare eclipse problem for balls, we consider the slightly more general case of ellipsoids. Actually, this case is somewhat representative of the general problem with arbitrary convex sets. This can be seen by appealing to the following result of Paouris [19]:

Theorem 1 (Concentration of Volume). *There is an absolute constant $c > 0$ such that the following holds: Given a convex set $K \subseteq \mathbb{R}^N$, draw a random vector X*

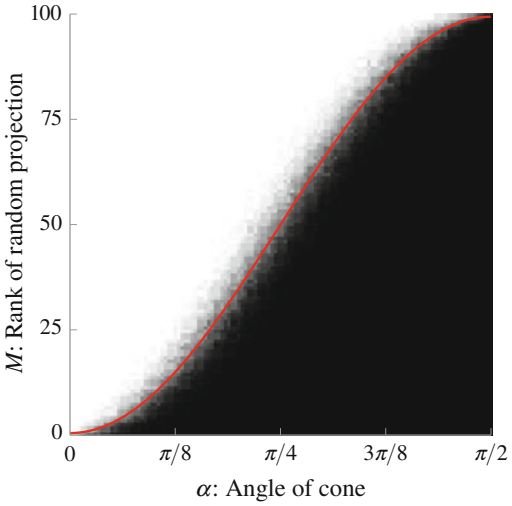


Fig. 2 Phase transition for a random null space to avoid a circular cone. Fixing the ambient dimension to be $N = 100$, then for each $\alpha = 1 : \pi/200 : \pi/2$ and $M = 1 : 100$, we randomly drew 100 $M \times N$ matrices with iid $\mathcal{N}(0, 1)$ entries and plotted the proportion whose null spaces avoided the circular cone with angle α. As expected, if α is large, then so must M so that the null space is small enough to avoid the cone. In red, we plot the curve $M = N \sin^2 \alpha + \cos 2\alpha$, which captures the phase transition by Theorem I and Proposition 4.3 in [2]. By Lemma 1, the circular cone is precisely the difference cone of two balls, and so this phase transition solves the rare eclipse problem in this special case

uniformly from K. Suppose K has the property that $\mathbb{E}[X] = 0$ *and* $\mathbb{E}[XX^\top] = I$. *Then*

$$\Pr\left(\|X\|_2 > r\right) \leq e^{-cr} \qquad \forall r \geq \sqrt{N}.$$

In words, the above theorem says that the volume of an isotropic convex set is concentrated in a round ball. The radius of the ball of concentration is $O(\sqrt{N})$, which corresponds to the fact that $\mathbb{E}\|X\|_2^2 = \mathbb{E}\,\mathrm{Tr}[XX^\top] = N$. This result can be modified to describe volume concentration of any convex set (isotropic or not). To see this, consider any convex set $K \subseteq \mathbb{R}^N$ of full dimension (otherwise the volume is zero). Then taking Y to be a random vector drawn uniformly from K, we define the centroid $c := \mathbb{E}[Y]$. Also, since K has full dimension, the inertia matrix $\mathbb{E}[(Y - c)(Y - c)^\top]$ is symmetric and positive definite, and we can take $A_0 := (\mathbb{E}[(Y - c)(Y - c)^\top])^{1/2}$. It is straightforward to verify that $X := A_0^{-1}(Y - c)$ is distributed uniformly over $K' := A_0^{-1}(K - c)$ and that K' satisfies the hypotheses of Theorem 1. We claim that Y is concentrated in an ellipsoid defined by

$$S_r := \{c + rA_0 x : x \in \mathscr{B}\}$$

for some $r \geq \sqrt{N}$, where $\mathscr{B}$ denotes the ball centered at 0 of radius 1. Indeed, Theorem 1 gives

$$\Pr\Big(Y \notin S_r\Big) = \Pr\Big(\|A_0^{-1}(Y-c)\|_2 > r\Big) \leq e^{-cr}.$$

Overall, the vast majority of any convex set is contained in an ellipsoid defined by its centroid and inertia matrix, and so two convex sets are *nearly* linearly separable if the corresponding ellipsoids are linearly separable. (A similar argument relates the case of two ellipsoids to a mixture of two Gaussians.)

Note that any ellipsoid has the following convenient form:

$$\{c + Ax : x \in \mathscr{B}\},$$

where $c \in \mathbb{R}^N$ is the center of the ellipsoid, A is some $N \times N$ symmetric and positive semidefinite matrix, and $\mathscr{B}$ denotes the ball centered at the origin of radius 1. Intuitively, the difference cone of any two ellipsoids will not be circular in general, as it was in the case of two balls. Indeed, the oblong shape of each ellipsoid (determined by its shape matrix A) precludes most of the useful symmetries in the difference cone, and as such, the analysis of the size of the cone is more difficult. Still, we established the following upper bound on the Gaussian width in the general case, which by Corollaries 1 and 2 translates to a sufficient number of rows for a random projection to typically maintain separation:

Theorem 2. *For $i = 1, 2$, take ellipsoids $S_i := \{c_i + A_i x : x \in \mathscr{B}\}$, where $c_i \in \mathbb{R}^N$, A_i is symmetric and positive semidefinite, and $\mathscr{B}$ denotes the ball centered at 0 of radius 1. Let $w_\cap$ denote the Gaussian width of the intersection between the unit sphere in $\mathbb{R}^N$ and the cone generated by the Minkowski difference $S_1 - S_2$. Then*

$$w_\cap \leq \frac{\|A_1\|_F + \|A_2\|_F}{\zeta - \big(\|A_1 e\|_2 + \|A_2 e\|_2\big)} + \frac{1}{\sqrt{2\pi}}$$

provided $\zeta > \|A_1 e\|_2 + \|A_2 e\|_2$; here, $\zeta := \|c_2 - c_1\|$ and $e := (c_1 - c_2)/\|c_1 - c_2\|$.

The proof is technical and can be found in the appendix, but the ideas behind the proof are interesting. There are two main ingredients, the first of which is the following result:

Proposition 1 (Proposition 3.6 in [5]). *Let $\mathscr{C}$ be any nonempty convex cone in $\mathbb{R}^N$, and let g be an N-dimensional vector with iid $\mathscr{N}(0,1)$ entries. Then*

$$w(\mathscr{C} \cap \mathbb{S}^{N-1}) \leq \mathbb{E}_g\Big[\|g - \Pi_{\mathscr{C}^*}(g)\|_2\Big],$$

where $\Pi_{\mathscr{C}^*}$ denotes the Euclidean projection onto the dual cone $\mathscr{C}^*$ of $\mathscr{C}$.

Proposition 1 is essentially a statement about convex duality, and while it provides an upper bound on $w_\cap$, in our case, it is difficult to find a closed form expression for the right-hand side. However, the bound is in terms of distance to the dual cone, and so any point in this cone provides an upper bound on this distance.

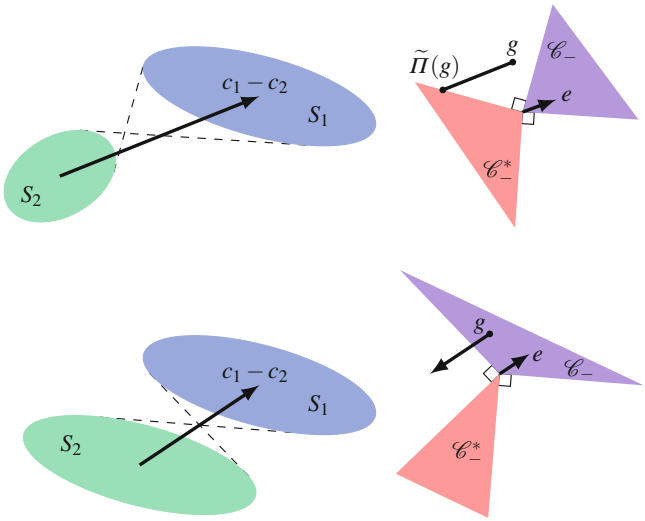


Fig. 3 Two examples of two ellipsoids along with their difference cone and dual cone. For each pair, on the left, the vector $c_1 - c_2$ is depicted, as are the extreme difference directions between the ellipsoids—these form the boundary of the difference cone $\mathscr{C}_-$, which is illustrated on the right along with its dual cone $\mathscr{C}_-^*$, i.e., the cone of separating hyperplanes. The vector $e \in \mathscr{C}_-$ is a normalized version of $c_1 - c_2$. In the first example, the pseudoprojection $\widetilde{\Pi}$ sends any point g to the closest point in the dual cone $\mathscr{C}_-^*$ along the line spanned by e. Interestingly, in cases where the ellipsoids are far apart, the cone $\mathscr{C}_-$ will be narrow, and so the boundary of the dual cone will essentially be the orthogonal complement of e. As such, the pseudoprojection is close to the true projection onto the polar cone in this limiting case. For this pseudoprojection to be well-defined, we require that for every g, the line which passes through g in the direction of e hits the dual cone at some point. This is not always the case, as the second example illustrates. It is straightforward to show that this pseudoprojection is well-defined if and only if the ellipsoids remain separated when projecting onto the line spanned by e

This leads to the second main ingredient in our analysis: We choose a convenient mapping $\widetilde{\Pi}$ that sends any vector g to a point in $\mathscr{C}^*$ (but not necessarily the closest point) while at the same time allowing the expectation of $\|g - \widetilde{\Pi}(g)\|_2$ to have a closed form. Since $\|g - \Pi_{\mathscr{C}^*}(g)\|_2 \leq \|g - \widetilde{\Pi}(g)\|_2$ for every possible instance of g, this produces a closed-form upper bound on the bound in Proposition 1.

Figure 3 illustrates how we chose the pseudoprojection $\widetilde{\Pi}$. Interestingly, this pseudoprojection behaves more like the true projection when the ellipsoids are more distant from each other. At the other extreme, note that Theorem 2 does not hold if the ellipsoids are too close, i.e., if $\|c_1 - c_2\| \leq \|A_1 e\|_2 + \|A_2 e\|_2$. This occurs, for example, if the two ellipsoids collide after projecting onto the span of e; indeed, taking x and y to be unit norm vectors such that $e^\top(c_1 + A_1 x) = e^\top(c_2 + A_2 y)$, then rearranging gives

$$\|c_1 - c_2\| = e^\top c_1 - e^\top c_2 = -e^\top A_1 x + e^\top A_2 y \leq |e^\top A_1 x| + |e^\top A_2 y| \leq \|A_1 e\|_2 + \|A_2 e\|_2.$$

As Figure 3 illustrates, our pseudoprojection even fails to be well-defined when the ellipsoids collide after projecting onto the span of e. So why bother using a random projection to maintain linear separability when there is a rank-1 projection available? There are two reasons: First, calculating this rank-1 projection requires access to the centers of the ellipsoids, which are not available in certain applications (e.g., unsupervised or semi-supervised learning or if the projection occurs blindly during the data collection step). Second, the use of a random projection is useful when projecting multiple ellipsoids simultaneously to preserve pairwise linear separability—as we will detail in the next subsection, randomness allows one to appeal to the union bound in a way that permits several ellipsoids to be projected simultaneously using particularly few projected dimensions.

At this point, we compare Theorem 2 to the better understood case of two balls. In this case, $A_1 = r_1 I$ and $A_2 = r_2 I$, and so Theorem 2 gives that

$$w_\cap \leq \sqrt{N} \cdot \frac{r_1 + r_2}{\|c_1 - c_2\|_2 - (r_1 + r_2)} + \frac{1}{\sqrt{2\pi}}.$$

If we consider the regime in which $r_1 + r_2 \leq \frac{1}{2}\|c_1 - c_2\|_2$, then we recover the case of two balls to within a factor of 2, suggesting that the analysis is tight (at least in this case). For a slightly more general lower bound, note that a projection maintains separation between two ellipsoids only if it maintains separation between balls contained in each ellipsoid. The radius of the largest ball in the ith ellipsoid is equal to the smallest eigenvalue $\lambda_{\min}(A_i)$ of the shape matrix A_i, and the center of this ball coincides with the center c_i of its parent ellipsoid. As such, we can again appeal to the case of two balls to see that Theorem 2 is reasonably tight for ellipsoids of reasonably small eccentricity $\lambda_{\max}(A_i)/\lambda_{\min}(A_i)$. Closed-form bounds for general ellipses with high eccentricity are unwieldy, but Figure 4 illustrates that our bound is far from tight when the ellipsoids are close to each other. Still, the bound improves considerably as the distance increases. As such, we leave improvements to Theorem 2 as an open problem (in particular, finding a closed-form characterization of the phase transition in terms of the c_is and A_is).

3.3 The Case of Multiple Convex Sets

Various classification tasks require one to distinguish between several different classes, and so one might ask for a random projection to maintain pairwise linear separability. For a fixed projection dimension M, let η_{ij} denote the probability that convex classes S_i and S_j collide after projection. Then the union bound gives that the probability of maintaining separation is $\geq 1 - \sum_{i,j:i<j} \eta_{ij}$.

This use of the union bound helps to illustrate the freedom which comes with a random projection. Recall that Theorem 2 requires that projecting the ellipsoids onto the line spanned by the difference $c_1 - c_2$ of their centers maintains separation. In the case of multiple ellipsoids, one might then be inclined to project onto the span of $\{c_i - c_j\}_{i,j:i<j}$. Generically, such a choice of projection puts $M = \binom{K}{2} = \Omega(K^2)$,

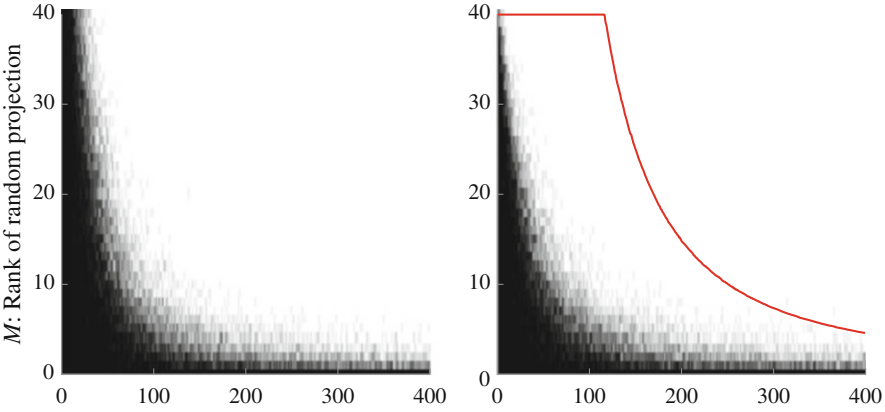


Fig. 4 Phase transition for a random projection to keep ellipsoids separated. (**a**) Fixing the ambient dimension to be $N = 40$, then for each $\zeta = 1 : 400$ and $M = 1 : 40$, we conducted *ten* trials. For each trial, we randomly drew A_1 and A_2 as iid standard Wishart-distributed $N \times N$ matrices with N degrees of freedom (i.e., $A_i = XX^\top$, where X is $N \times N$ with iid $\mathcal{N}(0,1)$ entries), along with an $M \times N$ matrix P with iid $\mathcal{N}(0,1)$ entries. Plotted is the proportion of trials for which the ellipsoids are disjoint after applying P (we did not record whether the ellipsoids were separated before projection). For each of the 160,000 trials, the shape matrices satisfied $\zeta \leq \|A_1 e\|_2 + \|A_2 e\|_2$, thereby rendering Theorem 2 irrelevant. (**b**) Next, we performed the same experiment, except we changed the distribution of A_1 and A_2 so that e is in the null space of both, and in the orthogonal complement of e, they are iid standard Wishart-distributed $(N-1) \times (N-1)$ matrices with $N-1$ degrees of freedom. As such, the corresponding ellipsoids resided in parallel hyperplanes, and $\|A_1 e\|_2 + \|A_2 e\|_2 = 0$ so that Theorem 2 applies. For each trial, we stored the bound on $w_\cap$ from Theorem 2 and calculated the sample average of the squares of these bounds corresponding to each $\zeta = 1 : 400$. The red curve plots these sample averages (or 40, whichever is smaller)—think of this as an upper bound on the phase transition. As one might expect, this bound appears to sharpen as the distance increases

where K is the total number of classes. On the other hand, suppose each pairwise distance $\|c_i - c_j\|$ is so large that the (i,j)th Gaussian width satisfies

$$w_\cap < \sqrt{2 \log \left(\frac{1}{p} \binom{K}{2} \right)}.$$

Then by Corollary 1, taking $M = 8 \log(\binom{K}{2}/p) + 1 = O_p(\log K)$ ensures that classes S_i and S_j collide after projection with probability $\eta_{ij} \leq p/\binom{K}{2}$, and so the probability of maintaining overall separation is $\geq 1 - p$. Of course, we will not save so much in the projection dimension when the convex bodies are closer to each other, but we certainly expect $M < K^2$ in reasonable cases.

At this point, we note the similarity between the performance $M = O(\log K)$ and what the Johnson–Lindenstrauss lemma guarantees when the classes are each

a single point. Indeed, a random projection of $M = \Omega_\epsilon(\log K)$ dimensions suffices to ensure that pairwise distances are preserved to within a factor of $1 \pm \epsilon$ with constant probability; this in turn ensures that pairwise separated points remain pairwise separated after projection. In fact, the proof technique for the Johnson–Lindenstrauss lemma is similar: First prove that a random projection typically preserves the norm of any vector, and then perform a union bound over all $\binom{K}{2}$ difference vectors. One might be inspired to use Johnson–Lindenstrauss ideas to prove a result analogous to Theorem 2 (this was actually an initial attempt by the authors). Unfortunately, since Johnson–Lindenstrauss does not account for the shape matrices A_i of the ellipsoids, one is inclined to consider worst-case orientations, and so terms like $\|A_i e\|_2$ are replaced by spectral norms $\|A_i\|_2$ in the analysis, thereby producing a strictly weaker result. Dasgupta [6] uses this Johnson–Lindenstrauss approach to project a mixture of Gaussians while maintaining some notion of separation.

4 Random Projection Versus Principal Component Analysis

In this section, we compare the performance of random projection and principal component analysis (PCA) for dimensionality reduction. First, we should briefly review how to perform PCA. Consider a collection of data points $\{x_i\}_{i=1}^p \subseteq \mathbb{R}^N$, and define the empirical mean by $\bar{x} := \frac{1}{p}\sum_{i=1}^p x_i$. Next, consider the empirical inertia matrix:

$$\widehat{\Sigma} := \frac{1}{p}\sum_{i=1}^p (x_i - \bar{x})(x_i - \bar{x})^\top = \frac{1}{p}\sum_{i=1}^p x_i x_i^\top - \bar{x}\bar{x}^\top.$$

The eigenvectors of $\widehat{\Sigma}$ with the largest eigenvalues are identified as the *principal components*, and the idea of PCA is to project $\{x_i\}_{i=1}^p$ onto the span of these components for dimensionality reduction.

In this section, we will compare random projection with PCA in a couple of ways. First, we observe some toy examples of data sets that illustrate when PCA is better and when random projection is better. Later, we make a comparison using a real-world hyperspectral data set.

4.1 Comparison Using Toy Examples

Here, we consider a couple of extreme data sets which illustrate when PCA outperforms random projection and vice versa. Our overarching model for the data sets will be the following: Given a collection of disjoint balls $\{S_i\}_{i=1}^K$ in $\mathbb{R}^N$, we independently draw p data points uniformly from $S := \bigcup_{i=1}^K S_i$. When p is large, we can expect $\widehat{\Sigma}$ to be very close to

$$\Sigma := \frac{1}{\mathrm{vol}(S)} \sum_{i=1}^{K} \int_{S_i} xx^\top dx - \mu\mu^\top$$

by the law of large numbers; here, $\mu \in \mathbb{R}^N$ denotes the mean of the distribution. Recall that the projection dimension for PCA is the number of large eigenvalues of $\widehat{\Sigma}$. Since the operator spectrum is a continuous function of the operator, we can count large eigenvalues of Σ to estimate this projection dimension. The following lemma will be useful to this end:

Lemma 2. *Consider a ball of the form $S := c + r\mathscr{B}$, where $\mathscr{B} \subseteq \mathbb{R}^N$ denotes the ball centered at 0 of radius 1. Define the operator:*

$$W := \int_S xx^\top dx.$$

Then the span of c and its orthogonal complement form the eigenspaces of W with eigenvalues:

$$\lambda_c = r^N \|c\|^2 \,\mathrm{vol}(\mathscr{B}) + Cr^{N+2}, \qquad \lambda_{c^\perp} = Cr^{N+2},$$

respectively, where C is some constant depending on N.

Proof. Pick any vector $v \in \mathbb{R}^N$ of unit norm. Then

$$v^\top W v = \int_{\mathscr{B}} v^\top (c + ry)(c + ry)^\top v r^N dy = (v^\top c)^2 \cdot r^N \,\mathrm{vol}(\mathscr{B}) + r^{N+2} v^\top \left(\int_{\mathscr{B}} yy^\top dy \right) v.$$

Notice that the operator $\int_{\mathscr{B}} yy^\top dy$ is invariant under conjugation by any rotation matrix. As such, this operator is a constant C multiple of the identity operator. Thus, $v^\top W v$ is maximized at λ_c when v is a normalized version of c and minimized at $\lambda_{c^\perp}$ whenever v is orthogonal to c. □

We start by considering the case where S is composed of two balls, namely, $S_1 := c + r\mathscr{B}$ and $S_2 := -c + r\mathscr{B}$. As far as random projection is concerned, in this case, we are very familiar with the required projection dimension: $\Omega_\eta(Nr^2/\|c\|^2)$. In particular, as $\|c\|$ approaches r, a random projection cannot provide much dimensionality reduction. To compare with PCA, note that in this case, Σ is a scalar multiple of $W_1 + W_2$, where

$$W_i := \int_{S_i} xx^\top dx.$$

Moreover, it is easy to show that $W_1 = W_2$. By Lemma 2, the dominant eigenvector of W_i is c, and so PCA would suggest to project onto the one-dimensional subspace spanned by c. Indeed, this projection always preserves separation, and so in this case, PCA provides a remarkable savings in projection dimension.

Now consider the case where S is composed of $2N$ balls $\{S_{n,1}\}_{n=1}^N \cup \{S_{n,2}\}_{n=1}^N$ defined by $S_{n,1} := e_n + r\mathscr{B}$ and $S_{n,2} := -e_n + r\mathscr{B}$, where e_n denotes the nth identity basis element. Then Σ is a scalar multiple of $\sum_{n=1}^N (W_{n,1} + W_{n,2})$, where

$$W_{n,i} := \int_{S_{n,i}} xx^\top dx.$$

Recall that $W_{n,1} = W_{n,2}$. Then Σ is simply a scalar multiple of $\sum_{n=1}^N W_{n,1}$. By Lemma 2, the $W_{n,1}$s are all diagonal, and their diagonals are translates of each other. As such, their sum (and therefore Σ) is a scalar multiple of the identity matrix—in this case, PCA would choose to not project down to fewer dimensions. On the other hand, if we take

$$M > \left(\sqrt{N\left(\frac{2r}{\sqrt{2}}\right)^2 + 1} + \sqrt{2\log\left(\frac{1}{p}\binom{2N}{2}\right)} \right)^2 + 1,$$

then by Corollary 1, a random projection maintains separation between any fixed pair of balls from $\{S_{n,1}\}_{n=1}^N \cup \{S_{n,2}\}_{n=1}^N$ with probability $\geq 1 - p/\binom{2N}{2}$, and so by the union bound, the balls are pairwise separated with probability $\geq 1 - p$. In particular, if $r = O(N^{-1/2})$, then we can take $M = O_p(\log N)$.

Overall, random projection performs poorly when the classes are close, but when there are multiple sufficiently separated classes, you can expect a dramatic dimensionality reduction. As for PCA, we have constructed a toy example for which PCA performs well (the case of two balls), but in general, the performance of PCA seems difficult to describe theoretically. Whereas the performance of random projection can be expressed in terms of "local" conditions (e.g., pairwise separation), as the last example illustrates, the performance of PCA can be dictated by more "global" conditions (e.g., the geometric configuration of classes). In the absence of theoretical guarantees for PCA, the following subsection provides simulations with real-world hyperspectral data to illustrate its performance compared to random projection.

4.2 Simulations with Hyperspectral Data

One specific application of dimensionality reduction is the classification of hyperspectral data. For this application, the idea is to distinguish materials by observing them across hundreds of spectral bands (like the red, green, and blue bands that the human eye detects). Each pixel of a hyperspectral image can be viewed as a vector of spectral information, capturing how much light of various frequencies is being reradiated from that portion of the scene. A hyperspectral image is naturally represented as a data cube with two spatial indices and one spectral index, and a common task is to identify the material observed at each pair of spatial indices.

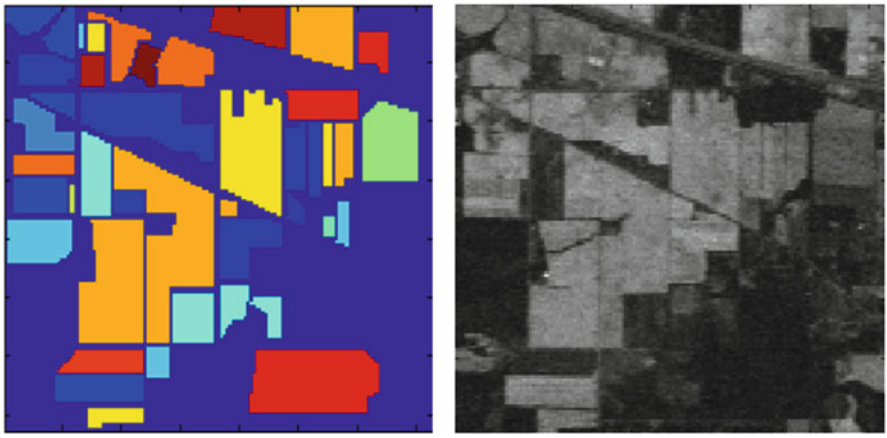


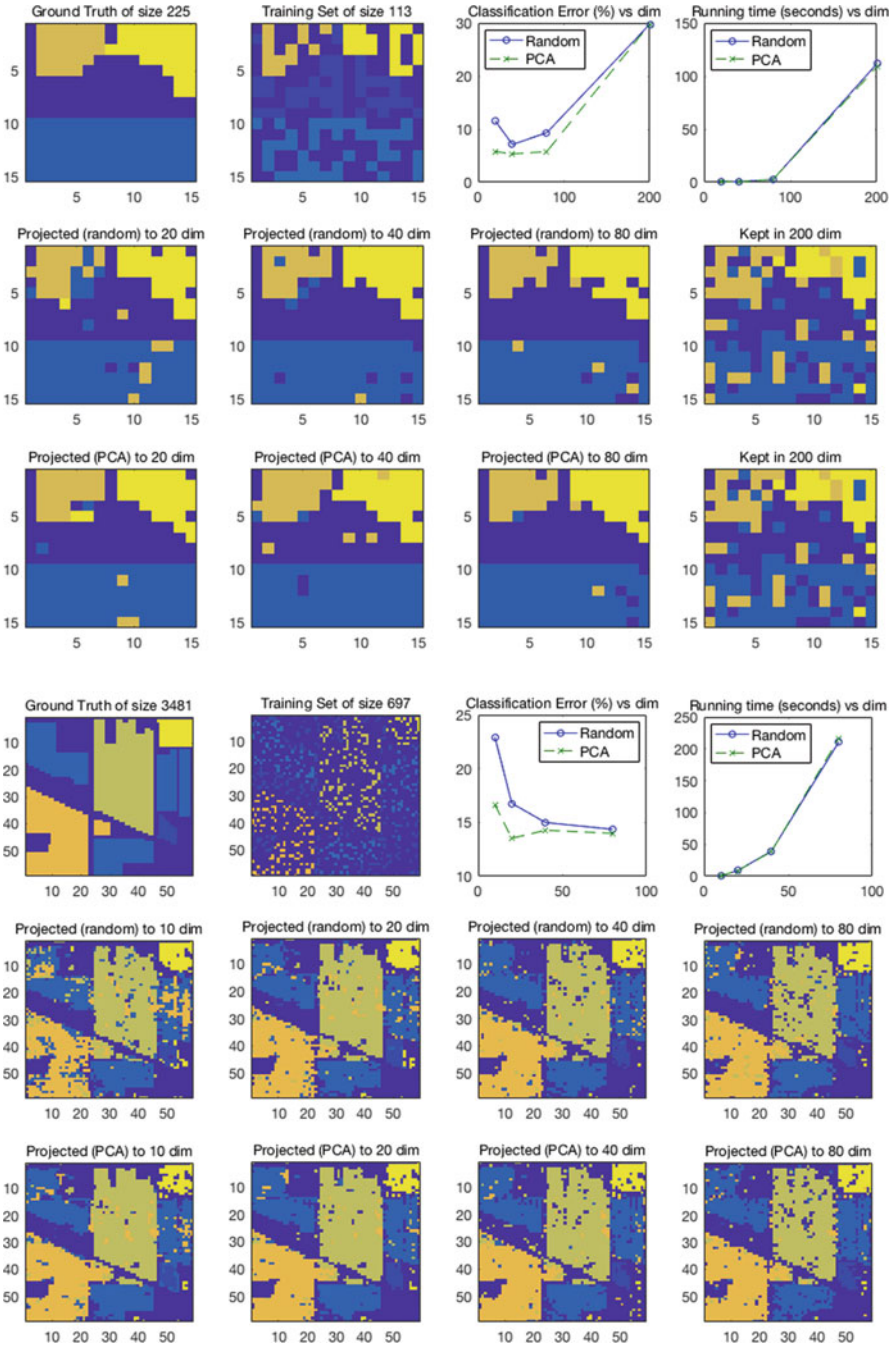
Fig. 5 The Indian Pines hyperspectral data set [13]. Each pixel corresponds to a different type of vegetation or crop. The ground truth image of labels is depicted on the left, and a sample band of the data set is displayed on the right

To do this, one might apply *per-pixel classification*, in which a classifier simply identifies the material in a given pixel from its spectral content, ignoring any spatial context. Since the spectral information is high-dimensional, it is natural to attempt dimensionality reduction before classification. A popular choice for this task is PCA [15, 22], and in this subsection, we provide some preliminary simulations to compare its performance with random projection.

All experiments described in this subsection were conducted using the Indian Pines hyperspectral data set [13]. This data set consists of a hyperspectral image with 145×145 pixels, each containing spectral reflectance data represented by a vector of length $N = 200$. Each pixel corresponds to a particular type of vegetation or crop, such as corn or wheat, with a total of 17 different classes (see Figure 5 for an illustration).

For our simulations, the task will consist of using the known labels of a training set (a small subset of the $21,025 = 145 \times 145$ pixels) to make accurate predictions for the remainder of the pixels. To keep the simulations fast, each simulation considers a small patch of pixels. More precisely, given a patch of p pixels and a prescribed training ratio r, we pick a random subset of the pixels of size rp to be the training set. We use the labels from this training set to train a classifier that will then attempt to guess the label of each of the other $(1 - r)p$ pixels from the location of its spectral reflectance in 200-dimensional space. The classifier we use is MATLAB's built-in implementation of multinomial logistic regression. Performance is measured by classification error and runtime.

Given this setting, for different values of projection dimension M, we draw an $M \times N$ matrix P with iid $\mathcal{N}(0, 1)$ entries and replace every spectral reflectance data point x by Px. In the degenerate case $M = N$, we simply take P to be the identity matrix. For comparison, we also use principal component analysis (PCA) for dimensionality reduction, which will interrogate the training set to identify


Ground Truth of size 225
Training Set of size 113
Classification Error (%) vs dim
Random
PCA
Running time (seconds) vs dim
Random
PCA
Projected (random) to 20 dim
Projected (random) to 40 dim
Projected (random) to 80 dim
Kept in 200 dim
Projected (PCA) to 20 dim
Projected (PCA) to 40 dim
Projected (PCA) to 80 dim
Kept in 200 dim
Ground Truth of size 3481
Training Set of size 697
Classification Error (%) vs dim
Random
PCA
Running time (seconds) vs dim
Random
PCA
Projected (random) to 10 dim
Projected (random) to 20 dim
Projected (random) to 40 dim
Projected (random) to 80 dim
Projected (PCA) to 10 dim
Projected (PCA) to 20 dim
Projected (PCA) to 40 dim
Projected (PCA) to 80 dim

M principal components before projecting the data set onto the span of these components. An immediate advantage of random projection is that it allows the sensing mechanism to blindly compress the data, as it does not need a training set to determine the compression function.

Figure uses different patches of the Indian Pines data set and different training ratios to compare both the classification accuracy and runtime of multinomial logistic regression when applied to various projections of the data set. The first experiment focuses on a small patch of 225 pixels, and the second considers a patch of 3481 pixels. These experiments reveal a few interesting phenomena. First of all, dimensionality reduction leads to impressive speedups in runtime. Perhaps more surprising is the fact that there seems to be an improvement in classification performance after projecting the data. We are far from completely understanding this behavior, but we suspect it has to do with regularization and overfitting.

It is also interesting how similar random projection and PCA perform. Note that the PCA method has an unfair advantage since it is data-adaptive, meaning that it uses the training data to select the projection, and in practical applications for which the sensing process is expensive, one might be interested in projecting in a nonadaptive way, thereby allowing for less sensing. Our simulations suggest that the expense is unnecessary, as a random projection will provide almost identical performance. As indicated in the previous subsection, random projection is also better understood as a means to maintain linear separability, and so there seems to be little benefit in choosing PCA over random projection (at least for this sort of classification task).

5 Future Work

One of the main points of this paper is that random projections can maintain separation between sufficiently separated sets, and this is useful for classification in the projected domain. Given the mindset of compressed sensing, it is impressive that the sets need not be low-dimensional to enjoy separation in the projected domain. What this suggests is a more general notion of simplicity that is at play, of

Fig. 6 The performance of classification by multinomial logistic regression after projecting onto subspaces of various dimensions M. Depicted are two particular patches of the entire Indian Pines data set—the top uses a patch of 225 pixels, while the bottom uses a patch of 3481 pixels. In each case, the first two plots in the first row depict the ground truth labels in the patch, as well as the random training set we selected. The third plot compares, for different values of projection dimension M, the classification error incurred with random projection and with principal component analysis. The fourth plot shows the runtime (in seconds) for different values of M. The second and third rows depict the classification outcomes when using random projection and PCA, respectively. One can see that dimensionality reduction not only speeds up the algorithm but also improves the classification performance by discouraging overfitting

which low-dimensionality and sufficient separation are mere instances. Obviously, understanding this general notion is a worthy subject of future work.

From a more applied perspective, it would be worth investigating alternative notions of distortion. Indeed, linear separability is the best-case scenario for classification, but it is not at all necessary. After identifying any worthy notion of distortion, one might study how much distortion is incurred by random projection, and hopefully some of the ideas contained in this paper will help.

One of our main results (Theorem 2) provides a sufficient number of rows for a random projection to maintain separation between ellipsoids. However, as illustrated in Figure 4, this bound is far from optimal. Considering this case of two ellipsoids is somewhat representative of the more general case of two convex sets (as we identified using Theorem 1), improvements to Theorem 2 would be rather interesting. In particular, it would be nice to characterize the phase transition in terms of the ellipsoids' parameters, as we already have in the case of two balls.

Finally, the random projections we consider here all have iid $\mathcal{N}(0,1)$ entries, but real-world sensing systems may not enjoy this sort of flexibility. As such, it would be interesting to extend the results of this paper to more general classes of random projections, in particular, random projections which can be implemented with a hyperspectral imager (say).

6 Appendix: Proofs

6.1 *Proof of Gordon's Escape Through a Mesh Theorem*

This proof is chiefly based on the following result, which appears as Corollary 1.2 in [10]:

Gordon's Comparison Theorem. Let S be a closed subset of $\mathbb{S}^{n-1}$. Draw an $M \times N$ matrix P with iid $\mathcal{N}(0,1)$ entries. Then

$$\mathbb{E}\left[\min_{x \in S} \|Px\|_2\right] \geq \lambda_M - w(S),$$

where $\lambda_M := \mathbb{E}\|g\|_2$ and g is a random M-dimensional vector with iid $\mathcal{N}(0,1)$ entries.

To prove the escape theorem, consider the function:

$$f_S : P \mapsto \min_{x \in S} \|Px\|_2.$$

Gordon's comparison theorem gives that $\mathbb{E}[f_S] \geq \lambda_M - w(S)$, and so

$$\Pr\Big(Y \cap S = \emptyset\Big) = \Pr\Big(\min_{x \in S} \|Px\|_2 > 0\Big)$$

$$= \Pr\Big(\min_{x\in S}\|Px\|_2 > \big(\lambda_M - w(S)\big) - \big(\lambda_M - w(S)\big)\Big)$$
$$\geq \Pr\Big(\min_{x\in S}\|Px\|_2 > \mathbb{E}[f_S] - \big(\lambda_M - w(S)\big)\Big). \tag{1}$$

Next, we note that f_S is Lipschitz with respect to the Frobenius norm with constant 1, and so we can appeal to (1.6) of [17] to get

$$\Pr\Big(f_S(P) > \mathbb{E}[f_S] - t\Big) \geq 1 - e^{-t^2/2} \qquad \forall t > 0. \tag{2}$$

Taking $t = \lambda_M - w(S)$ and applying (2) to (1) then gives the result.

6.2 *Proof of Lemma 1*

Let $\mathscr{C}_-$ denote the cone generated by the Minkowski difference $S_1 - S_2$. We will show $\mathscr{C}_- = \mathrm{Circ}(\alpha)$ by verifying both containments.

We begin by finding the smallest $\alpha \in [0, \pi/2]$ for which $\mathscr{C}_- \subseteq \mathrm{Circ}(\alpha)$. By the definition of $\mathrm{Circ}(\alpha)$, this containment is equivalent to

$$\cos\alpha \leq \inf_{z\in\mathscr{C}_-} \frac{\langle z, c_1 - c_2\rangle}{\|z\|\|c_1 - c_2\|} = \min_{z\in S_1 - S_2} \frac{\langle z, c_1 - c_2\rangle}{\|z\|\|c_1 - c_2\|}. \tag{3}$$

To find the smallest such α, we solve this optimization problem. Taking $d := c_1 - c_2$, then $S_1 - S_2 = (r_1 + r_2)\mathscr{B} + d$, and so we seek to

$$\text{minimize} \quad f(y) = \frac{\langle y + d, d\rangle}{\|y + d\|\|d\|} \quad \text{subject to} \quad \|y\| \leq r_1 + r_2.$$

Quickly note that the objective function is well-defined over the feasibility region due to the assumption $r_1 + r_2 < \|d\|$. We first claim that $f(y)$ is minimized on the boundary, i.e., where $\|y\| = r_1 + r_2$. To see this, suppose $\|y\| < r_1 + r_2$, and letting $P_{d^\perp}$ denote the orthogonal projection onto the orthogonal complement of the span of d, take $t > 0$ such that $\|y + tP_{d^\perp}y\| = r_1 + r_2$. Then $y + tP_{d^\perp}y$ lies on the boundary and

$$f(y + tP_{d^\perp}y) = \frac{\langle y + tP_{d^\perp}y + d, d\rangle}{\|y + tP_{d^\perp}y + d\|\|d\|} = \frac{\langle y + d, d\rangle}{\|y + tP_{d^\perp}y + d\|\|d\|}$$
$$< \frac{\langle y + d, d\rangle}{\|y + d\|\|d\|} = f(y).$$

As such, it suffices to minimize subject to $\|y\| = r_1 + r_2$. At this point, the theory of Lagrange multipliers can be applied since the equality constraint $g(y) := \|y\|^2 = (r_1 + r_2)^2$ is a level set of a function whose gradient $\nabla g(y) = 2y$ does not vanish

on the level set. Thus, the minimizers of f with $g(y) = (r_1 + r_2)^2$ satisfy $\nabla f(y) = -\lambda \nabla g(y) = -2\lambda y$ for some Lagrange multiplier $\lambda \in \mathbb{R}$.

To continue, we calculate $\nabla f(y)$. It is actually easier to calculate the gradient of $h(u) := \langle u, d\rangle / \|u\|\|d\|$:

$$\nabla h(u) = \frac{1}{\|u\|^2}\left(d - \left\langle \frac{u}{\|u\|}, d \right\rangle \frac{u}{\|u\|}\right).$$

Note that $\nabla h(u) = 0$ only if u is a nontrivial multiple of d, i.e., only if u maximizes h (by Cauchy–Schwarz). Also, it is easy to verify that $\langle u, \nabla h(u)\rangle = 0$. Overall, changing variables $u \leftarrow y + d$ gives that any minimizer $y^\natural$ of f subject to $\|y\| = r_1 + r_2$ satisfies

$$\nabla f(y^\natural) = -2\lambda y^\natural \qquad \text{for some } \lambda \in \mathbb{R}, \tag{4}$$

$$\nabla f(y^\natural) \neq 0, \tag{5}$$

$$\langle y^\natural + d, \nabla f(y^\natural)\rangle = 0. \tag{6}$$

At this point, (4) and (5) together imply that $\nabla f(y^\natural)$ is a nontrivial multiple of $y^\natural$, and so combining with (6) gives

$$\langle y^\natural + d, y^\natural \rangle = 0.$$

As such, 0, d, and $y^\natural + d$ form vertices of a right triangle with hypotenuse $\|d\|$, and the smallest α satisfying (3) is the angle between d and $y^\natural + d$. Thus, $\sin\alpha = \|y^\natural\|/\|d\| = (r_1 + r_2)/\|c_1 - c_2\|$.

It remains to prove the reverse containment, $\mathrm{Circ}(\alpha) \subseteq \mathscr{C}_-$, for this particular choice of α. Define

$$G := \{z : \langle z, d\rangle = \|z\|\|d\|\cos\alpha,\ \|z\| = \|y^\natural + d\|\}.$$

Then $\mathrm{Circ}(\alpha)$ is the cone generated by G, and so it suffices to show that $G \subseteq S_1 - S_2 = (r_1 + r_2)\mathscr{B} + d$. To this end, pick any $z \in G$, and consider the triangle with vertices 0, d, and z. By definition, the angle between d and z is α, and the side z has length $\|y^\natural + d\|$. As such, by the side-angle-side postulate, this triangle is congruent to the triangle with vertices at 0, d, and $y^\natural + d$. This implies that the side between z and d has length $\|z - d\| = \|y^\natural\| = r_1 + r_2$, and so $z = (z - d) + d \in (r_1 + r_2)\mathscr{B} + d$, as desired.

6.3 *Proof of Theorem 2*

This proof makes use of the following lemma:

Lemma 3. *Take an $n \times n$ matrix A and let g have iid $\mathscr{N}(0, 1)$ entries. Then*

$$\sqrt{\frac{2}{\pi}}\|A\|_F \le \mathbb{E}\|Ag\|_2 \le \|A\|_F.$$

Proof. Let $A = UDV$ be the singular value decomposition of A. Since the Gaussian is isotropic, $\mathbb{E}\|Ag\|_2 = \mathbb{E}\|Dg\|_2$, and since the function $x \mapsto x^2$ is convex, Jensen's inequality gives

$$\mathbb{E}\|Dg\|_2 \le \sqrt{\mathbb{E}\|Dg\|_2^2} = \sqrt{\sum_{i=1}^n D_{ii}^2 \mathbb{E} g_i^2} = \|D\|_F = \|A\|_F.$$

Similarly, since $x \mapsto \|x\|_2$ is convex, we can also use Jensen's inequality to get

$$\mathbb{E}\|Dg\|_2 = \mathbb{E}\sqrt{\sum_{i=1}^n D_{ii}^2 g_i^2} \ge \sqrt{\sum_{i=1}^n \left(\mathbb{E}|D_{ii}g_i|\right)^2} = \mathbb{E}|g_1| \sqrt{\sum_{i=1}^n D_{ii}^2} = \sqrt{\frac{2}{\pi}}\|A\|_F,$$

which completes the proof. □

To prove Theorem 2, let $\mathscr{C}_-$ denote the cone generated by the Minkowski difference $S_1 - S_2$. We will exploit Proposition 1, which gives the following estimate in terms of the polar cone $\mathscr{C}_-^* := \{w : \langle w, z\rangle \le 0 \ \forall z \in \mathscr{C}_-\}$:

$$w_\cap \le \mathbb{E}_g\Big[\|g - \Pi_{\mathscr{C}_-^*}(g)\|_2\Big],$$

where g has iid $\mathscr{N}(0,1)$ entries and $\Pi_{\mathscr{C}_-^*}$ denotes the Euclidean projection onto $\mathscr{C}_-^*$. Instead of directly computing the distance between g and its projection onto $\mathscr{C}_-^*$, we will construct a mapping $\widetilde{\Pi}$ which sends g to some member of $\mathscr{C}_-^*$, but for which distances are easier compute; indeed $\|g - \widetilde{\Pi}(g)\|_2$ will be an upper bound on $\|g - \Pi_{\mathscr{C}_-^*}(g)\|_2$. Consider the polar decomposition $c_2 - c_1 = \zeta e$, where $\zeta > 0$. Then we can decompose $g = g_1 e + g_2$, and we define $\widetilde{\Pi}(g)$ to be the point in $\mathscr{C}_-^*$ of the form $\alpha e + g_2$ which is closest to g. With this definition, we have

$$\|g - \Pi_{\mathscr{C}_-^*}(g)\|_2 \le \|g - \widetilde{\Pi}(g)\|_2 = \min|g_1 - \alpha| \quad \text{s.t.} \quad \alpha e + g_2 \in \mathscr{C}_-^*.$$

To simplify this constraint, we find a convenient representation of the polar cone:

$$\begin{aligned}
\mathscr{C}_-^* &= \{w : \langle w, z\rangle \le 0 \ \forall z \in \mathscr{C}_-\} \\
&= \{w : \langle w, u - v\rangle \le 0 \ \forall u \in S_1, v \in S_2\} \\
&= \{w : \langle w, c_2 - c_1\rangle \ge \langle w, A_1 x\rangle - \langle w, A_2 y\rangle \ \forall x, y \in \mathscr{B}\} \\
&= \Big\{w : \langle w, c_2 - c_1\rangle \ge \max_{x\in\mathscr{B}}\langle w, A_1 x\rangle + \max_{y\in\mathscr{B}}\langle w, -A_2 y\rangle\Big\}
\end{aligned}$$

$$= \left\{ w : \langle w, c_2 - c_1 \rangle \geq \max_{x \in \mathscr{B}} \langle A_1^\top w, x \rangle + \max_{y \in \mathscr{B}} \langle -A_2^\top w, y \rangle \right\}$$
$$= \{ w : \langle w, c_2 - c_1 \rangle \geq \|A_1 w\|_2 + \|A_2 w\|_2 \},$$

where the last step uses the fact that each A_i is symmetric. The constraint $\alpha e + g_2 \in \mathscr{C}_-^*$ is then equivalent to

$$\alpha \zeta \geq \|A_1(\alpha e + g_2)\|_2 + \|A_2(\alpha e + g_2)\|_2.$$

At this point, we focus on the case in which the projection $e^\top S_1$ is disjoint from $e^\top S_2$. In this case, we have the following strict inequality:

$$\max_{x \in \mathscr{B}} \langle c_1 + A_1 x, e \rangle = \max_{u \in S_1} \langle u, e \rangle < \min_{v \in S_2} \langle v, e \rangle = \min_{y \in \mathscr{B}} \langle c_2 + A_2 y, e \rangle,$$

and rearranging then gives

$$\zeta = \langle c_2 - c_1, e \rangle > \max_{x \in \mathscr{B}} \langle A_1 x, e \rangle + \max_{y \in \mathscr{B}} \langle -A_2 x, e \rangle$$
$$= \max_{x \in \mathscr{B}} \langle x, A_1^\top e \rangle + \max_{y \in \mathscr{B}} \langle x, -A_2^\top e \rangle = \|A_1 e\|_2 + \|A_2 e\|_2.$$

As such, taking

$$\alpha \geq \alpha^* := \frac{\|A_1 g_2\|_2 + \|A_2 g_2\|_2}{\zeta - \big(\|A_1 e\|_2 + \|A_2 e\|_2\big)} \tag{7}$$

produces a point $\alpha e + g_2 \in \mathscr{C}_-^*$, considering

$$\alpha \zeta \geq \alpha\big(\|A_1 e\|_2 + \|A_2 e\|_2\big) + \|A_1 g_2\|_2 + \|A_2 g_2\|_2 \geq \|A_1(\alpha e + g_2)\|_2 + \|A_2(\alpha e + g_2)\|_2,$$

where the last step follows from the triangle inequality. Note that if $g_1 \geq \alpha^*$, then we can take $\alpha = g_1$ to get $\|g - \widetilde{\Pi}(g)\|_2 = 0$. Otherwise, $\|g - \widetilde{\Pi}(g)\|_2 \leq |g_1 - \alpha^*| = \alpha^* - g_1$. Overall, we have

$$\|g - \Pi_{\mathscr{C}^*}(g)\|_2 \leq \|g - \widetilde{\Pi}(g)\|_2 \leq (\alpha^* - g_1)_+.$$

By the monotonicity of expectation, we then have

$$w_\cap \leq \mathbb{E}_g\Big[\|g - \Pi_{\mathscr{C}^*}(g)\|_2\Big] \leq \mathbb{E}_g(\alpha^* - g_1)_+ = \mathbb{E}_{g_2}\Big[\mathbb{E}_{g_1}\Big[(\alpha^* - g_1)_+ \Big| g_2\Big]\Big]. \tag{8}$$

To estimate the right-hand side, we first have

$$\mathbb{E}_{g_1}\Big[(\alpha^* - g_1)_+ \Big| g_2\Big] = \int_{-\infty}^{\infty} (\alpha^* - z)_+ d\Phi(z) = \alpha^* \Phi(\alpha^*) + \frac{1}{\sqrt{2\pi}} e^{-(\alpha^*)^2/2}, \tag{9}$$

which lies between $\alpha^*/2$ and $\alpha^* + 1/\sqrt{2\pi}$ since $\alpha \geq 0$.

Let $P_{e^\perp}$ denote the $n \times n$ orthogonal projection onto the orthogonal complement of the span of e. Appealing to Lemma 3 with $A := A_i P_{e^\perp}$ then gives

$$\mathbb{E}\|A_i g_2\|_2 = \mathbb{E}\|A_i P_{e^\perp} g\|_2 \leq \|A_i P_{e^\perp}\|_F \leq \|A_i\|_F,$$

where the last inequality follows from the fact that each row of $A_i P_{e^\perp}$ is a projection of the corresponding row in A_i and therefore has a smaller 2-norm. Considering (7), this implies

$$\mathbb{E}_{g_2}\alpha^* \leq \frac{\|A_1\|_F + \|A_2\|_F}{\zeta - \left(\|A_1 e\|_2 + \|A_2 e\|_2\right)},$$

which combined with (8) and (9) then gives

$$w_\cap \leq \mathbb{E}_{g_2}\left[\mathbb{E}_{g_1}\left[(\alpha^* - g_1)_+ \middle| g_2\right]\right] \leq \frac{\|A_1\|_F + \|A_2\|_F}{\zeta - \left(\|A_1 e\|_2 + \|A_2 e\|_2\right)} + \frac{1}{\sqrt{2\pi}}.$$

Acknowledgements The authors thank Matthew Fickus and Katya Scheinberg for insightful discussions. A. S. Bandeira was supported by AFOSR award FA9550-12-1-0317; D. G. Mixon was supported by NSF award DMS-1321779, AFOSR F4FGA06060J007, and AFOSR Young Investigator Research Program award F4FGA06088J001; and B. Recht was supported by ONR award N00014-11-1-0723 and NSF awards CCF-1139953 and CCF-11482. The views expressed in this article are those of the authors and do not reflect the official policy or position of the US Air Force, Department of Defense, or the US Government.

References

1. E. Abbe, N. Alon, A.S. Bandeira, Linear Boolean classification, coding and "the critical problem". arXiv:1401.6528 [cs.IT]
2. D. Amelunxen, M. Lotz, M.B. McCoy, J.A. Tropp, Living on the edge: a geometric theory of phase transitions in convex optimization. arXiv:1303.6672
3. K. Aryafar, S. Jafarpour, A. Shokoufandeh, Music genre classification using sparsity-eager support vector machines. Technical report, Drexel University (2012)
4. P.T. Boufounos, R.G. Baraniuk, 1-bit compressive sensing, in *42nd Annual Conference on Information Sciences and Systems, CISS* (2008), pp. 16–21
5. V. Chandrasekaran, B. Recht, P.A. Parrilo, A.S. Willsky, The convex geometry of linear inverse problems. Found. Comput. Math. **12**, 805–849 (2012)
6. S. Dasgupta, Learning mixtures of Gaussians, in *40th Annual IEEE Symposium on Foundations of Computer Science* (1999), pp. 634–644.
7. S. Dasgupta, Experiments with random projection, in *Proceedings of Unvertainty in Artificial Intelligence* (2000)
8. M.A. Davenport, M.F. Duarte, M.B. Wakin, J.N. Laska, D. Takhar, K.F. Kelly, R.G. Baraniuk, The smashed filter for compressive classification and target recognition, in *Proceedings of SPIE* (2007)

9. M.F. Duarte, M.A. Davenport, M.B. Wakin, J.N. Laska, D. Takhar, K.F. Kelly, R.G. Baraniuk, Multiscale random projections for compressive classification, in *IEEE International Conference on Image Processing (ICIP)*, pp. VI-161–164 (2007)
10. Y. Gordon, On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$, in *Geometric Aspects of Functional Analysis, Israel Seminar 1986–87*. Lecture Notes in Mathematics, vol. 1317 (1988), pp. 84–106
11. A. Gupta, R. Nowak, B. Recht, Sample complexity for 1-bit compressed sensing and sparse classification, in *IEEE International Symposium on Information Theory Proceedings (ISIT)* (2010), pp. 1553–1557
12. J. Haupt, R. Castro, R. Nowak, G. Fudge, A. Yeh, Compressive sampling for signal classification, in *40th Asilomar Conference on Signals, Systems and Computers* (2006)
13. Hyperspectral Remote Sensing Scenes, http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes
14. P. Indyk, R. Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality, in *Proceedings of 30th Symposium on Theory of Computing* (ACM, New York, 1998), pp. 604–613
15. R.J. Johnson, Improved feature extraction, feature selection, and identification techniques that create a fast unsupervised hyperspectral target detection algorithm, Master's Thesis, Air Force Institute of Technology, 2008
16. J.M. Kleinberg, Two algorithms for nearest-neighbor search in high dimensions, in *Proceedings of the 29th Annual ACM Symposium on Theory of Computing STOC* (1997), pp. 599–608
17. M. Ledoux, M. Talagrand, *Probability in Banach Spaces* (Springer, Berlin, 1991)
18. A. Majumdar, R.K. Ward, Robust classifiers for data reduced via random projections. IEEE Trans. Syst. Man Cybern. B Cybern. **40**, 1359–1371 (2010)
19. G. Paouris, Concentration of mass on convex bodies. Geom. Funct. Anal. **16**, 1021–1049 (2006)
20. Y. Plan, R. Vershynin, One-bit compressed sensing by linear programming. Commun. Pure Appl. Math. **66**, 1275–1297 (2013)
21. H. Reboredo, F. Renna, R. Calderbank, M.R.D. Rodrigues, Compressive classification of a mixture of Gaussians: analysis, designs and geometrical interpretation. arXiv:1401.6962
22. S.A. Robila, L. Maciak, New approaches for feature extraction in hyperspectral imagery, in *Proceedings of IEEE Long Island Systems, Applications and Technology Conference (IEEE LISAT)* (2006), pp. 1–7

Weak Phase Retrieval

Sara Botelho-Andrade, Peter G. Casazza, Dorsa Ghoreishi, Shani Jose, and Janet C. Tremain

Abstract Phase retrieval and phaseless reconstruction for Hilbert space frames is a very active area of research. Recently, it was shown that these concepts are equivalent. In this paper, we make a detailed study of a weakening of these concepts to weak phase retrieval and weak phaseless reconstruction. We will give several necessary and/or sufficient conditions for frames to have these weak properties. We will prove three surprising results: (1) Weak phaseless reconstruction is equivalent to phaseless reconstruction. That is, it never was *weak*; (2) weak phase retrieval is not equivalent to weak phaseless reconstruction; (3) weak phase retrieval requires at least $2m - 2$ vectors in an m-dimensional Hilbert space. We also give several examples illustrating the relationship between these concepts.

Keywords Phase retrieval · Weak phase retrieval · Norm retrieval · Spark · Complement property

1 Introduction

The problem of retrieving the phase of a signal, given a set of intensity measurements, has been studied by engineers for many years. Signals passing through linear systems often result in lost or distorted phase information. This partial loss of phase information occurs in various applications including speech recognition [4, 14, 15] and optics applications such as X-ray crystallography [3, 11, 12]. The concept of *phase retrieval* for Hilbert space frames was introduced in 2006 by Balan, Casazza, and Edidin [1], and since then it has become an active area of research [2, 5, 9, 13, 16]. Phase retrieval deals with recovering the phase of a signal given intensity measurements from a redundant linear system. In phaseless reconstruction the

S. Botelho-Andrade · P.G. Casazza (✉) · D. Ghoreishi · S. Jose · J.C. Tremain
Department of Mathematics, University of Missouri, Columbia, MO 65211, USA
e-mail: sandrade102087@gmail.com; casazzap@missouri.edu; dorsa.ghoreishi@gmail.com; shanijose@gmail.com; tremainjc@missouri.edu

H. Boche et al. (eds.), *Compressed Sensing and its Applications*,
Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-69802-1_7

unknown signal itself is reconstructed from these measurements. In recent literature, the two terms were used interchangeably. However, it is not obvious from the definitions that the two are equivalent. Recently, authors in [6] proved that phase retrieval is equivalent to phaseless reconstruction in both the real and complex case.

Phase retrieval has been defined for vectors as well as for projections. *Phase retrieval by projections* occur in real-life problems, such as crystal twinning [10], where the signal is projected onto some higher-dimensional subspaces and has to be recovered from the norms of the projections of the vectors onto the subspaces. We refer the reader to [8] for a detailed study of phase retrieval by projections. At times these projections are identified with their target spaces. Determining when subspaces $\{W_i\}_{i=1}^n$ and $\{W_i^\perp\}_{i=1}^n$ both do phase retrieval has given way to the notion of *norm retrieval* [7], another important area of research.

While investigating the relationship between phase retrieval and phaseless reconstruction, in [6] it was noted that if two vectors have the same phase, then they will be zero in the same coordinates. This gave way to a weakening of phase retrieval, known as weak phase retrieval. In this work, we study the weakened notions of phase retrieval and phaseless reconstruction. One limitation of current methods used for retrieving the phase of a signal is computing power. Recall that a generic family of $(2m-1)$-vectors in $\mathbb{R}^m$ does phaseless reconstruction; however, no set of $(2m-2)$-vectors can (see [1] for details). By generic we are referring to an open dense set in the set of $(2m-1)$-element frames in $\mathbb{H}^m$. We started with the motivation that weak phase retrieval could be done with $m+1$ vectors in $\mathbb{R}^m$. However, it will be shown that the cardinality condition can only be relaxed to $2m-2$. Nevertheless, the results we obtain in this work are interesting in their own right and contribute to the overall understanding of phase retrieval. We provide illustrative examples in the real and complex cases for weak phase retrieval.

The rest of the paper is organized as follows: In Section 2, we give basic definitions and certain preliminary results to be used in the paper. Weak phase retrieval is defined in Section 3. Characterizations are given in both real and complex case. Also, the minimum number of vectors needed for weak phase retrieval is obtained. In Section 4, we define weak phaseless reconstruction and prove that it is equivalent to phase retrieval in the real case. We conclude by providing certain illustrative examples in Section 5.

2 Preliminaries

In this section, we introduce some of the basic definitions and results from frame theory. Throughout this paper, $\mathbb{H}^m$ denotes an m-dimensional real or complex Hilbert space, and we will write $\mathbb{R}^m$ or $\mathbb{C}^m$ when it is necessary to differentiate between the two. We start with the definition of a frame in $\mathbb{H}^m$.

Definition 1. A family of vectors $\Phi = \{\phi_i\}_{i=1}^n$ in $\mathbb{H}^m$ is a **frame** if there are constants $0 < A \leq B < \infty$ so that for all $x \in \mathbb{H}^m$

$$A\|x\|^2 \leq \sum_{i=1}^{n} |\langle x, \phi_i \rangle|^2 \leq B\|x\|^2,$$

where A and B are the **lower and upper frame bounds** of the frame, respectively. The frame is called an **A-tight frame** if $A = B$ and is a **Parseval frame** if $A = B = 1$.

In addition, Φ is called an **equal norm frame** if $\|\phi_i\| = \|\phi_j\|$ for all i, j and is called a **unit norm frame** if $\|\phi_i\| = 1$ for all $i = 1, 2, \ldots n$.

Next, we give the formal definitions of phase retrieval, phaseless reconstruction, and norm retrieval. Note that, here, phase of vector $x = re^{it}$ is taken as e^{it}.

Definition 2. Let $\Phi = \{\phi_i\}_{i=1}^n \in \mathbb{H}^m$ be such that for $x, y \in \mathbb{H}^m$

$$|\langle x, \phi_i \rangle| = |\langle y, \phi_i \rangle|, \text{ for all } i = 1, 2, \ldots, n.$$

Φ yields

(i) **phase retrieval** with respect to an orthonormal basis $\{e_i\}_{i=1}^m$ if there is a $|\theta| = 1$ such that x and θy have the same phase. That is, $x_i = \theta y_i$, for all $i = 1, 2, \ldots, m$, where $x_i = \langle x, e_i \rangle$.
(ii) **phaseless reconstruction** if there is a $|\theta| = 1$ such that $x = \theta y$.
(iii) **norm retrieval** if $\|x\| = \|y\|$.

We note that tight frames $\{\phi_i\}_{i=1}^m$ for $\mathbb{H}^m$ do norm retrieval. Indeed, if

$$|\langle x, \phi_i \rangle| = |\langle y, \phi_i \rangle|, \text{ for all } i = 1, 2, \ldots, m,$$

then

$$A\|x\|^2 = \sum_{i=1}^{m} |\langle x, \phi_i \rangle|^2 = \sum_{i=1}^{m} |\langle y, \phi_i \rangle|^2 = A\|y\|^2.$$

Phase retrieval in $\mathbb{R}^m$ is classified in terms of a fundamental result called the complement property, which we define below:

Definition 3 ([1]). A frame $\Phi = \{\phi_i\}_{i=1}^n$ in $\mathbb{H}^m$ satisfies the **complement property** if for all subsets $I \subset \{1, 2, \ldots, n\}$, either $\text{span}\{\phi_i\}_{i \in I} = \mathbb{H}^m$ or $\text{span}\{\phi_i\}_{i \in I^c} = \mathbb{H}^m$.

A fundamental result from [1] is:

Theorem 1 ([1]). *If Φ does phaseless reconstruction, then it has complement property. In $\mathbb{R}^m$, if Φ has complement property, then it does phase retrieval.*

It follows that if $\Phi = \{\phi_i\}_{i=1}^n$ does phase retrieval in $\mathbb{R}^m$, then $n \geq 2m - 1$.

Full spark is another important notion of vectors in frame theory. A formal definition is given below:

Definition 4. Given a family of vectors $\Phi = \{\phi_i\}_{i=1}^n$ in $\mathbb{H}^m$, the **spark** of Φ is defined as the cardinality of the smallest linearly dependent subset of Φ. When

spark$(\Phi) = m + 1$, every subset of size m is linearly independent, and in that case, Φ is said to be **full spark**.

We note that from the definitions it follows that full spark frames with $n \geq 2m-1$ have the complement property and hence do phaseless reconstruction. Moreover, if $n = 2m - 1$, then the complement property clearly implies full spark.

3 Weak Phase Retrieval

In this section, we define the notion of weak phase retrieval and make a detailed study of it. We obtain the minimum number of vectors required to do weak phase retrieval. First we define the notion of vectors having weakly the same phase.

Definition 5. Two vectors in $\mathbb{H}^m$, $x = (a_1, a_2, \ldots, a_m)$ and $y = (b_1, b_2, \ldots, b_m)$ **weakly have the same phase** if there is a $|\theta| = 1$ so that

$$\text{phase}(a_i) = \theta \text{phase}(b_i), \text{ for all } i = 1, 2, \ldots, m, \text{ for which } a_i \neq 0 \neq b_i.$$

In the real case, if $\theta = 1$ we say x, y **weakly have the same signs**, and if $\theta = -1$ they **weakly have opposite signs.**

In the definition above, note that we are only comparing the phase of x and y for entries where both are non-zero. Hence, two vectors may *weakly* have the same phase but not have the same phase in the usual sense. We define weak phase retrieval formally as follows:

Definition 6. If for any $x = (a_1, a_2, \ldots, a_m)$ and $y = (b_1, b_2, \ldots, b_m)$ in $\mathbb{H}^m$, the equality

$$|\langle x, \phi_i \rangle| = |\langle y, \phi_i \rangle|, \text{ for all } i = 1, 2, \ldots, m,$$

implies that x, y weakly have the same phase, then the family of vectors $\{\phi_i\}_{i=1}^n$ in $\mathbb{H}^m$ does weak phase retrieval.

Observe that the difference between phase retrieval and weak phase retrieval is that in the latter it is possible for $a_i = 0$ but $b_i \neq 0$.

3.1 Real Case

Now we begin our study of weak phase retrieval in $\mathbb{R}^m$. The following proposition provides a useful criteria for determining when two vectors have weakly the same or opposite phases. In what follows, we use $[m]$ to denote the set $\{1, 2, \ldots, m\}$.

Proposition 1. *Let $x = (a_1, a_2, \ldots, a_m)$ and $y = (b_1, b_2, \ldots, b_m)$ in $\mathbb{R}^m$. The following are equivalent:*

1. *We have*

$$sgn\,(a_i a_j) = sgn\,(b_i b_j), \text{ for all } a_i a_j \neq 0 \neq b_i b_j.$$

2. Either x, y have weakly the same signs or they have weakly opposite signs.

Proof. (1) $\Rightarrow$ (2): Let

$$I = \{1 \leq i \leq m : a_i = 0\} \text{ and } J = \{1 \leq i \leq n : b_i = 0\}.$$

Let

$$K = [m] \setminus (I \cup J).$$

So $i \in K$ if and only if $a_i \neq 0 \neq b_i$. Let $i_0 = min\ K$. We examine two cases:

Case 1: $sgn\ a_{i_0} = sgn\ b_{i_0}$.
For any $i_0 \neq k \in K$, $sgn\ (a_{i_0}a_k) = sgn\ (b_{i_0}b_k)$ implies $sgn\ a_k = sgn\ b_k$. Since all other coordinates of either x or y are zero, it follows that x, y weakly have the same signs.

Case 2: $sgn\ a_{i_0} = -sgn\ b_{i_0}$.
For any $i_o \neq k \in K$, $a_{i_0}a_k = b_{i_0}b_k$ implies $sgn\ a_k = -sgn\ b_k$. Again, since all other coordinates of either x or y are zero, it follows that x, y weakly have opposite signs.

(2) $\Rightarrow$ (1): This is immediate.

The next lemma will be useful in the following proofs as it gives a criteria for showing when vectors do not weakly have the same phase.

Lemma 1. *Let $x = (a_1, a_2, \ldots, a_m)$ and $y = (b_1, b_2, \ldots, b_m)$ be vectors in $\mathbb{R}^m$. If there exists $i \in [m]$ such that $a_i b_i \neq 0$ and $\langle x, y\rangle = 0$, then x and y do not have weakly the same or opposite signs.*

Proof. We proceed by way of contradiction. If x and y weakly have the same signs, then $a_j b_j \geq 0$ for all $j \in [m]$, and in particular we arrive at the following contradiction:

$$\langle x, y\rangle = \sum_{j=1}^{n} a_j b_j \geq a_i b_i > 0.$$

If x and y weakly have opposite signs, then $a_j b_j \leq 0$ for all $j \in [m]$, and by reversing the inequalities in the expression above, we get the desired result.

The following result relates weak phase retrieval and phase retrieval. Recall that in the real case, it is known that phase retrieval, phaseless reconstruction, and the complement property are equivalent [1, 6].

Corollary 1. *Suppose $\Phi = \{\phi_i\}_{i=1}^n \in \mathbb{R}^m$ does weak phase retrieval but fails complement property. Then there exist two vectors $v, w \in \mathbb{R}^m$ such that $v \perp w$ and*

$$|\langle v, \phi_i\rangle| = |\langle w, \phi_i\rangle| \text{ for all } i. \tag{1}$$

Further, v and w are disjointly supported.

Proof. By the assumption, $\Phi = \{\phi_i\}_{i=1}^n$ fails complement property, so there exists $I \subset [n]$, s.t. $A = Span\{\phi_i\}_{i\in I} \neq \mathbb{R}^m$ and $B = Span\{\phi_i\}_{i\in I^c} \neq \mathbb{R}^m$. Choose $\|x\| = \|y\| = 1$ such that $x \perp A$ and $y \perp B$. Then

$$|\langle x + y, \phi_i\rangle| = |\langle x - y, \phi_i\rangle| \text{ for all i=1, 2, ..., n.}$$

Let $w = x + y$ and $v = x - y$. Then $v \perp w$. Observe

$$\langle w, v\rangle = \langle x + y, x - y\rangle = \|x\|^2 + \langle y, x\rangle - \langle x, y\rangle - \|y\|^2 = 0.$$

Moreover, the assumption that Φ does weak phase retrieval implies v and w have weakly the same or opposite phases. Then it follows from Lemma 1 that $v_i w_i = 0$ for all $i = 1, 2, \ldots, m$, and so v and w are disjointly supported.

Example 1. In $\mathbb{R}^2$ let $\phi_1 = (1, 1)$ and $\phi_2 = (1, -1)$. These vectors clearly fail complement property. But if $x = (a_1, a_2)$, $y = (b_1, b_2)$, and we have

$$|\langle x, \phi_i\rangle| = |\langle y, \phi_i\rangle|, \text{ for } i = 1, 2,$$

then

$$|a_1 + a_2|^2 = |b_1 + b_2|^2 \text{ and } |a_1 - a_2|^2 = |b_1 - b_2|^2.$$

By squaring these out and subtracting the result, we get:

$$4a_1a_2 = 4b_1b_2.$$

Hence, either x, y have the same signs or opposite signs. That is, these vectors do weak phase retrieval.

With some particular assumptions, the following proposition gives the specific form of vectors which do weak phase retrieval but not phase retrieval.

Proposition 2. *Let $\Phi = \{\phi_i\}_{i=1}^n \in \mathbb{R}^m$ be such that Φ does weak phase retrieval but fails complement property. Let $x = (a_1, a_2, \ldots, a_m)$, $y = (b_1, b_2, \ldots b_m) \in \mathbb{R}^m$ such that $x + y \perp x - y$ and satisfy equation (1). If $a_i b_i \neq 0$, $a_j b_j \neq 0$ for some i, j, and all other coordinates of x and y are zero, then*

$$|a_i| = |b_i|, \text{ for } i = 1, 2.$$

Proof. Without loss of generality, take $x = (a_1, a_2, 0, \ldots, 0)$ and $y = (b_1, b_2, 0, \ldots, 0)$. Observe that both $x + y$ and $x - y$ either weakly have the same phase or weakly have the opposite phase. Thus, by Corollary 1, $x+y$ and $x-y$ have disjoint support as these vectors are orthogonal. Also,

$$x + y = (a_1 + b_1, a_2 + b_2, 0, \ldots, 0) \text{ and } x - y = (a_1 - b_1, a_2 - b_2, 0, \ldots, 0).$$

Since $x + y$ and $x - y$ are disjointly supported, it reduces to the cases where either $a_1 = \pm b_1$ and $a_2 = \mp b_2$. In both cases, it follows from equation (1) that $|a_i| = |b_i|$ for all $i = 1, 2, \ldots, m$.

The next theorem gives the main result about the minimum number of vectors required to do weak phase retrieval in $\mathbb{R}^m$. Recall that phase retrieval requires $n \geq 2m - 1$ vectors.

Theorem 2. *If $\{\phi_i\}_{i=1}^n$ does weak phase retrieval on $\mathbb{R}^m$, then $n \geq 2m - 2$.*

Proof. For a contradiction assume $n \leq 2m - 3$ and choose $I \subset [n]$ with $I = [m - 2]$. Then $|I| = m - 2$ and $|I^c| \leq m - 1$. For this partition of $[n]$, let $x + y$ and $x - y$ be as in the proof of Corollary 1. Then $x + y$ and $x - y$ must be disjointly supported which follows from the Corollary 1. Therefore, for each $i = 1, 2, \ldots, m$, $a_i = \epsilon_i b_i$, where $\epsilon_i = \pm 1$ for each i and a_i, b_i are the i^{th} coordinates of x and y, respectively. Observe the conclusion holds for a fixed x and any $y \in (\text{span}\{\phi\}_{i\in I})^{\perp}$ and $\dim\ (\text{span}\{\phi_i\}_{i\in I})^{\perp} \geq 2$. However, this poses a contradiction since there are infinitely many distinct choices of y in this space, while our argument shows that there are at most 2^m possibilities for y.

Contrary to the initial hopes, the previous result shows that the minimal number of vectors doing weak phase retrieval is only one less than the number of vectors doing phase retrieval. However, it is interesting to note that a minimal set of vectors doing weak phase retrieval is necessarily full spark, as is true for the minimal number of vectors doing phase retrieval, as the next result shows.

Theorem 3. *If $\Phi = \{\phi_i\}_{i=1}^{2n-2}$ does weak phase retrieval in $\mathbb{R}^n$, then Φ is full spark.*

Proof. We proceed by way of contradiction. Assume Φ is not full spark. Then there exists $I \subset \{1, 2, \ldots, 2n-2\}$ with $|I| = n$ such that $\dim \text{span}\{\phi_i\}_{i\in I} \leq n-1$. Observe that the choice of I above implies $|I^c| = n - 2$. Now we arrive at a contradiction by applying the same argument used in (the proof of) Theorem 2.

It is important to note that the converse of Theorem 3 does not hold. For example, the canonical basis in $\mathbb{R}^2$ is trivially full spark but does not do weak phase retrieval.

If Φ is as in Theorem 3, then the following corollary guarantees it is possible to add a vector to this set and obtain a collection which does phaseless reconstruction.

Corollary 2. *If Φ is as in Theorem 3, then there exists a dense set of vectors F in $\mathbb{R}^n$ such that $\{\psi\} \cup \Phi$ does phaseless reconstruction for any $\psi \in \mathsf{F}$.*

Proof. We observe that the set of $\psi \in \mathbb{R}^n$ such that $\Phi \cup \{\psi\}$ is full spark is dense in $\mathbb{R}^n$. To see this let $\mathsf{G} = \bigcup_{\substack{I\subset[2n-2]\\|I|=n-1}} \text{span}\{\phi_i\}_{i\in I}$. Then G is the finite union of hyperplanes, so G^c is dense and $\{\psi\} \cup \Phi$ is full spark for any $\psi \in \mathsf{G}^c$. To verify that this collection of vectors is full spark. Note that either a subcollection of m-vectors is contained in Φ, then it spans $\mathbb{R}^n$, or the subcollection contains the vector ψ. In this case, denote $I \subset [2n - 2]$ with $|I| = n - 1$ and suppose $\sum_{i\in I} a_i\phi_i + a\psi = 0$. Therefore, $a\psi = -\sum_{i\in I} a_i\phi_i$ and if $a \neq 0$ then $a\psi \in \text{span}\{\phi_i\}_{i\in I}$, a contradiction. It follows $a = 0$ and since Φ is full spark (see Theorem 3), in particular, $\{\phi_i\}_{i\in I}$ are linearly independent; it follows that $a_i = 0$ for all $i \in I$.

3.2 Complex Case

An extension of Proposition 1 in the complex case is given below:

Proposition 3. *Let* $x = (a_1, a_2, \ldots, a_m)$ *and* $y = (b_1, b_2, \ldots, b_m)$ *in* $\mathbb{C}^m$. *The following are equivalent:*

1. *If there is a* $|\theta| = 1$ *such that phase* $(a_i) = \theta$*phase* (b_i), *for some i, then phase* $(a_i a_j) = \theta^2$*phase* $(b_i b_j)$, $i \neq j$ *and* $a_j \neq 0 \neq b_j$.
2. *x and y weakly have the same phase.*

Proof. (1) $\Rightarrow$ (2): Let the index sets I, J, and K be as in Proposition 1. By (1), there is a $|\theta| = 1$ such that *phase* $(a_i) = \theta$*phase* (b_i) for some $i \in K$.

Now, for any $j \in K$, $j \neq i$,

$$\textit{phase}\ (a_i a_j) = \textit{phase}\ (a_i)\ \textit{phase}\ (a_j) = \theta\textit{phase}\ (b_i)\ \textit{phase}\ (a_j).$$

But *phase* $(a_i a_j) = \theta^2$ *phase* $(b_i b_j) = \theta^2$ *phase* (b_i)*phase* (b_j). Thus, it follows that *phase* $(a_j) = \theta$ *phase* (b_j). Since all other coordinates of either x or y are zero, it follows that x, y weakly have the same phase.

(2) $\Rightarrow$ (1): By definition, there is a $|\theta| = 1$ such that *phase* $(a_i) = \theta$ *phase* (b_i) for all $a_i \neq 0 \neq b_i$. Now, (1) follows immediately as *phase* $(a_i a_j) =$ *phase* (a_i)*phase* (a_j).

4 Weak Phaseless Reconstruction

In this section, we define weak phaseless reconstruction and study its characterizations. A formal definition is given below:

Definition 7. A family of vectors $\{\phi_i\}_{i=1}^n$ in $\mathbb{H}^m$ does **weak phaseless reconstruction** if for any $x = (a_1, a_2, \ldots, a_m)$ and $y = (b_1, b_2, \ldots, b_m)$ in $\mathbb{H}^m$, with

$$|\langle x, \phi_i\rangle| = |\langle y, \phi_i\rangle|, \text{ for all } i = 1, 2, \ldots, n, \tag{2}$$

there is a $|\theta| = 1$ so that

$$a_i = \theta b_i, \text{ for all } i = 1, 2, \ldots, m, \text{ for which } a_i \neq 0 \neq b_i.$$

In particular, $\{\phi_i\}$ does phaseless reconstruction for vectors having all non-zero coordinates.

Note that if $\Phi = \{\phi_i\}_{i=1}^n \in \mathbb{R}^m$ does weak phaseless reconstruction, then it does weak phase retrieval. The converse is not true in general. Let $x = (a_1, a_2, \ldots, a_m)$ and $y = (b_1, b_2, \ldots, b_m)$. If $\Phi = \{\phi_i\}_{i=1}^n \in \mathbb{R}^m$ does weak phase retrieval and $|\{i | a_i b_i \neq 0\}| = 2$, then Φ may not do weak phaseless reconstruction. If $a_i b_i = a_j b_j$ where $a_i b_i \neq 0$ and $a_j b_j \neq 0$, then we certainly cannot conclude in general that $|a_i| = |b_i|$ (see Example 2).

Theorem 4. *Let* $x = (a_1, a_2, \ldots, a_m)$ *and* $y = (b_1, b_2, \ldots, b_m)$ *in* $\mathbb{R}^m$. *The following are equivalent:*

I. There is a $\theta = \pm 1$ *so that*

$$a_i = \theta b_i, \text{ for all } a_i \neq 0 \neq b_i.$$

II. We have $a_i a_j = b_i b_j$ *for all* $1 \leq i, j \leq m$, *and* $|a_i| = |b_i|$ *for all* i *such that* $a_i \neq 0 \neq b_i$.

III. The following hold:

A. Either x, y *have weakly the same signs or they have weakly the opposite signs.*

B. One of the following holds:

(i) There is a $1 \leq i \leq m$ *so that* $a_i = 0$ *and* $b_j = 0$ *for all* $j \neq i$.

(ii) There is a $1 \leq i \leq m$ *so that* $b_i = 0$ *and* $a_j = 0$ *for all* $j \neq i$.

(iii) If (i) and (ii) fail and $I = \{1 \leq i \leq m : a_i \neq 0 \neq b_i\}$, *then the following hold:*

(a) If $i \in I^c$ *then* $a_i = 0$ *or* $b_i = 0$.

(b) For all $i \in I$, $|a_i| = |b_i|$.

Proof. $(I) \Rightarrow (II)$: By (I) $a_i = \theta b_i$ for all i such that both are non-zero, so $a_i a_j = (\theta b_i)(\theta b_j)$ and so $a_i a_j = \theta^2 b_i b_j$. Since $\theta = \pm 1$ it follows that $a_i a_j = b_i b_j$ for all i, j (that are non-zero). The second part is trivial.

$(II) \Rightarrow (III)$:

(A) This follows from Proposition 1.

(B) (i) Assume $a_i = 0$ but $b_i \neq 0$. Then for all $j \neq i$, we have $a_i a_j = 0 = b_i b_j$ and so $b_j = 0$.

(ii) This is symmetric to (i).

(iii) If (i) and (ii) fail, then by definition, for any i, either both a_i and b_i are zero or they are both non-zero, which proves (A). (B) is immediate.

$(III) \Rightarrow (I)$: The existence of θ is clear by part A. In part B, (i) and (ii) trivially imply (I). Assume (iii); then for each i such that $a_i \neq 0 \neq b_i$ and $|a_i| = |b_i|$, then $a_i = \pm b_i$.

Corollary 3. *Let* Φ *be a frame for* $\mathbb{R}^m$. *The following are equivalent:*

1. Φ *does weak phaseless reconstruction.*
2. *For any* $x = (a_1, a_2, \ldots, a_m)$ *and* $y = (b_1, b_2, \ldots, b_m)$ *in* $\mathbb{R}^m$, *if*

$$|\langle x, \phi_i \rangle| = |\langle y, \phi_i \rangle| \text{ for all } i,$$

then each of the equivalent conditions in Theorem 4 holds.

The following theorems provide conditions under which weak phase retrieval is equivalent to weak phaseless reconstruction.

Proposition 4. *Let $\Phi = \{\phi_i\}_{i=1}^n$ do weak phase retrieval on vectors $x = (a_1, a_2, \ldots, a_m)$ and $y = (b_1, b_2, \ldots, b_m)$ in $\mathbb{H}^m$. If $|I| = |\{i : a_i b_i \neq 0\}| \geq 3$ and $a_i a_j = b_i b_j$ for all $i, j \in I$, then Φ does weak phaseless reconstruction.*

Proof. If i, j, k are three members of I with $a_i a_j = b_i b_j$, $a_i a_k = b_i b_k$, and $a_k a_j = b_k b_j$, then a short calculation gives $a_i^2 a_j a_k = b_i^2 b_j b_k$ and hence $|a_i| = |b_i|$. This computation holds for each $i \in I$, and since Φ does phase retrieval, there is a $|\theta| = 1$ so that *phase* $a_i = \theta$ *phase* b_i for all i. It follows that $a_i = \theta\, b_i$ for all $i = 1, 2, \ldots, m$.

It turns out that whenever a frame contains the canonical basis, then weak phase retrieval and phaseless reconstruction are the same.

Proposition 5. *Let the frame $\Phi = \{\phi_i\}_{i=1}^n \in \mathbb{R}^m$ does weak phase retrieval. If Φ contains the canonical basis, then Φ does phaseless reconstruction.*

Proof. Let $x = (a_1, a_2, \ldots, a_m)$ and $y = (b_1, b_2, \ldots, b_m) \in \mathbb{R}^m$. By definition of weak phase retrieval, Φ satisfies the equation 2. In particular, for $\phi_i = e_i$, the equation 2 implies that $|a_i| = |b_i|$, $\forall i = 1, 2, \ldots, m$, hence the theorem.

We conclude this section by showing the surprising result that weak phaseless reconstruction is same as phaseless reconstruction in $\mathbb{R}^m$, i.e., it is not really **weak**.

Theorem 5. *Frames which do weak phaseless reconstruction in $\mathbb{R}^m$ do phaseless reconstruction.*

Proof. For a contradiction assume $\Phi = \{\phi\}_{i=1}^n \subset \mathbb{R}^m$ does weak phaseless reconstruction but fails the complement property. Then there exists $I \subset [n]$ such that $\text{Span}_{i\in I}\, \phi_i \neq \mathbb{R}^m$ and $\text{Span}_{i\in I^c}\, \phi_i \neq \mathbb{R}^m$. Pick non-zero vectors $x, y \in \mathbb{R}^m$ such that $x \perp \text{Span}_{i\in I}\, \phi_i \neq \mathbb{R}^m$ and $y \perp \text{Span}_{i\in I^c}\, \phi_i \neq \mathbb{R}^M$. Then for any $c \neq 0$, we have

$$|\langle x + cy, \phi_i\rangle| = |\langle x - cy, \phi_i\rangle| \quad \text{for all } i \in [n].$$

Now we consider the following cases where x_i and y_i denotes the i^{th} coordinate of the vectors x and y.

Case 1: $\{i : x_i \neq 0\} \cap \{i : y_i \neq 0\} = \emptyset$

Set $c = 1$ and observe since $x \neq 0$ there exists some $i \in [n]$ such that $x_i \neq 0$ and $y_i = 0$ and similarly there exists $j \in [n]$ such that $y_j \neq 0$ but $x_j = 0$. Then $x + y$ and $x - y$ have the same sign in the i^{th}-coordinate but opposite signs in the j^{th} coordinate; this contradicts the assumption that Φ does weak phaseless reconstruction.

Case 2: There exists $i, j \subset [n]$ such that $x_i y_i \neq 0$ and $x_j = 0$, $y_j \neq 0$.

Without loss of generality, we may assume $x_i y_i > 0$; otherwise consider $-x$ or $-y$. If $0 < c \leq \frac{x_i}{y_i}$, then the i^{th} coordinates of $x + cy$ and $x - cy$ have the same sign, whereas the j^{th} coordinates have opposite signs which contradicts the assumption. By considering $y + cx$ and $y - cx$, this argument holds in the case that $y_j = 0$ and $x_j \neq 0$.

Case 3: $x_i = 0$ if and only if $y_i = 0$.

By choosing c small enough, we have that $x_i + cy_i \neq 0$ if and only if $x_i - cy_i \neq 0$. By weak phase retrieval, there is a $|d| = 1$ so that $x_i + cy_i = d(x_i - cy_i)$. But this forces either $x_i \neq 0$ or $y_i \neq 0$ but not both which contradicts the assumption for case 3.

It is known [7] that if $\Phi = \{\phi_i\}_{i=1}^n$ does phase retrieval or phaseless reconstruction in $\mathbb{H}^m$ and T is an invertible operator on $\mathbb{H}^m$, then $\{T\phi_i\}_{i=1}^n$ does phase retrieval. It now follows that the same result holds for weak phaseless reconstruction. However, this result does not hold for weak phase retrieval. Indeed, if $\phi_1 = (1, 1)$ and $\phi_2 = (1, -1)$, then we have seen that this frame does weak phase retrieval in $\mathbb{R}^2$. But the invertible operator $T(\phi_1) = (1, 0),\ T(\phi_2) = (0, 1)$ maps this frame to a frame which fails weak phase retrieval.

5 Illustrative Examples

In this section, we provide examples of frames that do weak phase retrieval in $\mathbb{R}^3$ and $\mathbb{R}^4$. As seen earlier, the vectors $(1, 1)$ and $(1, -1)$ do weak phase retrieval in $\mathbb{R}^2$ but fail phase retrieval.

Our first example is a frame which does weak phase retrieval but fails weak phaseless reconstruction.

Example 2. We work with the row vectors of

$$\Phi = \left[\begin{array}{c|ccc} \phi_1 & 1 & 1 & 1 \\ \phi_2 & -1 & 1 & 1 \\ \phi_3 & 1 & -1 & 1 \\ \phi_4 & 1 & 1 & -1 \end{array}\right]$$

Observe that the rows of this matrix form an equal norm tight frame Φ (and hence do norm retrieval). If $x = (a_1, a_2, a_3)$ the following is the coefficient matrix where the row E_i represents the coefficients obtained from the expansion $|\langle x, \phi_i\rangle|^2$

$$1/2 \left[\begin{array}{c|cccc} & a_1a_2 & a_1a_3 & a_2a_3 & \sum_{i=1}^3 a_i^2 \\ E_1 & 1 & 1 & 1 & 1/2 \\ E_2 & -1 & -1 & 1 & 1/2 \\ E_3 & -1 & 1 & -1 & 1/2 \\ E_4 & 1 & -1 & -1 & 1/2 \end{array}\right]$$

Then the following row operations give

$$1/2\left[\begin{array}{c|cccc} & a_1a_2 & a_1a_3 & a_2a_3 & \sum_{i=1}^3 a_i^2 \\ F_1 = E_1 - E_2 & 1 & 1 & 0 & 0 \\ F_2 = E_3 - E_4 & -1 & 1 & 0 & 0 \\ F_3 = E_1 - E_3 & 1 & 0 & 1 & 0 \\ F_4 = E_2 - E_4 & -1 & 0 & 1 & 0 \\ F_4 = E_1 - E_4 & 0 & 1 & 1 & 0 \\ F_5 = E_2 - E_3 & 0 & -1 & 1 & 0 \end{array}\right]$$

$$1/2\left[\begin{array}{c|cccc} & a_1a_2 & a_1a_3 & a_2a_3 & \sum_{i=1}^3 a_i^2 \\ F_1 - F_2 & 1 & 0 & 0 & 0 \\ F_3 + F_4 & 0 & 0 & 1 & 0 \\ F_5 - F_6 & 0 & 1 & 0 & 0 \end{array}\right]$$

Therefore, we have demonstrated a procedure to identify a_ia_j for all $1 \leq i \neq j \leq 3$. This shows that given $y = (b_1, b_2, b_3)$ satisfying $|\langle x, \phi_i\rangle|^2 = |\langle y, \phi_i\rangle|^2$, then by the procedure outlined above, we obtain

$$a_ia_j = b_ib_j, \text{ for all } 1 \leq i \neq j \leq 3.$$

By Proposition 1, these four vectors do weak sign retrieval in $\mathbb{R}^3$. However, this family fails to do weak phaseless reconstruction. Observe the vectors $x = (1, 2, 0)$ and $y = (2, 1, 0)$ satisfy $|\langle x, \phi_i\rangle| = |\langle y, \phi_i\rangle|$ however do not have the same absolute value in each coordinate.

Our next example is a frame which does weak phase retrieval but fails phaseless reconstruction.

Example 3. We provide a set of six vectors in $\mathbb{R}^4$ which does weak phase retrieval in $\mathbb{R}^4$. In this case our vectors are the rows of the matrix:

$$\Phi = \left[\begin{array}{c|cccc} \phi_1 & 1 & 1 & 1 & -1 \\ \phi_2 & -1 & 1 & 1 & 1 \\ \phi_3 & 1 & -1 & 1 & 1 \\ \phi_4 & 1 & 1 & -1 & -1 \\ \phi_5 & 1 & -1 & 1 & -1 \\ \phi_6 & 1 & -1 & -1 & 1 \end{array}\right]$$

Note that Φ fails to do phase retrieval as it requires seven vectors in $\mathbb{R}^4$ to do phase retrieval in $\mathbb{R}^4$. Given $x = (a_1, a_2, a_3, a_4)$, $y = (b_1, b_2, b_3, b_4)$ we assume

$$|\langle x, \phi_i\rangle|^2 = |\langle y, \phi_i\rangle|^2, \text{ for all } i = 1, 2, 3, 4, 5, 6. \tag{3}$$

Step 1: The following is the coefficient matrix obtained after expanding $|\langle x, \phi_i\rangle|^2$ for $i = 1, 2, \ldots, 6$.

$$\frac{1}{2}\left[\begin{array}{c|cccccc|c} & a_1a_2 & a_1a_3 & a_1a_4 & a_2a_3 & a_2a_4 & a_3a_4 & \sum_{i=1}^{4} a_i^2 \\ E_1 & 1 & 1 & -1 & 1 & -1 & -1 & \frac{1}{2} \\ E_2 & -1 & -1 & -1 & 1 & 1 & 1 & \frac{1}{2} \\ E_3 & -1 & 1 & 1 & -1 & -1 & 1 & \frac{1}{2} \\ E_4 & 1 & -1 & -1 & -1 & -1 & 1 & \frac{1}{2} \\ E_5 & -1 & 1 & -1 & -1 & 1 & -1 & \frac{1}{2} \\ E_6 & -1 & -1 & 1 & 1 & -1 & -1 & \frac{1}{2} \end{array}\right]$$

Step 2: Consider the following row operations, the last column becomes all zeroes, so we drop it and we get:

$$\left[\begin{array}{c|cccccc} F_1 = \frac{1}{2}(E_1 - E_4) & 0 & 1 & 0 & 1 & 0 & -1 \\ F_2 = \frac{1}{2}(E_2 - E_5) & 0 & -1 & 0 & 1 & 0 & 1 \\ F_3 = \frac{1}{2}(E_3 - E_6) & 0 & 1 & 0 & -1 & 0 & 1 \\ A_1 = \frac{1}{2}(F_1 + F_2) & 0 & 0 & 0 & 1 & 0 & 0 \\ A_2 = \frac{1}{2}(F_1 + F_3) & 0 & 1 & 0 & 0 & 0 & 0 \\ A_3 = \frac{1}{2}(F_2 + F_3) & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

Step 3: Subtracting out A_1, A_2, and A_3 from E_1, E_2, E_3, and E_4, we get:

$$\left[\begin{array}{c|cccccc} E_1' = & 1 & 0 & -1 & 0 & -1 & 0 \\ E_2' = & -1 & 0 & -1 & 0 & 1 & 0 \\ E_3' = & -1 & 0 & 1 & 0 & -1 & 0 \\ E_4' = & 1 & 0 & -1 & 0 & -1 & 0 \end{array}\right]$$

Step 4: We will show that $a_ia_j = b_ib_j$ for all $i \neq j$.
Performing the given operations, we get:

$$\left[\begin{array}{c|cccccc} D_1 = \frac{-1}{2}(E_2' + E_3') & 1 & 0 & 0 & 0 & 0 & 0 \\ A_2 & 0 & 1 & 0 & 0 & 0 & 0 \\ D_2 = \frac{-1}{2}(E_1' + E_2') & 0 & 0 & 1 & 0 & 0 & 0 \\ A_1 & 0 & 0 & 0 & 1 & 0 & 0 \\ D_3 = \frac{-1}{2}(E_3' + E_4') & 0 & 0 & 0 & 0 & 1 & 0 \\ A_3 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

Doing the same operations with $y = (b_1, b_2, b_3, b_4)$, we get:

$$a_ia_j = b_ib_j, \text{ for all } 1 \leq i \neq j \leq 4.$$

Remark 1. It should be noted that weak phase retrieval does not imply norm retrieval. We may use the previous example to illustrate this. Let $\Phi = \{\phi_i\}_{i=1}^6$ be as in Example 3. Suppose Φ does norm retrieval. Since there are only 6 vectors, Φ fails the complement property. Now, take $x = (1, 1, -1, 1) \perp \{\phi_1, \phi_2, \phi_3\}$ and $y = (1, 1, 1, 1) \perp \{\phi_4, \phi_5, \phi_6\}$. Then, we have $|\langle x + y, \phi_i\rangle| = |\langle x - y, \phi_i\rangle|$ for all $i = 1, 2, \ldots 6$. From the Definition 2 (iii), this implies $\|x + y\| = \|x - y\|$. Since $\|x\| = \|y\|$, this implies that $x \perp y$, which is a contradiction.

Acknowledgement The second through fifth authors were supported by NSF DMS 1609760, NSF ATD 1321779, and ARO W911NF-16-1-0008.

References

1. R. Balan, P.G. Casazza, D. Edidin, On signal reconstruction without phase. Appl. Comput. Harmon. Anal. **20**(3), 345–356 (2006)
2. A.S. Bandeira, J. Cahill, D. Mixon, A.A. Nelson, Saving phase: injectivity and stability for phase retrieval. Appl. Comput. Harmon. Anal. **37**(1), 106–125 (2014)
3. R.H. Bates, D. Mnyama, The status of practical Fourier phase retrieval. Adv. Electron. Electron Phys. **67**, 1–64 (1986)
4. C. Becchetti, L.P. Ricotti, *Speech Recognition Theory and C++ Implementation* (Wiley, Hoboken, 1999)
5. B. Bodmann, N. Hammen, Stable phase retrieval with low redundancy frames. Preprint. arXiv:1302.5487
6. S. Botelho-Andrade, P.G. Casazza, H. Van Nguyen, J.C. Tremain, Phase retrieval versus phaseless reconstruction. J. Math Anal. Appl. **436**(1), 131–137 (2016)
7. J. Cahill, P.G. Casazza, J. Jasper, L.M. Woodland, Phase retrieval and norm retrieval (2014). arXiv preprint arXiv:1409.8266
8. J. Cahill, P. Casazza, K. Peterson, L. Woodland, Phase retrieval by projections. Available online: arXiv:1305.6226
9. A. Conca, D. Edidin, M. Hering, C. Vinzant, An algebraic characterization of injectivity of phase retrieval. Appl. Comput. Harmon. Anal. **38**(2), 346–356 (2015)
10. J. Drenth, *Principles of Protein X-ray Crystallography* (Springer, New York, 2010)
11. J.R. Fienup, Reconstruction of an object from the modulus of its fourier transform. Opt. Lett. **3**, 27–29 (1978)
12. J.R. Fienup, Phase retrieval algorithms: a comparison. Appl. Opt. **21**(15), 2758–2768 (1982)
13. T. Heinosaari, L. Maszzarella, M.M. Wolf, Quantum tomography under prior information. Commun. Math. Phys. **318**(2), 355–374 (2013)
14. L. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series (Prentice-Hall International, Inc., Upper Saddle River, 1993)
15. J.M. Renes, R. Blume-Kohout, A.J. Scott, C.M. Caves, Symmetric informationally complete quantum measurements. J. Math. Phys. **45**, 2171–2180 (2004)
16. C. Vinzant, A small frame and a certificate of its injectivity. Preprint. arXiv:1502.0465v1

Cubatures on Grassmannians: Moments, Dimension Reduction, and Related Topics

Anna Breger, Martin Ehler, Manuel Gräf, and Thomas Peter

Abstract This chapter provides an overview of recent results on cubature points in Grassmannians. We address several topics such as moment reconstruction, dimension reduction, and cubature points in Grassmannians for approximation tasks. We also provide some new results on the connection between cubatures and the concept of frames for polynomial spaces.

Keywords Grassmannian · Cubatures · Moment reconstruction · Frames · Coverings

1 Introduction

Function approximation, integration, and inverse problems are just few examples of numerical fields that rely on efficient strategies for function sampling. As particular sampling rules, the concepts of cubatures in the Euclidean space and the sphere have been widely investigated to integrate polynomials by a finite sum of sampling values, cf. [24, 32, 41, 43, 49]. To some extent, cubatures are universal sampling strategies in the sense that they are highly efficient in many fields. In certain aspects of function approximation, covering, and integration, they have proved superior to the widely used random sampling [13, 56]. Recently, cubatures on compact manifolds have attracted attention, cf. [12, 34, 51]. Integration, covering, and polynomial approximation from cubatures on manifolds and homogeneous spaces

A. Breger • M. Ehler (✉) • T. Peter
Department of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria
e-mail: anna.breger@univie.ac.at; matin.ehler@univie.ac.at; petert@uni-osnabrueck.de

M. Gräf
Acoustics Research Institute, Wohllebengasse 12-14, A-1040 Vienna, Austria
e-mail: mgraef@kfs.oeaw.ac.at

H. Boche et al. (eds.), *Compressed Sensing and its Applications*,
Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-69802-1_8

have been extensively studied from a theoretical point of view, cf. [22, 26, 36, 45, 56] and references therein.

The Grassmannian manifold is the space of orthogonal projectors of fixed rank. While orthogonality is a leading concept in many fields, projectors are intrinsically tied together with dimension reduction, an important tool in many data analysis tasks. In this chapter we like to provide a brief overview of recent results on cubatures in Grassmannian manifolds.

Our starting point in Section 2 is the problem of reconstructing a sparse (i.e., finitely supported) probability measure μ in $\mathbb{R}^d$ from its first few moments. Sparse distributions are indeed uniquely determined by their first few moments, and Prony's method has recently been adapted to this reconstruction [8, 44]. According to the Johnson-Lindenstrauss lemma, low-dimensional projections of μ still capture essential information [21]. Taking the first few moments of low-dimensional projections only, we now aim to reconstruct the first few moments of μ, but we allow for general probability distributions in Section 3, not necessarily sparse ones, cf. [11]. A new construction of suitable projections is provided in Theorem 2. It turns out that the choice of projectors is closely related to cubatures in Grassmannians, i.e., the set of low-dimensional projectors should form a cubature (see Section 4, Theorem 3). Hence, the reconstruction of high-dimensional moments from lower-dimensional ones is naturally related to the concept of Grassmannian cubatures. We then discuss in Section 5.1 numerical constructions of cubatures in Grassmannians by minimizing the worst-case integration error of polynomials, cf. [2, 16]. In Section 5.2, we go beyond polynomials and briefly discuss sequences of low-cardinality cubatures that yield optimal worst-case integration error rates for Bessel potential functions, cf. [12]; see also [16]. The optimal integration errors of cubatures directly induce schemes for function approximation from samples by replacing the inner products of L_2 orthonormal basis expansions with cubature approximations; see Section 5.3. Intuitively, good samplings for function approximation should cover the underlying space well. Indeed in Section 5.4, we recapitulate that sequences of low-cardinality Grassmannian cubatures are asymptotically optimal coverings, cf. [15]. To further reflect on the versatility of Grassmannian cubatures, we also discuss their use in phase retrieval problems; see Section 5.5.

So far, we have outlined the use of Grassmannian cubatures for various topics in numerical mathematics. Within a single Grassmannian, the rank of the projectors is fixed. However, the use of projectors with varying ranks offers additional flexibility and may have benefits in practice; see [40, 59]. Therefore, the concept of cubatures on unions of Grassmannians is discussed in Section 6. The number of required cubature points is mainly steered by the dimension of the underlying polynomial space. By determining the dimensions of polynomial spaces on unions of Grassmannians, cf. [29], we address one of the necessary prerequisites for the aforementioned topics within unions of Grassmannians (i.e., approximation of integrals and functions, moment reconstruction, covering, and phase retrieval). For special cases, we provide elementary proofs. The general cases need deeper analysis, for which we refer to [29].

2 Reconstruction from Moments and Dimension Reduction

2.1 Reconstructing Sparse Distributions from Moments

Our starting point is a high-dimensional random vector $X \in \mathbb{R}^d$ with finite support $\{x_i\}_{i=1}^m \subset \mathbb{R}^d$, i.e., X is distributed according to a discrete probability measure μ on $\mathbb{R}^d$ with support $\{x_i\}_{i=1}^m$ and positive weights $\{a_i\}_{i=1}^m$ satisfying $\sum_{i=1}^m a_i = 1$, so that

$$\mu = \sum_{i=1}^m a_i \delta_{x_i},$$

where δ_{x_i} denotes the point measure at x_i. We now aim to reconstruct μ from knowledge of the moments

$$m_\mu(\lambda) := \mathbb{E}X^\lambda = \sum_{i=1}^m a_i x_i^\lambda, \quad \lambda \in \Lambda, \tag{1}$$

where $\Lambda \subset \mathbb{N}^d$ is some fixed subset. The nonlinear inverse problem of reconstructing μ means to identify its support $\{x_i\}_{i=1}^m$ and its weights $\{a_i\}_{i=1}^m$. The core idea of Prony's method is to determine an ideal $\mathcal{I}$ of polynomials on $\mathbb{R}^d$ just from the moments $m_\mu(\lambda)$, $\lambda \in \Lambda$, through a system of linear equations, such that its zero locus

$$\mathcal{V}(\mathcal{I}) = \{x \in \mathbb{R}^d : f(x) = 0,\ \forall f \in \mathcal{I}\}$$

is exactly the point set $\{x_i\}_{i=1}^m$. The one-dimensional case, expressed in terms of difference equations, was introduced in [8]; see also [10, 54, 55]; the multivariate case is treated in [44].

Once $\mathcal{I}$ is determined, its zero locus $\mathcal{V}(\mathcal{I}) = \{x_i\}_{i=1}^m$ can be determined by standard methods [9], and the weights $\{a_i\}_{i=1}^m$ are computed by a system of linear equations from the Vandermonde system (1).

More specifically, the zero locus $\mathcal{V}(\mathcal{I}_i)$ of each ideal

$$\mathcal{I}_i := \big((z - x_i)^\alpha : \alpha \in \mathbb{N}^d,\ |\alpha| = 1\big)$$

is $\mathcal{V}(\mathcal{I}_i) = \{x_i\}$, for $i = 1, \ldots, m$, so that $\{x_i\}_{i=1}^m = \mathcal{V}(\mathcal{I})$ with $\mathcal{I} := \mathcal{I}_1 \cdots \mathcal{I}_m$. Note that $\mathcal{I}$ coincides with

$$\mathcal{I} = \Big(\prod_{i=1}^m (z - x_i)^{\alpha_i} : \alpha_i \in \mathbb{N}^d,\ |\alpha_i| = 1,\ i = 1, \ldots, m\Big),$$

so that we have d^m many generators of the ideal that must now be determined from the moments $m_\mu(\lambda)$, $\lambda \in \Lambda$.

To simplify, let us now suppose that $d = 1$. In this case, the ideal $\mathcal{I}$ is generated by the single polynomial

$$p(z) = (z - x_1) \cdots (z - x_m) = \sum_{k=0}^{m} p_k z^k$$

of degree m. Its coefficient sequence $\{p_k\}_{k=0}^m$ satisfies

$$\sum_{k=0}^{m} p_k m_\mu(k + \lambda) = \sum_{i=1}^{m} x_i^\lambda a_i \sum_{k=0}^{m} p_k x_i^k = \sum_{i=1}^{m} x_i^\lambda a_i p(x_i) = 0. \tag{2}$$

Equation (2) holds for arbitrary values of λ. Thus, varying λ und using that $p_m = 1$ leads to the linear system of equations

$$\sum_{k=0}^{m-1} p_k m_\mu(k + \lambda) = -m_\mu(m + \lambda), \quad \lambda \in \Lambda', \tag{3}$$

where $\Lambda' \subset \Lambda$ such that $k + \Lambda' \subset \Lambda$, for all $k = 0, \ldots, m$. We now attempt to solve (3) for $p_0, \ldots, p_{m-1}$. Obviously, Λ must be sufficiently large, so that

$$H := \big(m_\mu(k + \lambda)\big)_{\substack{\lambda \in \Lambda' \\ k=0,\ldots,m-1}} \in \mathbb{R}^{|\Lambda'| \times m} \tag{4}$$

can have full rank m. From knowledge of p, the eigenvalues of its companion matrix yield its zeros $\{x_i\}_{i=1}^m$. Having determined $\{x_i\}_{i=1}^m$, (1) yields a Vandermonde system of linear equations to compute the weights $\{a_i\}_{i=1}^m$. Note that the rank condition in (4) is satisfied for $\Lambda = \{0, \ldots, 2m - 1\}$ and $\Lambda' = \{0, \ldots, m - 1\}$, cf. [52] and [53] for an overview of Prony's methods.

The case $d > 1$ is more involved but can essentially be treated similarly. In [44] it is shown that $\#\Lambda = \mathcal{O}(m^d)$ suffices to ensure reconstruction, while $\#\Lambda = \mathcal{O}(md)$ suffices if $\{x_i\}_{i=1}^m$ are in general position.

Concerning numerical stability, one has to differentiate between the idea of Prony's method as presented here and stable numerical variants for implementation as for example ESPRIT [57], MUSIC [58], and finite rate of innovation [60]. These algorithms perform excellent in many applications. If λ and k are chosen as proposed in (2), the system matrix (4) is a Hankel matrix that can be factored into

$$H = A^\top D A$$

with a diagonal matrix $D = \operatorname{diag}(a_i)_{i=1}^m$ and a Vandermonde matrix $A = (x_i^k)_{k=0,i=1}^{m-1,m}$. For $d > 1$, a similar factorization holds, where A is a generalized Vandermonde matrix. Due to the Vandermonde structure of A, its condition number tends to be large if the minimal separation distance ν_μ is small or if there are large corner deviations α_μ, i.e.,

$$\nu_\mu := \min_{\substack{i,j=1,\dots,m \\ i\neq j}} \|x_i - x_j\|_2 \approx 0 \quad \text{or} \quad \alpha_\mu := \max_{\substack{i=1,\dots,m \\ j=1,\dots,d}} |\log\left(|\langle x_i, e_j\rangle|\right)| \gg 0.$$

This pinpoints stable performances when the measure μ has well-separated support without large growing, respectively, damping factors and with well-behaved weights.

Note that the Prony method works beyond probability measures and can deal with $x_i \in \mathbb{C}^d$, $a_i \in \mathbb{C}$ and to this end also with $\lambda \in \mathbb{Z}^d$. Indeed, if Λ' in (2) is chosen as $\Lambda' \subset -\mathbb{N}^d$, then the resulting system matrix becomes a Toeplitz matrix, which is preferred in some literature on Prony's method.

2.2 Dimension Reduction

The idea of dimension reduction is that properties of interest of a high-dimensional random vector $X \in \mathbb{R}^d$ may still be captured within its orthogonal $k < d$ dimensional projection, i.e., in PX, where P is an element in the Grassmannian space

$$\mathcal{G}_{k,d} := \{P \in \mathbb{R}^{d\times d}_{\text{sym}} \ : \ P^2 = P;\ \text{trace}(P) = k\}.$$

Here, $\mathbb{R}^{d\times d}_{\text{sym}}$ is the set of symmetric matrices in $\mathbb{R}^{d\times d}$. Consider two sparsely distributed independent random vectors

$$X, Y \sim \sum_{i=1}^m a_i \delta_{x_i}. \tag{5}$$

Their difference $X - Y$ is distributed according to

$$X - Y \sim \sum_{i,j=1}^m a_i a_j \delta_{x_i - x_j}.$$

For $P \in \mathcal{G}_{k,d}$, the magnitude of the differences is distributed according to

$$\|PX - PY\|^2 \sim \sum_{i,j=1}^m a_i a_j \delta_{\|Px_i - Px_j\|^2}.$$

In fact, for $0 < \epsilon < 1$ and k with $d \geq k \geq \frac{4\log(m)}{\epsilon^2/2 - \epsilon^3/3}$, there is $P \in \mathcal{G}_{k,d}$, such that

$$(1-\epsilon)\|X - Y\|^2 \leq \frac{d}{k}\|PX - PY\|^2 \leq (1+\epsilon)\|X - Y\|^2 \tag{6}$$

holds with probability 1. This is the direct consequence of realizations of the Johnson-Lindenstrauss lemma applied to the deterministic point set $\{x_i\}_{i=1}^m$, cf. [21].

Note that (6) tells us that the dimension reduction still preserves essential information of X and Y. At this point though, we just know of its existence, and we have not yet specified any particular projector P such that (6) holds; see [1, 21, 46] for different types of random choices.

We should point out that PX and PY are contained in a k-dimensional subspace of $\mathbb{R}^d$ but still have d entries as vectors in d dimensions. The actual dimension reduction takes place by applying $Q \in \mathcal{V}_{k,d}$ with $Q^\top Q = P$, where

$$\mathcal{V}_{k,d} := \{Q \in \mathbb{R}^{k\times d} : QQ^\top = I_k\}$$

denotes the Stiefel manifold. The inequality (6) becomes

$$(1-\epsilon)\|X-Y\|^2 \leq \frac{d}{k}\|QX-QY\|^2 \leq (1+\epsilon)\|X-Y\|^2,$$

where $QX, QY \in \mathbb{R}^k$ are properly dimension-reduced random vectors still containing the information of the pairwise differences up to a factor $1 \pm \epsilon$.

3 High-Dimensional Moments from Lower-Dimensional Ones

3.1 Moments and Spanning Sets

We shall now combine dimension reduction with a modified problem, which is related to the reconstruction from moments. First, we drop the sparsity conditions and allow arbitrary probability measures μ on $\mathbb{R}^d$. Let $X \in \mathbb{R}^d$ be some random vector with unknown Borel probability distribution on $\mathbb{R}^d$. Suppose we do not have access to its moments, but we observe the first few moments of order T of low-dimensional linear projections, i.e., for $\{Q_j\}_{j=1}^n \subset \mathcal{V}_{k,d}$, we measure

$$\mathbb{E}(Q_jX)^s, \quad s \in \mathbb{N}^k, \ |s| \leq T. \tag{7}$$

We cannot reconstruct μ directly, but we aim to determine the first few high-dimensional moments

$$\mathbb{E}X^r, \quad r \in \mathbb{N}^d, \ |r| \leq T. \tag{8}$$

In other words, we know the first few moments of order T of the dimension-reduced random vectors $Q_jX \in \mathbb{R}^k$, $j = 1, \ldots, n$, and our task is to reconstruct the high-dimensional moments, cf. [11]. The idea is to interpret moments as algebraic polynomials and represent desired high-degree polynomials as products of polynomials of lower degree.

Polynomials of total degree T on $\mathbb{R}^d$, denoted by $\mathrm{Pol}_T(\mathbb{R}^d)$, are decomposed by

$$\mathrm{Pol}_T(\mathbb{R}^d) = \bigoplus_{t=0}^{T} \mathrm{Hom}_t(\mathbb{R}^d),$$

where $\mathrm{Hom}_t(\mathbb{R}^d)$ denotes the space of homogeneous polynomials of degree t on $\mathbb{R}^d$. Let $x \in \mathbb{R}^d$ be a vector of unknowns; then $(Q_j x)^s$ is a homogenous polynomial of degree $|s|$. If

$$\{(Q_j x)^s : j = 1, \ldots, n,\ s \in \mathbb{N}^k,\ |s| = t\} \tag{9}$$

spans $\mathrm{Hom}_t(\mathbb{R}^d)$, then each monomial of order t is a linear combination of elements in (9), i.e., for $r \in \mathbb{N}^d$ with $|r| = t$, there are coefficients $c_{j,s}$ such that

$$x^r = \sum_{j=1}^{n} \sum_{s \in \mathbb{N}^k,\, |s|=t} c_{j,s} (Q_j x)^s, \qquad \text{for all } x \in \mathbb{R}^d. \tag{10}$$

Hence, the linearity of the expectation yields that all high-dimensional moments of order t can be reconstructed from the low-dimensional moments

$$\mathbb{E}(Q_j X)^s, \quad j = 1, \ldots, n, \quad |s| = t.$$

Let us summarize the above discussion:

Theorem 1. *Let $X \in \mathbb{R}^d$ be a random vector, and, for $\{Q_j\}_{j=1}^n \subset \mathcal{V}_{k,d}$, suppose that (9) spans $\mathrm{Hom}_t(\mathbb{R}^d)$. Then any moment $\mathbb{E}X^r$, for $r \in \mathbb{N}^d$ with $|r| = t$, is a linear combination of*

$$\{\mathbb{E}(Q_j X)^s : j = 1, \ldots, n,\ s \in \mathbb{N}^k,\ |s| = t\},$$

where the coefficients are independent of the distribution of X and are taken from (10).

Thus, we aim to find $\{Q_j\}_{j=1}^n$, such that (9) spans $\mathrm{Hom}_t(\mathbb{R}^d)$ for each $t \le T$. This is the topic of the subsequent section. Note that spanning sets in finite dimensions are also called frames.

3.2 Frames for Polynomial Spaces

The most excessive dimension reduction in Theorem 1 corresponds to $k = 1$. In this case, we observe that we only need to take care of the maximal $t = T$:

Proposition 1. *Let $\{Q_j\}_{j=1}^n \subset \mathcal{V}_{1,d}$ and $x \in \mathbb{R}^d$ be a vector of unknowns.*

a) If $\{(Q_j x)^t\}_{j=1}^n$ *is a frame for* $\mathrm{Hom}_t(\mathbb{R}^d)$*, then* $\{(Q_j x)^{t-1}\}_{j=1}^n$ *is a frame for* $\mathrm{Hom}_{t-1}(\mathbb{R}^d)$.

b) If $\{(Q_j x)^{t-1}\}_{j=1}^n$ *is linearly independent in* $\mathrm{Hom}_{t-1}(\mathbb{R}^d)$*, then* $\{(Q_j x)^t\}_{j=1}^n$ *is linearly independent in* $\mathrm{Hom}_t(\mathbb{R}^d)$.

Proof.

a) Let f be an arbitrary element in $\mathrm{Hom}_{t-1}(\mathbb{R}^d)$. There is $g \in \mathrm{Hom}_t(\mathbb{R}^d)$ such that its first partial derivative $\partial_1 g$ coincides with f. Since $\{(Q_j x)^t\}_{j=1}^n$ is a frame for $\mathrm{Hom}_t(\mathbb{R}^d)$, there are coefficients $\{c_j\}_{j=1}^n$ such that $g = \sum_{j=1}^n c_j (Q_j x)^t$. Therefore, we obtain

$$f(x) = \sum_{j=1}^{n} c_j (Q_j e_1) t (Q_j x)^{t-1},$$

which verifies part a).

b) Suppose that $0 = \sum_{j=1}^n c_j (Q_j x)^t$. Applying all partial derivatives yields

$$0 = \sum_{j=1}^{n} c_j (Q_j e_i) t (Q_j x)^{t-1}, \qquad i = 1, \ldots, d.$$

The linear independence assumption implies $c_j (Q_j e_i) = 0$, for $i = 1, \ldots, d$, and, therefore, $c_j = 0$, for $j = 1, \ldots, n$, since $Q_j \neq 0$.

Part a) of Proposition 1 tells us that if $\{(Q_j x)^t\}_{j=1}^n$ is a frame for $\mathrm{Hom}_t(\mathbb{R}^d)$, then

$$\{(Q_j x)^s : j = 1, \ldots, n,\ s \in \mathbb{N},\ |s| \leq t\} \tag{11}$$

is a frame for $\mathrm{Pol}_t(\mathbb{R}^d)$. The proof directly shows that the first low-dimensional moments are sufficient to reconstruct the first high-dimensional moments.

Next, we provide a general construction recipe of $\{Q_j\}_{j=1}^n \subset \mathcal{V}_{1,d}$ that covers arbitrary d and t. Note that the dimension of $\mathrm{Hom}_t(\mathbb{R}^d)$ is $\binom{t+d-1}{d-1}$.

Theorem 2. *Let* $\{v_i\}_{i=1}^d$ *be pairwise different positive real numbers, let* $\{\alpha_j\}_{j=1}^{t+d-1}$ *be pairwise different nonnegative integers, and let* $V = (v_i^{\alpha_j})_{i,j}$ *denote the associated* $(t+d-1) \times d$*-Vandermonde-type matrix. Suppose that the* $\binom{t+d-1}{d-1} \times d$ *matrix* Q *is built from all minors of* V *of order* $d-1$*. We denote the rows of* Q *by* $Q_1, \ldots, Q_n$*, where* $n = \binom{t+d-1}{d-1}$*. Then* $\{(Q_j x)^t\}_{j=1}^n$ *is a basis for* $\mathrm{Hom}_t(\mathbb{R}^d)$.

Proof. We expand $(Q_j x)^t$ by the multivariate binomial formula

$$(Q_j x)^t = \sum_{\alpha \in \mathbb{N}^d,\ |\alpha| = t} \binom{t}{\alpha} Q_j^\alpha x^\alpha.$$

The coefficients are put into the j-th row of a matrix $M_1 \in \mathbb{R}^{n\times n}$, i.e.,

$$M_1 = \left(\binom{t}{\alpha} Q_j^\alpha \right)_{j,\alpha}.$$

We must now check that M_1 is invertible.

Dividing each column α by its respective binomial coefficient $\binom{t}{\alpha}$ yields the matrix $M_2 = \left(Q_j^\alpha\right)_{j,\alpha} \in \mathbb{R}^{n\times n}$, and M_1 is invertible if and only if M_2 is. Let c denote the product of all minors of order d of V. It follows from [61] that

$$\det(M_2) = c^{d-1}.$$

The Vandermonde structure yields that $c \neq 0$, so that M_2 and hence M_1 are invertible. Thus, $\{(Q_j x)^t\}_{j=1}^n$ is indeed a basis for $\mathrm{Hom}_t(\mathbb{R}^d)$.
Note that normalization of the rows of Q in Theorem 2 yields $\{Q_j\}_{j=1}^n \subset \mathcal{V}_{1,d}$, and $\{(Q_j x)^t\}_{j=1}^n$ is a basis for $\mathrm{Hom}_t(\mathbb{R}^d)$. Thus, for each $s \leq t$, $\{(Q_j x)^s\}_{j=1}^n$ is a frame for $\mathrm{Hom}_s(\mathbb{R}^d)$ according to Proposition 1.

4 Frames vs. Cubatures for Moment Reconstruction

4.1 Frames and Cubatures on the Sphere and Beyond

So far, we have seen that reconstruction of high-dimensional moments from low-dimensional ones is related to frames for $\mathrm{Hom}_t(\mathbb{R}^d)$. Next, we shall relate such frames to cubature points. Let $\mathrm{Hom}_t(\mathbb{S}^{d-1})$ denote the space of homogeneous polynomials $\mathrm{Hom}_t(\mathbb{R}^d)$ restricted to the sphere $\mathbb{S}^{d-1}$. For points $\{Q_j\}_{j=1}^n \subset \mathbb{S}^{d-1}$ and weights $\{\omega_j\}_{j=1}^n \subset \mathbb{R}$, we say that $\{(Q_j, \omega_j)\}_{j=1}^n$ is a cubature for $\mathrm{Hom}_t(\mathbb{S}^{d-1})$ if

$$\int_{\mathbb{S}^{d-1}} f(x)\mathrm{d}x = \sum_{j=1}^n \omega_j f(Q_j), \quad \text{for all } f \in \mathrm{Hom}_t(\mathbb{S}^{d-1}),$$

where $\mathrm{d}x$ denotes the standard measure on the sphere normalized to have mass one. Note that for $k = 1$, the Stiefel manifold $\mathcal{V}_{1,d}$ coincides with $\mathbb{S}^{d-1}$. It turns out that the frame property of $\{(Q_j x)^s\}_{j=1}^n$ is related to the concept of cubature points:

Theorem 3. *Let $\{Q_j\}_{j=1}^n \subset \mathcal{V}_{1,d}$ and $x \in \mathbb{R}^d$ be a vector of unknowns.*

a) If $\{(Q_j x)^t\}_{j=1}^n$ is a frame for $\mathrm{Hom}_t(\mathbb{R}^d)$, then there are weights $\{\omega_j\}_{j=1}^n \subset \mathbb{R}$, such that $\{(Q_j^\top, \omega_j)\}_{j=1}^n$ is a cubature for $\mathrm{Hom}_t(\mathbb{S}^{d-1})$.

b) If there are weights $\{\omega_j\}_{j=1}^n \subset \mathbb{R}$ such that $\{(Q_j^\top, \omega_j)\}_{j=1}^n$ is a cubature for $\mathrm{Hom}_{2t}(\mathbb{S}^{d-1})$, then $\{(Q_j x)^t\}_{j=1}^n$ is a frame for $\mathrm{Hom}_t(\mathbb{R}^d)$.

Proof. a) Since $\{(Q_j x)^t\}_{j=1}^n$ is a frame for $\mathrm{Hom}_t(\mathbb{R}^d)$, for each $a \in \mathbb{S}^{d-1}$, there are coefficients $\{c_j(a)\}_{j=1}^n \subset \mathbb{R}$ such that

$$(a^\top x)^t = \sum_{j=1}^n c_j(a)(Q_j x)^t.$$

Note that the mapping $a \mapsto c_j(a)$ can be chosen to be continuous, for each $j = 1, \dots, n$. Therefore, we derive

$$\int_{\mathbb{S}^{d-1}} (a^\top x)^t \mathrm{d}a = \sum_{j=1}^n (Q_j x)^t \int_{\mathbb{S}^{d-1}} c_j(a) \mathrm{d}a = \sum_{j=1}^n (Q_j x)^t \omega_j,$$

with $\omega_j = \int_{\mathbb{S}^{d-1}} c_j(a) \mathrm{d}a$. Since the above equality holds for all $x \in \mathbb{R}^d$, $\{(Q_j^\top, \omega_j)\}_{j=1}^n$ is a cubature for $\mathrm{Hom}_t(\mathbb{S}^{d-1})$.

b) Note that $\mathrm{Hom}_t(\mathbb{S}^{d-1})$ is a reproducing kernel Hilbert space, and let us denote its reproducing kernel with respect to the standard inner product by K_t. For now, we restrict x and a to the sphere. The reproducing property yields

$$(a^\top x)^t = \int_{\mathbb{S}^{d-1}} (z^\top x)^t K_t(z, a) \mathrm{d}z.$$

The mapping $z \mapsto (z^\top x)^t K_t(z, a)$ is contained in $\mathrm{Hom}_{2t}(\mathbb{S}^{d-1})$, so that the cubature property yields

$$(a^\top x)^t = \sum_{j=1}^n \omega_j (Q_j x)^t K_t(Q_j^\top, a) = \sum_{j=1}^n (Q_j x)^t c_j(a),$$

where $c_j(a) = \omega_j K_t(Q_j^\top, a)$. A homogeneity argument concludes the proof.
Note that the degree of the homogeneous polynomials in Part b) of Theorem 3 is not the same ($2t$ for the cubatures and t for the frame). The degree $2t$ is due to multiplication of two homogeneous polynomials of degree t, which is not just an artifact of the proof. There are indeed cubatures for $\mathrm{Hom}_t(\mathbb{S}^{d-1})$, whose cardinality is lower than the dimension of $\mathrm{Hom}_t(\mathbb{R}^d)$; see [37], for instance.

In fact, Theorem 3 holds in much more generality in suitable finite dimensional reproducing kernel Hilbert spaces. Let (Ω, σ) be a finite measure space, and let $\mathcal{F}$ be a linear subspace of continuous functions in $L_2(\Omega, \sigma)$. For points $\{q_j\}_{j=1}^n \subset \Omega$ and weights $\{\omega_j\}_{j=1}^n \subset \mathbb{R}$, we say that $\{(q_j, \omega_j)\}_{j=1}^n$ is a cubature for $\mathcal{F}$ if

$$\int_\Omega f(x) \mathrm{d}\sigma(x) = \sum_{j=1}^n \omega_j f(q_j), \quad \text{for all } f \in \mathcal{F}.$$

The following result generalizes Theorem 3:

Proposition 2. *Let $K : \Omega \times \Omega \to \mathbb{R}$ be a symmetric kernel that linearly generates $\mathcal{F}$, i.e., $K(x, y) = K(y, x)$ for $x, y \in \Omega$, and*

$$\mathcal{F} = \operatorname{span}\{K(a, \cdot) : a \in \Omega\}. \tag{12}$$

For $\{q_j\}_{j=1}^n \subset \Omega$, the following holds:

a) If $\{K(q_j, \cdot)\}_{j=1}^n$ is a frame for $\mathcal{F}$, then there are weights $\{\omega_j\}_{j=1}^n \subset \mathbb{R}$, such that $\{(q_j, \omega_j)\}_{j=1}^n$ is a cubature for $\mathcal{F}$.
b) If there are weights $\{\omega_j\}_{j=1}^n \subset \mathbb{R}$ such that $\{(q_j, \omega_j)\}_{j=1}^n$ is a cubature for the linear span of $\mathcal{F} \cdot \mathcal{F}$, then $\{K(q_j, \cdot)\}_{j=1}^n$ is a frame for $\mathcal{F}$.

The proof of Proposition 2 is structurally the same as for Theorem 3 with $K(x, y) = (x^\top y)^t$ and $\mathcal{F} = \operatorname{Hom}_t(\mathbb{S}^{d-1})$, so we omit the details.

Remark 1. Part b) of Proposition 2 implies $n \geq \dim(\mathcal{F})$. Analoguous results in [23], for instance, are restricted to positive weights.

4.2 Moment Reconstruction with Cubatures in Grassmannians

To switch to the Grassmannian setting, we first note that $Q \in \mathcal{V}_{1,d}$ if and only if $Q^\top Q \in \mathcal{G}_{1,d}$. Moreover, the kernel

$$K_{t,1} : \mathcal{G}_{1,d} \times \mathcal{G}_{1,d} \to \mathbb{R}, \quad (P, R) \mapsto \operatorname{trace}(PR)^t$$

linearly generates $\operatorname{Hom}_t(\mathcal{G}_{1,d})$. Let $\operatorname{Hom}_t(\mathcal{G}_{k,d})$ denote the restrictions of homogeneous polynomials of degree t on $\mathbb{R}^{d\times d}_{\text{sym}}$ to the Grassmannian $\mathcal{G}_{k,d}$. For $x \in \mathbb{S}^{d-1}$, it holds that

$$(Qx)^{2t} = K_{t,1}(Q^\top Q, xx^\top), \tag{13}$$

so that $\operatorname{Hom}_{2t}(\mathbb{S}^{d-1})$ corresponds to $\operatorname{Hom}_t(\mathcal{G}_{1,d})$. According to (13) we deduce that for $\{Q_j\}_{j=1}^n \subset \mathcal{V}_{1,d}$ the set $\{(Q_j x)^{2t}\}_{j=1}^n$ is a frame for $\operatorname{Hom}_{2t}(\mathbb{R}^d)$ if and only if $\{K_{t,1}(Q_j^\top Q_j, \cdot)\}_{j=1}^n$ is a frame for $\operatorname{Hom}_t(\mathcal{G}_{1,d})$. Similarly, $\{(Q_j^\top, \omega_j)\}_{j=1}^n$ is a cubature for $\operatorname{Hom}_{2t}(\mathbb{S}^{d-1})$ if and only if $\{(Q_j^\top Q_j, \omega_j)\}_{j=1}^n$ is a cubature for $\operatorname{Hom}_t(\mathcal{G}_{1,d})$. Therefore, we can switch to the Grassmannian setting to formulate the following moment reconstruction result:

Corollary 1 ([11]). *For $r \in \mathbb{N}^d$ with $|r| \leq t \leq d$, there are coefficients $a_s^r \in \mathbb{R}$, $s \in \mathbb{N}^d$, $|s| = |r|$, such that if $\{(P_j, \omega_j)\}_{j=1}^n$ is a cubature for $\operatorname{Hom}_t(\mathcal{G}_{1,d})$, then any random vector $X \in \mathbb{R}^d$ satisfies*

$$\mathbb{E}X^r = \sum_{j=1}^n \omega_j \sum_{s\in\mathbb{N}^d,\ |s|=|r|} a_s^r \mathbb{E}(P_j X)^s. \tag{14}$$

For $Q_j \in \mathcal{V}_{1,d}$ and $P_j \in \mathcal{G}_{1,d}$ with $P_j = Q_j^\top Q_j$, one can switch between the moments of Q_jX and P_jX by the formula

$$(P_jX)^s = Q_j^s (Q_jX)^{|s|}, \quad s \in \mathbb{N}^d.$$

It may depend on the context whether P_jX or Q_jX is preferred.

5 Cubatures in Grassmannians

Proposition 2 connects frames and cubatures beyond $\mathcal{G}_{1,d}$ and can be applied to the general Grassmannians $\mathcal{G}_{k,d}$ by $\Omega = \mathcal{G}_{k,d}$, $\mathcal{F} = \mathrm{Hom}_t(\mathcal{G}_{k,d})$, and the kernel $K = K_{t,k}$ given by

$$K_{t,k} : \mathcal{G}_{k,d} \times \mathcal{G}_{k,d} \to \mathbb{R}, \quad (P, R) \mapsto \mathrm{trace}(PR)^t,$$

cf. [11, 29]. In the following sections, we shall provide further examples for the usefulness of Grassmannian cubatures beyond moment reconstruction.

5.1 Numerical Construction of Cubatures

Cubatures on Grassmannians with constant weights are constructed in [4] from group orbits. Here we shall briefly present a method based on numerical minimization. The t-fusion frame potential for points $\{P_j\}_{j=1}^n \subset \mathcal{G}_{k,d}$ and weights $\{\omega_j\}_{j=1}^n \subset \mathbb{R}$ is

$$\mathrm{FFP}(\{(P_j, \omega_j)\}_{j=1}^n, t) := \sum_{i,j=1}^n \omega_j \omega_i \,\mathrm{trace}(P_i P_j)^t.$$

Note that $\mathcal{G}_{k,d}$ is naturally endowed with an orthogonally invariant probability measure $\mu_{k,d}$. Assuming that $\sum_{j=1}^n \omega_j = 1$, the fusion frame potential is lower bounded by

$$\mathrm{FFP}(\{(P_j, \omega_j)\}_{j=1}^n, t) \geq \int_{\mathcal{G}_{k,d}} \int_{\mathcal{G}_{k,d}} \mathrm{trace}(PR)^t \mathrm{d}\mu_{k,d}(P) \mathrm{d}\mu_{k,d}(R), \tag{15}$$

cf. [16] and also [2]. Since the constant functions are contained in $\mathrm{Hom}_t(\mathcal{G}_{k,d})$, any cubature must satisfy $\sum_{j=1}^n \omega_j = 1$.

Theorem 4 ([2, 16]). *If* $\sum_{j=1}^n \omega_j = 1$ *and* (15) *holds with equality, then* $\{(P_j, \omega_j)\}_{j=1}^n$ *is a cubature for* $\mathrm{Hom}_t(\mathcal{G}_{k,d})$.

In order to check for equality in (15), we require a more explicit expression for the right-hand side. In fact, it holds

$$\int_{\mathcal{G}_{k,d}} \int_{\mathcal{G}_{k,d}} \mathrm{trace}(PR)^t \mathrm{d}\mu_{k,d}(P)\mathrm{d}\mu_{k,d}(R) = \sum_{\substack{|\pi|=t,\\ \ell(\pi)\le d/2}} \frac{C_\pi^2(I_k)}{C_\pi(I_d)},$$

where I_d denotes the $d\times d$ identity matrix and π is an integer partition of t with $\ell(\pi)$ being the number of nonzero parts and C_π are the zonal polynomials, cf. [19, 38, 48]. Evaluation of C_π at I_k and I_d, respectively, yields

$$C_\pi(I_d) = 2^{|\pi|}|\pi|!\left(\tfrac{d}{2}\right)_\pi \prod_{1\le i<j\le \ell(\pi)} (2\pi_i - 2\pi_j - i + j)/\prod_{i=1}^{\ell(\pi)} (2\pi_i + \ell(\pi) - i)!,$$

cf. [27]. Here, $(a)_\pi$ denotes the generalized hypergeometric coefficient given by

$$(a)_\pi := \prod_{i=1}^{\ell(\pi)} \left(a - \tfrac{1}{2}(i-1)\right)_{\pi_i}, \qquad (a)_s := a(a+1)\dots(a+s-1). \tag{16}$$

Fixing the weights $\{\omega_j\}_{j=1}^n \subset \mathbb{R}$, say $\omega_j = 1/n$, for $j = 1, \dots, n$, we can now aim to numerically minimize the t-fusion frame potential $\mathrm{FFP}(\{(P_j, \omega_j)\}_{j=1}^n, t)$ over all sets of n points $\{P_j\}_{j=1}^n \subset \mathcal{G}_{k,d}$ and check for equality in (15), where the right-hand side can be computed explicitly. Equality can be achieved for suitable relations between n and t, cf. [33]. See [15, 16] for successful minimizations in $\mathcal{G}_{2,4}$.

5.2 Cubatures for Approximation of Integrals

Cubature points enable us to replace integrals over polynomials by finite sums. We now aim to go beyond polynomials and keep track of the integration error. Without loss of generality, we assume $k \le \frac{d}{2}$ throughout since $\mathcal{G}_{d-k,d}$ can be identified with $\mathcal{G}_{k,d}$.

The eigenfunctions $\{\varphi_\pi\}_{\ell(\pi)\le k}$ of the Laplace-Beltrami operator Δ on $\mathcal{G}_{k,d}$ are an orthonormal basis for $L_2(\mathcal{G}_{k,d})$ and are naturally indexed by integer partitions π of length at most k. Let $\{-\lambda_\pi\}_{\ell(\pi)\le k}$ be the corresponding eigenvalues, i.e.,

$$\lambda_\pi = 2|\pi|d + 4\sum_{i=1}^k \pi_i(\pi_i - i), \tag{17}$$

cf. [42, Theorem 13.2]. Without loss of generality, we choose each φ_π to be real valued, in particular, $\varphi_{(0)} \equiv 1$. Essentially following [12, 47], we formally define $(I-\Delta)^{s/2} f$ by

$$\langle (I-\Delta)^{s/2} f, \varphi_\pi \rangle := (1+\lambda_\pi)^{s/2} \langle f, \varphi_\pi \rangle, \qquad \text{for all } \ell(\pi) \le k.$$

The Bessel potential space $H_p^s(\mathcal{G}_{k,d})$, for $1 \le p \le \infty$ and $s \ge 0$, is

$$H_p^s(\mathcal{G}_{k,d}) := \{ f \in L_p(\mathcal{G}_{k,d}) : \|f\|_{H_p^s} < \infty \}, \quad \text{where}$$
$$\|f\|_{H_p^s} := \|(I-\Delta)^{s/2} f\|_{L_p},$$

i.e., $f \in H_p^s(\mathcal{G}_{k,d})$ if and only if $f \in L_p(\mathcal{G}_{k,d})$ and $(I-\Delta)^{s/2} f \in L_p(\mathcal{G}_{k,d})$.

The expected worst-case error of integration in Bessel potential spaces of n independent random points endowed with constant weights is of the order $n^{-\frac{1}{2}}$:

Proposition 3 ([13, 16, 37, 50]). *For $s > k(d-k)/2$, suppose $P_1, \dots, P_n$ are random points on $\mathcal{G}_{k,d}$, independently identically distributed according to $\mu_{k,d}$ then it holds*

$$\sqrt{\mathbb{E}\Big[\sup_{\substack{f \in H_2^s(\mathcal{G}_{k,d}) \\ \|f\|_{H_2^s} \le 1}} \Big| \int_{\mathcal{G}_{k,d}} f(P) \mathrm{d}\mu_{k,d}(P) - \frac{1}{n} \sum_{j=1}^n f(P_j) \Big|^2 \Big]} = c n^{-\frac{1}{2}}$$

with $c^2 = \sum_{1 \le \ell(\pi) \le k} (1+\lambda_\pi)^{-s}$.

The following result follows from [12, Theorem 2.12]:

Theorem 5 ([12]). *Let $s > k(d-k)/p$. Any sequence of cubatures $\{(P_j^{(t)}, \omega_j^{(t)})\}_{j=1}^{n_t}$ with nonnegative weights for $\mathrm{Hom}_t(\mathcal{G}_{k,d})$, $t = 1, 2, \dots$, satisfies, for $f \in H_p^s(\mathcal{G}_{k,d})$,*

$$\Big| \int_{\mathcal{G}_{k,d}} f(P) \mathrm{d}\mu_{k,d}(P) - \sum_{j=1}^{n_t} \omega_j^{(t)} f(P_j^{(t)}) \Big| \lesssim t^{-s} \|f\|_{H_p^s}.$$

Let us connect the cardinality n_t of the cubature sequence with the strength t. The lower bound $n_t \gtrsim t^{k(d-k)}$ follows from results in [23] that also relate to our Proposition 2. Therefore, the best we can hope for is $n_t \asymp t^{k(d-k)}$:

Definition 1. We call a sequence of cubatures $\{(P_j^{(t)}, \omega_j^{(t)})\}_{j=1}^{n_t}$ of strength t with $t = 1, 2, \dots$, satisfying

$$n_t \asymp t^{k(d-k)} \tag{18}$$

with $n_t \to \infty$, a *low-cardinality cubature sequence*.

Remark 2. For any $t = 1, 2, \ldots$, there exist cubatures $\{(P_j^{(t)}, \omega_j^{(t)})\}_{j=1}^{n_t}$ for $\mathrm{Hom}_t(\mathcal{G}_{k,d})$ with positive weights satisfying (18), cf. [23]. Grassmannian t-designs are cubatures for $\mathrm{Hom}_t(\mathcal{G}_{k,d})$ with constant weights $\omega_j = 1/n$, for $j = 1, \ldots, n$. According to [33] indeed there do exist Grassmannian t-designs that are low-cardinality cubature sequence, i.e., satisfying (18).
For low-cardinality cubature sequences, Theorem 5 leads to

$$\left| \int_{\mathcal{G}_{k,d}} f(P) \mathrm{d}\mu_{k,d}(P) - \sum_{j=1}^{n_t} \omega_j^{(t)} f(P_j^{(t)}) \right| \lesssim n_t^{-\frac{s}{k(d-k)}} \|f\|_{H_p^s}. \tag{19}$$

We should point out that analogous results for the sphere are contained in [13, 14]. By comparing Proposition 3 with Theorem 5 for $p = 2$, we observe that the condition $s > k(d-k)/2$ yields that the cubature points' error rate $n_t^{-\frac{s}{k(d-k)}}$ is better than the one for the random points. Given any sequence of points of cardinality n_t, it is noteworthy that the rate $n_t^{-\frac{s}{k(d-k)}}$ cannot be improved, cf. [12].

5.3 Cubatures for Function Approximation

The basic idea for applying cubature points in function approximation is quite simple. The standard expansion of any $f \in L_2(\mathcal{G}_{k,d})$ in the orthogonal basis $\{\varphi_\pi\}_{\ell(\pi)\le k}$ yields

$$f = \sum_{\ell(\pi)\le k} \langle f, \varphi_\pi \rangle \varphi_\pi \approx \sum_{\substack{|\pi|\le t \\ \ell(\pi)\le k}} \langle f, \varphi_\pi \rangle \varphi_\pi, \tag{20}$$

where the approximation is simply derived by truncating the infinite series at $|\pi| \le t$. The inner product $\langle f, \varphi_\pi \rangle$ is an integral that we approximate by the concept of cubatures, i.e., the error for approximating the integral by a finite sum is steered by (19):

$$\begin{aligned} \sum_{\substack{|\pi|\le t \\ \ell(\pi)\le k}} \langle f, \varphi_\pi \rangle \varphi_\pi &= \sum_{\substack{|\pi|\le t \\ \ell(\pi)\le k}} \int_{\mathcal{G}_{k,d}} f(P) \varphi_i(P) \mathrm{d}\sigma_{k,d}(P) \varphi_i \\ &\approx \sum_{\substack{|\pi|\le t \\ \ell(\pi)\le k}} \sum_{j=1}^{n} \omega_j f(P_j) \varphi_\pi(P_j) \varphi_\pi \\ &= \sum_{j=1}^{n} \omega_j f(P_j) \sum_{\substack{|\pi|\le t \\ \ell(\pi)\le k}} \varphi_\pi(P_j) \varphi_\pi. \end{aligned} \tag{21}$$

If we define $K_t(P,R) = \sum_{\substack{|\pi|\leq t \\ \ell(\pi)\leq k}} \varphi_\pi(P)\varphi_\pi(R)$, then we derive the approximation

$$f \approx \sum_{j=1}^{n} \omega_j f(P_j) K_t(P_j, \cdot). \tag{22}$$

The right-hand side of (22) is composed of two separate approximations, truncation of the series (20) and the approximation of the integral via cubatures (21). To obtain suitable error rates, it turns out that we better replace the sharp truncation by a smoothed version, i.e., we define the kernel K_t on $\mathcal{G}_{k,d} \times \mathcal{G}_{k,d}$ by

$$K_t(P,Q) = \sum_{\ell(\pi)\leq k} h(t^{-2}\lambda_\pi)\varphi_\pi(P)\varphi_\ell(Q), \tag{23}$$

where $h : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is an infinitely often differentiable and nonincreasing function with $h(x) = 1$, for $x \leq 1/2$, and $h(x) = 0$, for $x \geq 1$. The smoothed series truncation becomes the expression

$$\sigma_t(f) := \int_{\mathcal{G}_{k,d}} f(P) K_t(P, \cdot) \mathrm{d}\mu_{k,d}(P), \tag{24}$$

and $\sigma_t(f)$ approximates f with an error rate that matches the ones in Theorem 5:

Theorem 6 ([47]). *If $f \in H_p^s(\mathcal{G}_{k,d})$, then*

$$\|f - \sigma_t(f)\|_{L_p} \lesssim t^{-s} \|f\|_{H_p^s}.$$

To approximate f from finitely many samples, we combine the smoothed truncation with cubature points to replace the integral by a finite sum. For sample points $\{P_j\}_{j=1}^n \subset \mathcal{G}_{k,d}$ and weights $\{\omega_j\}_{j=1}^n$, we define

$$\sigma_t(f, \{(P_j, \omega_j)\}_{j=1}^n) := \sum_{j=1}^{n} \omega_j f(P_j) K_t(P_j, \cdot) \tag{25}$$

which coincides with (22) but with the kernel K_t from (23). Note that we must now consider functions f in Bessel potential spaces, for which point evaluation makes sense. The term $\sigma_{r(t)}(f, \{(P_j, \omega_j)\}_{j=1}^n)$ is contained in $\mathrm{Hom}_t(\mathcal{G}_{k,d})$ for $r(t) = \sqrt{\lceil \frac{4}{k} t^2 \rceil}$, cf. [16, Theorem 5]. The following approximation is a consequence of [47, Proposition 5.3]:

Theorem 7 ([16]). *If* $\{(P_j^{(t)}, \omega_j^{(t)})\}_{j=1}^{n_t}$ *is a sequence of cubatures with nonnegative weights for* $\mathrm{Hom}_{2t}(\mathcal{G}_{k,d})$, $t = 1, 2, \ldots$, *then, for* $f \in H^s_\infty(\mathcal{G}_{k,d})$,

$$\|f - \sigma_{r(t)}(f, \{(P_j^{(t)}, \omega_j^{(t)})\}_{j=1}^{n_t})\|_{L_\infty} \lesssim t^{-s}(\|f\|_{L_\infty} + \|f\|_{H^s_\infty}), \tag{26}$$

where $r(t) = \sqrt{\lceil \frac{4}{k} t^2 \rceil}$.
For low-cardinality cubatures, the inequality becomes

$$\|f - \sigma_{r(t)}(f, \{(P_j^{(t)}, \omega_j^{(t)})\}_{j=1}^{n_t})\|_{L_\infty} \lesssim n_t^{-\frac{s}{k(d-k)}}(\|f\|_{L_\infty} + \|f\|_{H^s(L_\infty)}), \tag{27}$$

so that we obtain error rates similar to (19).

5.4 Cubatures as Efficient Coverings

We have seen in the previous sections that cubatures relate to the approximation of integrals and are also useful to approximate functions from samples. Intuitively, good samplings for approximation need to cover the underlying space sufficiently well. Indeed, we shall connect cubatures with asymptotically optimal coverings.

Given any finite collection of points $\{P_j\}_{j=1}^n \subset \mathcal{G}_{k,d}$, we define the *covering radius* ρ by

$$\rho := \rho(\{P_j\}_{j=1}^n) := \sup_{P \in \mathcal{G}_{k,d}} \min_{1 \le j \le n} \|P - P_j\|,$$

where $\|\cdot\|$ denotes the Frobenius norm on the space of symmetric matrices. Note that the covering radius is simply the radius of the largest hole. Let $B_r(P)$ denote the closed ball of radius r centered at $P \in \mathcal{G}_{k,d}$. Since

$$\mathcal{G}_{k,d} = \bigcup_{j=1}^n B_\rho(P_j)$$

and $\mu_{k,d}(B_r(P)) \asymp r^{k(d-k)}$, for $P \in \mathcal{G}_{k,d}$ with $0 < r \le 1$, we deduce

$$1 = \mu_{k,d}(\mathcal{G}_{k,d}) \le \sum_{j=1}^n \mu_{k,d}(B_\rho(P_j)) \lesssim n\rho^{k(d-k)},$$

which leads to the lower bound $n^{-\frac{1}{k(d-k)}} \lesssim \rho$. Point sequences in $\mathcal{G}_{k,d}$ that match this lower bound asymptotically in n are referred to as asymptotically optimal coverings.

Theorem 8 ([15]). *Any low-cardinality cubature sequence* $\{(P_j^{(t)}, \omega_j^{(t)})\}_{j=1}^{n_t}$ *for* $\mathrm{Hom}_t(\mathcal{G}_{k,d})$ *with positive weights is covering asymptotically optimal, i.e., its covering radius* $\rho^{(t)}$ *satisfies* $n^{-\frac{1}{k(d-k)}} \asymp \rho^{(t)}$.

Analogous results for the sphere are contained in [14]. Theorem 7 with (27) and Theorem 8 show the efficiency of low-cardinality cubature points in approximation and covering.

5.5 Cubatures for Phase Retrieval

To reflect the versatility of Grassmannian cubatures, we now briefly discuss their use in phase retrieval. The problem of reconstructing vectors from phaseless magnitude measurements has attracted great attention in the recent literature; see, e.g., [7, 17, 18, 20, 25]. For $x \in \mathbb{R}^d$, the mapping

$$\hat{x} : \mathcal{G}_{k,d} \to \mathbb{R}, \quad P \mapsto \|Px\|^2$$

is a homogeneous polynomial of degree 2, hence contained in $\mathrm{Hom}_2(\mathcal{G}_{k,d})$. Notice that $\hat{x}$ is in a one-to-one correspondence with the rank-one matrix $xx^\top$ since

$$\hat{x}(P) = x^\top P x = \mathrm{trace}(Pxx^\top).$$

The problem of reconstructing $xx^\top$ from finitely many samples $\{\hat{x}(P_j)\}_{j=1}^n$, where $\{P_j\}_{j=1}^n \subset \mathcal{G}_{k,d}$, is known as the phase retrieval problem. Most publications deal with $k = 1$. For $k > 1$, we refer to [3, 6, 28, 30, 31] and references therein.

If there are weights $\{\omega_j\}_{j=1}^n$ such that $\{(P_j, \omega_j)\}_{j=1}^n$ is a cubature for $\mathrm{Hom}_2(\mathcal{G}_{k,d})$, then $xx^\top$ can be directly reconstructed via the closed formula

$$xx^\top = \frac{d}{k} \sum_{j=1}^n \omega_j \hat{x}(P_j) \Big[\frac{1}{\alpha} \sum_{j=1}^n \omega_j \hat{x}(P_j) P_j - \frac{\beta}{\alpha} I_d \Big], \tag{28}$$

where $\alpha = \frac{2k(d-k)}{d(d+2)(d-1)}$ and $\beta = \frac{k(kd+k-2)}{d(d+2)(d-1)}$, cf. [3]. However, cubatures for $\mathrm{Hom}_2(\mathcal{G}_{k,d})$ must have at least $d(d+1)/2$ many points. Thus, the number of samples grows quadratic with the ambient dimension d. We are seeking reconstruction from fewer samples at the expense of replacing the closed reconstruction formula with a feasibility problem of a semidefinite program. We consider the problem

$$\text{find } A \in \mathbb{R}^{d\times d}_{\succeq 0}, \quad \text{subject to} \quad \mathrm{trace}(P_j A) = \hat{x}(P_j),\ j = 1, \ldots, n, \tag{29}$$

where $\mathbb{R}^{d\times d}_{\succeq 0}$ denotes symmetric, positive semidefinite matrices in $\mathbb{R}^{d\times d}$.

Theorem 9 ([3]). *There are constants $c_1, c_2 > 0$ such that the following holds: if $n \geq c_1 d$ and $\{P_j\}_{j=1}^n \subset \mathcal{G}_{k,d}$ are chosen independently identically distributed according to $\mu_{k,d}$, then the matrix $xx^\top$ is the unique solution to (29) with probability $1 - e^{-c_2 n}$, for all $x \in \mathbb{R}^d$.*

This theorem generalizes results in [17, 18] from $k = 1$ to $k \geq 1$. If the projectors $\{P_j\}_{j=1}^n$ are sampled from the idealized perfect cubature $\mu_{k,d}$, the number of needed samples grows linearly with d. Next, we shall find a balance between the deterministic cubatures required for (28) and the full randomness invoked by $\mu_{k,d}$ used in Theorem 9.

From here on, we suppose that the length $\|x\|$ is known to us. To simplify notation, we make the convention that $P_0 = I_d$, hence $\langle xx^*, P_0 \rangle = \operatorname{trace}(xx^*) = \|x\|^2$; consider the problem

$$\text{find } A \in \mathbb{R}^{d\times d}_{\succeq 0}, \quad \text{subject to} \quad \operatorname{trace}(AP_j) = \hat{x}(P_j),\ j = 0, \ldots, n, \tag{30}$$

where $\{P_j\}_{j=1}^n \subset \mathcal{G}_{k,d}$ and $\hat{x}(I_d) := \|x\|^2$. For $k = 1$, the following result is essentially due to [39]. The extension to $k \geq 1$ has been derived in [31]:

Theorem 10 ([31]). *Suppose that $\|x\|^2$ is known and that $\{(\mathcal{P}_j, \omega_j)\}_{j=1}^n$ is a cubature with nonnegative weights for $\operatorname{Hom}_t(\mathcal{G}_{k,d})$, $t \geq 3$. Let $\mu = \sum_{j=1}^n \omega_j \delta_{\mathcal{P}_j}$ denote the corresponding discrete probability measure, where $\delta_{\mathcal{P}_j}$ is the point measure in $\mathcal{P}_j$. If $\{P_j\}_{j=1}^n \subset \mathcal{G}_{k,d}$ are independently sampled from μ, then with probability at least $1 - e^{-\gamma}$, the rank-one matrix xx^* is the unique solution to (30) provided that*

$$n \geq c_1 \gamma t d^{1+2/t} \log^2(d), \tag{31}$$

where $\gamma \geq 1$ is an arbitrary parameter and c_1 is a constant, which does not depend on d.

Hence, choosing random projectors distributed according to discrete probability measures allows us to reconstruct $xx^\top$ with less than d^2 many measurements.

6 Cubatures of Varying Ranks

In the previous sections, we were dealing with cubatures for Grassmannians of fixed rank. In order to allow more flexibility, we now aim to remove this restriction, i.e., we shall investigate cubatures for unions of Grassmannians. Our aim is to provide elementary proofs of some results in [29] that were derived by the use of representation theoretic concepts.

Given a non-empty set $\mathcal{K} \subset \{1, \ldots, d-1\}$, we define the corresponding union of Grassmannians by

$$\mathcal{G}_{\mathcal{K},d} := \bigcup_{k\in\mathcal{K}} \mathcal{G}_{k,d} = \{P \in \mathbb{R}^{d\times d}_{\mathrm{sym}} \,:\, P^2 = P,\ \operatorname{trace}(P) \in \mathcal{K}\}.$$

As for a single Grassmannian, the polynomials on $\mathcal{G}_{\mathcal{K},d}$ are given by multivariate polynomials in the matrix entries of a given projector $P \in \mathcal{G}_{\mathcal{K},d}$, i.e.,

$$\mathrm{Pol}_t(\mathcal{G}_{\mathcal{K},d}) := \{f|_{\mathcal{G}_{\mathcal{K},d}} : f \in \mathrm{Pol}_t(\mathbb{R}^{d\times d}_{\mathrm{sym}})\}. \tag{32}$$

The dimension of $\mathrm{Pol}_t(\mathcal{G}_{\mathcal{K},d})$ is an indicator of the number of points needed to obtain a cubature on $\mathcal{G}_{\mathcal{K},d}$, cf. Proposition 2 and [23]. To compute this dimension, we shall first derive a lower bound:

Proposition 4. *Let $\mathcal{K} = \{k_i\}_{i=1}^r \subset \{1,\dots,d-1\}$ and $t \in \mathbb{N}_0$ be given such that*

$$\min\{k_1, d-k_1\} \geq \cdots \geq \min\{k_r, d-k_r\}. \tag{33}$$

Then it holds

$$\dim(\mathrm{Pol}_t(\mathcal{G}_{\mathcal{K},d})) \geq \sum_{i=1}^{s} \dim(\mathrm{Pol}_{t-i+1}(\mathcal{G}_{k_i,d})), \quad s := \min\{t+1, |\mathcal{K}|\}. \tag{34}$$

Note that the dimension of each $\mathrm{Pol}_{t-i+1}(\mathcal{G}_{k_i,d})$ is known, i.e.,

$$\dim(\mathrm{Pol}_t(\mathcal{G}_{k,d})) = \sum_{\substack{|\pi|\leq t,\\ \ell(\pi)\leq \min\{k,d-k\}}} \mathcal{D}(d, 2\pi), \tag{35}$$

where

$$\mathcal{D}(d,\pi) = \prod_{1\leq i<j\leq \frac{d}{2}} \frac{(l_i+l_j)(l_i-l_j)}{(j-i)(d-i-j)} \cdot \begin{cases} \prod_{1\leq i\leq \frac{d}{2}} \frac{2l_i}{d-2i}, & d \text{ odd}, \\ 2, & d \text{ even and } \pi_{\lfloor \frac{d}{2} \rfloor} > 0, \\ 1, & d \text{ even and } \pi_{\lfloor \frac{d}{2} \rfloor} = 0, \end{cases} \tag{36}$$

with $l_i := \frac{d}{2} + \pi_i - i$, for $1 \leq i \leq \frac{d}{2}$, cf. [35, Formulas (24.29) and (24.41)] and [4, 5]. Thus, (34) is an explicit lower bound on the dimension of $\mathrm{Pol}_t(\mathcal{G}_{\mathcal{K},d})$.

Proof (of Proposition 4). We will show that the lower bound (34) is valid for any ordering of the indices $k_1,\dots,k_r$. In particular it holds for the ordering specified in (33), which maximizes the right-hand side over all such lower bounds.

For $t = 0$ or $r = 1$, the sum in (34) reduces to a single term, so that the lower bound indeed holds. For fixed $t \geq 1$, we verify the general case by induction over r, where we proceed from $r-1$ to r with $r \geq 2$.

Choose $\{f_i\}_{i=1}^m \subset \mathrm{Pol}_t(\mathbb{R}^{d\times d}_{\mathrm{sym}})$ and $\{g_j\}_{j=1}^n \subset \mathrm{Pol}_{t-1}(\mathbb{R}^{d\times d}_{\mathrm{sym}})$ such that $\{f_i|_{\mathcal{G}_{k_1,d}}\}_{i=1}^m$ and $\{g_j|_{\mathcal{G}_{\mathcal{K}\setminus\{k_1\},d}}\}_{j=1}^n$ are bases for the spaces $\mathrm{Pol}_t(\mathcal{G}_{k_1,d})$ and $\mathrm{Pol}_{t-1}(\mathcal{G}_{\mathcal{K}\setminus\{k_1\},d})$, respectively. We infer that any linear combination

$$h := \sum_{i=1}^{m} \alpha_i f_i|_{\mathcal{G}_{\mathcal{K},d}} + \sum_{j=1}^{n} \beta_j \big(\mathrm{trace}(\cdot) - k_1\big) g_j|_{\mathcal{G}_{\mathcal{K},d}}$$

is contained in $\mathrm{Pol}_t(\mathcal{G}_{\mathcal{K},d})$. Suppose now that h vanishes on $\mathcal{G}_{\mathcal{K},d}$. In particular, h vanishes on $\mathcal{G}_{k_1,d}$, so that $\alpha_i = 0$, $i = 1,\dots,m$. Vanishing on $\mathcal{G}_{\mathcal{K}\setminus\{k_1\},d}$ implies $\beta_j = 0, j = 1,\dots,n$. Hence, the function system

$$\{f_i|_{\mathcal{G}_{\mathcal{K},d}}\}_{i=1}^m \cup \{(\mathrm{trace}(\cdot) - k_1)g_j|_{\mathcal{G}_{\mathcal{K},d}}\}_{j=1}^n$$

is linearly independent in $\mathrm{Pol}_t(\mathcal{G}_{\mathcal{K},d})$. By using $(s-1) = \min\{t, r-1\}$, we infer by the induction hypothesis

$$\begin{aligned}
\dim(\mathrm{Pol}_t(\mathcal{G}_{\mathcal{K},d})) &\geq \dim(\mathrm{Pol}_t(\mathcal{G}_{k_1,d})) + \dim(\mathrm{Pol}_{t-1}(\mathcal{G}_{\mathcal{K}\setminus\{k_1\},d})) \\
&\geq \dim(\mathrm{Pol}_t(\mathcal{G}_{k_1,d})) + \sum_{i=1}^{(s-1)} \dim(\mathrm{Pol}_{(t-1)-i+1}(\mathcal{G}_{k_{i+1},d})) \\
&= \dim(\mathrm{Pol}_t(\mathcal{G}_{k_1,d})) + \sum_{i=2}^{s} \dim(\mathrm{Pol}_{t-i+1}(\mathcal{G}_{k_i,d})) \\
&= \sum_{i=1}^{s} \dim(\mathrm{Pol}_{t-i+1}(\mathcal{G}_{k_i,d})),
\end{aligned}$$

which proves the lower bound (34).
In the case $\mathcal{K} = \{k, d-k\}$, we can verify that the lower bound is matched by elementary methods:

Proposition 5. *Let $1 \leq k \leq d-1$ with $k \neq \frac{d}{2}$ and $t \geq 1$. Then it holds*

$$\mathrm{Pol}_t(\mathcal{G}_{k,d} \cup \mathcal{G}_{d-k,d}) \cong \mathrm{Pol}_t(\mathcal{G}_{k,d}) \oplus \mathrm{Pol}_{t-1}(\mathcal{G}_{d-k,d}). \tag{37}$$

Proof. We consider the restriction mapping

$$|_{\mathcal{G}_{k,d}} \;:\; \mathrm{Pol}_t(\mathcal{G}_{k,d} \cup \mathcal{G}_{d-k,d}) \longrightarrow \mathrm{Pol}_t(\mathcal{G}_{k,d}), \quad f \mapsto f|_{\mathcal{G}_{k,d}}$$

and shall verify that the dimension of its null-space satisfies

$$\mathrm{null}(|_{\mathcal{G}_{k,d}}) = (\mathrm{trace}(\cdot) - k)\,\mathrm{Pol}_{t-1}(\mathcal{G}_{k,d} \cup \mathcal{G}_{d-k,d}). \tag{38}$$

Since $|_{\mathcal{G}_{k,d}}$ is onto and $(\mathrm{trace}(\cdot) - k)\,\mathrm{Pol}_{t-1}(\mathcal{G}_{k,d} \cup \mathcal{G}_{d-k,d})$ is equivalent to $\mathrm{Pol}_{t-1}(\mathcal{G}_{d-k,d})$, this would imply (37).

It is obvious that the right-hand side in (38) is contained in $\mathrm{null}(|_{\mathcal{G}_{k,d}})$. The latter can also be deduced from the lower bounds (34). For the reverse set inclusion, let $f \in \mathrm{null}(|_{\mathcal{G}_{k,d}})$. We must now check that $f|_{\mathcal{G}_{d-k,d}} \in \mathrm{Pol}_{t-1}(\mathcal{G}_{d-k,d})$.

To proceed let us denote $n := \dim(\mathrm{Pol}_t(\mathcal{G}_{k,d} \cup \mathcal{G}_{d-k,d}))$. According to [29] (see also [2]), there are $\{X_j\}_{j=1}^n \subset \mathcal{G}_{k,d} \cup \mathcal{G}_{d-k,d}$ and $\{c_j\}_{j=1}^n \subset \mathbb{R}$ such that

$$f(P) = \sum_{j=1}^{n} c_j \,\mathrm{trace}(X_j P)^t|_{\mathcal{G}_{k,d} \cup \mathcal{G}_{d-k,d}}, \quad P \in \mathcal{G}_{k,d} \cup \mathcal{G}_{d-k,d}.$$

By applying the binomial formula, we observe that

$$f + (-1)^{t+1} f(I_d - \cdot) \in \mathrm{Pol}_{t-1}(\mathcal{G}_{k,d} \cup \mathcal{G}_{d-k,d}). \tag{39}$$

Therefore, the assumption $f|_{\mathcal{G}_{k,d}} \equiv 0$ implies that $f(I - \cdot)|_{\mathcal{G}_{k,d}} \in \mathrm{Pol}_{t-1}(\mathcal{G}_{k,d})$. Since $f \mapsto f(I - \cdot)$ is an isomorphism between $\mathrm{Pol}_{t-1}(\mathcal{G}_{k,d})$ and $\mathrm{Pol}_{t-1}(\mathcal{G}_{d-k,d})$, we derive $f|_{\mathcal{G}_{d-k,d}} \in \mathrm{Pol}_{t-1}(\mathcal{G}_{d-k,d})$. Thus, we have verified (38), which concludes the proof. Proposition 5 shows that for $\mathcal{K} = \{k, d-k\}$ the inequality in Proposition 4 becomes an equality. It turns out that equality holds in the general cases as well:

Theorem 11 ([29]). *Let $\mathcal{K} = \{k_i\}_{i=1}^r \subset \{1, \ldots, d-1\}$ and $t \in \mathbb{N}_0$ be given such that*

$$\min\{k_1, d - k_1\} \geq \cdots \geq \min\{k_r, d - k_r\}.$$

Then it holds

$$\mathrm{Pol}_t(\mathcal{G}_{\mathcal{K},d}) \cong \bigoplus_{i=1}^{s} \mathrm{Pol}_{t-i+1}(\mathcal{G}_{k_i,d}), \quad s := \min\{t+1, |\mathcal{K}|\}. \tag{40}$$

Compared to our elementary proofs of Propositions 4 and 5, the proof of Theorem 11 presented in [29] is more involved, using representation theoretic concepts in combination with orthogonally invariant reproducing kernel decompositions.

Note that Theorem 11 implies that each $f \in \mathrm{Pol}_t(\mathcal{G}_{\mathcal{K},d})$, which vanishes on $\mathrm{Pol}_t(\mathcal{G}_{k_1,d})$, must contain a factor $(\mathrm{trace}(\cdot) - k_1)\big|_{\mathcal{G}_{\mathcal{K},d}}$, i.e., the restriction mapping $|_{\mathcal{G}_{k_1,d}}$ from $\mathrm{Pol}_t(\mathcal{G}_{\mathcal{K},d})$ to $\mathrm{Pol}_t(\mathcal{G}_{k_1,d})$, for $t \geq 1$, satisfies

$$\mathrm{null}(|_{\mathcal{G}_{k_1,d}}) = (\mathrm{trace}(\cdot) - k_1)\,\mathrm{Pol}_{t-1}(\mathcal{G}_{\mathcal{K},d}),$$

cf. [29].

Our better understanding of the space $\mathrm{Pol}_t(\mathcal{G}_{\mathcal{K},d})$ enables the study of cubatures on unions of Grassmannians in the areas of the previous sections. This shall be addressed in future work.

Acknowledgements Thomas Peter was funded by the German Academic Exchange Service (DAAD) through P.R.I.M.E. 57338904. All authors have been supported by the Vienna Science and Technology Fund (WWTF) through project VRG12-009.

References

1. D. Achlioptas, Database-friendly random projections: Johnson-lindenstrauss with binary coins. J. Comput. Syst. Sci. **66**(4), 671–687 (2003)
2. C. Bachoc, M. Ehler, Tight *p*-fusion frames. Appl. Comput. Harmon. Anal. **35**(1), 1–15 (2013)
3. C. Bachoc, M. Ehler, Signal reconstruction from the magnitude of subspace components. IEEE Trans. Inf. Theory **61**(7), 1–13 (2015)
4. C. Bachoc, R. Coulangeon, G. Nebe, Designs in Grassmannian spaces and lattices. J. Algebr. Combin. **16**, 5–19 (2002)
5. C. Bachoc, E. Bannai, R. Coulangeon, Codes and designs in Grassmannian spaces. Discret. Math. **277**, 15–28 (2004)
6. S. Bahmanpour, J. Cahill, P.G. Casazza, J. Jasper, L.M. Woodland, Phase retrieval and norm retrieval, in *Trends in Harmonic Analysis and Its Applications*. Contemporary Mathematics, vol. 650 (American Mathematical Society, Providence, RI, 2015), pp. 3–14
7. A.S. Bandeira, J. Cahill, D.G. Mixon, A.A. Nelson, Saving phase: injectivity and stability for phase retrieval. Appl. Comput. Harmon. Anal. **37**(1), 106–125 (2014)
8. G. Baron, De prony, "essai expérimental et analytique sur les lois de la dilatabilité des fluides élastique et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, à différentes températures". J. de l'École Polytechnique **1**(2), 24–76 (1795)
9. D.J. Bates, J.D. Hauenstein, A.J. Bellommese, C.W. Wampler, *Numerically Solving Polynomial Systems with Bertini*, vol. 25 (SIAM, Philadelphia, PA, 2013)
10. R. Beinert, G. Plonka, Sparse phase retrieval of one-dimensional signals by Prony's method. Front. Appl. Math. Stat. **3**(5) (2017)
11. B. Bodman, M. Ehler, M. Gräf, From low to high-dimensional moments without magic. J. Theor. Probab. (2017). https://doi.org/10.1007/s10959-017-0785-x
12. L. Brandolini, C. Choirat, L. Colzani, G. Gigante, R. Seri, G. Travaglini, Quadrature rules and distribution of points on manifolds, Annali della Scuola Normale Superiore di Pisa - Classe di Scienze **13**(4), 889–923 (2014)
13. J. Brauchart, E. Saff, I.H. Sloan, R. Womersley, QMC designs: optimal order quasi Monte Carlo integration schemes on the sphere. Math. Comput. **83**, 2821–2851 (2014)
14. J.S. Brauchart, J. Dick, E.B. Saff, I.H. Sloan, Y.G. Wang, R.S. Womersley, Covering of spheres by spherical caps and worst-case error for equal weight cubature in Sobolev spaces. J. Math. Anal. Appl. **431**(2), 782–811 (2015)
15. A. Breger, M. Ehler, M. Gräf, Points on manifolds with asymptotically optimal covering radius (2016). arXiv: 1607.06899
16. A. Breger, M. Ehler, M. Gräf, Quasi Monte Carlo integration and kernel-based function approximation on Grassmannians, in *Frames and Other Bases in Abstract and Function Spaces: Novel Methods in Harmonic Analysis*, vol. 1 (Birkhauser/Springer, Berlin, 2017)
17. 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)
18. E.J. Candès, T. Strohmer, V. Voroninski, Phaselift: exact and stable signal recovery from magnitude measurements via convex programming. Commun. Pure Appl. Math. **66**(8), 1241–1274 (2013)
19. Y. Chikuse, *Statistics on Special Manifolds*. Lecture Notes in Statistics (Springer, New York, 2003)
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 (2015)
21. S. Dasgupta, A. Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algoritm. **22**(1), 60–65 (2003)
22. C. de Boor, K. Hölling, S.D. Riemenschneider, *Box Splines* (Springer, New York, 1993)
23. P. de la Harpe, C. Pache, Cubature formulas, geometrical designs, reproducing kernels, and Markov operators, in *Infinite Groups: Geometric, Combinatorial and Dynamical Aspects (Basel)*, vol. 248 (Birkhäuser, Basel, 2005), pp. 219–267

24. P. Delsarte, J.M. Goethals, J.J. Seidel, Spherical codes and designs. Geom. Dedicata **6**, 363–388 (1977)
25. L. Demanet, P. Hand, Stable optimizationless recovery from phaseless linear measurements. J. Fourier Anal. Appl. **20**, 199–221 (2014)
26. R.A. DeVore, G.G. Lorentz, *Constructive Approximation* (Springer, Berlin, 1993)
27. I. Dumitriu, A. Edelman, G. Shuman, MOPS: multivariate orthogonal polynomials (symbolically). J. Symb. Comput. **42**(6), 587–620 (2007)
28. D. Edidin, Projections and phase retrieval. Appl. Comput. Harmon. Anal. **42**(2), 350–359 (2017)
29. M. Ehler, M. Gräf, Reproducing kernels for the irreducible components of polynomial spaces on unions of Grassmannians (2017). arXiv:1411.5865
30. M. Ehler, M. Fornasier, J. Sigl, Quasi-linear compressed sensing. SIAM Multiscale Model. Simul. **12**(2), 725–754 (2014)
31. M. Ehler, F. Kiraly, M. Gräf, Phase retrieval using cubatures of positive semidefinite matrices. Waves, Wavelets and Fractals - Adv. Anal. **1**(1), 32–50 (2015)
32. H. Engels, Numerical quadrature and cubature, in *Computational Mathematics and Applications* (Academic Press, London, 1980)
33. U. Etayo, J. Marzo, J. Ortega-Cerdà, Asymptotically optimal designs on compact algebraic manifolds, arXiv: 1612.06729 (2016)
34. F. Filbir, H.N. Mhaskar, A quadrature formula for diffusion polynomials corresponding to a generalized heat kernel. J. Fourier Anal. Appl. **16**(5), 629–657 (2010)
35. W. Fulton, J. Harris, *Representation Theory, A First Course* (Springer, Berlin, 1991)
36. D. Geller, I.Z. Pesenson, Band-limited localized Parseval frames and Besov spaces on compact homogeneous manifolds. J. Geom. Anal. **21**(2), 334–371 (2011)
37. M. Gräf, *Efficient Algorithms for the Computation of Optimal Quadrature Points on Riemannian Manifolds* (Universitätsverlag Chemnitz, Chemnitz, 2013)
38. K. Gross, D. St. P. Richards, Special functions of matrix argument. I: algebraic induction, zonal polynomials and hypergeometric functions. Trans. Am. Math. Soc. **301**, 781–811 (1987)
39. D. Gross, F. Krahmer, R. Kueng, A partial derandomization of PhaseLift using spherical designs. J. Fourier Anal. Appl. **21**(2), 229–266 (2015)
40. M.T. Harandi, C. Sanderson, S. Shirazi, B.C. Lovell, Kernel analysis on Grassmann manifolds for action recognition. Pattern Recogn. Lett. **34**(15), 1906–1915 (2013)
41. S.G. Hoggar, *t*-designs in projective spaces. Eur. J. Comb. **3**, 233–254 (1982)
42. A.T. James, A.G. Constantine, Generalized Jacobi polynomials as spherical functions of the Grassmann manifold. Proc. Lond. Math. Soc. **29**(3), 174–192 (1974)
43. H. König, Cubature formulas on spheres. Adv. Multivar. Approx. Math. Res. **107**, 201–211 (1999)
44. S. Kunis, T. Peter, T. Römer, U. von der Ohe, A multivariate generalization of Prony's method. Linear Algebra Appl. **490**, 31–47 (2016)
45. M. Maggioni, H.N. Mhaskar, Diffusion polynomial frames on metric measure spaces. Appl. Comput. Harmon. Anal. **24**(3), 329–353 (2008)
46. J. Matousek, On variants of the Johnson-Lindenstrauss lemma. Random Struct. Algoritm. **33**(2), 142–156 (2008)
47. H.N. Mhaskar, Eignets for function approximation on manifolds. Appl. Comput. Harmon. Anal. **29**, 63–87 (2010)
48. R.J. Muirhead, *Aspects of Multivariate Statistical Theory* (Wiley, New York, 1982)
49. A. Neumaier, J.J. Seidel, Discrete measures for spherical designs, eutactic stars and lattices. Indag. Math. **91**(3), 321–334 (1988)
50. E. Novak, H. Wozniakowski, *Tractability of Multivariate Problems. Volume II*. EMS Tracts in Mathematics, vol. 12 (EMS Publishing House, Zürich, 2010)
51. I.Z. Pesenson, D. Geller, Cubature formulas and discrete fourier transform on compact manifolds, in *From Fourier Analysis and Number Theory to Radon Transforms and Geometry*, vol. 28 (Springer, New York, 2012), pp. 431–453

52. T. Peter, G. Plonka, A generalized Prony method for reconstruction of sparse sums of eigenfunctions of linear operators. Inverse Prob. **29**(2), 025001 (2013)
53. G. Plonka, M. Tasche, Prony methods for recovery of structured functions. GAMM-Mitteilungen **37**(2), 239–258 (2014)
54. D. Potts, M. Tasche, Parameter estimation for exponential sums by approximate Prony method. Signal Process. **90**, 1631–1642 (2010)
55. D. Potts, M. Tasche, Parameter estimation for nonincreasing exponential sums by Prony-like methods. Linear Algebra Appl. **439**, 1024–1039 (2013)
56. A. Reznikov, E.B. Saff, The covering radius of randomly distributed points on a manifold. Int. Math. Res. Not. **2016**(19), 6065–6094 (2016)
57. R. Roy, A. Paulraj, T. Kailath, ESPRIT – A subspace rotation approach to estimation of parameters of cisoids in noise. IEEE Trans. Acoust. Speech Signal Process. **34**(5), 1340–1342 (1986)
58. R. Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)
59. P. Turaga, A. Veeraraghavan, A. Srivastava, R. Chellappa, Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition. IEEE Trans. Pattern Anal. Mach. Intell. **33**(11), 2273–2286 (2011)
60. M. Vetterli, P. Marziliano, T. Blu, Sampling signals with finite rate of innovation. IEEE Trans. Signal Process. **50**(6), 1417–1428 (2002)
61. I.B. Yaacov, The Vandermonde determinant identity in higher dimension (2017). arXiv: 1405.0993

A Randomized Tensor Train Singular Value Decomposition

Benjamin Huber, Reinhold Schneider, and Sebastian Wolf

Abstract The hierarchical SVD provides a quasi-best low-rank approximation of high-dimensional data in the hierarchical Tucker framework. Similar to the SVD for matrices, it provides a fundamental but expensive tool for tensor computations. In the present work, we examine generalizations of randomized matrix decomposition methods to higher-order tensors in the framework of the hierarchical tensor representation. In particular we present and analyze a randomized algorithm for the calculation of the hierarchical SVD (HSVD) for the tensor train (TT) format.

Keywords Tensors · Tensor train format · Tensor product approximation · Randomized algorithm · Random matrix · Randomized singular value decomposition · TT-SVD · TT-Decomposition · Randomized decomposition · Low-rank approximation

1 Introduction

Low-rank matrix decompositions, such as the singular value decomposition (SVD) and the QR decomposition, are principal tools in data analysis and scientific computing. For matrices with small rank, both decompositions offer a tremendous reduction in computational complexity and can expose the underlying problem structure. In recent years generalizations of these low-rank decompositions to higher-order tensors have proven to be very useful and efficient techniques as well. In particular the hierarchical Tucker [1] and the tensor train [2] format made quite an impact, as both formats allow to circumvent the notorious *curse of dimensionality*, i.e., the exponential scaling of the ambient spaces with respect to the order of the tensors. Applications of these formats are as various as high-dimensional PDE's

B. Huber • R. Schneider (✉) • S. Wolf
Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany
e-mail: schneidr@math.tu-berlin.de

H. Boche et al. (eds.), *Compressed Sensing and its Applications*, Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-69802-1_9

like the Fokker-Planck equations and the many particle Schrödinger equations, applications in neuroscience, graph analysis, signal processing, computer vision, and computational finance; see the extensive survey of Grasedyck et al. [3]. Also in a recent paper in machine learning, Cohen et al. [4] showed a connection between these tensor formats and deep neural networks and used this to explain the much higher power of expressiveness of deep neural networks over shallow ones.

One of the main challenges when working with these formats is the calculation of low-rank decompositions of implicitly or explicitly given tensors, i.e., the high-dimensional analog of the classical SVD calculation. For matrices there exists a wide range of methods, which allow these calculations with high efficiency and precision. One particular branch is randomized methods which appear often in the literature, mostly as efficient heuristics to calculate approximate decompositions. It was only recently that thanks to new results from *random matrix theory*, a rigorous analysis of these procedures became possible; see [5]. In this work we aim to extend some of these results for randomized matrix decompositions to the high-dimensional tensor case. To this end we present an algorithm which allows the efficient calculation of the tensor train SVD (TT-SVD) for general higher-order tensors. Especially for sparse tensors, this algorithm exhibits a superior complexity, scaling only linear in the order, compared to the exponential scaling of the naive approach. Extending the results of [5], we show that stochastic error bounds can also be obtained for these higher-order methods.

This work focuses on the theoretical and algorithmic aspects of this randomized (TT-)SVD. However a particular application on our mind is the work in [6, 7], where we treat the tensor completion problem. That is, in analogy to matrix completion (see, e.g., [8–10]), we want to reconstruct a tensor from N measurements using a low-rank assumption. We use an iterative (hard) thresholding procedure, which requires the (approximate) calculation of a low-rank decomposition in each iteration of the algorithm. As the deterministic TT-SVD is already a fundamental tool, there are of course many further possible applications for our randomized variant; see, for example, [11–13].

We start with a brief recap of tensor product spaces and introduce the notation used in the remainder of this work. In Section 2 we give an overview of different tensor decompositions, generalizing the singular value decomposition from matrices to higher-order tensors. In the second part, a detailed introduction of the tensor train format is provided. Section 3.1 summarizes results for randomized matrix decompositions which are important for this work. In Section 3.2 we introduce our randomized TT-SVD scheme and prove stochastic error bounds for this procedure. An interesting relation between the proposed algorithm and the popular alternating least squares (ALS) algorithm is examined in Section 4. Section 5 collects several numerical experiments showing the performance of the proposed algorithms. Section 6 closes with some concluding remarks.

1.1 Tensor Product Spaces

Let us begin with some preliminaries on tensors and tensor spaces. For an exhaustive introduction, we refer to the monograph of Hackbusch [14].

Given Hilbert spaces $V_1, \dots, V_d$, the tensor product space of order d

$$\mathscr{V} = \bigotimes_{i=1}^{d} V_i \,,$$

is defined as the closure of the span of all elementary tensor products of vectors from V_i, i.e.,

$$\mathscr{V} := \overline{\text{span } \{\mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \ldots \otimes \mathbf{v}_d \mid \mathbf{v}_i \in V_i\}} \,.$$

The elements $\mathbf{x} \in \mathscr{V}$ are called tensors of order d. If each space V_i is supplied with an orthonormal basis $\{\varphi^i_{\mu_i} : \mu_i \in \mathbb{N}\}$, then any $\mathbf{x} \in \mathscr{V}$ can be represented as

$$\mathbf{x} = \sum_{\mu_1=1}^{\infty} \dots \sum_{\mu_d=1}^{\infty} \mathbf{x}[\mu_1, \dots, \mu_d]\, \varphi^1_{\mu_1} \otimes \cdots \otimes \varphi^d_{\mu_d} \,.$$

Using this basis, with a slight abuse of notation, we can identify $\mathbf{x} \in \mathscr{V}$ with its representation by a d-variate function, often called hyper matrix,

$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_d) \mapsto \mathbf{x}[\mu_1, \dots, \mu_d] \in \mathbb{K} \,,$$

depending on discrete variables, usually called indices $\mu_i \in \mathbb{N}$. Analogous to vectors and matrices, we use square brackets $\mathbf{x}[\mu_1, \dots, \mu_d]$ to index the entries of this hypermatrix. Of course, the actual representation of $\mathbf{x} \in \mathscr{V}$ depends on the chosen bases φ^i of V_i. The index μ_i is said to correspond to the μ-th mode or equivalently the μ-th dimension of the tensor.

In the remainder of this article, we confine to finite dimensional real linear spaces $V_i := \mathbb{R}^{n_i}$; however most parts are easy to extend to the complex case as well. For these, the tensor product space

$$\mathscr{V} = \bigotimes_{i=1}^{d} \mathbb{R}^{n_i} = \mathbb{R}^{n_1 \times n_2 \times \ldots \times n_d} := \text{span } \{\mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \ldots \otimes \mathbf{v}_d \mid \mathbf{v}_i \in \mathbb{R}^{n_i}\}$$

is easily defined. If it is not stated explicitly, the $V_i = \mathbb{R}^{n_i}$ are supplied with the canonical basis $\{\mathbf{e}^i_1, \dots, \mathbf{e}^i_{n_i}\}$ of the vector spaces $\mathbb{R}^{n_i}$. Then every $\mathbf{x} \in \mathscr{V}$ can be represented as

$$\mathbf{x} = \sum_{\mu_1=1}^{n_1} \dots \sum_{\mu_d=1}^{n_d} \mathbf{x}[\mu_1, \dots, \mu_d]\, \mathbf{e}^1_{\mu_1} \otimes \cdots \otimes \mathbf{e}^d_{\mu_d} \,. \tag{1}$$

We equip the finite dimensional linear space $\mathscr{V}$ with the inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{\mu_1=1}^{n_1} \cdots \sum_{\mu_d=1}^{n_d} \mathbf{x}[\mu_1, \ldots, \mu_d]\, \mathbf{y}[\mu_1, \ldots, \mu_d] \,.$$

and the corresponding l_2-norm $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. We use the fact that for a Hilbert space V_i, the dual space V^* is isomorphic to V_i and use the identification $V \simeq V^*$. For the treatment of reflexive Banach spaces, we refer to [15, 16].

The number of possibly non-zero entries in the representation of $\mathbf{x}$ is $n_1 \cdots n_d = \Pi_{i=1}^{d} n_i$, and with $n = \max\{n_i : i = 1, \ldots, d\}$, the dimension of the space $\mathscr{V}$ scales exponentially in d, i.e., $\mathscr{O}(n^d)$. This is often referred to as the *curse of dimensions* and presents the main challenge when working with higher-order tensors.

1.2 Tensor Contractions and Diagrammatic Notation

Important concepts for the definitions of tensor decompositions are so-called matricizations and contractions introduced in this section.

The matricization or flattening of a tensor is the reinterpretation of the given tensor as a matrix, by combining a subset of modes to a single mode and combining the remaining modes to a second one.

Definition 1 (Matricization or Flattening)**.** Let $[n] = \{1, 2, \ldots, n\}$ and $\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ be a tensor of order d. Furthermore let $\alpha \subseteq [d]$ be a subset of the modes of $\mathbf{x}$, and let $\beta = [d] \backslash \alpha$ be its complement. Given two bijective functions $\mu_\alpha : [n_{\alpha_1}] \times [n_{\alpha_2}] \times \ldots \to [n_{\alpha_1} \cdot n_{\alpha_2} \cdots]$ and μ_β, respectively.

The α-matricization or α-flattening

$$\begin{aligned} \hat{M}_\alpha : \mathbb{R}^{n_1 \times \ldots \times n_d} &\to \mathbb{R}^{m_\alpha \times m_\beta} \\ \mathbf{x} &\mapsto \hat{M}_\alpha(\mathbf{x}) \end{aligned}$$

of $\mathbf{x}$ is defined entry-wise as

$$\mathbf{x}[i_1, \ldots, i_d] =: \hat{M}_\alpha(\mathbf{x}) \left[\mu_\alpha(i_{\alpha_1}, i_{\alpha_2} \ldots), \mu_\beta(i_{\beta_1}, i_{\beta_2}, \ldots) \right]. \tag{2}$$

A common choice for μ_α and μ_β is $\mu(i_1, i_2, \ldots) = \sum_k i_k \prod_{j>k} n_j$. The actual choice is of no significance though, as long as it stays consistent. The matrix dimensions are given as $m_\alpha = \prod_{j \in \alpha} n_j$ and $m_\beta = \prod_{j \in \beta} n_j$.

The inverse operation is the de-matricization or unflattening $\hat{M}^{-1}$. In principle it is possible to define de-matricization for any kind of matrix, typically called tensorization. However this requires to give the dimensions of the resulting tensor and the details of the mapping alongside with the operator. Instead, in this work the de-matricization is only applied to matrices where at least one mode of the matrix encodes a tensor structure through a former matricization, in which the details of

the mapping are clear from the context. For all other modes, the de-matricization is simply defined to be the identity.

The second important tool is tensor contractions, which are generalizations of the matrix-vector and matrix-matrix multiplications to higher-order tensors.

Definition 2 (Tensor Contraction)**.** Let $\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ and $\mathbf{y} \in \mathbb{R}^{m_1 \times \ldots \times m_e}$ be two tensors of order d and e, respectively, with $n_k = m_l$. The contraction of the k-th mode of $\mathbf{u}$ with the l-th mode of $\mathbf{v}$

$$\mathbf{z} := \mathbf{x} \circ_{k,l} \mathbf{y} \tag{3}$$

is defined entry-wise as

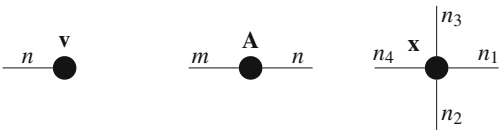
$$\begin{aligned} &\mathbf{z}[i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_d, j_1, \ldots, j_{l-1}, j_{l+1}, \ldots, j_e] \\ &= \sum_{p=0}^{n_k} \mathbf{x}[i_1, \ldots, i_{k-1}, p, i_{k+1}, \ldots, i_d]\ \mathbf{y}[j_1, \ldots, j_{l-1}, p, j_{l+1}, \ldots, j_e] \end{aligned}$$

or via the matricizations

$$\mathbf{z} = \hat{M}^{-1}\left(\hat{M}_{\{k\}}(\mathbf{x})^T \hat{M}_{\{l\}}(\mathbf{y})\right) .$$

The resulting tensor $\mathbf{z} \in \mathbb{R}^{n_1 \times \ldots \times n_{k-1} \times n_{k+1} \times \ldots \times n_d \times m_1 \times \ldots m_{k-1} \times m_{k+1} \times \ldots \times m_e}$ is of order $d + e - 2$. Note that in order for this operation to be well-defined, $n_k = m_l$ must hold.

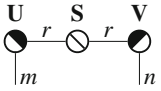
If no indices are specified, i.e., only $\circ$, a contraction of the last mode of the left operand and the first mode of the right operand is assumed. If tuples of indices are given, e.g., $\circ_{(i,j,k),(l,p,q)}$, a contraction of the respective mode pairs (i/l, j/p k/q) is assumed.[1] As writing this for larger tensor expressions quickly becomes cumbersome, we also use a diagrammatic notation to visualize the contraction. In this notation a tensor is depicted as a dot or box with edges corresponding to each of its modes. If appropriate the cardinality of the corresponding index set is given as well. From left to right, the following shows this for an order one tensor (vector) $\mathbf{v} \in \mathbb{R}^n$, an order two tensor (matrix) $\mathbf{A} \in \mathbb{R}^{m \times n}$, and an order four tensor $\mathbf{x} \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$.



If a contraction is performed between the modes of two tensors, the corresponding edges are joined. The following shows this exemplary for the inner product of two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ and a matrix-vector product with $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{v} \in \mathbb{R}^n$.

[1] As one can easily show the order of the contractions does not matter.

There are two special cases concerning orthogonal and diagonal matrices. If a specific matricization of a tensor yields an orthogonal or diagonal matrix, the tensor is depicted by a half-filled circle (orthogonal) or a circle with a diagonal bar (diagonal), respectively. The half filling and the diagonal bar both divide the circle in two halves. The edges joined to either half correspond to the mode sets of the matricization, which yields the orthogonal or diagonal matrix. As an example the diagrammatic notation can be used to depict the singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank r, as shown in the following.



2 Low-Rank Tensor Decompositions

In this section we give an introduction to the low-rank tensor decomposition techniques used in the remainder of this work. As there are in fact quite different approaches to generalize the singular value decomposition, and thereby also the definition of the rank, to higher-order tensors, we start with an overview of the most popular formats. For an in-depth overview including application, we refer to the survey of Grasedyck et al. [3]. In the second part, we provide a detailed introduction of the tensor train format, which is used in the remainder of this work.

The probably best known and classical tensor decomposition is the representation by a sum of elementary tensor products, i.e.,

$$\mathbf{x} = \sum_{i=1}^{r} \mathbf{u}_{1,i} \otimes \mathbf{u}_{2,i} \otimes \ldots \otimes \mathbf{u}_{d,i} \tag{4}$$

where $\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ and $\mathbf{u}_{k,i} \in \mathbb{R}^{n_k}$ are vectors from the respective vector spaces. This format is mainly known as the *canonical format* but also appears in the literature under the names canonical polyadic (CP) format, CANDECOMP, and PARAFAC. The canonical or CP rank is defined as the minimal r such that a decomposition as in (4) exists. Note that in general, there is no unique CP representation with minimal rank. This is somewhat expected, since even for matrices, the SVD is not unique if two or more singular values coincide. Some discussion on the uniqueness can be found in the paper of Kolda and Bader [17]. For tensors with small canonical rank, (4) offers a very efficient representation, requiring only $\mathcal{O}(rdn)$ storage instead of $\mathcal{O}(n^d)$ for the direct representation. Unfortunately the canonical format suffers from several difficulties and instabilities. First of all the

task of determining the canonical rank of a tensor with order $d > 2$ is, in contrast to matrices, highly nontrivial. In fact it was shown by [18] that even for order $d = 3$, the problem of deciding whether a rational tensor has CP-rank r is NP-hard (and NP-complete for finite fields). Consequently also the problem of calculating low-rank approximations proves to be challenging. That is, given a tensor $\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ and a CP-rank r, finding the best CP-rank r approximation

$$\mathbf{x}^* = \underset{\mathbf{y} \in \mathbb{R}^{n_1 \times \ldots \times n_d},\ \text{CP-rank}(\mathbf{y}) \leq r}{\operatorname{argmin}} (\|\mathbf{x} - \mathbf{y}\|) \ . \tag{5}$$

The norm $\|\cdot\|$ used may differ depending on the application. In the matrix case, the Eckart-Young theorem provides that for the Frobenius and spectral norm, this best approximation can be straightforwardly calculated by a truncated SVD. In contrast, De Silva and Lim [19] proved that the problem of the best CP-rank r approximation, as formulated in (5), is ill-posed for many ranks $r \geq 2$ and all orders $d > 2$ regardless of the choice of the norm $\|\cdot\|$. Furthermore they showed that the set of tensors that do not have a best CP-rank r approximation is a non-null set, i.e., there is a strictly positive probability that a randomly chosen tensor does not admit a best CP-rank r approximation. Finally it was shown by De Silva and Lim [19] that neither the set $\{\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_d} \mid \text{CP-rank}(\mathbf{x}) = r\}$ of all tensors with CP-rank r nor the set $\{\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_d} \mid \text{CP-rank}(\mathbf{x}) \leq r\}$ of all tensors with CP-rank at most r is closed for $d > 2$. These are some severe difficulties for both the theoretical and practical works with the canonical format.

The second classical approach to generalize the SVD to higher-order tensors is the subspace-based Tucker decomposition. It was first introduced by Tucker [20] in 1963 and has been refined later on in many works; see, e.g., [14, 17, 21]. Given a tensor $\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$, the main idea is to find minimal subspaces $U_i \subseteq \mathbb{R}^{n_i}$, such that $\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ is an element of the induced tensor space

$$\mathscr{U} = \bigotimes_{i=1}^{d} U_i \subseteq \bigotimes_{i=1}^{d} \mathbb{R}^{n_i} = \mathbb{R}^{n_1 \times \ldots \times n_d} \ . \tag{6}$$

Let $r_i = \dim(U_i)$ denote the dimension of the i-th subspace and let $\{\mathbf{u}_{i,j}, j = 0, \ldots, r_i\}$ be an orthonormal basis of U_i. If the subspaces are chosen such that $\mathbf{x} \in \mathscr{U}$, then (1) states that there is a $\mathbf{c}$ such that $\mathbf{x}$ can be expressed as

$$\mathbf{x} = \sum_{\nu_1=1}^{r_1} \ldots \sum_{\nu_d}^{r_d} \mathbf{c}[\nu_1, \nu_2, \ldots, \nu_d] \cdot \mathbf{u}_{1,\nu_1} \otimes \ldots \otimes \mathbf{u}_{d,\nu_d} \ . \tag{7}$$

Usually the basis vectors are combined to orthogonal matrices $\mathbf{U}_i = (\mathbf{u}_{i,1}, \ldots, \mathbf{u}_{i,r_i})$, called *basis matrices*. This leads to the following common form of the Tucker format

$$\mathbf{x}[\mu_1, \ldots, \mu_d] = \sum_{\nu_1=1}^{r_1} \ldots \sum_{\nu_d}^{r_d} \mathbf{c}[\nu_1, \ldots, \nu_d] \mathbf{U}_1[\mu_1, \nu_1] \ldots \mathbf{U}_d[\mu_d, \nu_d] \ . \tag{8}$$

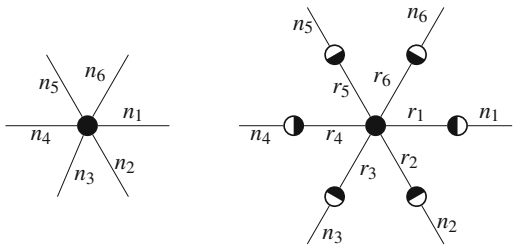


Fig. 1 Left: A tensor $\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_6}$ of order 6. Right: Its Tucker decomposition

The order d tensor $\mathbf{c} \in \mathbb{R}^{r_1 \times \ldots \times r_d}$ of the prefactors is usually called *core tensor*. The d-tuple $\mathbf{r} = (r_1, r_2, \ldots, r_d)$ of the subspace dimensions is called the representation rank and is associated with the particular representation. The *Tucker rank* (T-rank) of $\mathbf{x}$ is defined as the unique minimal d-tuple $\mathbf{r}^* = (r_1^*, \ldots, r_d^*)$, such that there exists a Tucker representation of $\mathbf{x}$ with rank $\mathbf{r}^*$. Equation (8) consists of d tensor contractions that can be visualized in the diagrammatic notation, which is exemplarily shown in Figure 1 for $d = 6$. Note that even for the minimal T-rank, the Tucker decomposition is not unique, as for any orthogonal matrix $\mathbf{Q}_i \in \mathbb{R}^{r_i \times r_i}$, one can define a matrix $\tilde{\mathbf{U}}_i = \mathbf{U}_i \mathbf{Q}_i$ and the tensor

$$\tilde{\mathbf{c}}\,[\mu_1, \ldots, \mu_d] = \sum_{\nu=1}^{r_i} \mathbf{c}\,[\mu_1, \ldots, \nu, \ldots \mu_d]\, \mathbf{Q}^T\,[\nu, \mu_i]$$

such that the tensor $\mathbf{x}$ can also be written as

$$\mathbf{u}\,[\mu_1, \ldots, \mu_d] = \sum_{\nu_1=1}^{r_1} \ldots \sum_{\nu_d=1}^{r_d} \tilde{\mathbf{c}}\,[\nu_1, \ldots, \nu_d]\, \mathbf{U}_1[\mu_1, \nu_1] \ldots \tilde{\mathbf{U}}_i[\mu_i, \nu_i] \ldots \mathbf{U}_d[\mu_d, \nu_d],$$

which is a valid Tucker decomposition with the same rank.

It is shown by De Lathauwer et al. [21] that the Tucker rank as the minimal d-tuple is indeed well defined and that the entries r_i of the Tucker rank correspond to the rank of the i-th mode matricization of the tensor. That is

$$\text{T-rank}(\mathbf{x}) = \left(\text{rank}(\hat{M}_{\{1\}}(\mathbf{x})), \ldots, \text{rank}(\hat{M}_{\{d\}}(\mathbf{x}))\right) . \tag{9}$$

The proof is tightly linked to the fact that a Tucker representation of a tensor $\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ with minimal representation rank can be obtained by successive singular value decompositions. This procedure is referred to as the higher-order singular value decomposition (HOSVD); see [21] for the details. Using truncated SVDs, an approximation of $\mathbf{x}$ by a tensor $\mathbf{x}^*$ with lower T-rank $\mathbf{r}^* = (r_1^*, \ldots r_d^*) \preceq (r_1, \ldots, r_d)$ can be obtained. Where the symbol $\preceq$ denotes an entry-wise $\leq$, i.e., $(r_1, \ldots r_d) \preceq (r_1^*, \ldots, r_d^*) \iff r_i \leq r_i^* \forall i$. In contrast to the Eckart-Young theorem for matrices, the approximation $\mathbf{x}^*$ obtained in this way is *not* the best T-rank $\mathbf{r}^*$ approximation of $\mathbf{x}$. However, it is a quasi-best approximation by a factor $\sqrt{d}$, i.e.,

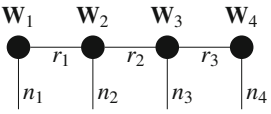
$$\|\mathbf{x} - \mathbf{x}^*\|_F \leq \sqrt{d} \min_{\mathbf{y} \,:\, \text{T-rank}(\mathbf{y}) \leq \mathbf{r}^*} (\|\mathbf{x} - \mathbf{y}\|_F) \ . \tag{10}$$

For many applications this quasi-best approximation is sufficient. As for the canonical format, finding the true best approximation is at the very least NP-hard in general, as it is shown by [22] that finding the best rank $(1, \ldots, 1)$ approximation is already NP-hard. To store a tensor in the Tucker format, only the core tensor and the basis matrices have to be stored. This amounts to a storage requirement of $\mathcal{O}(r^d + dnr)$, where $r := \max(r_1, \ldots, r_d)$ and $n := \max(n_1, \ldots, n_d)$. Compared to the $\mathcal{O}(n^d)$, this is a major reduction but does not break the curse of dimensionality as the exponential scaling in d remains.

A more recent development is the *hierarchical Tucker (HT)* format, introduced by Hackbusch and Kühn [1]. It inherits most of the advantages of the Tucker format, in particular a generalized higher-order SVD; see [23]. But in contrast to the Tucker format, the HT format allows a linear scaling with respect to the order for the storage requirements and common operations for tensors of fixed rank. The main idea of the HT format is to extend the subspace approach of the Tucker format by a multilayer hierarchy of subspaces. For an in-depth introduction of the hierarchical Tucker format, we refer to the pertinent literature, e.g., [1, 14, 23]. In this work we will instead focus on the *tensor train (TT)* format, as introduced by Oseledets [2]. The TT format offers mostly the same advantages as the more general HT format while maintaining a powerful simplicity. In fact it can to some extent be seen as a special case of the HT format; see Grasedyck and Hackbusch [24] for details on the relation between the TT and HT format.

2.1 Tensor Train Format

In this section we give a detailed introduction to the *tensor train (TT)* format. In the formulation used in this work, the TT format was introduced by Oseledets [2]; however an equivalent formulation was known in quantum physics for quite some time; see, e.g., [25] for an overview. The idea of the TT format is to separate the modes of a tensor into d order two and three tensors. This results in a tensor network that is exemplary shown for an order four tensor $\mathbf{x} = \mathbf{W}_1 \circ \mathbf{W}_2 \circ \mathbf{W}_3 \circ \mathbf{W}_4$ in the following diagram.



Formally it can be defined as follows.

Definition 1 (Tensor Train Format)**.** Let $\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ be a tensor of order d. A factorization

$$\mathbf{x} = \mathbf{W}_1 \circ \mathbf{W}_2 \circ \ldots \circ \mathbf{W}_{d-1} \circ \mathbf{W}_d \,, \tag{11}$$

of $\mathbf{x}$, into component tensors $\mathbf{W}_1 \in \mathbb{R}^{n_1 \times r_1}$, $\mathbf{W}_i \in \mathbb{R}^{r_{i-1} \times n_i \times r_i}$ $(i = 2, \ldots, d-1)$ and $\mathbf{W}_d \in \mathbb{R}^{r_{d-1} \times n_d}$, is called a tensor train (TT) representation of $\mathbf{x}$. Equivalently (11) can be given entry-wise as

$$\mathbf{x}[i_1, \ldots, i_d] = \sum_{j_1} \ldots \sum_{j_{d-1}} \mathbf{W}_1[i_1, j_1]\, \mathbf{W}_2[j_1, i_2, j_2] \ldots \mathbf{W}_{d-1}[j_{d-2}, i_{d-1}, j_{d-1}]\, \mathbf{W}_d[j_{d-1}, i_d] \,.$$

The tuple of the dimensions $\mathbf{r} = (r_1, \ldots, r_{d-1})$ of the component tensors is called the representation rank and is associated with the specific representation. In contrast the tensor train rank (TT-rank) of $\mathbf{x}$ is defined as the minimal rank tuple $\mathbf{r}^* = (r_1^*, \ldots, r_{d-1}^*)$ such that there exists a TT representation of $\mathbf{x}$ with representation rank equal to $\mathbf{r}^*$.

As for the Tucker format, the TT-rank is well defined and linked to the rank of specific matricizations via

$$\text{TT-Rank}(\mathbf{x}) = \left(\text{rank}(\hat{M}_{\{1\}}(\mathbf{x})), \text{rank}(\hat{M}_{\{1,2\}}(\mathbf{x})), \ldots, \text{rank}(\hat{M}_{\{1,2,\ldots,d-1\}}(\mathbf{x}))\right) \,.$$

The proof is again closely linked to the fact that a tensor train decomposition of an arbitrary tensor can be calculated using successive singular value decompositions. This procedure is commonly referred to as the TT-SVD. For this work the TT-SVD is of particular importance as it constitutes the deterministic baseline for our randomized approach in Section 3.2. In the following we therefore provide a complete step-by-step description of this procedure.

Tensor Train Singular Value Decomposition (TT-SVD)
The intuition of the TT-SVD is that in every step, a (matrix) SVD is performed to detach one open mode from the tensor. Figure 2 shows this process step-by-step for an order four tensor and is frequently referred to in the following description. The TT-SVD starts by calculating an SVD of the matricization of $\mathbf{x} = \mathbf{x}_0$, where all modes but the first one are combined (Figure 2(a)–(c)):

$$\mathbf{U}_1 \mathbf{S}_1 \mathbf{V}_1^T := \text{SVD}\left(\hat{M}_{\{1\}}(\mathbf{x}_0)\right) \,, \tag{12}$$

with $\mathbf{U}_1 \in \mathbb{R}^{n_1 \times r_1}$, $\mathbf{S}_1 \in \mathbb{R}^{r_1 \times r_1}$, $\mathbf{V}_1^T \in \mathbb{R}^{r_1 \times (n_2 \cdot \ldots \cdot n_d)}$. The dimension r_1 is equal to the rank of $\hat{M}_{\{1\}}(\mathbf{x}_0)$. The resulting matrices $\left(\mathbf{S}_1 \mathbf{V}_1^T\right)$ and $\mathbf{U}_1$ are each dematricized, which is trivial for $\mathbf{U}_1$ in the first step (Figure 2(d)–(e))

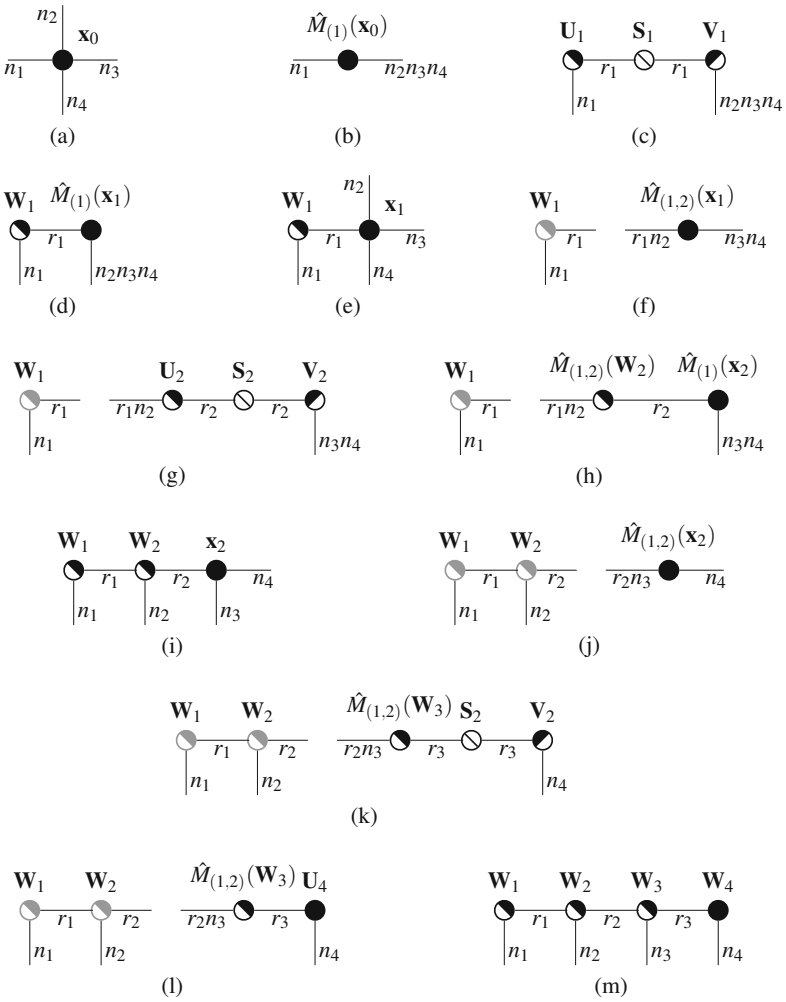


Fig. 2 Step-by-step depiction of the TT-SVD by example for an order *four* tensor

$$\mathbf{W}_1 := \hat{M}^{-1}(\mathbf{U}_1) \qquad \mathbf{W}_1 \in \mathbb{R}^{n_1 \times r_1} \tag{13}$$

$$\mathbf{x}_1 := \hat{M}^{-1}\left(\mathbf{S}_1 \mathbf{V}_1^T\right) \qquad \mathbf{x}_1 \in \mathbb{R}^{r_1 \times n_2 \times \ldots \times n_d} \tag{14}$$

Note that there holds

$$\mathbf{W}_1 \circ \mathbf{x}_1 = \hat{M}^{-1}\left(\mathbf{U}_1 \mathbf{S}_1 \mathbf{V}_1^T\right) = \mathbf{x}_0 = \mathbf{x}\,. \tag{15}$$

In the next step, a matricization of the newly acquired tensor $\mathbf{x}_1$ is performed. The first dimension of the matricization is formed by the first two modes of $\mathbf{x}_1$,

corresponding to the new dimension introduced by the prior SVD and the second original dimension. The second dimension of the matricization is formed by all remaining modes of $\mathbf{x}_1$ (Figure 2(f)). From this matricization another SVD is calculated (Figure 2(g)):

$$\mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^T := \text{SVD}\left(\hat{M}_{\{1,2\}}(\mathbf{x}_1)\right) , \tag{16}$$

with $\mathbf{U}_2 \in \mathbb{R}^{(r_1 \cdot n_2) \times r_2}$, $\mathbf{S}_2 \in \mathbb{R}^{r_2 \times r_2}$, $\mathbf{V}_2^T \in \mathbb{R}^{r_2 \times (n_3 \cdot \ldots \cdot n_d)}$. As in the first step, $\mathbf{U}_2$ and $\left(\mathbf{S}_2\mathbf{V}_2^T\right)$ are then dematricized (Figure 2 (i))

$$\mathbf{W}_2 := \hat{M}^{-1}(\mathbf{U}_2) \qquad \mathbf{W}_2 \in \mathbb{R}^{r_1 \times n_2 \times r_2} \tag{17}$$

$$\mathbf{x}_2 := \hat{M}^{-1}(\mathbf{S}_2\mathbf{V}_2^T) \qquad \mathbf{x}_2 \in \mathbb{R}^{r_2 \times n_3 \times \ldots \times n_d} \tag{18}$$

and again there holds

$$\mathbf{W}_2 \circ \mathbf{x}_2 = \hat{M}^{-1}\left(\mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^T\right) = \mathbf{x}_1 \tag{19}$$

$$\Rightarrow \mathbf{W}_1 \circ \mathbf{W}_2 \circ \mathbf{x}_2 = \mathbf{x} . \tag{20}$$

The obtained rank r_2 is equal to the rank of the matricization $\hat{M}_{\{1,2\}}(\mathbf{x})$, which can be shown as follows. First note that $\mathbf{Q} := \hat{M}_{\{1,2\}}(\mathbf{U}_1 \circ \mathbf{U}_2) \in \mathbb{R}^{(n_1 \cdot n_2) \times r_2}$ is an orthogonal matrix, because

$$\left(\mathbf{Q}\mathbf{Q}^T\right)[j,j'] = \sum_i \hat{M}_{\{1,2\}}(\mathbf{W}_1 \circ \mathbf{W}_2)[i,j]\, \hat{M}_{(1,2)}(\mathbf{W}_1 \circ \mathbf{W}_2)[i,j'] \tag{21}$$

$$= \sum_{i_1,i_2,k,k'} \mathbf{W}_1[i_1,k]\, \mathbf{W}_2[k,i_2,j]\, \mathbf{W}_1[i_1,k']\, \mathbf{W}_2[k',i_2,j'] \tag{22}$$

$$= \sum_{i_2,k,k'} \underbrace{\sum_{i_1} \mathbf{W}_1[i_1,k]\, \mathbf{W}_1[i_1,k']}_{\mathbf{I}[k,k']}\, \mathbf{W}_2[k,i_2,j]\, \mathbf{W}_2[k',i_2,j'] \tag{23}$$

$$= \sum_{i_2,k} \mathbf{W}_2[k,i_2,j]\, \mathbf{W}_2[k,i_2,j'] \tag{24}$$

$$= \mathbf{I}[j,j'] . \tag{25}$$

Then consider that

$$\hat{M}_{\{1,2\}}(\mathbf{x}) = \hat{M}_{\{1,2\}}(\mathbf{W}_1 \circ \mathbf{W}_2 \circ \mathbf{x}_3) \tag{26}$$

$$= \hat{M}_{\{1,2\}}(\mathbf{W}_1 \circ \mathbf{W}_2)\, \hat{M}_{\{1\}}(\mathbf{x}_2) \tag{27}$$

$$= \mathbf{Q}\mathbf{S}_2\mathbf{V}_2^T \tag{28}$$

holds. This is a valid SVD of $\hat{M}_{\{1,2\}}(\mathbf{x})$, and since the matrix of singular values is unique, it follows that in fact $\operatorname{rank}(\hat{M}_{\{1,2\}}(\mathbf{x})) = r_2$.

This procedure is continued for a total of $d-2$ steps and in each step the order of $\mathbf{x}_i \in \mathbb{R}^{r_i \times n_{i+1} \times \ldots \times n_d}$ shrinks by one. Furthermore there holds

$$\mathbf{W}_1 \circ \mathbf{W}_2 \circ \ldots \circ \mathbf{W}_i \circ \mathbf{x}_i = \mathbf{x} \tag{29}$$

$\mathbf{W}_i \in \mathbb{R}^{r_{i-1} \times n_i \times r_i}$ and $r_i = \operatorname{rank}\left(\hat{M}_{\{1,\ldots,i\}}(\mathbf{x})\right)$ in every step. The $(d-1)$-st step (Figure 2 (k))

$$\mathbf{U}_{d-1}\mathbf{S}_{d-1}\mathbf{V}_{d-1}^T = \operatorname{SVD}\left(\hat{M}_{\{1,2\}}(\mathbf{x}_{d-1})\right) , \tag{30}$$

with $\mathbf{U}_{d-1} \in \mathbb{R}^{(r_{d-2}\cdot n_{d-1}) \times r_{d-1}}$, $\mathbf{S}_{d-1} \in \mathbb{R}^{r_{d-1} \times r_{d-1}}$, $\mathbf{V}_{d-1}^T \in \mathbb{R}^{r_{d-1} \times n_d}$, is special since the de-matricization of $(\mathbf{S}_{d-1}\mathbf{V}_{d-1}^T)$ yields an order two tensor that is named $\mathbf{W}_d$ instead of $\mathbf{x}_d$ (Figure 2 (l)-(m))

$$\mathbf{W}_{d-1} = \hat{M}^{-1}(\mathbf{U}_{d-1}) \qquad \mathbf{W}_{d-1} \in \mathbb{R}^{r_{d-2} \times n_{d-1} \times r_{d-1}} \tag{31}$$

$$\mathbf{W}_d = \mathbf{S}_{d-1}\mathbf{V}_{d-1}^T \qquad \mathbf{x}_d \in \mathbb{R}^{r_{d-1} \times n_d} . \tag{32}$$

Finally

$$\mathbf{W}_1 \circ \mathbf{W}_2 \circ \ldots \circ \mathbf{W}_{d-1} \circ \mathbf{W}_d = \mathbf{x} \tag{33}$$

is a valid TT representation of $\mathbf{x}$ with TT-rank $\mathbf{r} = (r_1, \ldots, r_{d-1})$, whose entries $r_i = \operatorname{rank}(\hat{M}_{\{1,\ldots,i\}}(\mathbf{x}))$ are exactly the ranks of the matricizations as asserted.

The same algorithm can also be used to calculate a rank $\mathbf{r}^* = (r_1^*, \ldots, r_{d-1}^*)$ approximation of a tensor $\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ with TT-rank $\mathbf{r} \succeq \mathbf{r}^*$. To this end the normal SVDs are replaced by truncated rank r_i^* SVDs, yielding a tensor $\mathbf{x}^*$ of TT-rank $\mathbf{r}^*$. In contrast to the matrix case, $\mathbf{x}^*$ is in general not the best rank $\mathbf{r}^*$ approximation of $\mathbf{x}$. However as shown by [2], it is a quasi-best approximation with

$$\|\mathbf{x} - \mathbf{x}^*\|_2 \leq \sqrt{d-1} \min_{\mathbf{y} \,:\, \text{TT-rank}(\mathbf{y}) \preceq \mathbf{r}^*} (\|\mathbf{x} - \mathbf{y}\|) . \tag{34}$$

The computational complexity of the TT-SVD is dominated by the $d-1$ matrix singular value decompositions, with all other contributions being asymptotically negligible. With $n := \max(n_1, \ldots, n_d)$ and $r := \max(r_1, \ldots r_d)$, the cost scales as $\mathcal{O}(n^{d+1} + \sum_{i=1}^{d-1} r^2 n^{d-i}) \subset \mathcal{O}(dn^{d+1})$, i.e., still exponential in the order. This is somewhat expected because there are in general n^d entries in the original tensor that have to be considered. Unfortunately $\mathbf{x}$ being sparse or otherwise structured incurs no dramatic change because the structure is generally lost after the first SVD.

Apart from the generalized singular value decomposition, the TT format offers several further beneficial properties. In particular it is able to break the curse of

dimensionality, in the sense that the storage complexity of a tensor $\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ with TT-rank $\mathbf{r} = (r_1, \ldots, r_{d-1})$ in a minimal TT representation scales as $\mathcal{O}(dnr^2)$, i.e., linearly in the order. Here $n := \max_i(n_1, \ldots, n_d)$ and $r := \max_i(r_1, \ldots, r_{d-1})$. Additionally also the computational complexity of common operations as additions and scalar products scales only linearly in the order for fixed ranks; see [2, 14]. Another desirable property is that the set of tensors with rank at most $\mathbf{r}$ form a closed set, and as shown by Holtz et al. [26], the set of tensor with exact rank $\mathbf{r}$ forms a smooth manifold, allowing the application of Riemannian optimization techniques [27, 28] and dynamical low-rank approximation [29, 30]; see also the review article [12]. Especially for numerical applications, these properties made the tensor train one of, if not, the most popular tensor decomposition of recent years.

3 Randomized SVD for Higher-Order Tensors

As shown in the previous section, calculating a low-rank representation or approximation of a given higher-order tensor is a challenging task, as the complexity of the tensor train SVD (TT-SVD) scales exponentially in the order. For dense tensors this is of course somewhat expected as there is an exponential number of entries that have to be incorporated. Nevertheless also for sparse and structured matrices, the two decomposition techniques exhibit an exponential scaling. In this section we look at randomized methods for the calculation of approximate matrix factorizations. For sparse or structured matrices, these techniques allow for a very efficient calculation of common matrix factorizations such as the SVD or QR decomposition while offering rigorous stochastic error bounds. In the second part of this section, we apply these results to formulate randomized TT-SVD algorithms. We show that there hold stochastic error bounds similar to the matrix setting. We also show that this randomized TT-SVD has only linear complexity with respect to the order when applied to sparse tensors.

3.1 Randomized SVD for Matrices

Randomized techniques for the calculation of SVD or QR factorizations of matrices have been proposed many times in the literature. However it was only recently that, thanks to the application of new results from *random matrix theory*, these procedures could be analyzed rigorously. We start this section by presenting some results from the work of Halko et al. [5], which will provide a solid basis for the randomized tensor factorization methods of the second part of this section. In this part we restrict ourself to standard Gaussian random matrices, i.e., matrices whose entries are i.i.d. standard Gaussian random variables. The usage of structured random matrices is discussed in Section 6.

In the formulation of Halko et al. [5], the basis of all decompositions is a randomized method to calculate an approximate low-rank subspace projection

$$\mathbf{A} \approx \mathbf{Q}\mathbf{Q}^T\mathbf{A} \tag{35}$$

where $\mathbf{A} \in \mathbb{R}^{n\times m}$ is a given matrix and $\mathbf{Q} \in \mathbb{R}^{n\times s}$ is an orthogonal matrix approximately spanning the range of $\mathbf{A}$. Here $s = r + p$, where r is the desired rank and p is an oversampling parameter. With this projection at hand, numerous different low-rank decompositions can be calculated deterministically at low costs. For example, the singular value decomposition of $\mathbf{A}$ can be calculated by forming $\mathbf{B} := \mathbf{Q}^T\mathbf{A}$ and calculating the deterministic SVD $\mathbf{U}\mathbf{S}\mathbf{V}^T = \mathbf{B}$ of $\mathbf{B}$. Using $\tilde{\mathbf{U}} = \mathbf{Q}\mathbf{U}$

$$\tilde{\mathbf{U}}\mathbf{S}\mathbf{V}^T = \mathbf{Q}\mathbf{Q}^T\mathbf{A} \approx \mathbf{A} \tag{36}$$

is an approximate SVD of $\mathbf{A}$ containing only the approximation error incurred by the subspace-projection. The computational costs of the involved operations scale as $\mathcal{O}\left(sT_{\text{mult}} + s^2(m+n)\right)$, where T_{mult} is the cost to calculate the matrix-vector product with $\mathbf{A}$, which is $\mathcal{O}(mn)$ for a general matrix but can be much lower for structured or sparse matrices. In a similar way, other matrix factorizations can also be computed with low costs if the projection (35) is given.

The main challenge is the calculation of the approximate range $\mathbf{Q}$ through random techniques. For this [5] present the following prototype algorithm. Given a matrix $\mathbf{A} \in \mathbb{R}^{n_1\times n_2}$.

Listing 1 Randomized Range Approximation
Input: $\mathbf{A}$, r, p Output: $\mathbf{Q}$
Create a standard Gaussian random matrix $\mathbf{G} \in \mathbb{R}^{n_2\times(r+p)}$
Calculate the intermediate matrix $\mathbf{B} := \mathbf{A}\mathbf{G} \in \mathbb{R}^{n_1\times s}$.
Compute the factorization $\mathbf{Q}\mathbf{R} = \mathbf{B}$.

The following theorem proves that the $\mathbf{Q}$ obtained in this manner is indeed an approximation of the range of $\mathbf{A}$ in the sense of (35).

Theorem 1 (Halko et al. [5]). *Given $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $s = r + p$ with $p \geq 2$. For the projection $\mathbf{Q}$ obtained by procedure 1, there holds the following error bounds.*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^T\mathbf{A}\| \leq \left[1 + 11\sqrt{(r+p)\cdot\min(m,n)}\right]\sigma_{r+1} \tag{37}$$

with probability at least $1 - 6p^{-p}$ and for $p \geq 4$ and any $u, t \geq 1$

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^T\mathbf{A}\| \leq \left[1 + t\sqrt{\frac{12r}{p}}\right]\left(\sum_{k>r}\sigma_k^2\right)^{1/2} + ut\frac{e\sqrt{r+p}}{p+1}\sigma_{r+1} \tag{38}$$

with probability at least $1 - 5t^{-p} - 2e^{-u^2/2}$.

Let us highlight furthermore that for the operator norm, we have that $\sigma_{r+1} = \inf_{\text{rank }\mathbf{B}=r} \|\mathbf{A}-\mathbf{B}\|_{op}$ and for the Frobenius norm there holds

$$\left(\sum_{k>r} \sigma_k^2\right)^{\frac{1}{2}} = \inf_{\text{rank}(\mathbf{B})\leq r} \|\mathbf{A}-\mathbf{B}\| \,.$$

3.2 Randomized TT-SVD

In this section we show how the same idea of the randomized range approximation for matrices can be used to formulate a randomized algorithm that calculates an approximate TT-SVD of arbitrary tensors. We show that stochastic error bounds analogous to the matrix case can be obtained. Furthermore we show that for sparse tensors, this randomized TT-SVD can be calculated in *linear* complexity with respect to the order of the tensor, instead of the exponential complexity of the deterministic TT-SVD.

The idea of our randomized TT-SVD procedure is to calculate nested range approximations increasing by one mode at a time. The corresponding projector is composed of separated orthogonal parts, which are calculated using procedure 1. This is visualized in Figure 3. These orthogonal parts will become the component tensors $\mathbf{W}_2, \ldots, \mathbf{W}_d$ of the final TT decomposition. The first component tensor $\mathbf{W}_1$ is given by contracting the initial $\mathbf{x}$ with all orthogonal components, i.e.,

$$\mathbf{W}_1 = \mathbf{x} \circ_{(2,\ldots,d),(2,\ldots,d)} (\mathbf{W}_2 \circ \ldots \circ \mathbf{W}_d)$$

The exact procedure calculating the orthogonal components and this final contraction is given in Listing 2.

Listing 2 Randomized TT-SVD

Input: $\mathbf{x}$, Output: $\mathbf{W}_1, \ldots, \mathbf{W}_d$

Set $\mathbf{b}_{d+1} := \mathbf{x}$
For $j = d, \ldots, 2$:
  Create a Gaussian random tensor $\mathbf{g} \in \mathbb{R}^{s_{j-1} \times n_1 \times \ldots \times n_{j-1}}$
  Calculate $\mathbf{a}_j := \mathbf{g} \circ_{(2,\ldots,j),(1,\ldots,j-1)} \mathbf{b}_{j+1}$
  Calculate the factorization $\mathbf{R}_j\mathbf{Q}_j := \hat{M}_{\{1\}}(\mathbf{a}_j)$
  Set $\mathbf{W}_j := \hat{M}^{-1}(\mathbf{Q}_j)$
  if j = d:
    Calculate $\mathbf{b}_j = \mathbf{b}_{j+1} \circ_{(j),(2)} \mathbf{W}_j$
  else
    Calculate $\mathbf{b}_j = \mathbf{b}_{j+1} \circ_{(j,j+1),(2,3)} \mathbf{W}_j$

Set $\mathbf{W}_1 = \mathbf{b}_2$

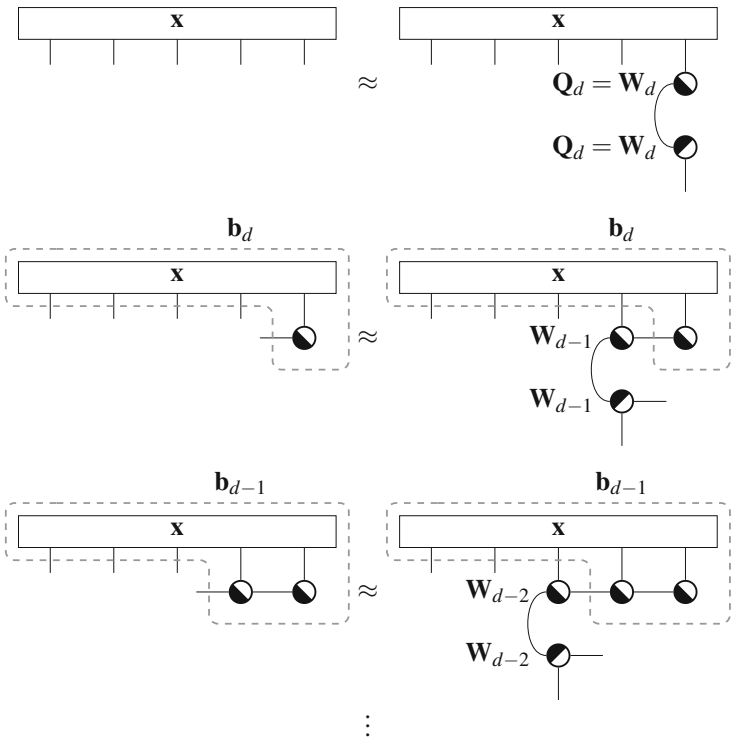


Fig. 3 Iterative construction of the tensor $\mathbf{b}_i$ by subsequent range approximations

At the end of the procedure in Listing 2

$$\mathbf{x} \approx \mathbf{W}_1 \circ \mathbf{W}_2 \circ \mathbf{W}_3 \circ \ldots \circ \mathbf{W}_d$$

is an approximate TT decomposition of rank $\mathbf{s} = (s_1, \ldots, s_{d-1})$. This final composition can also be given in terms of contractions with the orthogonal parts, i.e.,

$$\begin{aligned}
\mathbf{x} &\approx \mathbf{W}_1 \circ \mathbf{W}_2 \circ \mathbf{W}_3 \circ \ldots \circ \mathbf{W}_d && (39)\\
&= \left(\mathbf{x} \circ_{(2,\ldots,d),(2,\ldots,d)} (\mathbf{W}_2 \circ \ldots \circ \mathbf{W}_d)\right) \circ \mathbf{W}_2 \circ \mathbf{W}_3 \circ \ldots \circ \mathbf{W}_d && (40)\\
&= \mathbf{x} \circ_{(2,\ldots,d),(2,\ldots,d)} ((\mathbf{W}_2 \circ \ldots \circ \mathbf{W}_d) \circ_{1,1} (\mathbf{W}_2 \circ \ldots \circ \mathbf{W}_d)) && (41)\\
&=: \hat{P}_{2,\ldots,d}(\mathbf{x}) && (42)
\end{aligned}$$

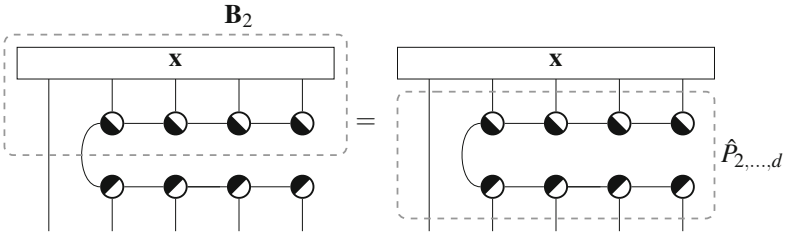


Fig. 4 Depiction of the randomized TT-SVD as the action of the projection operator $\hat{P}_{2,\ldots,d}$

where the effect of the orthogonal parts can also be seen as the action of an projector $\hat{P}_{2,\ldots,d}$. Note that since all parts are orthogonal, this is indeed an orthogonal projector. This relation is visualized in Figure 4. In the following, it will be useful to also define the orthogonal projections

$$\hat{P}_{i,\ldots,d}(\mathbf{x}) := \mathbf{x} \circ_{(i,\ldots,d),(i,\ldots,d)} ((\mathbf{W}_i \circ \ldots \circ \mathbf{W}_d) \circ_{1,1} (\mathbf{W}_i \circ \ldots \circ \mathbf{W}_d)) \ . \tag{43}$$

The following theorem shows that there exists an stochastic error bound for this randomized TT-SVD.

Theorem 2 (Error Bound)**.** *Given* $\mathbf{x} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$, $s = r + p$ *with* $p \geq 4$. *For every* $u, t \geq 1$, *the error of the randomized TT-SVD, as given in Listing 2, fulfills*

$$\|\mathbf{x} - P_{2..d}(\mathbf{x})\| \leq \sqrt{d-1}\ \eta(r,p) \min_{TT\text{-}rank(\mathbf{y}) \leq \mathbf{r}} \|\mathbf{x} - \mathbf{y}\| \tag{44}$$

with probability at least $(1 - 5t^{-p} - 2e^{-u^2/2})^{d-1}$. *The parameter* η *is given as*

$$\eta = 1 + t\sqrt{\frac{12r}{p}} + ut\frac{e\sqrt{r+p}}{p+1} \ . \tag{45}$$

Proof. For syntactical convenience let us define $\mathbf{B}_i := \left(\hat{M}_{\{1,\ldots,i-1\}}(\mathbf{b}_i)\right)^T$. Then as $\hat{P}_{2,\ldots,d}$ is an orthogonal projector, we have

$$\begin{aligned}
\|\mathbf{x} - P_{2..d}(\mathbf{x})\|^2 &= \|\mathbf{x}\|^2 - \|P_{2..d}(\mathbf{x})\|^2 && (46)\\
&= \|\mathbf{x}\|^2 - \langle \mathbf{B}_2, \mathbf{B}_2 \rangle && (47)\\
&= \|\mathbf{x}\|^2 - \langle \mathbf{Q}_2\mathbf{B}_3, \mathbf{Q}_2\mathbf{B}_3 \rangle && (48)\\
&= \|\mathbf{x}\|^2 - \langle \mathbf{B}_3, \mathbf{Q}_2^T\mathbf{Q}_2\mathbf{B}_3 \rangle \ . && (49)
\end{aligned}$$

For all $2 \leq i \leq d$, there holds

$$\left\langle \mathbf{B}_{i+1}, \mathbf{Q}_i^T \mathbf{Q}_i \mathbf{B}_{i+1} \right\rangle = \left\langle \mathbf{B}_{i+1}, \mathbf{B}_{i+1} - (\mathbf{I} - \mathbf{Q}_i^T \mathbf{Q}_i) \mathbf{B}_{i+1} \right\rangle \tag{50}$$

$$= \|\mathbf{B}_{i+1}\|^2 - \left\langle \mathbf{B}_{i+1}, (\mathbf{I} - \mathbf{Q}_i^T \mathbf{Q}_i) \mathbf{B}_{i+1} \right\rangle \tag{51}$$

$$= \|\mathbf{B}_{i+1}\|^2 - \|(\mathbf{I} - \mathbf{Q}_i^T \mathbf{Q}_i) \mathbf{B}_{i+1}\|^2 \tag{52}$$

$$= \left\langle \mathbf{B}_{i+2}, \mathbf{Q}_{i+1}^T \mathbf{Q}_{i+1} \mathbf{B}_{i+2} \right\rangle - \|(\mathbf{I} - \mathbf{Q}_i^T \mathbf{Q}_i) \mathbf{B}_{i+1}\|^2 , \tag{53}$$

where we used that $\mathbf{Q}_i^T \mathbf{Q}_i$ is an orthogonal projector as well. Inserting this iteratively into (46) gives

$$\|\mathbf{x} - P_{2..d}(\mathbf{x})\|^2 = \|\mathbf{x}\|^2 - \|\mathbf{B}_{d+1}\|^2 + \sum_{i=2}^{d} \|(\mathbf{I} - \mathbf{Q}_i^T \mathbf{Q}_i) \mathbf{B}_{i+1}\|^2 \tag{54}$$

$$= \sum_{i=2}^{d} \|(\mathbf{I} - \mathbf{Q}_i^T \mathbf{Q}_i) \mathbf{B}_{i+1}\|^2 , \tag{55}$$

where we used that $\mathbf{B}_{d+1}$ as a matricization of $\mathbf{b}_{d+1} := \mathbf{x}$ has the same norm as $\mathbf{x}$ itself. As $\mathbf{Q}_i$ is obtained in the exact setting of theorem 1, we know that for all i

$$\|(\mathbf{I} - \mathbf{Q}_i \mathbf{Q}_i^T) \mathbf{B}_{i+1}\|^2 \tag{56}$$

$$\leq \left[\left(1 + t \sqrt{\frac{12r}{p}} \right) \left(\sum_{k>r} \sigma_k^2(\mathbf{B}_{i+1}) \right)^{1/2} + ut \frac{e\sqrt{r+p}}{p+1} \sigma_{r+1}(\mathbf{B}_{i+1}) \right]^2 \tag{57}$$

$$\leq \left[\left(1 + t \sqrt{\frac{12r}{p}} + ut \frac{e\sqrt{r+p}}{p+1} \right) \left(\sum_{k>r} \sigma_k^2(\mathbf{B}_{i+1}) \right)^{1/2} \right]^2 \tag{58}$$

$$\leq \eta^2 \sum_{k>r} \sigma_k^2(\mathbf{B}_{i+1}) \tag{59}$$

holds with probability at least $1 - 5t^{-p} - 2e^{-u^2/2}$. Note that the singular values of $\mathbf{B}_{i+1}$ are the same as of $\hat{M}_{\{1,\ldots,i-1\}}(\hat{P}_{i+1..d}(\mathbf{x}))$; see, e.g., Figure 4. As shown by Hochstenbach and Reichel [31], the application of an orthogonal projection can only decrease the singular values. Thereby it follows that

$$\|(\mathbf{I} - \mathbf{Q}_i \mathbf{Q}_i^T) \mathbf{B}_{i+1}\|^2 \leq \eta^2 \sum_{k>r} \sigma_k^2(\mathbf{B}_{i+1}) \tag{60}$$

$$= \eta^2 \sum_{k>r} \sigma_k^2(\hat{M}_{\{1,\ldots,i-1\}}(\hat{P}_{i+1..d}(\mathbf{x}))) \tag{61}$$

$$\leq \eta^2 \sum_{k>r} \sigma_k^2(\hat{M}_{\{1,\dots,i-1\}}(\mathbf{x})) \tag{62}$$

$$\leq \eta^2 \min_{\operatorname{rank}(\hat{M}_{\{1,\dots,i-1\}}(\mathbf{y}))\leq r} \|\mathbf{x}-\mathbf{y}\|^2 \tag{63}$$

$$\leq \eta^2 \min_{\text{TT-rank}(\mathbf{y})\leq \mathbf{r}} \|\mathbf{x}-\mathbf{y}\|^2 . \tag{64}$$

As the random tensors $\mathbf{g}$ are sampled independently in each step, the combined probability that the above holds for all i is at least $\rho \geq (1 - 5t^{-p} - 2e^{-u^2/2})^{d-1}$, as asserted.

Note that if the tensor $\mathbf{x}$ actually has TT-rank $\mathbf{r}$ or smaller, that is, if $\min_{\text{TT-rank}(\mathbf{y})\leq \mathbf{r}} \|\mathbf{x}-\mathbf{y}\| = 0$, then the randomized TT-SVD is exact with probability one. This follows directly from theorem 2 by using $t \to \infty, u \to \infty$.

Using standard Gaussian random tensors, the computational complexity of the randomized TT-SVD is bounded by $\mathcal{O}(dsn^d)$, which is very similar to the deterministic TT-SVD presented in Section 2.1. However, as we show in the following proposition 1, for sparse tensors the complexity scales only *linearly* in the order, which is a dramatic reduction compared to the exponential scaling of the deterministic TT-SVD.

Proposition 1. *Assume that* $\mathbf{x} \in \mathbb{R}^{n_1 \times \dots \times n_d}$ *contains at most* N *non-zero entries. Then the computational complexity of the randomized TT-SVD given in Listing 2 scales as* $\mathcal{O}(d(s^2N + s^3n))$.

Proof. First note that if $\mathbf{x}$ has at most N non-zero entries, then each $\mathbf{b}_i$ has at most $s_{i-1}N$ non-zero entries. The fact that $\mathbf{x}$ has at most N non-zero entries implies that, independent of j, there are at most N tuples $(k_1, \dots, k_j)$ such that the subtensor $\mathbf{x}[k_1, \dots, k_j, \cdot, \dots, \cdot]$ is not completely zero. Now each $\mathbf{b}_i$ can be given as $\mathbf{b}_i = \mathbf{x} \circ_{(i,\dots,d),(2,\dots,d-i+1)} (\mathbf{W}_i \circ \dots \circ \mathbf{W}_d)$. As any contraction involving a zero tensors results in a zero tensor, $\mathbf{b}_i$ is non-zero only if the first $d - i + 1$ modes take values according to the at most N tuples. As there is only one further mode of dimension s_{i-1}, there can in total be only $s_{i-1}N$ non-zero entries in $\mathbf{b}_i$.

Creating only the, at most $s_{j-1}s_jN$, entries of $\mathbf{g}$ actually needed to perform the product $\mathbf{a}_j := \mathbf{g} \circ_{(2,\dots j),(1,\dots,j-1)} \mathbf{b}_{j+1}$, this calculation can be done in $\mathcal{O}(s_{j-1}s_jN)$. Calculating the QR of $\hat{M}_{(1)}(\mathbf{a}_j)$ has complexity $\mathcal{O}(s_{j-1}^2 n_j s_j)$. The involved (de-)matrification actually does not incur any computational costs. Finally the product $\mathbf{b}_j = \mathbf{b}_{j+1} \circ_{(j,j+1),(2,3)} \mathbf{W}_{d-j}$ has complexity $\mathcal{O}(s_{j-1}s_jN)$. These steps have to be repeated $d - 1$ times. Adding it all up, this gives an asymptotic cost bounded by $\mathcal{O}(d(s^2N + s^3n))$, where $s := \max(s_1, \dots, s_d)$. □

4 Relation to the Alternating Least Squares (ALS) Algorithm

There is an interesting connection between the proposed randomized TT-SVD and the popular alternating least squares (ALS) algorithm, which is examined in this section. Most of this section is still work in progress, but we consider sharing the ideas worthwhile nevertheless. The ALS itself is a general optimization algorithm, highly related to the very successful DMRG algorithm known in quantum physics. We provide only a minimal introduction and refer to the literature for an exhaustive treatment; see, e.g., [32, 33].

The ALS is used to solve optimization problems on the set of tensors with fixed TT-rank $\mathbf{r}$, for general objective functionals $\mathscr{J} : \mathbb{R}^{n_1 \times \ldots \times n_d} \to \mathbb{R}$. The special case interesting in this work is

$$\mathscr{J}(\mathbf{x}) := \|\mathbf{f} - \mathbf{x}\|^2$$

for a given tensor $\mathbf{f} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$. The global optimum is then exactly the best rank $\mathbf{r}$ approximation

$$\mathbf{x}^* := \operatorname{argmin}_{\text{TT-rank}(\mathbf{y})=\mathbf{r}} \|\mathbf{y} - \mathbf{f}\| \,. \tag{65}$$

Observing that the parametrization $\mathbf{x} = \tau(\mathbf{W}_1, \ldots, \mathbf{W}_d) = \mathbf{W}_1 \circ \ldots \circ \mathbf{W}_d$ is multilinear in the parameters $\mathbf{W}_i \in \mathbb{R}^{r_{i-1} \times n_i \times r_i}$. Hence fixing all components $\mathbf{W}_1$ except the i-th component $\mathbf{U}_i$ provides a parametrization of $\mathbf{x}$ which is linear in $\mathbf{U}_i \in \mathbb{R}^{r_{i-1} \times n_i \times r_i}$

$$\mathbf{x} := \mathbf{x}(\mathbf{U}_i) := \mathbf{W}_1 \circ \ldots \circ \mathbf{U}_i \circ \ldots \circ \mathbf{W}_d \tag{66}$$

Therefore the original optimization problem in the large ambient space $\mathbb{R}^{n_1 \times \ldots \times n_d}$ is restricted or projected onto a relatively small subspace $\mathbb{R}^{r_{i-1} \times n_i \times r_i}$, where it can be easily solved,

$$\mathbf{W}_i := \operatorname{argmin}_{\mathbf{U}_i \in \mathbb{R}^{r_{i-1} \times n_i \times r_i}} \left(\|\mathbf{W}_1 \circ \ldots \circ \mathbf{U}_i \circ \ldots \circ \mathbf{W}_d - \mathbf{f}\|^2 \right) \,. \tag{67}$$

This procedure is then continued iteratively by choosing another component $\mathbf{W}_j$ to be optimized next, resulting in a nonlinear Gauß Seidel iteration. The process of optimizing each component exactly once is often called a half-sweep.

Although these ideas can also be applied for the canonical format and general tensor networks as well, the tensor train and hierarchical Tucker format admit the possibility to use an orthonormal bases, which can be directly derived from the components $\mathbf{W}_i$ by corresponding (left/right) orthogonalization. With this simple post processing step, the ALS algorithm performs much better and more stable than without orthogonalization; *see* ,.g., [32]. To get started, the ALS algorithm requires an initial guess, i.e., it needs $d - 1$ (left /right) orthogonal components. A usual choice is to use (Gaussian) random tensors for the $d - 1$ components, possibly

orthogonalized. The interesting observation is that using this random initialization, one half-sweep can almost be cast to our proposed randomized TT-SVD and vice versa, in the sense that numerically exactly the same operations are performed. The only difference is that, in the picture of the randomized TT-SVD, the random tensor $\mathbf{g}$ is not chosen as a Gaussian random tensor in each step but as the first $d - i$ (contracted) random components of the ALS initialization. Note that this means that for the matrix case $d = 2$, the two methods actually coincide completely. Would it be possible to extend our error bounds to the setting of using structured random tensors $\mathbf{g}$ and also to cope with the stochastic dependence implied by the fact that $\mathbf{g}$ of different steps are not sampled independently, one could, for example, prove stochastic error bounds for the first sweep of the ALS, possible not only for the low-rank approximation setting but also for more general objective functionals $\mathscr{J}$. Numerical results indeed do suggest that such extensions might be possible. However, this is devoted to forthcoming research.

5 Numerical Experiments

In order to provide practical proof of the performance of the presented randomized TT-SVD, we conducted several numerical experiments. All calculations were performed using the *xerus* C++ toolbox [34], which also contains our implementation of the randomized TT-SVD. The random tensors used in the following experiments are created as follows. Standard Gaussian random tensors are created by sampling each entry independently from $\mathscr{N}(0, 1)$. Sparse random tensors are created by sampling N entries independently from $\mathscr{N}(0, 1)$ and placing them at positions sampled independently and evenly distributed from $[n_1] \times [n_2] \times \ldots \times [n_d]$. The low-rank tensors are created by sampling the entries of the component tensors $\mathbf{W}_1, \ldots, \mathbf{W}_d$ in representation (11), independently from $\mathscr{N}(0, 1)$, i.e., all components $\mathbf{W}_i$ are independent standard Gaussian random tensors. In some experiments we impose a certain decay for the singular values of the involved matrifications. To this end we create all random components as above; then for i from 1 to $d - 1$, we contract $\mathbf{W}_i \circ W_{i+1}$ and then re-separate them by calculating the SVD $\mathbf{USV}^T$ but replacing the $\mathbf{S}$ with a matrix $\tilde{\mathbf{S}}$ in which the singular values decay in the desired way. For a quadratical decay that is $\tilde{\mathbf{S}} := \mathrm{diag}(1, \frac{1}{2^2}, \frac{1}{3^2}, \ldots, \frac{1}{250^2}, 0, \ldots, 0)$, 250 is a cutoff used in all experiments below. Then set $\mathbf{W}_i = \hat{M}^{-1}(\mathbf{U})$ and $\mathbf{W}_{i+1} = \hat{M}^{-1}(\mathbf{SV}^T)$. Note that since the later steps change the singular values of the earlier matrification, the singular values of the resulting tensor do not obey the desired decay exactly. However empirically we observed that this method yields a sufficiently well approximation for most applications, even after a single sweep.

A general problem is that, as described in Section 2.1, the calculation of the actual best rank $\mathbf{r}$ approximation of higher-order tensors is NP-hard in general. Moreover to the author's knowledge, there are no nontrivial higher-order tensors for which this best approximation is known in advance. Therefore a direct check of our stochastic error bound (44) using the actual best approximation error is unfeasible.

Instead most of the numerical experiments use the error of the deterministic TT-SVD introduced in Section 2.1 for comparison, which gives a quasi-best approximation. The factor of $\sqrt{d-1}$ is present in both error bounds, but the remaining error dependence given by (44) is verifiable in this way.

If not stated otherwise, we use the same values for all dimensions $n_i = n$, target ranks $r_i = r$, and (approximate) ranks of the solution $r_i^* = r^*$. In derogation from this rule, the ranks are always chosen within the limits of the dimensions of the corresponding matrifications. For example, for $d = 8, n = 4, r = 20$, the actual TT-rank would be $\mathbf{r} = (4, 16, 20, 20, 20, 16, 4)$. 256 samples are calculated for each point, unless specified otherwise. The results for the randomized TT-SVD are obtained by calculating a rank $r + p$ approximation as described in Section 3.2 and then using the deterministic TT-SDV to truncate this to rank r. This is done so that in all experiments, the randomized and the deterministic approximations have identical final ranks $\mathbf{r}$.

5.1 Approximation Quality for Nearly Low-Rank Tensors

In this experiment we examine the approximation quality of the randomized TT-SVD for almost exact low-rank tensors, i.e., we create random TT-rank $\mathbf{r}^*$ tensors $\mathbf{x}_{\text{exact}} \in \mathbb{R}^{n \times \ldots \times n}$ and standard Gaussian random tensors $\mathbf{n} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$. The target tensor is then created as

$$\mathbf{x} := \frac{\mathbf{x}_{\text{exact}}}{\|\mathbf{x}_{\text{exact}}\|} + \tau \frac{\mathbf{n}}{\|\mathbf{n}\|}, \tag{68}$$

for some noise level τ. Subsequently rank $\mathbf{r}$ approximations $\mathbf{y}_{\text{det}}$ and $\mathbf{y}_{\text{rnd}}$ of $\mathbf{x}$ are calculated using the randomized and deterministic TT-SVD. Finally we examine the relative errors:

$$\epsilon_{\text{det}} := \frac{\|\mathbf{x} - \mathbf{y}_{\text{det}}\|}{\|\mathbf{x}\|} \qquad \epsilon_{\text{rnd}} := \frac{\|\mathbf{x} - \mathbf{y}_{\text{rnd}}\|}{\|\mathbf{x}\|}. \tag{69}$$

Figure 5 show these errors for different noise levels τ. The parameters are chosen as $d = 10, n = 4, r^* = 10, r = 10$, and $p = 5$, with 256 samples calculated for each method and noise level.

As expected the error of the classical TT-SVD almost equals the noise τ for all samples, with nearly no variance. Independent of the noise level, the error ϵ_{rnd} of the randomized TT-SVD is larger by a factor of approximately 1.6. The only exception is in the case $\tau = 0$ where both methods are exact up to numerical precision. In contrast to the classical TT-SVD, there is some variance in the error ϵ_{rnd}. Notably this variance continuously decreases to zero with the noise level τ. These observations are in agreement with the theoretical expectations, as theorem 2 states that the approximation error of the randomized TT-SVD is with high probability smaller than a factor times the error of the best approximation.

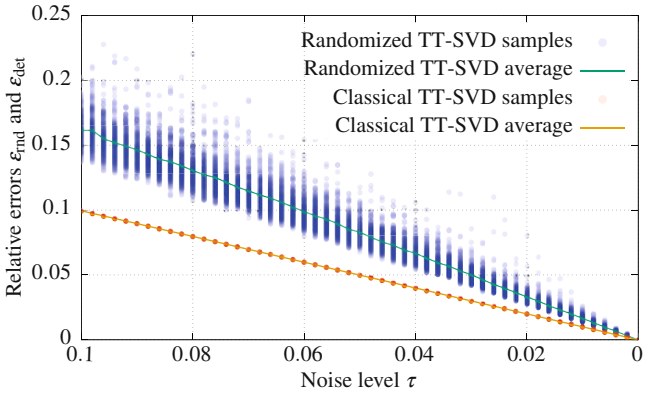


Fig. 5 Approximation error of deterministic and randomized TT-SVD in dependency to the noise level. Parameters are $d = 10,\ n = 4,\ r^* = 10,\ r = 10,\ p = 5$

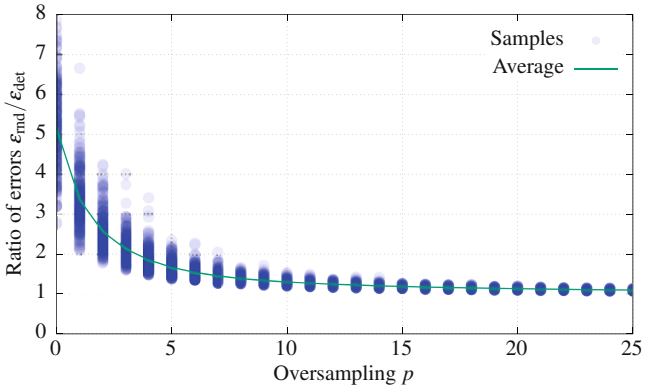


Fig. 6 Approximation error of the randomized TT-SVD in terms of the deterministic error in dependency of the oversampling for nearly low-rank tensors. Parameters are $d = 10,\ n = 4,\ r^* = 10,\ r = 10,\ \tau = 0.05$

5.2 Approximation Quality with Respect to Oversampling

In this experiment we examine the influence of the oversampling parameter p on the approximation quality. The first setting is similar to the one in experiment 5.1, i.e., nearly low-rank tensors with some noise. In contrast to experiment 5.1, the noise level $\tau = 0.05$ is fixed and the oversampling parameter p is varied. For each sample we measure the error of the approximation obtained by the randomized TT-SVD ϵ_{rnd} in terms of the error of the deterministic TT-SVD ϵ_{det}. The results for the parameters $d = 10, n = 4, r^* = 10, r = 10, \tau = 0.05$ are shown in Figure 6. For small p a steep decent of the error factor is observed which slowly saturates toward a factor of approximately one for larger p. The variance decreases at a similar pace.

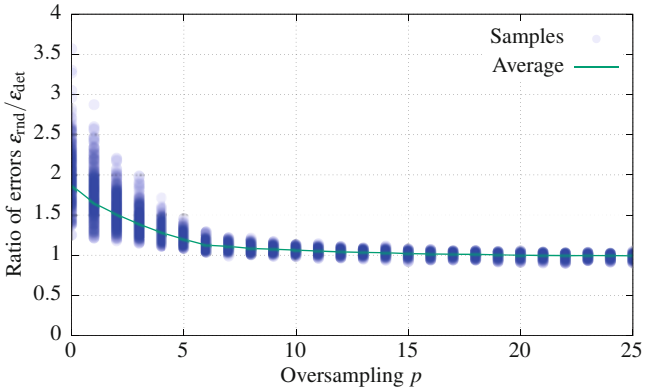


Fig. 7 Approximation error of the randomized TT-SVD in terms of the deterministic error in dependency of the oversampling for tensors with quadratically decaying singular values. Parameters are $d = 10,\ n = 4,\ r = 10$

In the second setting, tensors with quadratically decaying singular values are used; see the general remarks at the beginning of the section for the details of the creation. The behavior of the error factor is visualized in Figure 7 for the same parameters $d = 10, n = 4, r = 10$. There are several differences compared to the first setting. Most obvious for all p the factor is much smaller, i.e., the second setting is more favorable to the randomized TT-SVD. The same is also true for the variance. A more subtle difference, at least in the measured range of p, is that there are many samples for which the error factor is smaller than one, i.e., the randomized approximation is actually better than the deterministic one.

Very loosely speaking theorem 2 predicts a $1 + \frac{c}{\sqrt{p}}$ dependency of the error factor with respect to p, which is also roughly what is observed in both experiments.

5.3 *Approximation Quality with Respect to the Order*

In this third experiment, the impact of the order on the approximation quality is investigated. Again the first setting uses nearly low-rank tensors with some noise. The parameters are chosen as $d = 10, n = 4, r^* = 10, r = 10, \tau = 0.05$. The result is shown in Figure 8. As in experiment 5.2 in the second setting, target tensors with quadratically decaying singular values are used. The results for the parameters $d = 10, n = 4, r = 10$ are shown in Figure 9. For both settings the factor slightly increases from $d = 4$ to $d = 7$ but then stabilizes to constant values of approximately 1.65 and 1.95, respectively. The same qualitative behavior is observed for the variance. This is somewhat better than expected from the theoretical results. The factor $\sqrt{d-1}$ in the error term of Theorem 2 is not visible as it is also present in the error bound of the deterministic TT-SVD. However, the order also

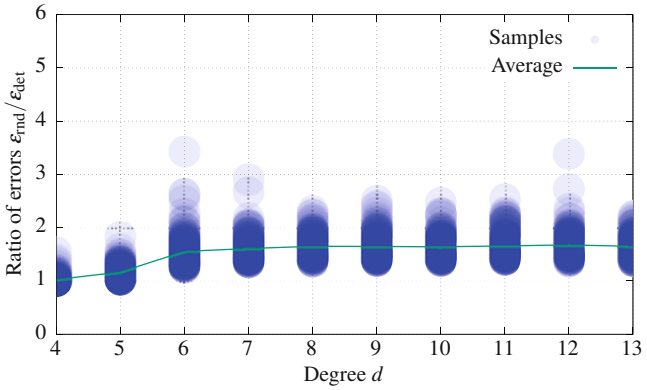


Fig. 8 Approximation error of the randomized TT-SVD in terms of the deterministic error in dependency of the order for nearly low-rank tensors. Parameters are $d = 10$, $n = 4$, $r^* = 10$, $r = 10$, $\tau = 0.05$

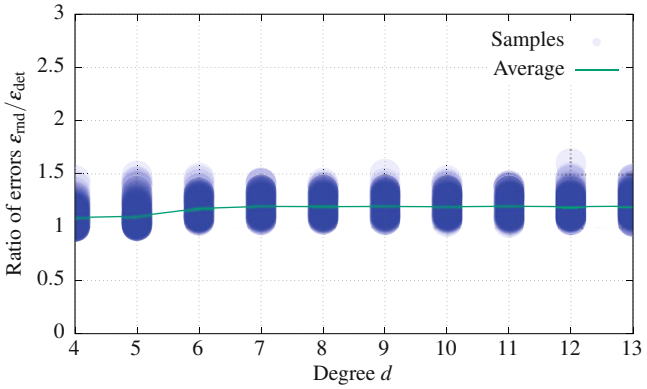


Fig. 9 Approximation error of the randomized TT-SVD in terms of the deterministic error in dependency of the order for tensors with quadratically decaying singular values. Parameters are $d = 10$, $n = 4$, $r = 10$

appears as an exponent in the probability, which should be observed in these results. The fact that it is not suggest that a refinement of theorem 2 is possible in which this exponent does not appear.

5.4 Computation Time

In this experiment we verify the computational complexity of the randomized TT-SVD algorithm, in particular the linear scaling with respect to the order for sparse tensors. To this end we create random sparse tensors with varying order and and a fixed number $N = 500$ entries and measure the computation time of the TT-SVD.

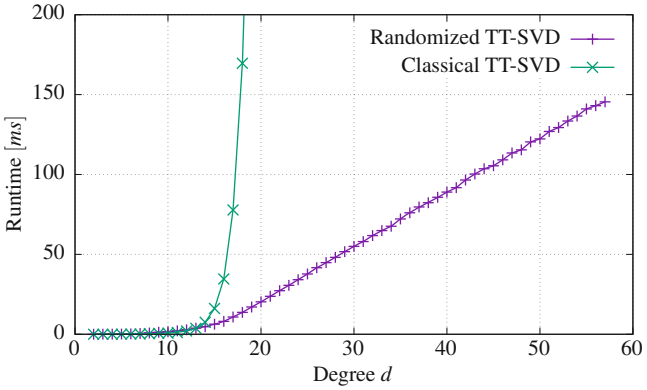


Fig. 10 Runtime of the deterministic and randomized TT-SVD algorithms for different orders. Parameters are $n = 2,\ r = 10,\ p = 10$

The other parameters are chosen as $n = 2, r = 10, p = 10$. Figure 10 shows the results which clearly confirm the linear scaling of the randomized TT-SVD. As a comparison also, the runtime of the classical TT-SVD is given for the smaller orders. While the absolute numbers are of course hardware and implementation depended, the dramatic edge of the randomized approach is obvious.

5.5 Approximation Quality Using Low-Rank Random Tensors

This final experiment uses the low-rank random tensor approach to the TT-SVD discussed in Section 4, i.e., instead of the proposed randomized TT-SVD, one half-sweep of the ALS algorithm with random initialization is performed. The remainder of the setting is the same as the second one of experiment 5.2, i.e., tensors with quadratically decaying singular values and parameters $d = 10, n = 4, r = 10$. Figure 11 shows the results and also as a comparison the average errors from experiment 5.2. Apparently the error factor using the ALS half-sweep is somewhat larger than the one of the randomized TT-SVD, but otherwise exhibits the same behavior with respect to the oversampling. While there are no theoretical results on this method yet, this result is encouraging as it suggest that error bounds similar to theorem 2 are possible for the ALS with random initialization.

6 Conclusions and Outlook

We have shown theoretically and practically that the randomized TT-SVD algorithm introduced in this work provides a robust alternative to the classical deterministic TT-SVD algorithm at low computational expenses. In particular the randomized

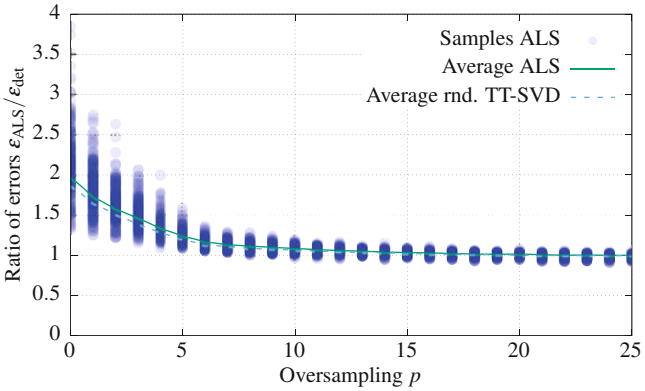


Fig. 11 Approximation error of one half-sweep of the ALS with random initialization in terms of the error of the deterministic TT-SVD in dependency of the oversampling. Tensors with quadratically decaying singular values are used along the with the parameters $d = 10$, $n = 4$, $r = 10$

TT-SVD is provably exact if applied to a tensor with TT-rank smaller or equal to the target rank. For the case of actual low-rank approximations, stochastic error bounds hold. The numerical experiments suggest that these proven bounds are somewhat pessimistic as the observed error is mostly significantly smaller than expected. Especially we do not observe a significant deterioration of the error bound with increased order, as suggested by the current theoretical results. This leaves room for improvements and we believe that enhanced versions of our theorem are possible. On the computational side, we have provided efficient implementations of the proposed algorithm, available in the *xerus* C++ toolbox. For sparse tensors, we have shown that the randomized TT-SVD algorithm dramatically outperforms the deterministic algorithm, scaling only linearly instead of exponentially in the order, which was verified by measurement of the actual implementation. We believe that these results show that the randomized TT-SVD algorithm is a useful tool for low-rank approximations of higher-order tensors.

In order to avoid repetition, we presented our randomized TT-SVD algorithm only for the popular tensor train format. Let us note however that the very same ideas can straight forwardly be applied to obtain an algorithm for a randomized HOSVD for the Tucker format. We expect that they can also be extended to obtain a randomized HSVD for the more general hierarchical Tucker format, but this is still work in progress. While an extension to the canonical polyadic format would certainly be desirable as well, we expect such an extension to be much more evolved, if possible at all.

A topic of further investigations is the use of structured random tensors in the randomized TT-SVD. For the matrix case, several choices of structured random matrices are already discussed in the work of Halko et al. [5]. Transferring their results to the high-dimensional case could allow choices of randomness which lead

to reduced computational cost if the given tensor is dense, as it is the case for matrices. The even more interesting choice however is to use random low-rank tensors, as already discussed in Section 4. On the one hand, an analysis of this setting directly benefits the alternating least squares algorithm, as it would result in error bound for the first half-sweep for a random initial guess. This can be of major importance as there are mainly local convergence theories for the ALS, which is why the starting point matters a lot. On the other hand, having error bounds also for this setting allows computationally fast application of the randomized TT-SVD to tensors given in various data-sparse formats, e.g., in the canonical, the TT or HT format, and also combination of those. This is, for example, important for the iterative hard thresholding algorithm for tensor completion, discussed in the introduction. Here in each iteration an SVD of a low rank plus a sparse tensor has to be calculated.

References

1. W. Hackbusch, S. Kühn, A new scheme for the tensor representation. J. Fourier Anal. Appl. **15**(5), 706–722 (2009)
2. I.V. Oseledets, Tensor-train decomposition. SIAM J. Sci. Comput. **33**(5), 2295–2317 (2011)
3. L. Grasedyck, D. Kressner, C. Tobler, A literature survey of low-rank tensor approximation techniques. GAMM-Mitteilungen **36**(1), 53–78 (2013)
4. N. Cohen, O. Sharir, A. Shashua, On the expressive power of deep learning: a tensor analysis (2015). arXiv preprint arXiv: 1509.05009 554
5. N. Halko, P.-G. Martinsson, J.A. Tropp, Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. **53**(2), 217–288 (2011)
6. H. Rauhut, R. Schneider, Ž. Stojanac, Tensor completion in hierarchical tensor representations, in *Compressed Sensing and its Applications* (Springer, Berlin, 2015), pp. 419–450
7. H. Rauhut, R. Schneider, Z. Stojanac, Low rank tensor recovery via iterative hard thresholding (2016). arXiv preprint arXiv: 1602.05217
8. E.J. Candès, B. Recht, Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717 (2009)
9. B. Recht, A simpler approach to matrix completion. J. Mach. Learn. Res. **12**, 3413–3430 (2011)
10. J.-F Cai, E.J. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**(4), 1956–1982 (2010)
11. M. Bachmayr, W. Dahmen, Adaptive near-optimal rank tensor approximation for high-dimensional operator equations. Found. Comput. Math. **15**(4), 839–898 (2015)
12. M. Bachmayr, R. Schneider, A. Uschmajew, Tensor networks and hierarchical tensors for the solution of high-dimensional partial differential equations. Found. Comput. Math. **16**(6), 1423–1472 (2016)
13. M. Bachmayr, R. Schneider, Iterative methods based on soft thresholding of hierarchical tensors. Found. Comput. Math. **17**(4), 1037–1083 (2017)
14. W. Hackbusch, *Tensor Spaces and Numerical Tensor Calculus*, vol. 42 (Springer Science & Business Media, New York, 2012)
15. A. Falcó, W. Hackbusch, On minimal subspaces in tensor representations. Found. Comput. Math. **12**(6), 765–803 (2012)
16. A. Falcó, W. Hackbusch, A. Nouy, Geometric structures in tensor representations (Final Release) (2015). arXiv preprint arXiv: 1505.03027

17. T.G. Kolda, B.W. Bader, Tensor decompositions and applications. SIAM Rev. **51**(3), 455–500 (2009)
18. J.Håstad, Tensor rank is NP-complete. J. Algorithms **11**(4), 644–654 (1990)
19. V. De Silva, L.-H. Lim, Tensor rank and the ill-posedness of the best low-rank approximation problem. SIAM J. Matrix Anal. Appl. **30**(3), 1084–1127 (2008). arXiv: math/0607647 [math.NA]
20. L.R. Tucker, Some mathematical notes on three-mode factor analysis. Psychometrika **31**(3), 279–311 (1966)
21. L. De Lathauwer, B. De Moor, J. Vandewalle, A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. **21**(4), 1253–1278 (2000)
22. C.J. Hillar, L.-H. Lim, Most tensor problems are NP-hard. J. ACM (JACM) **60**(6), 45 (2013). arXiv: 0911.1393 [cs.CC]
23. L. Grasedyck, Hierarchical singular value decomposition of tensors. SIAM J. Matrix Anal. Appl. **31**(4), 2029–2054 (2010)
24.] L. Grasedyck, W. Hackbusch, An introduction to hierarchical (H-) rank and TT-rank of tensors with examples. Comput. Methods Appl. Math. **11**(3), 291–304 (2011)
25. D. Perez-Garcia, F Verstraete, M.M. Wolf, J.I. Cirac. Matrix product state representations (2006). Preprint . arXiv: quant-ph/0608197 [quant-ph]
26. S. Holtz, T Rohwedder, R. Schneider, On manifolds of tensors of fixed TT-rank. Numer. Math. **120**(4), 701–731 (2012)
27. D. Kressner, M. Steinlechner, B. Vandereycken, Preconditioned low-rank Riemannian optimization for linear systems with tensor product structure. SIAM J. Sci. Comput. **38**(4), A2018–A2044 (2016)
28. M.M. Steinlechner, Riemannian optimization for solving high-dimensional problems with low-rank tensor structure. Ph.D Thesis. École polytechnique fédérale de Lausanne (2016)
29. C. Lubich, T. Rohwedder, R. Schneider, B. Vandereycken, Dynamical approximation by hierarchical Tucker and tensor-train tensors. SIAM J. Matrix Anal. Appl. **34**(2), 470–494 (2013)
30. C. Lubich, I.V. Oseledets, B. Vandereycken, Time integration of tensor trains. SIAM J. Numer. Anal. **53**(2), 917–941 (2015)
31. M.E. Hochstenbach, L. Reichel, Subspace-restricted singular value decompositions for linear discrete ill-posed problems. J. Comput. Appl. Math. **235**(4), 1053–1064 (2010)
32. S. Holtz, T. Rohwedder, R. Schneider, The alternating linear scheme for tensor optimization in the tensor train format. SIAM J. Sci. Comput. **34**(2), A683–A713 (2012)
33. M. Espig, W. Hackbusch, A. Khachatryan, On the convergence of alternating least squares optimisation in tensor format representations (2015). arXiv preprint arXiv: 1506.00062
34. B. Huber, S. Wolf, Xerus a general purpose tensor library (2014–2017). https://libxerus.org/

Versatile and Scalable Cosparse Methods for Physics-Driven Inverse Problems

Srđan Kitić, Siouar Bensaid, Laurent Albera, Nancy Bertin, and Rémi Gribonval

Abstract Solving an underdetermined inverse problem implies the use of a regularization hypothesis. Among possible regularizers, the so-called *sparsity* hypothesis, described as a *synthesis* (generative) model of the signal of interest from a low number of elementary signals taken in a dictionary, is now widely used. In many inverse problems of this kind, it happens that an alternative model, the *cosparsity* hypothesis (stating that the result of some linear *analysis* of the signal is sparse), offers advantageous properties over the classical synthesis model. A particular advantage is its ability to intrinsically integrate physical knowledge about the observed phenomenon, which arises naturally in the remote sensing contexts through some underlying partial differential equation. In this chapter, we illustrate on two worked examples (acoustic source localization and brain source imaging) the power of a generic cosparse approach to a wide range of problems governed by physical laws, how it can be adapted to each of these problems in a very versatile fashion, and how it can scale up to large volumes of data typically arising in applications.

Srđan Kitić contributed to the results reported in this chapter when he was with Univ Rennes, Inria, CNRS, IRSIA. The chapter was written while he was a postdoc with Technicolor, Rennes.

S. Kitić
Technicolor, Cesson-Sévigné, France
e-mail: srdan.kitic@orange.com

S. Bensaid · L. Albera
Université de Rennes 1 (LTSI), Rennes, France
e-mail: siouar.bensaid@univ-rennes1.fr; laurent.albera@univ-rennes1.fr

N. Bertin · R. Gribonval (✉)
Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France
e-mail: nancy.bertin@irisa.fr; remi.gribonval@inria.fr

H. Boche et al. (eds.), *Compressed Sensing and its Applications*,
Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-69802-1_10

Keywords Inverse problem · Sparse regularization · Cosparsity · Partial differential equation · Acoustic source localization · Brain source imaging · Electroencephalography

1 Introduction

Inverse source problems consist in inferring information from an object in an indirect manner, through the signals it emits or scatters. This covers in particular *remote sensing*, a term coined in Earth sciences [12], to indicate acquisition of shape or structure data of the Earth's crust, using signal processing techniques. More generally, remote sensing considers any method that collects distant observations from an object, with a variety of possible signal modalities and imaging modes (active or passive) that usually determine the applied processing technique.

Remote sensing and inverse problems encompass a wide range of practical applications, many of which play essential roles in various parts of modern lifestyle: medical ultrasound tomography, electroencephalography (EEG), magnetoencephalography (MEG), radar, seismic imaging, radio astronomy, etc.

To address inverse problems, an important issue is the apparent shortage of observed data compared to the ambient dimensionality of the objects of interest. A common thread to address this issue is to design low-dimensional models able to capture the intrinsic low dimensionality of these objects while allowing the design of scalable and efficient algorithms to infer them from partial observations.

The sparse data model has been particularly explored in this context [9, 75, 78]. It is essentially a generative *synthesis* model describing the object of interest as a sparse superposition of elementary objects (*atoms*) from a so-called *dictionary*. Designing the atoms in a given application scenario can be challenging. As documented in this chapter, exploiting dictionaries in the context of large-scale inverse problems can also raise serious computational issues.

An alternative model is the so-called *analysis* sparse model, or *cosparse* model, whereby the sparsity assumption is expressed on an analysis version of the object of interest, resulting from the application of a (differential) linear operator. As we will see, this alternative approach is natural in the context of many inverse problems where the objects of interest are physical quantities with properties driven by conservation or propagation laws. Indeed, the fact that such laws are expressed in terms of partial differential equations (PDEs) has several interesting consequences. First, using standard discretization schemes, the model (which is embodied by an *analysis operator* rather than a dictionary) can be directly deduced from the knowledge of these PDEs. Moreover, the resulting model description is often very concise, and the associated linear analysis operator is very sparse, leading to efficient computations. The framework thus fits very well into iterative algorithms for sparse regularization and large-scale convex optimization. Finally, the framework is adaptable to difficult settings where, besides the object of interest, some other "nuisance" parameters are unknown: uncalibrated sensors, partially known impedances, etc.

To demonstrate the scalability and versatility of this framework, this chapter uses as worked examples two practical scenarios involving two types of signals (namely, acoustic and electroencephalographic) for illustration purposes. The generic modeling and algorithmic framework of physics-driven cosparse methods which is described here has however the potential to adapt to many other remote sensing situations as well.

2 Physics-Driven Inverse Problems

Many quantities of interest that are measured directly or indirectly by sensors are intimately related to propagation phenomena governed by certain laws of physics. Linear *partial differential equations (PDEs)* are widely used to model such laws including sound propagation in gases (acoustic wave equation), electrodynamics (Maxwell equations), electrostatic fields (Poisson's equation), thermodynamics (heat equation), and even option pricing (Black-Scholes equation), among many others. When coupled with a sparsity assumption on the underlying sources, these lead to a number of feasible approaches to address physics-driven inverse problems.

2.1 Linear PDEs

Hereafter, ω denotes the coordinate parameter (e.g., space r and/or time t) of an open *domain* Θ. Linear PDEs take the following form:

$$\mathfrak{D}\boldsymbol{x}(\omega) := \sum_{|d|\leq\zeta} \boldsymbol{a}_d(\omega) D^d \boldsymbol{x}(\omega) = \boldsymbol{c}(\omega), \quad \omega \in \Theta, \tag{1}$$

where $\boldsymbol{a}_d$, $\boldsymbol{x}$, and $\boldsymbol{c}$ are functions of the variable ω. Typically one can think of $\boldsymbol{x}(\omega)$ as the propagated field and $\boldsymbol{c}(\omega)$ as the source contribution. The function $\boldsymbol{a}_d$ denotes coefficients that may (or may not) vary with ω.

Above, d is the multi-index variable with $|d| = \mathsf{d}_1 + \ldots + \mathsf{d}_\mathsf{l}$, $\mathsf{d}_\mathsf{i} \in \mathbb{N}_0$. For a given $d = (\mathsf{d}_1, \ldots, \mathsf{d}_\mathsf{l})$, $D^d\boldsymbol{x}(\omega)$ denotes the d^{th} partial differential of $\boldsymbol{x}$ with respect to ω, defined as:

$$D^d \boldsymbol{x}(\omega) := \frac{\partial^{|d|}\boldsymbol{x}}{\partial\omega_1^{\mathsf{d}_1}\partial\omega_2^{\mathsf{d}_2}\ldots\partial\omega_\mathsf{l}^{\mathsf{d}_\mathsf{l}}}.$$

In order for continuous $D^d\boldsymbol{x}(\omega)$ to exist, one needs to restrict the class of functions to which $\boldsymbol{x}(\omega)$, $\omega \in \Theta$ belongs. Such function spaces are called *Sobolev spaces*. In this chapter, functions are denoted by boldface italic lowercase (*e.g.,* $\boldsymbol{f}$), while linear operators acting on these are denoted by uppercase fraktur font (*e.g.,* $\mathfrak{D}, \mathfrak{L}$). For linear PDEs, linear *initial conditions* and/or *boundary conditions* are also considered, and we denote them as $\mathfrak{B}\boldsymbol{x} = \boldsymbol{b}$. Finally, we compactly write (1), equipped with appropriate boundary/initial conditions, in linear operator form:

$$\mathfrak{L}\boldsymbol{x} = \boldsymbol{c}, \tag{2}$$

where $\mathfrak{L} := (\mathfrak{D}, \mathfrak{B})$ and $\boldsymbol{c} := (\boldsymbol{c}, \boldsymbol{b})$, by abuse of notation. For simplicity, we consider only *self-adjoint* operators[1] $\mathfrak{L}$. With regard to remote sensing, $\mathfrak{L}$, $\boldsymbol{x}$, and $\boldsymbol{c}$ represent the "channel," the propagated, and the "source" signal, respectively.

2.2 Green's Functions

While our final goal is to find $\boldsymbol{c}$ given partial observations of $\boldsymbol{x}$, let us assume, for now, that $\boldsymbol{c}$ is given and that we want to infer the solution $\boldsymbol{x}$.

The existence and uniqueness of solutions $\boldsymbol{x}$ of PDEs, in general, is an open problem. These are subject to certain boundary and/or initial conditions, which constrain the behavior of the solution at the "edge" $\partial\Theta$ of the domain Θ. Fortunately, for many types of PDEs, the required conditions are known, such as those provided by Cauchy-Kowalevski theorem, for PDEs with analytic coefficients. Hence, we do not dwell on this issue; instead, we assume that the unique solution exists (albeit, it can be very unstable – PDEs represent archetypal ill-posed problems [40]).

Looking at (2), one may ask whether there exist an inverse operator $\mathfrak{L}^{-1}$, such that we can compute the solution as $\boldsymbol{x} = \mathfrak{L}^{-1}\boldsymbol{c}$. Indeed, in this setting such operator exists and is the gist of the method of *Green's functions* for solving PDEs. The operator is (as expected) of integral form, and its kernel is given by the Green's functions $g(\omega, s)$, defined as follows:

$$\begin{aligned} \mathfrak{D}\boldsymbol{g}(\omega, s) &= \delta(\omega - s), \quad s \in \Theta, \\ \mathfrak{B}\boldsymbol{g}(\omega, s) &= 0, \quad s \in \partial\Theta, \end{aligned} \tag{3}$$

where $\delta(\cdot)$ represents Dirac's delta distribution. In signal processing language, the Green's function can be seen as the response of the system (2) to the impulse centered at $s \in \Theta$.

If we assume that $\boldsymbol{b} = 0$ on $\partial\Theta$, it is easy to show that the solution of a linear PDE is obtained by integrating the right-hand side of (2). Namely, since

$$\mathfrak{L}\boldsymbol{x}(\omega) = c(\omega) = \int_{\Theta} \delta(s - \omega)\boldsymbol{c}(s)\boldsymbol{d}s = \mathfrak{L}\int_{\Theta} \boldsymbol{g}(s, \omega)\boldsymbol{c}(s)\boldsymbol{d}s, \tag{4}$$

we can identify $\boldsymbol{x}(\omega)$ with the integral. Note that $\boldsymbol{g}(s, \omega) = \boldsymbol{g}(\omega, s)$ for a self-adjoint operator $\mathfrak{L}$, and the latter can be put in front of integration since it acts on ω. When the boundary conditions are inhomogeneous ($\boldsymbol{b} \neq 0$), the same approach can be taken except that one needs two types of Green's functions: one defined as in (3), and

[1] The operators for which $\langle \mathfrak{L}\boldsymbol{p}_1, \boldsymbol{p}_2 \rangle = \langle \boldsymbol{p}_1, \mathfrak{L}\boldsymbol{p}_2 \rangle$ holds. Otherwise, setting the adjoint boundary conditions would be required.

another with $\delta(\omega - s)$ placed at the boundary (but with $\mathfrak{D}\boldsymbol{g} = 0$, otherwise). Then, one obtains $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ from (4), and the solution $\boldsymbol{x}$ is superposition: $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$.

Since integration is again a linear operation, we can compactly represent (4) in the form of a linear operator:

$$\boldsymbol{x} = \mathfrak{G}\boldsymbol{c}, \tag{5}$$

and, by comparing it with (2), we can deduce that $\mathfrak{G} = \mathfrak{L}^{-1}$.

Green's functions are available in analytic form only for a restricted set of combinations of domain geometries/initial/boundary conditions. In such a case, evaluating $\boldsymbol{g}(\cdot, \cdot)$ is direct, but the integral (4) may still be difficult to evaluate to obtain the integral operator $\mathfrak{G}$. Functional approximations, such as using a *parametrix* [32], can be helpful; however in the most general case, one will have to resort to numerical approximations, as will be discussed in Section 4.

2.3 Linear Inverse Problem

The inverse problem we are interested in is the *estimation of the field* $\boldsymbol{x}$ (or, equivalently, of the source component $\boldsymbol{c}$) from a limited set of (noisy) field measurements acquired at a sensor array. In the case of a spatial field, such as a heat map, the coordinate parameter is spatial $\omega = r$, and each measurement is typically a scalar estimate of the field at m given locations, $\mathrm{y_j} \approx \boldsymbol{x}(r_\mathrm{j})$, perhaps after some spatial smoothing. In the case of a time series, the coordinate parameter is time $\omega = t$, and the measurements are obtained by analog-to-digital sampling (usually at a fixed sampling rate) at t time instants, corresponding to $\mathrm{y}_\ell \approx (\boldsymbol{h} \star \boldsymbol{x})(\mathrm{t}_\ell)$, where $\boldsymbol{h}(\mathrm{t})$ is a temporal filter, optionally applied for temporal smoothing. In the case of a spatiotemporal field, such as an acoustic pressure field, $\omega = (r, \mathrm{t})$ and the acquired measurement consist of multiple *time series* obtained by analog-to-digital sampling (at a fixed sampling rate) at a number of spatial locations, corresponding to $\mathrm{y}_{\mathrm{j},\ell} \approx (\boldsymbol{h} \star_t \boldsymbol{x})(r_\mathrm{j}, \mathrm{t}_\ell)$ with $\star_t$ a convolution along the temporal dimension. Except when the nature of ω is essential for discussion, we use below the generic notation $\mathrm{y_j} \approx \boldsymbol{x}(\omega_\mathrm{j})$.

Now, we consider a vector $y \in \mathbb{R}^{\mathsf{m}}$ (resp. $\in \mathbb{R}^{\mathsf{t}}$ or $\in \mathbb{R}^{\mathsf{m}\times\mathsf{t}}$) of measurements as described above. Without additional assumptions, recovering $\boldsymbol{c}$ or $\boldsymbol{x}$ given the *measurement vector* y only is impossible. Understanding that (2) and (5) are dual representations of the same physical phenomenon, we term such problems *physics-driven inverse problems*.

Algebraic Methods Rigorous algebraic methods for *particular* instances of inverse source problems have been thoroughly investigated in the past and are still a very active area of research. The interested reader may consult, *e.g.,* [40], and the references therein. A more generic technique (which doesn't prescribe a specific PDE), closely related to the numerical framework we will discuss in next sections, is proposed in [60]. However, its assumptions are strong (although not

completely realistic), including in particular the availability of analytic expressions for the Green's functions. In addition, they must respect the approximate Strang-Fix conditions [26], and the source signal must follow the *finite rate of innovation (FRI)* [82] model, *i.e.,* it should be a weighted sum of Dirac impulses. In practice, it also requires a high *signal-to-noise ratio (SNR)*, or sophisticated numerical methods in order to combat sensitivity to additive noise, and possibly a large number of sensors.

Sparse Regularization Fortunately, in reality, the support of the source contribution $\boldsymbol{c}$ in (1) is often confined to a much smaller subset $\Theta_0 \subset \Theta$ (*i.e.,* $\boldsymbol{c}(\omega) = 0,\ \forall \omega \in \Theta_0^c$), representing *sources*, *sinks*, or other singularities of a physical field. Two practical examples will be given soon in Section 3. This crucial fact is exploited in many regularization approaches for inverse source problems, including the one described in this chapter. Sparsity-promoting regularizers often perform very well in practice (at least empirically) [52, 66–68]. Even though there exist pathological inverse source problems so severely ill-posed that the source sparsity assumption alone is not sufficient [20, 23], such cases seem to rarely occur in practice. Hence, sparse regularization can be considered as an effective heuristics for *estimating* solutions of various inverse source problems, although not as an all-purpose rigorous methodology for all physics-driven inverse problems.

3 Worked Examples

As an illustration, in the rest of this chapter we consider two physics-driven inverse problems: acoustic source localization (driven by the acoustic wave equation) and brain source localization (driven by Poisson's equation).

3.1 Acoustic Source Localization from Microphone Measurements

The problem we are concerned with is determining the position of one or more sources of sound based solely on microphone recordings. The problem arises in different fields, such as speech and sound enhancement [34], speech recognition [4], acoustic tomography [58], robotics [80], and aeroacoustics [45]. Traditional approaches based on time difference of arrival (TDOA) dominate the field [8], but these usually only provide the direction of arrival of the sound sources and are generally sensitive to reverberation effects. We take a different approach and use the physics of acoustic wave propagation to solve the localization problem.

The Wave Equation Sound is the manifestation of acoustic pressure, which is a function of position r and time t. Acoustic pressure $\boldsymbol{x} := \boldsymbol{x}(r, \mathrm{t})$, in the presence of a sound source, respects the inhomogeneous acoustic wave equation:

$$\Delta \boldsymbol{x} - \frac{1}{\boldsymbol{v}^2} \frac{\partial^2}{\partial \mathrm{t}^2} \boldsymbol{x} = \boldsymbol{c}, \quad (r, \mathrm{t}) \in \Theta := \Gamma \times (0, \tau), \tag{6}$$

where Γ denotes the spatial domain, Δ is the Laplacian operator with respect to the spatial variable r, and $\boldsymbol{v}$ is the speed of sound (around $334\text{m} \cdot \text{s}^{-1}$ at room temperature, it may depend on space and/or time but is often approximated as constant). The right-hand side $\boldsymbol{c} := \boldsymbol{c}(r, \text{t})$ represents the pressure emitted by a sound source at position r and time t, if any (if a source is not emitting at some time instant t, then $\boldsymbol{c}(r, \text{t})$ is zero at this source position; as an important consequence, $\boldsymbol{c} = 0$ everywhere, but at the source positions).

Initial and Boundary Conditions To ensure self-adjointness, we impose homogeneous initial and boundary conditions,

$$\forall r \in \Gamma,\ \boldsymbol{x}(r, 0) = 0,\ \frac{\partial}{\partial \text{t}} \boldsymbol{x}(r, 0) = 0, \tag{7}$$

i.e., the acoustic field is initially at rest.

In addition, we may impose *Dirichlet* ($\boldsymbol{x} \mid_{\partial\Gamma} = 0$) or *Neumann* ($\nabla \boldsymbol{x} \cdot \boldsymbol{n} \mid_{\partial\Gamma} = 0$) boundary conditions, where $\nabla \boldsymbol{x}$ is the spatial gradient (with respect to r) and $\boldsymbol{n}$ is the outward normal vector to the boundary $\partial\Gamma$. A generalization is the so-called *Robin* boundary condition, which models reflective boundaries, or *Mur's* boundary condition [59]:

$$\forall r \in \partial\Gamma, \forall \text{t}:\ \frac{\partial \boldsymbol{x}}{\partial \text{t}} + \boldsymbol{v}\xi \nabla \boldsymbol{x} \cdot \boldsymbol{n} = 0, \tag{8}$$

where $\xi \geq 0$ is the specific acoustic impedance (again, possibly dependent on space and time but reasonably considered as fixed over time). For $\xi \approx 1$, Mur's condition *approximates* an absorbant boundary.

Inverse Problem An array consisting of m omnidirectional microphones, with a known geometry, outputs the measurements assembled into the vector $y \in \mathbb{R}^{\mathsf{mt}}$, where t is the number of time samples. Thus, we assume that the microphones output discrete signals, with an antialiasing filter and sampler applied beforehand. The goal is, ideally, to recover the fields $\boldsymbol{x}$ or $\boldsymbol{c}$ from the data y, using prior information that the sound sources are sparse in the spatial domain Γ.

Related Work Sound source localization through wavefield extrapolation and low-complexity regularization was first introduced by Malioutov et al. in [54]. They assumed a free-field propagation model, which allowed them to analytically compute the associated Green's functions. The narrowband sound sources were estimated by applying sparse synthesis or low-rank regularizations. A wideband extension was proposed in [53], which is, however, a two-stage approach that implicitly depends on solving the narrowband problem. The free space assumption was first abandoned by Dokmanić and Vetterli [24, 25], for source localization in the frequency domain. They used the Green's function dictionary numerically computed by solving the Helmholtz equation with Neumann boundary conditions, by the *finite element method (FEM)*. The wideband scenario was tackled as a jointly sparse problem, to which, in order to reduce computational cost, a modification of the

OMP algorithm was applied. However, as argued in [17], this approach is critically dependent on the choice of frequencies and can fail if modal frequencies are used. Le Roux et al. [50] proposed to use the CoSaMP algorithm for solving the sparse synthesis problem in the same spirit.

3.2 Brain Source Localization from EEG Measurements

Electrical potentials produced by neuronal activity can be measured at the surface of the head using electroencephalography (EEG). The localization of sources of this neuronal activity (during either cognitive or pathological processes) requires a so-called *head model*, aiming at representing geometrical and electrical properties of the different tissues composing the volume conductor, as well as a *source model*.

Poisson's Equation It is commonly admitted that the electrical potential $\boldsymbol{x} := \boldsymbol{x}(\boldsymbol{r})$ at location $\boldsymbol{r}$ within the human head mostly reflects the activity of pyramidal cells located in the gray matter and oriented perpendicularly to the cortical surface. This activity is generally modeled by current dipoles. Given the geometry and the scalar field $\{\boldsymbol{\sigma}(\boldsymbol{r})\}$ of electrical conductivities at location $\boldsymbol{r}$ within the head, Poisson's equation [10, 55] relates the electrical potential $\boldsymbol{x}$ and the current density j:

$$-\nabla \cdot (\boldsymbol{\sigma}\, \nabla \boldsymbol{x}) = \nabla \cdot j, \qquad \boldsymbol{r} \in \Theta \tag{9}$$

where Θ is the spatial domain (interior of the human head) and $\nabla \cdot j$ is the volume current. The operators $\nabla\cdot$ and ∇, respectively, denote the divergence and gradient with respect to the spatial variable $\boldsymbol{r}$.

Boundary Condition We assume the Neumann boundary condition,

$$\boldsymbol{\sigma} \nabla \boldsymbol{x} \cdot \boldsymbol{n} = 0, \qquad \boldsymbol{r} \in \partial\Theta, \tag{10}$$

which reflects the absence of current outside the human head.

Inverse Problem An array consisting of m electrodes located on the scalp (see Figure 2) captures the EEG signal $y = [\boldsymbol{x}(r_1), \ldots, \boldsymbol{x}(r_{\mathsf{m}})]^{\mathsf{T}} \in \mathbb{R}^{\mathsf{m}}$. The brain source localization problem consists in recovering from y the electrical field inside the head (with respect to some reference electrical potential), $\boldsymbol{x}$, or the current density, j, under a sparsity assumption on the latter.

Related Work Numerous methods for brain source localization were developed to localize equivalent current dipoles from EEG recordings. Among them, beamforming techniques [65], subspace approaches [1, 57, 72], and sparse methods [79] are the most popular. Regarding dictionary-based sparse techniques, the most famous is MCE (minimum current estimate) [79], which computes minimum ℓ_1-norm estimates using a so-called *leadfield matrix*, corresponding to discretized Green's functions sampled at the electrode locations.

4 Discretization

A key preliminary step in the deployment of numerical methods to address inverse problems lies in the discretization of the quantities at hand, which amounts to convert the continuous PDE model (2) into a finite-dimensional linear model. A priori, any discretization method could be used within the regularization framework we propose; here, we limit ourselves to two families of them among the most common.

4.1 Finite-Difference Methods (FDM)

The simplest way to discretize the original continuous PDE is to replace the derivatives by *finite differences* obtained from their Taylor's expansion at a certain order, after discretization of the variable domain Θ itself on a (generally regular) grid. Consider a grid of discretization nodes $\{\omega^\ell\}_{\ell \in I}$ for the domain Θ and its boundary $\partial\Theta$. For each (multidimensional) index ℓ corresponding to the interior of the domain, the partial derivative $D^d \boldsymbol{x}(\omega^\ell)$ is approximated by finite linear combination of values of the vector $x = [\boldsymbol{x}(\omega^{\ell'})]_{\ell' \in I}$ associated to indices ℓ' such that $\omega^{\ell'}$ is in the neighborhood of ω^ℓ. The *stencil* defining these positions, as well as the order of the approximation, characterizes a particular FDM. A similar approach defines approximations to partial derivatives at the boundary and/or initial points.

Example: The Standard Leapfrog Method (LFM) As an example, we describe here the standard *Leapfrog Method (LFM)* applied to the discretization of a 2D, isotropic acoustic wave equation (6). Here, the domain Θ is three-dimensional, with variables r_x, r_y (two spatial coordinates) and t (time). The corresponding PDE to be discretized only involves second-order derivatives of $\boldsymbol{x}$ with respect to these variables. By denoting $\mathrm{x}^\tau_{\mathrm{i,j}} := \boldsymbol{x}(\omega^\tau_{\mathrm{i,j}})$ the field value at grid position $\omega^\tau_{\mathrm{i,j}} := (r_x(\mathrm{i}), r_y(\mathrm{j}), \tau)$, the LFM approximation is (for $\tau > 2$ and excluding the boundaries):

$$\mathfrak{D}\boldsymbol{x}(\omega^\tau_{\mathrm{i,j}}) = \left(\frac{\partial^2}{\partial r_x^2} + \frac{\partial^2}{\partial r_y^2} - \frac{1}{\mathrm{v}^2}\frac{\partial^2}{\partial t^2}\right)\boldsymbol{x}(\omega^\tau_{\mathrm{i,j}}) \approx \frac{x^\tau_{\mathrm{i-1,j}} - 2x^\tau_{\mathrm{i,j}} + x^\tau_{\mathrm{i+1,j}}}{\mathrm{d_x^2}} + \frac{x^\tau_{\mathrm{i,j-1}} - 2x^\tau_{\mathrm{i,j}} + x^\tau_{\mathrm{i,j+1}}}{\mathrm{d_y^2}} - \frac{1}{\mathrm{v}^2}\frac{x^{\tau+1}_{\mathrm{i,j}} - 2x^{\mathrm{t}}_{\mathrm{i,j}} + x^{\tau-1}_{\mathrm{i,j}}}{\mathrm{d_\tau^2}}, \tag{11}$$

where $\mathrm{d_x}, \mathrm{d_y}$ and d_τ denote the discretized spatial and temporal step sizes, respectively. This FDM scheme can be summarized as the use of a 7-point stencil centered at $x^\tau_{\mathrm{i,j}}$ in this case. It is associated to a finite-dimensional linear operator D such that Dx approximates the discretized version of $\mathfrak{D}\boldsymbol{x}(\omega)$ in the interior of the domain. The approximation error is of the order of $O(\max(\mathrm{d_x}, \mathrm{d_y}, \mathrm{d_t})^2)$.

Similar formulas for boundary nodes are obtained by substituting a nonexistent spatial point in the scheme (11) by the expressions obtained from discretized boundary conditions. For instance, for the frequency-independent acoustic, absorbing boundary condition (8), proposed in [48], *e.g.,* the missing point $x^{\tau}_{\mathsf{i}+1,\mathsf{j}}$ behind the right "wall," is evaluated as:

$$x^{\tau}_{\mathsf{i}+1,\mathsf{j}} = x^{\tau}_{\mathsf{i}-1,\mathsf{j}} + \frac{\mathsf{d_x}}{\mathsf{v}\mathrm{d}_\tau \xi_{\mathsf{i},\mathsf{j}}} \left(x^{\tau-1}_{\mathsf{i},\mathsf{j}} - x^{\tau+1}_{\mathsf{i},\mathsf{j}} \right). \tag{12}$$

When corners (and edges in 3D) are considered, the condition (8) is applied to all directions where the stencil points are missing. Combining (12) and (11) yields a linear operator B such that $Bx = b$ approximates the discretized version of $\mathfrak{B}\boldsymbol{x}(\omega) = \boldsymbol{b}(\omega)$ on the initial/boundary points of the domain.

Concatenating D and B yields a square matrix Ω (of size $\mathsf{n} = \mathsf{st}$ where s is the size of the spatial grid and t the number of time samples).

Using LFM to Solve the Discretized Forward Problem While we are interested to use the above discretization to solve *inverse* problems, let us recall how it serves to address the *forward* problem, i.e., to estimate $\boldsymbol{x}$ when $\boldsymbol{c}$ is given. Under the assumption $Dx = c$, the leapfrog relation (11) allows to compute $\mathsf{x}^{\tau+1}_{\mathsf{i},\mathsf{j}}$ using $c^{\tau}_{\mathsf{i},\mathsf{j}}$ and values of x at two previous discrete time instants ($x^{\tau}_{(\cdot,\cdot)}$ and $x^{\tau-1}_{(\cdot,\cdot)}$). Similarly, homogeneous boundary conditions ($b = 0$) translate into relations between neighboring values of x on the boundaries and over time. For example, the above described discretization of Mur's boundary condition yields an explicit expression of $x^{\tau+1}_{\mathsf{i},\mathsf{j}}$ at the boundary (see Equation (49) for details.) Neglecting approximation errors, LFM discretization thus yields a convenient *explicit* scheme [51] to solve $\Omega x = c$. An example of a 2D acoustic field discretized by this method is presented in Figure 1. Numerical stability of explicit FDM schemes, such as LFM, can only be ensured if the step sizes respect some constraining condition, such as the Courant-Friedrich-Lewy condition for hyperbolic PDEs [51]. In the abovementioned example, for instance, this condition translates into $\mathsf{v}\mathrm{d}_\tau / \min(\mathsf{d_x}, \mathsf{d_y}) \leq 1/\sqrt{2}$. This limits the resolution (for instance in space and time) achievable by these methods.

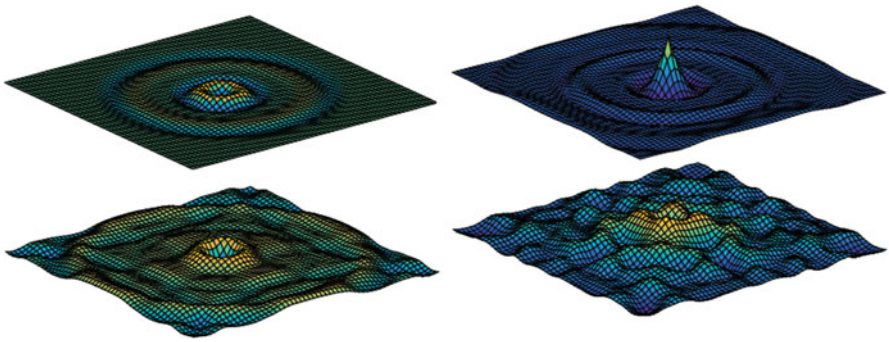


Fig. 1 Example of a discretized 2D acoustic pressure field at different time instants

4.2 Finite Element Methods (FEM)

The finite element method (FEM) is a numerical approximation used to solve boundary value problems when the solution is intractable analytically due to geometric complexities and inhomogeneities. Among several variants, the Galerkin FEM is famous as it is both well-rooted in theory and simple in derivation [35]. In the Galerkin FEM, a solution is computed in three main steps: 1) the formulation of the problem in its *weak/variational* form, 2) the discretization of the formulated problem, and 3) the choice of the approximating subspace. As an illustrative example, let's consider the well-known problem of Poisson's equation (9) with Neumann boundary condition (10).

Weak/Variational Formulation The first step aims at expressing the aforementioned PDE in an algebraic form. For a given test function $\boldsymbol{w}(r)$ in some (to be specified) Hilbert space of regular functions $\boldsymbol{H}$, we have, denoting $\boldsymbol{c} = \nabla.j$ the volume current which serves as a source term,

$$\begin{aligned}\int_{\Theta} \boldsymbol{c}(r)\boldsymbol{w}(r)dr &= -\int_{\Theta} \nabla \cdot (\sigma(r)\nabla \boldsymbol{x}(r))\boldsymbol{w}(r)dr \\ &= -\int_{\partial\Theta} n \cdot (\sigma(r)\nabla \boldsymbol{x}(r))\boldsymbol{w}(r)dr + \int_{\Theta} \nabla \boldsymbol{w}(r) \cdot (\sigma(r)\nabla \boldsymbol{x}(r))dr \\ &= \int_{\Theta} \nabla \boldsymbol{w}(r) \cdot (\sigma(r)\nabla \boldsymbol{x}(r))dr. \end{aligned} \tag{13}$$

The second line in (13) is derived using Green's identity which is the multidimensional analogue of integration by parts [35], whereas, the last line is deduced from the Neumann boundary condition (10). Notice that the resulting equality in (13) can be written as

$$\mathfrak{a}(\boldsymbol{x}, \boldsymbol{w}) = \mathfrak{b}(\boldsymbol{w}) \qquad \forall \boldsymbol{w} \in \boldsymbol{H} \tag{14}$$

where $\mathfrak{a}(.,.)$ is a symmetric bilinear form on $\boldsymbol{H}$ and $\mathfrak{b}(.)$ is a linear function on $\boldsymbol{H}$. The equality in (14) is referred to as the *weak* or the *variational* formulation of the PDE in (9)–(10), a name that stems from the less stringent requirements put on the functions $\boldsymbol{x}$ and $\boldsymbol{c}$. In fact, the former should be differentiable only once (vs. twice in the *strong* formulation), whereas the latter needs to be integrable (vs. continuous over $\bar{\Theta}$ in the *strong* formulation). These relaxed constraints in the *weak* form make it relevant to a broader collection of problems.

Discretization with the Galerkin Method In the second step, the Galerkin method is applied to the *weak* form. This step aims at discretizing the continuous problem in (14) by projecting it from the infinite-dimensional space $\boldsymbol{H}$ onto a finite-dimensional subspace $\boldsymbol{H}_h \subset \boldsymbol{H}$ (h refers to the precision of the approximation).

Denoting $\{\boldsymbol{\phi}_\ell\}_{\ell \in I}$ a basis of $\boldsymbol{H}_h$, any function $\boldsymbol{x}$ in the finite-dimensional subspace $\boldsymbol{H}_h$ can be written in a unique way as a linear combination $\boldsymbol{x} = \sum_{\ell \in I} x_\ell \boldsymbol{\phi}_\ell$. Therefore, given $\boldsymbol{x}$ a solution to the problem and if we take as a test function $\boldsymbol{w}$ a basis function $\boldsymbol{\phi}_\mathsf{i}$, the discrete form of (14) is then expressed as

$$\sum_{\ell \in I} \mathfrak{a}_h(\boldsymbol{\phi}_\mathsf{i}, \boldsymbol{\phi}_\ell) x_\ell = \mathfrak{b}_h(\boldsymbol{\phi}_\mathsf{i}) \qquad \forall \mathsf{i} \in I \tag{15}$$

where

$$\mathfrak{a}_h(\boldsymbol{\phi}_\mathsf{i}, \boldsymbol{\phi}_\ell) := \int_{\Theta_h} \sigma(r) \nabla \boldsymbol{\phi}_\mathsf{i}(r) \cdot \nabla \boldsymbol{\phi}_\ell(r) dr \tag{16}$$

$$\mathfrak{b}_h(\boldsymbol{\phi}_\mathsf{i}) := \int_{\Theta_h} \boldsymbol{c}(r) \boldsymbol{\phi}_\mathsf{i}(r) dr \tag{17}$$

with Θ_h a discretized solution domain (see next). The discretization process thus results in a linear system of $\mathsf{n} := card(I)$ equations with n unknowns $\{x_\ell\}_{\ell \in I}$, which can be rewritten in matrix form $\Omega x = c$. The so-called *global stiffness matrix* Ω is a symmetric matrix of size $\mathsf{n} \times \mathsf{n}$ with elements $\Omega_{ij} = \mathfrak{a}_h(\boldsymbol{\phi}_\mathsf{i}, \boldsymbol{\phi}_\mathsf{j})$, and c is the *load vector* of length n and elements $c_\mathsf{i} = \mathfrak{b}_h(\boldsymbol{\phi}_\mathsf{i})$. Notice that, in the case where $\sigma(r)$ is a positive function (as in EEG problem for instance), the matrix Ω is also positive semidefinite. This property can be easily deduced from the bilinear form $\mathfrak{a}(.,.)$ where $\mathfrak{a}(\boldsymbol{x}, \boldsymbol{x}) = \int_\Theta \sigma(r)(\nabla \boldsymbol{x}(r))^2 dr \geq 0$ for any function $\boldsymbol{x} \in \boldsymbol{H}$ and from the relationship $\mathfrak{a}(\boldsymbol{x}, \boldsymbol{x}) = \boldsymbol{x}^T \Omega \boldsymbol{x}$.

For the considered Poisson's equation, the stiffness matrix Ω is also rank deficient by one. This comes from the fact that $\boldsymbol{x}$ can only be determined up to an additive constant (corresponding to an arbitrary choice of reference for the electrical potential it represents), since only the gradient of $\boldsymbol{x}$ appears in Poisson's equation with Neumann boundary condition.

Choice of the Approximating Subspace and Discretization Basis The construction of the discretized solution domain Θ_h and the choice of basis functions $\{\boldsymbol{\phi}_\ell\}_{\ell \in I}$ are two pivotal points in FEM since they deeply affect the accuracy of the approximate solution obtained by solving $\Omega x = c$. They also impact the sparsity and conditioning of Ω, hence the computational properties of numerical schemes to solve this linear system.

In FEM, the domain is divided uniformly or nonuniformly into discrete elements composing a mesh, either of triangular shape (tetrahedral in 3D) or rectangular shape (hexahedral in 3D). The triangular (tetrahedral) mesh is often more adapted when dealing with complex geometries (see example in Figure 2 for a mesh of a human head to be used in the context of the EEG inverse problem).

Given the mesh, basis functions are typically chosen as piecewise polynomials, where each basis function is nonzero only on a small part of the domain around a given basic element of the mesh, and satisfy some interpolation condition.

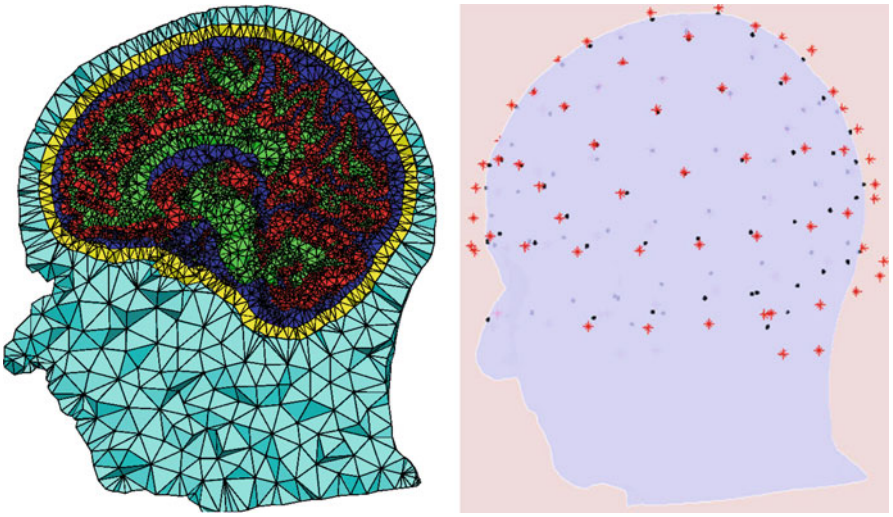


Fig. 2 Left: a sagittal cross section of tetrahedral mesh generated used iso2mesh software [29] for a segmented head composed of five tissues: gray matter (red), white matter (green), cerebrospinal fluid (blue), skull (yellow), and scalp (cyan). Right: profile view of a human head wearing an electrode helmet ($\mathsf{m} = 91$ electrodes). Red stars indicate the real positions of the electrodes on the helmet, while black dots refer to their projection onto the scalp $\{r_\mathsf{j}\}_{\mathsf{j}=1:\mathsf{m}}$.

Typical choices lead to families of basis functions whose spatial support overlaps a little: the support of $\boldsymbol{\phi}_\mathsf{i}$ and $\boldsymbol{\phi}_\ell$ only intersects if they correspond to close mesh elements. As a result $\mathfrak{a}_h(\boldsymbol{\phi}_\mathsf{i}, \boldsymbol{\phi}_\ell)$ is zero for the vast majority of pairs i, ℓ, and the *stiffness* matrix Ω is sparse with $\|\Omega\|_0 = O(\mathsf{n})$.

Using FEM to Solve the Forward EEG Problem Once again, while our ultimate goal is to exploit FEM for *inverse* problems, its use for *forward* problems is illustrative. In bioelectric field problems, a well-known example of problem modeled by (9)–(10) and solved by FEM is the forward EEG problem that aims at computing the electric field within the brain and on the surface of the scalp using a known current source within the brain and the discretized medium composed of the brain and the surrounding layers (skull, scalp, etc.) [38, 43].

4.3 Numerical Approximations of Green's Functions

Discretization methods such as FDM or FEM somehow "directly" discretize the PDE at hand, leading to a matrix Ω which is a discrete equivalent of the operator $\mathfrak{L}$. While (2) implicitly defined $\boldsymbol{x}$ given $\boldsymbol{c}$, in the discretized world, the matrix Ω allows to implicitly define x given c as

$$\Omega x = c, \tag{18}$$

with $x \in \mathbb{R}^\mathsf{n}$ the discretized representation of $\boldsymbol{x}$, and similarly for c and $\boldsymbol{c}$.

We now turn to the discretization of the (potentially more explicit) integral representation of $\boldsymbol{x}$ using Green's functions (5), associated to the integral operator $\mathfrak{G}$, which, as noted in Section 4.3, is often a necessity. One has firstly to discretize the domain Θ, the PDE $\mathfrak{L}$, the field $\boldsymbol{x}$, and the source $\boldsymbol{c}$ and secondly to numerically solve (3) and (4).

Assuming that the equation $\mathfrak{L}\boldsymbol{x} = \boldsymbol{c}$ has a unique solution, we expect the discretized version of $\mathfrak{L}$, the matrix $\Omega \in \mathbb{R}^{\mathsf{n}\times\mathsf{n}}$, to be full rank. Under this assumption, we can write

$$x = \Psi c, \text{ with } \Psi = \Omega^{-1}. \tag{19}$$

In compressive sensing terminology, Ψ is a *dictionary* of discrete Green's functions.

Not surprisingly, the discretized version of the integral operator $\mathfrak{G}$ is the matrix inverse of the discretized version of the differential operator $\mathfrak{L}$. Hence, c and x can be seen as dual representations of the same discrete signal, with linear transformations from one signal space to another. Yet, as we will see, there may be significant differences in sparsity between the matrices Ψ and Ω: while Ψ is typically a dense matrix (Green's functions are often delocalized, *e.g.,* in the context of propagation phenomena), with $\|\Psi\|_0$ of the order of n^2, the analysis operator is typically very sparse, with $\|\Omega\|_0 = O(\mathsf{n})$. In the context of linear inverse problems where one only observes $y \approx Ax$, algorithms may thus have significantly different computational properties whether they are designed with one representation in mind or the other.

4.4 Discretized Inverse Problem

We now have all elements in place to consider the discretized version of the inverse problems expressed in Section 2.3. The signals and measurement vectors are, respectively, denoted by $x \in \mathbb{R}^{\mathsf{s}}$, $c \in \mathbb{R}^{\mathsf{s}}$, and $y \in \mathbb{R}^{\mathsf{m}}$, in the case of a spatial field, or $x \in \mathbb{R}^{\mathsf{st}}$, $c \in \mathbb{R}^{\mathsf{st}}$, and $y \in \mathbb{R}^{\mathsf{mt}}$, in the case of the spatiotemporal field. We denote n the dimension of x and c, which is $\mathsf{n} = \mathsf{s}$ in the former case and $\mathsf{n} = \mathsf{st}$ in the latter.

The vector of measurements y can be seen as a subsampled version of x, possibly contaminated by some additive noise e. In the case of a spatial field, $y = Ax + e$, where A is an $\mathsf{m} \times \mathsf{s}$ spatial subsampling matrix (row-reduced identity) and e is a discrete representation of additive noise $\boldsymbol{e}$. In the case of a spatiotemporal field, the same holds where A is an $(\mathsf{mt}) \times (\mathsf{st})$ block-diagonal concatenation of identical (row-reduced identity) spatial subsampling matrices. Overall, we have to solve

$$y = Ax + e, \tag{20}$$

where A is a given subsampling matrix. Given Ω (and, therefore, $\Psi = \Omega^{-1}$), equivalently to (20), we can write

$$y = A\Psi c + e, \tag{21}$$

where x and c satisfy (18)–(19).

Sparsity or Cosparsity Assumptions Since $A\Psi \in \mathbb{R}^{\mathsf{m}\times\mathsf{s}}$ (resp $\in \mathbb{R}^{\mathsf{mt}\times\mathsf{st}}$), and $\mathsf{m} < \mathsf{s}$, it is obvious that one cannot recover every possible source signal c from the measurements y, hence the need for a low-dimensional model on x or on c. As discussed in Section 2.3, a typical assumption is the sparsity of the source field $\boldsymbol{c}$, which in the discretized setting translates into c being a very sparse vector, with $\|c\|_0 \ll \mathsf{n}$ (or well approximated by a sparse vector), possibly with an additional structure. This gives rise to *sparse synthesis* regularization, usually tackled by convex relaxations or greedy methods that exploit the mode $x = \Psi c$ with sparse c. Alternatively, this is expressed as a *sparse analysis* or *cosparse* model on x asserting that Ωx is sparse.

5 Sparse and Cosparse Regularization

Previously discussed techniques for solving inverse source problems suffer from two serious practical limitations, *i.e.,* algebraic methods (Section 2.3) impose strong, often unrealistic assumptions, whereas sparse synthesis approaches based on numerical Green's function approaches (Section 3) do not scale gracefully for nontrivial geometries. Despite the fact that physical fields are not perfectly sparse in any finite basis, as demonstrated in one of the chapters of the previous issue of this monograph [64], it becomes obvious that we can escape discretization only for very restricted problem setups. Therefore, we focus on the second issue using the *analysis* version of sparse regularization.

5.1 Optimization Problems

Following traditional variational approaches [70], estimating the unknown parameters x and c corresponds to an abstract optimization problem, which we uniformly term *physics-driven (co)sparse regularization*:

$$\min_{x,c} f_d(Ax - y) + f_r(c) \tag{22}$$

$$\text{subject to } \Omega x = c,\ Cx = h.$$

Here, f_d is the data fidelity term (enforcing consistency with the measured data), whereas f_r is the regularizer (promoting (structured) sparse solutions c). The matrix C and vector h represent possible additional constraints, such as source support restriction or specific boundary/initial conditions, as we will see in Section 7.

We restrict both f_d and f_r to be convex, lower semicontinuous functions, *e.g.,* f_d can be the standard sum of squares semimetric and f_r can be the ℓ_1 norm. In some cases the constraints can be omitted. Obviously, we can solve (22) for either c or x and recover another using (18) or (19). The former gives rise to sparse synthesis

$$\min_c f_d(A\Psi c - y) + f_r(c) \quad \text{subject to} \quad C\Psi c = h. \tag{23}$$

or *sparse analysis (*aka *cosparse)* optimization problem

$$\min_x f_d(Ax - y) + f_r(\Omega x) \quad \text{subject to} \quad Cx = h. \tag{24}$$

The discretized PDE encoded in Ω is the analysis operator. As mentioned, the two problems are equivalent in this context [28], but as we will see in Section 6, their computational properties are very different. Additionally, note that the operator Ω is obtained by explicit discretization of (2), while the dictionary Ψ of Green's functions is discretized implicitly (in general, since analytic solutions of (3) are rarely available), *i.e.,* by inverting Ω, which amounts to computing numerical approximations to Green's functions (see Section 4.3).

5.2 Optimization Algorithm

Discretization can produce optimization problems of huge scale (see Sections 6–7 for examples), some of which can be even intractable. Since (23) and (24) are nominally equivalent, the question is whether there is a computational benefit in solving one or another. Answering this question is one of the goals of the present chapter.

Usually, problems of such scale are tackled by first-order optimization algorithms that require only the objective and the (sub)gradient oracle at a given point [62]. The fact that we allow both f_d and f_r to be non-smooth forbids using certain popular approaches, such as *fast iterative soft thresholding algorithm* [5]. Instead we focus on the *alternating direction method of multipliers (ADMM)* algorithm [27, 33], which has become a popular scheme due to its scalability and simplicity. Later in this subsection, we discuss two variants of the ADMM algorithm, convenient for tackling different issues related to the physics-driven framework.

Alternating Direction Method of Multipliers (ADMM) For now, consider a convex optimization problem of the form[2]

$$\min_{z} f(z) + g(Kz - b), \tag{25}$$

where the functions f and g are convex, proper, and lower semicontinuous [62]. Either of these can account for hard constraints, if given as a *characteristic function* $\chi_S(z)$ of a convex set S:

$$\chi_S(z) := \begin{cases} 0 & z \in S, \\ +\infty & \text{otherwise.} \end{cases} \tag{26}$$

An equivalent formulation of the problem (25) is

$$\min_{z_1, z_2} f(z_1) + g(z_2) \quad \text{subject to} \quad Kz_1 - b = z_2, \tag{27}$$

for which the (scaled) *augmented Lagrangian* [11] writes:

$$L_\rho(z_1, z_2, u) = f(z_1) + g(z_2) + \frac{\rho}{2}\|Kz_1 - z_2 - b + u\|_2^2 - \frac{\rho}{2}\|u\|_2^2 \tag{28}$$

with ρ a positive constant. Note that the augmented Lagrangian is equal to the standard (unaugmented) Lagrangian plus the quadratic penalty on the constraint residual $Kz_1 - b - z_2$.

The ADMM algorithm consists in iteratively minimizing the augmented Lagrangian with respect to z_1 and z_2 and maximizing it with respect to u. If the standard Lagrangian has a saddle point [11, 15], iterating the following expressions yields a global solution of the problem[3]:

$$z_1^{(\mathrm{j}+1)} = \arg\min_{z_1} f(z_1) + \frac{\rho}{2}\|Kz_1 - b - z_2^{(\mathrm{j})} + u^{(\mathrm{j})}\|_2^2 \tag{29}$$

$$z_2^{(\mathrm{j}+1)} = \operatorname{prox}_{\frac{1}{\rho}g}\left(Kz_1^{(\mathrm{j}+1)} - b + u^{(\mathrm{j})}\right) \tag{30}$$

$$u^{(\mathrm{j}+1)} = u^{(\mathrm{j})} + Kz_1^{(\mathrm{j}+1)} - b - z_2^{(\mathrm{j}+1)}. \tag{31}$$

The iterates $z_1^{(\mathrm{j})}$ and $z_2^{(\mathrm{j})}$ update the primal variables, and $u^{(\mathrm{j})}$ updates the dual variable of the convex problem (27). The expression $\operatorname{prox}_{f/\rho}(v)$ denotes the well-known *proximal operator* [56] of the function f/ρ applied to some vector v:

[2] The change of notation, in particular from x/c to z for the unknown, is meant to cover both cases in a generic framework.

[3] j denotes an iteration index.

$$\operatorname{prox}_{\frac{1}{\rho}f}(v) = \arg\min_{w} f(w) + \frac{\rho}{2}\|w - v\|_2^2. \tag{32}$$

Proximal operators of many functions of our interest are computationally efficient to evaluate (linear or linearithmic in the number of multiplications, often admitting closed-form expressions). Such functions are usually termed "simple" in the optimization literature [14].

Weighted Simultaneous Direction Method of Multipliers (SDMM) The first ADMM variant we consider is *weighted SDMM (simultaneous direction method of multipliers)* [18]. It refers to an application of ADMM to the case where more than two functions are present in the objective:

$$\min_{z, z_1 \dots z_{\mathrm{f}}} \sum_{\mathrm{i}=1}^{\mathrm{f}} f_{\mathrm{i}}(z_{\mathrm{i}}) \quad \text{subject to} \quad K_{\mathrm{i}} z - b_{\mathrm{i}} = z_{\mathrm{i}}. \tag{33}$$

Such an extension can be written [18] as a special case of the problem (27), for which the iterates are given as follows:

$$z^{(\mathrm{j}+1)} = \arg\min_{z} \sum_{\mathrm{i}=1}^{\mathrm{f}} \frac{\rho_{\mathrm{i}}}{2} \|K_{\mathrm{i}} z - b_{\mathrm{i}} + u_{\mathrm{i}}^{(\mathrm{j})} - z_{\mathrm{i}}^{(\mathrm{j})}\|_2^2, \tag{34}$$

$$z_{\mathrm{i}}^{(\mathrm{j}+1)} = \operatorname{prox}_{\frac{1}{\rho_{\mathrm{i}}} f_{\mathrm{i}}} \left(K_{\mathrm{i}} z^{(\mathrm{j}+1)} - b_{\mathrm{i}} + u_{\mathrm{i}}^{(\mathrm{j})} \right), \tag{35}$$

$$u_{\mathrm{i}}^{(\mathrm{j}+1)} = u_{\mathrm{i}}^{(\mathrm{j})} + K_{\mathrm{i}} z^{(\mathrm{j}+1)} - b_{\mathrm{i}} - z_{\mathrm{i}}^{(\mathrm{j}+1)}. \tag{36}$$

We can now instantiate (33), with I denoting the identity matrix:

- for the sparse synthesis problem: $K_1 = I$, $K_2 = A\Psi$ and $K_3 = C\Psi$;
- for the sparse analysis problem, by $K_1 = \Omega$, $K_2 = A$ and $K_3 = C$.

In both cases $b_1 = 0$, $b_2 = y$, and $b_3 = h$, and the functions f_{i} are $f_{\mathrm{r}}, f_{\mathrm{d}}$, and 0.

Choice of the multipliers. The multipliers ρ_{i} only need to be strictly positive, but a suitable choice can be helpful for the overall convergence speed of the algorithm. In our experiments, we found that assigning larger values for ρ_{i}'s corresponding to hard constraints (*e.g.,* $\|Az - y\|_2 \leq \varepsilon$ or $C\Psi c = h$) and proportionally smaller values for other objectives was beneficial.

The weighted SDMM is convenient for comparison, since it can be easily shown that it yields iteration-wise numerically identical solutions for both the synthesis and analysis problems, if the intermediate evaluations are exact. However, solving a large system of normal equations per iteration of the algorithm seems wasteful in practice. For an improved efficiency, another ADMM variant is more convenient, known as the *preconditioned ADMM* or the *Chambolle-Pock (CP)* algorithm [14].

Chambolle-Pock (CP) For simplicity, we demonstrate the idea on the setting involving only two objectives, as in (27). The potentially expensive step is the

ADMM iteration (29), due to the presence of a matrix K in the square term. Instead, as proposed in [14] and analysed in [74], an additional term is added to the subproblem

$$z_1^{(j+1)} = \arg\min_{z_1} f(z_1) + \frac{\rho}{2}\|Kz_1 - b - z_2^{(j)} + u^{(j)}\|_2^2 + \frac{\rho}{2}\|z_1 - z_1^{(j)}\|_P^2, \tag{37}$$

where $\|v\|_P = v^\mathsf{T} P v$. A clever choice is $P = \frac{1}{\tau\sigma} I - K^\mathsf{T} K$, which, after some manipulation, yields:

$$z_1^{(j+1)} = \text{prox}_{\sigma f}\left(z_1^{(j)} + \sigma K^\mathsf{T} u^{(j)} - \sigma\tau K^\mathsf{T}\left(Kz_1^{(j)} - b - z_2^{(j)}\right)\right). \tag{38}$$

Thus, P acts as a preconditioner and simplifies the subproblem.

In the original formulation of the algorithm [14], the z_2 and u updates are merged together. The expression for $u^{(j+1)}$, along with a straightforward application of Moreau's identity [15], leads to

$$\begin{aligned} z^{(j+1)} &= \text{prox}_{\sigma f}\left(z^{(j)} + \sigma K^\mathsf{T}\left(2u^{(j)} - u^{(j-1)}\right)\right) \\ u^{(j+1)} &= \text{prox}_{\tau g^*}\left(u^{(j)} - \tau\left(Kz^{(j+1)} - b\right)\right), \end{aligned} \tag{39}$$

where the primal variable index has been dropped (z instead of z_1), since the auxiliary variable z_2 does not appear explicitly in the iterations any longer. The function g^* represents the *convex conjugate* [4] of g, and the evaluation of its associated proximal operator $\text{prox}_{g^*}(\cdot)$ is of the same computational complexity as of $\text{prox}_g(\cdot)$, again thanks to Moreau's identity.

As mentioned, different flavors of the CP algorithm [14, 19] can easily lend to settings where a sum of multiple objectives is given, but, for the purpose of demonstration, in the present chapter, we use this simple formulation involving only two objectives (the boundary conditions are presumably absorbed by the regularizers in (23) and (24)). We instantiate (39):

- in the synthesis case, with $K = A\Psi$, $z = c, f = f_r$ and $g = f_d$.
- in the analysis case, we exploit the fact that A has a simple structure and set $K = \Omega, z = x, f(\cdot) = f_d(A\cdot)$, and $g(\cdot) = f_r(\cdot)$. Since A is a row-reduced identity matrix, and thus a tight frame, evaluation of the proximal operators of the type $\text{prox}_{f_d}(A\cdot)$ is usually as efficient as evaluating $\text{prox}_{f_d}(\cdot)$, *i.e.,* without composition with the measurement operator [11]. Moreover, if $\text{prox}_{f_d}(\cdot)$ is separable component-wise (*i.e.,* can be evaluated for each component independently), so is the composed operator $\text{prox}_{f_d}(A\cdot)$.

Accelerated Variants If the objective has additional regularity, such as strong convexity, accelerated variants of ADMM algorithms are available [22, 36]. Thus, since the evaluation of proximal operators is assumed to be computationally

[4] $g^*(\lambda) := \sup_z g(z) - z^\mathsf{T}\lambda$.

"cheap," the main computational burden comes from matrix-vector multiplications (both in CP and SDMM) and from solving the linear least squares subproblem (in SDMM only). For the latter, in the large-scale setting, one needs to resort to iterative algorithms to approximate the solution (ADMM is robust to inexact computations of intermediate steps, as long as the accumulated error is finite [27]). These iterative algorithms can often be initialized (*warm-started*) using a previous iterations' estimate, which may greatly help their convergence. We can also control the accuracy, thus ensuring that there is no large drift between the sparse and cosparse versions.

5.3 Computational Complexity

The overall computational complexity of the considered algorithms results from a combination of their iteration cost and their convergence rate.

Iteration Cost It appears that the iteration cost of ADMM is driven by that of the multiplication of vectors with matrices and their transposes. In practice, most discretization schemes, such as finite-difference (FD) or finite element method (FEM), are *locally supported* [51, 76]. By this we mean that the number of nonzero coefficients required to approximate $\mathfrak{L}$, *i.e.,* $\mathtt{nnz}(\Omega)$, is linear with respect to n, the dimension of the discretized space. In turn, applying Ω and its transpose is in the order of $O(\mathsf{n})$ operations, thanks to the sparsity of the analysis operator. This is in stark contrast with synthesis minimization, whose cost is dominated by much heavier $O(\mathsf{mn})$ multiplications with the dense matrix $A\Psi$ and its transpose. The density of the dictionary Ψ is not surprising; it stems from the fact that the physical quantity modeled by $\boldsymbol{x}$ is spreading in the domain of interest (otherwise, we would not be able to obtain remote measurements). As a result, and as will be confirmed experimentally in the following sections, *the analysis minimization is computationally much more efficient*.

Convergence Rate In [14], the authors took a different route to develop the CP algorithm, where they considered a saddle point formulation of the original problem (25) directly. The asymptotic convergence rate of the algorithm was discussed for various regimes. In the most general setting considered, it can be shown that, for $\tau\sigma\|K\| \leq 1$ and any pair (z, u), the *weak primal-dual gap* is bounded and that, when $\tau\sigma\|K\| < 1$, the iterates $z^{(\mathsf{j})}, u^{(\mathsf{j})}$ converge *("ergodic convergence")* to saddle points of the problem (25). Thus, it can be shown that the algorithm converges with a rate of $O(\frac{1}{\mathsf{j}})$. In terms of the order of iteration count, this convergence rate cannot be improved in the given setting, as shown by Nesterov [62]. However, considering the bounds derived from [14], we can conclude that the rate of convergence is proportional to:

- the operator norm $\|K\|$ (due to the constraint on the product $\tau\sigma$);
- the distance of the initial points $(z^{(0)}, u^{(0)})$ from the optimum (z^*, u^*).

In both cases, a lower value is preferred. Concerning the former, the unfortunate fact is that the ill-posedness of PDE-related problems is reflected in the conditioning of Ω and Ψ. Generally, the rule of thumb is that the finer the discretization, the larger the condition number, since either (or both) $\|\Omega\|$ or $\|\Psi\|$ can grow unbounded [31].

Multiscale Acceleration A potential means for addressing the increasing condition numbers of Ω and Ψ is to apply multiscale schemes, in the spirit of widely used *multigrid methods* for solutions of PDE-generated linear systems. The multigrid methods are originally exploiting smoothing capabilities of Jacobi and Gauss-Seidel iterations [69] and are based on hierarchical discretizations of increasing finesses. Intuitively, the (approximate) solution at a lower level is interpolated, and forwarded as the initial point for solving a next-in-hierarchy higher-resolution problem, until the target (very) high-resolution problem (in practice, more sophisticated schemes are often used). In the same spirit, one could design a hierarchy of discretizations for problems (23) or (24) and exploit the fact that $\|\Omega\|$ and $\|A\Psi\|$ are reducing proportionally to lowering discretization finesse. At the same time, matrix-vector multiplications become much cheaper to evaluate.

Initialization Strategy Finally, for the synthesis optimization problem (23), we often expect the solution vector c^* to be sparse, *i.e.,* to mostly contain zero components. Therefore, a natural initialization point would be $c^{(0)} = 0$, and we expect $\|c^* - c^{(0)}\|$ to be relatively small. However, as mentioned, the synthesis version is not generally preferable, due to high per-iteration cost and memory requirements. On the other hand, we do not have such a simple intuition for initializing $z^{(0)}$ for the cosparse problem (24). Fortunately, we can leverage the multiscale scheme described in the previous paragraph: we would solve the analysis version of the regularized problem at all levels in hierarchy, except at the coarsest one, where the synthesis version with $c^{(0)} = 0$ would be solved instead. The second problem in hierarchy would be initialized by the interpolated version of $z^* = \Psi c^*$, with c^* being the solution at the coarsest level. Ideally, such a scheme would inherit good properties of both the analysis- and synthesis-based physics-driven regularization. In Section 6.2, we empirically investigate this approach, to confirm the predicted performance gains.

6 Scalability

In this section we empirically investigate differences in computational complexity of the synthesis and analysis physics-driven regularization, through simulations based on the weighted SDMM (34), and the multiscale version of the Chambolle-Pock algorithm (39). First, we explore the scalability of the analysis compared to the synthesis physics-driven regularization, applied to the acoustic source localization problem (results and discussions are adopted from [47]). Next, we demonstrate the effectiveness of the mixed synthesis-analysis multiscale approach on a problem governed by Poisson's equation.

6.1 Analysis vs Synthesis

Let us recall that, for acoustic source localization, we use the *finite-difference time-domain (FDTD) standard leapfrog (SLF)* method [51, 76] for discretization of the acoustic wave equation (6) with imposed initial/boundary conditions. This yields a discretized spatiotemporal pressure field $x \in \mathbb{R}^{\mathsf{nt}}$ and a discretized spatiotemporal source component $c \in \mathbb{R}^{\mathsf{nt}}$, built by vectorization and sequential concatenation of t corresponding n-dimensional scalar fields. The matrix operator Ω is a banded, lower triangular, sparse matrix with a very limited number of nonzeros per row (*e.g.,* maximum seven in the 2D case). Note that the Green's function dictionary $\Psi = \Omega^{-1}$ cannot be sparse, since it represents the truncated impulse responses of an infinite impulse response ("reverberation") filter. Finally, the measurements are presumably discrete and can be represented as $y \approx Ax$, where $A \in \mathbb{R}^{\mathsf{mt}\times\mathsf{nt}}$ is a block diagonal matrix, where each block is an identical spatial subsampling matrix.

Optimization Problems To obtain $\hat{x}$ and $\hat{c}$, we first need to solve *one* of the two optimization problems:

$$\hat{x} = \arg\min_x f_r(\Omega_\Theta x) + f_d(Ax - y) \;\text{ subject to }\; \Omega_{\partial\Theta} x = 0 \tag{40}$$

$$\hat{c} = \arg\min_c f_r(c_\Theta) + f_d(A\Psi c - y) \;\text{ subject to }\; c_{\partial\Theta} = 0, \tag{41}$$

where the matrix $\Omega_{\partial\Theta}$ is formed by extracting rows of Ω corresponding to initial conditions (7), and boundary conditions (8), while Ω_Θ is its complement corresponding to (the interior of) the domain itself. Analogously, the vector $c_{\partial\Theta}$ corresponds to components of c such that $\Omega_{\partial\Theta} x = \Omega_{\partial\Theta}\Psi c = c_{\partial\Theta}$ (due to $\Omega_{\partial\Theta} \perp \Psi$), while c_Θ is the vector built from complementary entries of c. The data fidelity term is the ℓ_2 norm constraint on the residual, *i.e.,* $f_d = \chi_{\{u \,|\, \|u\|_2 \le \varepsilon\}}(A \cdot -y)$, where χ is the characteristic function defined in Equation (26).

Source Model and Choice of the Penalty Function for Source Localization Assuming a small number of sources that remain at fixed locations, the true source vector is group sparse: denoting by $\{c_{\mathsf{j}} \in \mathbb{R}^{\mathsf{t}}\}_{\mathsf{j}=1\ldots\mathsf{s}}$ the subvectors of c corresponding to the s discrete spatial locations in Γ, we assume that only few of these vectors are nonzero. As a consequence the regularizer f_r is chosen as the joint $\ell_{2,1}$ group norm [42] with non-overlapping groups associated to this partition of c.

Detection of Source Locations Given the estimated $\hat{x}$, or equivalently $\hat{c}$, the localization task becomes straightforward. Denoting $\{\hat{c}_{\mathsf{j}} \in \mathbb{R}^{\mathsf{t}}\}_{\mathsf{j}=1\ldots\mathsf{s}}$ the subvectors of $\hat{c}$ corresponding to discrete locations in Γ, estimated source positions can be retrieved by setting an energy threshold on each $\hat{c}_{\mathsf{j}}$. Conversely, if the number of sound sources k is known beforehand (for simplicity this is our assumption in the rest of the text), one can consider the k spatial locations with highest magnitude $\|\hat{c}_{\mathsf{j}}\|_2$ to be the sound source positions.

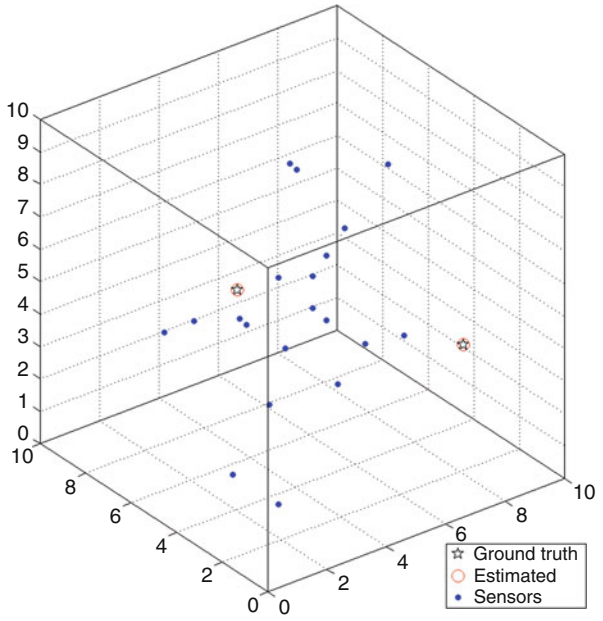


Fig. 3 Localization of two simulated sources (stars) by a 20-microphone random array (dots) in 3D

Results An example localization result of this approach is presented in Figure 3. The simulated environment is a reverberant 3D acoustic chamber, with boundaries modeled by the Neumann (hard wall) condition, corresponding to highly reverberant conditions that are difficult for traditional TDOA methods cf. [8]. The problem dimension is $\mathsf{n} = \mathsf{st} \approx 3 \times 10^6$.

Empirical Computational Complexities To see how the two regularizations scale in the general setting, we explicitly compute the matrix $A\Psi$ and use it in computations. The SDMM algorithm (34) requires solving a system of normal equations, with a coefficient matrix of size $\mathsf{n} \times \mathsf{n}$ with $\mathsf{n} = \mathsf{st}$. Its explicit inversion is infeasible in practice, and we use the *least squares minimum residual (LSMR)* [30] iterative method instead. This method only evaluates matrix-vector products; thus its per-iteration cost is driven by the (non)sparsity of the applied coefficient matrix, whose number of nonzero entries is $O(\mathsf{st})$, in the analysis, and $O(\mathsf{smt}^2)$, in the synthesis case. In order to ensure there is no bias towards any of the two approaches, an oracle stopping criterion is used: SDMM iterations stop when the objective function $f_r(c^{(\mathrm{j})})$ falls below a predefined threshold, close to the ground truth value. Given this criterion, and by setting the accuracy of LSMR sufficiently high, the number of SDMM iterations remains equal for both the analysis and synthesis regularizations.

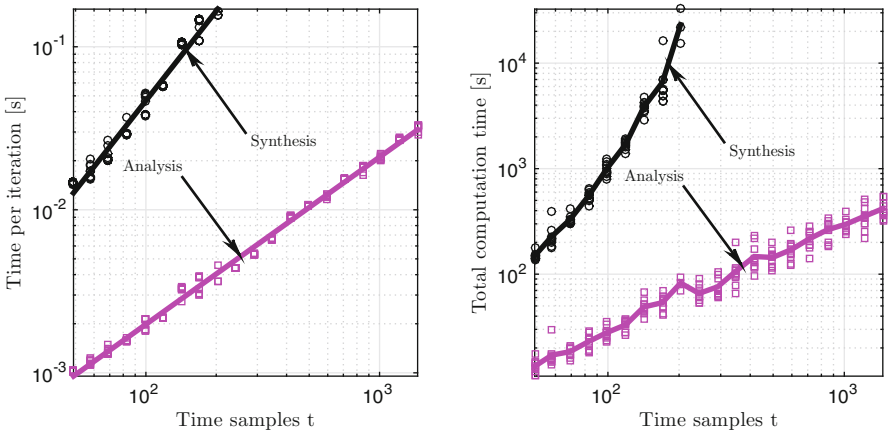


Fig. 4a Computation time *vs* problem size: (left) per inner iteration, (right) total. Solid line, average; dots, individual realizations

In Figure 4a, the results with varying number of time samples t are presented,[5] verifying that the two approaches scale differently with respect to problem size. Indeed, the per-iteration cost of the LSMR solver grows linearly with t, in the analysis case, while being nearly quadratic for the synthesis counterpart. The difference between the two approaches becomes striking when the total computation time is considered, since the synthesis-based problem exhibits cubic growth (in fact, above a certain size, it becomes infeasible to scale the synthesis problem due to high memory requirements and computation time).

Keeping the problem size $\mathsf{n} = \mathsf{st}$ fixed, we now vary the number of microphones m (corresponding to a number of measurements mt). We expect the per-iteration complexity of the analysis regularization to be almost independent of m, while the cost of the synthesis version should grow linearly. The results in the left part of Figure 4b confirm this behavior. However, we noticed that the number of SDMM iterations *decreases* with m for both models, at the same pace. The consequence is that the total computation time increases in the synthesis case, but this *computation time decreases when the number of microphones increases* in the analysis case, as shown in the right graph. While perhaps a surprise, this is in line with recent theoretical studies [16, 73] suggesting that the availability of more data may enable the acceleration of certain machine learning tasks. Here the acceleration is only revealed when adopting the analysis viewpoint rather than the synthesis one.

[5]The spatial dimensions remain fixed to ensure solvability of the inverse problem.

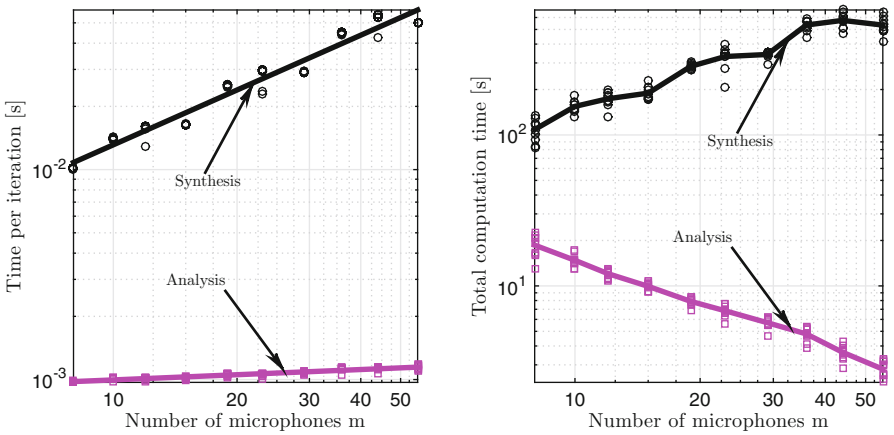


Fig. 4b Computation time *vs* number of measurements: (left) per inner iteration, (right) total. Solid line, average; dots, individual realizations

6.2 Multiscale Acceleration

The back-to-back comparison of the analysis and synthesis regularizations reveals that the former is a preferred option for large-scale problems, when a *numerically identical* SDMM algorithm (34) is used. We are now interested to understand how the two approaches behave when more suitable, but nonidentical versions of the CP algorithm (39) are used instead. To investigate this question, let us consider a very simple one-dimensional differential equation:

$$-\frac{d^2\boldsymbol{x}(r)}{dr^2} = \boldsymbol{c}(r), \tag{42}$$

with $\boldsymbol{x}(0) = \boldsymbol{x}(\phi) = 0$ (*e.g.,* modeling a potential distribution of a grounded thin rod, with sparse "charges" $\boldsymbol{c}(r)$).

Optimization Problems Given the discretized analysis operator Ω and the dictionary Ψ, and assuming, for simplicity, noiseless measurements $y = Ax^*$, physics-driven regularization boils down to solving either of the following two problems:

$$\hat{x} = \arg\min_x \|\Omega x\|_1 \ \text{ subject to } \ Ax = y \tag{43}$$

$$\hat{c} = \arg\min_c \|c\|_1 \ \text{ subject to } \ A\Psi c = y \tag{44}$$

As noted in Section 5.3, the operator norms $\|\Omega\|$ and $\|\Psi\|$ are key quantities for convergence analysis. To obtain Ω and Ψ, we apply finite (central) differences,

here at n points with the dicretization step $\delta r = 1/\mathsf{n}$. We end up with the well-known symmetric tridiagonal Toeplitz[6] matrix Ω, *i.e.,* the 1D discrete Laplacian operator, with a "stencil" defined as $\delta r^2[-1, \ 2, \ -1]$. Its singular values admit simple analytical formula [76]:

$$\sigma_{\mathsf{i}} = 4\mathsf{n}^2 \sin^2\left(\frac{\pi \mathsf{i}}{2\mathsf{n}}\right), \quad \mathsf{i} = 1 \ldots \mathsf{n}. \tag{45}$$

We can immediately deduce $\|\Omega\| \approx 4\mathsf{n}^2$ and $\|\Psi\| \approx 1/\pi^2$, which is very unfavorable for the analysis approach, but appreciated in the synthesis case.[7] The situation is opposite if a discretization with unit step size is applied. Note that we can safely scale each term in the objective (24) by a constant value, without affecting the optimal solution x^*. Provided that f_r can be factored – for the ℓ_1 norm in (43) we have $\|\Omega x\|_1 = |\mathrm{w}| \|\frac{1}{\mathrm{w}} \Omega x\|_1$, $\mathrm{w} \neq 0$ – we can normalize the problem by multiplying with $1/\delta r^2$, which yields $\|\frac{1}{\delta r^2}\Omega\| \approx 4$, irrespective of the problem dimension n.

Numerical Experiments Considering the multiscale scheme described in Section 5.3, we would preferably solve the non-normalized synthesis problem at the coarsest scale, and consequently solve the normalized analysis problems from the second level in hierarchy onward. However, to see the benefits of the multiscale approaches more clearly, here we turn a blind eye on this fact and use the non-normalized finite-difference discretization for both approaches (thereby crippling the analysis approach from the start). To investigate the influence of different aspects discussed in Section 5.3, we set the target problem dimension to $\mathsf{n} = 10^4$ and build a multiscale pyramid with 5 levels of discretization, the coarsest using only 500 points to approximate (42).

Optimization Algorithms Six variants of the CP algorithm (39) are considered:

- ***Analysis***: the matrices Ω and A are built for the target (high-resolution) problem, and the algorithm is initialized by an all-zero vector ($x^{(0)} = 0$).
- ***Analysis multiscale***: A set of analysis operators and measurement matrices is built for each of the five scales in hierarchy. At the coarsest scale, the algorithm is initialized by an all-zero vector; at subsequent scales, we use a (linearly) interpolated estimate $\hat{x}$ from the lower hierarchical level as a starting point.
- ***Synthesis (zero init)***: Analogous to the first, single-scale analysis approach, the target resolution problem is solved by the synthesis version of CP initialized by an all-zero vector $z^{(0)} = 0$.
- ***Synthesis (random init)***: Same as above, but initialized by a vector whose components are sampled from a high-variance univariate normal distribution.

[6]Note that, in this simplistic setting, a fast computation of $\Omega^{-1}c$ using the Thomas algorithm [77] could be exploited. The reader is reminded that this is not a generally available commodity, which is the main incentive for considering the analysis counterpart.

[7]The value of $\|A\Psi\|$ is actually somewhat lower than $\|\Psi\|$; it depends on the number of microphones m and their random placement.

- ***Synthesis multiscale***: Analogous to analysis multiscale approach, a set of reduced dictionary matrices $A\Psi$ is built for each scale in hierarchy. The algorithm at the coarsest scale is initialized by an all-zero vector, and the estimation-interpolation scheme is continued until the target resolution.
- ***Mixed multiscale***: We use the solution of the synthesis multiscale approach *at the coarsest scale* to initialize the second level in hierarchy of the analysis version. Then, the analysis multiscale proceeds as before.

Performance Metrics Even with an oracle stopping criterion, the number of iterations between different versions of the CP algorithm may vary. To have comparable results, we fix the number of iterations to 10^4, meaning that the full-resolution (single-scale) approaches are given an unfair advantage, due to their higher per-iteration cost. Therefore, we output two performance metrics: i) a relative error, $\epsilon = \|\hat{x} - x^*\| / \|x^*\|$, $\hat{x}$ and x^* being respectively the estimated and the ground truth (propagated) signal[8] and ii) processing time for the given number of iterations. The experiments are conducted for different values of m, the number of measurements.

Results The results presented in Figure 5 (left) confirm our predictions: the synthesis approach initialized with all-zeros and the proposed mixed synthesis-analysis approach perform better than the rest in terms of the relative error metric. It is clear that improper initialization significantly degrades performance – notably, for the synthesis algorithm initialized randomly and the two analysis approaches. The single-scale analysis version is the slowest to converge, due to its large Lipschitz constant $\|\Omega\|^2$ at $\delta r = 1/\mathsf{n}$ and trivial initialization. However, processing time results on the right graph of Figure 5 reveal that synthesis-based approaches imply much higher computational cost than the analysis ones, even if the multiscale scheme is applied. In addition their computational performance suffers when the number of measurements increases – which is, naturally, beneficial with regard to the relative error – due to the increased cost of matrix-vector products with $G = A\Psi$ (where G is precomputed once and for all before iterating the algorithm). Fortunately, the mixed approach is mildly affected, since only the computational cost at the coarsest scale increases with m.

7 Versatility

In this section we demonstrate the versatility of physics-driven cosparse regularization. First, we discuss two notions of "blind" acoustic source localization enabled by the physics-driven approach. All developments and experiments in this part refer to 2D spatial domains; however, the principles are straightforwardly extendable

[8]This metric is more reliable than the corresponding one with respect to the source signal, since small defects in support estimation should not disproportionately affect performance.

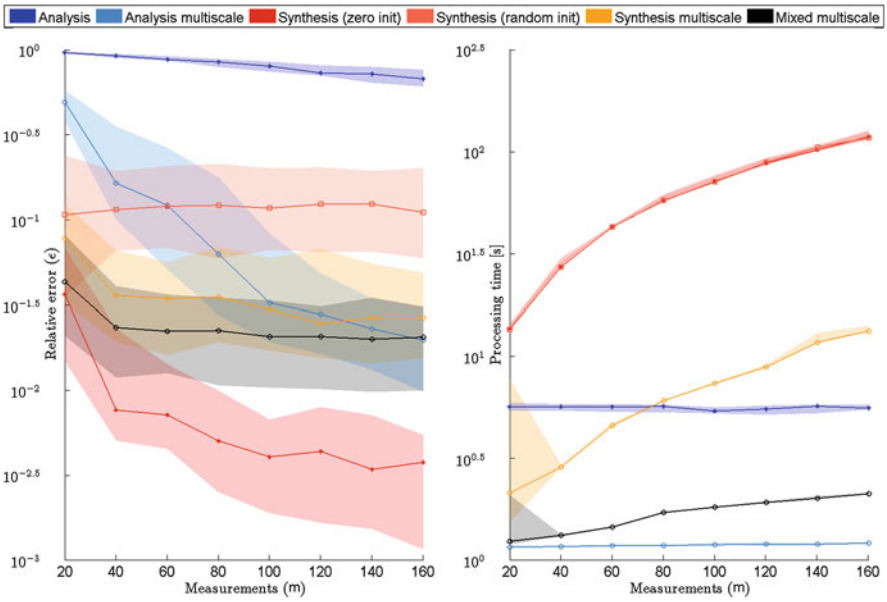


Fig. 5 Median performance metrics (logarithmic scale) wrt number of measurements (shaded regions correspond to 25%–75% percentiles). Left, relative error ϵ; right, processing time

to three spatial dimensions. In the second subsection, we apply the regularization to another problem, using a different discretization method: source localization in electroencephalography with FEM. There we consider a three-dimensional problem, with physically relevant domain geometry (real human head).

7.1 Blind Acoustic Source Localization

The attentive reader may have noticed that so far no explicit assumption has been made on the shape of the spatial domain under investigation. In fact, it has been shown in [46] that the proposed regularization facilitates acoustic source localization in spatial domains of exotic shape, even if there is no line of sight between the sources and microphones. This is an intriguing capability, as such a scenario prevents the use of more traditional methods based on TDOA. One example is presented in Figure 6 (left), termed "hearing behind walls." Here the line of sight between the sources and microphones is interrupted by a soundproof obstacle; hence the acoustic waves can propagate from the sources to the microphones only by reverberation. Yet, as shown with the empirical probability (of exactly localizing all sources) results in Figure 6 (right), the physics-driven localization is still possible.

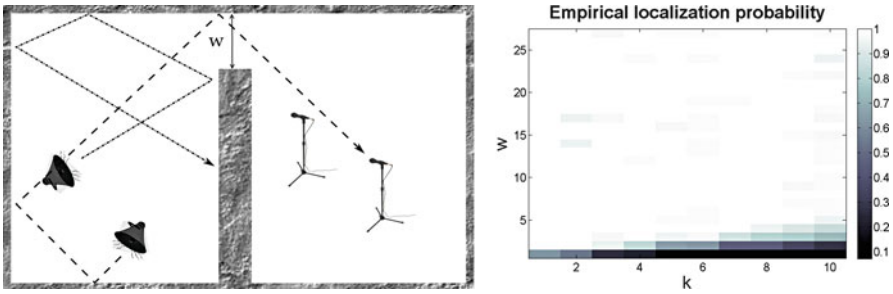


Fig. 6 Left, hearing behind walls scenario; right, empirical probability of accurate localization given the gap width w and number of sources k (results from [46])

However, an issue with applying physics-driven regularization is that it comes with the strong assumption of knowing the parametrized physical model almost perfectly. In reality, such knowledge is not readily available; hence there is an interest in inferring certain unknown physical parameters directly from the data. Ideally, such estimation would be done simultaneously with solving the original inverse problem, which is the second notion of "blind" localization in this subsection. For the acoustic wave equation (6) and boundary conditions (8), various parameters could be unknown. In this section, we consider two of them: sound speed $\boldsymbol{v}$ and the specific acoustic impedance ξ. Note that imposing a parameter model is necessary in this case; otherwise, these blind estimation problems would be ill-posed.

Blind Localization and Estimation of Sound Speed (BLESS) The speed of sound $\boldsymbol{v}$ is usually a slowly varying function of position and time, *e.g.,* due to a temperature gradient of space caused by an air conditioner or radiator. Provided that the temperature is in steady state and available, one could approximate $\boldsymbol{v}$ as constant. However, if such approximation is very inaccurate, the physical model embedded in the analysis operator will be wrong. The effects of such model inaccuracies have been exhaustively investigated [13, 39] and are known to significantly alter regularization performance. Therefore, our goal here is to simultaneously recover the pressure signal (in order to localize sound sources) and estimate the sound speed function $\boldsymbol{v}$. For demonstrational purpose, we regard $\boldsymbol{v} := \boldsymbol{v}(r)$, *i.e.,* a function that varies only in space.

To formalize the problem, consider the FDM leapfrog discretization scheme presented in (11). Instead of a scalar sound speed parameter v, we now have a vector unknown corresponding to the sampled function $\mathrm{v_{ij}} = \boldsymbol{v}(r_x(\mathrm{i}), r_y(\mathrm{j})) > 0$. Denoting $q \in \mathbb{R}^{\mathsf{n}}$ the vector with stacked entries $\mathrm{q_{i,j}} = \mathrm{v_{i,j}^{-2}}$, we can represent the analysis operator Ω as follows:

$$\Omega = \Omega_1 + \mathtt{diag}\,(q)\,\Omega_2, \tag{46}$$

where the singular matrices Ω_1 and Ω_2 are obtained by factorizing wrt v in (11).

Assume the entries of q are in some admissible range $[\mathrm{v}_{\max}^{-2}, \mathrm{v}_{\min}^{-2}]$, e.g., given by the considered temperature range in a given environment. Moreover, assume that v and q are slowly varying functions. We model this smoothness by a vector space of polynomials of degree $\mathsf{r}-1$ in the space variables (constant over the time dimension), which leads to the model $q = Fa$, where F is a dictionary of sampled polynomials and a is a weight vector [7].

Adding a as an unknown in (40) (instantiated, *e.g.*, with f_d a simple quadratic penalty), and introducing the auxiliary sparse variable c, yields the optimization problem:

$$\begin{aligned} &\min_{x,c,a} \; f_r(c_\Theta) + \lambda \|Ax - y\|_2^2 \\ &\text{subject to} \quad \Omega = \Omega_1 + \mathtt{diag}\,(Fa)\,\Omega_2, \quad \mathrm{v}_{\max}^{-2} \preceq Fa \preceq \mathrm{v}_{\min}^{-2} \\ &\qquad\qquad \Omega x = c, \quad c_{\partial\Theta} = 0. \end{aligned} \tag{47}$$

Unfortunately, due to the presence of the bilinear term $\mathtt{diag}\,(Fa)\,\Omega_2 x$ relating optimization variables x and a, (47) is not a convex problem. However, it is biconvex – fixing either of these two makes the modified problem convex again; thus its global solution is attainable. This enables us to design an ADMM-based heuristic, by developing an augmented Lagrangian (28) comprising the three variables:

$$\begin{aligned} L_{\rho_1,\rho_2}(c,x,a,u_1,u_2) = {} & f_r(c_\Theta) + \chi_{\cdot=0}\,(c_{\partial\Theta}) \\ & + \frac{\rho_1}{2} \| \left(\Omega_1 + \mathtt{diag}\,(Fa)\,\Omega_2\right) x - c + u_1 \|_2^2 + \frac{\rho_2 \lambda}{2} \|Ax - y + u_2\|_2^2 \\ & + \chi_{\mathrm{v}_{\max}^{-2} \preceq \cdot \preceq \mathrm{v}_{\min}^{-2}}\,(Fa) - \frac{\rho_1}{2}\|u_1\|_2^2 - \frac{\rho_2}{2}\|u_2\|_2^2. \end{aligned} \tag{48}$$

From here, the ADMM iterates are straightforwardly derived, similar to (29)–(31). We skip their explicit formulation to reduce the notational load of the chapter.

In order to demonstrate the joint estimation performance of the proposed approach, we vary the number of sources k and microphones m. First, a vector $\tilde{a}$ is randomly generated from centered Gaussian distribution of unit variance. Then, a is computed as the Euclidean projection of $\tilde{a}$ to a set $\left\{a \mid \mathrm{u}_{\max}^{-2} \preceq F_{\mathtt{null}}^{[\mathsf{r}]} a \preceq \mathrm{u}_{\min}^{-2}\right\}$. We let $\mathrm{u}_{\min} = 300$m/s and $\mathrm{u}_{\max} = 370$m/s and use Neumann boundary conditions. The performance is depicted as an empirical localization probability graph in Figure 7a, for two values of the degree r of the polynomials used to model the smoothness of q. One can notice that the performance deteriorates with q less smooth (*i.e.,* with larger r), since the dimension of the model polynomial space increases. When localization is successful, q is often perfectly recovered, as exemplified in Figure 7b.

[8] Vertical axis, k/m, the ratio between the number of sources and sensors; horizontal axis, m/s, the proportion of the discretized space occupied by sensors

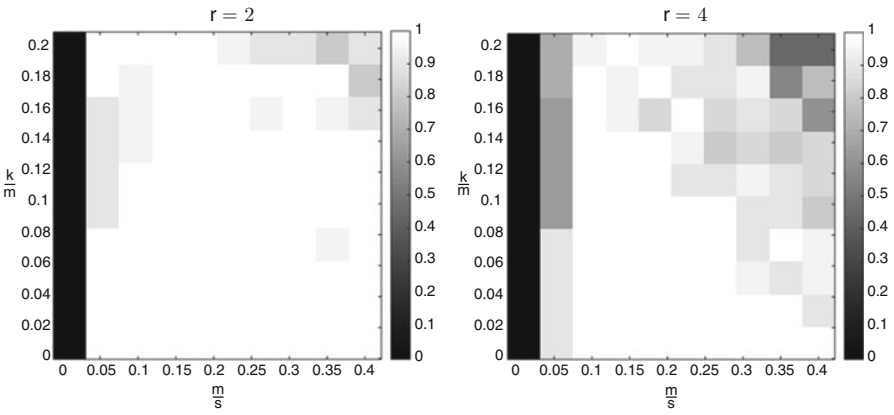


Fig. 7a Empirical localization probability with estimated sound speed. Vertical axis, k/m, the ratio between the number of sources and sensors; horizontal axis, m/s, the proportion of the discretized space occupied by sensors

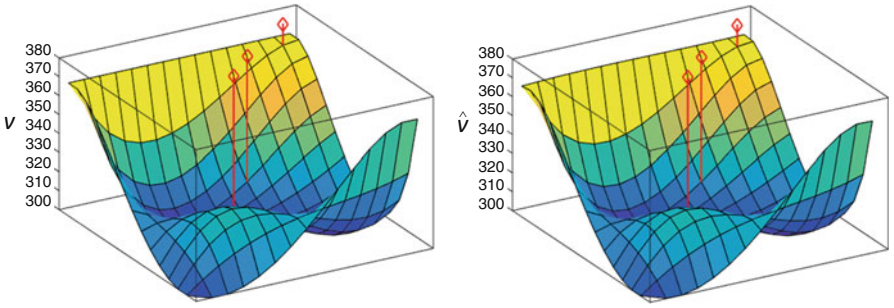


Fig. 7b The original sound speed (left) v and the estimate (right) $\hat{v}$ (the diamond markers indicate the spatial position of the sources)

Cosparse Acoustic Localization, Acoustic Impedance Estimation and Signal Recovery (CALAIS) A perhaps even more critical acoustic parameter is the specific boundary impedance ξ in (8). While we may have an approximate guess of the sound speed, the impedance varies more abruptly, as it depends on the type of material composing a boundary of the considered enclosed space [49]. The approach recently proposed in [3] relies on the training phase using a known sound source, allowing one to calibrate the acoustic model for later use with unknown sources in the same environment. Here we present a method [6] to avoid the calibration phase, and, as for the sound speed, to simultaneously infer the unknown parameter ξ and the acoustic pressure x.

Now we consider discretization of the spatial domains' boundary. Let $\Omega_{\partial\Gamma}$ represent the subset of rows of the analysis operator Ω corresponding to the boundary conditions only, and let Ω_0 denote the matrix corresponding to initial conditions

only (we have $\Omega_{\partial\Theta} = [\Omega^{\mathsf{T}}{}_0\ \Omega^{\mathsf{T}}{}_{\partial\Gamma}]^{\mathsf{T}}$, up to a row permutation). To account for Mur's boundary conditions, FDM discretization can be explicitly written as:

$$x_{\mathsf{i},\mathsf{j}}^{\tau+1}(1+\tfrac{\lambda}{\xi_{\mathsf{i},\mathsf{j}}})- \\ \left[2(1-2\lambda^2)x_{\mathsf{i},\mathsf{j}}^{\tau}+\lambda^2(x_{\mathsf{i},\mathsf{j}+1}^{\tau}+x_{\mathsf{i},\mathsf{j}-1}^{\tau})+2\lambda^2 x_{\mathsf{i}-1,\mathsf{j}}^{\tau}-\left(1-\tfrac{\lambda}{\xi_{\mathsf{i},\mathsf{j}}}\right)x_{\mathsf{i},\mathsf{j}}^{\tau-1}\right]=0, \tag{49}$$

where $\lambda = \mathrm{v}\mathrm{d}_\tau/\mathrm{d}_\mathsf{x} = \mathrm{v}\mathrm{d}_\tau/\mathrm{d}_\mathsf{y}$. Denote $\eta = \left[\xi_{1,1}\ \xi_{2,1}\ \ldots\ \xi_{\mathsf{i},\mathsf{j}}\ \ldots\right]^{-\mathsf{T}}$ the vector of inverse acoustic impedances, *i.e.,* of *specific acoustic admittances*, which we assume does not change in time. We introduce the matrix S which distributes the admittances stored in η at appropriate positions in discretized space and repeats these across all time instances $[1,\mathsf{t}]$. Factorizing (49) with respect to **"**, we can represent $\Omega_{\partial\Gamma}$ (up to an adequate row permutation) as:

$$\Omega_{\partial\Gamma} = \underbrace{\begin{bmatrix} \underline{\Omega}_{\partial\Gamma_1} & & \\ & \underline{\Omega}_{\partial\Gamma_1} & \\ & & \ddots \\ & & \underline{\Omega}_{\partial\Gamma_1} \end{bmatrix}}_{\Omega_{\partial\Gamma_1}} + \mathtt{diag}\,(S\eta) \underbrace{\begin{bmatrix} \underline{\Omega}_{\partial\Gamma_2} & & \\ & \underline{\Omega}_{\partial\Gamma_2} & \\ & & \ddots \\ & & \underline{\Omega}_{\partial\Gamma_2} \end{bmatrix}}_{\Omega_{\partial\Gamma_2}}. \tag{50}$$

where the rows of each block $\underline{\Omega}_{\partial\Gamma_1}$ (resp.) $\underline{\Omega}_{\partial\Gamma_2}$ are indexed by the space coordinate, while the blocks themselves are indexed by time.

Note that, for standard rooms, the boundaries are composed of walls, floor, ceiling, windows, etc. At least on macroscopic scale, each of these structures is approximately homogeneous. Hence, we suppose that η admits a *piecewise constant* model, provided we take care of the ordering of elements within η. This weak assumption usually holds in practice, unless the discretization is very crude. To promote such a signal model, the discrete total variation norm $\|\eta\|_{\mathrm{TV}} = \|\nabla\eta\|_1$ is commonly used.

This model, along with the assumption that the initial/boundary conditions are homogeneous, inspires the following optimization problem:

$$\min_{x,\eta} f_r(\Omega_\Theta x) + \|\eta\|_{\mathrm{TV}} + \lambda\|\Omega_{\partial\Gamma}x\|_2^2, \\ \text{subject to } \Omega_0 x = 0,\ Ax = y,\ \eta \succeq 0 \\ \Omega_{\partial\Gamma} = \Omega_{\partial\Gamma_1} + \mathtt{diag}\,(S\eta)\,\Omega_{\partial\Gamma_2}, \tag{51}$$

where λ is a user-defined positive constant. Therefore, we end up with another biconvex problem and propose to address it again by an ADMM heuristics. As in the previous case, one proceeds by defining the augmented Lagrangian which determines the algorithm. We do not develop these steps here, due to spatial constraints.

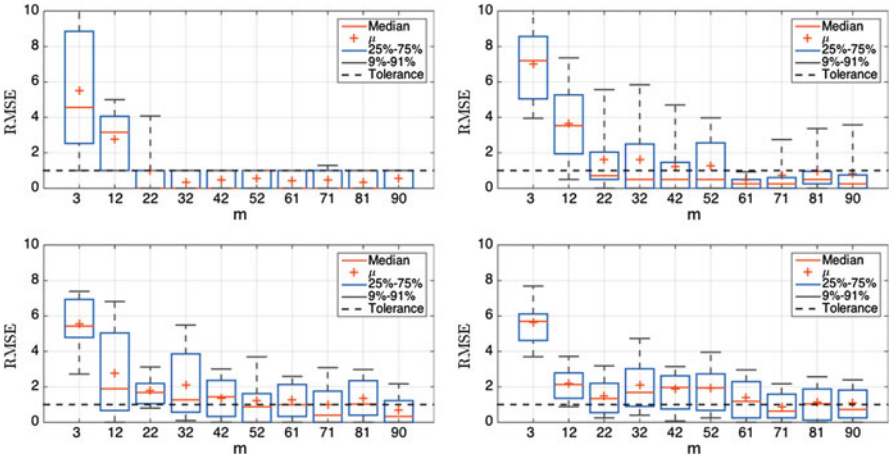


Fig. 8 RMSE for $\mathsf{k} = 1$, $\mathsf{k} = 2$ (top); and $\mathsf{k} = 3$, $\mathsf{k} = 4$ (bottom) simulated sources

To illustrate the performance and robustness of the approach, we propose a somewhat challenging experimental setup. Within a simulated 2D spatial domain, modeled by high-resolution LFM discretization, we randomly distribute k sources and m microphones. The ground truth admittance is generated such that it approximately satisfies the assumed model – a white Gaussian noise is added to a piecewise constant vector η to account for model inaccuracy. The measurements y are also corrupted by noise, such that *per-sample* SNR is 20dB. Finally, the matrices Ω_{Θ}, $\Omega_{\partial\Gamma_1}$, $\Omega_{\partial\Gamma_2}$, and S, used in the optimization problem (51), are obtained by discretizing the physical model by a low-resolution LFM. This embeds some inaccuracy at the PDE modeling level, making simulations more physically relevant.

The results in Figure 8 are with respect to average Euclidean distance between ground truth source locations (empirical *root-mean-square error (RMSE)*) and the estimated ones. The dashed line denotes the spatial step size of the coarse grid; thus the errors below this threshold are tolerated. The median RMSE values indicate that localization is possible provided that the number of microphones is sufficiently high. The error increases with the number of sources to localize, but usually remains lower than the double of spatial step size, suggesting that the sources are localized in their true, immediate neighborhoods.

7.2 Cosparse Brain Source Localization

Most source localization algorithms use one of the two following source models: the point source model, which explains the data with a small number of equivalent current dipoles, and the distributed source model, which uses thousands of dipoles. Whereas the latter allows for an estimation of the spatial extent of the source,

it requires to make assumptions about the spatial source distribution, which may lead to blurred (or even distorted) solutions [37]. On the other hand, the former often gives helpful first approximations and superior performance in environments where there are few sources which are clustered [37]. Regarding head models, they aim at representing geometrical and electrical properties of the different tissues composing the volume conductor. Various models were proposed going from concentric homogeneous spheres with isotropic conductivities for which analytic computations of Green's functions are possible, to realistically shaped models with refined tissue conductivity values [81].

FEM Discretization and Head Model As seen in Section 4.2, FEM can be used to discretize Poisson's equation with Neumann boundary condition (9)–(10) and derive an equation of the form $\Omega\, x = c$ where the so-called linear analysis operator [61] Ω is the *stiffness matrix* and vectors x and c respectively contain the potential and total current flow values at the different nodes of the mesh. A realistic head model obtained from anatomical imaging modalities, such as computed tomography (CT) and structural magnetic resonance imaging (sMRI), is segmented, and a linear tetrahedral mesh of n nodes is then generated to divide it into small elements where a unique conductivity value is assigned to each one. In Figure 9, we illustrate the different steps performed to come to the sought algebraic system.

As previously explained in Section 4.2, the $(\mathsf{n} \times \mathsf{n})$ matrix Ω is symmetric, positive semidefinite, rank deficient by one and sparse with only few components in each row [69]. Generally, instead of considering the singular linear system $\Omega x = c$, another possibility is to transform it into a regular one and solve this instead. The regular system is chosen such that its unique solution belongs to the set of solutions of the original singular system. As described in [10], the easiest approach is to fix the value of the potential to zero in one node. The special structure of the matrix Ω then allows us to cancel the corresponding row and column in Ω and also the respective entry in the right-hand side vector c. This leads to a system for which the $(\mathsf{n}-1 \times \mathsf{n}-1)$ resulting matrix is symmetric, sparse, and positive definite, as it can be

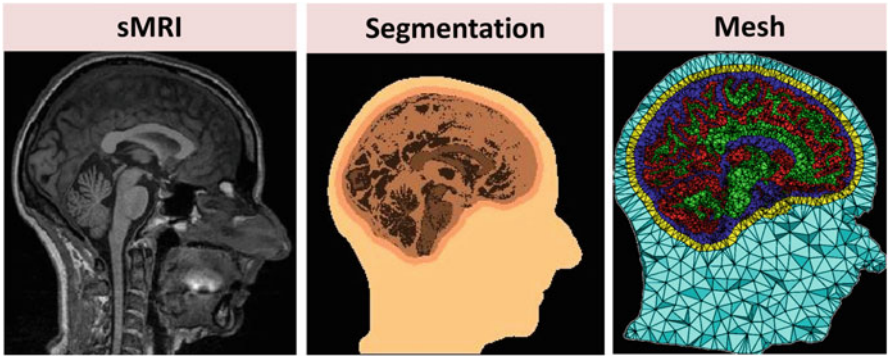


Fig. 9 The different steps of preprocessing in discretizing realistic head volume using FEM

derived from a bilinear form satisfying the same properties as $\mathfrak{a}(.,.)$ in Section 4.2. By abuse of notation, we still denote it Ω. The solution of this system solves the initial system with a zero potential value in the reference node.

Source Model We consider the following assumptions:

A1. There are g possible current sources that cover the volume of the gray matter $\mathscr{G}$.

A2. Each current source is modeled as a "pure" dipole, consisting in a point dipole characterized by its position ρ_{q} and moment $p_{\mathsf{q}} = \|p_{\mathsf{q}}\|_2 \; n_{\mathsf{q}}$ where $\|p_{\mathsf{q}}\|_2$ and n_{q} correspond to the activity level and the orientation of the dipole, respectively. In this model, the current density is expressed as $j = \sum_{\mathsf{q}=1}^{\mathsf{g}} p_{\mathsf{q}} \; \delta(\rho_{\mathsf{q}})$ [71].

A3. Each current dipole is oriented orthogonally to the cortical surface (p_{q} is the normal to the cortical surface at position ρ_{q}).

A4. At most k ($\mathsf{k} < \mathsf{m}$) current dipoles have non-negligible activity.

In addition to the "pure" dipole, there are other numerical methods used to model the current dipole, such as the subtraction potential method and the direct potential approach using Saint Venant's principle [71, 83]. Though the "pure" dipole is a mathematical idealization of the real "physical" dipole that has finite current and separation between its monopoles, it is frequently used for its simple derivation.

According to the variational form introduced in (15), the component c_{i} in the *load vector* is

$$c_{\mathsf{i}} = \int_{\Theta_h} \nabla \cdot j \; \boldsymbol{\phi}_{\mathsf{i}}(\boldsymbol{r}) \; d\boldsymbol{r}. \tag{52}$$

By applying Green's identity to (52), using the fact that no current sources are present in the scalp $\partial\Theta$, considering the expression of the domain Θ_h as a union of d tetrahedra Θ_h^{e} present in the mesh, and eventually assuming that each dipole position ρ_{q} coincides with a node position in $\mathscr{G}$, appropriate calculations lead to rewrite the entry in the *load vector* as:

$$c_{\mathsf{i}} = \begin{cases} -p_{\mathsf{q}} \cdot \sum_{\mathsf{e} \in \mathsf{d}_{\mathsf{i}}} \nabla^{\mathsf{e}} \boldsymbol{\phi}_{\mathsf{i}}(\rho_{\mathsf{q}}) & \text{if} \quad \boldsymbol{r}_{\mathsf{i}} = \rho_{\mathsf{q}} \\ 0 \quad \text{otherwise,} \end{cases} \tag{53}$$

where $\nabla^{\mathsf{e}} \boldsymbol{\phi}_{\mathsf{i}}$ is the gradient of function $\boldsymbol{\phi}_{\mathsf{i}}$ over element Θ_h^{e}. By injecting the expression of dipole moment in (53) and in the case of nonzero c_{i}, the latter can be expressed as the product $c_{\mathsf{i}} = \|p_{\mathsf{q}}\|_2 \;\; n_{\mathsf{q}} \cdot \sum_{\mathsf{e} \in \mathsf{d}_{\mathsf{i}}} \nabla^{\mathsf{e}} \boldsymbol{\phi}_{\mathsf{i}}(\rho_{\mathsf{q}})$, which allows us to factorize vector c as $c = B\,z$ where B is a $(\mathsf{n} - 1 \times \mathsf{n} - 1)$ diagonal matrix defined by

$$B_{\mathsf{i},\mathsf{i}} = \begin{cases} -n_{\mathsf{q}} \cdot \sum_{\mathsf{e} \in \mathsf{d}_{\mathsf{i}}} \nabla^{\mathsf{e}} \boldsymbol{\phi}_{\mathsf{i}}(\rho_{\mathsf{q}}) & \text{if} \quad \boldsymbol{r}_{\mathsf{i}} = \rho_{\mathsf{q}} \\ 1 \quad \text{otherwise.} \end{cases} \tag{54}$$

In addition, z is an $(\mathsf{n}-1)$-dimensional sparse vector with $\mathsf{n}-1-\mathsf{g}$ known zero elements defined as

$$z_{\mathsf{i}} = \begin{cases} \|p_{\mathsf{q}}\|_2 & \text{if} \quad \boldsymbol{r}_{\mathsf{i}} = \rho_{\mathsf{q}} \\ 0 & \text{otherwise.} \end{cases} \tag{55}$$

Consequently, matrix B conveys our knowledge about the orientation of the g dipoles of the gray matter $\mathscr{G}$, while the nonzero elements in vector z represent the activity of dipoles restricted to the cortical volume. It is noteworthy that, even when dipoles positions do not coincide with node positions, the factorization of vector c is still possible. However, in that case, the matrix B is no longer a square matrix but rather a tall matrix (not left invertible), which makes the computation of Ω more complicated. In addition, the assumption on the dipole position is still realistic and affordable by using a dense mesh in $\mathscr{G}$.

Overall Model Combining the properties $\Omega x = c$, the source model $c = Bz$, and the observation model $y = Ax$, the brain source localization defined above can finally be reformulated as a cosparse analysis model fitting problem given by:

$$\begin{cases} \tilde{\Omega}\, x = z \\ y = Ax, \end{cases} \tag{56}$$

where the analysis operator $\tilde{\Omega}$ is given by $\tilde{\Omega} = B^{-1}\,\Omega$ and the sensing matrix A is an $\mathsf{m} \times \mathsf{n}-1$ row-reduced identity matrix. As B is diagonal, $\tilde{\Omega}$ is still sparse.

Optimization Problem To address the cosparse analysis model fitting problem (56), we express the following convex optimization problem:

$$\min_{x} \|\tilde{\Omega}_1\, x\|_1 + \lambda \|\tilde{\Omega}_2\, x\|_2^2 \tag{57}$$

$$\text{subject to } A\, x = y.$$

where $\tilde{\Omega}_1$ is the $(\mathsf{g} \times \mathsf{n}-1)$ submatrix of $\tilde{\Omega}$ obtained by extracting the rows of $\tilde{\Omega}$ corresponding to the support set of the gray matter $\mathscr{G}$, whereas $\tilde{\Omega}_2$ corresponds to the rows indicated by the complementary set $\bar{\mathscr{G}}$. By choosing the appropriate weight λ, the cosparse solution of the optimization problem (57) will fulfill the assumptions (**A1**) to (**A4**). Namely, $\|\tilde{\Omega}_1\, x\|_1$ will promote sparsity at the surface of the cortex, while $\lambda \|\tilde{\Omega}_2\, x\|_2^2$ will attenuate the signal in the other regions. The linear constraints $A\, x = y$ ensure that the model fits the electrode measurements. Depending on the resolution of the cubic grid tuned by n, the problem can reach considerably large scale. Therefore, we use the Chambolle-Pock method as described in Section 5.2.

Experiments and Performance Criterion One scenario was considered for a comparison of performance between the analysis and synthesis approaches. It aims at studying the influence of the SNR.

More particularly, $\mathsf{k} = 2$ synchronous epileptic dipoles were placed in $\mathscr{G}$ at $\rho_1 = [-71, 31, 92]^\mathsf{T}$ and $\rho_2 = [-70, 27, 92]^\mathsf{T}$, respectively, (locations are given in centimeters). Note that the origin (O) of the head model was defined as the intersection of the O-Cz axis (z-axis), the O-T4 axis (x-axis), and the O-Fpz axis (y-axis). A physiologically relevant model [41] was used to generate the time series corresponding to epileptic activity. It is noteworthy that this activity was the same for both epileptic dipoles, leading to synchronous epileptic sources. On the other hand, the background activity, i.e., the activity of non-epileptic dipoles of $\mathscr{G}$, was generated as Gaussian and as temporally and spatially white. Its power was controlled by a multiplicative coefficient in order to get different SNR values.

As far as the head model is concerned, we used a realistic head model obtained from sMRI. Ninety-one electrodes ($\mathsf{m} = 91$) were placed on the scalp using the 10-5 system [63]. In addition, in order to apply the FEM and compute the analysis operator Ω, we created a linear tetrahedral mesh of $\mathsf{n} = 180585$ nodes. Consequently, the size of $\tilde{\Omega}$ and the number of dipoles of $\mathscr{G}$ were $(\mathsf{n}-1)\times(\mathsf{n}-1) = 180584 \times 180584$ and $\mathsf{g} = 3110$, respectively.

The quality of the source localization was quantified for each method by means of the average root-mean-square error (RMSE), which is defined by:

$$RMSE = \frac{1}{\mathsf{k}\,\mathsf{mc}} \sum_{\mathsf{q}=1}^{\mathsf{k}} \sum_{\mathsf{i}=1}^{\mathsf{mc}} \min_{1 \le \mathsf{j} \le \mathsf{k}} \|\rho_{\mathsf{q}} - \widehat{\rho}_{\mathsf{j}}\| \tag{58}$$

where mc is the number of realizations fixed to 71, where ρ_{q} is the ground truth position of the q-th epileptic dipole, and where $\widehat{\rho}_{\mathsf{j}}$ is the j-th dipole location estimated during the i-th Monte Carlo trial. It is noteworthy that from one realization to another, the temporal dynamics of the g dipoles of $\mathscr{G}$ were changed while the location of the three epileptic dipoles stayed unchanged.

Figure 10 shows the RMSE criterion at the output of both algorithms as a function of the SNR. It appears that the analysis method is more robust with respect to the presence of noise than the synthesis one. Indeed, it succeeds in localizing perfectly both epileptic dipoles beyond 12 dB, while the synthesis-based method does not manage to do it perfectly.

Note that in such a practical context for which the brain sources are synchronous, the analysis method was also shown to overcome the RapMUSIC (recursively applied MUSIC) [57] and FO-D-MUSIC (fourth-order deflationary MUSIC) [1] algorithms [2]. In fact, RapMUSIC and FO-D-MUSIC are sequential versions of the subspace approach MUSIC (multiple signal classification) [72] based on second-order (SO) and fourth-order (FO) statistics, respectively.

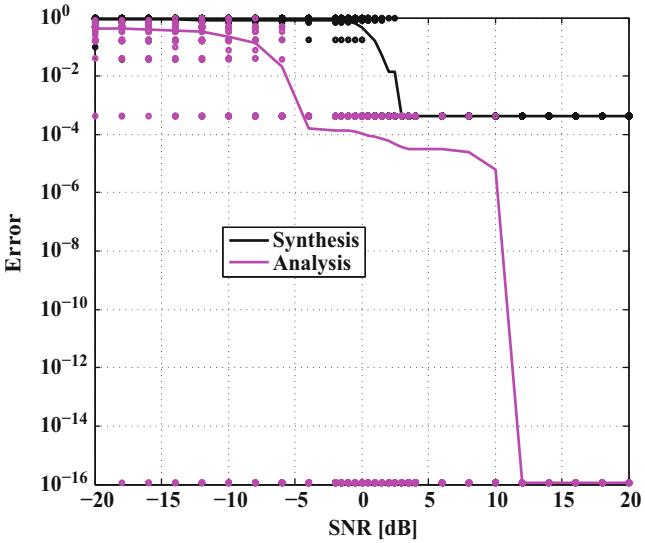


Fig. 10 Behavior of the analysis and synthesis approaches as a function of the SNR for $\mathsf{k} = 2$ epileptic dipoles and $\mathsf{m} = 91$ electrodes (solid line, average; dots, individual realizations)

8 Summary and Conclusion

In many physics-driven inverse problems, one can leverage both a sparsity hypothesis on some source term and the properties of the underlying PDE. A classical approach to combine these ingredients is to build a dictionary of Green's function of the PDE and to exploit traditional techniques (ℓ_1 regularization) for sparse synthesis reconstruction to estimate the source field. Yet, for high spatial and/or temporal resolutions, precomputing Green's functions can be challenging, and the synthesis version of the ℓ_1 optimization problem may become numerically intractable due to polynomial complexities of too high degree in the overall size of the discretization.

An alternative is to discretize the PDE itself, *e.g.,* through a finite-difference scheme or the finite element methods, which naturally leads to very sparse *analysis operators* rather than dictionaries. While the two approaches (synthesis and analysis) are formally equivalent, a primary advantage of the *cosparse analysis* regularization is a much smaller iteration cost. Although demonstration of the full potential of the existing cosparse approaches on *real* acoustic or EEG data remains to be done, results shown on simulated data allow to support our claims on their interest. Overall, as illustrated in this chapter, a promising approach to achieve precision and scalability is to combine the synthesis approach and the analysis one in a multiscale optimization strategy. Besides scalability, the cosparse analysis approach opens interesting perspectives regarding the ability to solve extended inverse problems where some physical parameters such as impedance or speed of

sound may be unknown. Using FEM, it allows to handle complex geometries, and as demonstrated on some brain source localization problems, it offers competitive robustness to noise.

Beyond model-based methods, an intensive research in machine learning has recently inspired several training-based approaches, *e.g.,* [3, 21, 44], which, however, either focus on a specific aspect of the problem at hand or even neglect its explicit physical nature. Instead, we feel that a promising research avenue are pretrained physics-driven cosparse models, potentially leading to "fully learnable" methods, unrestricted by parameterization or the geometry of the environment.

Acknowledgements This work was supported in part by the European Research Council, PLEASE project (ERC-StG-2011-277906).

References

1. L. Albera, A. Ferreol, D. Cosandier-Rimele, I. Merlet, F. Wendling, Brain source localization using a fourth-order deflation scheme. IEEE Trans. Biomed. Eng. **55**(2), 490–501 (2008)
2. L. Albera, S. Kitic, N. Bertin, G. Puy, R. Gribonval, Brain source localization using a physics-driven structured cosparse representation of EEG signals, in *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, September 21–24 (2014)
3. N. Antonello, T. van Waterschoot, M. Moonen, P.A. Naylor, Identification of surface acoustic impedances in a reverberant room using the FDTD method, in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 114–118 (IEEE, New York, 2014)
4. J. Barker, R. Marxer, E. Vincent, S. Watanabe, The third chime speech separation and recognition challenge: dataset, task and baselines, in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec (2015), pp. 504–511
5. A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**(1), 183–202 (2009)
6. N. Bertin, S. Kitić, R. Gribonval, Joint estimation of sound source location and boundary impedance with physics-driven cosparse regularization, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York, 2016), pp. 6340–6344
7. C. Bilen, S. Kitić, N. Bertin, R. Gribonval, Sparse acoustic source localization with blind calibration for unknown medium characteristics, in *iTwist-2nd international-Traveling Workshop on Interactions Between Sparse Models and Technology* (2014)
8. C. Blandin, A. Ozerov, E. Vincent, Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. Signal Process. **92**(8), 1950–1960 (2012)
9. T. Blumensath, M.E. Davies, Sampling theorems for signals from the union of finite-dimensional linear subspaces. IEEE Trans. Inf. Theory **55**(4), 1872–1882 (2009)
10. P. Bochev, R.B. Lehoucq, On the finite element solution of the pure Neumann problem. SIAM Rev. **47**(1), 50–66 (2005)
11. S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
12. J. Campbell, *Introduction to Remote Sensing* (Guilford Press, New York, 1996)
13. E.J. Candès, J.K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
14. A. Chambolle, T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imag. Vis. **40**(1), 120–145 (2011)

15. A. Chambolle, T. Pock, An introduction to continuous optimization for imaging. Acta Numer. **25**, 161–319 (2016)
16. V. Chandrasekaran, M.I. Jordan, Computational and statistical tradeoffs via convex relaxation. Proc. Natl. Acad. Sci. **110**(13), E1181–E1190 (2013)
17. G. Chardon, T. Nowakowski, J. De Rosny, L. Daudet, A blind dereverberation method for narrowband source localization. J. Select. Topics Signal Process. **9**, 815–824 (2015)
18. P.L. Combettes, J.-C. Pesquet, Proximal splitting methods in signal processing, in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering* (Springer, New York, 2011), pp. 185–212
19. L. Condat, A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl. **158**(2), 460–479 (2013)
20. G. Dassios, A. Fokas, The definite non-uniqueness results for deterministic eeg and meg data. Inv. Prob. **29**(6), 065012 (2013)
21. A. Deleforge, F. Forbes, R. Horaud, Acoustic space learning for sound-source separation and localization on binaural manifolds. Int. J. Neural Syst. **25**(01), 1440003 (2015)
22. W. Deng, W. Yin, On the global and linear convergence of the generalized alternating direction method of multipliers. J. Sci. Comput. **66**(3), 889–916 (2016)
23. A. Devaney, G. Sherman, Nonuniqueness in inverse source and scattering problems. IEEE Trans. Antennas Propag. **30**(5), 1034–1037 (1982)
24. I. Dokmanić, Listening to Distances and Hearing Shapes. PhD thesis, EPFL (2015)
25. I. Dokmanić, M. Vetterli, Room helps: acoustic localization with finite elements, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012* (IEEE, New York, 2012), pp. 2617–2620
26. P.L. Dragotti, M. Vetterli, T. Blu, Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang-fix. IEEE Trans. Signal Process. **55**(5), 1741–1757 (2007)
27. J. Eckstein, D.P. Bertsekas, On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**(1), 293–318 (1992)
28. M. Elad, P. Milanfar, R. Rubinstein, Analysis versus synthesis in signal priors. Inv. Prob. **23**(3), 947 (2007)
29. Q. Fang, D.A. Boas, Tetrahedral mesh generation from volumetric binary and grayscale images. in *2009 IEEE International Symposium on Biomedical Imaging from Nano to Macro* (IEEE, New York, 2009), pp. 1142–1145
30. D.C.-L. Fong, M. Saunders, LSMR: an iterative algorithm for sparse least-squares problems. SIAM J. Sci. Comput. **33**(5), 2950–2971 (2011)
31. I. Fried, Bounds on the spectral and maximum norms of the finite element stiffness, flexibility and mass matrices. Int. J. Solids Struct. **9**(9), 1013–1034 (1973)
32. A. Friedman, *Partial Differential Equations of Parabolic Type* (Courier Dover Publications, New York, 2008)
33. D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**(1), 17–40 (1976)
34. S. Gannot, D. Burshtein, E. Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Signal Process. **49**(8), 1614–1626 (2001)
35. M.S. Gockenbach, *Understanding and Implementing the Finite Element Method* (SIAM, Philadelphia, 2006)
36. T. Goldstein, B. O'Donoghue, S. Setzer, R. Baraniuk, Fast alternating direction optimization methods. SIAM J. Imag. Sci. **7**(3), 1588–1623 (2014)
37. R. Grech, T. Cassar, J. Muscat, K.P. Camilleri, S.G. Fabri, M. Zervakis, P. Xanthopoulos, V. Sakkalis, B. Vanrumste, Review on solving the inverse problem in EEG source analysis. J. NeuroEng. Rehabil. **7**, 5–25 (2008)
38. H. Hallez, B. Vanrumste, R. Grech, J. Muscat, W. De Clercq, A. Vergult, Y. D 'asseler, K.P. Camilleri, S.G. Fabri, S.V. Huffel, I. Lemahieu, Review on solving the forward problem in EEG source analysis. J. Neuroeng. Rehabil. **4**(4) (2007). https://doi.org/10.1186/1743-0003-4-46

39. M. Herman, T. Strohmer et al., General deviants: an analysis of perturbations in compressed sensing. IEEE J. Sel. Top. Sign. Proces. **4**(2), 342–349 (2010)
40. V. Isakov, *Inverse Problems for Partial Differential Equations*, vol. 127 (Springer Science & Business Media, Berlin, 2006)
41. B.H. Jansen, V.G. Rit, Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns? Biol. Cybern. **73**(4), 357–366 (1995)
42. R. Jenatton, J.-Y. Audibert, F. Bach, Structured variable selection with sparsity-inducing norms. J. Mach. Learn. Res. **12**, 2777–2824 (2011)
43. C.R. Johnson, Computational and numerical methods for bioelectric field problems. Crit. Rev. Biomed. Eng. **25**(1), 1–81 (1997)
44. U.S. Kamilov, H. Mansour, Learning optimal nonlinearities for iterative thresholding algorithms. IEEE Signal Process. Lett. **23**(5), 747–751 (2016)
45. Y. Kim, P. Nelson, Optimal regularisation for acoustic source reconstruction by inverse methods. J. Sound Vib. **275**(3–5), 463–487 (2004)
46. S. Kitić, N. Bertin, R. Gribonval, Hearing behind walls: localizing sources in the room next door with cosparsity, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York, 2014), pp. 3087–3091
47. S. Kitić, L. Albera, N. Bertin, R. Gribonval, Physics-driven inverse problems made tractable with cosparse regularization. IEEE Trans. Signal Process. **64**(2), 335–348 (2016)
48. K. Kowalczyk, M. van Walstijn, Modeling frequency-dependent boundaries as digital impedance filters in FDTD and K-DWM room acoustics simulations. J. Audio Eng. Soc. **56**(7/8), 569–583 (2008)
49. H. Kuttruff, *Room Acoustics* (CRC Press, Boca Raton, 2016)
50. J. Le Roux, P.T. Boufounos, K. Kang, J.R. Hershey, Source localization in reverberant environments using sparse optimization, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013* (IEEE, New York, 2013), pp. 4310–4314
51. R.J. LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems* (SIAM, Philadelphia, 2007)
52. T.T. Lin, F.J. Herrmann, Compressed wavefield extrapolation. Geophysics **72**(5), SM77–SM93 (2007)
53. D. Malioutov, A sparse signal reconstruction perspective for source localization with sensor arrays. Master's thesis (2003)
54. D. Malioutov, M. Çetin, A.S. Willsky, A sparse signal reconstruction perspective for source localization with sensor arrays. IEEE Trans. Signal Process. **53**(8), 3010–3022 (2005)
55. M. Mohr, B. Vanrumste, Comparing iterative solvers for linear systems associated with the finite difference discretisation of the forward problem in electro-encephalographic source analysis. Med. Biol. Eng. Comput. **41**, 75–84 (2003)
56. J.-J. Moreau, Proximité et dualité dans un espace hilbertien. Bull. de la Société mathématique de France, **93**, 273–299 (1965)
57. J.C. Mosher, R.M. Leahy, Source localization using recursively applied and projected (RAP) music. IEEE Trans. Signal Process. **47**(2), 332–340 (1999)
58. W. Munk, P. Worcester, C. Wunsch, *Ocean Acoustic Tomography* (Cambridge University Press, Cambridge, 2009)
59. G. Mur, Absorbing boundary conditions for the finite-difference approximation of the time-domain electromagnetic-field equations. IEEE Trans. Electromagn. Comput. **EMC-23**(4), 377–382 (1981)
60. J. Murray-Bruce, P.L. Dragotti, Solving physics-driven inverse problems via structured least squares, in *2016 24th European Signal Processing Conference (EUSIPCO)* (IEEE, New York, 2016), pp. 331–335
61. S. Nam, M.E. Davies, M. Elad, R. Gribonval, The cosparse analysis model and algorithms. Appl. Comput. Harmon. Anal. **34**(1), 30–56 (2013)
62. Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, vol. 87 (Springer Science & Business Media, New York, 2013)

63. R. Oostenvelda, P. Praamstra, The five percent electrode system for high-resolution EEG and ERP measurements. Clin. Neurophysiol. Elsevier, **112**(4), 713–719 (2001)
64. A. Pezeshki, Y. Chi, L.L. Scharf, E.K. Chong, Compressed sensing, sparse inversion, and model mismatch, in *Compressed Sensing and Its Applications* (Springer, New York, 2015), pp. 75–95
65. T. Piotrowski, D. Gutierrez, I. Yamada, J. Zygierewicz, Reduced-rank neural activity index for EEG/MEG multi-source localization, in *ICASSP'14, IEEE International Conference on Acoustics Speech and Signal Processing*, Florence, May 4–9 (2014)
66. L.C. Potter, E. Ertin, J.T. Parker, M. Cetin, Sparsity and compressed sensing in radar imaging. Proc. IEEE **98**(6), 1006–1020 (2010)
67. J. Provost, F. Lesage, The application of compressed sensing for photo-acoustic tomography. IEEE Trans. Med. Imag. **28**(4), 585–594 (2009)
68. A. Rosenthal, D. Razansky, V. Ntziachristos, Fast semi-analytical model-based acoustic inversion for quantitative optoacoustic tomography. IEEE Trans. Med. Imag. **29**(6), 1275–1285 (2010)
69. Y. Saad, *Iterative Methods for Sparse Linear Systems* (SIAM, Philadelphia, 2003)
70. O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, F. Lenzen, *Variational Methods in Imaging*, vol. 320 (Springer, Berlin, 2009)
71. P. Schimpf, C. Ramon, J. Haueisen, Dipole models for the EEG and MEG. IEEE Trans. Biomed. Eng. **49**(5), 409–418 (2002)
72. R. Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986). Reprint of the original 1979 paper from the RADC Spectrum Estimation Workshop
73. S. Shalev-Shwartz, N. Srebro, SVM optimization: inverse dependence on training set size, in *Proceedings of the 25th international conference on Machine learning* (ACM, New York, 2008), pp. 928–935
74. R. Shefi, M. Teboulle, Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. SIAM J. Optim. **24**(1), 269–297 (2014)
75. J.-L. Starck, J.M. Fadili, An overview of inverse problem regularization using sparsity, in *IEEE ICIP*, Cairo, Nov (2009), pp. 1453–1456
76. G. Strang, *Computational Science and Engineering*, vol. 791 (Wellesley-Cambridge Press, Wellesley, 2007)
77. L. Thomas, Using a computer to solve problems in physics. Appl. Digital Comput. **458**, 44–45 (1963)
78. J.A. Tropp, S.J. Wright, Computational methods for sparse solution of linear inverse problems. Proc. IEEE **98**(6), 948–958 (2010)
79. K. Uutela, M. Hamalainen, E. Somersalo, Visualization of magnetoencephalographic data using minimum current estimates. Elsevier NeuroImage **10**(2), 173–180 (1999)
80. J.-M. Valin, F. Michaud, J. Rouat, D. Létourneau, Robust sound source localization using a microphone array on a mobile robot, in *Proceedings. 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2003. (IROS 2003)*, vol. 2, pp. 1228–1233 (IEEE, New York, 2003)
81. F. Vatta, F. Meneghini, F. Esposito, S. Mininel, F. Di Salle, Realistic and spherical head modeling for EEG forward problem solution: a comparative cortex-based analysis. Comput. Intell. Neurosci. **2010**, 11pp. (2010)
82. M. Vetterli, P. Marziliano, T. Blu, Sampling signals with finite rate of innovation. IEEE Trans. Signal Process. **50**(6), 1417–1428 (2002)
83. C. Wolters, H. Köstler, C. Möller, J. Härdtlein, A. Anwander, Numerical approaches for dipole modeling in finite element method based source analysis. Int. Congr. Ser. **1300**, 189–192 (2007)

Total Variation Minimization in Compressed Sensing

Felix Krahmer, Christian Kruschel, and Michael Sandbichler

Abstract This chapter gives an overview over recovery guarantees for total variation minimization in compressed sensing for different measurement scenarios. In addition to summarizing the results in the area, we illustrate why an approach that is common for synthesis sparse signals fails and different techniques are necessary. Lastly, we discuss a generalization of recent results for Gaussian measurements to the subgaussian case.

Keywords Compressed sensing · Total variation minimization · Gradient sparsity

1 Introduction

The central aim of compressed sensing (CS) [4, 8] is the recovery of an unknown vector from very few linear measurements. Put formally, we would like to recover $x \in \mathbb{R}^n$ from $y = Ax + e \in \mathbb{R}^m$ with $m \ll n$, where e denotes additive noise.

For general x, recovery is certainly not possible; hence additional structural assumptions are necessary in order to be able to guarantee recovery. A common assumption used in CS is that the signal is *sparse*. Here for x we assume

F. Krahmer (✉)
Department of Mathematics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching/Munich, Germany
e-mail: felix.krahmer@tum.edu

C. Kruschel
Georg-August-University Göttingen, Institute for Numerical and Applied Mathematics, Lotzestr. 16-18, 37083 Göttingen, Germany

IAV GmbH, Development Center, Rockwellstr. 16, 38518 Gifhorn, Germany

M. Sandbichler
Department of Mathematics, University of Innsbruck, Technikerstraße 13, 6020 Innsbruck, Austria

H. Boche et al. (eds.), *Compressed Sensing and its Applications*, Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-69802-1_11

$$\|x\|_0 := |\{k \in [n]\colon x_k \neq 0\}| \leq s,$$

that is, there are only very few nonzero entries of x. And say that x is s-sparse for some given sparsity level $s \ll n$. We call a vector *compressible* if it can be approximated well by a sparse vector. To quantify the quality of approximation, we let

$$\sigma_s(x)_q := \inf_{\|z\|_0 \leq s} \|z - x\|_q$$

denote the error of the best s-sparse approximation of x.

In most cases, the vector x is not sparse in the standard basis, but there is a basis Ψ, such that $x = \Psi z$ and z is sparse. This is also known as *synthesis sparsity* of x. To find an (approximately) synthesis sparse vector, we can instead solve the problem of recovering z from $y = A\Psi z$. A common strategy in CS is to solve a basis pursuit program in order to recover the original vector. For a fixed noise level ε, it is given by

$$\text{minimize} \|z\|_1 \text{ such that } \|Az - y\|_2 \leq \varepsilon. \tag{1}$$

While this and related approaches of convex regularization have been studied in the inverse problems and statistics literature long before the field of compressed sensing developed, these works typically assumed the measurement setup was given. The new paradigm arising in the context of compressed sensing was to attempt to use the remaining degrees of freedom of the measurement system to reduce the ill-posedness of the system as much as possible. In many measurement systems, the most powerful known strategies will be based on randomization, i.e., the free parameters are chosen at random.

Given an appropriate amount of randomness (i.e., for various classes of random matrices A, including some with structure imposed by underlying applications), one can show that the minimizer $\hat{x}$ of (1) recovers the original vector x with error

$$\|x - \hat{x}\|_2 \leq c \left(\frac{\sigma_s(x)_1}{\sqrt{s}} + \varepsilon \right); \tag{2}$$

see, e.g., [1] for an elementary proof in the case of subgaussian matrices without structure and [16] for an overview, including many references, of corresponding results for random measurement systems with additional structure imposed by applications. Note that (2) entails that if x is s-sparse and the measurements are noiseless, the recovery is exact.

For many applications, however, the signal model of sparsity in an orthonormal basis has proven somewhat restrictive. Two main lines of generalization have been proposed. The first line of work, initiated by [31], is the study of sparsity in redundant representation systems, at first under incoherence assumptions on the dictionary. More recently, also systems without such assumptions have been analyzed [5, 20]. The main idea of these works is that even when one cannot recover

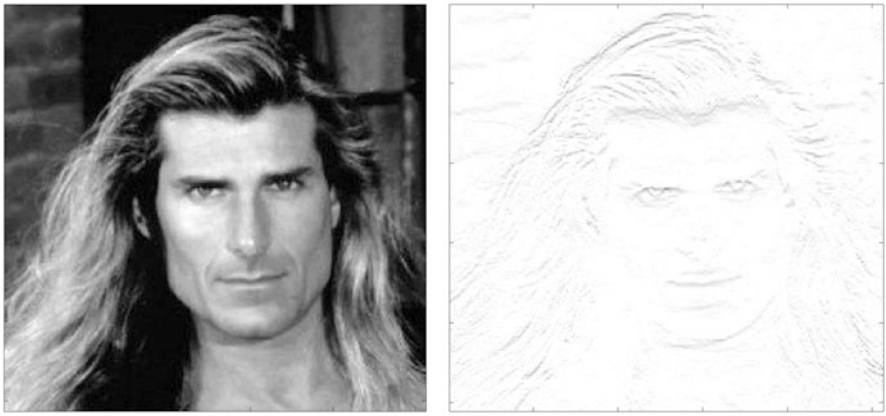


Fig. 1 The original Fabio image (left) and the absolute values after application of a discrete gradient operator (right).

the coefficients correctly due to conditioning problems, one may still hope for a good approximation of the signal.

The second line of work focuses on signals that are sparse after the application of some transform; one speaks of *cosparsity* or *analysis sparsity* [24]; see, e.g., [13] for an analysis of the Gaussian measurement setup in this framework. A special case of particular importance, especially for imaging applications, is that of sparse gradients. Namely, as it turns out, natural images often admit very sparse approximations in the gradient domain; see, e.g., Figure 1. Here the discrete gradient at location $i = (i_1, \ldots, i_n)$ is defined as the vector with its n entries given by $\big((\nabla z)_i\big)_j = z_{i+e_j} - z_i, j = 1, \ldots, n$, where e_j is the j-th standard basis vector.

A first attempt to recover a gradient sparse signal is to formulate a compressed sensing problem in terms of the sparse gradient. When this is possible (for instance, in the example of Fourier measurements [4]), applying (1) will correspond to minimizing $\|\nabla z\|_1 =: \|z\|_{TV}$, the *total variation seminorm*. Then (under some additional assumptions) compressed sensing recovery guarantees of the form (2) can apply. This proof strategy, however, only allows for showing that the gradient can be approximately recovered, not the signal. When no noise is present and the gradient is exactly sparse (which is not very realistic), this allows for signal recovery via integrating the gradient, but in case of noisy measurements, this procedure is highly unstable.

Nevertheless, the success motivates to minimize the total variation seminorm if one attempts to recover the signal directly, not the gradient. In analogy with (1), this yields the following minimization problem.

$$\text{minimize } \|z\|_{TV} = \|\nabla z\|_1 \text{ such that } \|Az - y\|_2 \leq \varepsilon.$$

For A the identity (i.e., not reducing the dimension), this relates to the famous Rudin-Osher-Fatemi functional, a classical approach for signal and image denoising [34]. Due to its high relevance for image processing, this special case of analysis

sparsity has received a lot of attention recently also in the compressed sensing framework where A is dimension reducing. The purpose of this chapter is to give an overview of recovery results for total variation minimization in this context of compressed sensing (Section 2) and to provide some geometric intuition by discussing the one-dimensional case under Gaussian or subgaussian measurements (to our knowledge, a generalization to the latter case does not appear yet in the literature) with a focus on the interaction between the high-dimensional geometry and spectral properties of the gradient operator (Section 3).

2 An Overview over TV Recovery Results

In this section, we will give an overview of the state-of-the-art guarantees for the recovery of gradient sparse signals via total variation minimization. We start by discussing in Section 2.1 sufficient conditions for the success of TV minimization.

Subsequently, we focus on recovery results for random measurements. Interestingly, the results in one dimension differ severely from the ones in higher dimensions. Instead of obtaining a required number of measurements roughly on the order of the sparsity level s, we need $\sqrt{sn}$ measurements for recovery. We will see this already in Section 2.2, where we present the results of Cai and Xu [3] for recovery from Gaussian measurements. In Section 3, we will use their results to obtain refined results for noisy measurements as well as guarantees for subgaussian measurements, combined with an argument of Tropp [37]. In Section 2.3 we will present results by Ward and Needell for dimensions larger or equal than two showing that recovery can be achieved from Haar-incoherent measurements.

2.1 Sufficient Recovery Conditions

Given linear measurements $Ax = y$ for an arbitrary $A \in \mathbb{R}^{m\times n}$ and a signal x with $\|\nabla x\|_0 \leq s$, a natural way to recover x is by solving

$$\text{minimize } \|\nabla z\|_1 \text{ such that } Az = y. \tag{3}$$

For $I \subset [n]$ we denote A_I as the columns of A indexed by I, and for a consecutive notation, we denote $\mathcal{I}_I^T \nabla$ as the rows of ∇ indexed by I and $\mathcal{I}$ as the identity matrix. The following results can also be easily applied to *analysis ℓ_1-minimization*, where any arbitrary matrix $D \in \mathbb{R}^{p\times n}$ replaces ∇ in (3), as well as to any real Hilbert space setting [21].

In many applications it is important to verify whether there is exactly one solution of (3). Since ∇ is not injective here, we cannot easily use the well-known recovery results in compressed sensing [9] for the matrix $A\nabla^\dagger$. However, a necessary condition can be given since x can only satisfy $Ax = y$ and $(\nabla x)_{I^c} = 0$ if

$$\ker(\mathcal{I}_{I^c}^T \nabla) \cap \ker(A) = \{0\}.$$

If ∇ is replaced by the identity, this is equivalent to A_I being injective. Since this injectivity condition is unavoidable, we assume for the rest of this section that it is satisfied.

The paper [24] provides sufficient and necessary conditions for uniform recovery via (3). The conditions rely on the null space of the measurements and are hard to verify similar to the classical compressed sensing setup [36]. The following result is a corollary of these conditions. It no longer provides a necessary condition but is more manageable.

Corollary 2.1. *[24] For all $x \in \mathbb{R}^n$ with $s := \|\nabla x\|_0$, the solution of (3) with $y = Ax$ is unique and equal to x if for all $I \subset [n]$ with $|I| \leq s$ it holds that*

$$\forall w \in \ker(A)\backslash\{0\}: \ \|(\nabla w)_I\|_1 < \|(\nabla w)_{I^c}\|_1.$$

To consider measurements for specific applications, where it is difficult to prove whether uniform recovery is guaranteed, one can empirically examine whether specific elements x solve (3) uniquely. For computed tomography measurements, a *Monte Carlo Experiment* is considered in [12] to approximate the fraction of all gradient s-sparse vectors to uniquely solve (3). The results prompt that there is a sharp transition between the case that every vector with a certain gradient sparsity is uniquely recoverable and the case that TV minimization will find a different solution than the desired vector. This behavior empirically agrees with the phase transition in the classical compressed sensing setup with Gaussian measurements [7].

To efficiently check whether many specific vectors x can be uniquely recovered via (3), one needs to establish characteristics of x which must be easily verifiable. Such a nonuniform recovery condition is given in the following theorem.

Theorem 2.1. *[12] It holds that $x \in \mathbb{R}^n$ is a unique solution of (3) if and only if there exists $w \in \mathbb{R}^m$ and $v \in \mathbb{R}^{n-1}$ such that*

$$\nabla^T v = A^T w, v_I = sign(\nabla x)_I, \|v_{I^c}\|_\infty < 1. \tag{4}$$

The basic idea of the proof is to use the optimality condition for convex optimization problems [33]. Equivalent formulations of the latter theorem can be found in [13, 39] where the problem is considered from a geometric perspective. However, verifying the conditions in Theorem 2.1 still requires solving a linear program where an optimal v for (4) needs to be found. In classical compressed sensing, the *Fuchs Condition* [10] is known as a weaker result as it suggests a particular w in (4) and avoids solving the consequential linear program. The following result generalizes this result to general analysis ℓ_1-minimization.

Corollary 2.2. *If $x \in \mathbb{R}^n$ satisfies*

$$\|(\mathcal{I}_{I^c}^T \nabla (\nabla^T \mathcal{I}_{I^c} \mathcal{I}_{I^c}^T \nabla + A^T A)^{-1} \nabla sign(\nabla x))_I\|_\infty < 1$$

then x is the unique solution of (3).

2.2 Recovery from Gaussian Measurements

As discussed above, to date no deterministic constructions of compressed sensing matrices are known that get anywhere near an optimal number of measurements. Also for the variation of aiming to recover approximately gradient sparse measurements, the only near-optimal recovery guarantees have been established for random measurement models. Both under (approximate) sparsity and gradient sparsity assumptions, an important benchmark is that of a measurement matrix with independent standard Gaussian entries. Even though such measurements are hard to realize in practice, they can be interpreted as the scenario with maximal randomness, which often has particularly good recovery properties. For this reason, the recovery properties of total variation minimization have been analyzed in detail for such measurements. Interestingly, as shown by the following theorem, recovery properties in the one-dimensional case are significantly worse than for synthesis sparse signals and also for higher-dimensional cases. That is why we focus on this case in Section 3, providing a geometric viewpoint and generalizing the results to subgaussian measurements.

Theorem 2.2. *[3] Let the entries of $A \in \mathbb{R}^{m\times n}$ be i.i.d. standard Gaussian random variables, and let $\hat{x}$ be a solution of (3) with input data $y = Ax_0$. Then*

1. *There exist constants $c_1, c_2, c_3, c_4 > 0$, such that for $m \geq c_1\sqrt{sn}(\log n + c_2)$*

$$\mathbb{P}(\forall x_0 \colon \|\nabla x_0\|_0 \leq s \colon \hat{x} = x_0) \geq 1 - c_3 \mathrm{e}^{-c_4\sqrt{m}}.$$

2. *For any $\eta \in (0,1)$, there are constants $\tilde{c}_1, \tilde{c}_2 > 0$ and a universal constant $c_2 > 0$, such that for $s \geq \tilde{c}_0$ and $(s+1) < \frac{n}{4}$. If $m \leq \tilde{c}_1\sqrt{sn} - \tilde{c}_2$, there exist infinitely many $x_0 \in \mathbb{R}^n$ with $\|\nabla x_0\|_0 \leq s$, such that $\mathbb{P}(\hat{x} \neq x_0) \geq 1 - \eta$.*

This scaling is notably different from what is typically obtained for synthesis sparsity, where the number of measurements scales linearly with s up to log factors. Such a scaling is only obtained for higher-dimensional signals, e.g., images. Indeed, in [3], it is shown that for dimensions at least two, the number of Gaussian measurements sufficient for recovery is

$$m \geq \begin{cases} c_2 s \log^3 n, & \text{if } d = 2 \\ c_d s \log n, & \text{if } d \geq 3, \end{cases}$$

where the constant c_d depends on the dimension.

Furthermore, as we can see in Theorem 2.5, this is also the scaling one obtains for dimensions larger than 1 and Haar-incoherent measurements. Thus the scaling of $\sqrt{sn}$ is a unique feature of the *one*-dimensional case. Also note that the square root factor in the upper bound makes the result meaningless for a sparsity level on the order of the dimension. This has been addressed in [14], showing that a dimension

reduction is also possible if the sparsity level is a (small) constant multiple of the dimension.

The proof of Theorem 2.2 uses Gordon's escape through the mesh Theorem [11]. We will elaborate on this topic in Section 3.

In case we are given noisy measurements $y = Ax_0 + e$ with $\|e\|_2 \leq \varepsilon$, we can instead of solving (3) consider

$$\text{minimize } \|\nabla z\|_1 \text{ such that } \|Az - y\|_2 \leq \varepsilon. \tag{5}$$

If ∇x_0 is not exactly, but approximately sparse, and our measurements are corrupted with noise, the following result can be established.

Theorem 2.3. *[3] Let the entries of $A \in \mathbb{R}^{m\times n}$ be i.i.d. standard Gaussian random variables, and let $\hat{x}$ be a solution of (5) with input data y satisfying $\|Ax_0 - y\|_2 \leq \varepsilon$. Then for any $\alpha \in (0,1)$, there are positive constants $\delta, c_0, c_1, c_2, c_3$, such that for $m = \alpha n$ and $s = \delta n$*

$$\mathbb{P}\left(\|x_0 - \hat{x}\|_2 \leq c_2 \frac{\min_{|S|\leq s} \|(\nabla x_0)_{S^c}\|_1}{\sqrt{n}} + c_3 \frac{\varepsilon}{\sqrt{n}}\right) \geq 1 - c_0 \mathrm{e}^{-c_1 n}.$$

This looks remarkably similar to the recovery guarantees obtained for compressed sensing; note however that the number of measurements needs to be proportional to n, which is not desirable. We will present a similar result with improved number of measurements in Section 3.5.

Theorem 2.4 (Corollary of Theorem 3.4). *Let $x_0 \in \mathbb{R}^n$ be such that $\|\nabla x_0\| \leq s$ for $s > 0$ and $A \in \mathbb{R}^{m\times n}$ with $m \geq C\sqrt{ns}\log(2n)$ be a standard Gaussian matrix. Furthermore, set $y = Ax_0 + e$, where $\|e\| \leq \varepsilon$ denotes the (bounded) error of the measurement and for some absolute constants $c, \tilde{c} > 0$ the solution $\hat{x}$ of (12) satisfies*

$$\mathbb{P}\left(\|\hat{x} - x_0\| > \frac{2\varepsilon}{c\sqrt[4]{ns}(\sqrt{\log(2n)} - 1)}\right) \leq \mathrm{e}^{-\tilde{c}\sqrt{ns}}.$$

Note, however, that in contrast to Theorem 2.3, this theorem does not cover the case of gradient compressible vectors but on the other hand Theorem 3.4 also incorporates the case of special subgaussian measurement ensembles. Also, if we set $s = \delta n$, we reach a similar conclusion as in Theorem 2.3.

2.3 Recovery from Haar-Incoherent Measurements

For dimensions $d \geq 2$, Needell and Ward [25, 26] derived recovery results for measurement matrices having the restricted isometry property (RIP) when composed of the Haar wavelet transform. Here we say that a matrix Φ has the RIP

of order k and level δ if for every k-sparse vector x it holds that

$$(1-\delta)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1-\delta)\|x\|_2^2.$$

The results of [25, 26] build upon a connection between a signal's wavelet representation and its total variation seminorm first noted by Cohen, Dahmen, Daubechies, and DeVore [6].

Their theorems yield stable recovery via TV minimization for N^d-dimensional signals. For $d = 2$, notably these recovery results concern images of size $N \times N$.

Several definitions are necessary in order to be able to state the theorem. The d-dimensional discrete gradient is defined via $\nabla: \mathbb{R}^{C^d} \to \mathbb{C}^{N^d \times d}$ and maps $x \in \mathbb{C}^{N^d}$ to its discrete derivative which, for each $\alpha \in [N]^d$ is a vector $(\nabla x)_\alpha \in \mathbb{C}^d$ composed of the derivatives in all d directions. Up to now, we have always used the anisotropic version of the TV seminorm, which can be seen as taking the ℓ_1 norm of the discrete gradient. The isotropic TV seminorm is defined via a combination of ℓ_2 and ℓ_1 norms. It is given by $\|z\|_{TV_2} := \sum_{\alpha \in [N]^d} \|(\nabla z)_\alpha\|_2$. The result in [25] is given in terms of the isotropic TV seminorm but can also be formulated for the anisotropic version.

Furthermore, we will need to concatenate several measurement matrices in order to be able to state the theorem. This will be done via the concatenation operator $\oplus: \mathrm{Lin}(\mathbb{C}^n, \mathbb{C}^{k_1}) \times \mathrm{Lin}(\mathbb{C}^n, \mathbb{C}^{k_2}) \to \mathrm{Lin}(\mathbb{C}^n, \mathbb{C}^{k_1+k_2})$, which "stacks" two linear maps.

Finally, we need the notion of shifted operators. For an operator $\mathcal{B}: \mathbb{C}^{N^{l-1}} \times (N-1) \times N^{d-l} \to \mathbb{C}^q$, these are defined as the operators $\mathcal{B}_{0_l}: \mathbb{C}^{N^d} \to \mathbb{C}^q$ and $\mathcal{B}^{0_l}: \mathbb{C}^{N^d} \to \mathbb{C}^q$ concatenating a column of zeros to the end or beginning of the l-th component, respectively.

Theorem 2.5 ([25]). *Let $N = 2^n$ and fix integers p and q. Let $\mathcal{A}: \mathbb{C}^{N^d} \to \mathbb{C}^p$ be a map that has the restricted isometry property of order $2ds$ and level $\delta < 1$ if it is composed of the orthonormal Haar wavelet transform. Furthermore let $\mathcal{B}_1, \ldots, \mathcal{B}_d$ with $\mathcal{B}_j: \mathbb{C}^{(N-1)N^{d-1}} \to \mathbb{C}^q$ be such that $\mathcal{B} = \mathcal{B}_1 \oplus \mathcal{B}_2 \oplus \cdots \oplus \mathcal{B}_d$ has the restricted isometry property of order $5ds$ and level $\delta < \frac{1}{3}$. Consider the linear operator $\mathcal{M} = \mathcal{A} \oplus [\mathcal{B}_1]_{0_1} \oplus [\mathcal{B}_1]^{0_1} \oplus \cdots \oplus [\mathcal{B}_d]_{0_d} \oplus [\mathcal{B}_d]^{0_d}$. Then $\mathcal{M}: \mathbb{C}^{N^d} \to \mathbb{C}^m$ with $m = 2dq + p$, and for all $x \in \mathbb{C}^{N^d}$ we have the following. Suppose we have noisy measurements $y = \mathcal{M}(x) + e$ with $\|e\|_2 \leq \varepsilon$, then the solution to*

$$\hat{x} = \mathrm{argmin}_z \|z\|_{TV_2} \quad \text{such that } \|\mathcal{M}(z) - y\|_2 \leq \varepsilon$$

satisfies

1. $\|\nabla(x - \hat{x})\|_2 \leq c_1 \left(\frac{\|\nabla x - (\nabla x)_S\|_{1,2}}{\sqrt{s}} + \sqrt{d}\varepsilon \right)$,
2. $\|x - \hat{x}\|_{TV_2} \leq c_2 \left(\|\nabla x - (\nabla x)_S\|_{1,2} + \sqrt{sd}\varepsilon \right)$,
3. $\|x - \hat{x}\|_2 \leq c_3 d \log N \left(\frac{\|\nabla x - (\nabla x)_S\|_{1,2}}{\sqrt{s}} + \sqrt{d}\varepsilon \right)$,

for some absolute constants c_1, c_2, c_3.

From the last point of the previous theorem, we see that for noiseless measurements and gradient sparse vectors x, perfect recovery can be achieved provided the RIP assumption holds. Subgaussian measurement matrices, for example, will have the RIP, also when composed of the Haar wavelet transform H (this is a direct consequence of rotation invariance). Moreover, as shown in [17], randomizing the column signs of an RIP matrix will, with high probability, also yield a matrix that has the RIP when composed of H. An important example is a subsampled Fourier matrix with random column signs, which relates to spread spectrum MRI (cf. [30]).

2.4 Recovery from Subsampled Fourier Measurements

Fourier measurements are widely used in many applications. Especially in medical applications as parallel-beam tomography and magnetic resonance imaging, it is desirable to reduce the number of samples to spare patients' burden. In Section 2.1, this is a motivation for introducing algorithmic checks for unique solutions of (3). In this section, we consider a probabilistic approach where an incomplete measurement matrix $A \in \mathbb{C}^{m\times n}$ chosen from the discrete Fourier transform on $\mathbb{C}^N$ is considered. Therefore we consider a subset Ω of the index set $\{-\lfloor n/2\rfloor + 1, \ldots, \lceil n/2\rceil\}$, where Ω consists of m integers chosen uniformly at random and, additionally, $0 \in \Omega$. Hence, we want to recover a signal, sparse in the gradient domain, with a measurement matrix $A = (e^{2\pi ikj/n})_{k\in\Omega, j\in[n]}$. In [4] the optimal sampling cardinality for s-sparse signals in the gradient domain was given and enables to recover one-dimensional signals from $\mathcal{O}(k\log(n))$ Fourier samples. It naturally extends to two dimensions.

Theorem 2.6. *[4] With probability exceeding $1-\eta$, a signal z, which is k-sparse in the gradient domain, is the unique solution of* (3) *if*

$$m \gtrsim k(\log(n) + \log(\eta^{-1})).$$

As already discussed in the introduction, the proof of this result proceeds via recovering the gradient and then using that the discrete gradient (with periodic boundary conditions) is injective. Due to the poor conditioning of the gradient, however, this injectivity results do not directly generalize to recovery guarantees for noisy measurements. For two (and more) dimensions, such results can be obtained via the techniques discussed in the previous subsection.

These techniques, however, do not apply directly. Namely, the Fourier (measurement) basis is not incoherent to the Haar wavelet basis; in fact, the constant vector is contained in both, which makes them maximally coherent. As observed in [29], this incoherence phenomenon only occurs for low frequencies; the high-frequency Fourier basis vectors exhibit small inner products to the Haar wavelet basis. This can be taken into account using a *variable density* sampling scheme with sampling density that is larger for low frequencies and smaller for high frequencies. For such

a sampling density, one can establish the restricted isometry for the corresponding randomly subsampled discrete Fourier matrix combined with the Haar wavelet transform with appropriately rescaled rows [18]. This yields the following recovery guarantee.

Theorem 2.7 ([18]). *Fix integers $N = 2^p$, m, and s such that $s \gtrsim \log(N)$ and*

$$m \gtrsim s \log^3(s) \log^5(N). \tag{6}$$

Select m frequencies $\{(\omega_1^j, \omega_2^j)\}_{j=1}^m \subset \{-N/2+1, \dots, N/2\}^2$ i.i.d. according to

$$\begin{aligned} \mathbb{P}\big[(\omega_1^j, \omega_2^j) = (k_1, k_2)\big] &= C_N \min\left(C, \frac{1}{k_1^2 + k_2^2}\right) \\ &=: \eta(k_1, k_2), \quad -N/2+1 \le k_1, k_2 \le N/2, \end{aligned} \tag{7}$$

where C is an absolute constant and C_N is chosen such that η is a probability distribution.
Consider the weight vector $\rho = (\rho_j)_{j=1}^m$ with $\rho_j = (1/\eta(\omega_1^j, \omega_2^j))^{1/2}$, and assume that the noise vector $\xi = (\xi_j)_{j=1}^m$ satisfies $\|\rho \circ \xi\|_2 \le \varepsilon\sqrt{m}$, for some $\epsilon > 0$. Then with probability exceeding $1 - N^{-C\log^3(s)}$, the following holds for all images $f \in \mathbb{C}^{N\times N}$:
Given noisy partial Fourier measurements $y = \mathcal{F}_\Omega f + \xi$, the estimation

$$f^\# = \operatorname{argmin}_{g\in\mathbb{C}^{N\times N}} \|g\|_{TV} \quad \text{such that} \quad \|\rho \circ (\mathcal{F}_\Omega g - y)\|_2 \le \varepsilon\sqrt{m}, \tag{8}$$

where $\circ$ denotes the Hadamard product, approximates f up to the noise level and best s-term approximation error of its gradient:

$$\|f - f^\#\|_2 \lesssim \frac{\|\nabla f - (\nabla f)_s\|_1}{\sqrt{s}} + \varepsilon. \tag{9}$$

A similar optimality result is given in [28], also for noisy data and inexact sparsity. In contrast to the previous result, this result includes the one-dimensional case. The key to obtaining such a result is showing that the stable gradient recover implies the stable signal recovery, i.e.,

$$\|z\|_2 \lesssim \gamma + \|z\|_{TV} \text{ with } \|Az\|_2 \le \gamma. \tag{10}$$

Again the sampling distribution is chosen as a combination of the uniform distribution and a decaying distribution. The main idea is to use this sampling to establish (10) via the RIP. We skip technicalities for achieving the optimality in the following theorem and refer to the original article for more details.

Theorem 2.8. *[28] Let $z \in \mathbb{C}^n$ be fixed and x be a minimizer of (5) with $\varepsilon = \sqrt{m}\delta$ for some $\delta > 0$, $m \gtrsim k \log(n)(1 + \log(\eta^{-1}))$, and an appropriate sampling distribution. Then with probability exceeding $1 - \eta$, it holds that*

$$\|\nabla z - \nabla x\|_2 \lesssim \left(\delta\sqrt{k} + C_1 \frac{\|P\nabla z\|_1}{\sqrt{k}}\right), \frac{\|z - x\|_2}{\sqrt{n}} \lesssim C_2 \left(\frac{\delta}{\sqrt{s}} + C_1 \frac{\|P\nabla z\|_1}{k}\right),$$

where P is the orthogonal projection onto a k-dimensional subspace,

$$C_1 = \log(k) \log^{1/2}(m), \text{ and } C_2 = \log^2(k) \log(n) \log(m).$$

In the two-dimensional setting, the result changes to

$$\|\nabla z - \nabla x\|_2 \lesssim \left(\delta\sqrt{k} + C_3 \frac{\|P\nabla z\|_1}{\sqrt{k}}\right), \|z - x\|_2 \lesssim C_2 \left(\delta + C_3 \frac{\|P\nabla z\|_1}{k}\right),$$

with remaining C_2 and

$$C_3 = \log(k) \log(n^2/k) \log^{1/2}(n) \log^{1/2}(m).$$

These results are optimal since the best error one can archive [26] is $\|z - x\|_2 \lesssim \|P\nabla z\|_1 k^{-1/2}$.

The optimality in the latter theorems is achieved by considering a combination of uniform random sampling and variable density sampling. Uniform sampling on its own can achieve robust and stable recovery. However, the following theorem shows that the signal error is no longer optimal, but the bound on the gradient error is still optimal up to log factors. Here (10) is obtained by using the Poincaré inequality.

Theorem 2.9 ([28]). *Let $z \in \mathbb{C}^n$ be fix and x be a minimizer of (5) with $\varepsilon = \sqrt{m}\delta$ for some $\delta > 0$ and $m \gtrsim k \log(n)(1 + \log(\eta^{-1}))$ with random uniform sampling. Then with probability exceeding $1 - \eta$, it holds that*

$$\|\nabla z - \nabla x\|_2 \lesssim \left(\delta\sqrt{k} + C\frac{\|P\nabla z\|_1}{\sqrt{k}}\right), \frac{\|z - x\|_2}{\sqrt{n}} \lesssim (\delta\sqrt{s} + C\|P\nabla z\|_1),$$

where P is the orthogonal projection onto a k-dimensional subspace and $C = \log(k) \log^{1/2}(m)$.

3 TV Recovery from Subgaussian Measurements in 1D

In this section, we will apply the geometric viewpoint discussed in [38] to the problem, which will eventually allow us to show the TV recovery results for noisy subgaussian measurements mentioned in Section 2.2.

As in the original proof of the 1D recovery guarantees for Gaussian measurements [3], the *Gaussian mean width* will play an important role in our considerations.

Definition 3.1. The (Gaussian) mean width of a bounded subset K of $\mathbb{R}^n$ is defined as

$$w(K) := \mathbb{E} \sup_{x \in K-K} \langle g, x \rangle,$$

where $g \in \mathbb{R}^n$ is a vector of i.i.d. $\mathcal{N}(0, 1)$ random variables.

In [3], the mean width appears in the context of the *Gordon's escape through the mesh* approach [11] (see Section 3.4), but as we will see, it will also be a crucial ingredient in applying the Mendelson small-ball method [15, 22].

The mean width has some nice (and important) properties; it is, for example, invariant under taking the convex hull, i.e.,

$$w(\mathrm{ch}(K)) = w(K).$$

Furthermore, it is also invariant under translations of K, as $(K-x_0)-(K-x_0) = K-K$. Due to the rotational invariance of Gaussian random variables, that is, $Ug \sim g$, we also have that $w(UK) = w(K)$. Also, it satisfies the inequalities

$$w(K) = \mathbb{E} \sup_{x \in K-K} \langle g, x \rangle \leq 2\mathbb{E} \sup_{x \in K} \langle g, x \rangle \leq 2\mathbb{E} \sup_{x \in K} |\langle g, x \rangle|,$$

which are equalities if K is symmetric about 0, because then $K = -K$ and hence $K - K = 2K$.

3.1 M^* Bounds and Recovery

In order to highlight the importance of the Gaussian mean width in signal recovery, we present some arguments from [38]. Thus in this section we present a classical result, the M^* bound, which connects the mean width to recovery problems, cf. [38]. Namely, recall that due to rotational invariance, the kernel of a Gaussian random matrix $A \in \mathbb{R}^{m \times n}$ is a random subspace distributed according to the uniform distribution (the Haar measure) on the Grassmannian

$$G_{n,n-m} := \{V \leq \mathbb{R}^n : \dim(V) = n - m\}.$$

Consequently, the set of all vectors that yield the same measurements directly correspond to such a random subspace.

The average size of the intersection of this subspace with a set reflecting the minimization objective now gives us an average bound on the worst-case error.

Theorem 3.1 (M^* Bound, Theorem 3.12 in [38]). *Let K be a bounded subset of $\mathbb{R}^n$ and E be a random subspace of $\mathbb{R}^n$ of drawn from the Grassmannian $G_{n,n-m}$ according to the Haar measure. Then*

$$\mathbb{E}\,\mathrm{diam}(K \cap E) \leq C\frac{w(K)}{\sqrt{m}}, \tag{11}$$

where C is absolute constant.

Given the M^* bound, it is now straightforward to derive bounds on reconstructions from linear observations. We first look at feasibility programs – which in turn can be used to obtain recovery results for optimization problems. For that, let $K \subset \mathbb{R}^n$ be bounded and $x \in K$ be the vector we seek to reconstruct from measurements $Ax = y$ with a Gaussian matrix $A \in \mathbb{R}^{m\times n}$.

Corollary 3.1 ([23]). *Choose $\hat{x} \in \mathbb{R}^n$, such that*

$$\hat{x} \in K \text{ and } A\hat{x} = y,$$

then one has, for an absolute constant C',

$$\mathbb{E}\sup_{x\in K}\|\hat{x} - x\|_2 \leq C'\frac{w(K)}{\sqrt{m}}.$$

This corollary directly follows by choosing $C' = 2C$, observing that $\hat{x}-x \in K-K$ and that the side constraint enforces $A(\hat{x} - x) = 0$.

Via a standard construction in functional analysis, the so-called Minkowski functional, one can now cast an optimization problem as a feasibility program so that Corollary 3.1 applies.

Definition 3.2. The Minkowski functional of a bounded, symmetric set $K \subset \mathbb{R}^n$ is given by

$$\|\cdot\|_K \colon \mathbb{R}^n \to \mathbb{R} \colon x \mapsto \inf\{t > 0 \colon x \in tK\}.$$

So the Minkowski functional tells us how much we have to "inflate" our given set K in order to capture the vector x. Clearly, from the definition we have that if K is closed

$$K = \{x \colon \|x\|_K \leq 1\}.$$

If a convex set K is closed and symmetric, then $\|\cdot\|_K$ defines a norm on $\mathbb{R}^n$.

Recall that a set K is star shaped, if there exists a point $x_0 \in K$, which satisfies that for all $x \in K$, we have $\{tx_0 + (1 - t)x \colon t \in [0, 1]\} \subset K$. It is easy to see that convex sets are star shaped, but, for example, unions of subspaces are not convex, but star shaped.

For bounded, star shaped K, the notion of $\|\cdot\|_K$ now allows to establish a direct correspondence between norm minimization problems and feasibility problems. With this observation, Corollary 3.1 translates to the following result.

Corollary 3.2. *For K bounded, symmetric, and star shaped, let $x \in K$ and $y = Ax$. Choose $\hat{x} \in \mathbb{R}^n$, such that it solves*

$$\min \|z\|_K \text{ with } Az = y,$$

then

$$\mathbb{E} \sup_{x \in K} \|\hat{x} - x\|_2 \leq C' \frac{w(K)}{\sqrt{m}}.$$

Here $\hat{x} \in K$ is due to the fact that the minimum satisfies $\|\hat{x}\|_K \leq \|x\|_K \leq 1$, as $x \in K$ by assumption.

This result directly relates recovery guarantees to the mean width; it thus remains to calculate the mean width for the sets under consideration. In the following subsections, we will discuss two cases. The first one directly corresponds to the desired signal model, namely, gradient sparse vectors. These considerations are mainly of theoretical interest, as the associated minimization problem closely relates to support size minimization, which is known to be NP hard in general. The second case considers the TV minimization problem introduced above, which then also yields guarantees for the (larger) set of vectors with bounded total variation.

Note, however, that the M^* bound only gives a bound for the expected error. We can relate this result to a statement about tail probabilities using Markov's inequality, namely,

$$\mathbb{P}(\sup_{x \in K} \|x - \hat{x}\|_2 > t) \leq t^{-1} \mathbb{E} \sup_{x \in K} \|x - \hat{x}\|_2 \leq C' \frac{w(K)}{t\sqrt{m}}.$$

In the next section, we compute the mean width for the set of gradient sparse vectors, that is, we now specify the set K in Corollary 3.1 to be the set of all vectors with energy bounded by one that only have a small number of jumps.

3.2 The Mean Width of Gradient Sparse Vectors in 1D

Here [27] served as an inspiration, as the computation is very similar for the set of sparse vectors.

Definition 3.3. The jump support of a vector x is given via

$$\text{Jsupp}(x) := \{i \in [n-1] : x_{i+1} - x_i \neq 0\}.$$

The jump support captures the positions, in which a vector x changes its values. With this, we now define the set

$$K_0^s := \{x \in \mathbb{R}^n : \|x\|_2 \leq 1, |\mathrm{Jsupp}(x)| \leq s\}.$$

The set K_0^s consists of all s-gradient sparse vectors, which have *two*-norm smaller than one. We will now calculate the mean width of K_0^s in order to apply Corollary 3.1 or 3.2.

Note that we can decompose the set K_0^s into smaller sets $K_J \cap B_2^n$ with $K_J = \{x : \mathrm{Jsupp}(x) \subset J\}$, $|J| = s$ and $B_2^n = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$. As we can't add any jumps within the set K_J, it is a subspace of $\mathbb{R}^n$. We can even quite easily find an orthonormal basis for it, if we define

$$(e_{[i,j]})_k := \frac{1}{\sqrt{j-i+1}} \begin{cases} 1, & \text{if } k \in [i,j] \\ 0, & \text{else} \end{cases}.$$

As we can align all elements of $J = \{j_1, j_2, \ldots, j_s\}$ with $1 \leq j_1 < j_2 < \ldots < j_s = n$, we see that $\{e_{[1,j_1]}, e_{[j_1+1,j_2]}, e_{[j_2+1,j_3]}, \ldots, e_{[j_{s-1}+1,j_s]}\}$ forms an ONB of K_J. Now, we can write all elements $x \in K_J \cap B_2^n$ as $x = \sum_{i=1}^s \alpha_i e_{[j_{i-1}+1,j_i]}$ by setting $j_0 := 0$. The property that $x \in B_2^n$ now enforces (ONB) that $\|\alpha\|_2 \leq 1$. Now, note that $K_0^s = -K_0^s$, so we have

$$w(K_0^s) = \mathbb{E} \sup_{x \in K_0^s - K_0^s} \langle g, x \rangle = 2\mathbb{E} \sup_{x \in K_0^s} \langle g, x \rangle.$$

Using the decomposition $K_0^s = \bigcup_{|J|=s} \left(K_J \cap B_2^n\right)$, we get

$$w(K_0^s) = 2\mathbb{E} \sup_{|J|=s} \sup_{x \in K_J \cap B_2^n} \langle g, x \rangle.$$

Now

$$\sup_{x \in K_J \cap B_2^n} \langle g, x \rangle \leq \sup_{\alpha \in B_2^s} \sum_{i=1}^s \alpha_i \langle g, e_{[j_{i-1}+1,j_i]} \rangle = \sup_{\alpha \in B_2^s} \sum_{i=1}^s \alpha_i \underbrace{\sum_{k=j_{i-1}+1}^{j_i} \frac{g_k}{\sqrt{j_i - j_{i-1}}}}_{=:G_i^J}.$$

Note that G_i^J is again a Gaussian random variable with mean 0 and variance 1. Furthermore, the supremum over α is attained, if α is parallel to G^J, so we have $\sup_{x \in K_J \cap B_2^n} \langle g, x \rangle = \|G^J\|_2$. Also note that G^J has i.i.d. entries, but for different J_1, J_2, the random vectors G^{J_1} and G^{J_2} may be dependent. Our task is now to calculate $\mathbb{E} \sup_{|J|=s} \|G^J\|_2$. As it has been shown, for example, in [9], we have that

$$\sqrt{\frac{2}{\pi}}\sqrt{s} \leq \mathbb{E}\|G^J\|_2 \leq \sqrt{s},$$

and from standard results for Gaussian concentration (cf. [27]), we get

$$\mathbb{P}(\|G^J\|_2 \geq \sqrt{s} + t) \leq \mathbb{P}(\|G^J\|_2 \geq \mathbb{E}\|G_J\|_2 + t) \leq \mathrm{e}^{-t^2/2}.$$

By noting that $|\{J \subset [n]: |J| = s\}| = \binom{n}{s}$, we see by a union bound that

$$\mathbb{P}(\sup_{|J|=s} \|G^J\|_2 \geq \sqrt{s} + t) \leq \binom{n}{s}\mathbb{P}(\|G^J\|_2 \geq \sqrt{s} + t) \leq \binom{n}{s}\mathrm{e}^{-t^2/2}.$$

For the following calculation, set $X := \sup_{|J|=s} \|G^J\|_2$. By Jensen's inequality and rewriting the expectation, we have that

$$\mathrm{e}^{\lambda \mathbb{E}X} \leq \mathbb{E}\mathrm{e}^{\lambda X} = \int_0^\infty \mathbb{P}(\mathrm{e}^{\lambda X} \geq \tau)\mathrm{d}\tau.$$

Now, the previous consideration showed that

$$\mathbb{P}(\mathrm{e}^{\lambda X} \geq \underbrace{\mathrm{e}^{\lambda(\sqrt{s}+t)}}_{=:\tau}) = \mathbb{P}(X \geq \sqrt{s} + t) \leq \binom{n}{s}\mathrm{e}^{-t^2/2} = \binom{n}{s}\mathrm{e}^{-(\log(\tau)/\lambda - \sqrt{s})^2/2},$$

Computing the resulting integrals yields

$$\mathrm{e}^{\lambda \mathbb{E}X} \leq \binom{n}{s}\mathrm{e}^{-s/2}\lambda\sqrt{2\pi}\mathrm{e}^{(\sqrt{s}+\lambda)^2/2}.$$

Using a standard bound for the binomial coefficients, namely, $\binom{n}{s} \leq \mathrm{e}^{s\log(\mathrm{e}n/s)}$, we see

$$\mathrm{e}^{\lambda \mathbb{E}X} \leq \mathrm{e}^{s\log(\mathrm{e}n/s) - s/2 + (\sqrt{s}+\lambda)^2/2 + \log(\lambda) + \log(\sqrt{2\pi})},$$

or equivalently

$$\lambda \mathbb{E}X \leq s\log(\mathrm{e}n/s) - s/2 + (\sqrt{s} + \lambda)^2/2 + \log(\lambda) + \log(\sqrt{2\pi})$$

By setting $\lambda = \sqrt{s\log(\mathrm{e}n/s)}$ and assuming (reasonably) large n, we thus get

$$\mathbb{E}X \leq 5\sqrt{s\log(\mathrm{e}n/s)}.$$

From this, we see that

$$w(K_0^s) \leq 10\sqrt{s\log(en/s)}.$$

It follows that the Gaussian mean width of the set of gradient sparse vectors is the same as the mean width of sparse vectors due to the similar structure. If we want to obtain accuracy δ for our reconstruction, according to Theorem 3.1, we need to take

$$m = \mathcal{O}\left(\frac{s\log(en/s)}{\delta^2}\right)$$

measurements.

In compressed sensing, the squared mean width of the set of s-sparse vectors (its so-called statistical dimension) already determines the number of required measurements in order to recover a sparse signal with basis pursuit. This is the case because the convex hull of the set of sparse vectors can be embedded into the ℓ_1-ball inflated by a constant factor.

In the case of TV minimization, as we will see in the following section, this embedding yields a (rather large) constant depending on the dimension.

3.3 The Extension to Gradient Compressible Vectors Needs a New Approach

In the previous subsection, we considered exactly gradient sparse vectors. However searching all such vectors x that satisfy $Ax = y$ is certainly not a feasible task. Instead, we want to solve the convex program

$$\min \|z\|_{TV} \text{ with } Az = y,$$

with $\|z\|_{TV} = \|\nabla z\|_1$ the total variation seminorm. Now if we have that $x \in K_0^s$, we get that

$$\|x\|_{TV} \leq 2\|\alpha\|_1 \leq 2\sqrt{s}\|\alpha\|_2 = 2\sqrt{s},$$

with α as in Section 3.2, so $K_0^s \subset K_{TV}^{2\sqrt{s}} := \{x \in B_2^n : \|x\|_{TV} \leq 2\sqrt{s}\}$. As $K_{TV}^{2\sqrt{s}}$ is convex, we even have $\mathrm{ch}(K_0^s) \subset K_{TV}^{2\sqrt{s}}$. We can think of the set $K_{TV}^{2\sqrt{s}}$ as "gradient-compressible" vectors.

In the proof of Theorem 3.3 in [3], the Gaussian width of the set $K_{TV}^{4\sqrt{s}}$ has been calculated via a wavelet-based argument. One obtains that $w(K_{TV}^{2\sqrt{s}}) \leq C\sqrt{\sqrt{ns}\log(2n)}$ with $C \leq 20$ being an absolute constant. In this section we illustrate that proof techniques different from the ones used in the case of synthesis sparsity are indeed necessary in order to obtain useful results. In the synthesis case,

the *one*-norm ball of radius $\sqrt{s}$ is contained in the set of s-sparse vectors inflated by a constant factor. This in turn implies that the mean width of the compressible vectors is bounded by a constant times the mean width of the s-sparse vectors.

We will attempt a similar computation, that is, to find a constant, such that the set $K_{TV}^{2\sqrt{s}}$ is contained in the "inflated" set $c_{n,s}\mathrm{ch}(K_0^s)$. Then $w(K_{TV}^{2\sqrt{s}}) \leq c_{n,s} w(K_0^s)$. Although this technique works well for sparse recovery, where $c_{n,s} = 2$, it pitiably fails in the case of TV recovery as we will see below.

Let us start with $x \in K_{TV}^{2\sqrt{s}}$. Now we can decompose $J := \mathrm{Jsupp}(x) = J_1 \uplus J_2 \uplus \ldots J_p$ with $|J_k| \leq s$ in an ascending manner, i.e., for all $k \in J_i, l \in J_{i+1}$, we have that $\alpha_k < \alpha_l$. Note that the number p of such sets satisfies $p \leq \frac{n}{s}$. Similarly as above, we now write $x = \sum_{i=1}^{|J|} \alpha_i e_{[j_{i-1}+1,j_i]} = \sum_{k=1}^{p} \sum_{i \in J_k} \alpha_i e_{[j_{i-1}+1,j_i]}$. From this, we see that

$$x = \sum_{k=1}^{p} \|\alpha_{J_k}\|_2 \underbrace{\sum_{i \in J_k} \frac{\alpha_i}{\|\alpha_{J_k}\|_2} e_{[j_{i-1}+1,j_i]}}_{\in K_0^s}.$$

The necessary factor $c_{n,s}$ can be found by bounding the size of $\|\alpha_{J_k}\|_2$, namely,

$$\max(\|\alpha_{J_k}\|_2) \leq \sum_{k=1}^{p} \|\alpha_{J_k}\|_2 \overset{C-S}{\leq} \underbrace{\|\alpha\|_2}_{\leq 1} \sqrt{p} \leq \sqrt{\frac{n}{s}}.$$

From this, we see that $K_{TV}^{2\sqrt{s}} \subset \sqrt{\frac{n}{s}}\mathrm{ch}(K_0^s)$. To see that this embedding constant is optimal, we construct a vector, for which it is needed.

To simplify the discussion, suppose that n and s are even and $s|n$. For even n, the vector $x_1 = (\sqrt{\frac{1-(-1)^k\varepsilon}{n}})_k$ has unity norm, lies in $K_{TV}^{2\sqrt{s}}$ for $\varepsilon < \frac{2\sqrt{s}}{n}$ and has jump support on all of $[n]$!

For a vector $x \in \mathbb{R}^n$ and an index set $I \subset [n]$, we define the restriction of x to I by

$$(x|_I)_j := \begin{cases} x_j, \text{ if } j \in I \\ 0, \text{ else.} \end{cases}$$

By splitting $\mathrm{Jsupp}(x_1)$ into sets $J_1, \ldots, J_{n/s}$ and setting $a_k = \sqrt{\frac{n}{s}} x_1|_{J_k} \in K_0^s$, we see that $x_1 = \sum_{k=1}^{n/s} \sqrt{\frac{s}{n}} a_k$, and in order for this to be elements of $c_{n,s}\mathrm{ch}(K_0^s)$, we have to set $c_{n,s} = \sqrt{\frac{n}{s}}$. This follows from

$$x_1 = \sum_{k=1}^{n/s} x_1|_{J_k} = \sum_{k=1}^{n/s} \sqrt{\frac{s}{n}} \frac{p}{p} a_k = \sum_{k=1}^{n/s} \frac{1}{p} \underbrace{\left(\sqrt{\frac{n}{s}} a_k\right)}_{\in \sqrt{\frac{n}{s}} K_0^s} \in \sqrt{\frac{n}{s}}\mathrm{ch}(K_0^s)$$

and no smaller inflation factor than $\sqrt{\frac{n}{s}}$ can suffice.

So from the previous discussion, we get

Lemma 3.1. *We have the series of inclusions*

$$\mathrm{ch}(K_0^s) \subset K_{TV}^{2\sqrt{s}} \subset \sqrt{\frac{n}{s}}\mathrm{ch}(K_0^s).$$

In view of the results obtainable for sparse vectors and the ℓ_1-ball, this is very disappointing, because Lemma 3.1 now implies that the width of $K_{TV}^{2\sqrt{s}}$ satisfies

$$w(K_{TV}^{2\sqrt{s}}) \leq w\left(\sqrt{\frac{n}{s}}\mathrm{ch}(K_0^s)\right) = \sqrt{\frac{n}{s}}w(K_0^s) \leq 10\sqrt{n\log(\mathrm{e}(n-1)/s)},$$

which is highly suboptimal.

Luckily, the results in [3] suggest that the factor n in the previous equation can be replaced by $\sqrt{sn}$. However, they have to resort to a direct calculation of the Gaussian width of $K_{TV}^{2\sqrt{s}}$. The intuition why the Gaussian mean width can be significantly smaller than the bound given in Lemma 3.1 stems from the fact that in order to obtain an inclusion, we need to capture all "outliers" of the set – no matter how small their measure is.

3.4 Exact Recovery

For exact recovery, the M^* bound is not suitable anymore, and, as suggested in [38], we will use "Gordon's escape through the mesh" in order to find conditions on exact recovery. Exact recovery for TV minimization via this approach has first been considered in [3].

Suppose we want to recover $x \in K_0^s$ from Gaussian measurements $Ax = y$. Given that we want our estimator $\hat{x}$ to lie in a set K, exact recovery is achieved, if $K \cap \{z : Az = y\} = \{x\}$. This is equivalent to requiring

$$(K - x) \cap \underbrace{\{z - x : Az = y\}}_{=\ker(A)} = \{0\}.$$

With the descent cone $D(K, x) = \{t(z - x) : t \geq 0, z \in K\}$, we can rewrite this condition as

$$D(K, x) \cap \ker(A) = \{0\},$$

by introducing the set $S(K, x) = D(K, x) \cap B_2^n$, we see that if

$$S(K,x) \cap \ker(A) = \emptyset,$$

we get exact recovery. The question, when a section of a subset of the sphere with a random hyperplane is empty, is answered by Gordon's escape through a mesh.

Theorem 3.2 ([11]). *Let $S \subset \mathbb{S}^{n-1}$ be fixed and $E \in G_{n,n-m}$ be drawn at random according to the Haar measure. Assume that $\hat{w}(S) = \mathbb{E}\sup_{u\in S}\langle g,u\rangle < \sqrt{m}$ and then $S \cap E = \emptyset$ with probability exceeding*

$$1 - 2.5\exp\left(-\frac{(m/\sqrt{m+1} - \hat{w}(S))^2}{18}\right).$$

So we get exact recovery with high probability from a program given in Theorem 3.1 or 3.2, provided that $m > \hat{w}(S(K,x_0))^2$.

Let's see how this applies to TV minimization. Suppose we are given $x \in K_0^s$ and Gaussian measurements $Ax = y$. Solving

$$\min \|z\|_{TV} \text{ with } Az = y,$$

amounts to using the Minkowski functional of the set $K = \{z \in \mathbb{R}^n : \|z\|_{TV} \leq \|x\|_{TV}\}$, which is a scaled TV ball.

In [3], the null space property for TV minimization given in Corollary 2.1 has been used in order to obtain recovery guarantees.

They consider the set, where this condition is not met

$$\mathcal{S} := \{x' \in B_2^n : \exists J \subset [n], |J| \leq s, \|(\nabla x')_J\|_1 \geq \|(\nabla x')_{J^c}\|_1\},$$

and apply Gordon's escape through the mesh to see that with high probability, its intersection with the kernel of A is empty, thus proving exact recovery with high probability. Their estimate to the mean width of the set $\mathcal{S}$

$$\hat{w}(\mathcal{S}) \leq c\sqrt[4]{ns}\sqrt{\log(2n)}$$

with $c < 19$ is essentially optimal (up to logarithmic factors), as they also show that $w(\mathcal{S}) \geq C\sqrt[4]{ns}$. So uniform exact recovery can only be expected for $m = \mathcal{O}(\sqrt{sn}\log n)$ measurements.

Let us examine some connections to the previous discussion about the descent cone.

Lemma 3.2. *We have that for $K = \{z \in \mathbb{R}^n : \|z\|_{TV} \leq \|x\|_{TV}\}$ defined as above and $x \in K_0^s$, it holds that $S(K,x) \subset \mathcal{S}$.*

Proof. Let $y \in S(K,x)$. Then there exists a $x \neq z \in K$, such that $y = \frac{z-x}{\|z-x\|_2}$. Set $J = \mathrm{Jsupp}(x)$; then, as $z \in K$, we have that $\|z\|_{TV} \leq \|x\|_{TV}$, or

$$\sum_{i\in J} |(\nabla x)_i| \geq \sum_{i\in J} |(\nabla z)_i| + \sum_{i\notin J} |(\nabla z)_i|$$

Now, by the triangle inequality and this observation, we have

$$\sum_{i\in J} |(\nabla x)_i - (\nabla z)_i| \geq \sum_{i\in J} |(\nabla x)_i| - |(\nabla z)_i| \geq \sum_{i\notin J} |(\nabla z)_i| = \sum_{i\notin J} |(\nabla x)_i - (\nabla z)_i|.$$

The last equality follows from the fact that ∇x is zero outside of the gradient support of x. Multiplying both sides with $\frac{1}{\|z-x\|_2}$ gives the desired result

$$\begin{aligned}\|(\nabla y)_J\|_1 &= \frac{1}{\|z-x\|_2} \sum_{i\in J} |(\nabla x)_i - (\nabla z)_i| \\ &\geq\geq \frac{1}{\|z-x\|_2} \sum_{i\notin J} |(\nabla x)_i - (\nabla z)_i| = \|(\nabla y)_{J^c}\|_1.\end{aligned}$$

The previous lemma shows that the recovery guarantees derived from the null space property and via the descent cone are actually connected in a very simple way.

Clearly, now if we do not intersect the set $\mathcal{S}$, we also do not intersect the set $S(K, x)$, which yields exact recovery, for example, with the same upper bounds on m as for $\mathcal{S}$. Even more specifically, in the calculation of $\hat{w}(\mathcal{S})$ given in [3], an embedding into a slightly larger set $\tilde{\mathcal{S}} = \{x \in B_2^n : \|x\|_{TV} \leq 4\sqrt{s}\}$ is made. This embedding can also quite easily be done if we note that $\|x\|_{TV} \leq 2\sqrt{s}$, as we showed above, and $\|z\|_{TV} \leq \|x\|_{TV}$.

Note that the same discussion also holds for higher-dimensional signals, such that the improved numbers of measurements as given in Section 2.2 can be applied.

3.5 Subgaussian Measurements

Up to this point, all our measurement matrices have been assumed to consist of i.i.d. Gaussian random variables. We will reduce this requirement in this section to be able to incorporate also subgaussian measurement matrices into our framework.

Definition 3.4. A real valued random variable X is called *subgaussian*, if there exists a number $t > 0$, such that $\mathbb{E}e^{tX^2} < \infty$. A real valued random vector is called subgaussian, if all of its one-dimensional marginals are subgaussian.

An obvious example of subgaussian random variables is Gaussian random variables, as the expectation in Definition 3.4 exists for all $t < 1$. Also, all bounded random variables are subgaussian.

Here, we rely on results given by Tropp in [37] using the results of Mendelson [15, 22]. We will consider problems of the form

$$\min \|z\|_{TV} \text{ such that } \|Az - y\| \leq \varepsilon, \tag{12}$$

where A is supposed to be a matrix with independent subgaussian rows. Furthermore, we denote the exact solution by x_0, i.e., $Ax_0 = y$. We pose the following assumptions on the distribution of the rows of A.

(M1) $\mathbb{E}A_i = 0$,
(M2) There exists $\alpha > 0$, such that for all $u \in \mathbb{S}^{n-1}$ it holds that $\mathbb{E}|\langle A_i, u\rangle| \geq \alpha$,
(M3) There is a $\sigma > 0$, such that for all $u \in \mathbb{S}^{n-1}$ it holds that $\mathbb{P}(|\langle A_i, u\rangle| \geq t) \leq 2\exp(-t^2/(2\sigma^2))$,
(M4) The constant $\rho := \frac{\sigma}{\alpha}$ is small.

Then the small-ball methods yield the following recovery guarantee (we present the version of [37]).

Theorem 3.3. *Let $x_0 \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$ be a subgaussian matrix satisfying (M1)-(M4) above. Furthermore, set $y = Ax_0 + e$, where $\|e\| \leq \varepsilon$ denotes the (bounded) error of the measurement. Then the solution $\hat{x}$ of (12) satisfies*

$$\|\hat{x} - x_0\| \leq \frac{2\varepsilon}{\max\{c\alpha\rho^{-2}\sqrt{m} - C\sigma w(S(K, x_0)) - \alpha t, 0\}}$$

with probability exceeding $1 - \mathrm{e}^{-ct^2}$. $D(K, x_0)$ denotes the descent cone of the set K at x_0, as defined in the previous section.

From this we see that, provided

$$m \geq \tilde{C}\rho^6 w^2(S(K, x_0)),$$

we obtain stable reconstruction of our original vector from (12). Note that the theorem is only meaningful for $t = \mathcal{O}(\sqrt{m})$, as otherwise the denominator vanishes.

In the previous section, we have shown the inclusion $S(K, x_0) \subset \mathcal{S}$ for $x_0 \in K_s^0$, and hence we have that

$$w(S(K, x_0)) \leq w(\mathcal{S}) \leq c\sqrt[4]{ns}\sqrt{\log(2n)}.$$

So we see that for $m \geq \tilde{C}\rho^6\sqrt{ns}\log(2n)$, we obtain the bound

$$\begin{aligned}\|\hat{x} - x_0\| &\leq \frac{2\varepsilon}{\max\{c\alpha\rho^{-2}\sqrt{\tilde{C}}\rho^3\sqrt[4]{ns}\sqrt{\log(2n)} - C\sigma\sqrt[4]{ns}\sqrt{\log(2n)} - \alpha t, 0\}} \\ &= \frac{2\varepsilon}{\max\{\sigma(c\sqrt{\tilde{C}} - C)\sqrt[4]{ns}\sqrt{\log(2n)} - \alpha t, 0\}}\end{aligned}$$

with high probability. We conclude that, given the absolute constants c, C, we need to set $\tilde{C} \geq \frac{C^2}{c^2}$ in order to obtain a meaningful result. Combining all our previous discussions with Theorem 3.3, we get

Theorem 3.4. *Let $x_0 \in \mathbb{R}^n$, $m \geq \tilde{C}\rho^6 \sqrt{ns}\log(2n)$ and $A \in \mathbb{R}^{m\times n}$ be a subgaussian matrix satisfying (M1)–(M4). Furthermore, set $y = Ax_0 + e$, where $\|e\| \leq \varepsilon$ denotes the (bounded) error of the measurement, constants $c, C, \tilde{C} > 0$ as above, and $t \leq \frac{\sigma(c\sqrt{\tilde{C}}-C)\sqrt[4]{ns}\sqrt{\log(2n)}}{\alpha}$. Then the solution $\hat{x}$ of (12) satisfies*

$$\mathbb{P}\left(\|\hat{x} - x_0\| > \frac{2\varepsilon}{\sigma(c\sqrt{\tilde{C}} - C)\sqrt[4]{ns}\sqrt{\log(2n)} - \alpha t}\right) \leq \mathrm{e}^{-ct^2}.$$

We can, for example, set $t = \rho(c\sqrt{\tilde{C}} - C)\sqrt[4]{ns}$ (for $n \geq 2$) to obtain the bound

$$\mathbb{P}\left(\|\hat{x} - x_0\| > \frac{2\varepsilon}{\sigma(c\sqrt{\tilde{C}} - C)\sqrt[4]{ns}(\sqrt{\log(2n)} - 1)}\right) \leq \mathrm{e}^{-\tilde{c}\rho\sqrt{ns}}.$$

For example, for i.i.d. standard Gaussian measurements, the constant $\rho = \sqrt{\frac{2}{\pi}}$.

Note that in the case of noise-free measurements $\varepsilon = 0$, Theorem 3.4 gives an exact recovery result for a wider class of measurement ensembles with high probability. Furthermore with a detailed computation of $w(S(K, x_0))$, one may be able to improve the number of measurements for nonuniform recovery. It also remains open, whether the lower bounds of Cai and Xu for the case of Gaussian measurements can be generalized to the subgaussian case. In fact, our numerical experiments summarized in Figure 2 suggest a better scaling in the ambient dimension, around $N^{1/4}$, in the average case. We consider it an interesting problem for future work to explore whether this is due to a difference between Rademacher and Gaussian matrix entries, between uniform and nonuniform recovery, or between the average and the worst case. Also, it is not clear whether the scaling is in fact $N^{1/4}$ or if the observed slope is just a linearization of, say, a logarithmic dependence.

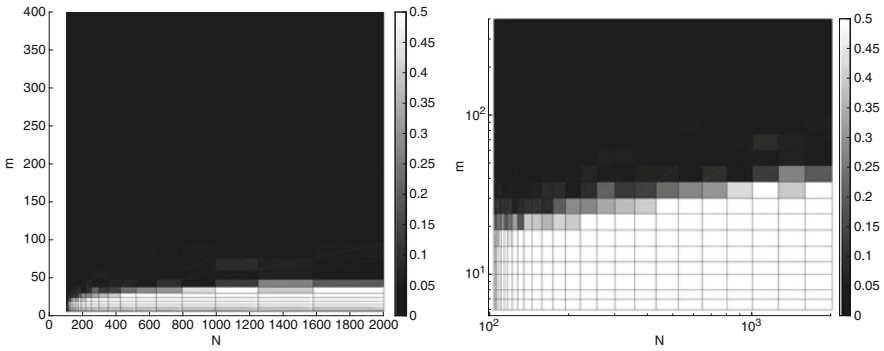


Fig. 2 Average error of recovery from Rademacher measurements in $1d$ with m measurements and ambient dimension N for fixed cosparsity level $s = 5$. Left: linear axis scaling, Right: logarithmic axis scaling. The slope of the phase transition in the log-log plot is observed to be about $\frac{1}{4}$.

4 Discussion and Open Problems

As the considerations in the previous sections illustrate, the mathematical properties of total variation minimization differ significantly from algorithms based on synthesis sparsity, especially in one dimension. For this reason, there are a number of questions that have been answered for synthesis sparsity, which are still open for the framework of total variation minimization. For example, the analysis provided in [19, 32] for deterministically subsampled partial random circulant matrices, as they are used to model measurement setups appearing in remote sensing or coded aperture imaging, could not be generalized to total variation minimization. The difficulty in this setup is that the randomness is encoded by the convolution filter, so it is not clear what the analogy of variable density sampling would be.

Another case of practical interest is that of sparse $0/1$ measurement matrices. Recently it has been suggested that such measurements increase efficiency in photoacoustic tomography, while at the same time, the signals to be recovered (after a suitable temporal transform) are approximately gradient sparse. This suggests the use of total variation minimization for recovery, and indeed empirically, this approaches yields good recovery results [35]. Theoretical guarantees, however, (as they are known for synthesis sparse signals via an expander graph construction [2]) are not available to date for this setup.

Acknowledgments FK and MS acknowledge support by the Hausdorff Institute for Mathematics (HIM), where part of this work was completed in the context of the HIM trimester program "Mathematics of Signal Processing"; FK and CK acknowledge support by the German Science Foundation in the context of the Emmy Noether Junior Research Group "Randomized Sensing and Quantization of Signals and Images" (KR 4512/1-1) and by the German Ministry of Research and Education in the context of the joint research initiative ZeMat. MS has been supported by the Austrian Science Fund (FWF) under Grant no. Y760 and the DFG SFB/TRR 109 "Discretization in Geometry and Dynamics."

References

1. R.G. Baraniuk, M. Davenport, R.A. DeVore, M. Wakin, A simple proof of the restricted isometry property for random matrices. Constr. Approx. **28**(3), 253–263 (2008)
2. R. Berinde, A. Gilbert, P. Indyk, H. Karloff, M. Strauss, Combining geometry and combinatorics: a unified approach to sparse signal recovery, in *46th Annual Allerton Conference on Communication, Control, and Computing*, 2008 (IEEE, New York, 2008), pp. 798–805
3. J.-F. Cai, W. Xu, Guarantees of total variation minimization for signal recovery. Inf. Infer. **4**(4), 328–353 (2015)
4. E.J. Candes, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(3), 489–509 (2006)
5. E.J. Candès, Y.C. Eldar, D. Needell, P. Randall, Compressed sensing with coherent and redundant dictionaries. Appl. Comput. Harmon. Anal. **31**(1), 59–73 (2010)
6. A. Cohen, W. Dahmen, I. Daubechies, R. DeVore, Harmonic analysis of the space BV. Rev. Mat. Iberoam. **19**(1), 235–263 (2003)

7. D. Donoho, High-dimensional centrally-symmetric polytopes with neighborliness proportional to dimension. Technical report, Department of Statistics, Stanford University (2004)
8. D.L. Donoho, Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
9. S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Springer, Berlin, 2013)
10. J.J. Fuchs, On sparse representations in arbitrary redundant bases. IEEE Trans. Inf. Theory **50**(6), 1341–1344 (2004)
11. Y. Gordon, *On Milman's Inequality and Random Subspaces which Escape Through a Mesh in* $\mathbb{R}^n$ (Springer, Berlin, 1988)
12. J. Jørgensen, C. Kruschel, D. Lorenz, Testable uniqueness conditions for empirical assessment of undersampling levels in total variation-regularized x-ray CT. Inverse Prob. Sci. Eng. **23**(8), 1283–1305 (2015)
13. M. Kabanava, H. Rauhut, Analysis ℓ_1-recovery with frames and Gaussian measurements. Acta Appl. Math. **140**(1), 173–195 (2015)
14. M. Kabanava, H. Rauhut, H. Zhang, Robust analysis ℓ_1-recovery from Gaussian measurements and total variation minimization. Eur. J. Appl. Math. **26**(06), 917–929 (2015)
15. V. Koltchinskii, S. Mendelson, Bounding the smallest singular value of a random matrix without concentration. Int. Math. Res. Not. **2015**(23), 12991–13008 (2015)
16. F. Krahmer, H. Rauhut, Structured random measurements in signal processing. GAMM-Mitteilungen **37**(2), 217–238 (2014)
17. F. Krahmer, R. Ward, New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. SIAM J. Math. Anal. **43**(3), 1269–1281 (2011)
18. F. Krahmer, R. Ward, Stable and robust sampling strategies for compressive imaging. IEEE Trans. Image Process. **23**(2), 612–622 (2014)
19. F. Krahmer, S. Mendelson, H. Rauhut, Suprema of chaos processes and the restricted isometry property. Commun. Pure Appl. Math. **67**(11), 1877–1904 (2014)
20. F. Krahmer, D. Needell, R. Ward, Compressive sensing with redundant dictionaries and structured measurements. SIAM J. Math. Anal. **47**(6), 4606–4629 (2015)
21. C. Kruschel, Geometrical interpretations and algorithmic verification of exact solutions in compressed sensing. PhD thesis, TU Braunschweig (2015)
22. S. Mendelson, Learning without concentration, in *COLT* (2014), pp. 25–39
23. S. Mendelson, A. Pajor, N. Tomczak-Jaegermann, Reconstruction and subgaussian operators in asymptotic geometric analysis. Geom. Funct. Anal. **17**(4), 1248–1282 (2007)
24. S. Nam, M. Davies, M. Elad, R. Gribonval, The cosparse analysis model and algorithms. Appl. Comput. Harmon. Anal. **34**(1), 30–56 (2013)
25. D. Needell, R. Ward, Near-optimal compressed sensing guarantees for total variation minimization. IEEE Trans. Image Process. **22**(10), 3941–3949 (2013)
26. D. Needell, R. Ward, Stable image reconstruction using total variation minimization. SIAM J. Imag. Sci. **6**(2), 1035–1058 (2013)
27. Y. Plan, R. Vershynin, Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. IEEE Trans. Inf. Theory **59**(1), 482–494 (2013)
28. C. Poon, On the role of total variation in compressed sensing. SIAM J. Imag. Sci. **8**(1), 682–720 (2015)
29. G. Puy, P. Vandergheynst, Y. Wiaux, On variable density compressive sampling. IEEE Signal Process. Lett. **18**, 595–598 (2011)
30. G. Puy, J. Marques, R. Gruetter, J.-P. Thiran, D. Van De Ville, P. Vandergheynst, Y. Wiaux, Spread spectrum magnetic resonance imaging. IEEE Trans. Med. Imag. **31**(3), 586–598 (2012)
31. H. Rauhut, K. Schnass, P. Vandergheynst, Compressed sensing and redundant dictionaries. IEEE Trans. Inf. Theory **54**(5), 2210–2219 (2008)
32. H. Rauhut, J. Romberg, J.A. Tropp, Restricted isometries for partial random circulant matrices. Appl. Comp. Harmon. Anal. **32**(2), 242–254 (2012)
33. R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1972)
34. L.I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms. Physica D **60**(1–4), 259–268 (1992)

35. M. Sandbichler, F. Krahmer, T. Berer, P. Burgholzer, M. Haltmeier, A novel compressed sensing scheme for photoacoustic tomography. SIAM J. Appl. Math. **75**(6), 2475–2494 (2015)
36. A. Tillmann, M. Pfetsch, The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. IEEE Trans. Inf. Theory **60**(2), 1248–1259 (2014)
37. J. Tropp, Convex recovery of a structured signal from independent random linear measurements, in *Sampling Theory, a Renaissance* (Springer, Berlin, 2015), pp. 67–101
38. R. Vershynin, Estimation in high dimensions: a geometric perspective, in *Sampling Theory, a Renaissance* (Springer, Berlin, 2015), pp. 3–66
39. H. Zhang, Y. Ming, W. Yin, One condition for solution uniqueness and robustness of both ℓ_1-synthesis and ℓ_1-analysis minimizations. Adv. Comput. Math. **42**(6), 1381–1399 (2016)

Compressed Sensing in Hilbert Spaces

Yann Traonmilin, Gilles Puy, Rémi Gribonval, and Mike E. Davies

Abstract In many linear inverse problems, we want to estimate an unknown vector belonging to a high-dimensional (or infinite-dimensional) space from few linear measurements. To overcome the ill-posed nature of such problems, we use a low-dimensional assumption on the unknown vector: it belongs to a low-dimensional model set. The question of whether it is possible to recover such an unknown vector from few measurements then arises. If the answer is yes, it is also important to be able to describe a way to perform such a recovery. We describe a general framework where appropriately chosen random measurements guarantee that recovery is possible. We further describe a way to study the performance of recovery methods that consist in the minimization of a regularization function under a data-fit constraint.

Keywords Inverse problem · Low-complexity model · Regularization · Dimension reduction · Compressed sensing · Random projection

Gilles Puy contributed to the results reported in this chapter when he was with Univ Rennes, Inria, CNRS, IRISA.

Y. Traonmilin (✉) • R. Gribonval
Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France
e-mail: yann.traonmilin@math.u-bordeaux.fr; remi.gribonval@inria.fr

G. Puy
Technicolor , 975 Avenue des Champs Blancs, 35576 Cesson-Sévigné, France
e-mail: Gilles.Puy@technicolor.com

M.E. Davies
Institute for Digital Communications (IDCom), University of Edinburgh,
The King's Buildings, Edinburgh EH9 3JL, UK
e-mail: mike.davies@ed.ac.uk

H. Boche et al. (eds.), *Compressed Sensing and its Applications*,
Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-69802-1_12

1 Introduction

Many signal processing tasks aim at estimating a signal x from its observation y. The signal x can often be described by a continuous physical phenomenon, and the observations y are made of a finite collection of scalar measurements. The most basic example of such observations is a sampled version of the signal x (e.g., for a sound recorded at a given sampling rate, the continuous x is the electrical signal produced by the microphone over time). More generally, we consider observations y modeled as

$$y = Ax + e \tag{1}$$

where $x \in \mathcal{H}$, $y \in \mathcal{F}$, and $\mathcal{H}, \mathcal{F}$ are Hilbert spaces of finite or infinite dimension. The operator A is a linear map, and e is a noise whose energy $\|e\|_{\mathcal{F}}$ is bounded. In most cases, the operator A models a finite number of measurements m. This Hilbert space setting is a way to have a general view of signal recovery problems in classical finite- or infinite-dimensional spaces where signals (in a wide sense: time series, images, videos, ...) are modeled, e.g., the space of continuous signals with finite energy $\mathcal{L}^2(\mathbb{R}^d)$, the space of bandlimited signals with finite energy or its equivalent after sampling, $\ell^2(\mathbb{R}^d)$, or the finite-dimensional vector space $\mathbb{R}^d$.

1.1 Observation Model and Low-Complexity Signals

Observing a continuous signal with finitely many linear measurements induces an information loss. If no further prior information on the signal is available, recovering x from y is generally not possible. However, if an (approximate) hypothesis of "low complexity" on x is available, enforcing the hypothesis in the recovery process can ensure that we are able to estimate x with reasonable accuracy. Low complexity can be defined in several ways. It often means that the signal lives in a "low-dimensional model" or can be described by few parameters. Two classical examples where low complexity helps to recover the signal are:

- Sampling of periodic bandlimited signals in $\mathcal{H} = \mathcal{L}^2(\mathbb{R})$: if the signal is known to be bandlimited with cutoff frequency B, it is possible to recover it perfectly provided it is sampled at a rate at least $2B$.
- Compressed sensing in $\mathcal{H} = \mathbb{R}^n$: if the signal is known to have at most k non-zero samples in $\mathbb{R}^n$, it can be recovered with high probability from m random Gaussian (or Fourier) observations provided $m \gtrsim k \log(n)$ [11]. (we use the symbol $\gtrsim$ to say that there is an absolute constant C such that if $m \geq Ck \log(n)$, recovery is possible with high probability.) Similarly, if the signal is an $n \times n$ matrix with rank at most r, in the space $\mathcal{H} = \mathbb{R}^{n \times n}$, it can be recovered with high probability from $m \gtrsim rn$ random Gaussian observations [14].

In the following, the notion of low complexity is summarized by the fact that x is well approximated by an element of a so-called model set Σ, where $\Sigma \subset \mathcal{H}$

is low-dimensional according to a notion of dimension that will be specified. The considered notion of dimension will be defined in Section 3.1 and is related to the number of unknowns we need to estimate to characterize the signal. In the context of linear inverse problems with such low-dimensional models, the first objective is to obtain conditions on the linear operator A and the model set Σ that guarantee a possible recovery. From this perspective, the analysis of "low-complexity recovery" is an extension of classical analyses of sparse recovery or low-rank matrix recovery. A second objective, related to the field of compressed sensing, is dimension reduction, where the goal is to design a linear operator A (often with randomness) so that low-complexity recovery is possible, with an emphasis on allowing the dimension m of the observation to be small.

1.2 Decoders

As the ultimate task is to recover x from y, the analysis of the observation of x must be held together with the study of the methods used to recover x from y, which we call *decoders*. In this chapter, we consider the general class of decoders which consist in minimizing a regularizer under a data-fit constraint. We study estimates x^* of x of the form

$$x^* \in \operatorname*{argmin}_{z \in \mathcal{H}} f(z) \ s.t. \ \|Az - (Ax + e)\|_{\mathcal{F}} \leq \epsilon. \tag{2}$$

Formulation (2) covers many decoders proposed in the literature, even though other formulations exist (e.g., minimizing $\|Az - (Ax + e)\|_{\mathcal{F}}$ under a constraint on $f(z)$ or using a Lagrangian formulation). The study presented in this chapter does not require x^* to be the unique minimizer of (2). It must be noted that this formulation somehow emphasizes practical signal processing applications because an estimation ϵ of the observation noise energy $\|e\|_{\mathcal{F}}$ is often available (e.g., in photography the noise level can be estimated using aperture time and light conditions; in image processing more advanced techniques allow to estimate noise level from an image [37]).

The main parameter of the decoder is the regularizer f. Its role is to force the estimate belonging to the chosen model set. The form of the data-fit constraint ($\|\cdot\|_{\mathcal{F}}$) influences the types of noise that the decoder can robustly manage. This raises interesting questions that are, however, out of the scope of this chapter. The main qualities required for a decoder are (1) to provide *exact recovery of vectors $x \in \Sigma$ in the noiseless setting* and (2) to be *stable to observation noise and robust to modeling error*.

We emphasize the role of two classes of decoders: "ideal" decoders and convex decoders.

- Given a problem with a model set Σ, the *ideal decoder* corresponds to minimizing (2) using $f := \iota_\Sigma$, the characteristic function of Σ, i.e., $\iota_\Sigma(x) = 0$ if $x \in \Sigma$, $\iota_\Sigma(x) = \infty$ otherwise. This decoder is called ideal, as it enforces perfectly the

fact that the solution must belong to Σ (the prior on the unknown). Unfortunately, it is generally hard to calculate efficiently as the function to minimize is both non-convex and non-smooth. Consequently, we often use a heuristic for the minimization or turn to a convex proxy to this minimization.

- The decoder is said to be a *convex decoder* when f is convex. Such a decoder is often easier to compute as the minimization problem has no local minimum other than the global minima even if this does not guarantee that the minimization can be efficiently performed; see, e.g., tensor recovery problems [28]. State of the art shows that having some additional hypothesis on the linear operator A enables to guarantee stability and robustness of certain convex decoders for classical model sets Σ.

1.3 The RIP: A Tool for the Study of Signal Recovery

As we just saw, studying signal recovery amounts to studying the interactions between the model Σ, the regularization f, and the measurement operator A. We propose here to use a tool that enables us to separate the study of A with respect to Σ from the study of f with respect to Σ: the restricted isometry property (RIP). It is generally defined in our setting for a linear observation operator A on the so-called secant set $\Sigma - \Sigma := \{x - x' : x \in \Sigma, x' \in \Sigma\}$.

Definition 1 (RIP). The linear operator $A : \mathcal{H} \to \mathcal{F}$ satisfies the RIP on the secant set $\Sigma - \Sigma$ with constant δ if for all $x \in \Sigma - \Sigma$:

$$(1 - \delta)\|x\|_{\mathcal{H}}^2 \leq \|Ax\|_{\mathcal{F}}^2 \leq (1 + \delta)\|x\|_{\mathcal{H}}^2 \tag{3}$$

where $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{F}}$ are Euclidean norms on $\mathcal{H}$ and $\mathcal{F}$.

This property is a famous sufficient condition on A to guarantee the success of convex decoders (2) in the case of sparse and low-rank signal recovery for appropriately chosen regularization f [13, 15, 17, 21, 25, 34]. Intuitively, the RIP requires the operator A to preserve the distance between any two elements of Σ (see Figure 1). Moreover, a lower RIP is a necessary condition for the existence of stable and robust decoders: given A and Σ, if a stable and robust decoder exists then, up to a global rescaling, A satisfies a lower RIP on the secant set $\Sigma - \Sigma$ [8, 18].

For example, in the case of sparse recovery, it is possible to show two facts.

- Fact 1: Random Gaussian matrices of size $m \times n$ satisfy the RIP on the set of $2k$-sparse vectors (the secant set of the set of k-sparse vectors) with constant $\delta < 1$ with high probability, provided $m \gtrsim \delta^{-2} k \log(n)$.
- Fact 2: As soon as A satisfies this RIP with constant $\delta < 1/\sqrt{2}$, it is guaranteed that minimization (2) with $f(\cdot) = \|\cdot\|_1$, the ℓ^1-norm, yields stable and robust recovery of all k-sparse vectors [9].

We see that the study of recovery guaranteed in this case is separated in two steps: (1) a study of the behavior of the linear operator A with respect to the model set Σ

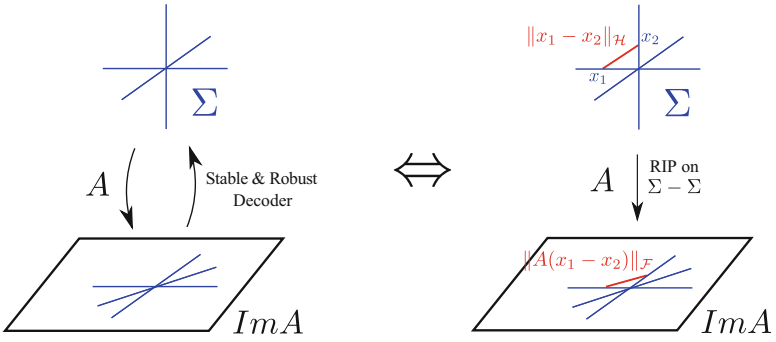
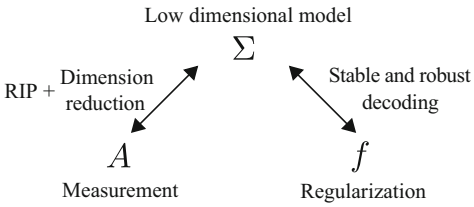


Fig. 1 A graphical representation of the equivalence between the existence of stable robust decoders and the RIP on the secant set. Operators satisfying the RIP approximately preserve distances between elements of Σ.

Fig. 2 Structure of the framework: The RIP framework allows to separate the study of dimension reduction and of decoding.



(in terms of RIP property) and (2) a study of the behavior of the regularizer f with respect to the model set Σ that has consequences for all operators satisfying a RIP with a small enough constant.

The framework presented in the following generalizes these features in order to manage not only the classical sparse recovery/low-rank recovery and related compressed sensing theory but much beyond to many sorts of low-dimensional model sets.

1.4 A General Compressed Sensing Framework

The remaining part of this chapter shows how it is possible to generalize the steps we just mentioned. The proposed framework[1] consists in answering the following questions (summarized in Figure 2):

- Low-dimensional model: when is Σ "low-dimensional"? (Section 2)
- Dimension reduction: given Σ, is there an operator A that satisfies the RIP on $\Sigma - \Sigma$? What level of dimension reduction can it achieve? (Section 3)

[1]This chapter gives a unified view of the latest developments in the area found in [33] and [38].

- What is a good regularizer? Given Σ and f, does a RIP of A on $\Sigma - \Sigma$ guarantee that f recovers the elements of Σ? (Section 4)

Section 5 mentions generalizations that were left out of the main of the chapter in order to keep the exposition accessible and discusses what challenges we face to go beyond this general compressed sensing framework in Hilbert spaces.

2 Low-Dimensional Models

We begin by precisely describing the low-dimensional models that will be considered in this chapter. We then focus on a model of *structured sparsity in levels*, which we will use as a running example to illustrate the different concepts used in this chapter.

2.1 Definition and Examples

The results presented in [33] show that one can always construct a linear operator A that satisfies the RIP on $\Sigma - \Sigma$ if its normalized secant set $\mathcal{S}(\Sigma)$ has a finite intrinsic dimension. The *normalized secant set* of Σ is defined as

$$\mathcal{S}(\Sigma) := \left\{ z = \frac{y}{\|y\|_{\mathcal{H}}} : y \in (\Sigma - \Sigma) \setminus \{\mathbf{0}\} \right\}.$$

We substitute $\mathcal{S}$ for $\mathcal{S}(\Sigma)$ hereafter to simplify notations. We illustrate in Figure 3 the RIP on the normalized secant set which is equivalent to the RIP on the secant set.

In this chapter, we measure the intrinsic dimension of $\mathcal{S}$ using the upper box-counting dimension, which is linked to the notion of covering number.

Definition 2 (Covering Number). Let $\alpha > 0$ and $\mathcal{S} \subset \mathcal{H}$. The covering number $N(\mathcal{S}, \alpha)$ of $\mathcal{S}$ is the minimum number of closed balls (with respect to the norm $\|\cdot\|_{\mathcal{H}}$) of radius α, with centers in $\mathcal{S}$, needed to cover $\mathcal{S}$.

The upper box-counting dimension is then defined as follows.

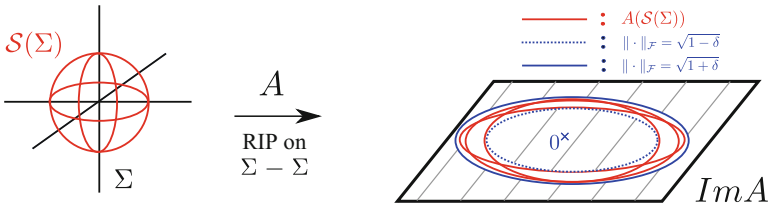


Fig. 3 A characterization of the RIP: the RIP on the normalized secant set. The image of the secant set must lie within a distance δ of the unit sphere.

Definition 3 (Upper Box-Counting Dimension). The upper box-counting dimension of $\mathcal{S}$ is

$$\operatorname{boxdim}(\mathcal{S}) := \limsup_{\alpha \to 0} \log[N(\mathcal{S}, \alpha)] / \log[1/\alpha].$$

Hence, as soon as $k > \operatorname{boxdim}(\mathcal{S})$, there exists a model-set dependent constant $\alpha_{\mathcal{S}} \in (0, 1/2)$ such that $N(\mathcal{S}, \alpha) \leq \alpha^{-k}$ for all $\alpha \leq \alpha_{\mathcal{S}}$. Further, if the covering number satisfies

$$N(\mathcal{S}, \alpha) \leq \left(\frac{C}{\alpha}\right)^k \tag{4}$$

then $\operatorname{boxdim}(\mathcal{S}) \leq k$.

We choose this definition of intrinsic dimension for two reasons. First, for many useful signal models – e.g., sparse vectors, low-rank matrices, and smooth manifolds – the upper box-counting dimension of the normalized secant set is known. The results presented in this chapter can thus be directly applied to these sets, without additional work. Second, one should be careful with the definition of intrinsic dimension used in an infinite-dimensional space. Indeed, for some definitions of dimension, there are examples where it is impossible to perform dimension reduction on vectors belonging to a set having a finite dimension (i.e., the set cannot be linearly and stably embedded in a finite-dimensional space [35, Chapter 6.1]). The upper box-counting dimension of the normalized secant set does not suffer from this issue.

In the following, we will say informally that a model Σ is *low-dimensional* if $\operatorname{boxdim}(\mathcal{S}(\Sigma))$ is small compared to the ambient dimension of the Hilbert space $\mathcal{H}$ (which may be infinite). In many examples, the dimension $\operatorname{boxdim}(\mathcal{S})$ is of the order of the number of parameters needed to describe elements of the model, as in the case of classical sparsity or low-rank matrices. For k-sparse vectors, the dimension of the normalized secant set $\mathcal{S}$ is of the order of k [25, Section C.2]. For $n \times n$ matrices of rank lower than r, the dimension of the normalized secant set $\mathcal{S}$ is of the order of rn.

2.2 *Structured Sparsity ...*

As a running example, we use a refinement of the notion of sparsity as a way to introduce the general framework: we consider a model of *structured sparsity in levels*.

We start by describing *structured sparsity*, a now classical generalization of the plain sparsity model. In many applications, signals are not only sparse but also clustered in groups of significant coefficients in a transformed domain (Fourier domain, Radon domain,...). Structured sparsity (also called group sparsity) is the assumption that the signal is supported on a few groups of coefficients [5, 24, 27].

Formally, we consider an orthonormal Hilbert basis $(e_i)_{i\in\mathbb{N}}$ of $\mathcal{H}$ and a finite collection G of non-overlapping finite *groups* of indices, i.e., subsets $g \subset \mathbb{N}$ with $|g| < \infty$ and $g \cap g' = \emptyset$ whenever $g \neq g'$. The restriction of the vector $x \in \mathcal{H}$ to the group g is $x_g := \sum_{i\in g}\langle x, e_i\rangle e_i$. A *group support* is a subset $T \subset G$, and the restriction of x to the group support T is $x_T := \sum_{g\in T} x_g$. The group support of $x \in \mathcal{H}$, denoted gsupp(x), is the smallest $T \subset G$ such that $x_T = x$. The size of the group support of x, denoted $|\text{gsupp}(x)|$, is the cardinality of gsupp(x) (to be distinguished from the number of non-zero coordinates in x).

Given an integer k, the *k-group-sparse model* is defined as

$$\Sigma_k := \{x \in \mathcal{H},\ |\text{gsupp}(x)| \leq k\}. \tag{5}$$

Let d be the size of the biggest group. We have the following covering of $\mathcal{S}(\Sigma_k)$:

$$N(\mathcal{S}(\Sigma_k), \alpha) \leq \left(\frac{C}{\alpha}\right)^{dk} \tag{6}$$

where C is a constant depending on d.

2.3 ...in Levels

Consider a collection of J orthogonal spaces $\mathcal{H}_j \subset \mathcal{H}$ each equipped with a k_j-group-sparse model Σ_j as defined in (5) (each with its Hilbert basis and its set G_j of groups). Since the subspaces are orthogonal, there is a natural isomorphism between their direct sum and their Cartesian product. It is simpler to work with the latter, and *structured sparsity in levels* is associated with the model (see Figure 4)

$$\Sigma := \left\{ x \in \mathcal{H}, x = \sum_{j=1}^{J} x_j, x_j \in \Sigma_{k_j} \right\}, \tag{7}$$

which is identified in the Cartesian product of the models $\Sigma_{k_1} \times \Sigma_{k_2} \times \ldots \times \Sigma_{k_j}$.

Two examples where this model is useful are medical imaging (MRI) and simultaneous signal and noise sparse modeling [2, 36, 39]:

- In MRI, the different levels where the signal is sparse are wavelet scales. MRI images are generally sparser at fine wavelet scales than large wavelet scales. This allows for more flexibility in the modeling of the signal than the simple sparsity model.

Fig. 4 A representation of structured sparsity in levels in $\mathcal{H}$. A structured sparsity in level model is formed by different structured sparsity models in orthogonal subspaces.

- Simultaneous signal and noise sparse modeling is a convenient setting for the separation of a signal sparse in some domain from noise that is sparse in another domain. An observed signal y is modeled as the superimposition of two components, $y = A_1x_1 + A_2x_2$, where A_1x_1 is the signal of interest, x_1 lives in the (structured) sparse model Σ_{k_1}, A_2x_2 is noise, and x_2 lives in the (structured) sparse model Σ_{k_2}. This model is also related to the separation of transients from stationary parts in audio or for the decomposition of images into cartoon and texture [32]. As $y = [A_1\ A_2]x$ with $x = [x_1^T,\ x_2^T]^T$, this corresponds to a two-level (structured) sparse model for x.

For structured sparsity in levels, we have [38]

$$\begin{aligned} N(\mathcal{S}, \alpha) &\leq N(\mathcal{S}(\Sigma_{k_1}, \alpha) \times \ldots \times N(\mathcal{S}(\Sigma_{k_j}), \alpha) \\ &\leq \left(\tfrac{C_1}{\alpha}\right)^{d_1 k_1} \times \ldots \times \left(\tfrac{C_J}{\alpha}\right)^{d_J k_J} \end{aligned} \tag{8}$$

where C_j are constants that are of the order of the dimension of each level times the maximum size of groups d_j in level j. Hence, up to log factors, the upper box-counting dimension of $\mathcal{S}$ in this case is of the order of $\sum d_j k_j$.

3 Dimension Reduction with Random Linear Operators

Now that we have defined the notion of dimension of a model Σ that we work with, and the desirable RIP property of a linear operator A, the remaining question is: how to construct a *dimension-reducing* linear operator $A : \mathcal{H} \to \mathbb{R}^m$ that satisfies the RIP on $\Sigma - \Sigma$?

Consider an MRI-like scenario with a sparsity in level signal model $\Sigma_{k_1} \times \ldots \times \Sigma_{k_J}$ in a wavelet basis. The fact that the signals in Σ have a support restricted to the first J wavelet scales implies that their energy decreases at high frequencies. Intuitively, it thus seems unnecessary to probe very high frequencies in the measurement process for this type of signals [1]. A good approximation of the signals can be obtained by probing all frequencies up to a certain bandlimit B. This process corresponds to a projection from the infinite-dimensional space $\mathcal{H}$ to a finite-dimensional space of size B. However, the dimension B, though finite, might still be reduced. Indeed, the signals are not just concentrated in the first J wavelet scales, but they are also sparse in levels. A dimension-reducing step can thus be envisioned after the projection onto the first B Fourier coefficients with, e.g., a random Gaussian matrix. Ideally the final dimension m should satisfy, up to log factors, $m = O\left(\sum_{j=1}^{J} k_j\right)$ (of the order of the number of parameters describing the model). Intuition thus suggests to build the operator A in two steps: a projection onto a finite (but high)-dimensional space followed by a multiplication with a random matrix. In fact, the authors of [33] present such a construction in the general setting that first projects the signal onto a subspace $H \subset \mathcal{H}$ of finite (but potentially large) dimension and then reduces the dimension using a random linear operator on H (see Figure 5).

	Projection		Reduction	
$\Sigma \subset \mathcal{H}$	$\longrightarrow$	Linear space $H \subset \mathcal{H}$	$\longrightarrow$	Linear space $= \mathbb{R}^m$
Low-dimension		Finite dimension		Low dimension
$\text{boxdim}(\mathcal{S}(\Sigma)) < \infty$		$\dim(H) = d < \infty$		$m \ll d$

Fig. 5 Strategy for dimension reduction: We aim at reducing the dimension of vectors belonging to Σ leaving an infinite-dimensional space $\mathcal{H}$.

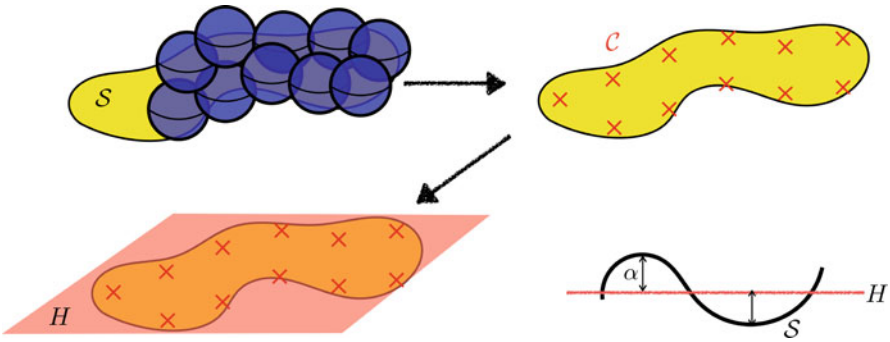


Fig. 6 Construction of H. *Top left*: cover of $\mathcal{S}$ with $N(\mathcal{S}, \alpha)$ balls of radius α. *Top right*: the centers of the balls, indicated by the red crosses, form an α-cover, denoted by $\mathcal{C}$, for $\mathcal{S}$. *Bottom left*: H is defined as the linear span of the vectors in $\mathcal{C}$. *Bottom right*: H approximates $\mathcal{S}$ with precision α.

3.1 Projection on a Finite-Dimensional Subspace

Assuming that $\text{boxdim}(\mathcal{S})$ is finite, we will see that, given $0 < \alpha < 1$, there always exists a finite-dimensional subspace $H \subset \mathcal{H}$ such that

$$(1-\alpha)\|x\|_{\mathcal{H}} \leq \|P_H x\|_{\mathcal{H}} \leq \|x\|_{\mathcal{H}} \tag{9}$$

for all $x \in \Sigma - \Sigma$, where P_H denotes the orthogonal projection onto H.

In the example of Fourier sampling of signal sparse in a Haar basis, it is possible to directly exhibit such a projection P_H by sampling low Fourier frequencies. However, one can generally construct H as follows. First, build an α-cover of the normalized secant set $\mathcal{S}$. As $\text{boxdim}(\mathcal{S})$ is finite, $N(\alpha, \mathcal{S}) < +\infty$ balls are sufficient to build this cover. Let now $\mathcal{C}$ be the set containing the center of these balls. It is then sufficient to take $H = \text{span}\ \mathcal{C}$; see Figure 6 and [33]. We remark that in the worst case, the cardinality of $\mathcal{C}$ is exponential in the dimension of $\mathcal{S}$; hence, H can have a dimension of the order of $e^{c(\alpha)\times\text{boxdim}(\mathcal{S})}$. Yet the important message to take away at this stage is that:

If the normalized secant set $\mathcal{S}(\Sigma)$ has a finite upper box-counting dimension, then there exists a finite-dimensional subspace $H \subset \mathcal{H}$ that approximates all vectors in $\mathcal{S}$ with precision α.

In the next section, we describe how to further reduce the dimension to $m = O(\text{boxdim}(\mathcal{S}))$ after this first projection.

3.2 Dimension Reduction Step

After the projection onto the finite-dimensional space H of the previous section, the goal is now to reduce the dimension down to $O(\text{boxdim}(\mathcal{S}))$. As most compressive sensing techniques use random observations to reduce the dimension, it seems natural to follow this route.

Denote by d the dimension of the subspace H and $(e_1, \ldots, e_d)$ an arbitrary orthonormal basis of H. By abuse of notation, identify the projection onto H with the linear operator $P_H : \mathcal{H} \to \mathbb{R}^d$ that returns the coordinates of the orthogonal projection onto H in the basis $(e_1, \ldots, e_d)$. The idea is now to compose P_H with a random matrix $M \in \mathbb{R}^{m \times d}$ to build $A : \mathcal{H} \to \mathbb{R}^m$, i.e., $A = MP_H$. Ideally, we would like A to satisfy the RIP, and $m \approx O(\text{boxdim}(\mathcal{S}))$: a number of measurements of the order of the dimension of the model. In this case, we would be assured that the ideal decoder is stable and robust and that the reduction of dimension is close to optimal.

3.2.1 Randomized Dimension Reduction

To exhibit a linear operator A satisfying the RIP with constant δ, one can first identify a finite-dimensional subspace $H \subset \mathcal{H}$ such that (9) holds with α small enough and then *build* a random $M : \mathbb{R}^d \to \mathbb{R}^m$ satisfying a RIP with small enough constant δ'. Sometimes one is directly provided with a random linear operator from $\mathcal{H} \to \mathbb{R}^m$ and needs to *check* whether the RIP holds with high probability. The approach described in [33] makes it possible to handle both cases. With a slight abuse of notation, in this subsection, $\mathcal{H}$ stands either for the original Hilbert space (case of a given random operator) or for $\mathbb{R}^d$ (two-step construction considered above).

Consider M a *random* linear operator from $\mathcal{H}$ to $\mathbb{R}^m$. An example results from the independent draw of m identically distributed random vectors $a_i \in \mathcal{H}$, so that for $x \in \mathcal{H}$, $Mx := (\langle a_i, x \rangle)_{i=1}^m$. A convenient way to help M satisfy the RIP is to choose its probability distribution so that, for any vector $x \in \mathcal{H}$,

$$\mathbb{E}_M \|Mx\|_2^2 = \|x\|_{\mathcal{H}}^2. \tag{10}$$

With the above *isotropy assumption*, a draw M of the random linear operator satisfies the RIP on $\Sigma - \Sigma$ if, and only if,

$$\left| \|Mx\|_2^2 - \mathbb{E}_{\tilde{M}} \|\tilde{M}x\|_2^2 \right| \leq \delta \|x\|_{\mathcal{H}}^2, \tag{11}$$

for all $x \in \Sigma - \Sigma$, where we emphasize with the $\tilde{M}$ notation that the expectation $\mathbb{E}$ is with respect to a linear operator with the same *distribution* as the one from which the particular M is drawn. As discussed in Section 5, even without the isotropy assumption (10), one can establish dimension reduction results using (11) as a generalized definition of the RIP [33].

To prove that M satisfies the RIP, the authors in [33] require it to satisfy two concentration inequalities. Define

$$\begin{aligned} h_M \colon \mathcal{H} &\longrightarrow \mathbb{R} \\ x &\longmapsto \|Mx\|_2^2 - \|x\|_{\mathcal{H}}^2. \end{aligned}$$

The assumption is that there exist two constants $c_1, c_2 \in (0, \infty]$ such that for any fixed $y, z \in \mathcal{S}(\Sigma) \cup \{0\}$,

$$\mathbb{P}_M \left\{ |h_M(y) - h_M(z)| \geq \lambda \, \|y - z\|_{\mathcal{H}} \right\} \leq 2e^{-c_1 m \lambda^2}, \quad \text{for } 0 \leq \lambda \leq c_2/c_1 \tag{12}$$

$$\mathbb{P}_M \left\{ |h_M(y) - h_M(z)| \geq \lambda \, \|y - z\|_{\mathcal{H}} \right\} \leq 2e^{-c_2 m \lambda}, \quad \text{for } \lambda \geq c_2/c_1. \tag{13}$$

By taking $z = 0$ in (12) and (13), we see that the above properties imply that, for any *given* vector in the normalized secant set, $y \in \mathcal{S}$, with high probability on the draw of M, the quantity $\|My\|_2^2$ stays close to its expected value $\mathbb{E}_M \|My\|_2^2 = \|y\|_{\mathcal{H}}^2 = 1$. Proving that the RIP holds consists in showing that, with high probability on the draw of M, this property actually holds uniformly for *all* vectors in $\mathcal{S}$, not just for any fixed vector $y \in \mathcal{S}$. Among other properties, this generalization to the entire set $\mathcal{S}$ is proved by using the fact that for any fixed $y, z \in \mathcal{S}$, if $\|y - z\|_{\mathcal{H}}$ is small then the difference between $\|My\|_2^2 - \|Mz\|_2^2$ and $\mathbb{E}_{\tilde{M}} \|\tilde{M}y\|_2^2 - \mathbb{E}_{\tilde{M}} \|\tilde{M}z\|_2^2$ is also small with high probability.

These concentration inequalities together with the finite dimension of $\mathcal{S}$ suffice to conclude on a sufficient number of measurements for M to satisfy the RIP [33, Theorem II.2].

Theorem 1. *Let $M : \mathcal{H} \to \mathbb{R}^m$ be a random linear map that satisfies (12) and (13). Assume that* $\operatorname{boxdim}(\mathcal{S}) < s$ *(there exists $0 < \alpha_{\mathcal{S}} < \frac{1}{2}$ such that $N(\mathcal{S}, \alpha) \leq \alpha^{-s}$ for all $0 < \alpha < \alpha_{\mathcal{S}}$).*

Then for any $\xi, \delta_0 \in (0, 1)$, M satisfies the RIP on $\Sigma - \Sigma$ with constant $\delta \leq \delta_0$ with probability of at least $1 - \xi$ provided that

$$m \geq \frac{1}{\delta_0^2} \frac{C}{\min(c_1, c_2)} \max \left\{ s \log \left(\frac{1}{\alpha_{\mathcal{S}}} \right), \log \left(\frac{6}{\xi} \right) \right\}, \tag{14}$$

where $C > 0$ is an absolute constant.

This theorem states that if the random operator M satisfies appropriate concentration inequalities and the set $\mathcal{S}$ has finite upper box-counting dimension, then reducing the dimension of the vectors in Σ is possible (recall that these vectors

possibly live in an infinite-dimensional space). A number of measurements m of the order of the dimension of the secant set $\mathcal{S}$ (only) are sufficient to be able to recover elements of Σ, whatever the ambient dimension of $\mathcal{H}$ (which can be infinite). We remark that the sufficient number of measurements grows as the RIP constant decreases (the closer A is to an isometry for elements in $\mathcal{S}$). In particular, the typical $\log n$ factor appearing in standard results for compressed sensing of k-sparse vectors in $\mathcal{H} = \mathbb{R}^n$ is in fact related to $\alpha_{\mathcal{S}}$ rather than the ambient dimension. Related results, independent of the ambient dimension, have been achieved for manifold embedding [20, 23].

For a fixed dimension of $\mathcal{S}$, if we wish to ensure an arbitrarily small probability that the RIP fails to hold, $\xi \leq 6\,(\alpha_{\mathcal{S}})^s$, then the number of measurements m also grows as ξ approaches zero. Vice versa, as the ratio between m and its minimum value $m_0 = \frac{1}{\delta^2} \frac{C}{\min(c_1, c_2)}\, s \log\left(\frac{1}{\alpha_{\mathcal{S}}}\right)$ grows, the RIP holds with probability exponentially close to 1.

Remark 1. The sufficient condition $m \geq m_0$ is *not* necessary. There are actually pathological sets Σ whose normalized secant set has an *infinite upper box-counting dimension* and for which some operators $M : \mathcal{H} \to \mathbb{R}^m$ *with only* $m = 1$ measurement satisfy the RIP [33].

3.2.2 Some Examples

When given a random linear operator $A : \mathcal{H} \to \mathbb{R}^m$, one can leverage the above result to check whether A satisfies the RIP with high probability. Alternatively, one can *construct* such an operator by pursuing the strategy described at the beginning of this section. We now need to choose the matrix $M \in \mathbb{R}^{m \times d}$. Examples of matrices $M \in \mathbb{R}^{m \times d}$ such that the operator $A = MP_H$ satisfies (12) and (13) are:

- matrices with independent random Gaussian entries with mean 0 and variance $1/m$;
- matrices whose entries are independent random Bernoulli variables $\pm 1/\sqrt{m}$;
- matrices whose rows are independently drawn from the Euclidean sphere of radius $\sqrt{d/m}$ in $\mathbb{R}^d$ using the uniform distribution.

If M is one of the above matrices (or more generally a matrix with independent subgaussian rows), considering the orthogonally projected model set $\Sigma' = P_H \Sigma$, its normalized secant set $\mathcal{S}' = \mathcal{S}(\Sigma')$, and $s > \text{boxdim}(\mathcal{S}') = \text{boxdim}(\mathcal{S})$, we have [33] M satisfies the RIP on $\Sigma' - \Sigma'$ with constant $\delta' < \delta_0$ with high probability provided

$$m \geq \frac{C'}{\delta_0^2} \max\left\{ s \log\left(\frac{1}{\alpha_{\mathcal{S}'}}\right), \log\left(\frac{6}{\xi}\right) \right\}, \tag{15}$$

where C' is a constant that depends on the distribution of M.

3.3 Summary

To summarize, a generic strategy (a way to implement the strategy in Figure 5) to build a compressive sensing measurement operator for a set Σ that has a normalized secant set $\mathcal{S}$ of finite upper box-counting dimension is:

1. Find a (potentially high-dimensional) finite-dimensional space H whose orthogonal projection operator satisfies (9). A generic construction of such a space is presented in Section 3.1.
2. Compose this projection operator with a random projection operator M (a random matrix) such that (12) and (13) hold.

Now that we have described how we can build operators preserving low-complexity models, we can turn to the study of the performance of methods used to recover x from y.

4 Performance of Regularizers for the Recovery of Low-Dimensional models

As they satisfy the RIP, the linear operators A built with the technique just described in Section 3 preserve the low-dimensional model Σ in the sense that stable reconstruction of vectors from Σ is possible with the so-called ideal decoder. Yet, this decoder is often intractable in practice as it involves possibly non-convex and/or non-smooth optimization. We now turn to general decoders, with an emphasis on convex decoders: minimization algorithms are well known for such decoders, and they are often possible to implement with off-the-shelf algorithms, as in the classical cases of basis pursuit (ℓ^1-norm minimization) or nuclear norm minimization.

4.1 Convex Decoders and Atomic Norms

In the framework of minimization (2), it is interesting to consider a particular class of convex functions: atomic norms with atoms included in the model set Σ [17]. Considering a set $\mathcal{A} \subset \mathcal{H}$, commonly called the set of *atoms*, the corresponding *atomic "norm"* is built using the convex hull of $\mathcal{A}$.

Definition 4 (Convex Hull). The convex hull of a set $\mathcal{A}$ is

$$\operatorname{conv}(\mathcal{A}) := \left\{ x = \sum c_i a_i : a_i \in \mathcal{A}, c_i \in \mathbb{R}_+, \sum c_i = 1 \right\} \tag{16}$$

Definition 5 (Atomic Norm). The atomic "norm" induced by the set $\mathcal{A}$ is defined as

$$\|x\|_{\mathcal{A}} := \inf \{ t \in \mathbb{R}_+ : x \in t \cdot \overline{\operatorname{conv}}(\mathcal{A}) \} \tag{17}$$

where $\overline{\text{conv}}(\mathcal{A})$ is the closure of $\text{conv}(\mathcal{A})$ in $\mathcal{H}$. The function $\|x\|_{\mathcal{A}}$ is a convex gauge that is not always a norm. It is a norm if $\mathcal{A}$ is symmetrical and bounded. We will keep the term atomic norm in the general case as an abuse of notation. This norm is finite only on the set

$$\mathcal{E}(\mathcal{A}) := \mathbb{R}_+ \cdot \overline{\text{conv}}(\mathcal{A}) = \{x = t \cdot y, t \in \mathbb{R}_+, y \in \overline{\text{conv}}(\mathcal{A})\} \subset \mathcal{H}. \tag{18}$$

It can be extended to $\mathcal{H}$ by setting $\|x\|_{\mathcal{A}} := +\infty$ if $x \notin \mathcal{E}(\mathcal{A})$.

Atoms are often *normalized*: a vector u is *normalized* if $\|u\|_{\mathcal{H}} = 1$.

Remark 2. Atomic norms are interesting because given any convex regularization function, it is always possible to find an atomic norm that performs noiseless recovery better (in the sense that it permits recovery for more measurement operators A [38]).

4.1.1 Classical Examples of Atomic Norms

As pointed out in [17], many well-known norms used for low-complexity recovery are atomic norms:

- ℓ^1-norm in $\mathbb{R}^n$: $\mathcal{A}$ is the set of canonical orthonormal basis vectors multiplied by a real scalar with modulus 1, i.e., the normalized 1-sparse vectors.
- Nuclear norm: $\mathcal{A}$ is the set of normalized rank-one matrices.
- Gauge generated by a finite polytope: $\mathcal{A}$ is composed of the vertices of a polytope.
- Spectral norm: $\mathcal{A}$ is the set of normalized orthogonal matrices.

4.1.2 Group Norms in Levels

For our running example, the model $\Sigma = \Sigma_1 \times \ldots \Sigma_J$ associated with structured sparsity in levels, we consider a similar class of atomic norms: the group norms in levels.

Given the subspace $\mathcal{H}_j$ associated with the jth level, $\mathcal{S}_j(1) \subset \mathcal{H}_j$ its unit sphere, G_j its set of groups, and $\Sigma_{1,j}$ the associated *one*-group-sparse model, consider the collection of atoms of the jth level

$$\mathcal{A}_j := \Sigma_{1,j} \cap \mathcal{S}_j(1). \tag{19}$$

The corresponding atomic norm is associated with the finite-dimensional space

$$\mathcal{E}(\mathcal{A}_j) = \text{span}(\{e_i\}_{i \in \cup_{g \in G_j}})$$

and simply given by

$$\|x\|_{\mathcal{A}_j} = \begin{cases} \sum_{g\in G} \|x_g\|_{\mathcal{H}}, & x \in \mathcal{E}(\mathcal{A}); \\ +\infty, & x \notin \mathcal{E}(\mathcal{A}) \end{cases} \tag{20}$$

The norm $\|x\|_{\mathcal{A}_j}$ is called a *group norm*, a *structured norm*, or a *mixed* $\ell^1 - \ell^2$*-norm* [41].

A natural regularizer for the *structured sparsity in level* model is defined as follows in $\mathcal{H}_1 \times \ldots \times \mathcal{H}_J$:

$$f_w : (x_1, \ldots x_J) \mapsto w_1 \|x_1\|_{\mathcal{A}_1} + \ldots + w_J \|x_J\|_{\mathcal{A}_J} \tag{21}$$

with weights $w_j > 0$. We will show in the next sections that setting appropriately the weights in each level can yield recovery guarantees of various strengths.

4.1.3 Atomic Norm Associated with a Union of Subspace Model

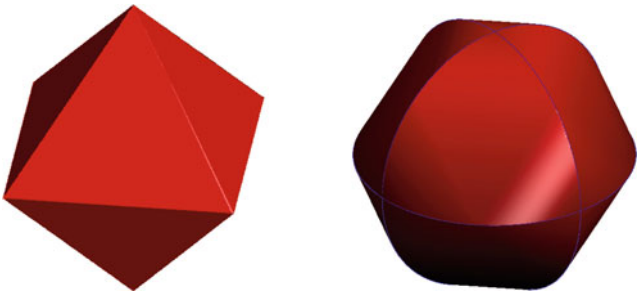
Many classical model sets Σ (the set of sparse vectors, the set of low-rank matrices, etc.) are homogeneous: if $x \in \Sigma$ then $\alpha x \in \Sigma$ for any scalar. As such they are (finite or infinite) unions of subspaces. Given any union of subspaces $\Sigma \subset \mathcal{H}$, the norm associated to its *normalized atoms*

$$\mathcal{A}(\Sigma) := \Sigma \cap \mathcal{S}(1) \tag{22}$$

will be of particular interest for the RIP analysis described in the next sections. As a shorthand notation, we define

$$\|\cdot\|_{\Sigma} := \|\cdot\|_{\Sigma \cap \mathcal{S}(1)}. \tag{23}$$

This norm is sometimes useful as a regularizer to perform recovery (i.e., by choosing $f(z) = \|z\|_{\Sigma}$ in minimization (2)). For the particular case where Σ is the set of k-sparse vectors, $\|\cdot\|_{\Sigma}$ is known as the k-support norm [3]. It is known to yield stable recovery guarantees for certain k-sparse vectors [3]; however, it has been shown that these results cannot be made uniform for all k-sparse vectors (and consequently similar negative results hold for structured sparsity in levels) [38]. We show in Figure 7 a representation of the ℓ^1-norm and of the k-support norm $\|\cdot\|_{\Sigma}$ for $k = 2$ in 3D ($\mathcal{H} = \mathbb{R}^3$), which are two atomic norms induced by normalized atoms included in the model set Σ.



(a) $\|\cdot\|_1$ (b) $\|\cdot\|_\Sigma$

Fig. 7 The unit ball of $\|\cdot\|_1$ (left) and the unit ball of $\|\cdot\|_\Sigma$ (k-support norm) for $\Sigma = \Sigma_2$ the set of two-sparse vectors in 3D (right).

4.2 Stable and Robust Recovery of Unions of Subspaces

The main result from [38] states that the stability of *any* decoder of the form (2) is guaranteed provided the linear operator A satisfies a RIP on the secant set $\Sigma - \Sigma$ with a constant $\delta < \delta_\Sigma(f)$ holds, where $\delta_\Sigma(f)$ is a constant that depends only on the regularizer f and the model set Σ (we give and discuss the definition of $\delta_\Sigma(f)$ in Section 4.3).

4.2.1 Stable Recovery in the Presence of Noise

Elements of the model can be stably recovered [38, Theorem 1.2]:

Theorem 2 (RIP Condition for Stable Recovery of a Union of Subspaces). *Assume that Σ is a union of subspaces. Then, for any continuous linear operator A on $\mathcal{H}$ that satisfies the RIP on the secant set $\Sigma - \Sigma$ with constant $\delta < \delta_\Sigma(f)$, we have for all $x \in \Sigma$, $e \in \mathcal{F}$ such that $\|e\|_\mathcal{F} \leq \epsilon$ (recall that ϵ is an* estimation *of the noise level used as a parameter of the decoder), with x^* the result of minimization* (2)*,*

$$\|x^* - x\|_\mathcal{H} \leq C_\Sigma(f, \delta) \cdot (\|e\|_\mathcal{F} + \epsilon) \tag{24}$$

where $C_\Sigma(f, \delta) < +\infty$.

We refer the reader to [38, Theorem 1.2] for an explicit expression of $C_\Sigma(f, \delta)$. It is increasing with respect to the RIP constant δ : the worse the RIP constant is, the worse the stability constant is (see, e.g., its expression for structured sparsity in levels in Theorem 4).

4.2.2 Robustness to Modeling Error

Regarding robustness to modeling error, generic results often use the so-called A-norm [8] (not to be confused with the atomic norm, here the A refers to the measurement operator) as an intermediate tool to measure the distance from a vector x to the model set Σ. Given a constant C, the A-norm is defined by

$$\|\cdot\|_{A,C} := C \cdot \|A \cdot\|_{\mathcal{F}} + \|\cdot\|_{\mathcal{H}}. \tag{25}$$

It is more convenient to express robustness results with respect to a norm that does not depend on the measurement operator A. We provide here a robustness result where the modeling error with respect to the regularizer f is used (this is more in-line with the classical literature for ℓ^1 minimization of nuclear norm minimization). Consider the (symmetrized) distance with respect to f:

$$d_f(x, \Sigma) = \inf_{\tilde{x} \in \Sigma} \frac{f(x - \tilde{x}) + f(\tilde{x} - x)}{2}. \tag{26}$$

When f is a positively homogeneous, nonnegative, and convex regularizer that bounds the A-norm, robustness with respect to d_f also generally holds [38, Theorem 3.2]:

Theorem 3. *Let Σ be union of subspaces. Let f be positively homogeneous, nonnegative, and convex with $f(x) < +\infty$ for $x \in \Sigma$. Consider a continuous linear operator A satisfying the RIP on $\Sigma - \Sigma$ with constant $\delta < \delta_\Sigma(f)$ and a noise level $\|e\|_{\mathcal{F}} \leq \epsilon$. Denote C_Σ the constant from Theorem 2, and assume that for all $u \in \mathcal{H}$, $\|u\|_{A,C_\Sigma} \leq C_{f,A,\Sigma} \cdot f(u)$ for some $C_{f,A,\Sigma} < \infty$. Then, for all $x \in \mathcal{H}$, $e \in \mathcal{F}$, such that $\|e\|_{\mathcal{H}} \leq \eta \leq \epsilon$, any minimizer x^* of (2) satisfies*

$$\|x^* - x\|_{\mathcal{H}} \leq C_\Sigma \cdot (\|e\|_{\mathcal{F}} + \epsilon) + 2C_{f,A,\Sigma} \cdot d_f(x_0, \Sigma). \tag{27}$$

Remark 3. To apply this theorem, we need $C_{f,A,\Sigma} < \infty$. This is the case for most classical examples (sparse recovery with ℓ^1-norm, low-rank matrix recovery with the nuclear norm). It is also true for the case where f being a convex gauge induced by a bounded closed convex set containing 0 and $\mathcal{H}$ is of finite dimension.

Remark 4. Both Theorems 2 and 3 can be extended to the case where Σ is a *cone* instead of a union of subspaces, with a definition of $\delta_\Sigma(f)$ adapted compared to the one given later in Section 4.3 (see Section 5).

4.2.3 Example: The Case of Sparsity in Levels

Consider the model set Σ corresponding to our running example of structured sparsity in levels, and choose as a regularizer the weighted atomic norm $f_w(\cdot)$ defined in (21). One can show [38, Theorem 4.1] that $\delta_\Sigma(f_w) \geq \frac{1}{\sqrt{2}}$ for $J = 1$ and

$$\delta_\Sigma(f_w) \geq \frac{1}{\sqrt{2 + J\kappa_w^2}}$$

for $J \geq 2$, where $\kappa_w := \max(w_j\sqrt{k_j})/\min(w_j\sqrt{k_j})$. In particular, for the particular weights $w_j = 1/\sqrt{k_j}$, we have $\delta_\Sigma(f_w) \geq \frac{1}{\sqrt{2+J}}$ for $J \geq 2$.

In comparison, Ayaz et al. [4] gave a uniform recovery result with the mixed $\ell^1 - \ell^2$-norm for structured compressed sensing under a RIP hypothesis. They showed that a RIP constant $\delta < \sqrt{2} - 1$ for vectors in the secant set guarantees the recovery of vectors from the model. The above result shows that the RIP constant of Ayaz et al. can be improved to $\frac{1}{\sqrt{2}}$. In [2], a model of sparsity in levels was introduced: it is in fact a structured sparsity in level model with classical sparsity (each group is reduced to a single coordinate) in each level. In [6], Bastounis et al. showed that when the model Σ is sparse in levels and $f(\cdot) = \sum_j \|\cdot\|_{\mathcal{A}_j} = \|\cdot\|_1$ (i.e., with weights $w_j = 1$, in this case, $\kappa_w^2 = \kappa_1^2$ is the maximum ratio of sparsity between levels), the RIP with constant $\delta = 1/\sqrt{J(\kappa_1 + 0.25)^2 + 1)}$ on $\Sigma - \Sigma$ guarantees recovery. This constant is improved to the constant $\delta_\Sigma(f_w) \geq 1/\sqrt{2 + J}$ when weighting the norm of each level with $w_j = 1/\sqrt{k_j}$. The above result further extends the work of Bastounis et al. to general structured sparsity. The following theorem [38, Theorem 4.3] summarizes the result with this optimal weighting:

Theorem 4. *Let Σ be the model set associated with structured sparsity in levels, and consider $f = f_w$ as a regularizer, with the adapted weights $w_j = 1/\sqrt{k_j}$. Suppose the continuous linear operator A satisfies the RIP with constant $\delta < \delta_\Sigma(f)$ on the secant set $\Sigma - \Sigma$. Then for all $x \in \mathcal{H}$, $e \in \mathcal{F}$ such that $\|e\|_{\mathcal{F}} \leq \epsilon$, and x^* the result of minimization (2), we have*

$$\|x^* - x\|_{\mathcal{H}} \leq C_\Sigma(f, \delta)(\|e\|_{\mathcal{F}} + \epsilon) + D_\Sigma(f, \delta) \cdot d_f(x, \Sigma) \tag{28}$$

where :

- *For $J = 1$, $\delta_0 = \frac{1}{\sqrt{2}}$, $C_\Sigma(f, \delta) \leq \frac{2\sqrt{1+\delta}}{1-\delta\sqrt{2}}$ and $D_\Sigma(f, \delta) = 2(1 + \sqrt{1+\delta}C_\Sigma(f, \delta))/\sqrt{k}$.*
- *For $J \geq 2$, $\delta_0 = \sqrt{\frac{1}{2+J}}$, $C_\Sigma(f, \delta) \leq \frac{(1+\sqrt{1+J})\sqrt{1+\delta}}{1-\delta\sqrt{2+J}}$ and $D_\Sigma(f, \delta) = 2\sqrt{2}(1 + \sqrt{1+\delta}C_\Sigma(f, \delta))$.*

This result recovers classical guarantees with ℓ^1 minimization for sparse recovery. Since $\delta_\Sigma(f) \geq 1/\sqrt{2 + J}$ for $J \geq 2$, combining Theorem 4 for $\delta < 1/\sqrt{2 + J}$ with results from Section 3 yields [38] that

$$m \geq O\left(J \sum_{j=1}^{J} \left(k_j d_j + k_j \log\left(\frac{3e|G_j|}{k_j}\right)\right)\right).$$

subgaussian measurements are sufficient to guarantee stable and robust recovery with f_w, where $w_j = 1/\sqrt{k_j}$.

Remark 5. The factor J might seem pessimistic, and we attribute its presence to the generality of the result. Should the structure of the observation matrix A be taken into account, better results can be achieved. In fact, if A is a block diagonal matrix where each block A_j has size $m_j \times n_j$, uniform recovery guarantees with the ℓ^1-norm hold if and only if uniform recovery holds on each block: this is possible as soon as each block A_j of A satisfies the RIP with some constant $\delta_j < \frac{1}{\sqrt{2}}$ on $\Sigma_j - \Sigma_j$, which is in turn exactly equivalent to the RIP with constant $\delta < \frac{1}{\sqrt{2}}$ on $\Sigma - \Sigma$.

Remark 6. To make sense of Theorem 4 in the *infinite-dimensional setting*, the domain where the regularizer f is finite must be extended outside of $\mathcal{E}(\Sigma)$ while keeping a finite constant $D_\Sigma(f, \delta)$. This can be done on a case-by-case basis when properties of A and f allow to conclude. For example, as Adcock and Hansen in [2], consider the following setting: $\mathcal{H} = \ell^2(\mathbb{N})$ with Hilbert basis $(e_i)_{i=1,+\infty}$. Consider Σ a sparsity in level model in $(e_1, .., e_N)$. Let $f = \|\cdot\|_1$. Then f is an extension of the definition of f_w in $\mathcal{E}(\Sigma)$ to the whole space $\mathcal{H}$ (with $w_j = 1$ for all j). In [2], the measurement operator A is a collection of (Fourier) measurements that have *a strong balancing property*. The important fact here is that this property requires $\|A^H A\|_\infty \leq C'$ where $\|\cdot\|_\infty$ is the maximum of the ℓ^∞-norms of the coefficients of $A^H A$ (where A^H is the Hermitian conjugate of A). With such a hypothesis, for any $u \in \mathcal{H}$, we have $\|Au\|_2^2 = |\langle u, A^H Au\rangle| \leq \|A^H Au\|_\infty \|u\|_1 \leq \|A^H A\|_\infty \|u\|_1 \|u\|_1 \leq C' \|u\|_1^2$. Thus in this case the A-norm is bounded by the ℓ^1-norm: $\|\cdot\|_{A,C} \leq (1 + C\sqrt{C'})\|u\|_1$.

4.3 Definition and Calculation of $\delta_\Sigma(f)$

When Σ is a union of subspaces, the sufficient RIP constant for recovery of elements of Σ with f is defined as

$$\delta_\Sigma(f) := \inf_{z \in \mathcal{T}_f(\Sigma)\setminus\{0\}} \sup_{x \in \Sigma} \delta_\Sigma(x, z). \tag{29}$$

where

$$\delta_\Sigma(x, z) := \frac{-\mathcal{R}e\langle x, z\rangle}{\|x\|_\mathcal{H} \sqrt{\|x + z\|_\Sigma^2 - \|x\|_\mathcal{H}^2 - 2\mathcal{R}e\langle x, z\rangle}}. \tag{30}$$

And $\mathcal{T}_f(\Sigma)$ is the set of descent vectors of f at points of Σ:

$$\mathcal{T}_f(\Sigma) := \{z \in \mathcal{H} : \exists x \in \Sigma \ / \ f(x + z) \leq f(x)\} \tag{31}$$

It is important to note that the constant $\delta_\Sigma(f)$ only depends on the geometry of Σ and f. This constant measures the quality of f as regularizer to recover elements of Σ under a RIP assumption: the larger the $\delta_\Sigma(f)$, the weaker the assumption on the linear operator A to ensure stable and robust recovery in Theorem 3.

To obtain concrete results, one needs to lower bound the above expression. As the supremum in the expression of $\delta_\Sigma(f)$ is a priori hard to compute explicitly, for $z \in \mathcal{T}_f(\Sigma)$, one can intuitively seek an element $x \in \Sigma$ that maximizes the correlation with $-z$ (i.e., such that $-\mathcal{R}e\langle x, z\rangle$ is maximized). For the model associated with structured sparsity in levels, with the regularizer $f = f_w$ and weights $w_j = 1/\sqrt{k_j}$, this consists in taking $x = -z_T$ where T is the support such that $z_T \in \Sigma$ and z_T concentrate the most energy of z. With such an x, denoting $z_{T_c} = z - z_T$, one can show that

$$\delta(-z_T, z) = \frac{1}{\sqrt{\frac{\|z_{T_c}\|_\Sigma^2}{\|z_T\|_\mathcal{H}^2} + 1}} \tag{32}$$

Since $z \in \mathcal{T}_f(\Sigma)$, one shows that the fact that z_T concentrates the most energy implies that $f(z_{T_c}) \leq f(z_T)$, which in turn allows one to conclude that $\|z_{T_c}\|_\Sigma^2 / \|z_T\|_\mathcal{H}^2 \leq 1 + J$ is bounded using a control of $\|z_{T_c}\|_\Sigma^2$ obtained by extending Cai's sparse decomposition of polytopes [9]. This leads to the bound $\delta_\Sigma(f_w) \geq 1/\sqrt{2+J}$ mentioned in Section 4.2.3.

5 Generality of the Whole Framework

The proof of the existence of random linear maps that reduce dimension while satisfying the RIP is valid for any finite-dimensional model set Σ in any Hilbert space. The guarantees for convex decoders from Section 4 allow to define a critical RIP value for any union of subspaces Σ and any regularizer f (for some pairs, this may yield $\delta_\Sigma(f) = 0$, e.g., the right-hand side of Figure 7). Overall, the compressive sensing framework described in this chapter is thus very general, and we give here an overview of examples where it applies.

5.1 A Flexible Way to Guarantee Recovery

The following list summarizes the results of the combined framework of Sections 3 and 4 for classical pairs of model Σ and regularizer f. It states the model Σ, the considered regularizer f, a lower bound on our sufficient RIP constant $\delta_\Sigma(f)$, and a sufficient number of (random subgaussian) measurements m to guarantee recovery using the construction from Section 3.

- Σ = Linear subspace of dimension n, f = indicator function ι_Σ or $\|\cdot\|_\Sigma$.
  - $\delta_\Sigma(f) = 1$: This sufficient RIP constant was already known, e.g., [8].
  - Sufficient number of measurements: $m \gtrsim n$.

- $\Sigma = k$-sparse vectors in dimension $n, f = \ell^1$-norm.
  - $\delta_\Sigma(f) \geq 1/\sqrt{2}$. This is the sharp RIP constant of Cai et al. [9] (sharpness will be discussed in the next section).
  - Sufficient number of measurements: $m \gtrsim klog(n)$.
- $\Sigma =$ Matrices of rank lower than r in dimension $n \times n, f =$ nuclear norm.
  - $\delta_\Sigma(f) \geq 1/\sqrt{2}$. This is also the sharp RIP constant of Cai et al. [9].
  - Sufficient number of measurements: $m \gtrsim rn$.
- $\Sigma =$ Finite union of k 1D-half-spaces with coherence $\mu(\Sigma), f = \|\cdot\|_\Sigma$.
  - $\delta_\Sigma(f) \geq \frac{2(1-\mu(\Sigma))}{3+2\mu(\Sigma)}$.
  - Sufficient number of measurements: $m \gtrsim \log(k)/\delta_\Sigma(f)^2$.
- $\Sigma =$ Permutation matrices of dimension $n \times n, f = \|\cdot\|_\Sigma$.
  - $\delta_\Sigma(f) \geq \frac{2}{3}$.
  - Sufficient number of measurements: $m \gtrsim n\log(n)$.

Combining the formalism of [33] with that of [38], these models can also be considered in an infinite-dimensional space $\mathcal{H}$, which is convenient to handle analog compressive sensing scenarios. Stable recovery guarantees are still valid in this infinite-dimensional setting. For robustness, one must make sure that the constant $C_{f,A,\Sigma}$ is finite. For convex f, this might need some further assumptions on the behavior of f and A outside of the space $\mathcal{E}(\Sigma)$ (the subspace spanned by Σ) as mentioned in Section 4.2.3.

5.2 Uniform vs Nonuniform Recovery Guarantees

The framework described in this chapter focuses on *uniform* recovery guarantees for arbitrary linear operators. Another trend of general framework for compressive sensing focuses on nonuniform guarantees for Gaussian observations. In particular, Chandrasekaran et al. [17] studied the general nonuniform recovery from Gaussian observations with atomic norms. In this case, the goal is to show that, for any element x of the model, atomic norm minimization will recover x from Ax with high probability on the draw of A. In contrast, in the framework presented in this chapter, we established conditions so that (with high probability) *the same* linear operator A (i.e., a particular draw of a random operator) allows to stably and robustly recover *all elements of the model* with arbitrary regularizers. Moreover, these results are proved for general random matrices (typically, subgaussian matrices).

5.3 Extensions

The guarantees for convex decoders for unions of subspaces from Section 4 have further been extended to the case where the model set is a cone (positively homogeneous sets) [38, Theorem 3.1]. This covers models such as (subsets of) the cone of positive semi-definite matrices or that of nonnegative matrices.

Beyond the pure Hilbert norm setting described in this chapter, the generalized definition of the RIP from equation (11) or its further generalizations to arbitrary norms in $\mathcal{H}$ can be used [33] to establish dimension reduction results for structured acquisition. An example is the use of random rank-one projections (which are a subset of sub-exponential random matrices) [10], which offer a computationally efficient way to gather linear observations of a matrix, thus making them an interesting observation method for algorithmic purposes in the low-rank matrix recovery problem. While the RIP constant $\delta_\Sigma(f)$ has not been extended to such settings yet, such developments seem accessible.

5.4 Sharpness of Results?

In [33] the finite dimension of the normalized secant set allows one to conclude on the possibilities in terms of dimension reduction. Only a number of measurements of the order of the dimension is sufficient. However, this hypothesis is not necessary. It is possible to find a model Σ whose normalized secant set $\mathcal{S}$ has infinite upper box-counting dimension such that there exists a measurement operator with the RIP on $\mathcal{S}$. Hence, a weaker necessary and sufficient condition on the "dimension" of $\mathcal{S}$ could exist to guarantee the existence of measurement operators with stable dimension reduction capabilities.

In terms of recovery for arbitrary regularizers, it has been shown that a sufficient RIP constant can be provided. For classical families of models and regularizers (sparse recovery with the ℓ^1-norm and low-rank matrix recovery with the nuclear norm), as well as for structured sparsity and the associated group norm, the constant $\delta_\Sigma(f)$ is sharp in the following way: we know that there exist RIP matrices with constant arbitrarily close to $\delta_\Sigma(f)$ (here $1/\sqrt{2}$) which do not permit uniform recovery [10, 19] for some dimension of $\mathcal{H}$ and some sparsity k (or rank r). Considering sparsity in levels, we observe that $\delta_\Sigma(f)$ complies with the necessary dependency on the ratios of sparsity between levels and the number of levels J [6]. These sharpness results all consider *families* of models and regularizers: it is a worst-case sharpness among these families of regularizer. However, one can consider the question of strong sharpness: for a given model Σ and regularizer f, what is the biggest RIP constant sufficient to guarantee recovery?

5.5 New Frontiers: Super-Resolution and Compressive Learning

Much of the algorithmic and mathematical techniques revolving around the notion of sparsity in the context of inverse problems and compressive sensing have been developed with finite-dimensional models, involving, e.g., a discretization of the time domain or of the frequency domain. However, the physical phenomena underlying the acquisition of modern data from the analog world are rather

intrinsically continuous [40]. The generic framework for inverse problems and dimension reduction presented in this chapter is directly set up in an arbitrary Hilbert space setting, and as such it opens new perspectives for handling the analog nature of many problems.

Super-resolution is one such problem. In super-resolution, one aims at recovering spikes combinations of few spikes from their low-pass observation. While spikes are usually modeled with Dirac measures, which can be considered as belonging to certain Banach spaces of measures (e.g., equipped with the total variation norm), one way of bringing super-resolution close to the content of this chapter is to consider a kernel metric, which will bring a Hilbert structure to such Banach spaces [26]. Intuitively, this amounts to choosing a high resolution at which we will measure energy in the signal space. In this context, all the results on recovery guarantees and dimension reduction hold. Several questions remain standing: is it possible to find a sufficient RIP constant $\delta_\Sigma(f)$ that also holds in this context? Do usual models in Banach spaces have a normalized secant set with finite dimension? With the work of [12, 16, 22], we already know that low-pass filtering allows to recover spikes up to some resolution with a convex decoder.

Another related problem is compressive learning. In [7, 29] it is shown empirically that Gaussian mixtures can be recovered from a so-called sketch of the data, which can be considered as random Fourier measurements of their probability density. Recent works suggest that for an appropriately chosen kernel metric, the secant set of sufficiently separated mixtures of Diracs is of finite dimension for appropriately chosen kernel metric [30]. It is then possible to guarantee the success of the ideal decoder with random observations. Practical results have been obtained using a greedy heuristic approach to the problem [31]. These results seem to indicate a possible generalization of the theory of dimension reduction and convex recovery to these problems.

Acknowledgements This work was supported in part by the European Research Council PLEASE project (ERC-StG-2011-277906) and the European Research Council C-SENSE project (ERC-ADG-2015-694888). M.E. Davies would like to acknowledge the support of EPSRC grant EP/J015180/1.

References

1. B. Adcock, A.C. Hansen, Generalized sampling and infinite-dimensional compressed sensing. Found. Comput. Math. **16**(5), 1263–1323 (2016)
2. B. Adcock, A. Hansen, B. Roman, G. Teschke, Generalized sampling: stable reconstructions, inverse problems and compressed sensing over the continuum. Adv. Imag. Electr. Phys. **182**(1), 187–279 (2014)
3. A. Argyriou, R. Foygel, N. Srebro, Sparse prediction with the k-support norm, in *Advances in Neural Information Processing Systems*, vol. 25, ed. by F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Curran Associates, Inc., Red Hook, 2012), pp. 1457–1465
4. U. Ayaz, S. Dirksen, H. Rauhut, Uniform recovery of fusion frame structured sparse signals. Appl. Comput. Harmon. Anal. **41**(2), 341–361 (2016)

5. R.G. Baraniuk, V. Cevher, M.F. Duarte, C. Hegde, Model-based compressive sensing. IEEE Trans. Inf. Theory **56**(4), 1982–2001 (2010)
6. A. Bastounis, A.C. Hansen, On random and deterministic compressed sensing and the restricted isometry property in levels, in *2015 International Conference on Sampling Theory and Applications (SampTA)* (IEEE, Piscataway, 2015), pp. 297–301
7. A. Bourrier, R. Gribonval, P. Pérez, Compressive gaussian mixture estimation, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (2013), pp. 6024–6028
8. A. Bourrier, M. Davies, T. Peleg, P. Perez, R. Gribonval, Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. IEEE Trans. Inf. Theory **60**(12), 7928–7946 (2014)
9. T. Cai, A. Zhang, Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. IEEE Trans. Inf. Theory **60**(1), 122–132 (2014)
10. T.T. Cai, A. Zhang, et al., Rop: matrix recovery via rank-one projections. Ann. Stat. **43**(1), 102–138 (2015)
11. E.J. Candès, The restricted isometry property and its implications for compressed sensing. C. R. Math. **346**(9-10), 589–592 (2008)
12. E.J. Candès, C. Fernandez-Granda, Super-resolution from noisy data. J. Four. Anal. Appl. **19**(6), 1229–1254 (2013)
13. E.J. Candes, Y. Plan, Matrix completion with noise. Proc. IEEE **98**(6), 925–936 (2010)
14. E.J. Candes, Y. Plan, Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. IEEE Trans. Inf. Theory **57**(4), 2342–2359 (2011)
15. E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
16. Y.D. Castro, F. Gamboa, D. Henrion, J.B. Lasserre, Exact solutions to super resolution on semi-algebraic domains in higher dimensions. IEEE Trans. Inf. Theory **63**(1), 621–630 (2017)
17. V. Chandrasekaran, B. Recht, P. Parrilo, A. Willsky, The convex geometry of linear inverse problems. Found. Comput. Math. **12**(6), 805–849 (2012)
18. A. Cohen, W. Dahmen, R. DeVore, Compressed sensing and best k-term approximation. J. Am. Math. Soc **22**(1), 211–231 (2009)
19. M.E. Davies, R. Gribonval, Restricted isometry constants where ℓ^p sparse recovery can fail for $0 < p \leq 1$. IEEE Trans. Inf. Theory **55**(5), 2203–2214 (2009)
20. S. Dirksen, Dimensionality reduction with subgaussian matrices: a unified theory. Found. Comput. Math. **16**(5), 1367–1396 (2016)
21. D.L. Donoho, For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. Commun. Pure Appl. Math. **59**(6), 797–829 (2006)
22. V. Duval, G. Peyré, Exact support recovery for sparse spikes deconvolution. Found. Comput. Math. **15**(5), 1315–1355 (2015)
23. A. Eftekhari, M.B. Wakin, New analysis of manifold embeddings and signal recovery from compressive measurements. Appl. Comput. Harmon. Anal. **39**(1), 67–109 (2015)
24. Y.C. Eldar, P. Kuppinger, H. Bolcskei, Block-sparse signals: uncertainty relations and efficient recovery. IEEE Trans. Signal Process. **58**(6), 3042–3054 (2010)
25. S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Springer, New York, 2013)
26. A. Gretton, Introduction to rkhs, and some simple kernel algorithms, in *Adv. Top. Mach. Learn. Lecture Conducted from University College London* (2013)
27. R. Gribonval, M. Nielsen, Beyond sparsity: recovering structured representations by ℓ^1 minimization and greedy algorithms. Adv. Comput. Math. **28**(1), 23–41 (2008)
28. C.J. Hillar, L.-H. Lim, Most tensor problems are np-hard. J. ACM (JACM) **60**(6), 45 (2013)
29. N. Keriven, A. Bourrier, R. Gribonval, P. Pérèz, Sketching for large-scale learning of mixture models, in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)* (2015)
30. N. Keriven, A. Bourrier, R. Gribonval, P. Pérez, Sketching for large-scale learning of mixture models (June 2016). Preprint. https://hal.inria.fr/hal-01329195

31. N. Keriven, N. Tremblay, Y. Traonmilin, R. Gribonval, Compressive k-means, in *ICASSP 2017* (2016, to appear) https://hal.archives-ouvertes.fr/hal-01386077
32. G. Kutyniok, Clustered sparsity and separation of cartoon and texture. SIAM J. Imag. Sci. **6**(2), 848–874 (2013)
33. G. Puy, M.E. Davies, R. Gribonval, Recipes for stable linear embeddings from hilbert spaces to $\mathbb{R}^m$ (2017, to appear). Trans. Inf. Theory. Preprint. https://hal.inria.fr/hal-01203614
34. B. Recht, M. Fazel, P. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. **52**(3), 471–501 (2010)
35. J.C. Robinson, *Dimensions, Embeddings, and Attractors*, vol. 186 (Cambridge University Press, Cambridge, 2010)
36. C. Studer, R.G. Baraniuk, Stable restoration and separation of approximately sparse signals. Appl. Comput. Harmon. Anal. **37**(1), 12–35 (2014)
37. C. Sutour, C.-A. Deledalle, J.-F. Aujol, Estimation of the noise level function based on a nonparametric detection of homogeneous image regions. SIAM J. Imag. Sci. **8**(4), 2622–2661 (2015)
38. Y. Traonmilin, R. Gribonval, Stable recovery of low-dimensional cones in Hilbert spaces: One RIP to rule them all. Appl. Comput. Harmon. Anal. (2016). https://doi.org/10.1016/j.acha.2016.08.004
39. Y. Traonmilin, S. Ladjal, A. Almansa, Robust multi-image processing with optimal sparse regularization. J. Math. Imag. Vis. **51**(3), 413–429 (2015)
40. M. Vetterli, P. Marziliano, T. Blu, Sampling signals with finite rate of innovation. IEEE Trans. Signal Process. **50**(6), 1417–1428 (2002)
41. M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **68**(1), 49–67 (2006)

Applied and Numerical Harmonic Analysis (87 Volumes)

A. Saichev and W.A. Woyczyński: *Distributions in the Physical and Engineering Sciences* (ISBN 978-0-8176-3924-2)
C.E. D'Attellis and E.M. Fernandez-Berdaguer: *Wavelet Theory and Harmonic Analysis in Applied Sciences* (ISBN 978-0-8176-3953-2)
H.G. Feichtinger and T. Strohmer: *Gabor Analysis and Algorithms* (ISBN 978-0-8176-3959-4)
R. Tolimieri and M. An: *Time-Frequency Representations* (ISBN 978-0-8176-3918-1)
T.M. Peters and J.C. Williams: *The Fourier Transform in Biomedical Engineering* (ISBN 978-0-8176-3941-9)
G.T. Herman: *Geometry of Digital Spaces* (ISBN 978-0-8176-3897-9)
A. Teolis: *Computational Signal Processing with Wavelets* (ISBN 978-0-8176-3909-9)
J. Ramanathan: *Methods of Applied Fourier Analysis* (ISBN 978-0-8176-3963-1)
J.M. Cooper: *Introduction to Partial Differential Equations with MATLAB* (ISBN 978-0-8176-3967-9)
A. Procházka, N.G. Kingsbury, P.J. Payner, and J. Uhlir: *Signal Analysis and Prediction* (ISBN 978-0-8176-4042-2)
W. Bray and C. Stanojevic: *Analysis of Divergence* (ISBN 978-1-4612-7467-4)
G.T. Herman and A. Kuba: *Discrete Tomography* (ISBN 978-0-8176-4101-6)
K. Gröchenig: *Foundations of Time-Frequency Analysis* (ISBN 978-0-8176-4022-4)
L. Debnath: *Wavelet Transforms and Time-Frequency Signal Analysis* (ISBN 978-0-8176-4104-7)
J.J. Benedetto and P.J.S.G. Ferreira: *Modern Sampling Theory* (ISBN 978-0-8176-4023-1)
D.F. Walnut: *An Introduction to Wavelet Analysis* (ISBN 978-0-8176-3962-4)
A. Abbate, C. DeCusatis, and P.K. Das: *Wavelets and Subbands* (ISBN 978-0-8176-4136-8)

H. Boche et al. (eds.), *Compressed Sensing and its Applications*,
Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-69802-1

O. Bratteli, P. Jorgensen, and B. Treadway: *Wavelets Through a Looking Glass* (ISBN 978-0-8176-4280-80
H.G. Feichtinger and T. Strohmer: *Advances in Gabor Analysis* (ISBN 978-0-8176-4239-6)
O. Christensen: *An Introduction to Frames and Riesz Bases* (ISBN 978-0-8176-4295-2)
L. Debnath: *Wavelets and Signal Processing* (ISBN 978-0-8176-4235-8)
G. Bi and Y. Zeng: *Transforms and Fast Algorithms for Signal Analysis and Representations* (ISBN 978-0-8176-4279-2)
J.H. Davis: *Methods of Applied Mathematics with a MATLAB Overview* (ISBN 978-0-8176-4331-7)
J.J. Benedetto and A.I. Zayed: *Sampling, Wavelets, and Tomography* (ISBN 978-0-8176-4304-1)
E. Prestini: *The Evolution of Applied Harmonic Analysis* (ISBN 978-0-8176-4125-2)
L. Brandolini, L. Colzani, A. Iosevich, and G. Travaglini: *Fourier Analysis and Convexity* (ISBN 978-0-8176-3263-2)
W. Freeden and V. Michel: *Multiscale Potential Theory* (ISBN 978-0-8176-4105-4)
O. Christensen and K.L. Christensen: *Approximation Theory* (ISBN 978-0-8176-3600-5)
O. Calin and D.-C. Chang: *Geometric Mechanics on Riemannian Manifolds* (ISBN 978-0-8176-4354-6)
J.A. Hogan: *Time-Frequency and Time-Scale Methods* (ISBN 978-0-8176-4276-1)
C. Heil: *Harmonic Analysis and Applications* (ISBN 978-0-8176-3778-1)
K. Borre, D.M. Akos, N. Bertelsen, P. Rinder, and S.H. Jensen: *A Software-Defined GPS and Galileo Receiver* (ISBN 978-0-8176-4390-4)
T. Qian, M.I. Vai, and Y. Xu: *Wavelet Analysis and Applications* (ISBN 978-3-7643-7777-9)
G.T. Herman and A. Kuba: *Advances in Discrete Tomography and Its Applications* (ISBN 978-0-8176-3614-2)
M.C. Fu, R.A. Jarrow, J.-Y. Yen, and R.J. Elliott: *Advances in Mathematical Finance* (ISBN 978-0-8176-4544-1)
O. Christensen: *Frames and Bases* (ISBN 978-0-8176-4677-6)
P.E.T. Jorgensen, J.D. Merrill, and J.A. Packer: *Representations, Wavelets, and Frames* (ISBN 978-0-8176-4682-0)
M. An, A.K. Brodzik, and R. Tolimieri: *Ideal Sequence Design in Time-Frequency Space* (ISBN 978-0-8176-4737-7)
S.G. Krantz: *Explorations in Harmonic Analysis* (ISBN 978-0-8176-4668-4)
B. Luong: *Fourier Analysis on Finite Abelian Groups* (ISBN 978-0-8176-4915-9)
G.S. Chirikjian: *Stochastic Models, Information Theory, and Lie Groups, Volume 1* (ISBN 978-0-8176-4802-2)
C. Cabrelli and J.L. Torrea: *Recent Developments in Real and Harmonic Analysis* (ISBN 978-0-8176-4531-1)
M.V. Wickerhauser: *Mathematics for Multimedia* (ISBN 978-0-8176-4879-4)

B. Forster, P. Massopust, O. Christensen, K. Gröchenig, D. Labate, P. Vandergheynst, G. Weiss, and Y. Wiaux: *Four Short Courses on Harmonic Analysis* (ISBN 978-0-8176-4890-9)
O. Christensen: *Functions, Spaces, and Expansions* (ISBN 978-0-8176-4979-1)
J. Barral and S. Seuret: *Recent Developments in Fractals and Related Fields* (ISBN 978-0-8176-4887-9)
O. Calin, D.-C. Chang, and K. Furutani, and C. Iwasaki: *Heat Kernels for Elliptic and Sub-elliptic Operators* (ISBN 978-0-8176-4994-4)
C. Heil: *A Basis Theory Primer* (ISBN 978-0-8176-4686-8)
J.R. Klauder: *A Modern Approach to Functional Integration* (ISBN 978-0-8176-4790-2)
J. Cohen and A.I. Zayed: *Wavelets and Multiscale Analysis* (ISBN 978-0-8176-8094-7)
D. Joyner and J.-L. Kim: *Selected Unsolved Problems in Coding Theory* (ISBN 978-0-8176-8255-2)
G.S. Chirikjian: *Stochastic Models, Information Theory, and Lie Groups, Volume 2* (ISBN 978-0-8176-4943-2)
J.A. Hogan and J.D. Lakey: *Duration and Bandwidth Limiting* (ISBN 978-0-8176-8306-1)
G. Kutyniok and D. Labate: *Shearlets* (ISBN 978-0-8176-8315-3)
P.G. Casazza and P. Kutyniok: *Finite Frames* (ISBN 978-0-8176-8372-6)
V. Michel: *Lectures on Constructive Approximation* (ISBN 978-0-8176-8402-0)
D. Mitrea, I. Mitrea, M. Mitrea, and S. Monniaux: *Groupoid Metrization Theory* (ISBN 978-0-8176-8396-2)
T.D. Andrews, R. Balan, J.J. Benedetto, W. Czaja, and K.A. Okoudjou: *Excursions in Harmonic Analysis, Volume 1* (ISBN 978-0-8176-8375-7)
T.D. Andrews, R. Balan, J.J. Benedetto, W. Czaja, and K.A. Okoudjou: *Excursions in Harmonic Analysis, Volume 2* (ISBN 978-0-8176-8378-8)
D.V. Cruz-Uribe and A. Fiorenza: *Variable Lebesgue Spaces* (ISBN 978-3-0348-0547-6)
W. Freeden and M. Gutting: *Special Functions of Mathematical (Geo-)Physics* (ISBN 978-3-0348-0562-9)
A. Saichev and W.A. Woyczyński: *Distributions in the Physical and Engineering Sciences, Volume 2: Linear and Nonlinear Dynamics of Continuous Media* (ISBN 978-0-8176-3942-6)
S. Foucart and H. Rauhut: *A Mathematical Introduction to Compressive Sensing* (ISBN 978-0-8176-4947-0)
G. Herman and J. Frank: *Computational Methods for Three-Dimensional Microscopy Reconstruction* (ISBN 978-1-4614-9520-8)
A. Paprotny and M. Thess: *Realtime Data Mining: Self-Learning Techniques for Recommendation Engines* (ISBN 978-3-319-01320-6)
A. Zayed and G. Schmeisser: *New Perspectives on Approximation and Sampling Theory: Festschrift in Honor of Paul Butzer's 85th Birthday* (ISBN 978-3-319-08800-6)

R. Balan, M. Begue, J. Benedetto, W. Czaja, and K.A Okoudjou: *Excursions in Harmonic Analysis, Volume 3* (ISBN 978-3-319-13229-7)
H. Boche, R. Calderbank, G. Kutyniok, J. Vybiral: *Compressed Sensing and its Applications* (ISBN 978-3-319-16041-2)
S. Dahlke, F. De Mari, P. Grohs, and D. Labate: *Harmonic and Applied Analysis: From Groups to Signals* (ISBN 978-3-319-18862-1)
A. Aldroubi, *New Trends in Applied Harmonic Analysis* (ISBN 978-3-319-27871-1)
M. Ruzhansky: *Methods of Fourier Analysis and Approximation Theory* (ISBN 978-3-319-27465-2)
G. Pfander: *Sampling Theory, a Renaissance* (ISBN 978-3-319-19748-7)
R. Balan, M. Begue, J. Benedetto, W. Czaja, and K.A Okoudjou: *Excursions in Harmonic Analysis, Volume 4* (ISBN 978-3-319-20187-0)
O. Christensen: *An Introduction to Frames and Riesz Bases, Second Edition* (ISBN 978-3-319-25611-5)
E. Prestini: *The Evolution of Applied Harmonic Analysis: Models of the Real World, Second Edition* (ISBN 978-1-4899-7987-2)
J.H. Davis: *Methods of Applied Mathematics with a Software Overview, Second Edition* (ISBN 978-3-319-43369-1)
M. Gilman, E. M. Smith, S. M. Tsynkov: *Transionospheric Synthetic Aperture Imaging* (ISBN 978-3-319-52125-1)
S. Chanillo, B. Franchi, G. Lu, C. Perez, E.T. Sawyer: *Harmonic Analysis, Partial Differential Equations and Applications* (ISBN 978-3-319-52741-3)
R. Balan, J. Benedetto, W. Czaja, M. Dellatorre, and K.A Okoudjou: *Excursions in Harmonic Analysis, Volume 5* (ISBN 978-3-319-54710-7)
I. Pesenson, Q.T. Le Gia, A. Mayeli, H. Mhaskar, D.X. Zhou: *Frames and Other Bases in Abstract and Function Spaces: Novel Methods in Harmonic Analysis, Volume 1* (ISBN 978-3-319-55549-2)
I. Pesenson, Q.T. Le Gia, A. Mayeli, H. Mhaskar, D.X. Zhou: *Recent Applications of Harmonic Analysis to Function Spaces, Differential Equations, and Data Science: Novel Methods in Harmonic Analysis, Volume 2* (ISBN 978-3-319-55555-3)
F. Weisz: *Convergence and Summability of Fourier Transforms and Hardy Spaces* (ISBN 978-3-319-56813-3)
C. Heil: *Metrics, Norms, Inner Products, and Operator Theory* (ISBN 978-3-319-65321-1)
S. Waldron: *An Introduction to Finite Tight Frames: Theory and Applications.* (ISBN: 978-0-8176-4814-5)
D. Joyner and C.G. Melles: *Adventures in Graph Theory: A Bridge to Advanced Mathematics.* (ISBN: 978-3-319-68381-2)
B. Han: *Framelets and Wavelets: Algorithms, Analysis, and Applications* (ISBN: 978-3-319-68529-8)
H. Boche, G. Caire, R. Calderbank, M. März, G. Kutyniok, R. Mathar: *Compressed Sensing and Its Applications* (ISBN: 978-3-319-69801-4)

For an up-to-date list of ANHA titles, please visit http://www.springer.com/series/4968