

Holger Boche
Robert Calderbank
Gitta Kutyniok
Jan Vybíral
Editors

Compressed Sensing and its Applications

MATHEON Workshop 2013



Birkhäuser



Applied and Numerical Harmonic Analysis

Series Editor

John J. Benedetto

University of Maryland
College Park, MD, USA

Editorial Advisory Board

Akram Aldroubi

Vanderbilt University
Nashville, TN, USA

Douglas Cochran

Arizona State University
Phoenix, AZ, USA

Hans G. Feichtinger

University of Vienna
Vienna, Austria

Christopher Heil

Georgia Institute of Technology
Atlanta, GA, USA

Stéphane Jaffard

University of Paris XII
Paris, France

Jelena Kovačević

Carnegie Mellon University
Pittsburgh, PA, USA

Gitta Kutyniok

Technische Universität Berlin
Berlin, Germany

Mauro Maggioni

Duke University
Durham, NC, USA

Zuowei Shen

National University of Singapore
Singapore, Singapore

Thomas Strohmer

University of California
Davis, CA, USA

Yang Wang

Michigan State University
East Lansing, MI, USA

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

Holger Boche • Robert Calderbank

Gitta Kutyniok • Jan Vybíral

Editors

Compressed Sensing and its Applications

MATHEON Workshop 2013



Birkhäuser

Editors

Holger Boche
Lehrstuhl für Theoretische
Informationstechnik
Technische Universität München
München, Germany

Robert Calderbank
Department of Electrical
and Computer Engineering
Duke University
Durham, NC, USA

Gitta Kutyniok
Institut für Mathematik
Technische Universität Berlin
Berlin, Germany

Jan Vybíral
Faculty of Mathematics and Physics
Charles University
Prague, Czech Republic

ISSN 2296-5009

ISSN 2296-5017 (electronic)

Applied and Numerical Harmonic Analysis

ISBN 978-3-319-16041-2

ISBN 978-3-319-16042-9 (eBook)

DOI 10.1007/978-3-319-16042-9

Library of Congress Control Number: 2015933373

Mathematics Subject Classification (2010): 94A12, 65F22, 94A20, 68U10, 90C25, 15B52

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

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

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:

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

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

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

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

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

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

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

University of Maryland
College Park

John J. Benedetto
Series Editor

Preface

Compressed sensing is the idea that it should be possible to capture attributes of a signal using very few measurements. Since publication of the initial papers in 2006, it has captured the imagination of the international signal processing community, and the mathematical foundations are nowadays quite well understood. Key to compressed sensing is the surprising fact that high-dimensional signals, which allow a sparse representation by a suitable basis or, more generally, a frame, can be recovered from what were very few linear measurements by using efficient algorithms such as convex optimization approaches.

From the very beginning, the area profited from a fruitful interaction of applied mathematicians with engineers, with new applications leading to new mathematical methods and vice versa. Applications of compressed sensing to communication theory, imaging sciences, optics, radar technology, sensor networks, and tomography have been developed, with the technology in some areas more advanced than in others.

In December 2013, an international workshop was organized by the editors of this volume at the Technische Universität Berlin focusing specifically on application aspects of compressed sensing. In this sense, it was the first meeting with a focus on the application side of this novel methodology. The workshop was supported by the MATHEON, which is a research center in Berlin for “Mathematics for Key Technologies,” as well as by the German Research Foundation (DFG). The workshop was attended by about 150 researchers from 13 different countries. Experts in a variety of research areas besides electrical engineering and mathematics were present among which were biologists, chemists, computer scientists, or material scientists. This mixture of people with different backgrounds oft led to particularly fruitful and inspiring discussions.

This book features contributions by three plenary speakers, namely Babak Hassibi (California Institute of Technology), Ali Pezeshki (Colorado State University), and Guillermo Shapiro (Duke University), and by thirteen invited speakers, namely Petros Boufounos (Mitsubishi Research Lab, USA), Volkan Cevher (EPFL), Shmuel Friedland (University of Illinois), Remi Gribonval (INRIA, Rennes), Anders Hansen (University of Cambridge), Peter Jung (Technische

Universität Berlin), Felix Krahmer (Technische Universität München, Germany), Dustin Mixon (Air Force Institute of Technology), Holger Rauhut (RWTH Aachen), Miguel Rodrigues (University College London), Rayan Saab (University of California, San Diego), Reinhold Schneider (Technische Universität Berlin, Germany), and Philipp Walk (Technische Universität München, Germany). It is the first monograph devoted to applications of compressed sensing. It is aimed at a broad readership including graduate students and researchers in the areas of mathematics, computer science, and engineering. However, it is also accessible to researchers working in any other field requiring methodologies for data science. Hence this volume can be used both as a state-of-the-art monograph on applications of compressed sensing and as a textbook for graduate students. Here is a brief outline of the contents of each chapter.

Chapter 1 is written by the editors and provides an introduction as well as a self-contained overview of the main results on the theory and applications of compressed sensing. It also serves to unify the notation throughout the whole book. Chapters 2–4 contain the contributions of the plenary speakers, whereas Chapters 5–15 feature the presentations of the invited speakers. Several chapters focus on problems one is facing when applying compressed sensing such as the problem of model mismatch (Chapter 3), quantization (Chapter 7), and unknown sparsifying dictionary (Chapter 8). Other chapters analyze specific constraints compressed sensing has to be adapted to for specific applications, in particular, structured sparsity, for which optimal sampling strategies (Chapter 5), algorithmic aspects (Chapters 4 + 12), and the specific structure of tensors (Chapters 9 + 14) are analyzed. Two chapters study theoretical obstacles, which, if overcome, would increase the impact of compressed sensing; Chapter 13 explores specific deterministic measurement matrices, and Chapter 11 explores co-sparsity-based reconstruction. The other chapters introduce and discuss the application of compressed sensing to areas such as acoustic imaging (Chapter 6), temporal color imaging (Chapter 2), and wireless communications (Chapters 10 + 15).

Finally, we would like to thank all members of the research group “Applied Functional Analysis” at Technische Universität Berlin, namely Martin Genzel, Mijail Guillemard, Anja Hedrich, Sandra Keiper, Anton Kolleck, Wang-Q Lim, Jackie Ma, Victoria Paternostro, Philipp Petersen, Friedrich Philipp, Rafael Reisenhofer, Martin Schäfer, Irena Bojarovska, and Yizhi Sun, without whom this MATHEON workshop would not have been possible.

München, Germany
Durham, NC, USA
Berlin, Germany
Prague, Czech Republic
December 2014

Holger Boche
Robert Calderbank
Gitta Kutyniok
Jan Vybíral

Contents

1	A Survey of Compressed Sensing	1
	Holger Boche, Robert Calderbank, Gitta Kutyniok, and Jan Vybíral	
2	Temporal Compressive Sensing for Video	41
	Patrick Llull, Xin Yuan, Xuejun Liao, Jianbo Yang, David Kittle, Lawrence Carin, Guillermo Sapiro, and David J. Brady	
3	Compressed Sensing, Sparse Inversion, and Model Mismatch	75
	Ali Pezeshki, Yuejie Chi, Louis L. Scharf, and Edwin K.P. Chong	
4	Recovering Structured Signals in Noise: Least-Squares Meets Compressed Sensing	97
	Christos Thrampoulidis, Samet Oymak, and Babak Hassibi	
5	The Quest for Optimal Sampling: Computationally Efficient, Structure-Exploiting Measurements for Compressed Sensing	143
	Ben Adcock, Anders C. Hansen, and Bogdan Roman	
6	Compressive Sensing in Acoustic Imaging	169
	Nancy Bertin, Laurent Daudet, Valentin Emiya, and Rémi Gribonval	
7	Quantization and Compressive Sensing	193
	Petros T. Boufounos, Laurent Jacques, Felix Krahmer, and Rayan Saab	
8	Compressive Gaussian Mixture Estimation	239
	Anthony Bourrier, Rémi Gribonval, and Patrick Pérez	
9	Two Algorithms for Compressed Sensing of Sparse Tensors	259
	Shmuel Friedland, Qun Li, Dan Schonfeld, and Edgar A. Bernal	

10	Sparse Model Uncertainties in Compressed Sensing with Application to Convolutions and Sporadic Communication	283
	Peter Jung and Philipp Walk	
11	Cosparsity in Compressed Sensing	315
	Maryia Kabanava and Holger Rauhut	
12	Structured Sparsity: Discrete and Convex Approaches	341
	Anastasios Kyrillidis, Luca Baldassarre, Marwa El Halabi, Quoc Tran-Dinh, and Volkan Cevher	
13	Explicit Matrices with the Restricted Isometry Property: Breaking the Square-Root Bottleneck	389
	Dustin G. Mixon	
14	Tensor Completion in Hierarchical Tensor Representations	419
	Holger Rauhut, Reinhold Schneider, and Željka Stojanac	
15	Compressive Classification: Where Wireless Communications Meets Machine Learning	451
	Miguel Rodrigues, Matthew Nokleby, Francesco Renna, and Robert Calderbank	
	Applied and Numerical Harmonic Analysis (69 volumes)	469

Chapter 1

A Survey of Compressed Sensing

Holger Boche, Robert Calderbank, Gitta Kutyniok, and Jan Vybíral

Abstract Compressed sensing was introduced some ten years ago as an effective way of acquiring signals, which possess a sparse or nearly sparse representation in a suitable basis or dictionary. Due to its solid mathematical backgrounds, it quickly attracted the attention of mathematicians from several different areas, so that the most important aspects of the theory are nowadays very well understood. In recent years, its applications started to spread out through applied mathematics, signal processing, and electrical engineering. The aim of this chapter is to provide an introduction into the basic concepts of compressed sensing. In the first part of this chapter, we present the basic mathematical concepts of compressed sensing, including the Null Space Property, Restricted Isometry Property, their connection to basis pursuit and sparse recovery, and construction of matrices with small restricted isometry constants. This presentation is easily accessible, largely self-contained, and includes proofs of the most important theorems. The second part gives an overview of the most important extensions of these ideas, including recovery of vectors with sparse representation in frames and dictionaries, discussion of (in)coherence and its implications for compressed sensing, and presentation of other algorithms of sparse recovery.

H. Boche (✉)

Technische Universität München, Theresienstr. 90/IV, München, Germany

e-mail: boche@tum.de

R. Calderbank

Duke University, 317 Gross Hall, Durham NC, USA

e-mail: robert.calderbank@duke.edu

G. Kutyniok

Technische Universität Berlin, Straße des 17. Juni 136, Berlin, Germany

e-mail: kutyniok@math.tu-berlin.de

J. Vybíral

Faculty of Mathematics and Physics, Charles University, Sokolovska 83,
186 00 Prague 8, Czech Republic

e-mail: vybiral@karlin.mff.cuni.cz

1.1 Introduction

Compressed sensing is a novel method of signal processing, which was introduced in [25] and [15, 16] and which profited from its very beginning from fruitful interplay between mathematicians, applied mathematicians, and electrical engineers. The mathematical concepts are inspired by ideas from a number of different disciplines, including numerical analysis, stochastic, combinatorics, and functional analysis. On the other hand, the applications of compressed sensing range from image processing [29], medical imaging [52], and radar technology [5] to sampling theory [56, 69], and statistical learning.

The aim of this chapter is twofold. In Section 1.3 we collect the basic mathematical ideas from numerical analysis, stochastic, and functional analysis used in the area of compressed sensing to give an overview of basic notions, including the Null Space Property and the Restricted Isometry Property, and the relations between them. Most of the material in this section is presented with a self-contained proof, using only few simple notions from approximation theory and stochastic recalled in Section 1.2. We hope that this presentation will make the mathematical concepts of compressed sensing appealing and understandable both to applied mathematicians and electrical engineers. Although it can also be used as a basis for a lecture on compressed sensing for a wide variety of students, depending on circumstances, it would have to be complemented by other subjects of the lecturers choice to make a full one-semester course. Let us stress that the material presented in this section is by no means new or original, actually it is nowadays considered classical, or “common wisdom” throughout the community.

The second aim of this chapter is to give (without proof) an overview of the most important extensions (Section 1.4). In this part, we refer to original research papers or to more extensive summaries of compressed sensing [23, 35, 40] for more details and further references.

1.2 Preliminaries

As the mathematical concepts of compressed sensing rely on the interplay of ideas from linear algebra, numerical analysis, stochastic, and functional analysis, we start with an overview of basic notions from these fields. We shall restrict ourselves to the minimum needed in the sequel.

1.2.1 Norms and quasi-norms

In the most simple setting of discrete signals on finite domain, signals are modeled as (column) vectors in then n -dimensional Euclidean space, denoted by \mathbb{R}^n . We shall

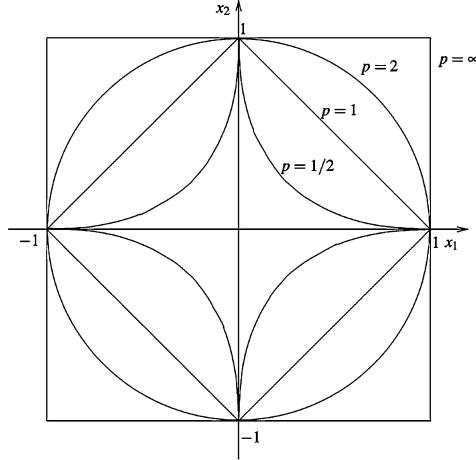


Fig. 1.1 Shape of the ℓ_p^2 unit ball for $p = 1/2, p = 1, p = 2$, and $p = \infty$

use different ways how to measure the size of such a vector. The most typical way, however, is to consider its ℓ_p^n -norm, which is defined for $x = (x_1, \dots, x_n)^T$ and $p \in (0, \infty]$ as (Fig. 1.1)

$$\|x\|_p = \begin{cases} \left(\sum_{j=1}^n |x_j|^p \right)^{1/p}, & p \in (0, \infty); \\ \max_{j=1, \dots, n} |x_j|, & p = \infty. \end{cases} \quad (1.1)$$

If $p < 1$, this expression does not satisfy the triangle inequality. Instead of that the following inequalities hold

$$\begin{aligned} \|x+z\|_p &\leq 2^{1/p-1} (\|x\|_p + \|z\|_p), \\ \|x+z\|_p^p &\leq \|x\|_p^p + \|z\|_p^p \end{aligned}$$

for all $x \in \mathbb{R}^n$ and all $z \in \mathbb{R}^n$. If $p = 2$, ℓ_2^n is a (real) Hilbert space with the scalar product

$$\langle x, z \rangle = z^T x = \sum_{i=1}^n x_i z_i.$$

If $x \in \mathbb{R}^n$, we can always find a permutation $\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$, such that the nonincreasing rearrangement $x^* \in [0, \infty)^n$ of x , defined by $x_j^* = |x_{\sigma(j)}|$ satisfies

$$x_1^* \geq x_2^* \geq \dots \geq x_n^* \geq 0.$$

If $T \subset \{1, \dots, n\}$ is a set of indices, we denote by $|T|$ the number of its elements. We shall complement this notation by denoting the size of the support of $x \in \mathbb{R}^n$ by

$$\|x\|_0 = |\text{supp}(x)| = |\{j : x_j \neq 0\}|.$$

Note that this expression is not even a quasinorm. The notation is justified by the observation, that

$$\lim_{p \rightarrow 0} \|x\|_p^p = \|x\|_0 \quad \text{for all } x \in \mathbb{R}^n.$$

Let k be a natural number at most equal to n . A vector $x \in \mathbb{R}^n$ is called k -sparse, if $\|x\|_0 \leq k$ and the set of all k -sparse vectors is denoted by

$$\Sigma_k = \{x \in \mathbb{R}^n : \|x\|_0 \leq k\}.$$

Finally, if $k < n$, the best k -term approximation $\sigma_k(x)_p$ of $x \in \mathbb{R}^n$ describes, how well can x be approximated by k -sparse vectors in the ℓ_p^n -norm. This can be expressed by the formula

$$\sigma_k(x)_p = \inf_{z \in \Sigma_k} \|x - z\|_p = \begin{cases} \left(\sum_{j=k+1}^n (x_j^*)^p \right)^{1/p}, & p \in (0, \infty); \\ x_{k+1}^*, & p = \infty. \end{cases} \quad (1.2)$$

The notions introduced so far can be easily transferred to n -dimensional complex spaces. Especially, the scalar product of $x, y \in \mathbb{C}^n$ is defined by

$$\langle x, y \rangle = \sum_{j=1}^n x_j \bar{y}_j,$$

where \bar{z} is the complex conjugate of $z \in \mathbb{C}$.

Linear operators between finite-dimensional spaces \mathbb{R}^n and \mathbb{R}^m can be represented with the help of matrices $A \in \mathbb{R}^{m \times n}$. The entries of A are denoted by a_{ij} , $i = 1, \dots, m$ and $j = 1, \dots, n$. The transpose of a matrix $A \in \mathbb{R}^{m \times n}$ is a matrix $A^T \in \mathbb{R}^{n \times m}$ with entries $(A^T)_{ij} = a_{ji}$. The identity matrix in $\mathbb{R}^{n \times n}$ or $\mathbb{C}^{n \times n}$ will be denoted by I .

1.2.2 Random Variables

As several important constructions from the field of compressed sensing rely on randomness, we recall the basic notions from probability theory.

We denote by $(\Omega, \Sigma, \mathbb{P})$ a probability space. Here stands Ω for the sample space, Σ for a σ -algebra of subsets of Ω , and \mathbb{P} is a probability measure on (Ω, Σ) . The sets $B \in \Sigma$ are called events, and their probability is denoted by

$$\mathbb{P}(B) = \int_B d\mathbb{P}(\omega).$$

A random variable X is a measurable function $X : \Omega \rightarrow \mathbb{R}$ and we denote by

$$\mu = \mathbb{E}X = \int_{\Omega} X(\omega) d\mathbb{P}(\omega)$$

its expected value, or mean, and by $\sigma^2 = \mathbb{E}[(X - \mu)^2] = \mathbb{E}(X^2) - (\mathbb{E}X)^2$ its variance. We recall Markov's inequality, which states

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|}{t} \quad \text{for all } t > 0. \quad (1.3)$$

A random variable X is called *normal* (or *Gaussian*), if it has a density function

$$f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right), \quad t \in \mathbb{R}$$

for some real μ and positive σ^2 , i.e. if $\mathbb{P}(a < X \leq b) = \int_a^b f(t) dt$ for all real $a < b$. In that case, the expected value of X is equal to μ and its variance to σ^2 and we often write $X \sim \mathcal{N}(\mu, \sigma^2)$. If $\mu = 0$ and $\sigma^2 = 1$, the normal variable is called *standard* and its density function is

$$f(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right), \quad t \in \mathbb{R}.$$

A random variable X is called *Rademacher* if

$$\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2. \quad (1.4)$$

Random variables X_1, \dots, X_N are called *independent*, if for every real t_1, \dots, t_N the following formula holds

$$\mathbb{P}(X_1 \leq t_1, \dots, X_N \leq t_N) = \prod_{j=1}^N \mathbb{P}(X_j \leq t_j).$$

In that case,

$$\mathbb{E}\left[\prod_{j=1}^N X_j\right] = \prod_{j=1}^N \mathbb{E}(X_j). \quad (1.5)$$

If the random variables X_1, \dots, X_N are independent and have the same distribution, we call them *independent identically distributed*, which is usually abbreviated as *i.i.d.*

1.3 Basic ideas of compressed sensing

There is a number of ways how to discover the landscape of compressed sensing. The point of view, which we shall follow in this section, is that we are looking for sparse solutions $x \in \mathbb{R}^n$ of a system of linear equations $Ax = y$, where $y \in \mathbb{R}^m$ and the $m \times n$ matrix A are known. We shall be interested in underdetermined systems, i.e. in the case $m \leq n$. Intuitively, this corresponds to solving the following optimization problem

$$\min_z \|z\|_0 \quad \text{subject to} \quad y = Az. \quad (P_0)$$

We will first show that this problem is numerically intractable if m and n are getting larger. Then we introduce the basic notions of compressed sensing, showing that for specific matrices A and measurement vectors y , one can recover the solution of (P_0) in a much more effective way.

1.3.1 Basis pursuit

The minimization problem (P_0) can obviously be solved by considering first all index sets $T \subset \{1, \dots, n\}$ with one element and employing the methods of linear algebra to decide if there is a solution x to the system with support included in T . If this fails for all such index sets, we continue with all index sets with two, three, and more elements. The obvious drawback is the rapidly increasing number of these index sets. Indeed, there are $\binom{n}{k}$ index sets $T \subset \{1, \dots, n\}$ with k elements and this quantity grows (in some sense) exponentially with k and n .

We shall start our tour through compressed sensing by showing that even every other algorithm solving (P_0) suffers from this drawback. This will be formulated in the language of complexity theory as the statement, that the (P_0) problem is NP-hard. Before we come to that, we introduce the basic terms used in the sequel. We refer for example to [2] for an introduction to computational complexity.

The *P-class* (“polynomial time”) consists of all decision problems that can be solved in polynomial time, i.e. with an algorithm, whose running time is bounded from above by a polynomial expression in the size of the input.

The *NP-class* (“nondeterministic polynomial time”) consists of all decision problems, for which there is a polynomial-time algorithm V (called verifier), with

the following property. If, given an input α , the right answer to the decision problem is “yes”, then there is a proof β , such that $V(\alpha, \beta) = \text{yes}$. Roughly speaking, when the answer to the decision problem is positive, then the proof of this statement can be verified with a polynomial-time algorithm.

Let us reformulate (P_0) as a decision problem. Namely, if the natural numbers k, m, n , $m \times n$ matrix A and $y \in \mathbb{R}^m$ are given, decide if there is a k -sparse solution x of the equation $Ax = y$. It is easy to see that this version of (P_0) is in the NP-class. Indeed, if the answer to the problem is “yes” and a certificate $x \in \mathbb{R}^n$ is given, then it can be verified in polynomial time if x is k -sparse and $Ax = y$.

A problem is called *NP-hard* if any of its solving algorithms can be transformed in polynomial time into a solving algorithm of any other NP-problem. We shall rely on a statement from complexity theory, that the following problem is both NP and NP-hard.

Exact cover problem

Given as the input a natural number m divisible by 3 and a system $\{T_j : j = 1, \dots, n\}$ of subsets of $\{1, \dots, m\}$ with $|T_j| = 3$ for all $j = 1, \dots, n$, decide, if there is a subsystem of mutually disjoint sets $\{T_j : j \in J\}$, such that $\bigcup_{j \in J} T_j = \{1, \dots, m\}$. Such a subsystem is frequently referred to as *exact cover*.

Let us observe that for any subsystem $\{T_j : j \in J\}$ it is easy to verify (in polynomial time) if it is an exact cover or not. So the problem is in the NP-class. The non-trivial statement from computational complexity is that this problem is also NP-hard. The exact formulation of (P_0) looks as follows.

ℓ_0 -minimization problem

Given natural numbers m, n , an $m \times n$ matrix A and a vector $y \in \mathbb{R}^m$ as input, find the solution of

$$\min_z \|z\|_0 \quad \text{s.t.} \quad y = Az.$$

Theorem 1. *The ℓ_0 -minimization problem is NP-hard.*

Proof. It is sufficient to show that any algorithm solving the ℓ_0 -minimization problem can be transferred in polynomial time into an algorithm solving the exact cover problem. Let therefore $\{T_j : j = 1, \dots, n\}$ be a system of subsets of $\{1, \dots, m\}$ with $|T_j| = 3$ for all $j = 1, \dots, n$. Then we construct a matrix $A \in \mathbb{R}^{m \times n}$ by putting

$$a_{ij} := \begin{cases} 1 & \text{if } i \in T_j, \\ 0 & \text{if } i \notin T_j, \end{cases}$$

i.e. the j th column of A is the indicator function of T_j (denoted by $\chi_{T_j} \in \{0, 1\}^m$) and

$$Ax = \sum_{j=1}^n x_j \chi_{T_j}. \quad (1.6)$$

The construction of A can of course be done in polynomial time.

Let now x be the solution to the ℓ_0 -minimization problem with the matrix A and the vector $y = (1, \dots, 1)^T$. It follows by (1.6) that $m = \|y\|_0 = \|Ax\|_0 \leq 3\|x\|_0$, i.e. that $\|x\|_0 \geq m/3$. We will show that the exact cover problem has a positive solution if, and only if, $\|x\|_0 = m/3$.

Indeed, if the exact cover problem has a positive solution, then there is a set $J \subset \{1, \dots, n\}$ with $|J| = m/3$ and

$$\chi_{\{1, \dots, m\}} = \sum_{j \in J} \chi_{T_j}.$$

Hence $y = Ax$ for $x = \chi_J$ and $\|x\|_0 = |J| = m/3$. If, on the other hand, $y = Ax$ and $\|x\|_0 = m/3$, then $\{T_j : j \in \text{supp}(x)\}$ solves the exact cover problem. \square

The ℓ_0 -minimization problem is NP-hard, if all matrices A and all measurement vectors y are allowed as inputs. The theory of compressed sensing shows nevertheless that for special matrices A and for $y = Ax$ for some sparse x , the problem can be solved efficiently.

In general, we replace the $\|z\|_0$ in (P0) by some $\|z\|_p$ for $p > 0$. To obtain a convex problem, we need to have $p \geq 1$. To obtain sparse solutions, $p \leq 1$ is necessary, cf. Figure 1.2.

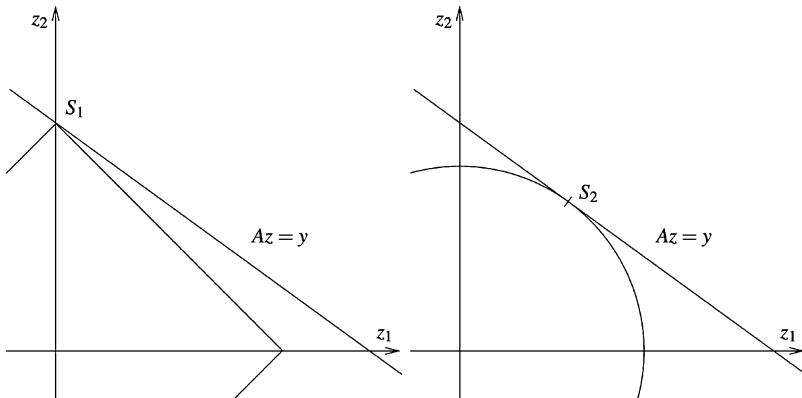


Fig. 1.2 Solution of $S_p = \operatorname{argmin}_{z \in \mathbb{R}^2} \|z\|_p$ s.t. $y = Az$ for $p = 1$ and $p = 2$

We are therefore naturally led to discuss under which conditions the solution to (P_0) coincides with the solution of the following convex optimization problem called *basis pursuit*

$$\min_z \|z\|_1 \quad \text{s.t.} \quad y = Az, \quad (P_1)$$

which was introduced in [19]. But before we come to that, let us show that in the real case this problem may be reformulated as a linear optimization problem, i.e. as the search for the minimizer of a linear function over a set given by linear constraints, whose number depends polynomially on the dimension. We refer to [42] for an introduction to linear programming.

Indeed, let us assume that (P_1) has a unique solution, which we denote by $x \in \mathbb{R}^n$. Then the pair (u, v) with $u = x^+$ and $v = x^-$, i.e. with

$$u_j = \begin{cases} x_j, & x_j \geq 0, \\ 0, & x_j < 0, \end{cases} \quad \text{and} \quad v_j = \begin{cases} 0, & x_j \geq 0, \\ -x_j, & x_j < 0, \end{cases}$$

is the unique solution of

$$\min_{u, v \in \mathbb{R}^n} \sum_{j=1}^n (u_j + v_j) \quad \text{s.t.} \quad Au - Av = y \quad \text{and} \quad u_j \geq 0 \quad \text{and} \quad v_j \geq 0 \quad \text{for all } j = 1, \dots, n. \quad (1.7)$$

If namely (u', v') is another pair of vectors admissible in (1.7), then $x' = u' - v'$ satisfies $Ax' = y$ and x' is therefore admissible in (P_1) . As x is the solution of (P_1) , we get

$$\sum_{j=1}^n (u_j + v_j) = \|x\|_1 < \|x'\|_1 = \sum_{j=1}^n |u'_j - v'_j| \leq \sum_{j=1}^n (u'_j + v'_j).$$

If, on the other hand, the pair (u, v) is the unique solution of (1.7), then $x = u - v$ is the unique solution of (P_1) . If namely z is another admissible vector in (P_1) , then $u' = z^+$ and $v' = z^-$ are admissible in (1.7) and we obtain

$$\|x\|_1 = \sum_{j=1}^n |u_j - v_j| \leq \sum_{j=1}^n (u_j + v_j) < \sum_{j=1}^n (u'_j + v'_j) = \|z\|_1.$$

Very similar argument works also in the case when (P_1) has multiple solutions.

1.3.2 Null Space Property

If $T \subset \{1, \dots, n\}$, then we denote by $T^c = \{1, \dots, n\} \setminus T$ the complement of T in $\{1, \dots, n\}$. If furthermore $v \in \mathbb{R}^n$, then we denote by v_T either the vector in $\mathbb{R}^{|T|}$, which contains the coordinates of v on T , or the vector in \mathbb{R}^n , which equals v on T and is zero on T^c . It will be always clear from the context, which notation is being used.

Finally, if $A \in \mathbb{R}^{m \times n}$ is a matrix, we denote by A_T the $m \times |T|$ sub-matrix containing the columns of A indexed by T . Let us observe that if $x \in \mathbb{R}^n$ with $T = \text{supp}(x)$, that $Ax = A_T x_T$.

We start the discussion of the properties of basis pursuit by introducing the notion of Null Space Property, which first appeared in [20].

Definition 1. Let $A \in \mathbb{R}^{m \times n}$ and let $k \in \{1, \dots, n\}$. Then A is said to have the *Null Space Property* (NSP) of order k if

$$\|v_T\|_1 < \|v_{T^c}\|_1 \quad \text{for all } v \in \ker A \setminus \{0\} \text{ and all } T \subset \{1, \dots, n\} \text{ with } |T| \leq k. \quad (1.8)$$

- Remark 1.* (i) The condition (1.8) states that vectors from the kernel of A are well spread, i.e. not supported on a set of small size. Indeed, if $v \in \mathbb{R}^n \setminus \{0\}$ is k -sparse and $T = \text{supp}(v)$, then (1.8) shows immediately, that v cannot lie in the kernel of A .
(ii) If we add $\|v_{T^c}\|_1$ to both sides of (1.8), we obtain $\|v\|_1 < 2\|v_{T^c}\|_1$. If then T are the indices of the k largest coordinates of v taken in the absolute value, this inequality becomes $\|v\|_1 < 2\sigma_k(v)_1$.

Theorem 2. Let $A \in \mathbb{R}^{m \times n}$ and let $k \in \{1, \dots, n\}$. Then every k -sparse vector x is the unique solution of (P_1) with $y = Ax$ if, and only if, A has the NSP of order k .

Proof. Let us assume that every k -sparse vector x is the unique solution of (P_1) with $y = Ax$. Let $v \in \ker A \setminus \{0\}$ and let $T \subset \{1, \dots, n\}$ with $|T| \leq k$ be arbitrary. Then v_T is k -sparse, and is therefore the unique solution of

$$\min_z \|z\|_1, \quad \text{s.t.} \quad Az = Av_T. \quad (1.9)$$

As $A(-v_{T^c}) = A(v - v_{T^c}) = A(v_T)$, this gives especially $\|v_T\|_1 < \|v_{T^c}\|_1$ and A has the NSP of order k .

Let us, on the other hand, assume that A has the NSP of order k . Let $x \in \mathbb{R}^n$ be a k -sparse vector and let $T = \text{supp}(x)$. We have to show that $\|x\|_1 < \|z\|_1$ for every $z \in \mathbb{R}^n$ different from x with $Az = Ax$. But this follows easily by using (1.8) for the vector $(x - z) \in \ker A \setminus \{0\}$

$$\begin{aligned} \|x\|_1 &\leq \|x - z_T\|_1 + \|z_T\|_1 = \|(x - z)_T\|_1 + \|z_T\|_1 < \|(x - z)_{T^c}\|_1 + \|z_T\|_1 \\ &= \|z_{T^c}\|_1 + \|z_T\|_1 = \|z\|_1. \end{aligned}$$

□

Remark 2. Theorem 2 states that the solutions of (P_0) may be found by (P_1) , if A has the NSP of order k and if $y \in \mathbb{R}^m$ is such that, there exists a k -sparse solution x of the equation $Ax = y$. Indeed, if in such a case, \hat{x} is a solution of (P_0) , then $\|\hat{x}\|_0 \leq \|x\|_0 \leq k$. Finally, it follows by Theorem 2 that \hat{x} is also a solution of (P_1) and that $x = \hat{x}$.

In the language of complexity theory, if we restrict the inputs of the ℓ_0 -minimization problem to matrices with the NSP of order k and to vectors y , for which there is a k -sparse solution of the equation $Ax = y$, the problem belongs to the P-class and the solving algorithm with polynomial running time is any standard algorithm solving (P_1) , or the corresponding linear problem (1.7).

1.3.3 Restricted Isometry Property

Although the Null Space Property is equivalent to the recovery of sparse solutions of underdetermined linear systems by basis pursuit in the sense just described, it is somehow difficult to construct matrices satisfying this property. We shall therefore present a sufficient condition called Restricted Isometry Property, which was first introduced in [15], and which ensures that the Null Space Property is satisfied.

Definition 2. Let $A \in \mathbb{R}^{m \times n}$ and let $k \in \{1, \dots, n\}$. Then the *restricted isometry constant* $\delta_k = \delta_k(A)$ of A of order k is the smallest $\delta \geq 0$, such that

$$(1 - \delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta)\|x\|_2^2 \quad \text{for all } x \in \Sigma_k. \quad (1.10)$$

Furthermore, we say that A satisfies the *Restricted Isometry Property* (RIP) of order k with the constant δ_k if $\delta_k < 1$.

Remark 3. The condition (1.10) states that A acts nearly isometrically when restricted to vectors from Σ_k . Of course, the smaller the constant $\delta_k(A)$ is, the closer is the matrix A to isometry on Σ_k . We will be therefore later interested in constructing matrices with small RIP constants. Finally, the inequality $\delta_1(A) \leq \delta_2(A) \leq \dots \leq \delta_k(A)$ follows trivially.

The following theorem shows that RIP of sufficiently high order with a constant small enough is indeed a sufficient condition for NSP.

Theorem 3. Let $A \in \mathbb{R}^{m \times n}$ and let k be a natural number with $k \leq n/2$. If $\delta_{2k}(A) < 1/3$, then A has the NSP of order k .

Proof. Let $v \in \ker A$ and let $T \subset \{1, \dots, n\}$ with $|T| \leq k$. We shall show that

$$\|v_T\|_2 \leq \frac{\delta_{2k}}{1 - \delta_k} \cdot \frac{\|v\|_1}{\sqrt{k}}. \quad (1.11)$$

If $\delta_k \leq \delta_{2k} < 1/3$, then Hölder's inequality gives immediately $\|v_T\|_1 \leq \sqrt{k}\|v_T\|_2 < \|v\|_1/2$ and the NSP of A of order k follows.

Before we come to the proof of (1.11), let us make the following observation. If $x, z \in \Sigma_k$ are two vectors with disjoint supports and $\|x\|_2 = \|z\|_2 = 1$, then $x \pm z \in \Sigma_{2k}$ and $\|x \pm z\|_2^2 = 2$. If we now combine the RIP of A

$$2(1 - \delta_{2k}) \leq \|A(x \pm z)\|_2^2 \leq 2(1 + \delta_{2k})$$

with the polarization identity, we get

$$|\langle Ax, Az \rangle| = \frac{1}{4} \left| \|Ax + Az\|_2^2 - \|Ax - Az\|_2^2 \right| \leq \delta_{2k}.$$

Using this formula for $x' = x/\|x\|_2$ and $z' = z/\|z\|_2$, we see that if A has the RIP of order $2k$ and $x, z \in \Sigma_k$ have disjoint supports, then

$$|\langle Ax, Az \rangle| \leq \delta_{2k} \|x\|_2 \|z\|_2. \quad (1.12)$$

To show (1.11), let us assume that $v \in \ker A$ is fixed. It is enough to consider $T = T_0$ the set of the k largest entries of v taken in the absolute value. Furthermore, we denote by T_1 the set of k largest entries of $v_{T_0^c}$ in the absolute value, by T_2 the set of k largest entries of $v_{(T_0 \cup T_1)^c}$ in the absolute value, etc. Using $0 = Av = A(v_{T_0} + v_{T_1} + v_{T_2} + \dots)$ and (1.12), we arrive at

$$\begin{aligned} \|v_{T_0}\|_2^2 &\leq \frac{1}{1 - \delta_k} \|Av_{T_0}\|_2^2 = \frac{1}{1 - \delta_k} \langle Av_{T_0}, A(-v_{T_1}) + A(-v_{T_2}) + \dots \rangle \\ &\leq \frac{1}{1 - \delta_k} \sum_{j \geq 1} |\langle Av_{T_0}, Av_{T_j} \rangle| \leq \frac{\delta_{2k}}{1 - \delta_k} \sum_{j \geq 1} \|v_{T_0}\|_2 \cdot \|v_{T_j}\|_2. \end{aligned}$$

We divide this inequality by $\|v_{T_0}\|_2 \neq 0$ and obtain

$$\|v_{T_0}\|_2 \leq \frac{\delta_{2k}}{1 - \delta_k} \sum_{j \geq 1} \|v_{T_j}\|_2.$$

The proof is then completed by the following simple chain of inequalities, which involve only the definition of the sets T_j , $j \geq 0$.

$$\begin{aligned} \sum_{j \geq 1} \|v_{T_j}\|_2 &= \sum_{j \geq 1} \left(\sum_{l \in T_j} |v_l|^2 \right)^{1/2} \leq \sum_{j \geq 1} \left(k \max_{l \in T_j} |v_l|^2 \right)^{1/2} \\ &= \sum_{j \geq 1} \sqrt{k} \max_{l \in T_j} |v_l| \leq \sum_{j \geq 1} \sqrt{k} \min_{l \in T_{j-1}} |v_l| \leq \sum_{j \geq 1} \sqrt{k} \cdot \frac{\sum_{l \in T_{j-1}} |v_l|}{k} \\ &= \sum_{j \geq 1} \frac{\|v_{T_{j-1}}\|_1}{\sqrt{k}} = \frac{\|v\|_1}{\sqrt{k}}. \end{aligned} \quad (1.13)$$

□

Combining Theorems 2 and 3, we obtain immediately the following corollary.

Corollary 1. Let $A \in \mathbb{R}^{m \times n}$ and let k be a natural number with $k \leq n/2$. If $\delta_{2k}(A) < 1/3$, then every k -sparse vector x is the unique solution of (P_1) with $y = Ax$.

1.3.4 RIP for random matrices

From what was said up to now, we know that matrices with small restricted isometry constants fulfill the null space property, and sparse solutions of underdetermined linear equations involving such matrices can be found by ℓ_1 -minimization (P_1) . We discuss in this chapter a class of matrices with small RIP constants. It turns out that the most simple way is to construct these matrices by taking its entries to be independent standard normal variables.

We denote until the end of this section

$$A = \frac{1}{\sqrt{m}} \begin{pmatrix} \omega_{1,1} & \dots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{m1} & \dots & \omega_{mn} \end{pmatrix}, \quad (1.14)$$

where $\omega_{ij}, i = 1, \dots, m, j = 1, \dots, n$, are i.i.d. standard normal variables. We shall show that such a matrix satisfies the RIP with reasonably small constants with high probability.

1.3.4.1 Concentration inequalities

Before we come to the main result of this chapter, we need some properties of independent standard normal variables.

- Lemma 1.** (i) Let ω be a standard normal variable. Then $\mathbb{E}(e^{\lambda\omega^2}) = 1/\sqrt{1-2\lambda}$ for $-\infty < \lambda < 1/2$.
(ii) (2-stability of the normal distribution) Let $m \in \mathbb{N}$, let $\lambda = (\lambda_1, \dots, \lambda_m) \in \mathbb{R}^m$ and let $\omega_1, \dots, \omega_m$ be i.i.d. standard normal variables. Then $\lambda_1\omega_1 + \dots + \lambda_m\omega_m \sim (\sum_{i=1}^m \lambda_i^2)^{1/2} \cdot \mathcal{N}(0, 1)$, i.e. it is equidistributed with a multiple of a standard normal variable.

Proof. The proof of (i) follows from the substitution $s := \sqrt{1-2\lambda} \cdot t$ in the following way.

$$\begin{aligned} \mathbb{E}(e^{\lambda\omega^2}) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda t^2} \cdot e^{-t^2/2} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{(\lambda-1/2)t^2} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-s^2/2} \cdot \frac{ds}{\sqrt{1-2\lambda}} = \frac{1}{\sqrt{1-2\lambda}}. \end{aligned}$$

Although the property (ii) is very well known (and there are several different ways to prove it), we provide a simple geometric proof for the sake of completeness. It is enough to consider the case $m = 2$. The general case then follows by induction.

Let therefore $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^2, \lambda \neq 0$, be fixed and let ω_1 and ω_2 be i.i.d. standard normal random variables. We put $S := \lambda_1 \omega_1 + \lambda_2 \omega_2$. Let $t \geq 0$ be an arbitrary non-negative real number. We calculate

$$\begin{aligned}\mathbb{P}(S \leq t) &= \frac{1}{2\pi} \int_{(u,v): \lambda_1 u + \lambda_2 v \leq t} e^{-(u^2+v^2)/2} du dv = \frac{1}{2\pi} \int_{u \leq c; v \in \mathbb{R}} e^{-(u^2+v^2)/2} du dv \\ &= \frac{1}{\sqrt{2\pi}} \int_{u \leq c} e^{-u^2/2} du.\end{aligned}$$

We have used the rotational invariance of the function $(u, v) \rightarrow e^{-(u^2+v^2)/2}$. The value of c is given by the distance of the origin from the line $\{(u, v) : \lambda_1 u + \lambda_2 v = t\}$. It follows by elementary geometry and Pythagorean theorem that (cf. $\Delta OAP \simeq \Delta BAO$ in Figure 1.3)

$$c = |OP| = |OB| \cdot \frac{|OA|}{|AB|} = \frac{t}{\sqrt{\lambda_1^2 + \lambda_2^2}}.$$

We therefore get

$$\mathbb{P}(S \leq t) = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{\lambda_1^2 + \lambda_2^2} \cdot u \leq t} e^{-u^2/2} du = \mathbb{P}\left(\sqrt{\lambda_1^2 + \lambda_2^2} \cdot \omega \leq t\right).$$

The same estimate holds for negative t 's by symmetry and the proof is finished. \square

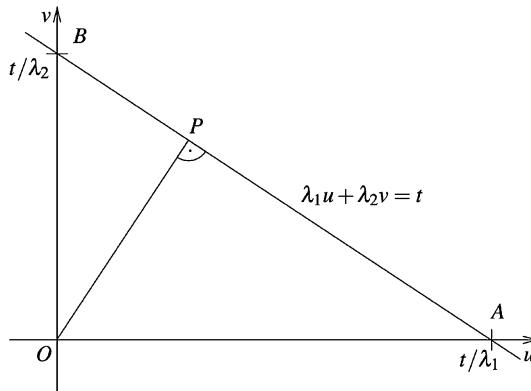


Fig. 1.3 Calculating $c = |OP|$ by elementary geometry for $\lambda_1, \lambda_2 > 0$

If $\omega_1, \dots, \omega_m$ are (possibly dependent) standard normal random variables, then $\mathbb{E}(\omega_1^2 + \dots + \omega_m^2) = m$. If $\omega_1, \dots, \omega_m$ are even independent, then the value of $\omega_1^2 + \dots + \omega_m^2$ concentrates very strongly around m . This effect is known as *concentration of measure*, cf. [49, 50, 55].

Lemma 2. *Let $m \in \mathbb{N}$ and let $\omega_1, \dots, \omega_m$ be i.i.d. standard normal variables. Let $0 < \varepsilon < 1$. Then*

$$\mathbb{P}(\omega_1^2 + \dots + \omega_m^2 \geq (1 + \varepsilon)m) \leq e^{-\frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]}$$

and

$$\mathbb{P}(\omega_1^2 + \dots + \omega_m^2 \leq (1 - \varepsilon)m) \leq e^{-\frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]}.$$

Proof. We prove only the first inequality. The second one follows in exactly the same manner. Let us put $\beta := 1 + \varepsilon > 1$ and calculate

$$\begin{aligned} \mathbb{P}(\omega_1^2 + \dots + \omega_m^2 \geq \beta m) &= \mathbb{P}(\omega_1^2 + \dots + \omega_m^2 - \beta m \geq 0) \\ &= \mathbb{P}(\lambda(\omega_1^2 + \dots + \omega_m^2 - \beta m) \geq 0) \\ &= \mathbb{P}(\exp(\lambda(\omega_1^2 + \dots + \omega_m^2 - \beta m)) \geq 1) \\ &\leq \mathbb{E} \exp(\lambda(\omega_1^2 + \dots + \omega_m^2 - \beta m)), \end{aligned}$$

where $\lambda > 0$ is a positive real number, which shall be chosen later on. We have used the Markov's inequality (1.3) in the last step. Further we use the elementary properties of exponential function and (1.5) for the independent variables $\omega_1, \dots, \omega_m$. This leads to

$$\mathbb{E} \exp(\lambda(\omega_1^2 + \dots + \omega_m^2 - \beta m)) = e^{-\lambda\beta m} \cdot \mathbb{E} e^{\lambda\omega_1^2} \dots e^{\lambda\omega_m^2} = e^{-\lambda\beta m} \cdot (\mathbb{E} e^{\lambda\omega_1^2})^m$$

and with the help of Lemma 1 we get finally (for $0 < \lambda < 1/2$)

$$\mathbb{E} \exp(\lambda(\omega_1^2 + \dots + \omega_m^2 - \beta m)) = e^{-\lambda\beta m} \cdot (1 - 2\lambda)^{-m/2}.$$

We now look for the value of $0 < \lambda < 1/2$, which would minimize the last expression. Therefore, we take the derivative of $e^{-\lambda\beta m} \cdot (1 - 2\lambda)^{-m/2}$ and put it equal to zero. After a straightforward calculation, we get

$$\lambda = \frac{1 - 1/\beta}{2},$$

which obviously satisfies also $0 < \lambda < 1/2$. Using this value of λ we obtain

$$\begin{aligned}\mathbb{P}(\omega_1^2 + \dots + \omega_m^2 \geq \beta m) &\leq e^{-\frac{1-1/\beta}{2} \cdot \beta m} \cdot (1 - (1 - 1/\beta))^{-m/2} = e^{-\frac{\beta-1}{2}m} \cdot \beta^{m/2} \\ &= e^{-\frac{\varepsilon m}{2}} \cdot e^{\frac{m}{2} \ln(1+\varepsilon)}.\end{aligned}$$

The result then follows from the inequality

$$\ln(1+t) \leq t - \frac{t^2}{2} + \frac{t^3}{3}, \quad -1 < t < 1. \quad \square$$

Using 2-stability of the normal distribution, Lemma 2 shows immediately that A defined as in (1.14) acts with high probability as isometry on one fixed $x \in \mathbb{R}^n$.

Theorem 4. *Let $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$ and let A be as in (1.14). Then*

$$\mathbb{P}\left(\left|\|Ax\|_2^2 - 1\right| \geq t\right) \leq 2e^{-\frac{m}{2}[t^2/2-t^3/3]} \leq 2e^{-Cmt^2} \quad (1.15)$$

for $0 < t < 1$ with an absolute constant $C > 0$.

Proof. Let $x = (x_1, x_2, \dots, x_n)^T$. Then we get by the 2-stability of normal distribution and Lemma 2

$$\begin{aligned}\mathbb{P}\left(\left|\|Ax\|_2^2 - 1\right| \geq t\right) &= \mathbb{P}\left(\left|(\omega_{1,1}x_1 + \dots + \omega_{1n}x_n)^2 + \dots + (\omega_{m1}x_1 + \dots + \omega_{mn}x_n)^2 - m\right| \geq mt\right) \\ &= \mathbb{P}\left(\left|\omega_1^2 + \dots + \omega_m^2 - m\right| \geq mt\right) \\ &= \mathbb{P}\left(\omega_1^2 + \dots + \omega_m^2 \geq m(1+t)\right) + \mathbb{P}\left(\omega_1^2 + \dots + \omega_m^2 \leq m(1-t)\right) \\ &\leq 2e^{-\frac{m}{2}[t^2/2-t^3/3]}.\end{aligned}$$

This gives the first inequality in (1.15). The second one follows by simple algebraic manipulations (for $C = 1/12$). \square

Remark 4. (i) Observe that (1.15) may be easily rescaled to

$$\mathbb{P}\left(\left|\|Ax\|_2^2 - \|x\|_2^2\right| \geq t\|x\|_2^2\right) \leq 2e^{-Cmt^2}, \quad (1.16)$$

which is true for every $x \in \mathbb{R}^n$.

- (ii) A slightly different proof of (1.15) is based on the rotational invariance of the distribution underlying the random structure of matrices defined by (1.14). Therefore, it is enough to prove (1.15) only for one fixed element $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$. Taking $x = e_1 = (1, 0, \dots, 0)^T$ to be the first canonical unit vector allows us to use Lemma 2 without the necessity of applying the 2-stability of normal distribution.

1.3.4.2 RIP for random Gaussian matrices

The proof of restricted isometry property of random matrices generated as in (1.14) is based on two main ingredients. The first is the concentration of measure phenomenon described in its most simple form in Lemma 2, and reformulated in Theorem 4. The second is the following entropy argument, which allows to extend Theorem 4 and (1.15) from one fixed $x \in \mathbb{R}^n$ to the set Σ_k of all k -sparse vectors.

Lemma 3. *Let $t > 0$. Then there is a set $\mathcal{N} \subset \mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ with*

- (i) $|\mathcal{N}| \leq (1 + 2/t)^n$ and
- (ii) for every $z \in \mathbb{S}^{n-1}$, there is a $x \in \mathcal{N}$ with $\|x - z\|_2 \leq t$.

Proof. Choose any $x^1 \in \mathbb{S}^{n-1}$. If $x^1, \dots, x^j \in \mathbb{S}^{n-1}$ were already chosen, take $x^{j+1} \in \mathbb{S}^{n-1}$ arbitrarily with $\|x^{j+1} - x^l\|_2 > t$ for all $l = 1, \dots, j$. This process is then repeated as long as possible, i.e. until we obtain a set $\mathcal{N} = \{x^1, \dots, x^N\} \subset \mathbb{S}^{n-1}$, such that for every $z \in \mathbb{S}^{n-1}$ there is a $j \in \{1, \dots, N\}$ with $\|x^j - z\|_2 \leq t$. This gives the property (ii).

We will use volume arguments to prove (i). It follows by construction that $\|x^i - x^j\|_2 > t$ for every $i, j \in \{1, \dots, N\}$ with $i \neq j$. By triangle inequality, the balls $B(x^j, t/2)$ are all disjoint and are all included in the ball with the center in the origin and radius $1 + t/2$. By comparing the volumes we get

$$N \cdot (t/2)^n \cdot V \leq (1 + t/2)^n \cdot V,$$

where V is the volume of the unit ball in \mathbb{R}^n . Hence, we get $N = |\mathcal{N}| \leq (1 + 2/t)^n$. □

With all these tools at hand, we can now state the main theorem of this section, whose proof follows closely the arguments of [4].

Theorem 5. *Let $n \geq m \geq k \geq 1$ be natural numbers and let $0 < \varepsilon < 1$ and $0 < \delta < 1$ be real numbers with*

$$m \geq C\delta^{-2} \left(k \ln(en/k) + \ln(2/\varepsilon) \right), \quad (1.17)$$

where $C > 0$ is an absolute constant. Let A be again defined by (1.14). Then

$$\mathbb{P}(\delta_k(A) \leq \delta) \geq 1 - \varepsilon.$$

Proof. The proof follows by the concentration inequality of Theorem 4 and the entropy argument described in Lemma 3. By this lemma, there is a set

$$\mathcal{N} \subset Z := \{z \in \mathbb{R}^n : \text{supp}(z) \subset \{1, \dots, k\}, \|z\|_2 = 1\},$$

such that

- (i) $|\mathcal{N}| \leq 9^k$ and
- (ii) $\min_{x \in \mathcal{N}} \|z - x\|_2 \leq 1/4$ for every $z \in Z$.

We show that if $|\|Ax\|_2^2 - 1| \leq \delta/2$ for all $x \in \mathcal{N}$, then $|\|Az\|_2^2 - 1| \leq \delta$ for all $z \in Z$.

We proceed by the following bootstrap argument. Let $\gamma > 0$ be the smallest number, such that $|\|Az\|_2^2 - 1| \leq \gamma$ for all $z \in Z$. Then $|\|Au\|_2^2 - \|u\|_2^2| \leq \gamma \|u\|_2^2$ for all $u \in \mathbb{R}^n$ with $\text{supp}(u) \subset \{1, \dots, k\}$. Let us now assume that $\|u\|_2 = \|v\|_2 = 1$ with $\text{supp}(u) \cup \text{supp}(v) \subset \{1, \dots, k\}$. Then we get by polarization identity

$$\begin{aligned} |\langle Au, Av \rangle - \langle u, v \rangle| &= \frac{1}{4} \left| (\|A(u+v)\|_2^2 - \|A(u-v)\|_2^2) - (\|u+v\|_2^2 - \|u-v\|_2^2) \right| \\ &\leq \frac{1}{4} \left| \|A(u+v)\|_2^2 - \|u+v\|_2^2 \right| + \frac{1}{4} \left| \|A(u-v)\|_2^2 - \|u-v\|_2^2 \right| \\ &\leq \frac{\gamma}{4} \|u+v\|_2^2 + \frac{\gamma}{4} \|u-v\|_2^2 = \frac{\gamma}{2} (\|u\|_2^2 + \|v\|_2^2) = \gamma. \end{aligned}$$

Applying this inequality to $u' = u/\|u\|_2$ and $v' = v/\|v\|_2$, we obtain

$$|\langle Au, Av \rangle - \langle u, v \rangle| \leq \gamma \|u\|_2 \|v\|_2 \quad (1.18)$$

for all $u, v \in \mathbb{R}^n$ with $\text{supp}(u) \cup \text{supp}(v) \subset \{1, \dots, k\}$.

Let now again $z \in Z$. Then there is an $x \in \mathcal{N}$, such that $\|z-x\|_2 \leq 1/4$. We obtain by triangle inequality and (1.18)

$$\begin{aligned} |\|Az\|_2^2 - 1| &= |\|Ax\|_2^2 - 1 + \langle A(z+x), A(z-x) \rangle - \langle z+x, z-x \rangle| \\ &\leq \delta/2 + \gamma \|z+x\|_2 \|z-x\|_2 \leq \delta/2 + \gamma/2. \end{aligned}$$

As the supremum of the left-hand side over all admissible z 's is equal to γ , we obtain that $\gamma \leq \delta$ and the statement follows.

Equipped with this tool, the rest of the proof follows by a simple union bound.

$$\begin{aligned}
\mathbb{P}(\delta_k(A) > \delta) &\leq \sum_{\substack{T \subset \{1, \dots, n\} \\ |T| \leq k}} \mathbb{P}\left(\exists z \in \mathbb{R}^n : \text{supp}(z) \subset T, \|z\|_2 = 1 \text{ and } |\|Az\|_2^2 - 1| > \delta\right) \\
&= \binom{n}{k} \mathbb{P}\left(\exists z \in Z \text{ with } |\|Az\|_2^2 - 1| > \delta\right) \\
&\leq \binom{n}{k} \mathbb{P}\left(\exists x \in \mathcal{N} : |\|Ax\|_2^2 - 1| > \delta/2\right).
\end{aligned}$$

By Theorem 4, the last probability may be estimated from above by $2e^{-C'm\delta^2}$. Hence we obtain

$$\mathbb{P}(\delta_k(A) > \delta) \leq 9^k \binom{n}{k} \cdot 2e^{-C'm\delta^2}$$

Hence it is enough to show that the last quantity is at most ε if (1.17) is satisfied. But this follows by straightforward algebraic manipulations and the well-known estimate

$$\binom{n}{k} \leq \frac{n^k}{k!} \leq \left(\frac{en}{k}\right)^k.$$

□

1.3.4.3 Lemma of Johnson and Lindenstrauss

Concentration inequalities similar to (1.15) play an important role in several areas of mathematics. We shall present their connection to the famous result from functional analysis called Johnson–Lindenstrauss lemma, cf. [1, 22, 46, 54]. The lemma states that a set of points in a high-dimensional space can be embedded into a space of much lower dimension in such a way that the mutual distances between the points are nearly preserved. The connection between this classical result and compressed sensing was first highlighted in [4], cf. also [47].

Lemma 4. *Let $0 < \varepsilon < 1$ and let m, N and n be natural numbers with*

$$m \geq 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \ln N.$$

Then for every set $\{x^1, \dots, x^N\} \subset \mathbb{R}^n$ there exists a mapping $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, such that

$$(1 - \varepsilon) \|x^i - x^j\|_2^2 \leq \|f(x^i) - f(x^j)\|_2^2 \leq (1 + \varepsilon) \|x^i - x^j\|_2^2, \quad i, j \in \{1, \dots, N\}. \quad (1.19)$$

Proof. We put $f(x) = Ax$, where again

$$Ax = \frac{1}{\sqrt{m}} \begin{pmatrix} \omega_{1,1} & \dots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{m1} & \dots & \omega_{mn} \end{pmatrix} x,$$

and $\omega_{ij}, i = 1, \dots, m, j = 1, \dots, n$ are i.i.d. standard normal variables. We show that with this choice f satisfies (1.19) with positive probability. This proves the existence of such a mapping.

Let $i, j \in \{1, \dots, N\}$ arbitrary with $x^i \neq x^j$. Then we put $z = \frac{x^i - x^j}{\|x^i - x^j\|_2}$ and evaluate the probability that the right-hand side inequality in (1.19) does not hold. Theorem 4 then implies

$$\begin{aligned} \mathbb{P}\left(\left|\|f(x^i) - f(x^j)\|_2^2 - \|x^i - x^j\|_2^2\right| > \varepsilon \|x^i - x^j\|_2^2\right) &= \mathbb{P}\left(\left|\|Az\|_2^2 - 1\right| > \varepsilon\right) \\ &\leq 2e^{-\frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]}. \end{aligned}$$

The same estimate is also true for all $\binom{N}{2}$ pairs $\{i, j\} \subset \{1, \dots, N\}$ with $i \neq j$. The probability that one of the inequalities in (1.19) is not satisfied is therefore at most

$$2 \cdot \binom{N}{2} \cdot e^{-\frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]} < N^2 \cdot e^{-\frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]} = \exp\left(2 \ln N - \frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]\right) \leq e^0 = 1$$

for $m \geq 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \ln N$. Therefore, the probability that (1.19) holds for all $i, j \in \{1, \dots, N\}$ is positive and the result follows. \square

1.3.5 Stability and Robustness

The ability to recover sparse solutions of underdetermined linear systems by quick recovery algorithms as ℓ_1 -minimization is surely a very promising result. On the other hand, two additional features are obviously necessary to extend this results to real-life applications, namely

- **Stability:** We want to be able to recover (or at least approximate) also vectors $x \in \mathbb{R}^n$, which are not exactly sparse. Such vectors are called *compressible* and mathematically they are characterized by the assumption that their best k -term approximation decays rapidly with k . Intuitively, the faster the decay of the best k -term approximation of $x \in \mathbb{R}^n$ is, the better we should be able to approximate x .
- **Robustness:** Equally important, we want to recover sparse or compressible vectors from noisy measurements. The basic model here is the assumptions that the

measurement vector y is given by $y = Ax + e$, where e is small (in some sense). Again, the smaller the error e is, the better we should be able to recover an approximation of x .

We shall show that the methods of compressed sensing can be extended also to this kind of scenario. There is a number of different estimates in the literature, which show that the technique of compressed sensing is stable and robust. We will present only one of them (with more to come in Section 1.4.3). Its proof is a modification of the proof of Theorem 3, and follows closely [11].

Inspired by the form of the noisy measurements just described, we will concentrate on the recovery properties of the following slight modification of (P_1) . Namely, let $\eta \geq 0$, then we consider the convex optimization problem

$$\min_{z \in \mathbb{R}^n} \|z\|_1 \quad \text{s.t.} \quad \|Az - y\|_2 \leq \eta. \quad (P_{1,\eta})$$

If $\eta = 0$, $(P_{1,\eta})$ reduces back to (P_1) .

Theorem 6. *Let $\delta_{2k} < \sqrt{2} - 1$ and $\|e\|_2 \leq \eta$. Then the solution \hat{x} of $(P_{1,\eta})$ satisfies*

$$\|x - \hat{x}\|_2 \leq \frac{C\sigma_k(x)_1}{\sqrt{k}} + D\eta, \quad (1.20)$$

where $C, D > 0$ are two universal positive constants.

Proof. First, let us recall that if A has RIP of order $2k$ and $u, v \in \Sigma_k$ are two vectors with disjoint supports, then we have by (1.12)

$$|\langle Au, Av \rangle| \leq \delta_{2k} \|u\|_2 \|v\|_2. \quad (1.21)$$

Let us put $h = \hat{x} - x$ and let us define the index set $T_0 \subset \{1, \dots, n\}$ as the locations of k largest entries of x taken in the absolute value. Furthermore, we define $T_1 \subset T_0^c$ to be the indices of k largest absolute entries of $h_{T_0^c}$, T_2 the indices of k largest absolute entries of $h_{(T_0 \cup T_1)^c}$, etc. As \hat{x} is an admissible point in $(P_{1,\eta})$, the triangle inequality gives

$$\|Ah\|_2 = \|A(x - \hat{x})\|_2 \leq \|Ax - y\|_2 + \|y - A\hat{x}\|_2 \leq 2\eta. \quad (1.22)$$

As \hat{x} is the minimizer of $(P_{1,\eta})$, we get $\|\hat{x}\|_1 = \|x + h\|_1 \leq \|x\|_1$, which we use to show that h must be small outside of T_0 . Indeed, we obtain

$$\begin{aligned} \|h_{T_0^c}\|_1 &= \|(x + h)_{T_0^c} - x_{T_0^c}\|_1 + \|(x + h)_{T_0} - h_{T_0}\|_1 - \|x_{T_0}\|_1 \\ &\leq \|(x + h)_{T_0^c}\|_1 + \|x_{T_0^c}\|_1 + \|(x + h)_{T_0}\|_1 + \|h_{T_0}\|_1 - \|x_{T_0}\|_1 \\ &= \|x + h\|_1 + \|x_{T_0^c}\|_1 + \|h_{T_0}\|_1 - \|x_{T_0}\|_1 \end{aligned}$$

$$\begin{aligned} &\leq \|x\|_1 + \|x_{T_0^c}\|_1 + \|h_{T_0}\|_1 - \|x_{T_0}\|_1 \\ &= \|h_{T_0}\|_1 + 2\|x_{T_0^c}\| \leq k^{1/2}\|h_{T_0}\|_2 + 2\sigma_k(x)_1. \end{aligned}$$

Using this together with the approach applied already in (1.13), we derive

$$\sum_{j \geq 2} \|h_{T_j}\|_2 \leq k^{-1/2}\|h_{T_0^c}\|_1 \leq \|h_{T_0}\|_2 + 2k^{-1/2}\sigma_k(x)_1. \quad (1.23)$$

We use the RIP property of A , (1.21), (1.22), (1.23) and the simple inequality $\|h_{T_0}\|_2 + \|h_{T_1}\|_2 \leq \sqrt{2}\|h_{T_0 \cup T_1}\|_2$ and get

$$\begin{aligned} (1 - \delta_{2k})\|h_{T_0 \cup T_1}\|_2^2 &\leq \|Ah_{T_0 \cup T_1}\|_2^2 = \langle Ah_{T_0 \cup T_1}, Ah \rangle - \langle Ah_{T_0 \cup T_1}, \sum_{j \geq 2} Ah_{T_j} \rangle \\ &\leq \|Ah_{T_0 \cup T_1}\|_2 \|Ah\|_2 + \sum_{j \geq 2} |\langle Ah_{T_0}, Ah_{T_j} \rangle| + \sum_{j \geq 2} |\langle Ah_{T_1}, Ah_{T_j} \rangle| \\ &\leq 2\eta\sqrt{1 + \delta_{2k}}\|h_{T_0 \cup T_1}\|_2 + \delta_{2k}(\|h_{T_0}\|_2 + \|h_{T_1}\|_2) \sum_{j \geq 2} \|h_{T_j}\|_2 \\ &\leq \|h_{T_0 \cup T_1}\|_2 \left(2\eta\sqrt{1 + \delta_{2k}} + \sqrt{2}\delta_{2k}\|h_{T_0}\|_2 + 2\sqrt{2}\delta_{2k}k^{-1/2}\sigma_k(x)_1 \right). \end{aligned}$$

We divide this inequality with $(1 - \delta_{2k})\|h_{T_0 \cup T_1}\|_2$, replace $\|h_{T_0}\|_2$ with the larger quantity $\|h_{T_0 \cup T_1}\|_2$ and subtract $\sqrt{2}\delta_{2k}/(1 - \delta_{2k})\|h_{T_0 \cup T_1}\|_2$ to arrive at

$$\|h_{T_0 \cup T_1}\|_2 \leq (1 - \rho)^{-1}(\alpha\eta + 2\rho k^{-1/2}\sigma_k(x)_1), \quad (1.24)$$

where

$$\alpha = \frac{2\sqrt{1 + \delta_{2k}}}{1 - \delta_{2k}} \quad \text{and} \quad \rho = \frac{\sqrt{2}\delta_{2k}}{1 - \delta_{2k}}. \quad (1.25)$$

We conclude the proof by using this estimate and (1.23)

$$\begin{aligned} \|h\|_2 &\leq \|h_{(T_0 \cup T_1)^c}\|_2 + \|h_{T_0 \cup T_1}\|_2 \leq \sum_{j \geq 2} \|h_{T_j}\|_2 + \|h_{T_0 \cup T_1}\|_2 \\ &\leq 2\|h_{T_0 \cup T_1}\|_2 + 2k^{-1/2}\sigma_k(x)_1 \leq C\frac{\sigma_k(x)_1}{\sqrt{k}} + D\eta \end{aligned}$$

with $C = 2(1 - \rho)^{-1}\alpha$ and $D = 2(1 + \rho)(1 - \rho)^{-1}$.

We shall give more details on stability and robustness of compressed sensing in Section 1.4.3.

1.3.6 Optimality of bounds

When recovering k -sparse vectors one obviously needs at least $m \geq k$ linear measurements. Even when the support of the unknown vector would be known, this number of measurements would be necessary to identify the value of the non-zero coordinates. Therefore, the dependence of the bound (1.17) on k can possibly only be improved in the logarithmic factor. We shall show that even that is not possible and that this dependence is already optimal as soon as a stable recovery of k -sparse vectors is requested. The approach presented here is essentially taken over from [40].

The proof is based on the following combinatorial lemma.

Lemma 5. *Let $k \leq n$ be two natural numbers. Then there are N subsets T_1, \dots, T_N of $\{1, \dots, n\}$, such that*

- (i) $N \geq \left(\frac{n}{4k}\right)^{k/2}$,
- (ii) $|T_i| = k$ for all $i = 1, \dots, N$ and
- (iii) $|T_i \cap T_j| < k/2$ for all $i \neq j$.

Proof. We may assume that $k \leq n/4$, otherwise one can take $N = 1$ and the statement becomes trivial. The main idea of the proof is straightforward (and similar to the proof of Lemma 3). We choose the sets T_1, T_2, \dots inductively one after another as long as possible, satisfying (ii) and (iii) on the way, and then we show that this process will run for at least N steps with N fulfilling (i).

Let $T_1 \subset \{1, \dots, n\}$ be any set with k elements. The number of subsets of $\{1, \dots, n\}$ with exactly k elements, whose intersection with T_1 has at least $k/2$ elements is bounded by the product of 2^k (i.e., the number of all subsets of T_1) and $\binom{n-k}{\lfloor k/2 \rfloor}$, which is the number of all subsets of T_1^c with at most $k/2$ elements. Therefore there are at least

$$\binom{n}{k} - 2^k \binom{n-k}{\lfloor k/2 \rfloor}$$

sets $T \subset \{1, \dots, n\}$ with k elements and $|T \cap T_1| < k/2$. We select T_2 to be any of them. After the j th step, we have selected sets T_1, \dots, T_j with (ii) and (iii) and there are still

$$\binom{n}{k} - j2^k \binom{n-k}{\lfloor k/2 \rfloor}$$

to choose from. The process stops if this quantity is not positive any more, i.e. after at least

$$\begin{aligned} N &\geq \frac{\binom{n}{k}}{2^k \binom{n-k}{\lfloor k/2 \rfloor}} \geq 2^{-k} \frac{\binom{n}{k}}{\binom{n-\lceil k/2 \rceil}{\lfloor k/2 \rfloor}} = 2^{-k} \frac{n!}{(n-k)!k!} \cdot \frac{(\lfloor k/2 \rfloor)!(n-k)!}{(n-\lceil k/2 \rceil)!} \\ &= 2^{-k} \frac{n(n-1)\dots(n-\lceil k/2 \rceil+1)}{k(k-1)\dots(k-\lceil k/2 \rceil+1)} \geq 2^{-k} \left(\frac{n}{k}\right)^{\lceil k/2 \rceil} \geq \left(\frac{n}{4k}\right)^{k/2} \end{aligned}$$

steps.

The following theorem shows that any stable recovery of sparse solutions requires at least m number of measurements, where m is of the order $k \ln(en/k)$.

Theorem 7. *Let $k \leq m \leq n$ be natural numbers, let $A \in \mathbb{R}^{m \times n}$ be a measurement matrix, and let $\Delta : \mathbb{R}^m \rightarrow \mathbb{R}^n$ be an arbitrary recovery map such that for some constant $C > 0$*

$$\|x - \Delta(Ax)\|_2 \leq C \frac{\sigma_k(x)_1}{\sqrt{k}} \quad \text{for all } x \in \mathbb{R}^n. \quad (1.26)$$

Then

$$m \geq C' k \ln(en/k) \quad (1.27)$$

with some other constant C' depending only on C .

Proof. We may assume that $C \geq 1$. Furthermore, if k is proportional to n (say $k \geq n/8$), then (1.27) becomes trivial. Hence we may also assume that $k \leq n/8$.

By Lemma 5, there exist index sets T_1, \dots, T_N with $N \geq (n/4k)^{k/2}$, $|T_i| = k$ and $|T_i \cap T_j| < k/2$ if $i \neq j$. We put $x_i = \chi_{T_i}/\sqrt{k}$. Then $\|x_i\|_2 = 1$, $\|x_i\|_1 = \sqrt{k}$ and $\|x_i - x_j\|_2 > 1$ for $i \neq j$.

Let

$$\mathcal{B} = \left\{ z \in \mathbb{R}^n : \|z\|_1 \leq \frac{\sqrt{k}}{4C} \quad \text{and} \quad \|z\|_2 \leq 1/4 \right\}.$$

Then $x_i \in 4C \cdot \mathcal{B}$ for all $i = 1, \dots, N$.

We claim that the sets $A(x_i + \mathcal{B})$ are mutually disjoint. Indeed, let us assume that this is not the case. Then there is a pair of indices $i, j \in \{1, \dots, n\}$ and $z, z' \in \mathcal{B}$ with $i \neq j$ and $A(x_i + z) = A(x_j + z')$. It follows that $\Delta(A(x_i + z)) = \Delta(A(x_j + z'))$ and we get a contradiction by

$$\begin{aligned} 1 &< \|x_i - x_j\|_2 = \|(x_i + z - \Delta(A(x_i + z)) - (x_j + z' - \Delta(A(x_j + z')) - z + z')\|_2 \\ &\leq \|(x_i + z - \Delta(A(x_i + z)))\|_2 + \|x_j + z' - \Delta(A(x_j + z'))\|_2 + \|z\|_2 + \|z'\|_2 \end{aligned}$$

$$\begin{aligned} &\leq C \frac{\sigma_k(x_i + z)_1}{\sqrt{k}} + C \frac{\sigma_k(x_j + z')_1}{\sqrt{k}} + \|z\|_2 + \|z'\|_2 \\ &\leq C \frac{\|z\|_1}{\sqrt{k}} + C \frac{\|z'\|_1}{\sqrt{k}} + \|z\|_2 + \|z'\|_2 \leq 1. \end{aligned}$$

Furthermore,

$$A(x_i + \mathcal{B}) \subset A((4C+1)\mathcal{B}), \quad i = 1, \dots, N$$

Let $d \leq m$ be the dimension of the range of A . We denote by $V \neq 0$ the d -dimensional volume of $A(\mathcal{B})$ and compare the volumes

$$\sum_{j=1}^N \text{vol}(A(x_j + \mathcal{B})) \leq \text{vol}(A((4C+1)\mathcal{B})).$$

Using linearity of A , we obtain

$$\left(\frac{n}{4k}\right)^{k/2} V \leq N \cdot V \leq (4C+1)^d V \leq (4C+1)^m V.$$

We divide by V and take the logarithm to arrive at

$$\frac{k}{2} \ln\left(\frac{n}{4k}\right) \leq m \ln(4C+1). \quad (1.28)$$

If $k \leq n/8$, then it is easy to check that there is a constant $c' > 0$, such that

$$\ln\left(\frac{n}{4k}\right) \geq c' \ln\left(\frac{en}{k}\right).$$

Putting this into (1.28) finishes the proof. \square

1.4 Extensions

Section 1.3 gives a detailed overview of the most important features of compressed sensing. On the other hand, inspired by many questions coming from application driven research, various additional aspects of the theory were studied in the literature. We present here few selected extensions of the ideas of compressed sensing, which turned out to be the most useful in practice. To keep the presentation reasonable short, we do not give any proofs, and only refer to relevant sources.

1.4.1 Frames and Dictionaries

We have considered in Section 1.3 vectors $x \in \mathbb{R}^n$, which are sparse with respect to the natural canonical basis $\{e_j\}_{j=1}^n$ of \mathbb{R}^n . In practice, however, the signal has a sparse representation with respect to a basis (or, more general, with respect to a frame or dictionary). Let us first recall some terminology.

A set of vectors $\{\phi_j\}_{j=1}^n$ in \mathbb{R}^n , which is linearly independent and which spans the whole space \mathbb{R}^n is called a basis. It follows easily that such a set necessarily has n elements. Furthermore, every $x \in \mathbb{R}^n$ can be expressed uniquely as a linear combination of the basis vectors, i.e. there is a unique $c = (c_1, \dots, c_n)^T \in \mathbb{R}^n$, such that

$$x = \sum_{j=1}^n c_j \phi_j. \quad (1.29)$$

A basis is called orthonormal, if it satisfies the orthogonality relations

$$\langle \phi_i, \phi_j \rangle = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \quad (1.30)$$

If $\{\phi_j\}_{j=1}^n$ is an orthonormal basis and $x \in \mathbb{R}^n$, then the decomposition coefficients c_j in (1.29) are given by $c_j = \langle x, \phi_j \rangle$. Furthermore, the relation

$$\|x\|_2^2 = \sum_{j=1}^n |c_j|^2 \quad (1.31)$$

holds true.

Equations (1.29)–(1.30) can be written also in matrix notation. If Φ is an $n \times n$ matrix with j -th column equal to ϕ_j , then (1.29) becomes $x = \Phi c$ and (1.30) reads $\Phi^T \Phi = I$, where I denoted the $n \times n$ identity matrix. As a consequence, $c = \Phi^T x$. We shall say that x has sparse or compressible representation with respect to the basis $\{\phi_j\}_{j=1}^n$ if the vector $c \in \mathbb{R}^n$ is sparse or compressible, respectively.

To allow for more flexibility in representation of signals, it is often useful to drop the condition of linear independence of the set $\{\phi_j\}_{j=1}^N \subset \mathbb{R}^n$. As before, we represent such a system of vectors by an $n \times N$ matrix Φ . We say that $\{\phi_j\}_{j=1}^N$ is a frame, if there are two positive finite constants $0 < A \leq B$, such that

$$A\|x\|_2^2 \leq \sum_{j=1}^N |\langle x, \phi_j \rangle|^2 \leq B\|x\|_2^2. \quad (1.32)$$

From $A > 0$, it follows that the span of the frame vectors is the whole \mathbb{R}^n and, therefore, that $N \geq n$. If one can choose $A = B$ in (1.32), then the frame is called tight. Dual frame of Φ is any other frame $\tilde{\Phi}$ with

$$\Phi \tilde{\Phi}^T = \tilde{\Phi} \Phi^T = I. \quad (1.33)$$

In general, for a given signal $x \in \mathbb{R}^n$ we can find infinitely many coefficients c , such that $x = \Phi c$. Actually, if $\tilde{\Phi}$ is a dual frame to Φ , one can take $c = \tilde{\Phi}^T x$. One is often interested in finding a vector of coefficients c with $x = \Phi c$, which is optimal in some sense. Especially, we shall say that x has a sparse or compressible representation with respect to the frame $\{\phi_j\}_{j=1}^N$ if c can be chosen sparse or compressible, cf. [33].

It can be shown that the smallest coefficient sequence in the ℓ_2^N sense is obtained by the choice $c = \Phi^\dagger x$, where Φ^\dagger is the Penrose pseudoinverse. In this context, Φ^\dagger is also called the canonical dual frame. Finally, let us note that (1.33) implies that

$$\sum_{j=1}^N \langle x, \phi_j \rangle \tilde{\phi}_j = \sum_{j=1}^N \langle x, \tilde{\phi}_j \rangle \phi_j = x$$

for every $x \in \mathbb{R}^n$.

The theory of compressed sensing was extended to the setting of sparse representations with respect to frames and dictionaries in [60]. The measurements now take the form $y = Ax = A\Phi c$, where c is sparse. Essentially, it turns out that if A satisfies the concentration inequalities from Section 1.3.4 and the dictionary Φ has small coherence, then the matrix $A\Phi$ has small RIP constants, and the methods of compressed sensing can be applied.

1.4.2 Coherence

We have provided in Section 1.3.4 a simple recipe how to construct matrices with small RIP constants - namely to choose each entry independently at random with respect to a correctly normalized standard distribution. On the other hand, if the matrix A is given beforehand, it is quite difficult to check if this matrix really satisfies the RIP, or to calculate its RIP constants. Another property of A , which is easily verifiable and which also ensures good recovery guarantees, is the coherence of A .

Definition 3. Let A be an $m \times n$ matrix and let $a_1, \dots, a_n \in \mathbb{R}^m$ be its columns. Then the coherence of A is the number $\mu(A)$ defined as

$$\mu(A) = \max_{1 \leq i < j \leq n} \frac{|\langle a_i, a_j \rangle|}{\|a_i\|_2 \|a_j\|_2}. \quad (1.34)$$

Due to Cauchy–Schwartz inequality, $\mu(A) \leq 1$ is always true. If $m \leq n$, then there is a lower bound (known as the Welch bound [71]) on the coherence given by $\mu(A) \geq \sqrt{\frac{n-m}{m(n-1)}}$. We give a particulary elegant proof of this bound, which has recently appeared in [45]. Without loss of generality, we may assume that the vectors a_1, \dots, a_n (which may be even complex) have unit norm and that $\mu = \max_{1 \leq i < j \leq n} |\langle a_i, a_j \rangle|$. Using the notion of the *trace* of a square matrix (which is just the sum of its diagonal entries) and some of its basic and very well-known properties, we obtain

$$\begin{aligned} 0 &\leq \text{tr}\left[\left(AA^* - \frac{n}{m}I\right)^2\right] = \text{tr}[(A^*A)^2] - \frac{n^2}{m} \\ &= \sum_{k,l=1}^n |\langle a_k, a_l \rangle|^2 - \frac{n^2}{m} \leq n + n(n-1)\mu^2 - \frac{n^2}{m}. \end{aligned}$$

Solving this inequality for μ gives the Welch bound.

Let us observe that if $n \gg m$, then this bound reduces to approximately $\mu(A) \geq 1/\sqrt{m}$. There is a lot of possible ways how to construct matrices with small coherence. Not surprisingly, one possible option is to consider random matrices A with each entry generated independently at random, cf. [58, Chapter 11]. Nevertheless the construction of matrices achieving the Welch bound exactly is still an active area of research, making use of ideas from algebra and number theory. On the other hand, it is easy to show that the Welch bound can not be achieved if n is much larger than m . It can be done only if $n \leq m(m+1)/2$ in the real case, and if $n \leq m^2$ in the complex case.

The connection of coherence to RIP is given by the following Lemma.

Lemma 6. *If A has unit-norm columns and coherence $\mu(A)$, then it satisfies the RIP of order k with $\delta_k(A) \leq (k-1)\mu(A)$ for all $k < 1/\mu(A)$.*

Combining this with Theorem 5, it gives recovery guarantees for the number of measurements m growing quadratically in the sparsity k .

1.4.3 Stability and Robustness

Basic discussion of stability and robustness of the methods of compressed sensing was given already in Section 1.3.5 with Theorem 6 being the most important representative of the variety of noise-aware estimates in the area. Its proof follows closely the presentation of [11]. The proof can be easily transformed to the spirit of Section 1.3.2 and 1.3.3 using the following modification of the Null Space Property.

Definition 4. We say that $A \in \mathbb{R}^{m \times n}$ satisfies the ℓ_2 -Robust Null Space Property of order k with constants $0 < \rho < 1$ and $\tau > 0$ if

$$\|v_T\|_2 \leq \frac{\rho \|v_{T^c}\|_1}{\sqrt{k}} + \tau \|Av\|_2 \quad (1.35)$$

for all $v \in \mathbb{R}^n$ and all sets $T \subset \{1, \dots, n\}$ with $|T| \leq k$.

The following theorem (which goes essentially back to [14]) is then the noise-aware replacement of Theorem 2.

Theorem 8. Let $A \in \mathbb{R}^{m \times n}$ with ℓ_2 -Robust Null Space Property of order k with constants $0 < \rho < 1$ and $\tau > 0$. Then for any $x \in \mathbb{R}^n$ the solution \hat{x} of $(P_{1,n})$ with $y = Ax + e$ and $\|e\|_2 \leq \eta$ satisfies

$$\|x - \hat{x}\|_2 \leq \frac{C}{\sqrt{k}} \sigma_k(x)_1 + D\eta \quad (1.36)$$

with constants $C, D > 0$ depending only on ρ and τ .

Finally, it turns out that the Restricted Isometry Property is also sufficient to guarantee the ℓ_2 -Robust Null Space Property and Theorem 3 can be extended to

Theorem 9. Let $A \in \mathbb{R}^{m \times n}$ and let k be a natural number with $k \leq n/2$. If $\delta_{2k}(A) < 1/3$, then A satisfies the ℓ_2 -Robust Null Space Property of order k with constants $0 < \rho < 1$ and $\tau > 0$ depending only on $\delta_{2k}(A)$.

Let us only point out, that the constant $1/3$ is by no means optimal, and that the same result (with more technical analysis) holds also if $\delta_{2k}(A) < 4/\sqrt{41}$, cf. [9, 10, 38, 39].

Theorems 6 and 8 are sufficient to analyze the situation, when the noise is bounded in the ℓ_2 -norm, no matter what the structure of the noise is. Unfortunately, it is not optimal for the analysis of measurements perturbed by Gaussian noise. To demonstrate this, let us assume that $e = (e_1, \dots, e_m)^T$, where e_i 's are independent normal variables with variance σ^2 , and that

$$y = Ax + e, \quad (1.37)$$

where the entries of $A \in \mathbb{R}^{m \times n}$ are independent standard normal variables. We divide this equation by \sqrt{m} and use that $A' = A/\sqrt{m}$ satisfies the RIP of order k with high probability for $m \geq Ck \ln(eN/k)$. As $\|e/\sqrt{m}\|_2 \leq 2\sigma$ with high probability, (1.36) becomes for a k -sparse $x \in \mathbb{R}^n$

$$\|x - \hat{x}\|_2 \leq D'\sigma. \quad (1.38)$$

We observe that increasing the number of (properly normalised) measurements does not lead to any decay of the approximation error.

To deal with this issue, the following recovery algorithm, called *Dantzig selector*

$$\min_{z \in \mathbb{R}^n} \|z\|_1 \quad \text{s.t.} \quad \|A^T(Az - y)\|_\infty \leq \tau, \quad (1.39)$$

was proposed and analyzed in [17]. It deals with the case, when $\|A^T e\|_\infty$ is small.

Theorem 10. *Let $A \in \mathbb{R}^{m \times n}$ be a matrix with RIP of order $2k$ and $\delta_{2k} < \sqrt{2} - 1$. Let the measurements y take the form $y = Ax + e$, where $\|A^T e\|_\infty \leq \tau$. Then the solution \hat{x} of (1.39) satisfies*

$$\|\hat{x} - x\|_2 \leq \frac{C}{\sqrt{k}} \sigma_k(x)_1 + D\sqrt{k}\tau, \quad (1.40)$$

where $C, D > 0$ depend only on $\delta_{2k}(A)$.

To see how this is related to measurements corrupted with Gaussian noise, let us assume again that the components of $e \in \mathbb{R}^m$ are i.i.d. normal variables with variance σ^2 . If the entries of A are again independent standard normal variables, then the 2-stability of normal variables gives that the coordinates of $A^T e$ are independent normal variables with mean zero and variance $\|e\|_2^2$. By simple union bound, we then obtain

$$\mathbb{P}(\|A^T e\|_\infty \geq t\|e\|_2) \leq 2n \exp(-t^2/2).$$

Combining this with the fact that $\mathbb{P}(\|e\|_2 \geq 2\sigma\sqrt{m}) \leq \exp(-m/2)$ and choosing $t = 2\sqrt{\ln(2n)}$, we finally get

$$\mathbb{P}(\|A^T e\|_\infty \geq 4\sigma\sqrt{m\ln(2n)}) \leq \exp(-m/2) + 2n \exp(-2\ln(2n)) \leq \frac{1}{n}. \quad (1.41)$$

Dividing (1.37) by \sqrt{m} again and applying Theorem 10, we obtain for the case of a k sparse vector $x \in \mathbb{R}^n$

$$\|x - \hat{x}\|_2 \leq D'\sigma\sqrt{\frac{k\ln(2n)}{m}} \quad (1.42)$$

if $m \geq Ck\ln(2n)$. The advantage of (1.42) over (1.38) is that (once $m \geq Ck\ln(2n)$) it decreases with m , i.e. taking more noisy measurements decreases the approximation error.

1.4.4 Recovery algorithms

Although we concentrated on ℓ_1 -minimization in the first part of this chapter, there is a number of different algorithms solving the problem of sparse signal recovery. Similarly to ℓ_1 -minimization, which was used successfully in machine learning

much before the advent of compressed sensing, many of these algorithms also predate the field of compressed sensing. We give an overview of some of these algorithms and refer to [40] for more extensive treatment.

1.4.4.1 ℓ_1 -minimization

The ℓ_1 -minimization problems (P_1) or $(P_{1,\eta})$ presented before form a backbone of the theory of compressed sensing. Their geometrical background allows for theoretical recovery guarantees, including corresponding stability and robustness extensions. They are formulated as convex optimization problems, which can be solved effectively by any general purpose numerical solver. Furthermore, several implementations dealing with the specific setting of compressed sensing are available nowadays.

Sometimes, it is more convenient to work with some of the equivalent reformulations of $(P_{1,\eta})$. Let us discuss two most important of them. Let $\eta \geq 0$ be given and let \hat{x} be a solution of the optimization problem $(P_{1,\eta})$

$$\hat{x} = \operatorname{argmin}_{z \in \mathbb{R}^n} \|z\|_1 \quad \text{s.t.} \quad \|Az - y\|_2 \leq \eta. \quad (P_{1,\eta})$$

Then there is a $\lambda \geq 0$, such that \hat{x} is also a solution of the non-constrained convex problem

$$\hat{x} = \operatorname{argmin}_{z \in \mathbb{R}^n} \frac{1}{2} \|Az - y\|_2^2 + \lambda \|z\|_1. \quad (1.43)$$

This version of ℓ_1 -minimization is probably the mostly studied one, see, for example, [34, 41, 51, 73]. On the other hand, if $\lambda > 0$ is given and \hat{x} is a solution to (1.43), then there is an $\eta > 0$, such that \hat{x} is also a solution of $(P_{1,\eta})$. In the same sense, $(P_{1,\eta})$ and (1.43) is also equivalent to *Lasso* (least absolute shrinkage and selection operator, cf. [64])

$$\hat{x} = \operatorname{argmin}_{z \in \mathbb{R}^n} \|Az - y\|_2^2 \quad \text{s.t.} \quad \|z\|_1 \leq \tau. \quad (1.44)$$

Unfortunately, the values of λ and $\tau > 0$ making these problems equivalent are a priori unknown.

The last prominent example of an optimization problem, which takes a form of ℓ_1 -minimization is the Dantzig selector (1.39). Let us also point out that [7] provides solvers for a variety of ℓ_1 -minimization problems.

1.4.4.2 Greedy algorithms

Another approach to sparse recovery is based on iterative identification/approximation of the support of the unknown vector x and of its components. For example,

one adds in each step of the algorithm one index to the support to minimize the mismatch to the measured data as much as possible. Therefore, such algorithms are usually referred to as greedy algorithms. For many of them, remarkable theoretical guarantees are available in the literature, sometimes even optimal in the sense of the lower bounds discussed above. Nevertheless, the techniques necessary to achieve these results are usually completely different from those needed to analyze ℓ_1 -minimization. We will discuss three of these algorithms, *Orthogonal Matching Pursuit*, *Compressive Sampling Matching Pursuit*, and *Iterative Hard Thresholding*.

Orthogonal Matching Pursuit (OMP)

Orthogonal Matching Pursuit [53, 65, 67] adds in each iteration exactly one entry into the support of \hat{x} . After k iterations, it therefore outputs a k -sparse vector \hat{x} .

The algorithm finds in each step the column of A most correlated with the residual of the measurements. Its index is then added to the support. Finally, it updates the target vector \hat{x}_i as the vector supported on T_i that best fits the measurements, i.e. which minimizes $\|y - Az\|_2$ among all $z \in \mathbb{R}^n$ with $\text{supp}(z) \subset T_i$. It is well known that this vector is given as the product of the Penrose pseudoinverse A^\dagger of A and y .

The formal transcription of this algorithm is given as follows.

Orthogonal Matching Pursuit (OMP)

Input: Compressed sensing matrix A , measurement vector y

Initial values: $\hat{x}_0 = 0, r = y, T_0 = \emptyset, i = 0$

Iteration step: Repeat until stopping criterion is met

$i := i + 1$	
$T_i \leftarrow T_{i-1} \cup \text{supp } H_1(A^T r)$	add largest residual entry to the support
$\hat{x}_i _{T_i} \leftarrow A_{T_i}^\dagger y$	update the estimate of the signal
$r \leftarrow y - A\hat{x}_i$	update the residual of the measurements

Output: \hat{x}_i

It makes use of the hard thresholding operator $H_k(x)$. If $x \in \mathbb{R}^n$ and $k \in \{0, 1, \dots, n\}$, then $H_k : x \rightarrow H_k(x)$ associates with x a vector $H_k(x) \in \mathbb{R}^n$, which is equal to x on the k entries of x with largest magnitude and zero otherwise. The stopping criteria can either limit the overall number of iterations (limiting also the size of the support of the output vector \hat{x}) or ensure that the distance between y and $A\hat{x}$ is small in some norm.

The simplicity of OMP is unfortunately connected with one of its weak points. If an incorrect index is added to the support in some step (which can happen in general and depends on the properties of the input parameters), it cannot be removed any more, and stays there until the end of OMP. We refer also to [26] for another variant of OMP.

Compressive Sampling Matching Pursuit (CoSaMP)

One attempt to overcome this drawback is presented in the following algorithm called *Compressive Sampling Matching Pursuit* [57]. It assumes that an additional input is given - namely the expected sparsity of the output. At each step it again enlarges the support, but in contrast to OMP, it will add at least k new entries. Afterwards, it again uses the Penrose pseudo-inverse to find the minimizer of $\|Az - y\|_2$ among all $z \in \mathbb{R}^n$ with $\text{supp}(z) \subset T_i$, but this time only the k largest of coordinates of this minimizer are stored.

The formal description is given by the following scheme.

Compressive Sampling Matching Pursuit (CoSaMP)

Input: Compressed sensing matrix A , measurement vector y , sparsity level k

Initial values: $\hat{x}_0 = 0, r = y, T_0 = \emptyset, i = 0$

Iteration step: Repeat until stopping criterion is met

$$i := i + 1$$

$$\begin{aligned} T_i &\leftarrow \text{supp}(\hat{x}_{i-1}) \cup \text{supp} H_{2k}(A^T r) && \text{update the support} \\ \hat{x}_i|_{T_i} &\leftarrow H_k(A_{T_i}^\dagger y) && \text{update the estimate of the signal} \\ r &\leftarrow y - A\hat{x}_i && \text{update the residual} \end{aligned}$$

Output: \hat{x}_i

Iterative Hard Thresholding (IHT)

The last algorithm [8] we shall discuss is also making use of the hard thresholding operator H_k . The equation $Az = y$ is transformed into $A^T Az = A^T y$, which again can be interpreted as looking for the fixed point of the mapping $z \rightarrow (I - A^T A)z + A^T y$. Classical approach is then to iterate this mapping and to put $\hat{x}_i = (I - A^T A)\hat{x}_{i-1} + A^T y = \hat{x}_{i-1} + A^T(y - A\hat{x}_{i-1})$. Iterative Hard Thresholding algorithm is doing exactly this, only combined with the hard thresholding operator H_k .

Iterative Hard Thresholding (IHT)

Input: Compressed sensing matrix A , measurement vector y , sparsity level k

Initial values: $\hat{x}_0 = 0, i = 0$

Iteration step: Repeat until stopping criterion is met

$$i := i + 1$$

$$\hat{x}_i = H_k(\hat{x}_{i-1} + A^T(y - A\hat{x}_{i-1})) \quad \text{update the estimate of the signal}$$

Output: \hat{x}_i

1.4.4.3 Combinatorial algorithms

The last class of algorithms for sparse recovery we shall review were developed mainly in the context of theoretical computer science and they are based on classical ideas from this field, which usually pre-date the area of compressed sensing. Nevertheless, they were successfully adapted to the setting of compressed sensing.

Let us present the basic idea on the example of Group Testing, which was introduced by Robert Dorfman [27] in 1943. One task of United States Public Health Service during the Second World War was to identify all syphilitic soldiers. However, syphilis test in that time was expensive and the naive approach of testing every soldier independently would have been very costly.

If the portion of infected soldiers would be large (say above 50 percent), then the method of individual testing would be reasonable (and nearly optimal). A realistic assumption however is that only a tiny fraction of all the soldiers is infected, say one in thousand, or one in ten thousand. The main idea of the area of Group Testing in this setting is that we can combine blood samples and test a combined sample to check if at least one soldier in the group has syphilis. Another example of this technique is the false coin problem from recreational mathematics, in which one is supposed to identify in a group of n coins a false coin weighting less than a real coin. We refer to [28] to an overview of the methods of Group Testing.

To relate this problem to compressed sensing, let us consider a vector $x = (x_1, \dots, x_n) \in \{0, 1\}^n$, where n is the number of soldiers, with $x_i = 0$ if the i th soldier is healthy, or $x_i = 1$ if he has syphilis. The grouping is then represented by an $m \times n$ matrix $A = (a_{ij})$, where $a_{ij} = 1$, if the blood sample of j th soldier was added to i th combined sample. The methods of Group Testing then allow to design efficient matrices A , such that the recovery of x can be done in a surprisingly small number of steps - even linear in the length of the sparse representation of x , i.e. in its sparsity k , cf. [43, 44].

1.4.5 Structured sparsity

In many applications, one has much more prior knowledge about the signal x , than just assuming that it possesses a sparse representation with respect to certain basis, frame, or dictionary.

For example, the image coder JPEG2000 exploits not only the fact that natural images have compressible representation in the wavelet basis (i.e., that most of their wavelet coefficients are small) but it also uses the fact that the values and locations of the large coefficients have a special structure. It turns out that they tend to cluster into a connected subtree inside the wavelet parent–child tree. Using this additional information can of course help to improve the properties of the coder and provide better compression rates [30, 31, 48].

Another model appearing frequently in practice is the model of block-sparse (or joint-sparse) signals. Assume that we want to recover N correlated signals

$x^1, \dots, x^N \in \mathbb{R}^n$ with (nearly) the same locations of their most significant elements. A simple example of such a situation are the three color channels of a natural RGB image, where we intuitively expect the important wavelet coefficients in all three channels to be on nearly the same locations. Furthermore, the same model often appears in the study of DNA microarrays, magnetoencephalography, sensor networks, and MIMO communication [6, 32, 63, 70]. It is usually convenient to represent the signals as columns of an $n \times N$ matrix $X = [x^1 \dots x^N]$. The recovery algorithms are then based on mixed matrix norms, which are defined for such an X as

$$\|X\|_{(p,q)} = \left(\sum_{i=1}^n \|\tilde{x}^i\|_p^q \right)^{1/q},$$

where $p, q \geq 1$ are real numbers and $\tilde{x}^i, i = 1, \dots, n$, are the rows of the matrix X . If A is again the sensing matrix and $Y = AX$ are the measurements, then the analogue of (P_1) in this setting is then

$$\hat{X} = \underset{Z \in \mathbb{R}^{n \times N}}{\operatorname{argmin}} \|Z\|_{(p,q)} \quad \text{s. t.} \quad Y = AZ$$

for a suitable choice of p and q , typically $(p, q) = (2, 1)$. We refer, for example, to [36, 66, 68] for further results.

Finally, let us point out that *model-based compressive sensing* [3] provides a general framework for many different kinds of structured sparsity.

1.4.6 Compressed Learning

In this last part, we will discuss applications of compressed sensing to a classical task of approximation theory, namely to learning of an unknown function f from a limited number of its samples $f(x^1), \dots, f(x^m)$. In its most simple form, treated already in [13] and elaborated in [59], one assumes that the function f is known to be a sparse combination of trigonometric polynomials of maximal order q in dimension d , i.e. that

$$f(x) = \sum_{l \in \{-q, -q+1, \dots, q-1, q\}^d} c_l e^{il \cdot x}$$

and $\|c\|_0 \leq k$, where $k \in \mathbb{N}$ is the level of sparsity. Theorem 2.1 of [59] then shows that, with probability at least $1 - \varepsilon$, f can be exactly recovered from samples $f(x^1), \dots, f(x^m)$, where $m \geq Ck \ln((2q+1)^d / \varepsilon)$ and x^1, \dots, x^m are uniformly and independently distributed in $[0, 2\pi]^d$. The recovery algorithm is given by

$$\underset{c}{\operatorname{argmin}} \|c\|_1 \quad \text{s. t.} \quad \sum_l c_l e^{il \cdot x^j} = f(x^j), \quad j = 1, \dots, m.$$

We refer to [12, 61] for further results and to [40, Chapter 12] for an overview on random sampling of functions with sparse representation in a bounded orthonormal system.

In another line of study, compressed sensing was used to approximate functions $f : [0, 1]^d \rightarrow \mathbb{R}$, which depend only on $k \ll d$ (unknown) *active variables* i_1, \dots, i_k , i.e.

$$f(x) = f(x_1, \dots, x_d) = g(x_{i_1}, \dots, x_{i_k}), \quad x \in [0, 1]^d.$$

In [24] and [72], the authors presented sophisticated combinatorial (adaptive and non-adaptive) constructions of sets of sampling points, which allowed for recovery of f to a precision of $1/L$ using only $C(k)(L+1)^k \ln d$ points. Observe that $(L+1)^k$ points would be necessary even if the location of the active coordinates would be known. The use of compressed sensing in this setting was then discussed in [62]. The algorithm developed there was based on approximation of directional derivatives of f at random points $\{x^1, \dots, x^{m_X}\}$ and random directions $\{\varphi^1, \dots, \varphi^{m_\Phi}\}$. Denoting the $m_\Phi \times m_X$ matrix of first order differences as Y and the $m_\Phi \times d$ matrix of random directions by Φ , it was possible to use direct estimates of probability concentrations to ensure that the k largest rows of $\Phi^T Y$ correspond to the k active coordinates of f with high probability. Again, only an additional $\ln d$ factor is paid for identifying the unknown active coordinates.

Finally, the paper [21] initiated a study of approximation of ridge functions of the type

$$f(x) = g(\langle a, x \rangle), \quad x \in [0, 1]^d, \quad (1.45)$$

where both the direction $a \in \mathbb{R}^d \setminus \{0\}$ and the univariate function g are unknown. Due to the assumption $a_j \geq 0$ for all $j = 1, \dots, d$, posed in [21], it was first possible to approximate g by sampling on grid points along the diagonal $\{\frac{i}{L}(1, \dots, 1)^T, i = 0, \dots, L\}$. Afterwards, the methods of compressed sensing were used in connection with the first order differences to identify the vector a . The importance of derivatives of f in connection with the assumption (1.45) is best seen from the simple formula

$$\nabla f(x) = g'(\langle a, x \rangle) \cdot a. \quad (1.46)$$

Hence, approximating the gradient of f at a point x gives actually also a scalar multiple of a .

Another algorithm to approximate the ridge functions was proposed in [37]. Similarly to [62], it was based on (1.46) and on approximation of the first order derivatives by first order differences. In contrary to [21], first the ridge direction a was recovered, and only afterwards the ridge profile g was approximated by any standard one-dimensional sampling scheme. Furthermore, no assumptions on signs of a was needed and it was possible to generalize the approach also for recovery of k -ridge functions of the type $f(x) = g(Ax)$, where $A \in \mathbb{R}^{k \times d}$ and g is a function of k variables. We refer also to [18] for further results.

Acknowledgements We would like to thank Sandra Keiper and Holger Rauhut for careful reading of a previous version of this Introduction and for useful comments. The last author was supported by the ERC CZ grant LL1203 of the Czech Ministry of Education.

References

1. Achlioptas, D.: Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.* **66**, 671–687 (2003)
2. Arora, S., Barak, B.: *Computational Complexity: A Modern Approach*. Cambridge University Press, Cambridge (2009)
3. Baraniuk, R., Cevher, V., Duarte, M.F., Hegde, C.: Model-based compressive sensing, *IEEE Trans. Inf. Theory* **56**, 1982–2001 (2010)
4. Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices. *Constr. Approx.* **28**, 253–263 (2008)
5. Baraniuk, R., Steeghs, P.: Compressive radar imaging. In: Proc. IEEE Radar Conf., Boston, pp. 128–133 (2007)
6. Baron, D., Duarte, M.F., Sarvotham, S., Wakin, M.B., Baraniuk, R.: Distributed compressed sensing of jointly sparse signals. In: Proc. Asilomar Conf. Signals, Systems, and Computers, Pacic Grove (2005)
7. Becker, S., Candès, E.J., Grant, M.: Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput.* **3**, 165–218 (2010)
8. Blumensath, T., Davies, M.: Iterative hard thresholding for compressive sensing. *Appl. Comput. Harmon. Anal.* **27**, 265–274 (2009)
9. Cai, T., Wang, L., Xu, G.: New bounds for restricted isometry constants. *IEEE Trans. Inf. Theory* **56**, 4388–4394 (2010)
10. Cai, T., Wang, L., Xu, G.: Shifting inequality and recovery of sparse vectors. *IEEE Trans. Signal Process.* **58**, 1300–1308 (2010)
11. Candès, E.J.: The restricted isometry property and its implications for compressed sensing. *C. R. Acad. Sci., Paris, Ser. I* **346**, 589–592 (2008)
12. Candès, E.J., Plan, Y.: A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inf. Theory* **57**, 7235–7254 (2011)
13. Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**, 489–509 (2006)
14. Candès, E.J., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59**, 1207–1223 (2006)
15. Candès, E.J., Tao, T.: Decoding by linear programming. *IEEE Trans. Inf. Theory* **51**, 4203–4215 (2005)
16. Candès, E.J., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inf. Theory* **52**, 5406–5425 (2006)
17. Candès, E.J., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n . *Ann. Stat.* **35**, 2313–2351 (2007)
18. Cevher, V., Tyagi, H.: Active learning of multi-index function models. In: Proc. NIPS (The Neural Information Processing Systems), Lake Tahoe, Reno (2012)
19. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20**, 33–61 (1998)
20. Cohen, A., Dahmen, W., DeVore, R.: Compressed sensing and best k -term approximation. *J. Am. Math. Soc.* **22**, 211–231 (2009)
21. Cohen, A., Daubechies, I., DeVore, R., Kerkyacharian, G., Picard, D.: Capturing ridge functions in high dimensions from point queries. *Constr. Approx.* **35**, 225–243 (2012)
22. Dasgupta, S., Gupta, A.: An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithm.* **22**, 60–65 (2003)

23. Davenport, M.A., Duarte, M.F., Eldar, Y.C., Kutyniok, G.: Introduction to compressed sensing. Compressed sensing, pp. 1–64. Cambridge University Press, Cambridge (2012)
24. DeVore, R., Petrova, G., Wojtaszczyk, P.: Approximation of functions of few variables in high dimensions. *Constr. Approx.* **33**, 125–143 (2011)
25. Donoho, D.L.: Compressed sensing. *IEEE Trans. Inf. Theory* **52**, 1289–1306 (2006)
26. Donoho, D.L., Tsaig, Y., Drori, I., Starck, J.-L.: Sparse solution of underdetermined systems of linear equations by stagewise Orthogonal Matching Pursuit. *IEEE Trans. Inf. Theory* **58**, 1094–1121 (2012)
27. Dorfman, R.: The detection of defective members of large populations. *Ann. Math. Stat.* **14**, 436–440 (1943)
28. Du, D., Hwang, F.: Combinatorial group testing and its applications. World Scientific, Singapore (2000)
29. Duarte, M., Davenport, M., Takhar, D., Laska, J., Ting, S., Kelly, K., Baraniuk R.: Single-pixel imaging via compressive sampling. *IEEE Signal Process. Mag.* **25**, 83–91 (2008)
30. Duarte, M., Wakin, M., Baraniuk, R.: Fast reconstruction of piecewise smooth signals from random projections. In: Proc. Work. Struc. Parc. Rep. Adap. Signaux (SPARS), Rennes (2005)
31. Duarte, M., Wakin, M., Baraniuk, R.: Wavelet-domain compressive signal reconstruction using a hidden Markov tree model. In: Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP), Las Vegas (2008)
32. Eldar, Y., Mishali, M.: Robust recovery of signals from a structured union of subspaces. *IEEE Trans. Inf. Theory* **55**, 5302–5316 (2009)
33. Elad, M.: Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer, New York (2010)
34. Figueiredo, M., Nowak, R., Wright, S.: Gradient projections for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE J. Select. Top. Signal Process.* **1**, 586–597 (2007)
35. Fornasier, M., Rauhut, H.: Compressive sensing. In: Scherzer, O. (ed.) *Handbook of Mathematical Methods in Imaging*, pp. 187–228. Springer, Heidelberg (2011)
36. Fornasier, M., Rauhut, H.: Recovery algorithms for vector valued data with joint sparsity constraints. *SIAM J. Numer. Anal.* **46**, 577–613 (2008)
37. Fornasier, M., Schnass, K., Vybíral, J.: Learning functions of few arbitrary linear parameters in high dimensions. *Found. Comput. Math.* **12**, 229–262 (2012)
38. Foucart, S.: A note on guaranteed sparse recovery via l_1 -minimization. *Appl. Comput. Harmon. Anal.* **29**, 97–103 (2010)
39. Foucart, S., Lai M.: Sparsest solutions of underdetermined linear systems via l_q -minimization for $0 < q \leq 1$. *Appl. Comput. Harmon. Anal.* **26**, 395–407 (2009)
40. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Birkhäuser/Springer, New York (2013)
41. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010)
42. Gártner, B., Matoušek, J.: Understanding and Using Linear Programming. Springer, Berlin (2006)
43. Gilbert, A., Li, Y., Porat, E., and Strauss, M.: Approximate sparse recovery: optimizaing time and measurements. In: Proc. ACM Symp. Theory of Comput., Cambridge (2010)
44. Gilbert, A., Strauss, M., Tropp, J., Vershynin, R.: One sketch for all: fast algorithms for compressed sensing. In: Proc. ACM Symp. Theory of Comput., San Diego (2007)
45. Jasper, J., Mixon, D.G., Fickus M.: Kirkman equiangular tight frames and codes. *IEEE Trans. Inf. Theory* **60**, 170–181 (2014)
46. Johnson, W.B., Lindenstrauss, J.: Extensions of Lipschitz mappings into a Hilbert space. In: Conf. in Modern Analysis and Probability, pp. 189–206 (1984)
47. Krahmer, F., Ward, R.: New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. *SIAM J. Math. Anal.* **43**, 1269–1281 (2011)
48. La, C., Do, M.N.: Tree-based orthogonal matching pursuit algorithm for signal reconstruction. In: IEEE Int. Conf. Image Processing (ICIP), Atlanta (2006)

49. Ledoux, M.: *The Concentration of Measure Phenomenon*. American Mathematical Society, Providence (2001)
50. Ledoux, M., Talagrand, M.: *Probability in Banach Spaces. Isoperimetry and Processes*. Springer, Berlin (1991)
51. Loris, I.: On the performance of algorithms for the minimization of ℓ_1 -penalized functions. *Inverse Prob.* **25**, 035008 (2009)
52. Lustig, M., Donoho, D., Pauly, J.M.: Sparse MRI: the application of compressed sensing for rapid MR imaging. *Magn. Reson. Med.* **58**, 1182–1195 (2007)
53. Mallat, S., Zhang, Z.: Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* **41**, 3397–3415 (1993)
54. Matoušek, J.: On variants of the Johnson-Lindenstrauss lemma. *Rand. Struct. Algorithm.* **33**, 142–156 (2008)
55. Milman, V.D., Schechtman, G.: *Asymptotic theory of finite-dimensional normed spaces*. Springer, Berlin (1986)
56. Mishali, M., Eldar, Y.: From theory to practice: Sub-nyquist sampling of sparse wideband analog signals. *IEEE J. Sel. Top. Signal Process.* **4**, 375–391 (2010)
57. Needell, D., Tropp, J.: CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.* **26**, 301–321 (2009)
58. Pietsch, A.: *Operator Ideals*. North-Holland, Amsterdam (1980)
59. Rauhut, H.: Random sampling of sparse trigonometric polynomials. *Appl. Comput. Harmon. Anal.* **22**, 16–42 (2007)
60. Rauhut, H., Schnass, K., Vandegeynst, P.: Compressed sensing and redundant dictionaries. *IEEE Trans. Inf. Theor.* **54**, 2210–2219 (2008)
61. Rauhut, H., Ward, R.: Sparse Legendre expansions via ℓ_1 -minimization. *J. Approx. Theory* **164**, 517–533 (2012)
62. Schnass, K., Vybráil, J.: Compressed learning of high-dimensional sparse functions. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3924–3927 (2011)
63. Stojnic, M., Parvaresh, F., Hassibi, B.: On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Trans. Signal Process.* **57**, 3075–3085 (2009)
64. Tibshirani, R.: Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996)
65. Tropp, J.: Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inf. Theor.* **50**, 2231–2242 (2004)
66. Tropp, J.: Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Process.* **86**, 589–602 (2006)
67. Tropp, J., Gilbert, A.: Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory* **53**, 4655–4666 (2007)
68. Tropp, J., Gilbert, A., Strauss, M.: Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Process.* **86**, 572–588 (2006)
69. Tropp, J., Laska, J., Duarte, M., Romberg, J., Baraniuk, R.: Beyond Nyquist: efficient sampling of sparse bandlimited signals. *IEEE Trans. Inf. Theor.* **56**, 520–544 (2010)
70. Wakin, M.B., Sarvotham, S., Duarte, M.F., Baron, D., Baraniuk, R.: Recovery of jointly sparse signals from few random projections. In: *Proc. Workshop on Neural Info. Proc. Sys. (NIPS)*, Vancouver (2005)
71. Welch, L.: Lower bounds on the maximum cross correlation of signals. *IEEE Trans. Inf. Theory* **20**, 397–399 (1974)
72. Wojtaszczyk, P.: Complexity of approximation of functions of few variables in high dimensions. *J. Complex.* **27**, 141–150 (2011)
73. Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman iterative algorithms for ℓ_1 -minimization with applications to compressed sensing. *SIAM J. Imag. Sci.* **1**, 143–168 (2008)

Chapter 2

Temporal Compressive Sensing for Video

Patrick Llull, Xin Yuan, Xuejun Liao, Jianbo Yang, David Kittle,
Lawrence Carin, Guillermo Sapiro, and David J. Brady

Abstract Video camera architects must design cameras capable of high-quality, dynamic event capture, while adhering to power and communications constraints. Though modern imagers are capable of both simultaneous spatial and temporal resolutions at micrometer and microsecond scales, the power required to sample at these rates is undesirable. The field of compressive sensing (CS) has recently suggested a solution to this design challenge. By exploiting physical-layer compression strategies, one may overlay the original scene with a coding sequence to sample at sub-Nyquist rates with virtually no additional power requirement. The underlying scene may be later estimated without significant loss of fidelity. In this chapter, we cover a variety of such strategies taken to improve an imager’s temporal resolution. Highlighting a new low-power acquisition paradigm, we show how a video sequence of high temporal resolution may be reconstructed from a single video frame taken with a low-framerate camera.

2.1 Introduction

The goal of image and video acquisition is, largely, to determine as much about the target scene as possible via remote methods. To this end, designers wish to maximize an imager’s normalized information capacity, which is proportional to $\frac{A^2}{\lambda^2}$, where A and λ are the optical aperture diameter and source wavelength, respectively. Real-life objects are fine, dynamic, and comprised of many materials; interesting features persist on transverse spatial (orthogonal to the optical axis) scales of λ , depth (parallel to the optical axis) scales of centimeters, temporal scales of microseconds, and spectral scales of 100λ [24]. With the advent of modern monocentric, multiscale lens design [7, 9, 36], the number of samples required to fully capture, transmit, and write the incident datastream is of the order of 10^{15} [24].

P. Llull (✉) • X. Yuan • X. Liao • J. Yang • D. Kittle • L. Carin • G. Sapiro • D.J. Brady
Department of Electrical and Computer Engineering, Duke University,
Durham, NC 27708, USA
e-mail: patrick.llull@duke.edu; xin.yuan@duke.edu; xjliao@duke.edu; jy118@duke.edu;
dsk12@duke.edu; lcarin@duke.edu; guillermo.sapiro@duke.edu; dbrady@duke.edu

Still-frame imaging systems, such as point-and-shoot cameras, are optimized for transverse spatial resolution for a given set of aperture characteristics. Terms familiar to photographers include f-number (F/%), focal length, exposure, International Standards Organization (ISO) sensitivity level, and field of view (FOV), among others. To combat motion blur inherent in dynamic scenes, photographers may increase the camera's temporal resolution by jointly adjusting the ISO sensitivity level, F/%, and exposure time. This results in a compromise between light efficiency, spatial resolution, image 'graininess' and temporal resolution; all constrained by power consumption.

Videographers must additionally cope with issues of temporal aliasing and data transfer. If F images are taken per second, a traditional video camera can, with high fidelity, capture and represent periodic objects moving with a maximum frequency of $F/2$ Hz. Any periodic motion residing outside this temporal passband is falsely reproduced during video playback. The result is motion that is aliased into the low temporal frequencies and is *independent* of the exposure time, which cannot exceed $1/F$. This effect often manifests in fast-moving automobile wheels that seem to spin slowly in the wrong direction. The main ways to improve a video camera's robustness to temporal aliasing are: (a) increase the framerate and (b) correctly model and account for the aliasing. Traditionally, approach (a) is taken due to simplicity, which forces higher data transfer rates to capture dynamic events.

A finite imaging bandwidth has the capacity to transmit ND_rF bits per second, where N represents the number of transverse pixels, F is the framerate, and D_r signifies the dynamic range. 10–20 megapixel sensors, which are standard quality for still-life cameras, already reach practical communication limits when recorded at a modest rate of ~ 30 fps. To increase the framerate given a maximum bandwidth, the video camera operator resorts to downsampling, reducing dynamic range, or reducing field of view. For a given video sequence, all of these solutions reduce the imager's sampling degrees of freedom and detectable voxel count (considering every spatial, temporal, and spectral channel). Given modern communication channels, full-scale spectral and volumetric video acquisition seems entirely beyond reach.

Several solutions have been proposed to improve resolution among various temporal, spatial, and spectral channels to partially address this bandwidth deficiency. In this chapter, we will focus on methods for improving the imager's resolution in time. The next section establishes a mathematical framework for the problem and describes several approaches of solving it.

2.2 Temporal resolution background

Any camera's temporal resolution is physically limited by the response time of the detector material, which ranges from microseconds to milliseconds. Detector pixels are arranged in charge-coupled device (CCD) and complementary metal oxide semiconductor (CMOS) technologies, which vary in performance and scalability. For an overview of these technologies, please visit [6].

The physical process of charge collection, distribution, readout, and quantization may be performed at amazing speeds of thousands of frames per second [22]. However, in addition to tackling the engineering challenges of writing and transmitting megapixel-scale data at these rates, camera designers consider additional physical characteristics (such as pixel spatial and temporal fill factors and quantum and charge transfer efficiency) when considering operating speed and everyday use.

2.2.1 *Temporal resolution of image and video cameras*

Although the optical system designer must choose between these readout strategies, we will focus in this text on the video capture process up to readout and ignore the contributions of electrical circuitry, analog-to-digital conversion, hardware-based image processing/denoising, data transmission, and data writing. Post-processing steps for image reconstruction are also described in this section.

Similar to spatial resolution, where pixel size and diffraction limit the potency of a camera, temporal resolution is limited by the sensor framerate and sensitivity. Parameters such as ISO level and bandwidth affect these characteristics, with further conventional dependencies on the pixel count and architecture.

Assuming rectangular pixels and a time-invariant optical impulse response h , a conventional imager forms an image on an $N_u \times N_v$ -pixel detector of a dynamic analog object $x(u, v, t)$ via discrete samples of the continuous transformation [6]

$$y(u', v', t') = \int \int \int x(u, v, t) h(u - u', v - v') \text{rect} \left(\frac{u - u'}{\Delta}, \frac{v - v'}{\Delta} \right) \times p_t(t, t', \Delta_t) dx dy dt, \quad (2.1)$$

where Δ and Δ_t , respectively, denote the detector pixel pitch and temporal integration time. p_t represents the temporal pixel sampling function, where t' varies more slowly than t . For a conventional imager, p_t is efficiently modeled via a rectangular pulse given by

$$p_t(t, t', \Delta_t) = \text{rect} \left(\frac{t - t'}{\Delta_t} \right). \quad (2.2)$$

Convolution of the object video with this temporal pixel sampling function equates to multiplication of the video's temporal spectrum with a bandlimited sinc function in the frequency domain. This sinc function has temporal nulls at multiples of $\frac{1}{\Delta_t}$ and, in practice, has finite support. Taking the 3-dimensional (space-time) Fourier transform of (2.1) and letting (ξ, η, w) , respectively, denote the transverse spatial and temporal frequency components, a conventional imager's spatiotemporal resolution is given by

$$\tilde{y}(\xi, \eta, w) = \text{sinc}(\xi\Delta, \eta\Delta)\text{sinc}(w\Delta_t)\tilde{x}(\xi, \eta, w)\tilde{h}(\xi, \eta, 0), \quad (2.3)$$

where \tilde{y} and \tilde{x} represent the 3D Fourier transforms of the image at the detector and of the object video, respectively. \tilde{h} , the optical transfer function (OTF), is presumably time-invariant, hence the temporal argument of 0. The object video is low-pass filtered by the OTF and the spatial and temporal pixel transfer functions. Additionally, spatial frequencies $\geq \frac{1}{2\Delta}$ and temporal frequencies $\geq \frac{1}{2\Delta_t}$ are aliased into the video's lower frequencies of (ξ, η) and w . Rapid motion is greatly attenuated due to the low values of the temporal pixel transfer function $\text{sinc}(w\Delta_t)$, resulting in motion blur.

Determination of the temporal integration time presents a fundamental tradeoff between an imager's framerate, temporal resolution, and SNR. This tradeoff begs the question: *what is the optimal time spent sampling (i.e., temporal fill factor) during a detected frame?* The temporal fill factor Γ of a sensor capturing with a framerate F is given by

$$\Gamma = \frac{\Delta_t}{F^{-1}}. \quad (2.4)$$

Shorter integration times Δ_t may mitigate the effects of motion blur (and hence result in higher temporal resolution) relative to larger Δ_t . Of course, decreasing the integration time reduces light collection efficiency and SNR. Additionally, temporal aliasing may still occur due to inefficient sampling rates. This effect is shown pictorially in Fig. 2.1 and proves similar to how altering a pixel's fill factor changes the pixel transfer function while maintaining the same aliasing limit [6] (and hence its resolution and light collection abilities).

The left drawing of Fig. 2.1 shows a low Γ which results in high temporal resolution but low light collection ability. The photographer could increase ISO to compensate for this at the cost of increased image noise. As previously discussed, the F/# could also be decreased to compensate for this at the cost of reduced FOV and a larger diffraction-limited spot size.

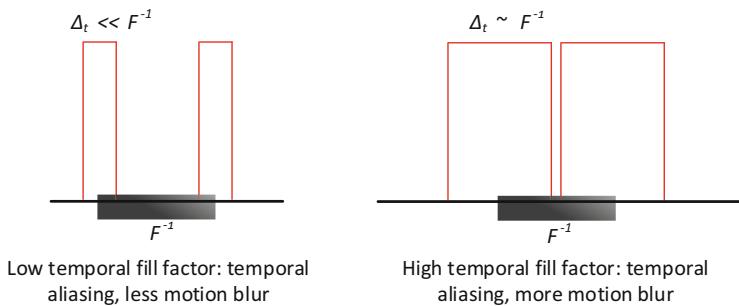


Fig. 2.1 Tradeoffs between different temporal fill factors. F denotes the camera framerate

Consider an imaged landscape impinging upon the focal plane with an optical irradiance $y(u', v', t')$ (in Watts per square meter). A camera's photoelectric conversion process depends on the sensor's responsivity R at a given wavelength λ by

$$R(\lambda) = \frac{\zeta(\lambda)q_e\lambda}{P_c C_v}, \quad (2.5)$$

where ζ is the sensor's quantum efficiency, q_e is an electron charge (in Coulombs), C_v is the speed of light in vacuum, and P_c is Planck's constant. Given a gain coefficient α (determined by the ISO setting), the number of electrons each pixel produces is given by

$$N(u', v', t') = \alpha R y(u', v', t') \Delta_t. \quad (2.6)$$

Assuming the impinging signal follows Poisson photon statistics and does not saturate the detector, the average signal to noise ratio (SNR) at the (u', v', t') focal plane coordinate as a function of exposure time is given by [6]

$$SNR(u', v', t') = \frac{N(u', v', t') \Delta_t}{\sqrt{N(u', v', t') \Delta_t + \kappa_c + \kappa_{dark}}}, \quad (2.7)$$

where κ_c and κ_{dark} are additive noise terms, respectively, induced by the camera ISO settings and sensor dark noise.

Photographers and videographers are routinely forced to optimize the SNR, spatial resolution, depth of field, and temporal resolution. Conventionally, some of these characteristics are compromised to relax design constraints and improve others. In the next section, we show how one may employ coded exposure strategies to dramatically improve a camera's temporal resolution while maintaining a total integration time sufficient for good light collection.

2.2.2 *Motion blur model and camera exposure modulation*

We keep in mind that the primary goal of improving a camera's temporal resolution is to reduce motion blur, thereby being able to capture faster motion. Here we model motion blur as a linear system of equations. Adopting the formulation of [31], an image exhibiting motion blur may be modelled as

$$Y_f = AX_f + \kappa, \quad (2.8)$$

where A is a 'blur matrix' of entries that correspond to convolution with a temporal blur kernel proportional to the temporal pixel transfer function p_t . A is circulant in the case of 1-dimensional motion and block-circulant in the case of

2-dimensional motion. Y_f and X_f are matrices that, respectively, represent the foreground (motion-blurred areas) of the two-dimensional measurement and that of the true (high temporal resolution) image. κ is the measurement noise.

Let us consider the 1-dimensional, vertical motion case for simplicity. In the case of a motion blur of k pixels for a conventional camera, A has circulant columns of k ones trailed by $N_v + k - 1$ zeros, where N_v is the vertical span of the unblurred object foreground X_f . The k ones model the conventional temporal pixel sampling function (2.2). The lack of spectral coverage and nulls in the temporal pixel transfer function (2.3) renders any inversion of (2.8) ill-posed.

To ensure a unique solution for the underlying image X_f , one must maintain a sufficient condition number for A . Flutter shutter (FS) cameras [31] improve the condition number of A by creating a more broadband temporal impulse response without modifying the focal plane array (FPA) architecture. Specifically, an FS camera is fitted with an electronic shutter in front of the objective lens. This shutter can block or transmit light much faster than conventional framerates. Quickly blocking and transmitting light sequences during one captured frame modulates the integration window Δ_t for every pixel in the FPA.

A camera employing FS hardware has its integration time Δ_t divided into M temporal bins of various durations. This results in many ‘chopped’ integration times $\Delta'_t = \frac{\Delta_t}{M}$ during the nominal exposure window Δ_t . If the resulting temporal convolution kernel is comprised of a sum of $N_w \leq M$ shorter exposures of temporal width in units of $l_i \Delta'_t$ ($i \in \{1, 2, \dots, N_w\}$) shifted by temporal offsets φ_i , the conventional pixel sampling function in (2.2) becomes

$$p_{t,Flutter}(t, t', \Delta'_t) = \sum_{i=1}^{N_w} \text{rect}\left(\frac{t - t' \varphi_i}{l_i \Delta'_t}\right), \quad (2.9)$$

where $p_{t,Flutter}$ is modelled in the columns of a new FS-modulated sensing matrix, A' . Columns of this matrix have concatenated binary features of length $\frac{l_i k}{M}$ pixels trailed by zeros.

The improved temporal impulse response is a superposition of shorter-windowed, shifted rectangle functions. Of course, one must jointly optimize the temporal shifts c_i , number of integration window divisions N_w , and integration subwindows $l_i \Delta'_t$ for the best-posed inversion of (2.8). [31, 38] explain the code optimization process in greater detail, but we will neglect it here for the sake of continuity.

Assuming an optimized code sequence, the FS-modulated imager now has a more broadband temporal spectrum (Fig. 2.2) given by

$$\tilde{y}(\xi, \eta, w) = \text{sinc}(\xi \Delta, \eta \Delta) \left(\sum_{i=1}^n \text{sinc}(w \Delta'_t l_i) e^{j c_i w} \right) \tilde{x}(\xi, \eta, w) \tilde{h}(\xi, \eta, 0). \quad (2.10)$$

The superposition of variable-width sinc functions results in an improved singular value spectrum of the new measurement matrix A' . Since A' now has greater invertibility, the pseudoinverse estimator

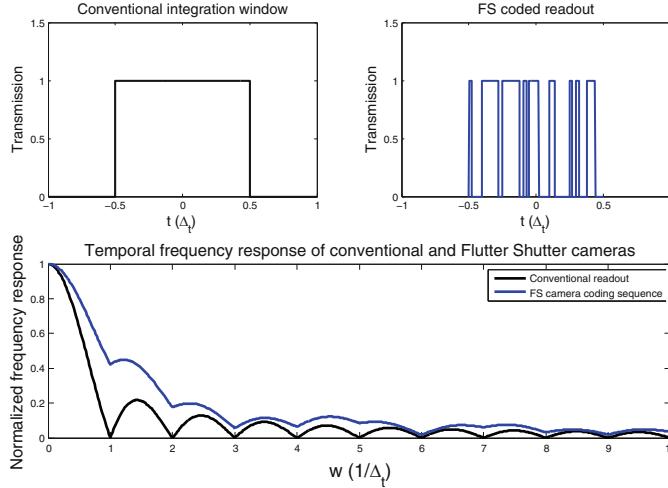


Fig. 2.2 Temporal spectra of conventional cameras and flutter shutter cameras with integration time Δ_t . The simulated coding sequence is shown in the upper right figure

$$X_{f,e} = A'^\dagger Y_f, \quad (2.11)$$

yields a reasonable result for the deblurred image $X_{f,e}$ under the assumption of additive Gaussian noise. Although FS systems may effectively recover fast motion lost from *motion blur*, signals of temporal frequencies $> \frac{1}{2\Delta_t}$ are still *temporally aliased* because of insufficient unique samples (i.e., $X_{f,e}$ still has the same framerate as Y_f) and an unknown aliasing model.

Consider the case of SNR for a desired (fixed) temporal resolution. A conventional imager may acquire M times as many individual frames, achieving $\sqrt{2}$ times the SNR (assuming a 50% duty cycle code, according to (2.7)).

In the case of a fixed operating bandwidth, however, flutter shutter cameras have an SNR advantage (by a factor of \sqrt{M}) over conventional imagers wishing to acquire at the same temporal resolution. In this likely scenario, both cameras suffer from temporal aliasing. One could see adaptive flutter shutter strategies playing an interesting role in future computational photographic cameras.

Coded exposure strategies all rely upon the same principle of exposure modulation to improve an imager's temporal resolution. Similar to the strategies employed in FS cameras, [1, 8, 35] stagger the exposure offsets and durations of *multiple* sensors (rather than divide the integration time) to achieve higher framerate (and hence greater robustness to temporal aliasing *and* motion blur) and better SNR. Of course, this strategy increases system volume and cost by a factor of the number of cameras used and demands image alignment and registration before post-processing. We neglect further discussion on these approaches due to their conceptual similarity to FS cameras.

2.3 Compressive sensing for video

So far, we've seen how reducing temporal fill factor Γ and modulating exposure can improve an imager's temporal resolution. Despite the success of these strategies, they sample and reconstruct images with the same bases (i.e., the canonical pixel basis), and are hence limited directly by the imager's spatiotemporal sampling structures. Since this canonical sample-to-image representation requires one sample per final estimated voxel of $x(u, v, t)$, full datacube capture of the information entering the optical imager (as mentioned in Section 2.1) requires a number of samples equal to the total desired voxel count of $x(u, v, t)$. One may expect this number to be ~ 1 exapixel per second for large-scale imagers [24], which using an 8-bit dynamic range requires an incredible bandwidth of 10^{15} bytes per second.

We can make use of compressed sensing theory established in Chapter 1 to reduce the data load from the imager. CS for high-dimensional signals residing in the space–time volume seems particularly lucrative as it applies to a variety of applications ranging from remote sensing to event capture. We refer to this field as CS-video. We formulate the basic CS-video problem in the following subsection.

2.3.1 CS-video model

One may formulate the CS-video problem by modelling a single measurement of a limited-framerate camera. Consider the low-framerate, motion-blurred images y as combinations of multiple subframes of the sharp video signal x . Intuitively, in order to find those subframes from the images y , we must have: (1) some way to uniquely identify the images (typically implemented via some transmission modulation as shown in Fig. 2.3); and (2) knowledge of our image formation process as a function of time. Here, we develop these principles in the context of CS-video.

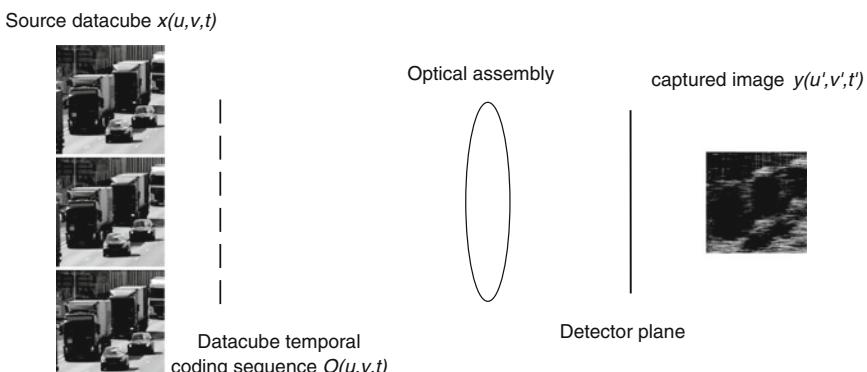


Fig. 2.3 Dynamic scene as imaged by a compressive video camera. The coding sequence O is implemented in an object or image plane of the optical system

Let indices (i, j, q) span the (u, v, t) axes, respectively. When presumably sensing N_F temporal channels (frames) within one detector frame, the $(i, j)^{th}$ pixel of the measurement y_{ij} may be written as a summation of the subframes of the high-framerate video $x_{ijq}, q = \{1, \dots, N_F\}$ along the temporal axis:

$$y_{ij} = \sum_{q=1}^{N_F} x_{ijq} O_{ijq} + \kappa_{ij}, \quad (2.12)$$

where O_{ijq} is the transmission function of the camera optics in space and time (Fig. 2.5). This can be expressed in matrix form as

$$y = Ax + \kappa, \quad (2.13)$$

where $y \in \mathbb{R}^{N_u N_v}$ is a vector consisting of the measurements y_{ij} , $A \in \mathbb{R}^{N_u N_v \times N_u N_v N_F}$, $x \in \mathbb{R}^{N_u N_v N_F}$, and $\kappa \in \mathbb{R}^{N_u N_v}$ are the imager's measurement matrix, vectorized form of the analog object video $x(u, v, t)$, and noise vector, respectively (Fig. 2.4). A is structured to account for effects such as the optical transmission O , the detector pixellation, and the optical impulse response. An isomorphic imager capturing data at a framerate equal to the desired temporal resolution would have $N_F = 1$ and has $A = I_{N_u N_v \times N_u N_v}$.

Herein, we consider the compressive case where $N_F \gg 1$. Since the measurements y have many fewer voxels than the video stream x , any inversion of (2.13) for x is an estimate of the *original video sequence* that impinged upon the detector, but at shorter temporal integration windows. The reconstructed video is expected to have temporal resolution similar to that of FS cameras while also effectively achieving temporal antialiasing.

To accomplish this, CS-video approaches capitalize upon temporal redundancies to improve the feasibility of reconstructing N_F distinct video frames from a single measurement given by (2.13). Doing this requires coding the data $y \in \mathbb{R}^{N_u N_v}$ that is represented as a linear combination of canonical pixel basis functions $\{h_i\}_{i=1}^{N_u N_v}$ with

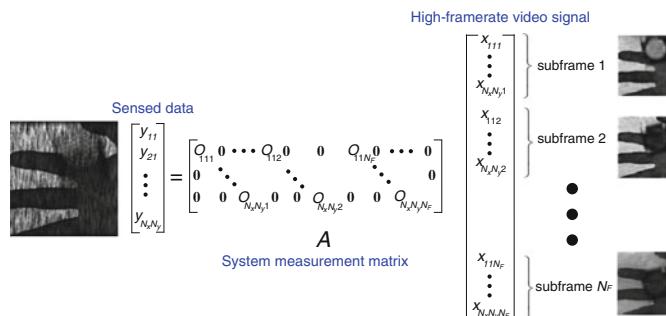


Fig. 2.4 Compressive capture of a high-framerate datastream x . $N_F \gg 1$, hence A is underdetermined

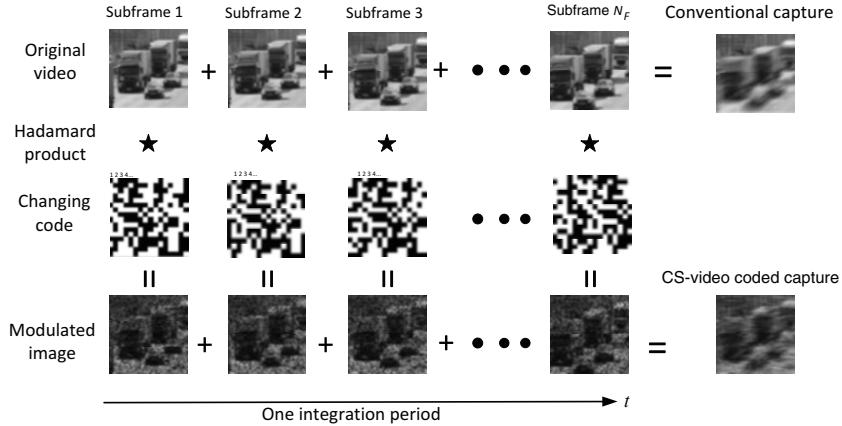


Fig. 2.5 Detection process for a CS-video system. The original video frames are pointwise-multiplied with a changing mask and summed along the temporal axis during the integration period. Taken from [48]

N_F different patterns corresponding to N_F frames one wishes to reconstruct for. To extrapolate the imager's temporal resolution, this process (shown in Fig. 2.5) must occur during a single integration period Δ_t . Doing so requires coding the transmission function O during the integration period (Fig. 2.3).

Following CS theory outlined in Chapter 1, these patterns should be as incoherent with the $k \ll M$ sparsified space-time basis functions $\{\psi_i\}_{i=1}^k$ as possible in order to uniquely preserve the high-dimensional structure of the datacube within the low-dimensional measurement subspace. The underlying goal is to take enough measurements (in the case of compressive video, one snapshot, or $N_u N_v$ measurements) on the coded pixel basis to implicitly capture the $k \ll N_u N_v N_F$ sparsifying basis coefficients needed to characterize x with high probability. One may imagine a transformation of the spatiotemporal bases into Fourier, Wavelet, or DCT bases to greatly reduce k . Users can choose different sparsifying bases based on the properties of the designed signal.

Performing this basis transformation, the coded measurement may be represented by

$$y = A\Psi c + \kappa, \quad (2.14)$$

where A is the underdetermined *forward matrix* representing the object- or image-space time-varying coding patterns applied to the video signal, $\Psi \in \mathbb{R}^{N_F N_u N_v \times N_F N_u N_v}$ is a square sparsifying matrix operator (with basis vectors as its columns) and c is a vector of the basis coefficients (Fig. 2.6).

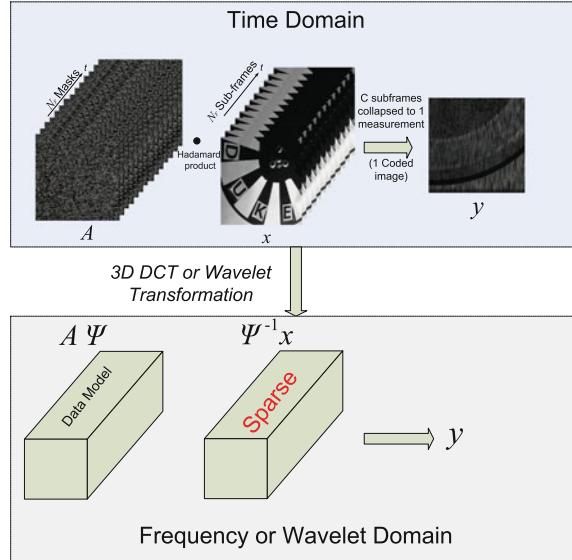


Fig. 2.6 Video-CS sparsifying transformations

Recall the continuous model of (2.13) to focus on coding strategies and determine optimal temporal sparsifying bases, and coding the image plane as a function of time modifies (2.1) to be

$$y(u', v', t') = \int \int \int x(u, v, t) O(u, v, t) h(u - u', v - v') \text{rect} \left(\frac{u - u'}{\Delta}, \frac{v - v'}{\Delta} \right) \times p_t(t, t', \Delta_t) du dv dt, \quad (2.15)$$

where h denotes the optical impulse response, $O(u, v, t)$ is the continuous equivalent of the discrete time-varying coding pattern $O_{i,j,k}$ at the (u, v) spatial coordinates of the video signal x and p_t follows the conventional case of rectangular sampling windows (of large temporal fill factor) in (2.2). Here, the O notation is used instead of modifying p_t since coding may be, in general, applied on a *per-pixel* basis as opposed to globally. $O(u, v, t)$ serves to multiplex a basis-transformed representation of the incident optical field onto the measurements y in an incoherent manner. Modulating O on the pixel level as a function of time increases the spatial incoherence of the measurements relative to the spatial sparsifying basis and is hence advantageous.

Since O varies with time, mechanical, electrical, and structural complexity become design constraints. We wish to implement O using as little additional power as possible, particularly as the number of image pixels scales. Several methods have been employed to do this, such as structured illumination and per-pixel coded exposure.

2.3.2 Active CS-video measurement techniques

Modulating the spatiotemporal datastream at N_F times the original resolution requires some form of physical encoding, which may be understood as a form of Code Division Multiple Access (CDMA), a protocol used to compress radio wave communications data transmitted over wired and wireless channels [33]. The idea of CS-video is to multiplex linearly independent projections of the space–time datacube onto the 2-dimensional sensor. If the multiplexing process is known and is consistent with the structural sparsity of datacube, the demultiplexing process is well-posed and the original temporal channels (frames) encoded onto the sensor may be reconstructed.

Since the encoding occurs in the presumably undersampled dimension (time, in this case), each spatiotemporal voxel must be encoded in the temporal dimension. Active modulation strategies change the state of the hardware used to generate the coding pattern ($O(u, v, t)$) N_F times per captured image. In the following subsection, we describe two active CS-video coding strategies.

2.3.2.1 Structured illumination

The modulation function $O(u, v, t)$ is implemented in the object or image space. Motivated by strobing effects (as often seen in entertainment or dancing events), one may project illumination onto an object in a known, rapid sequence rather than code the camera shutter to increase temporal acuity. We refer to object-space coding as *structured illumination* (Fig. 2.7). Structured illumination differs from conventional strobing in that the temporal spectrum of the exposure modulation

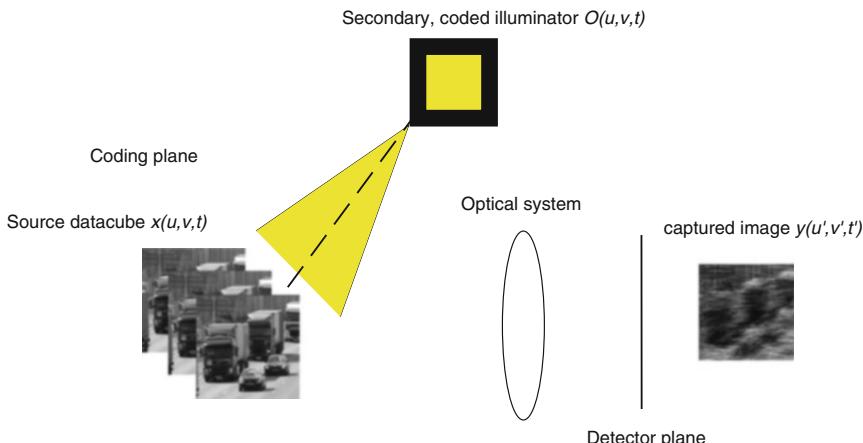


Fig. 2.7 Detection process for a compressive structured illumination system. The secondary light source modulates the image rapidly and at several known, distinct temporal frequencies

has an isolated (i.e., sparse) representation at *numerous* frequencies, as opposed to a single frequency.

[42] projected a coded illumination pattern on the image to capture a summation of realizations of a briefly illuminated scene within Δ_t . The duration of the coded exposure is sufficient for a low-framerate camera to acquire enough light to image dynamic events.

Like all CS-video implementations, structured illumination seeks to sparsify the scene on spatial and temporal bases. In this case, the authors show interest in detecting videos with motion that may be parameterized by

$$x(u, v, t) = x(u(t), v(t)), \quad (2.16)$$

where each spatial coordinate (both $u(t)$ and $v(t)$) varies periodically in time:

$$\{u(t), v(t)\} = \sum_{q=1}^Q a_q \cos(2\pi v q t) + b_q \sin(2\pi v q t), \quad (2.17)$$

where v is the frequency variable (inverse of the period) and Q is an integer. Such signals are bandlimited with a maximum frequency of Qv .

This representation is equivalent to observing signals comprised of a superposition of DFT basis vectors. Since such video signals only have energy in indexed frequencies corresponding to nv , they may be approximated with a maximum of $k = 2Q + 1 \ll N_u N_v$ nonzero DFT basis coefficients, rendering them sparse in time on the DFT basis [42].

The transmission function $O(u, v, t)$ may be expressed similarly to (2.9):

$$O(u, v, t) = \text{rect}\left(\frac{u}{2N_u}, \frac{v}{2N_v}\right) \sum_{q=1}^{N_F} \text{rect}\left(\frac{t - \xi_q}{l_q \Delta'_t}\right). \quad (2.18)$$

which corresponds to illuminating the conjugate object-space position of every detected pixel many times with durations given by $l_q \Delta'_t$ and with shifts ξ_q . Similar to the analysis provided in Sect. 2.2.2, the factor of temporal resolution increase of the coded data is largely dependent upon the shortest continuous integration window within the coding sequence.

Detection of such bandlimited signals suits $y = A\Psi c + \kappa$, with Ψ corresponding to a DFT basis matrix of the coding projections. The resulting measurement is that of a linear combination of a periodic signal's harmonics.

Although we've referred to structured illumination as object-space coding, homogeneous binary pupil coding has the same effect. We revisit the flutter shutter framework previously discussed in Sect. 2.2. Applying CS theory for sparse random sampling in the temporal dimension, CS-flutter shutter cameras [18] use the same coding strategies discussed in Sect. 2.2 but may now reconstruct a video rather than perform temporal deconvolution. Though suboptimal, this hardware configuration is readily implementable with inexpensive off-the-shelf parts.

2.3.2.2 Per-pixel spatial light modulator coded projections

Leaving the premise of coding the illumination, we revisit the concept of coding the camera’s shutter. We’ve seen how flutter shutter cameras can broaden the imager’s temporal bandpass by globally coding the imager. However, globally coding the shutter results in a 100% spatial correlation of the modulation structure imparted upon the signal. Coding the pixels locally, rather than globally, is thereby advantageous from a reconstruction quality standpoint [32]. Pixels may be coded locally through use of spatial light modulators, which are a broad class of devices that use microscale mirrors or liquid crystal technologies that change state or orientation at the pixel level. Two prominent types of spatial light modulators are deformable mirror devices (DMD) and liquid crystal on silicon (LCoS).

The single pixel camera [4] is the canonical example of spatial compression, whereby “codes” in the form of basis patterns are generated on a DMD prior to projection onto a single-pixel sensor. Since different projections are multiplexed at different instances in time, this strategy codes the measurements along the temporal dimension to achieve modest spatial resolution from a single pixel. This impactful work has found a home in areas where sensing hardware is costly, such as the short-wave-infrared (SWIR) regime.

As a successor to this work, [34] trades off spatial recovery for temporal recovery by using spatiotemporally compressive matrices. Despite this tradeoff, high-quality estimates of the spatiotemporal datacube are obtained due to the design of dual-scale [34] and sum-to-one [16] sensing matrices, which enable two recovery modalities: 1.) rapid, low-resolution video previewing via pseudoinverse inversion; 2.) High-spatiotemporal-resolution recovery of the space-time datacube. In the first modality, a pseudoinverse estimation is used to recover the low-resolution video rapidly. This result feeds into the second modality, which exploits the datacube’s spatiotemporal redundancy and uses state-of-the-art optical flow methods to aid in the high-quality video recovery.

Such approaches are very promising for SWIR and other spectral regions for which there exist sensitive modulators, but become difficult for hardware and computation as the desired recovered framerate increases.

[32] extends this strategy to sensors of arbitrary pixel count. We focus on monolithic sensors due to the ubiquitous use in everyday imaging. The authors use an LCoS to modulate the transmission function O in (2.15) as

$$O(u, v, t) = \text{rand}(N_u, N_v, N_F), \quad (2.19)$$

where $\text{rand}(m, n, p)$ denotes an $m \times n$ random binary matrix that is redrawn with Bernoulli(.5) for p distinct realizations. N_F of these random matrices are multiplexed onto the detector during one integration time. Similar to the framework outlined above, the high-framerate video signal x is multiplexed onto the low-framerate data y as expressed in (2.13). The measurement matrix $A \in \mathbb{R}^{N_u N_v \times N_u N_v N_F}$ may be written in augmented-diagonal form with each set of the per-pixel coded modulation function O_{ijq} on the q^{th} diagonal (Fig. 2.4).

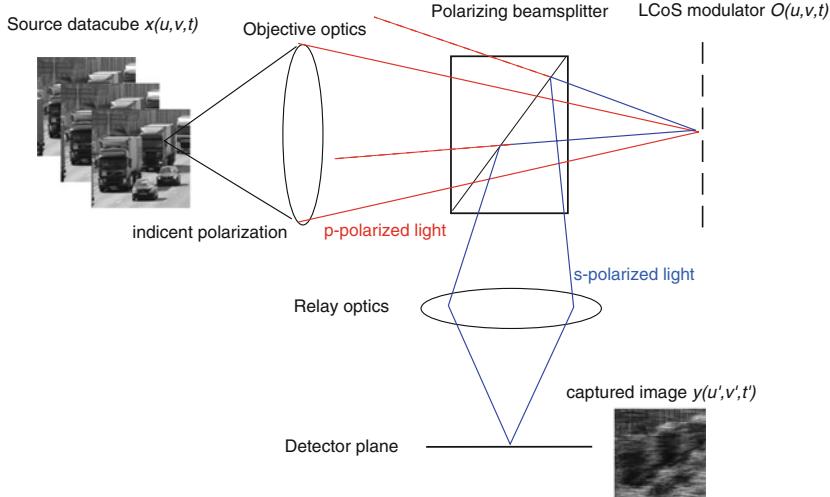


Fig. 2.8 Optical layout of per-pixel LCoS modulation. The LCoS modulator, placed in an intermediate image plane of the optical system, codes the pixel intensities by changing the polarization of the incident light. ‘On’ pixels maintain the current polarization and send the incident light back into object space; ‘off’ pixels change the polarization and transmit the light to the detector. Each pixel is modulated ‘on’ and ‘off’ many times during the integration window

LCoS hardware consists of an objective lens, polarizing beamsplitter, LCoS device, relay lens, and detector. The objective lens images the incident light onto the LCoS which operates by rotating the polarization state of the incident light via a retarding electro-optic modulator (for details on polarization and liquid crystal modulators, see [37]). Orthogonally polarized (s) light corresponds to ‘on’ LCoS pixels and is relayed by the relay lens onto the detector. ‘Off’ LCoS pixels maintain the incident (p) polarization state that travels back through the polarized beamsplitter and back into object space. See Fig. 2.8 for a pictorial view of the setup used by [32].

The temporal resolution of this strategy may be evaluated the same way as in Sect. 2.2.2, except each pixel now has an independent coding pattern. Assuming square pixels and code features, the resolution of the coded data acquired by such strategies for N_F LCoS modulations is given by the space-time Fourier transform of (2.15):

$$\begin{aligned} \tilde{y}(\xi, \eta, w) &= \text{sinc}(\xi \Delta, \eta \Delta) \text{sinc}(w \Delta_t) \tilde{h}(\xi, \eta, 0) \\ &\times \int \tilde{x}(\xi - \gamma, \eta - \gamma, w - N_F \gamma) \hat{O}(\xi, \eta) d\gamma \end{aligned} \quad (2.20)$$

where $\hat{O}(\xi, \eta)$ is the 2D spatial Fourier transform of the LCoS mask and γ is the temporal frequency component imparted by the LCoS mask on the object video. The reconstructed video’s temporal resolution is increased by a factor of N_F in

time and $\frac{1}{\Delta_c}$ in space, where Δ_c is a unitless quantity denoting the code feature size (in detector pixels). The temporal pixel transfer function $\text{sinc}(w\Delta_t)$ remains the same but is convolved in time with the product of the mask spectrum and a frequency-shifted version of the datastream. The time-varying modulation aliases high temporal frequencies of x into the passband of the detector sampling functions. Ideally, $\Delta_c = \Delta$ and no spatial resolution is lost.

Notably, spatiotemporally random pixel-wise modulation results in inhomogeneous temporal resolution across the detected image. For example, completely random codes will sample some spatial areas with more short sequences (corresponding to a very broadband frequency response); other areas may be ‘off’ (corresponding to temporal aliasing) or ‘on’ (corresponding to motion blur) for a long period of time due to low and high pixel-wise temporal fill factors, respectively. However, the hardware was designed such that each multiplexed data frame y received $\sim 50\%$ of the original signal and these errors mostly averaged out.

Beyond the $\sim 50\%$ light throughput penalty associated with the transmission function $O(u, v, t)$, the beamsplitter inherently results in an additional loss of $\sim 50\%$ of the incident light, assuming unpolarized incident illumination (this loss is reduced for incidentally p-polarized objects or scenes, which may arise naturally from interactions with polarizing materials such as windows, the sky, and water). Despite this loss of light, in the case of a fixed imaging bandwidth, the average SNR of the compressed video data is approximately improved by a factor of $\sqrt{N_F/4}$ relative to a conventional camera taking a single exposure of duration $\frac{\Delta_t}{N_F}$. These SNR and effective framerate benefits render such a coding strategy highly desirable.

Rather than projecting many completely random patterns onto the detected image during one frame, [17] used a pre-designed sampling structure and an overcomplete dictionary to approach the problem. The LCoS was configured such that each of N_F detector pixel within an $\sqrt{N_F} \times \sqrt{N_F}$ region was exposed one fixed, very brief duration $l_d\Delta_t/N_F$ during the integration time Δ_t . The factor of l_d was empirically chosen between 2 and 3 to improve light throughput while maintaining a broadband temporal frequency response. This was designed to mimic an actual hardware implementation of such a coded sampling sequence into CMOS detector hardware since most cameras currently do not have per-pixel frame buffers.

One unanswered question in the CS-video literature is ‘what is the optimal generic dictionary for sparse temporal representation’? [17] utilized a similar coding strategy as that just discussed, with the addition of constructing an overcomplete dictionary with 10^5 atoms to sparsify the coded data. Storing these dictionary atoms as columns of a matrix D , the underlying object video x may be written as $x = Dc$, where c is now a the vector of dictionary atom coefficients representing N_F high-speed underlying video frames.

These strategies successfully compressively capture the object video but require sending data to each modulator pixel N_F times during the exposure, resulting in binary data transmission to $N_F \times \frac{N_u}{L} \times \frac{N_v}{L}$ LCoS pixels per captured frame, where L is the magnification of the LCoS pixel size onto the detector. As previously mentioned, L should be kept as small as possible to increase the spatial incoherence of the transmission pattern with that of the landscape, and to decrease the spectral support of $\hat{O}(\xi, \eta)$.

The power consumed by LCoS modulators is proportional to the pixel count and operating framerate. Though such modulators can sustain incredible rates of > 1 kHz (binary masks stored in memory) for HD display, the power requirement becomes unsustainable for imagers of larger information capacities as in [9]. Ideally, such CS-video sampling schemes would be implemented directly on the focal plane to avoid additional hardware and power overheads [17].

2.3.3 *Coded aperture compressive temporal imaging: a passive CS-video technique*

We now arrive at the fundamental question: How should we compressively sample the space-time optical datastream in a way that is economical and retains high-fidelity representations of the dynamic scenes of interest? Rather than using an LCoS modulator, [24] employs a coded aperture to modulate the spatiotemporal structure of the datastream x . Coded apertures have been used in spectroscopy and mass spectroscopy [29], spectral imaging [15, 21, 40, 43, 44], x-ray tomography [26, 27], compressive radar [19, 45], and spatial superresolution [3, 4] applications. The coded aperture is widely applicable because transparency and opacity are physically feasible phenomena at all portions of the spectrum, in contrast to reflection and polarization modulation. At shorter wavelengths, a high-density material (such as lead) with holes cut inside can suffice as a coded aperture; visible wavelengths can use arrays of opaque and clear features (such as chrome and quartz); longer wavelength, coherent imaging systems can utilize arrays of metamaterial receivers as coded apertures.

As per the CS theory, one may recover information lying in dimensions the detector is not sensitive to, provided the code features are uncorrelated with a sparse representation of x in that dimension. A coded aperture-based modulation approach may be used to improve the temporal bandpass of the imager in the same way as the LCoS-based modulation strategy discussed in Sect. 2.3.2.2. Changing the code projections onto the sensor is accomplished by physically translating, via a low power method, the mask transverse to the detector plane as a function of time (Fig. 2.9(a)). A prototype implementation of this coding strategy, dubbed ‘CACTI’ (Coded Aperture Compressive Temporal Imaging), has demonstrated compressive temporal superresolution by a factor of ~ 10 [24].

Coded aperture-based modulation shares many commonalities with active modulation. Both CS-video approaches:

1. Modulate the exposure multiple times per exposure, and possibly on a per-pixel basis.
2. Exploit the temporal redundancy of ambient space-time datacubes.
3. May employ the same reconstruction algorithms for data estimation.
4. Tend to use binary codes to emulate the true states of open versus closed shutters.
5. Require additional hardware for implementation.

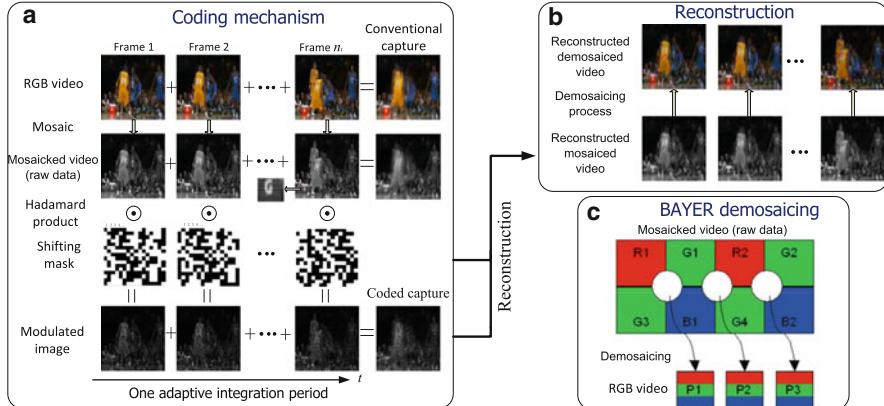


Fig. 2.9 CACTI image acquisition process. Demosaicing occurs after reconstruction to avoid interpolating coded pixels [52]

Despite these similarities, this proposed passive modulation strategy offers numerous advantages relative to reflective or active media:

1. Low power consumption. A typical micromirror device or liquid crystal modulator will consume $\sim 1\text{--}10\mu\text{W}$ per pixel - on the same level of the camera itself - which becomes unwieldy at higher pixel counts.
2. Coded aperture can be applied across a great portion of the electromagnetic spectrum:
 - a. Coded apertures exist in the form of masks with holes for x-rays and IR imaging, metamaterial receivers for THz imaging [19, 45], and chrome-on-quartz for visible wavelengths.
 - b. Active coding methods (LCoS/structured illumination) may/do not have materials or illumination readily available at other regions of the spectrum; micromirrors are generally optimized in terms of reflectivity and contrast for the visible and near infrared spectral regimes (250–2700nm) and degrade substantially outside it.
3. An in-line imaging system architecture. A reflective system by definition increases system volume. Active modulators are also preferably reflective due to their larger fill factor ($\sim 90\%$ versus $\sim 60\%$ for transmissive devices).
4. Greater light throughput than strategies employing beamsplitters to separate the light path. Beyond the $\sim 50\%$ light throughput penalty associated with the transmission function $O(u, v, t)$, the beamsplitter inherently results in an additional loss of $\sim 50\%$ of the incident light (as discussed in Sect. 2.3.2.2).

Table 2.1 Comparison of CS-video architectures

	Flutter Shutter	Active illumination	Per-Pixel Coding	CACTI
Cost	Low	Moderate	High	Low
Temporal bandwidth	High	Moderate	High	High
Power consumption	Low	Moderate	High	Moderate
Spatial pattern flexibility	Low	Moderate	High	Low

There are also drawbacks:

1. Reduced control over the coding pattern. Active strategies have direct control over every pixel within the pattern. We have not observed significant degradation of the video quality in our system.
2. Faster modulation rates. Several active architectures utilize onboard memory transmission to obviate streaming. Such modulators typically utilize a “non-video” display modality that enables binary pattern display at the kHz range - a substantial improvement over mechanical translation, which is limited by the response time and stroke of the translator.

Data sparsification proves the ultimate limit in obtainable temporal compression ratios. Typically this level has maximized at $\sim 10\text{--}30$ for “good” reconstructions, depending on the temporal redundancy of the scene [17, 32].

A table depicting the differences between the various CS-video architectures is shown in Table 2.1.

Bearing these similarities and differences in mind, we now focus on the physical process of this passive coding technique [25].

The CACTI image detection process is depicted in Fig. 2.9. The mosaiced RGGB color channels of the discrete space-time source datacube are multiplied at each of N_F temporal channels with a shifted version of a coded aperture pattern. Each detected frame y is the temporal sum of the coded temporal channels and contains the object’s spatiotemporal-multiplexed information. Reconstruction is performed directly on the mosaiced, coded data to avoid interpolation errors on dark areas of y that are attributed to the attenuation of the coding pattern. The RGGB reconstructed, mosaiced videos are interleaved to form the final estimate.

Instead of modulating each pixel of an LCoS modulator at $\frac{N_F}{\Delta_t}$ Hz, CACTI systems implement physical-layer compression via the physical motion of the coded aperture during the exposure period. This analog modulation strategy has the advantages of reduced power consumption, reduced system size, and a power requirement that remains constant with increasing camera pixel count. [24] showed that a translation of a fixed, random code proved nearly equivalent (in terms of reconstruction PSNR) to generating each code pattern independently. From this standpoint, having a fixed coding pattern does not translate to temporal redundancy along the spatial dimensions. The temporal incoherence of such masks with bases such as DCT and overcomplete dictionaries appears to be similar. To the best of the authors’ knowledge, optimal coding patterns for CS-video have yet to be found.

CACTI's forward detection process is similar to (2.15). Translating the mask with (u, v) motion parameterized by $(r(t), s(t))$ yields the model

$$y(u', v', t') = \int \int \int x(u, v, t) O(u - r(t), v - s(t)) h(u - u', v - v') \times \text{rect}\left(\frac{u - u'}{\Delta}, \frac{v - v'}{\Delta}\right) p_t(t, t', \Delta_t) du dv dt. \quad (2.21)$$

We assume the coded aperture moves (in one direction) linearly during Δ_t such that $r(t) = 0, s(t) = vt$, where v is in units of pixels per second.

The discrete formulation of (2.21) is equivalent to (2.12). The transmission function O is implemented as a random binary mask

$$O_{ijq} = \text{rand}(N_u, N_v, s_q), \quad (2.22)$$

that physically translates as a function of time to the q^{th} discrete position s_q . The mask has features randomly drawn from a Bernoulli(.5) distribution.

The resolution of the CACTI-coded data is given by the Fourier transform of (2.21), which produces the same result as that obtained by active per-pixel modulators in (2.20). The coded aperture is used to divide the integration time into $\frac{v}{\Delta_t}$ distinct time slices and alias temporal frequencies v times the original cutoff frequency $\frac{1}{\Delta_t}$ into the detector passband.

Importantly, the increase in temporal bandpass is directly proportional to the code translation speed. One may increase v simply by translating the code a greater distance during the nominal integration time Δ_t [24]. As the desired framerate increases, this process consumes substantially less power than modulating $\frac{N_F N_u N_v}{\Delta_t}$ pixels per second via an electronic modulator.

Like the per-pixel-modulated CS-video systems, CACTI systems image the spatiotemporal scene x onto the coding element, which translates while the scene action progresses. The spatiotemporally modulated scene at the code plane is subsequently imaged onto the detector. Fig. 2.10 shows the hardware used for the CACTI prototype, which includes an objective lens, the coded aperture mounted upon a piezoelectric stage, an achromatic relay optic, and a CCD detector.

The quality of the relay optic is of crucial importance since the code features should be as small as possible. An achromatic, multi-element relay lens is desired to mitigate the effects of optical aberrations and ensure the code features are imaged with high fidelity.

The piezo is driven with an arbitrary or periodic waveform during the integration time. Although linear motion has proven favorable due to its implementation simplicity and it has yielded high-quality reconstructions, designing other waveforms may be a topic of interest in future systems. Adapting the waveform amplitude to the temporal complexity of the scene has also shown promise [51] and is addressed in Sect. 2.4.3.

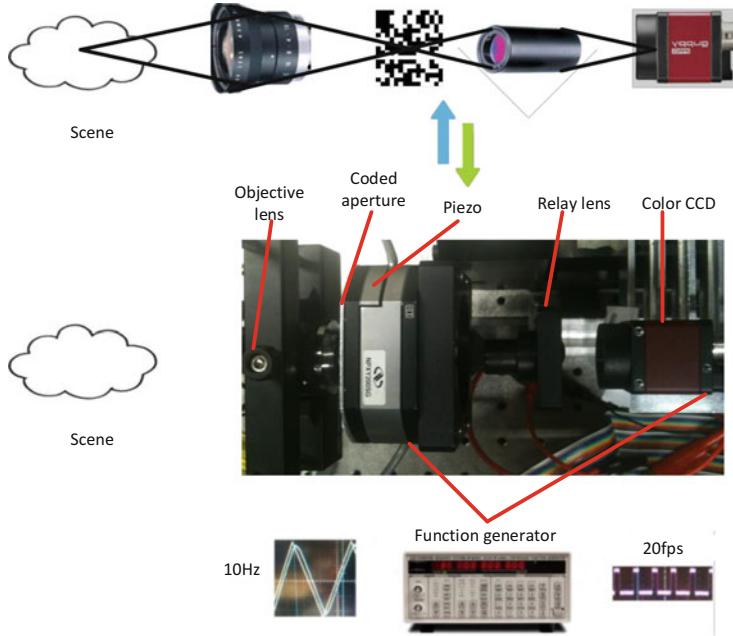


Fig. 2.10 Color CACTI prototype camera

Importantly, the discrete code motion as a function of time can be modelled at an arbitrary number of positions determined by the accuracy of the piezo and calibration procedure. [24] reported $N_F = 148$ reconstructed frames from a single image. Since the piezo velocity ultimately limits the increase in temporal bandpass, reconstructing $N_F > \frac{v}{\Delta t}$ frames per image results in temporal interpolation and smoothing of the perceived motion.

2.4 CS-video reconstruction algorithms and adaptive sensing frameworks

Now that we've provided an overview of the CS-video sensing process, we turn to the recovery of the compressed video data. This involves a series of steps detailed in Fig. 2.11. Importantly, there are two basic reconstruction modalities: 1.) individual measurements may be reconstructed independently; 2.) A sequence of measurements may be reconstructed within one pipeline. The latter strategy lends itself to adaptive reconstruction techniques that modify the initial prior information placed upon the video sequence x . Additional motion estimation via means of optical flow or other techniques can be added into this pipeline. [32, 34] use optical

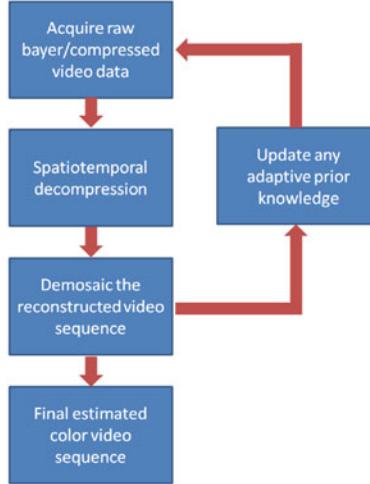


Fig. 2.11 Process undertaken for decompressive inference of compressed video data. A single measurement or multiple measurements may be reconstructed concurrently

flow to regularize the inversions, resulting in high-quality reconstructions. Although we omit specific details of optical flow here, we direct the interested reader to these texts.

If a binary coding pattern is applied at the pixel level in the image space, the individual color channels are aliased [6]; it is beneficial to demosaic the reconstructed datacube after decompressive inference for two reasons: 1.) The interpolation methods conventionally employed during demosaicking cannot faithfully reproduce the coding pattern among the three color channels; 2.) The computational complexity is reduced by a factor of ~ 3 . Since the color channels on a Bayer sensor generally contain quite redundant information, joint reconstruction and demosaicking of CS-video data may lead to substantial reconstruction improvements. For now, we focus on the first step: spatiotemporal decompression.

Sect. 2.3 describes several methods to compressively sample the video sequence x in space and time. Classically, CS applications establish a relationship between x and y through random projections [13, 20, 41]. Specifically, assuming x is compressible in the Ψ basis, the q^{th} compressive measurement $y_q, q \in \mathbb{R}^{N_x N_y \times 1}$ may be constituted by projecting each of the $N_x N_y N_F$ voxels of x onto a “random” basis (i.e., a random linear combination of the basis functions in Ψ). Enough (in the case of CS-video, much fewer than $N_x N_y N_F$) of these “random” basis measurements prove sufficient for recovery of the entire datacube x represented in the Ψ basis as a vector of coefficients c .

During decompressive inference of the coded data y , we must invert the linear transformation (2.14) for the basis coefficients. The solution is an approximation to the basis-transformed version of x . This relationship holds with high probability if RIP is satisfied (see Chapter 1).

Since, in general, y is an undersampled version of the underlying signal x , inversion for the entries in c (and hence x) is ill-posed. However, if one exploits the fact that c is sparse with respect to the Ψ basis, then one may estimate the basis coefficients c of the signal x accurately [13, 20]. If c is sufficiently sparse, a typical means of solving such an ill-posed problem is via an ℓ_1 -regularized formulation

$$\hat{c} = \arg \min_z \|y - A\Psi z\|_2^2 + \tau \|z\|_1, \quad (2.23)$$

where the decision variable z represents the sparse coefficients c as they vary by iteration and the scalar τ controls the relative importance of the fitting and penalty terms. This basic framework has been the starting point for several recent CS inversion algorithms, including linear programming [11] and greedy algorithms [39], for a point estimate of the basis-transformed vector c [20]. Importantly, other terms may be added to further regularize the problem as well; for example, [32] added a term to regularize the brightness constancy relationship for optical flow.

Gaussian mixture models [48], two-step iterative shrinkage thresholding [5], generalized alternating projection [23, 24] and linearized Bregman [10, 30] have been used to reconstruct CS-video data with high fidelity. Although most of these are global algorithms, the Gaussian mixture model (GMM) algorithm proposed in [48] is a patch-based algorithm and has advantages compared with the global algorithms. Section 2.4.2 describes the benefits in detail (Figures 2.12–2.15).

For CS-video results, we seek an algorithm that can reconstruct the data in nearly real time. The Generalized Alternating Projection (GAP) algorithm is fast and can invert the data in a set of transform bases. We detail GAP and its specific uses in CS-video below.

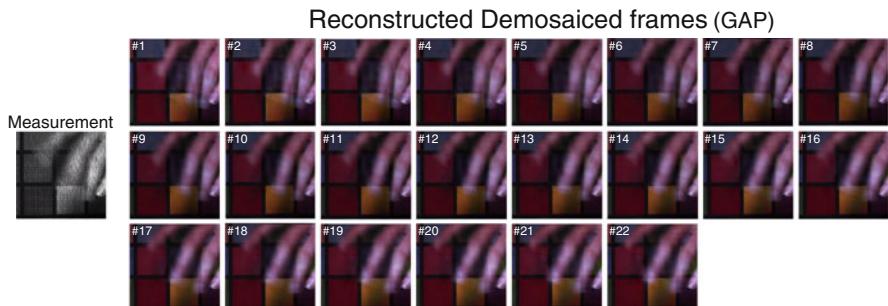


Fig. 2.12 Coded image and reconstructed space-time datacube of a hand moving rapidly in front of an X-Rite color checker chart

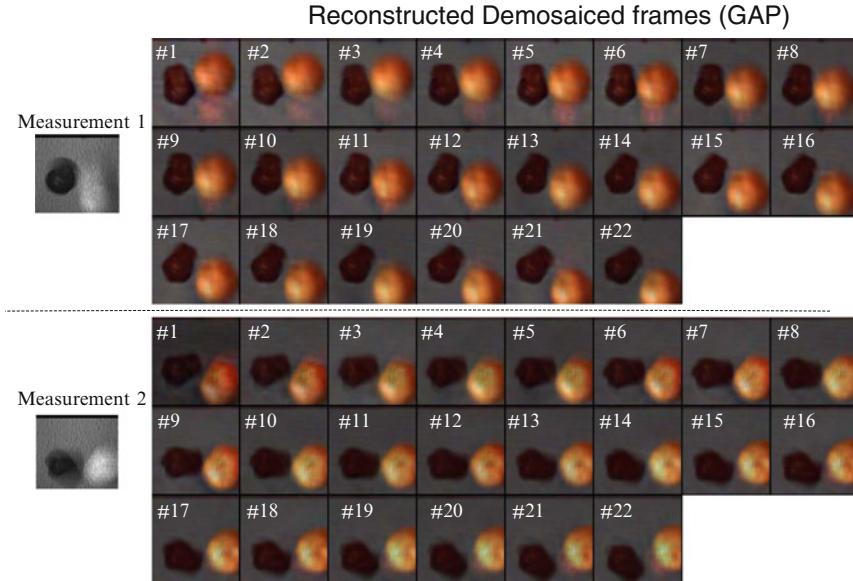


Fig. 2.13 Two fruits bouncing off the tabletop, reconstructed in high-speed. Taken from [52]

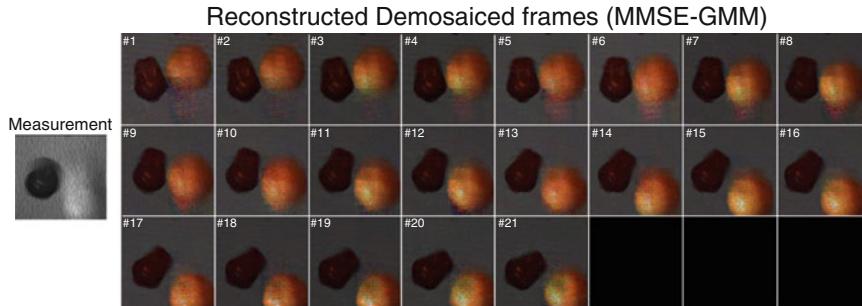


Fig. 2.14 The same high-speed video reconstructed using the GMM-MMSE estimator. $k = 20$ Gaussian kernels were used as dictionary atoms

2.4.1 Generalized alternating projection

A dynamic scene (video) is modulated by the coded aperture during the integration window Δ_t . An estimate of the dynamic video sequence discretized into N_F frames may be formed by solving

$$\min(C) \text{ s.t. } \|\Psi c\|_{\ell_{2,1}} \leq C \text{ and } A\Psi c = y, \quad (2.24)$$

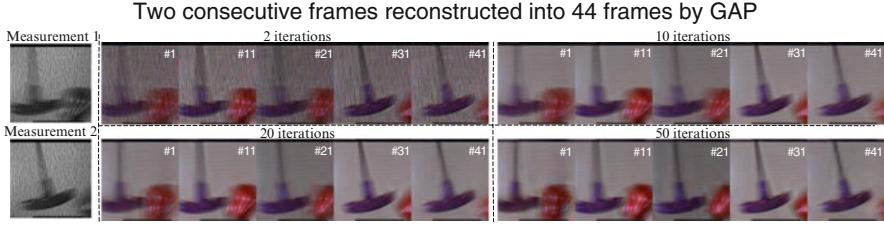


Fig. 2.15 Two separate measurements of a screwdriver hitting an apple. Note how the reconstructions improve as the number of iterations increases

where Ψ is an orthonormal basis transformation matrix with basis coefficients c . C represents the size of a weighted $\ell_{2,1}$ ball governing the sparsity of the coefficients in space and time [52]. GAP begins with a zero vector of $\ell_{2,1}$ group-wise shrinkage weights $\theta^{(0)} = \mathbf{0}$ of the basis coefficients c . This vector is first projected onto the linear manifold $y = A\Psi c$:

$$c^{(\beta)} = \operatorname{argmin}_{A\Psi c=y} \|c - \theta^{(\beta-1)}\|_2^2, \quad (2.25)$$

where $\beta \geq 1$ represents the current iteration number. The resulting coefficient vector $c^{(\beta)}$ is consequently re-projected onto the weighted $\ell_{2,1}$ ball of changing size (θ) as a form of shrinkage [23]:

$$\theta^{(\beta)} = \operatorname{argmin}_{\theta} \|c^{(\beta)} - \theta\|_2^2. \quad (2.26)$$

This process iterates between (2.25) and (2.26) until either $\|c^{(\beta)} - c^{(\beta-1)}\|_2 \leq \gamma$ (γ is a small constant) or the user stops the program manually. The final estimate is assigned to the last iteration's estimate: $\hat{c} = c^{(\beta)}$.

GAP makes use of structural sparsity by dividing the datacube to be estimated into $b_u \times b_v \times b_t$ groups (Fig. 2.19(a)). The wavelet basis, used to sparsify in the spatial domain, has coefficients grouped into neighborhoods of $b_u \times b_v$ that are assigned a weight based on their spatial scale (Fig. 2.19(b)). The DCT basis sparsifies the temporal dimension and is grouped into bins b_t units long. The grouped DCT coefficients are assigned weights based on their temporal frequency (Fig. 2.19(c)).

Fifty iterations of this process typically yield high-fidelity results. For our color datacube size of $512 \times 512 \times 3 \times 22$ voxels, unoptimized Matlab code runs the algorithm in ~ 40 seconds on an Intel Core-i5 CPU with 16GB RAM running at 3.3 GHz. GAP's monotonic convergence properties [23] guarantee that each iteration will improve the result.

2.4.2 Gaussian mixture models for CS-video

During the last decade, dictionary learning algorithms have emerged a powerful tool for image processing, including denoising [2, 14], inpainting [28], and CS [53]. Rather than using a generic dictionary basis transformation such as DCT or wavelets, these algorithms use patch-level basis vectors (or atoms) of candidate scenes obtained from given training data [2] or directly from the corrupted image itself [14]. These dictionary learning algorithms can achieve remarkably sparse representations of a scene provided that, for all patches within the image, at least one of the patches within the training data resembles the given patch.

More recently, Gaussian mixture models (GMM) have been used for dictionary learning [12, 46, 47, 49, 50]. By representing the image as a mixture of two-dimensional Gaussian distributions of adaptive mean and covariance matrices, the underlying signals may be approximated as “block-sparse”; a notion that has led to many state-of-the-art results in image processing.

In this section, we focus our attention on GMMs for CS video [46, 47]. We model the patch-level images as a mixture of k Gaussian distributions

$$x_l \sim \sum_{q=1}^k \lambda_q \mathcal{N}(x_l | \mu_q, \Sigma_q), \quad (2.27)$$

where λ_q , μ_q , and Σ_q denote the weight, mean, and covariance matrix of the q^{th} Gaussian distribution in the mixture, with $\sum_q \lambda_q = 1$. These parameters, which are constituted by the training data, can be estimated by the expectation-maximization algorithm.

Assuming the noise is distributed according to $\kappa \sim \mathcal{N}(0, R)$, the data conditioned on the mixture of Gaussians is distributed according to

$$y_l | x_l \sim \mathcal{N}(y_l | Ax_l, R). \quad (2.28)$$

Applying Bayes' rule, we obtain a posterior distribution that is also a GMM:

$$p(x_l | y_l) = \sum_{q=1}^k \tilde{\lambda}_q \mathcal{N}(x_l | \tilde{\mu}_q, \tilde{\Sigma}_q), \quad (2.29)$$

with parameters given by

$$\tilde{\lambda}_q = \frac{\lambda_q \mathcal{N}(y_l | Ax_l, R + A \Sigma_k A^T)}{\sum_{r=1}^k \lambda_r \mathcal{N}(y_l | x_r, R + A \Sigma_r A^T)}, \quad (2.30)$$

$$\begin{aligned} \tilde{\Sigma}_q &= (A^T R^{-1} A + \Sigma_q^{-1})^{-1}, \\ \tilde{\mu}_q &= \tilde{\Sigma}_q (A^T R^{-1} y_l + \Sigma_q^{-1} \mu_q). \end{aligned} \quad (2.31)$$

One may utilize the MMSE estimator

$$\mathbb{E}[x_l|y_l] = \int x_l p(x_l|y_l) dx_l = \sum_{q=1}^k \tilde{\lambda}_q \tilde{\mu}_q \equiv \hat{x}_l, \quad (2.32)$$

to invert every patch in the image.

This Bayesian framework has the key advantages of prior information on the video x and patch-level processing. Due to the temporal redundancy of ambient scenes, one expects reconstructions that can improve as the algorithm learns from past reconstructions.

Since the GMM prior (2.27) depends on the training data, we can use recent reconstructions to gradually replace unused portions of the fixed training data.

The eigenvectors of each GMM component's covariance matrix define the orthonormal basis used to sparsify the data. As the prior information is updated, the changing posterior distributions enable the algorithm to essentially *adapt* the sparsifying bases to the spatiotemporal structure of the scene, thereby improving the reconstructions. Fig. 2.17 demonstrates the quality of this online-learned GMM reconstruction algorithm. A simulated dataset of 240 high-speed video frames of traffic on a busy highway are compressed and coded via the CACTI framework into 30 frames ($N_F = 8$). The GMM algorithm improves over time; note the smaller errors in Fig. 2.17(d) versus those in Fig. 2.17(b)[48]. Importantly, when updated online, this algorithm can outperform fixed-basis algorithms such as GAP [23] and TwIST [5] after acquiring a few frames of data.

Additionally, the patch-based nature of this inversion strategy suggests great parallelizeability; implementation of such an approach on GPUs can enable unprecedented reconstruction speeds.

The aforementioned GMM method depends heavily on the training data. While the training data should ideally have identical statistical properties as the signals being recovered, one can only find ones that have similar statistics in practice. For some applications, it is relatively easy to find good training signals; in others (e.g., CS-video or CS hyperspectral imaging), however, finding good training signals is a great challenge. On account of this issue, [46] presented an alternative approach to learn the signals from the compressed measurements by maximizing the marginal likelihood of the GMM given only the measurement vectors $\{y_l\}$ with the true signals $\{x_l\}$ treated as latent random vectors and marginalized out of the likelihood. This approach unifies the task of learning the GMM and the task of signal reconstruction in a single model, and its inference is accomplished by pursuing a rigorous expectation maximization of the marginal likelihood. In experiments, we observe the better performance of [46] over that of [47, 48].

2.4.3 Adaptive CS-video strategies

The results indicate that inverting (2.13) is more feasible for smaller degrees of motion blur. For example, 1 of the 4 fingers of the hand was reconstructed sharply; the other 3 had residual motion blur. Scenes producing residual blur in the reconstructions have information at the higher end of the temporal spectrum and hence are still attenuated in the reconstructions despite the increase in bandpass.

To reconstruct such results with high fidelity, more measurements or a better sparsifying basis are required. To address the issue of number of measurements needed for a given reconstruction quality, [51] proposed the concept of varying the integration time of the camera given a fixed mask translation rate.

Most CS-video techniques guarantee that each frame of coded data y will receive the same spatial coding pattern when summed along the temporal axis. Because of this, block matching may be performed directly on the compressed data frames y to estimate the severity of the motion blur. If a fixed reconstruction peak signal to noise ratio (PSNR) is desired, higher the inter-frame block motion (in pixels) necessitates reduced compression ratios in order to faithfully reconstruct the image.

Figs. 2.16–2.18 demonstrate the concept of an adaptive integration time, with the measurement taken using the CACTI coding strategy. Note that this may be extended to the active per-pixel CS-video techniques very easily. A 360-frame video sequence of traffic (the same used in 2.4.2 but extended) on a busy highway is synthetically stopped in the middle and is captured and encoded at various levels of compression. The reconstructed frames and their statistics are shown in Figs. 2.18 and 2.16, respectively.

The number of compressed frames N_F depends on the velocity estimated by the block matching algorithm and on the desired PSNR. [51] employs a lookup table to determine the optimal integration period for the next few frames. In this demonstration, the compression ratio of the captured video is set to target a PSNR value of 22 dB.

Compression ratios as high as 16 can achieve this PSNR during periods of zero motion, resulting in a significant decrease in data load and a significant increase in measurement SNR attributed to the $16 \times$ increase in integration time.

This adaptively determined sampling strategy also benefits from an inversion algorithm capable of adapting to the properties of the data is also desirable. The Gaussian Mixture Model mentioned in 2.4.2 used an adaptive prior to improve results as time progressed.

The last adaptive component is the coding structure. [48] proposed a patch-based transmission function T and forward matrix A employing 8×8 pixel patches. (2.12) can be hereby applied at the patch level rather than globally.

The per-pixel CS-video frameworks benefit from a patch-level transmission function that is implemented on an LCoS modulator as in Sect. 2.3.2.2. Mounting an LCoS modulator onto a piezo enables maximum coding flexibility at little power consumption. Rather than rapidly changing the transmission function on the LCoS, one may instead use the CACTI coding strategy of mechanical translation to move

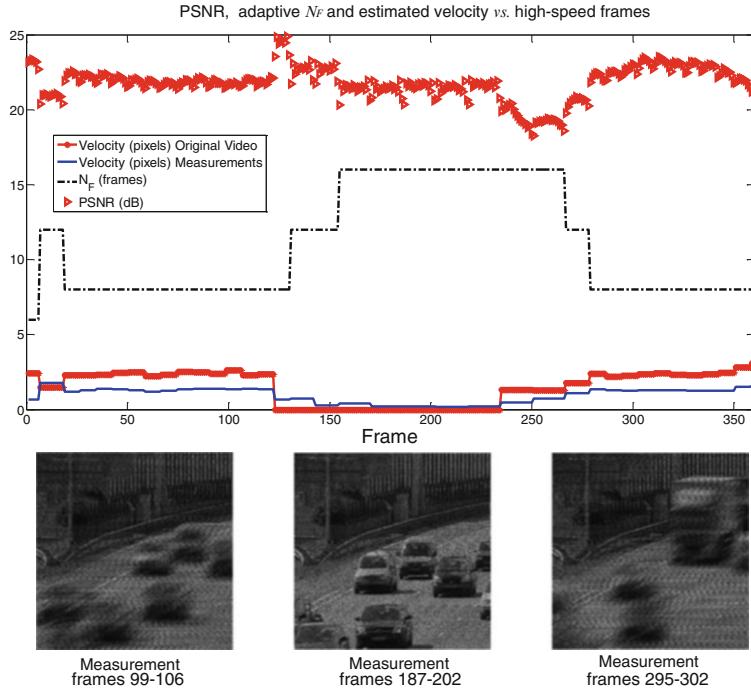


Fig. 2.16 Object velocity and reconstruction statistics for the simulated scene in Fig. 2.18. The dotted red and blue lines at the bottom represent the maximum velocities in each frame found via the block-matching algorithm. The dotted black line represents the number of frames N_F compressed into a single measurement. The target reconstruction PSNR is 22 dB [51]

a locally (in time) fixed LCoS during the exposure window. The LCoS may then be updated at a modest rate of 1 Hz to increase the diversity of projections of the object onto the detector (Fig. 2.19).

2.5 Video frames recovered from a snapshot

In this section we show images taken from the color CACTI prototype color camera. The setup (Fig. 2.10) is the same as that used in [24] with the exception of the detector, which has been replaced with an AVT Guppy PRO, a color camera with $5.6\mu\text{m}$ pixels. All images are taken ~ 3 feet away from the camera.

The coded data and reconstructed frames are shown side-by-side for a variety of scenes captured by the prototype camera at a rate of 20fps. $N_F = 22$ frames are reconstructed per snapshot. The factor of temporal compression here is $\frac{v}{\Delta_c} = 11$; hence, *the reconstructed data has an effective framerate of 220fps*. All reconstructions presented here are performed by GAP.

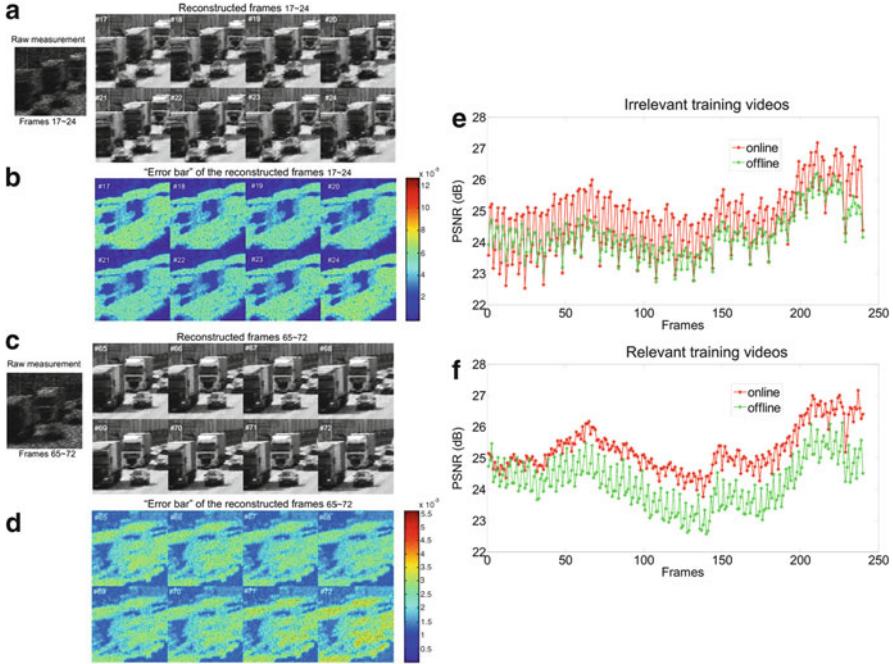


Fig. 2.17 (a-d) Online-learned GMM reconstructions and error bars (relative to original data) of a 240-frame simulated dataset of traffic. (e-f) Performance comparison of online-learned vs. offline reconstructions using irrelevant and relevant initial training data [48]

The three scenes of interest consist of: 1) A hand waving rapidly in front of a color checker background (Fig. 2.12); 2) two toy fruits rebounding after hitting the ground (Fig. 2.13); 3) A purple hammer hitting a toy apple (Fig. 2.15).

The severe motion blur (~ 70 pixels for the left index finger at the top) of the hand is reconstructed into discernable fingers. Note the background and nearly stationary pinky finger are kept sharp during the reconstruction.

Fig. 2.13 shows two consecutive 20fps measurements of fruit (dropped from a height of 1 foot above the ground) bouncing off the tabletop. Back-to-back, these measurements appear discontinuous, but the reconstructed frames paint a clear representation of what happened during the fall. The severe motion blur of the orange (> 200 pixels) renders some textural details lost, but the shape and major features are preserved.

Fig. 2.15 shows the dynamic event of a hammer striking a toy apple. GAP can produce a noisy estimate of x in two iterations. As mentioned before, 50 iterations is sufficient to converge to an estimate of the underlying video sequence.

Note how the detail from rapidly moving parts of the scene seems lost on the low-framerate measurements but is recovered in the reconstructed images. For example, the rough texture of the orange and the shape of the apple are disambiguated during the inversion process. The hand is likewise clearer. The moving code has aliased this information as low-frequency, attenuated signal into the coded data y .

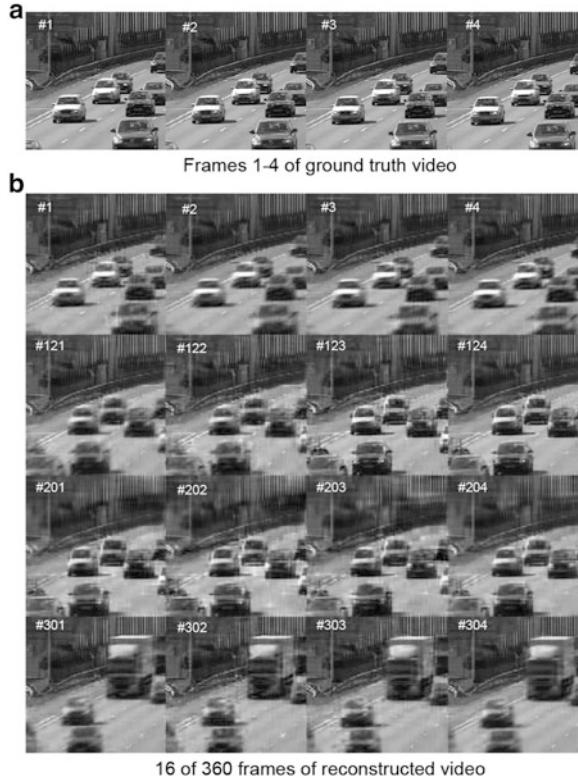


Fig. 2.18 Selected frames of a 360-frame simulated video reconstruction. (a) Four of 360 ground truth video frames. These frames are played back at various velocities prior to detection. (b) 16 of 360 estimated video frames reconstructed by GAP. The block-matching algorithm determined the ideal compression ratio given a fixed PSNR

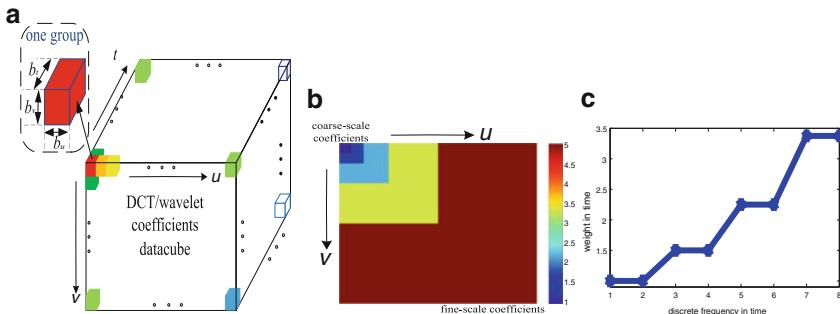


Fig. 2.19 (a) Decomposition of the video into wavelet and DCT coefficients. The large 3D cube represents the wavelet/DCT coefficients and the small 3D cubes in different color denote different groups, with the size $\{b_u, b_v, b_t\}$ shown by one example group on the top-left. (b) Wavelet-tree structure scale-weight in space. The groups in the same wavelet level (shown in the same color) share the same weight [52]. (c) DCT block structure time-weight. Each b_t (here $b_t = 2$) frames in time share the same weight

2.6 Conclusion and discussion

This book chapter highlights the important and growing topic of CS-video. The CACTI framework, which is of particular focus in this section, introduces a novel method to code the 2D spatial, temporal, and color dimensions using a simple 2D detector, piezoelectric actuator, and function generator. Importantly, a plethora of methods may be implemented to translate the code in a known manner (e.g., physical shaking or simple harmonic oscillation) to further reduce the system complexity.

Given the modest communications improvements over recent decades, the compressibility of the ambient world, and novel reconstruction algorithm advancements, it is plausible to see large-scale 4-dimensional compressive imaging systems establishing a foothold in the imaging world.

Despite the improvements made in CS in general, modern CS designs still sample the image uniformly in time. For CS-video, one may adapt the integration time, coding pattern, and reconstruction algorithm according to the temporal complexity of the scene as discussed in this chapter with a goal of maintaining a certain reconstruction fidelity.

In this chapter, we have shown how knowledge of a time-varying coding pattern can enhance the temporal resolution of an off-the-shelf camera. Spectral [21], spatial, and longitudinal CS all similarly employ joint hardware, coding, and inference designs to enhance the sampling dimensionality of cameras. Visualizing these recent advances into large-scale implementations [7] leads us to believe physical-layer compression can one day enable effective exapixel-scale capture of our dynamic world.

Acknowledgements The research, results, and theory presented here were supported by the Knowledge Enhanced Compressive Measurement Program at the Defense Advanced Research Projects Agency, grant N660011114002. Additional support from the ONR, NGA, ARO, and NSF is acknowledged.

References

1. Agrawal, A., Gupta, M., Veeraraghavan, A., Narasimhan S.G.: Optimal coded sampling for temporal super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 599–606, 2010
2. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.* **54**(11), 4311–4322 (2006)
3. Arguello, H., Rueda, H.F., Arce, G.R.: Spatial super-resolution in code aperture spectral imaging. In: SPIE Defense, Security, and Sensing. International Society for Optics and Photonics, pp. 83650A–83650A, 2012
4. Baraniuk, R.G.: Single-pixel imaging via compressive sampling. *IEEE Signal Process. Mag.* **25**(2), 83–91 (2008)
5. Bioucas-Dias, J.M., Figueiredo, M.A.T.: A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Trans. Image Process.* **16**(12), 2992–3004 (2007)

6. Brady D.J.: Optical Imaging and Spectroscopy. Wiley-Interscience, New York (2009)
7. Brady, D.J., Hagen, N.: Multiscale lens design. *Opt. Express* **17**(13), 10659–10674 (2009)
8. Brady, D.J., Feldman, M., Pitsianis, N., Guo, J.P., Portnoy, A., Fiddy, M.: Compressive optical montage photography. In: International Society for Optics and Photonics, *Optics & Photonics 2005*, pp. 590708–590708, 2005
9. Brady, D.J., Gehm, M.E., Stack, R.A., Marks, D.L., Kittle, D.S., Golish, D.R., Vera, E.M., Feller, D.: Multiscale gigapixel photography. *Nature* **486**(7403), 386–389 (2012)
10. Cai, J.-F., Osher, S., Shen Z.: Linearized bregman iterations for compressed sensing. *Math. Comput.* **78**(267), 1515–1536 (2009)
11. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20**(1), 33–61 (1998)
12. Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D., Carin, L.: Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: algorithm and performance bounds. *IEEE Trans. Signal Process.* **58**(12), 6140–6155 (2010)
13. Donoho, D.L.: Compressed sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006)
14. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.* **15**(12), 3736–3745 (2006)
15. Gehm, M.E., John, R., Brady, D.J., Willett, R.M., Schulz T.J., et al.: Single-shot compressive spectral imaging with a dual-disperser architecture. *Opt. Express* **15**(21), 14013–14027 (2007)
16. Goldstein, T., Xu, L., Kelly, K.F., Baraniuk, R.: The stone transform: Multi-resolution image enhancement and real-time compressive video. *arXiv preprint arXiv:1311.3405* (2013)
17. Hitomi, Y., Gu, J., Gupta, M., Mitsunaga, T., Nayar, S.K.: Video from a single coded exposure photograph using a learned over-complete dictionary. In: *IEEE International Conference on Computer Vision (ICCV)*, 2011 pp. 287–294, 2011
18. Holloway, J., Sankaranarayanan, A.C., Veeraraghavan, A., Tambe, S.: Flutter shutter video camera for compressive sensing of videos. In: *IEEE International Conference on Computational Photography (ICCP)*, pp. 1–9, 2012
19. Hunt, J., Driscoll, T., Mrozack, A., Lipworth, G., Reynolds, M., Brady, D., Smith, D.R.: Metamaterial apertures for computational imaging. *Science* **339**(6117), 310–313 (2013)
20. Ji, S., Xue, Y., Carin, L.: Bayesian compressive sensing. *IEEE Trans. Signal Process.* **56**(6), 2346–2356 (2008)
21. Kittle, D., Choi, K., Wagadarikar, A., Brady, D.J.: Multiframe image estimation for coded aperture snapshot spectral imagers. *Appl. Opt.* **49**(36), 6824–6833 (2010)
22. Kleinfelder, S., Lim, S.H., Liu, X., Gamal, A.E.: A 10,000 frames/s 0.18 μ m cmos digital pixel sensor with pixel-level memory. *ISSCC Dig. Tech. Papers* 88–89 (2001)
23. Liao, X., Li, H., Carin, L.: Generalized alternating projection for weighted- $\ell_{2,1}$ minimization with applications to model-based compressive sensing. *SIAM J. Imag. Sci.* **7**(2), 797–823 (2014)
24. Llull, P., Liao, X., Yuan, X., Yang, J., Kittle, D., Carin, L., Sapiro, G., Brady, D.J.: Coded aperture compressive temporal imaging. *Opt. Express* **21**(9), 10526–10545 (2013)
25. Llull, P., Yuan, X., Liao, X., Yang J., Carin, L., Sapiro, G., Brady, D.J.: Compressive extended depth of field using image space coding. In: *Computational Optical Sensing and Imaging*, pp. CM2D–3. Optical Society of America, 2014
26. MacCabe, K., Krishnamurthy, K., Chawla, A., Marks, D., Samei, E., Brady, D.J.: Pencil beam coded aperture x-ray scatter imaging. *Opt. Express* **20**(15), 16310–16320 (2012)
27. MacCabe, K.P., Holmgren, A.D., Tornai, M.P., Brady, D.J.: Snapshot 2d tomography via coded aperture x-ray scatter imaging. *Appl. Opt.* **52**(19), 4582–4589 (2013)
28. Mairal, J., Elad, M., Sapiro, G.: Sparse representation for color image restoration. *IEEE Trans. Image Process.* **17**, 53–69 (2008)
29. McCain, S.T., Gehm, M.E., Wang, Y., Pitsianis, N.P., Brady, D.J.: Coded aperture raman spectroscopy for quantitative measurements of ethanol in a tissue phantom. *Appl. Spectrosc.* **60**(6), 663–671 (2006)
30. Osher, S., Mao, Y., Dong, B., Yin, W.: Fast linearized bregman iteration for compressive sensing and sparse denoising. *arXiv preprint arXiv:1104.0262* (2011)

31. Raskar, R., Agrawal, A., Tumblin, J.: Coded exposure photography: motion deblurring using fluttered shutter. *ACM Trans. Graph.* **25**, 795–804 (2006)
32. Reddy, D., Veeraraghavan, A., Chellappa, R.: P2C2: Programmable pixel compressive camera for high speed imaging. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011 , pp. 329–336, 2011
33. Salehi, J.A.: Code division multiple-access techniques in optical fiber networks. i. fundamental principles. *IEEE Trans. Commun.* **37**(8), 824–833 (1989)
34. Sankaranarayanan, A.C., Studer, C., Baraniuk, R.G.: CS-MUVI: Video compressive sensing for spatial-multiplexing cameras. In: *IEEE International Conference on Computational Photography (ICCP)*, pp. 1–10, 2012
35. Shankar, M., Pitsianis, N.P., Brady, D.J.: Compressive video sensors using multichannel imagers. *Appl. Opt.* **49**(10), B9–B17 (2010)
36. Son, H., Marks, D.L., Tremblay, E.J., Ford, J., Hahn, J., Stack, R., Johnson, A., McLaughlin, P., Shaw, J., Kim, J.: A multiscale, wide field, gigapixel camera. In: *Computational Optical Sensing and Imaging*, p. 3. JTUE2, Optical Society of America, 2011
37. Teich, M.C., Saleh, B.: *Fundamentals of Photonics*, p. 3. Wiley Interscience, Canada (1991)
38. Tendero, Y., Rougé, B., Morel, J.-M.: A formalization of the flutter shutter. *J. Phys. Conf. Ser.* **386**(1), 012001 IOP Publishing (2012)
39. Tropp, J., Gilbert, A.C.: Signal recovery from partial information via orthogonal matching pursuit (2005)
40. Tsai, T.-H., Brady, D.J.: Coded aperture snapshot spectral polarization imaging. *Appl. Opt.* **52**(10), 2153–2161 (2013)
41. Tsaig, Y., Donoho, D.L.: Extensions of compressed sensing. *Signal Process.* **86**(3), 549–571 (2006)
42. Veeraraghavan, A., Reddy, D., Raskar, R.: Coded strobing photography: compressive sensing of high speed periodic videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(4), 671–686 (2011)
43. Wagadarikar, A., John, R., Willett, R., Brady, D.: Single disperser design for coded aperture snapshot spectral imaging. *Appl. Opt.* **47**(10), B44–B51 (2008)
44. Wagadarikar, A.A., Pitsianis, N.P., Sun, X., Brady, D.J.: Video rate spectral imaging using a coded aperture snapshot spectral imager. *Opt. Express* **17**(8), 6368–6388 (2009)
45. Watts, C.M., Shrekenhamer, D., Montoya, J., Lipworth, G., Hunt, J., Slesman, T., Krishna, S., Smith, D.R., Padilla, W.J.: Coded and compressive thz imaging with metamaterials. In: *International Society for Optics and Photonics SPIE OPTO*, pp. 89851N–89851N, 2014
46. Yang, J., Liao, X., Yuan, X., Llull, P., Brady, D.J., Sapiro, G., Carin, L.: Compressive Sensing by learning a gaussian mixture model from measurements. *IEEE Trans. Image Process* **24**(1), 106–119 (2015). doi: [10.1109/TIP.2014.2365720](https://doi.org/10.1109/TIP.2014.2365720)
47. Yang, J., Yuan, X., Liao, X., Llull, P., Brady, D.J., Sapiro, G., Carin, L.: Video compressive sensing using Gaussian mixture models. *IEEE Trans. Image Process* **23**(11), 4863–4878 (2014). doi: [10.1109/TIP.2014.2344294](https://doi.org/10.1109/TIP.2014.2344294)
48. Yang, J., Yuan, X., Liao, X., Llull, P., Sapiro, G., Brady, D.J., Carin, L.: Gaussian mixture model for video compressive sensing. In: *International Conference on Image Processing*, 2013
49. Yu, G., Sapiro, G.: Statistical compressed sensing of Gaussian mixture models. *IEEE Trans. Signal Process.* **59**(12), 5842–5858 (2011)
50. Yu, G., Sapiro, G., Mallat, S.: Solving inverse problems with piecewise linear estimators: from Gaussian mixture models to structured sparsity. *IEEE Trans. Image Process.* **21**(5), 2481–2499 (2012)
51. Yuan, X., Yang, J., Llull, P., Liao, X., Sapiro, G., Brady, D.J., Carin, L.: Adaptive temporal compressive sensing for video. In: *IEEE International Conference on Image Processing*, 2013
52. Yuan, X., Llull, P., Liao, X., Yang, J., Sapiro, G., Brady, D.J., Carin, L.: Low-cost compressive sensing for color video and depth. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014
53. Zhou, M., Chen, H., Paisley, J., Ren, L., Li, L., Xing, Z., Dunson, D., Sapiro, G., Carin, L.: Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images. *IEEE Trans. Image Process.* **21**(1), 130–144 (2012)

Chapter 3

Compressed Sensing, Sparse Inversion, and Model Mismatch

Ali Pezeshki, Yuejie Chi, Louis L. Scharf, and Edwin K.P. Chong

Abstract The advent of compressed sensing theory has revolutionized our view of imaging, as it demonstrates that subsampling has manageable consequences for image inversion, provided that the image is sparse in an a priori known dictionary. For imaging problems in spectrum analysis (estimating complex exponential modes), and passive and active radar/sonar (estimating Doppler and angle of arrival), this dictionary is usually taken to be a DFT basis (or frame) constructed for resolution of $2\pi/n$, with n a window length, array length, or pulse-to-pulse processing length. However, in reality no physical field is sparse in a DFT frame or in any a priori known frame. No matter how finely we grid the parameter space (e.g., frequency, delay, Doppler, and/or wavenumber) the sources may not lie in the center of the grid cells and consequently there is always mismatch between the assumed and the actual frames for sparsity. But what is the sensitivity of compressed sensing to mismatch between the physical model that generated the data and the

A. Pezeshki (✉) • E.K.P. Chong

Department of Electrical and Computer Engineering, and Department of Mathematics,
Colorado State University, Fort Collins, CO 80523, USA

e-mail: ali.pezeshki@colostate.edu; edwin.chong@colostate.edu

Y. Chi

Department of Electrical and Computer Engineering, and Department of Biomedical Informatics,
The Ohio State University, Columbus, OH 43210, USA

e-mail: chi.97@osu.edu

L.L. Scharf

Department of Mathematics, Colorado State University, Fort Collins, CO 80523, USA
e-mail: louis.scharf@colostate.edu

The authors were supported in part by the NSF by Grants CCF-1018472, CCF-1017431, CCF-0916314, CCF-0915299, and CCF-1422658.

©2011 IEEE. Reprinted, with permission, from Y. Chi, L. L. Scharf, A. Pezeshki, A. R. Calderbank, “Sensitivity to basis mismatch in compressed sensing,” *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2182–2195, May 2011.

©2011 IEEE. Reprinted, with permission, from L. L. Scharf, E. K. P. Chong, A. Pezeshki, and J. R. Luo, “Sensitivity considerations in compressed sensing,” *Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, 6–9 Nov. 2011, pp. 744–748.

mathematical model that is assumed in the sparse inversion algorithm? In this chapter, we study this question. The focus is on the canonical problem of DFT inversion for modal analysis.

3.1 Introduction

In a great number of fields of engineering and applied science the problem confronting the designer is to *invert an image*, acquired from a sensor suite, for the underlying field that produced the image. And typically the desired resolution for the underlying field exceeds the temporal or spatial resolution of the image itself. Certainly this describes the problem of identifying field elements from electromagnetic and acoustic images, multipath components in wireless communication, sources of information in signal intelligence, cyber attackers in data networks, and radiating sources in radar and sonar. Here we give *image* its most general meaning to encompass a time series, a space series, a space-time series, a 2-D image, and so on. Similarly, we give *field* its most general meaning to encompass complex-exponential modes, radiating modes, packetized voice or data, coded modulations, multipath components, and the like.

Broadly speaking there are two classical principles for inverting the kinds of images that are measured in optics, electromagnetics, and acoustics. The first principle is one of matched filtering, wherein a sequence of rank-one subspaces, or one-dimensional test images, is matched to the measured image by filtering or correlating or phasing. The sequence of test images is generated by scanning a prototype image (e.g., a waveform or a steering vector) through frequency, wavenumber, doppler, and/or delay. In time series analysis, this amounts to classical spectrum analysis to identify the frequency modes, and the corresponding mode amplitudes, of the signal (see, e.g., [23, 28]). In phased-array processing, it amounts to spectrum analysis in frequency and wavenumber to identify the frequency-wavenumber coordinates of sources impinging on a sensor array (see, e.g., [36] and [20]). In Space-Time Adaptive Processing (STAP) radar and sonar, it amounts to spectrum analysis in delay, frequency, and wavenumber to reconstruct the radar/sonar field (see, e.g., [37] and [21]). This matched filtering principle actually extends to subspace matching for those cases in which the model for the image is comprised of several dominant modes [28, 30]. When there is a known or estimated noise covariance, then this matched filter concept may be further extended to a whitened matched filter or to a Minimum Variance UnBiased (MVUB) filter (see, e.g., [28]). The MVUB estimator has the interpretation of a generalized sidelobe canceler that scans two subspaces, one matched to the signal subspace as in matched filtering, and the other to its orthogonal complement [27].

The second principle is one of parameter estimation in a separable nonlinear model, wherein a sparse *modal* representation for the field is posited and estimates of linear parameters (complex amplitudes of modes) and nonlinear mode parameters (frequency, wavenumber, delay, and/or doppler) are extracted, usually based on

maximum likelihood, or some variation on linear prediction (see, [28, 35], and [34]). There is a comprehensive literature in electrical engineering, mathematics, physics, and chemistry on parameter estimation and performance bounding in such models (see, e.g., [14, 22, 28, 32, 35], and [34]). One important limitation of the classical principles is that any subsampling of the measured image has consequences for resolution (or bias) and for variability (or variance).

Compressed sensing stands in contrast to these principles. It says that complex baseband data may be compressed before processing, when it is known *apriori* that the field to be imaged is sparse in a *known* basis or dictionary. For imaging problems in electromagnetics, acoustics, nuclear medicine, and active/passive sensing, this dictionary is usually taken to be a DFT basis or frame that is constructed for resolution of $2\pi/n$, with n a window length, array length, or pulse-to-pulse processing length. A great number of articles (see, e.g., [1, 6, 15, 17], and [13]) have considered the use of compressed sensing theory for active/passive imaging, when the sources of interest are taken to be on a regular grid in delay, Doppler, and wavenumber, and point to the potential of this theory as a new high resolution imaging principle. But no matter how large n is, the actual field will not place its sources on the grid points $\{2\pi\ell/n\}_{\ell=0}^{n-1}$ of the image model. This means the image is not actually sparse in the DFT frame, as any mode of the field that does not align with the DFT frame will produce spectral leakage into all DFT modes through the Dirichlet kernel, which decays only as $1/f$ in frequency f .

This observation raises the following questions. What is the sensitivity of compressed sensing for image inversion to *mismatch* between the assumed frame for sparsity and the actual basis or frame in which the image is sparse? How does the performance of compressed sensing in the presence of model mismatch compare with the performance of classical matched filtering for imaging in frequency, wavenumber, delay, and/or doppler, or with that of linear prediction or other approximations to maximum likelihood? These are the general questions that we discuss in this chapter. Our discussion follows our results and analysis in [11] and [29], where we have studied these questions in great detail. The reader is referred to [16, 18], and [12] for other studies on the treatment of model mismatch on compressed sensing.

The chapter is organized as follows. In Section 3.2, we discuss the main cause of model mismatch in compressed sensing, by contrasting overdetermined separable nonlinear models with underdetermined sparse linear models as their surrogates. In Section 3.3, we discuss the particularly problematic nature of model mismatch in compressed sensing for DFT inversion. In Section 3.4, we present numerical examples that show the degradation of compressed sensing inversions in the presence of DFT grid mismatch, even when the DFT grid is made very fine. In Section 3.5, we discuss general analytical results that show how guarantee bounds for sparse inversion degenerate in presence of basis mismatch. Finally, in Section 3.6, we briefly review a few promising new results that may provide an alternative to gridding.

3.2 Model Mismatch in Compressed Sensing

When viewed from the point of view of separable nonlinear models (see, e.g., [14] and [28]), most inverse imaging problems are decidedly *not* under-determined. However, when the problem is approximated with a high resolution linear model, where the unknown modal basis is replaced with a highly resolved apriori dictionary, then this approximating linear model *is* underdetermined. But this model is always mismatched to the actual model determined by the physics of the problem.

More specifically, in the separable nonlinear model, parameters such as complex scattering coefficients compose a small number of modes that are nonlinearly modulated by mode parameters such as frequency, wavenumber, and delay, and the problem is to estimate both the mode parameters and the scattering coefficients. In the approximating linear model, scattering coefficients compose a large number of modes whose frequencies, wavenumbers, and delays are assumed to be on a prespecified grid, and the problem is to determine which modes on this prespecified grid are active and to estimate the complex scattering coefficients of those modes. So the question of model mismatch in compressed sensing is really a question of quantifying the consequences of replacing an overdetermined separable nonlinear model, in which the basis or frame for sparsity is to be determined, with an underdetermined linear model in which a basis or frame is assumed to render a sparse model.

In order to frame our question more precisely, let us begin with two models for a measured image $x \in \mathbb{C}^n$. The first is the mathematical model that is *assumed* in the compressed sensing procedure, and the other is the physical model that has actually produced the image x from the field.

In the mathematical model, the image is composed as

$$x = \Psi c, \quad (3.1)$$

where the basis $\Psi \in \mathbb{C}^{n \times n}$ is known (apriori selected), and is typically a gridded imaging matrix (e.g., an n -point DFT matrix or an n -point discretized ambiguity matrix) and $c \in \mathbb{C}^n$ is assumed to be k -sparse.

But, as a matter of fact, the image x is composed by the physics as

$$x = \tilde{\Psi} \tilde{c}, \quad (3.2)$$

where the basis $\tilde{\Psi} \in \mathbb{C}^{n \times k}$ is determined by a point spread function, a Green's function, or an impulse response, and \tilde{c} is a k -dimensional complex vector. Typically $\tilde{\Psi}$ is determined by frequency, wavenumber, delay, and/or doppler parameters that are *unknown* apriori. These are the true but unknown modal parameters that nonlinearly parameterize the image x . More importantly, these parameters do *not* lie exactly on the gridding points corresponding to the columns of Ψ (e.g., DFT vectors). In other words, the columns of $\tilde{\Psi}$ almost never coincide with a subset of columns of Ψ . We call this *basis mismatch*, and note that it is present in all imaging problems, no matter how large n is, or equivalently no matter how fine-grained the gridding procedure is.

If m (with $k \ll m \ll n$) linear measurements $y = Ax$ are taken with a compressive measurement matrix $A \in \mathbb{C}^{m \times n}$, then we have two different inversion problems depending on which model we consider for x . With the physical model, we have an overdetermined separable nonlinear problem,

$$y = A\tilde{\Psi}\tilde{c}, \quad (3.3)$$

where we wish to invert the m -dimensional measurement vector y for the $2k < m$ unknown parameters in $\tilde{\Psi}$ and \tilde{c} . We refer to this case as a separable nonlinear case, where k linear parameters determine the elements of \tilde{c} and k nonlinear parameters determine the columns of $\tilde{\Psi}$.

With the mathematical model, we have an underdetermined linear problem,

$$y = A\Psi c, \quad (3.4)$$

which we typically regularize by assuming sparsity for c . In this latter case, the problem of estimating the nonlinear parameters in $\tilde{\Psi}$ is assumed away by replacing the unknown physical basis $\tilde{\Psi}$ by the preselected mathematical basis Ψ . The goal is then to extract the linear parameters c that are active (nonzero) in the assumed mathematical model from the relatively small $m < n$ number of linear measurements. This replacement of an overdetermined separable nonlinear model with an underdetermined linear one would be plausible if the parameter vector c would in fact be k -sparse and the k modes of the physical model are well-approximated by the k modes of the mathematical basis Ψ .

Typically, the sparse recovery of c from y is solved as a basis pursuit (BP) problem (see Chapter 1, Section 1.3.1):

$$\hat{c} = \arg \min_z \|z\|_1 \quad s.t. \quad y = \Phi z \quad (3.5)$$

where $\Phi = A\Psi$. If Φ satisfies the so-called *Restricted Isometry Property* (RIP) with restricted-isometry constant (RIC) $\delta_{2k}(\Phi) < \sqrt{2} - 1$, then the estimate \hat{c} for the presumably sparse c is exact (see Chapter 1, Theorem 1.6, and also [5] and [3]).

With noisy measurements, typically a basis pursuit denoising (BPDN) [3, 4] is considered:

$$\hat{c} = \arg \min_z \|z\|_1 \quad s.t. \quad \|y - \Phi z\|_2 \leq \varepsilon, \quad (3.6)$$

where ε upper-bounds the 2-norm of the noise. If Φ satisfies the RIP, then the 2-norm of the error in estimating c will be bounded above by a term that scales linearly with ε . The reader is referred to [3, 4] for details. Many other approaches such as orthogonal matching pursuit (OMP) (see Chapter 1, Section 1.4.4.2, and [33]), ROMP [26] and CoSaMP [24] for finding a sparse solution to (3.4) also exist.

But is it plausible to assume that c in (3.1) (and subsequently (3.4)) is k -sparse? This is the crux of the issue when comparing compressed sensing to the classical estimation methods such as linear prediction, exact and approximate maximum likelihood, subspace decompositions, and so on.

To answer this question, let us consider the coordinate transformation between the physical representation vector \tilde{c} and the mathematical representation vector c :

$$c = \Psi^{-1} \tilde{\Psi} \tilde{c}. \quad (3.7)$$

We see that unless $\Psi^{-1} \tilde{\Psi}$ is an $n \times k$ slice of the $n \times n$ identity matrix the parameter vector \tilde{c} is not sparse in the standard basis. But as we discussed, the columns of $\tilde{\Psi}$ almost never coincide with the columns of Ψ , which are posited on a preselected grid (e.g., an n -point DFT at resolution $2\pi/n$). As a result c is almost never sparse, and in an array of problems including line spectrum estimation, Doppler estimation, and direction of arrival estimation, where Ψ is an n -point DFT matrix, c is not even compressible.

So the question is, “what is the consequence of assuming that c is sparse in the identity basis, when in fact it is only sparse in an *unknown* basis $\Psi^{-1} \tilde{\Psi}$ that is determined by the mismatch between Ψ and $\tilde{\Psi}$?” In the rest of this chapter, we study this question in detail. Our study follows that of [11], where this question was originally posed and answered. Our emphasis is on the canonical problem of sparse DFT inversion for modal analysis.

3.3 A Canonical Problem

A mismatch case of particular interest arises in Fourier imaging when a sparse signal with arbitrary frequency components is taken to be sparse in a DFT basis. Our objective in this section is to highlight the particularly problematic nature of basis mismatch in this application.

Suppose the sparsity basis Ψ in the mathematical model (3.1), assumed by the compressed sensing procedure, is the unitary n -point DFT basis. Then, the ℓ th column of Ψ is a Vandermonde vector of the form

$$\psi_\ell = \begin{pmatrix} 1 \\ e^{j \frac{2\pi\ell}{n}} \\ \vdots \\ e^{j \frac{2\pi\ell(n-1)}{n}} \end{pmatrix} \quad (3.8)$$

and the basis Ψ is

$$\Psi = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{j\frac{2\pi}{n}} & \cdots & e^{j\frac{2\pi(n-1)}{n}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{j\frac{2\pi(n-1)}{n}} & \cdots & e^{j\frac{2\pi(n-1)^2}{n}} \end{bmatrix}. \quad (3.9)$$

Without loss of generality, suppose that the $\tilde{\ell}$ th column $\tilde{\psi}_{\tilde{\ell}}$ of the physical basis $\tilde{\Psi}$ is closest, in its (normalized) frequency, to ψ_{ℓ} of Ψ , but it is mismatched to ψ_{ℓ} by $\Delta\theta_{\ell,\tilde{\ell}}$ in (normalized) frequency, where $0 \leq \Delta\theta_{\ell,\tilde{\ell}} < \frac{2\pi}{n}$. In other words,

$$\tilde{\psi}_{\tilde{\ell}} = \begin{pmatrix} e^{j(0+\Delta\theta_{\ell,\tilde{\ell}})} \\ e^{j(\frac{2\pi\ell}{n}+\Delta\theta_{\ell,\tilde{\ell}})} \\ \vdots \\ e^{j(\frac{2\pi\ell}{n}+\Delta\theta_{\ell,\tilde{\ell}})(n-1)} \end{pmatrix}. \quad (3.10)$$

Then, it is easy to see that the (r,ℓ) th element of $\Psi^{-1}\tilde{\Psi} \in \mathbb{C}^{n \times k}$, in the coordinate transformation (3.7), is obtained by sampling the Dirichlet kernel

$$D_n(\theta) = \frac{1}{n} \sum_{n'=0}^{n-1} e^{jn'\theta} = \frac{1}{n} e^{j\frac{\theta(n-1)}{2}} \frac{\sin(\theta n/2)}{\sin(\theta/2)} \quad (3.11)$$

at $\theta = \Delta\theta_{\ell,\tilde{\ell}} - \frac{2\pi}{n}(r-\ell)$.

The Dirichlet kernel $D_n(\theta)$, shown in Fig. 3.1 (ignoring the unimodular phasing term) for $n = 64$, decays slowly as $|D_n(\theta)| \leq (n\theta/2\pi)^{-1}$ for $|\theta| \leq \pi$, with $D(0) = 1$. This decay behavior follows from the fact that $|\sin(\theta/2)| \geq 2|\theta/2\pi|$ for $|\theta| \leq \pi$, where the equality holds when $|\theta| = \pi$. This means that $(n\theta/2\pi)^{-1}$ is in fact the envelope of $|D_n(\theta)|$. Therefore, every mismatch between a physical frequency and the corresponding (closest in normalized frequency) DFT frequency produces a column in $\Psi^{-1}\tilde{\Psi}$ for which the entries vanish slowly as each column is traversed. The consequence of this is that the parameter vector c in the mathematical model (3.1), for which the compressed sensing procedure is seeking a sparse solution, is in fact not sparse, because the entries of \tilde{c} leak into all entries of $c = \Psi^{-1}\tilde{\Psi}\tilde{c}$.

In addition, to frequency mismatch, $\tilde{\psi}_{\tilde{\ell}}$ may also be mismatched to ψ_{ℓ} by damping factors $\lambda_{\ell,\tilde{\ell}} \geq 0$. In such a case, the (r,ℓ) th element of $\Psi^{-1}\tilde{\Psi}$ is

$$(\Psi^{-1}\tilde{\Psi})_{r,\ell} = \frac{1}{n} \sum_{n'=0}^{n-1} e^{n'[-\lambda_{\ell,\tilde{\ell}} + j(\Delta\theta_{\ell,\tilde{\ell}} - \frac{2\pi(r-\ell)}{n})]}. \quad (3.12)$$

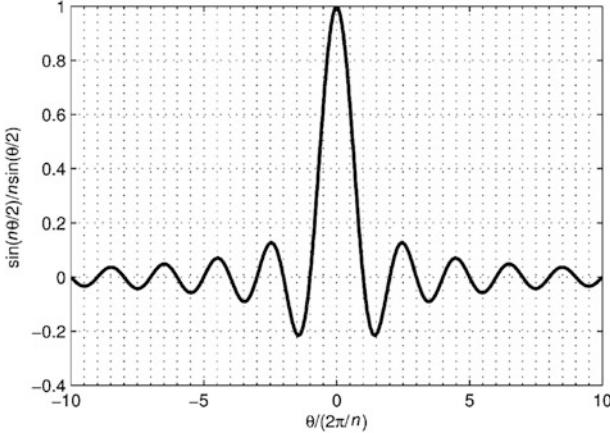


Fig. 3.1 The Dirichlet kernel $\frac{1}{n} \frac{\sin(n\theta/2)}{\sin(\theta/2)}$ vs. $\theta/(2\pi/n)$.

In general, the basis mismatch problem exists in almost all applications and is not limited to Fourier imaging. However, we have emphasized Fourier imaging in this section as a canonical problem, where basis mismatch seems to have a particularly destructive effect. Other problems such as multipath resolution, where discretely delayed Nyquist pulses are correlated with Nyquist pulses whose delays do not lie on these discrete delays, appear not as troublesome as the case of Fourier imaging.

3.4 Effect of Mismatch on Modal Analysis

We now present several numerical examples to study the effect of DFT grid mismatch on modal analysis based on compressed sensing measurements and compare these results with those obtained using classical image inversion principles, namely standard DFT imaging (matched filtering) and linear prediction (LP). The reader is referred to [28, 35], and [34] for a description of linear prediction.

3.4.1 Inversion in a DFT Basis

In all the numerical examples in this section, the dimension of the image x is $n = 64$. The number m of measurements y used for inversion is the same for all methods and we report results for $m = n/4 = 16$ to $n = n/2 = 32$ to $m = n/1 = 64$. For each inversion method (DFT, LP, and CS), we choose a compression matrix A that is typically used in that type of inversion. For DFT and LP inversions, we choose A as

the first m rows of the $n \times n$ identity matrix. For CS inversion (in a DFT basis), we choose A by drawing m rows from the $n \times n$ identity matrix uniformly at random.

Example 1 (no mismatch and noise free). We first consider the case where the field we wish to invert for contains only modes that are aligned with the DFT frequencies. This is to demonstrate that compressed sensing (from here on CS) and LP both provide perfect field recovery when there is no mismatch. BP is used as the inversion algorithm for CS. No noise is considered for now. The inversion results are shown in Fig. 3.2(a)–(c). In each subfigure (a) through (c) there are four panels. In the top-left panel the true underlying modes are illustrated with stems whose locations on the unit disc indicate the frequencies of the modes, and whose heights illustrate the mode amplitudes. The phases of the modes are randomly chosen, and not indicated on the figures. The frequencies at which modes are placed, and their amplitudes are $(9\frac{2\pi}{n}, 1)$, $(10\frac{2\pi}{n}, 1)$, $(20\frac{2\pi}{n}, .5)$, and $(45\frac{2\pi}{n}, .2)$. These frequencies are perfectly aligned with the DFT frequencies.

But what is the connection between the circular plots in Fig. 3.2 and the models Ψc and $\tilde{\Psi} \tilde{c}$? The top-left panel (actual modes) in each subplot is an “illustration” of $(\tilde{\Psi}, \tilde{c})$, with the locations of the bars on the unit disc corresponding to modes in $\tilde{\Psi}$ and the heights of the bars corresponding to the values of entries in \tilde{c} . The top-right panel (conventional DFT) illustrates (Ψ, \hat{c}) , where \hat{c} is the estimate of c obtained by DFT processing the measurement vector y . The bottom-left (CS) illustrates (Ψ, \hat{c}) , where \hat{c} is the solution of the basis pursuit (BP) problem (3.5). The bottom-right panel (LP) illustrates $(\hat{\Psi}, \hat{c})$, where $\hat{\Psi}$ and \hat{c} are, respectively, estimates of Ψ and c obtained by LP (order 8). We observe that both CS and LP provide perfect recovery, when there is no mismatch and noise. The DFT processing however has leakage according to the Dirichlet kernel unless the measurement dimension is increased to the full dimension $n = 64$. This was of course expected.

Example 2 (mismatched but noise free). We now introduce basis mismatch either by moving some of the modes off the DFT grid or by damping them. For frequency mismatch, the first two modes are moved to $(9.25\frac{2\pi}{n}, 1)$ and $(9.75\frac{2\pi}{n}, 1)$. For damping mismatch the mode at $(9\frac{2\pi}{n}, 1)$ is drawn off the unit circle to radius 0.95, so that the mode is damped as $(0.95)^{n'}$ at the n' sampling instance. The rest of the modes are the same as in the mismatch free case. Figs. 3.3 and 3.4 show the inversion results for DFT, CS, and LP (order 8) for $m = n/4 = 16$, $m = n/2 = 32$, and $m = n/1 = 64$, in the presence of frequency mismatch and damping mismatch. In all cases, DFT and CS result in erroneous inversions. The inaccuracy in inversion persists even when the number of measurements is increased to the full dimension. However, we observe that LP is always exact. These are all noise free cases.

Example 3 (noisy with and without mismatch). We now consider noisy observations for both mismatched and mismatch-free cases. In the mismatch-free case, the frequencies at which the modes are placed, and their amplitudes are $(9\frac{2\pi}{n}, 1)$, $(11\frac{2\pi}{n}, 1)$, $(20\frac{2\pi}{n}, .5)$, and $(45\frac{2\pi}{n}, .2)$. For frequency mismatch, the first two modes

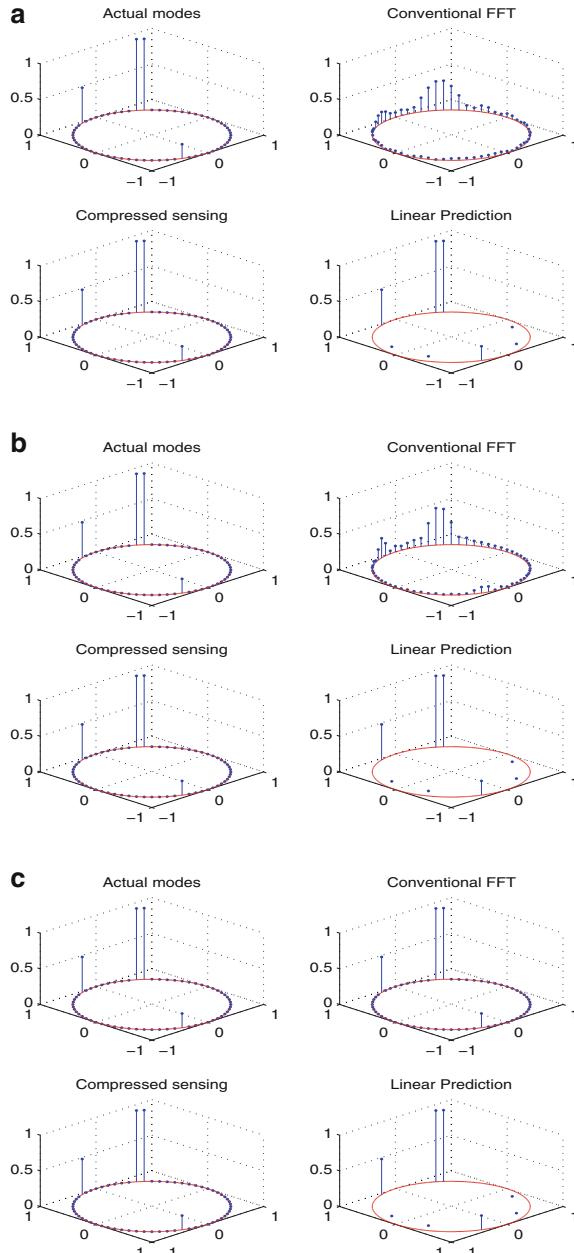


Fig. 3.2 Comparison of DFT, CS, and LP inversions in the absence of basis mismatch (a) $m = n/4 = 16$, (b) $m = n/2 = 32$, and (c) $m = n/1 = 64$.

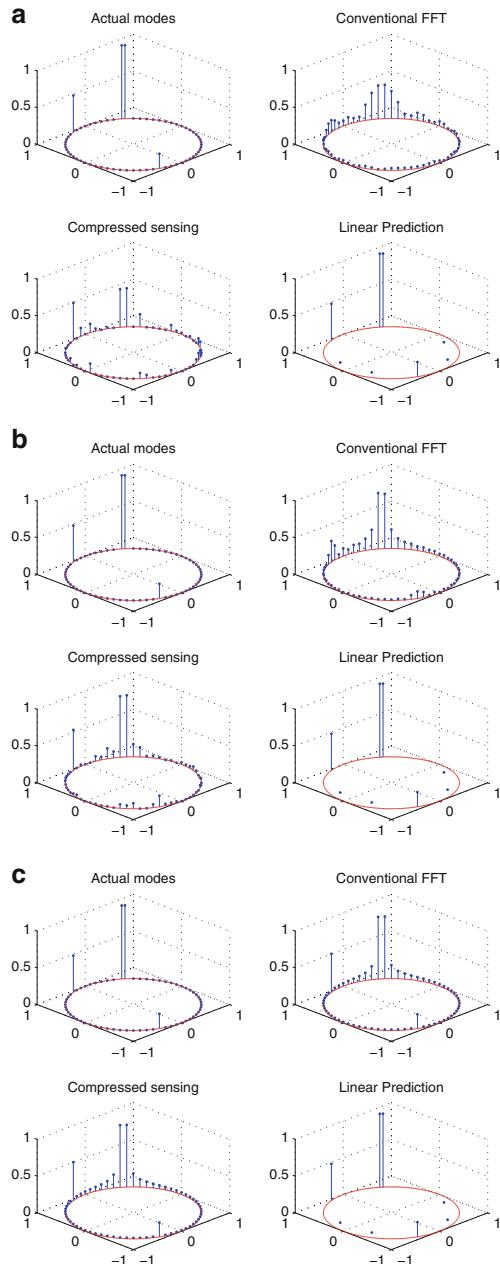


Fig. 3.3 Comparison of DFT, CS, and LP inversions in the presence of basis mismatch (frequency mismatch), for (a) $m = n/4 = 16$, (b) $m = n/2 = 32$, and (c) $m = n/1 = 64$.

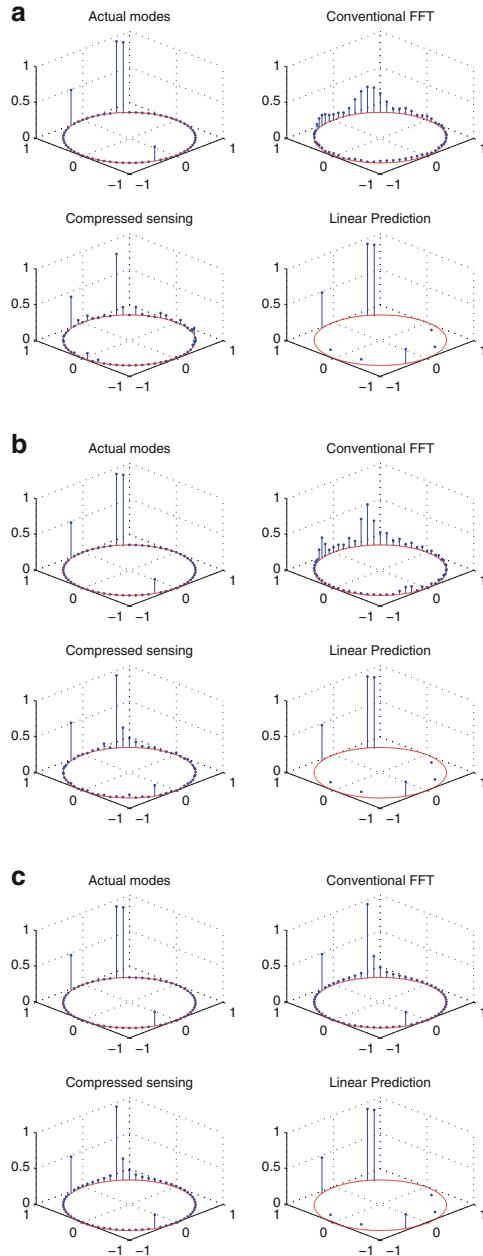


Fig. 3.4 Comparison of DFT, CS, and LP inversions in the presence of basis mismatch (damping mismatch), for (a) $m = n/4 = 16$, (b) $m = n/2 = 32$, and (c) $m = n/1 = 64$.

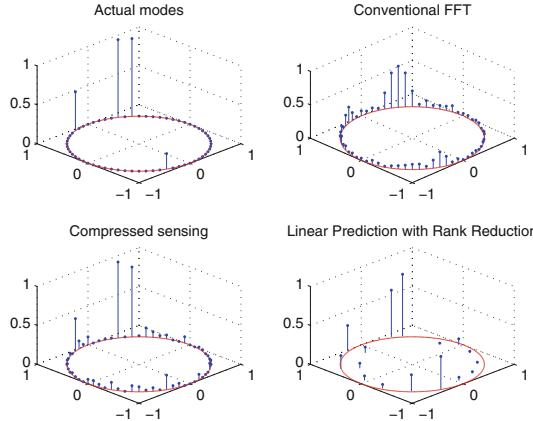


Fig. 3.5 Comparison of DFT, CS, and LP inversions with $m = n/2 = 32$ noisy observations but no mismatch.

are moved to $(9.25\frac{2\pi}{n}, 1)$ and $(10.75\frac{2\pi}{n}, 1)$. For damping mismatch the mode at $(9\frac{2\pi}{n}, 1)$ is drawn off the unit circle to radius 0.95. The number of measurements is $m = n/2 = 32$. The noise is additive and complex proper Gaussian with variance σ^2 . The signal-to-noise-ratio (SNR) is $10\log_{10}(1/\sigma^2) = 7\text{dB}$, relative to the unit amplitude modes. The LP order is changed to 16, but rank reduction (see [35] and [34]) is applied to reduce the order back to 8 as is typical in noisy cases. The inversion results are shown in Figs. 3.5 and 3.6 for the mismatch-free and mismatched cases, respectively. BPDN is used as the inversion method for CS. In the mismatch-free case in Fig. 3.5, CS provides relatively accurate estimates of the modes. However, in the mismatched case in Figs. 3.6(a),(b), CS breaks down and LP provides more reasonable estimates of the modes.

3.4.2 Inversion in an Overresolved DFT Dictionary

A natural question to ask is “can model mismatch sensitivities of CS be mitigated by overresolving the mathematical model (relative to Rayleigh limit) to ensure that mathematical modes are close to physical modes?” One might expect that over-resolution of a mathematical basis would produce a performance that only bottoms out at the quantization variance of the presumed frame for sparsity. But in fact, aggressive over resolution in a frame can actually produce worse performance at high SNR than a frame with lower resolution as we now show through an example. Our study in this section follows [29], where a detailed study of effects of model mismatch in an overresolved DFT dictionary has been presented.

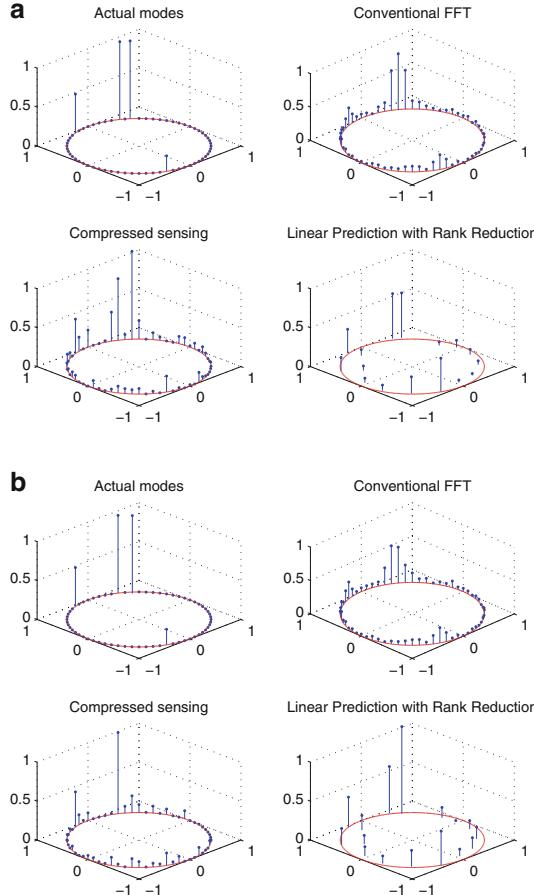


Fig. 3.6 Comparison of DFT, CS, and LP inversions with $m = n/2 = 32$ noisy observations: (b) frequency mismatch and (b) damping mismatch.

For our example, we consider the problem of estimating the frequency $f_1 = 0.5$ Hz of a unit-amplitude complex exponential from $m = 25$ samples, taken in presence of a unit-amplitude interfering complex exponential at frequency $f_2 = 0.52$ Hz and complex proper Gaussian noise of variance σ^2 . The 0.2 Hz separation between the two tones is a separation of half the Rayleigh limit of $1/25$ Hz. Both complex exponentials have zero phase. We consider both BPDN and OMP (see Chapter 1, Section 1.4.4.2) for estimating f_1 through sparse inversion of the measurements in an over-resolved DFT dictionary.

The measurement model to be inverted is $y = A\Psi c$, where here A is the $m \times m$ identity matrix and Ψ is the $m \times mL$ DFT frame at resolution $2\pi/mL$:

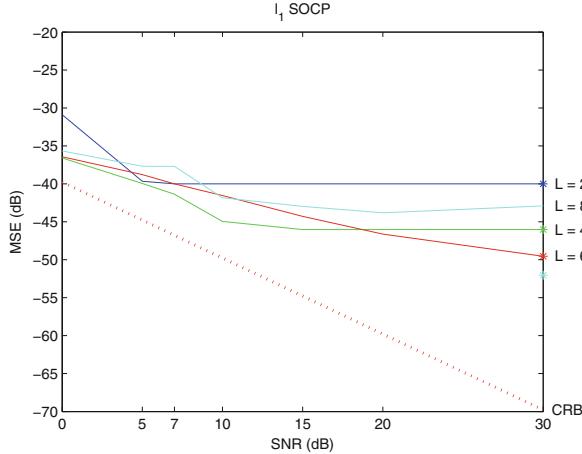


Fig. 3.7 Experimental mean-squared errors vs. SNR for estimating f_1 using BPDN inversion in a DFT frame with resolution $2\pi/mL$ for $m = 25$ and different values of L . The asterisks indicate the width of the half-cell ($1/2mL$ Hz) for the various L values ($L = 2, 4, 6, 8$).

$$\Psi = \frac{1}{\sqrt{m}} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{j\frac{2\pi}{mL}} & \cdots & e^{j\frac{2\pi(m-1)}{mL}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{j\frac{2\pi(m-1)}{mL}} & \cdots & e^{j\frac{2\pi(m-1)(mL-1)}{mL}} \end{bmatrix}. \quad (3.13)$$

The over-resolution factor L is used to grid the frequency space L times finer than the Rayleigh limit $2\pi/m$. The number of modes in the frame Ψ is $n = mL$, corresponding to complex exponentials spaced at $1/mL$ Hz. This may be viewed as the resolution of the DTFT frame, and BPDN and OMP can only return frequency estimates at multiples of $1/mL$.

Figs. 3.7 and 3.8 show experimental mean-squared errors (MSEs) in dB for estimating f_1 using BPDN and OMP, respectively, for different values of L . The values of L are selected such that the two frequencies f_1 and f_2 always lie half way between two grid points of the DFT frame. The half-cell widths $1/2mL$ Hz (half of distance between grid points) are shown in the plots with asterisks on the right side of each plot. These asterisks show the amounts of bias-squared ($1/2mL$)² in estimating f_1 , for various values of L , if the sparse inversion algorithms (BPDN or OMP) happen to choose the closest point on the DFT grid to f_1 . The linear dotted line is the Cramer-Rao Bound (see, e.g., [28]). The SNR in dB is defined to be $10\log_{10}(1/\sigma^2)$.

Fig. 3.7 demonstrates that at $L = 2$ the BPDN inversions are noise-defeated below 5 dB and resolution-limited above 5 dB, meaning they are limited by the resolution of the frame. That is, below 5 dB, mean-squared error is bias-squared plus variance, while above 5 dB, mean-squared error is bias-squared due to the resolution of the

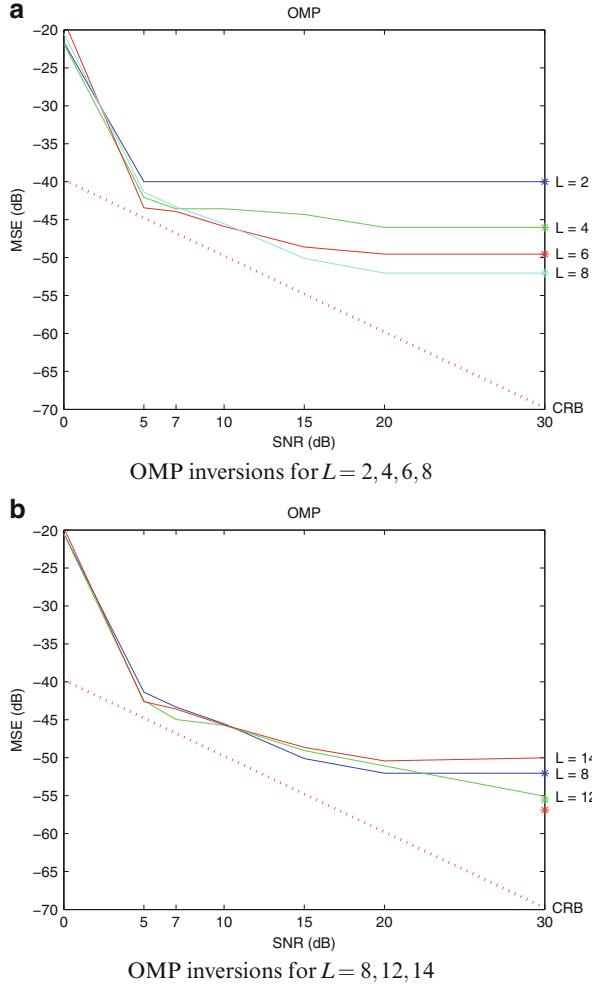


Fig. 3.8 Experimental mean-squared errors vs. SNR for estimating f_1 using OMP inversion in a DFT frame with resolution $2\pi/mL$ for $m = 25$ and different values of L . The asterisks indicate the width of the half-cell ($1/2mL$ Hz) for the various L values. (a) $L = 2, 4, 6, 8$ and (b) $L = 8, 12, 14$.

frame. At $L = 4$, the inversions are noise-defeated below 5 dB, noise-limited from 5 to 10 dB, and resolution-limited above 10 dB. As the expansion factor is increased, corresponding to finer- and finer-grained resolution in frequency, the frame loses its incoherence, meaning the dimension of the null space increases so much that there are many sparse inversions that meet a fitting constraint. As a consequence, we see that at $L = 6$ and $L = 8$ the mean-squared errors never reach their resolution limits, as variance overtakes bias-squared to produce mean-squared errors that are larger than

those of inversions that used smaller expansion factors. This suggests that there is a clear limit to how much bias-squared can be reduced with frame expansion, before variance overtakes bias-squared to produce degraded BPDN inversions.

Figs. 3.8(a) and (b) make these points for OMP inversions. The OMP inversions extend the threshold behavior of the inversions, they track the CRB more closely in the noise-limited region, and they reach their resolution limit for larger values of L before reaching their null-space limit at high SNRs. For example, as $L = 8$ the null-space limit has not yet been reached at SNR = 30 dB, whereas for $L = 14$, the null-space limit is reached before the resolution limit can be reached.

We note that the results in Figs. 3.7 and 3.8 are actually too optimistic, as the two mode amplitudes are equal. For a weak mode in the presence of a strong interfering mode, the results are much worse. Also, in all of these experiments, the fitting error is matched to the noise variance. With mismatch between fitting error and noise variance the results are more pessimistic. Nonetheless, they show that the consequence of over-resolution is that performance follows the Cramer-Rao bound more closely at low SNR, but at high SNR it departs more dramatically from the Cramer-Rao bound. This matches intuition that has been gained from more conventional spectrum analysis where there is a qualitatively similar trade-off between bias and variance. That is, bias may be reduced with frame expansion (over-resolution), but there is a penalty to be paid in variance.

Figs. 3.9(a)–(d) are scatter plots of BPDN and OMP inversions at 7 dB SNR, for $L = 2, 5, 9$. Fig. 3.9(a) scatters estimator errors for (f_1, f_2) for ℓ_1 inversions and Fig. 3.9(b) is the same data, plotted as normalized errors in estimating $((f_1 + f_2), (f_1 - f_2))$, the sum and difference frequencies. At $L = 2$ mean-squared error is essentially bias-squared, whereas for $L = 9$ it is essentially variance. The normalized errors in Fig. 3.9(b) demonstrate that the average frequency of the two tones is easy to estimate and the difference frequency is hard to estimate. (The vertical scale is nearly 10 times the horizontal scale.) This scatter plot demonstrates that BPDN inversions favor large negative differences over large positive differences, suggesting that the algorithm more accurately estimates the mode at frequency f_1 than it estimates the mode at frequency f_2 . The geometry of the scatter plots indicates that a concentration ellipse drawn from the inverse of the Fisher information matrix would be a poor descriptor of errors.

Figs. 3.9(c) and (d) make these points for OMP inversions as Figs. 3.9(a) and (b) made for BPDN inversions. However now the preference for large negative errors in estimating the difference frequency disappears. This suggests that the first dominant mode that is removed is equally likely to be near one or the other of the two modes in the data. The correlation between sum and difference errors reflects the fact that a large error in extracting the first mode will produce a large error in extracting the second. For OMP, it is possible that a concentration ellipse will accurately trap the solutions.

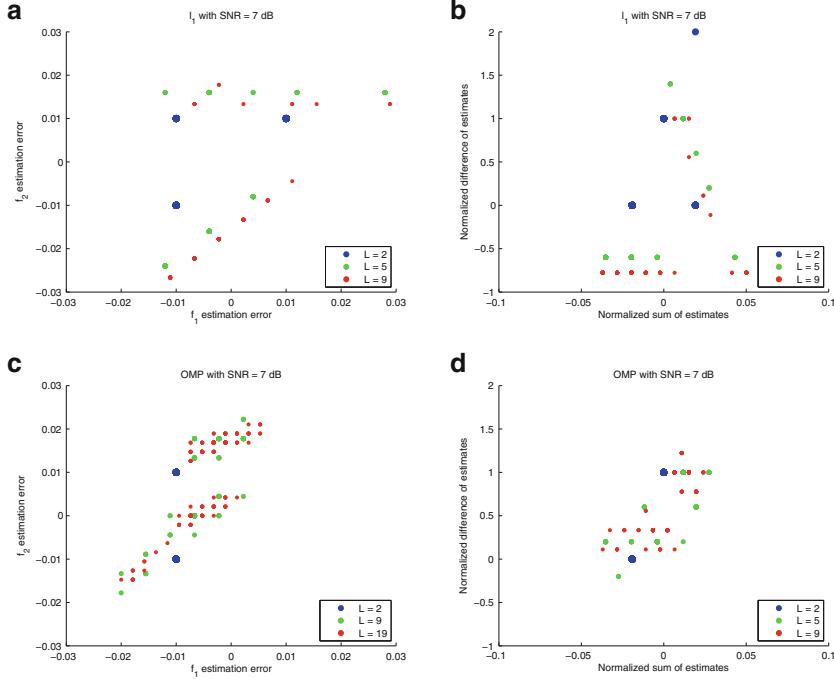


Fig. 3.9 Scatter plots of BP and OMP inversions at 7 dB SNR, for $L = 2, 5, 9$: (a) Scatter plot of estimator errors for (f_1, f_2) for BP inversions. (b) Scatter plot of normalized errors in estimating $((f_1 + f_2), (f_1 - f_2))$ using BP inversion. (c) Scatter plot of estimator errors of (f_1, f_2) for OMP inversions. (d) Scatter plot of normalized errors in estimating $((f_1 + f_2), (f_1 - f_2))$ using OMP.

3.5 Performance Bounds for Compressed Sensing with Basis Mismatch

When the RIC of $\Phi = A\Psi$ satisfies $\delta_{2k}(\Phi) < \sqrt{2} - 1$ for $2k$ -sparse signals, the solution \hat{c} to the basis pursuit problem (3.5) satisfies (see Chapter 1, Theorem 1.6, and also [4, 5], and [3])

$$\|\hat{c} - c\|_1 \leq C_0 \|c - \sigma_k(c)_1\|_1 \quad (3.14)$$

and

$$\|\hat{c} - c\|_2 \leq C_0 k^{-1/2} \|c - \sigma_k(c)_1\|_1, \quad (3.15)$$

where $\sigma_k(c)_1$ is the best k -term approximation to c in ℓ_1 norm and C_0 is a constant. These inequalities are at the core of compressed sensing theory, as they demonstrate that for sparse signals, where the best k -term approximation error $c - \sigma_k(c)_1$ vanishes, the signal recovery using basis pursuit is exact.

The analysis of [11], however, indicates that under basis mismatch the k -term approximation degenerates considerably and it fails to provide any guarantee for the solution of the mismatched basis pursuit problem. This is captured in the following theorem from [11].

Theorem 1 (best k -term approximation error). *Consider the coordinate transformation $c = \Psi^{-1}\tilde{\Psi}\tilde{c}$. Let $\beta = \max_{i,j} |\langle \psi_i, \tilde{\psi}_j \rangle|$ be the worst-case coherence between the columns of $\Psi = [\psi_0, \dots, \psi_{n-1}]$ and the columns of $\tilde{\Psi} = [\tilde{\psi}_0, \dots, \tilde{\psi}_{k-1}]$. Then,*

$$\|c - \sigma_k(c)_1\|_1 \leq (n - k)\beta\|\tilde{c}\|_1. \quad (3.16)$$

In [11], it is shown that under certain mismatch conditions the upper bound in (3.16) is achieved, resulting in the worst-case error. The bound in (3.16) can be combined with (3.14) and (3.15) to produce upper bounds on the inversion error for basis pursuit in the presence of basis mismatch. The error bounds scale linearly with n so that the bound increases with the size of the original image, independently of the compressed dimension, provided the RIP condition is maintained.

The bound in (3.16) is relevant to any sparse inversion algorithm that relies on best k -term approximation error bounds for its inversion guarantees. This class of algorithms includes regularized orthogonal matching pursuit (ROMP) [25] and CoSaMP [24] among others. However, we note that above theorem does not imply that the compressed sensing inversions will necessarily be bad. Rather it says that the inversions cannot be guaranteed to be good based on best k -term approximation reasoning.

3.6 Compressed Sensing Off The Grid

Recently, several approaches based on atomic norm minimization [7] have been proposed to eliminate the need for an apriori selected basis (or grid) in compressed sensing. However, the guaranteed theoretical resolution of these methods is a few Rayleigh limits, and sub-Rayleigh resolution of modes is not guaranteed. In particular, in [2] the authors show that a line spectrum with minimum frequency separation $\Delta_f > 4/k$ can be recovered from the first $2k$ Fourier coefficients via atomic norm minimization. In [31], the authors improve the resolution result by showing that a line spectrum with minimum frequency separation $\Delta_f > 4/n$ can be recovered from most subsets of the first n Fourier coefficients of size at least $m = \mathcal{O}(k \log(k) \log(n))$. This framework has been extended in [10] for 2-D line spectrum estimation. Another approach is proposed in [8] and [9], where the problem is reformulated as a structured matrix completion inspired by a classical work on matrix pencil [19]. These are all promising new directions that may provide pathways to high-resolution modal analysis in measurement deprived scenarios without the need to use prespecified grids for inversion. We refer the interested reader to [2, 7, 8, 10, 31], and [9] for details.

References

1. Baraniuk, R., Steeghs, P.: Compressive radar imaging. In: Proc. 2007 IEEE radar conf., pp. 128–133. Waltham, Massachusetts (2007)
2. Candes, E., Fernandez-Granda, C.: Towards a mathematical theory of super-resolution. preprint (Mar. 2012, arxiv:1203.5871.)
3. Candès, E.J.: The restricted isometry property and its implications for compressed sensing. *Académie des Sciences* **1**(346), 589–592 (2008)
4. Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory* **52**(2), 489–509 (2006)
5. Candès, E.J., Tao, T.: Decoding by linear programming. *IEEE Trans. Inform. Theory* **51**, 4203–4215 (2005)
6. Cevher, V., Gurbuz, A., McClellan, J., Chellappa, R.: Compressive wireless arrays for bearing estimation of sparse sources in angle domain. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 2497–2500. Las Vegas, Nevada (2008)
7. Chandrasekaran, V., Recht, B., Parrilo, P.A., Willsky, A.S.: The convex geometry of linear inverse problems. *Found. Comput. Math.* **12**(6), 805–849 (2012)
8. Chen, Y., Chi, Y.: Spectral compressed sensing via structured matrix completion. In: Proc. International Conference on Machine Learning (ICML) (2013)
9. Chen, Y., Chi, Y.: Robust spectral compressed sensing via structured matrix completion. preprint (2013, arXiv:1304.8126.)
10. Chi, Y., Chen, Y.: Compressive recovery of 2-D off-grid frequencies. In: Conf. rec. asilomar conf. signals, systems, and computers (2013)
11. Chi, Y., Scharf, L.L., Pezeshki, A., Calderbank, R.: Sensitivity to basis mismatch in compressed sensing. *IEEE Trans. Signal Process.* **59**(5), 2182–2195 (2011)
12. Duarte, M.F., Baraniuk, R.G.: Spectral compressive sensing. *Appl. Comput. Harmon. Anal.* **35**(1), 111–129 (2013)
13. Fannjiang, A., Yan, P., Strohmer, T.: Compressed remote sensing of sparse objects. *SIAM J. Imag. Sci.* **3**(3), 595–618 (2010)
14. Golub, G.H., Van Loan, C.F.: *Matrix Computations*, 3rd edn. John Hopkins Univ. Press, Baltimore, MD (1996)
15. Gurbuz, A., McClellan, J., Cevher, V.: A compressive beamforming method. In: Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), pp. 2617–2620. Las Vegas, NV (2008)
16. Herman, M.A., Needell, D.: Mixed operators in compressed sensing. In: Proc. 44th Annual Conference on Information Sciences and Systems (CISS). Princeton, NJ (2010)
17. Herman, M.A., Strohmer, T.: High-resolution radar via compressed sensing. *IEEE Trans. Signal Process.* **57**(6), 2275–2284 (2009)
18. Herman, M.A., Strohmer, T.: General deviants: an analysis of perturbations in compressed sensing. *IEEE J. Sel. Top. Sign. Proces.: Special Issue on Compressive Sens.* **4**(2), 342–349 (2010)
19. Hua, Y.: Estimating two-dimensional frequencies by matrix enhancement and matrix pencil. *IEEE Trans. Signal Process.* **40**(9), 2267–2280 (1992)
20. Karim, H., Viberg, M.: Two decades of array signal processing research: the parametric approach. *IEEE Signal Process. Mag.* **13**(4), 67–94 (1996)
21. Klemm, R.: *Space-Time Adaptive Processing*. IEEE Press, UK (1998)
22. McWhorter, L.T., Scharf, L.L.: Cramer-Rao bounds for deterministic modal analysis. *IEEE Trans. Signal Process.* **41**(5), 1847–1862 (1993)
23. Mullis, C.T., Scharf, L.L.: Quadratic estimators of the power spectrum. In: Haykin, S. (ed.) *Advances in Spectrum Estimation*, vol. 1, chap. 1, pp. 1–57. Prentice Hall, Englewood Cliffs, NJ (1990)
24. Needell, D., Tropp, J.: CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.* **26**, 301–321 (2008)

25. Needell, D., Vershynin, R.: Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Found. Comput. Math.* **9**, 317–334 (2009)
26. Needell, D., Vershynin, R.: Signal recovery from inaccurate and incomplete measurements via regularized orthogonal matching pursuit. *IEEE J. Sel. Top. Sign. Proces.* **4**(2), 310–316 (2010)
27. Pezeshki, A., Veen, B.D.V., Scharf, L.L., Cox, H., Lundberg, M.: Eigenvalue beamforming using a generalized sidelobe canceller and subset selection. *IEEE Trans. Signal Process.* **56**(5), 1954–1967 (2008)
28. Scharf, L.L.: *Statistical Signal Processing*. Addison-Wesley, MA (1991)
29. Scharf, L.L., Chong, E.K.P., Pezeshki, A., Luo, J.: Sensitivity considerations in compressed sensing. In: Conf. Rec. Forty-fifth Asilomar Conf. Signals, Syst. Pacific Grove, CA (2011)
30. Scharf, L.L., Friedlander, B.: Matched subspace detectors. *IEEE Trans. Signal Process.* **42**(8), 2146–2157 (1994)
31. Tang, G., Bhaskar, B.N., Shah, P., Recht, B.: Compressed sensing off the grid. preprint (Jul. 2012, arxiv:1207.6053.)
32. Trees, H.L.V., Bell, K.L.: *Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking*. IEEE Press, New York (2007)
33. Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory* **53**(12) (1992)
34. Tufts, D.W., Kumaresan, R.: Estimation of frequencies of multiple sinusoids: making linear prediction perform like maximum likelihood. *Proc. IEEE* **70**, 975–989 (1982)
35. Tufts, D.W., Kumaresan, R.: Singular value decomposition and improved frequency estimation using linear prediction. *IEEE Trans. Acoust. Speech Signal Process.* **30**(4), 671–675 (1982)
36. Van Trees, H.L.: *Optimum Array Processing*. Wiley Interscience, New York (2002)
37. Ward, J.: Maximum likelihood angle and velocity estimation with space-time adaptive processing radar. In: Conf. Rec. 1996 Asilomar Conf. Signals, Systs., Comput., pp. 1265–1267. Pacific Grove, CA (1996)

Chapter 4

Recovering Structured Signals in Noise: Least-Squares Meets Compressed Sensing

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi

Abstract The typical scenario that arises in most “big data” problems is one where the ambient dimension of the signal is very large (e.g., high resolution images, gene expression data from a DNA microarray, social network data, etc.), yet is such that its desired properties lie in some low dimensional structure (sparsity, low-rankness, clusters, etc.). In the modern viewpoint, the goal is to come up with efficient algorithms to reveal these structures and for which, under suitable conditions, one can give theoretical guarantees. We specifically consider the problem of recovering such a structured signal (sparse, low-rank, block-sparse, etc.) from noisy compressed measurements. A general algorithm for such problems, commonly referred to as generalized LASSO, attempts to solve this problem by minimizing a least-squares cost with an added “structure-inducing” regularization term (ℓ_1 norm, nuclear norm, mixed ℓ_2/ℓ_1 norm, etc.). While the LASSO algorithm has been around for 20 years and has enjoyed great success in practice, there has been relatively little analysis of its performance. In this chapter, we will provide a full performance analysis and compute, in closed form, the mean-square-error of the reconstructed signal. We will highlight some of the mathematical vignettes necessary for the analysis, make connections to noiseless compressed sensing and proximal denoising, and will emphasize the central role of the “statistical dimension” of a structured signal.

4.1 Introduction

Consider the standard linear regression setup, in which we are interested in recovering an unknown signal $\mathbf{x}_0 \in \mathbb{R}^n$ from a vector $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z} \in \mathbb{R}^m$ of noisy linear observations.

C. Thrampoulidis (✉) • S. Oymak • B. Hassibi
California Institute of Technology, Pasadena, CA, USA
e-mail: cthrampo@caltech.edu; soymak@caltech.edu; hassibi@caltech.edu

4.1.1 Ordinary Least-Squares

The most prominent method for estimating the unknown vector \mathbf{x}_0 is that of ordinary least-squares (OLS). The OLS estimate $\hat{\mathbf{x}}$ of \mathbf{x}_0 is obtained by minimizing the residual squared error

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{Ax}\|_2^2. \quad (4.1)$$

OLS has a long history originating in the early 1800s due to works by Gauss and Legendre [41, 53], and its behavior is by now very well understood. In particular, in the classical setting $m > n$, assuming \mathbf{A} is full column-rank, (4.1) has a unique solution which is famously given by

$$\hat{\mathbf{x}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{y}. \quad (4.2)$$

The squared error-loss of the OLS estimate in (4.2) is, thus, expressed as

$$\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2 = \mathbf{z}^T \mathbf{A} (\mathbf{A}^T \mathbf{A})^{-2} \mathbf{A}^T \mathbf{z}. \quad (4.3)$$

Starting from (4.3) and imposing certain generic assumptions on the measurement matrix \mathbf{A} and/or the noise vector \mathbf{z} , it is possible to conclude precise and simple formulae characterizing the estimation error $\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2$. As an example, when the entries of \mathbf{z} are drawn i.i.d. normal of zero-mean and variance σ^2 , then $\mathbb{E} \|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2 = \sigma^2 \text{trace}((\mathbf{A}^T \mathbf{A})^{-1})$. Furthermore, when the entries of \mathbf{A} are drawn i.i.d. normal of zero-mean and variance $1/m$, $\mathbf{A}^T \mathbf{A}$ is a Wishart matrix whose asymptotic eigendistribution is well known. Using this, and letting m, n grow, we find that the squared error concentrates around

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2} \approx \frac{n}{m-n}. \quad (4.4)$$

Such expressions serve as valuable insights regarding the behavior of the OLS estimator and are meant to provide guidelines for the effectiveness of the estimator in practical situations.

4.1.2 Structured Signals and Noisy Compressed Sensing

Gauss and Legendre proposed the OLS method in the context of traditional statistical data analysis, in which one assumes a large number m of observations and only a small number n of well-chosen variables corresponding to the entries of the desired signal \mathbf{x}_0 . That trend is rapidly changing in the today's era of "big-data" and of the collection of massively large data of all kinds. In a number of

disciplines and application domains including machine learning, signal processing, social networks, and DNA microarrays, we are increasingly confronted with *very large* data sets where we need to extract some signal-of-interest. In this setting, it is important to have signal recovery algorithms that are computationally efficient and that need not access the entire data directly, i.e. *can reliably estimate \mathbf{x}_0 from fewer number of observations m than the dimension n of the model to be estimated*.

This problem of compressed recovery is in general ill-posed. Even in the absence of noise, recovering a general signal \mathbf{x}_0 of dimension n from only $m < n$ linear observations $\mathbf{y} = \mathbf{Ax}_0$ is futile, since there is typically infinitely many solutions satisfying the measurement equations. Thus, on the face of it the limited availability of data relative to the ambient dimension of the signal to be estimated can lead to the *curse of dimensionality* [16]. Fortunately, in many applications, the signal of interest lives in a manifold of much lower dimension than that of the original ambient space.

We refer to such signals that are structurally constrained to only have very few degrees of freedom relative to the ambient dimension, as *structured signals*. The most celebrated example of structure in the compressed sensing literature is sparsity. To appreciate how knowledge of the structure of the unknown signal \mathbf{x}_0 can help alleviate the ill-posed nature of the problem, consider a desired signal \mathbf{x}_0 which is k -sparse i.e., has only $k < n$ (often $k \ll n$) non-zero entries. Suppose we make m noisy measurements of \mathbf{x}_0 using the $m \times n$ measurement matrix \mathbf{A} to obtain $\mathbf{y} = \mathbf{Ax}_0 + \mathbf{z}$ and further suppose each set of m columns of \mathbf{A} be linearly independent. Then, as long as $m > k$, we can always find the sparsest solution to

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{Ax}\|_2^2,$$

via exhaustive search of $\binom{n}{k}$ such least-squares problems. Under the same assumptions that led to (4.4), this gives a normalized squared error

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2} \approx \frac{k}{m - k}. \quad (4.5)$$

The catch here, of course, is that the computational complexity of the estimation procedure that was just described is exponential in the ambient dimension n , thus, making it intractable. So, the fundamental question to ask is whether we can estimate the *structured* (here, sparse) signal \mathbf{x}_0 in a computationally efficient way. And if so, how many measurements m are needed? *How does the squared error behave and how does it compare to (4.4) and (4.5)?*

4.1.3 LASSO

LASSO is a method for reconstructing the unknown signal \mathbf{x}_0 and is particularly useful in the undetermined setting $m < n$ and when one seeks *sparse* signals. It was originally proposed by Tibshirani in 1996 [61], and consists of solving the following constrained optimization problem

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{Ax}\|_2^2 \quad \text{s.t. } \|\mathbf{x}\|_1 \leq \|\mathbf{x}_0\|_1. \quad (4.6)$$

The program in (4.6) solves for an estimate $\hat{\mathbf{x}}$ that best fits the vector of observations \mathbf{y} , in the OLS sense, while at the same time is constrained to retain structure similar to that of \mathbf{x}_0 . At this point, recall that the ℓ_1 -norm is typically used in the compressed sensing literature to promote sparse solutions.

Solving (4.6) requires *a priori* knowledge of $\|\mathbf{x}_0\|_1$. When such knowledge is not available, one can instead solve *regularized* versions of it, like,

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{Ax}\|_2 + \lambda \|\mathbf{x}\|_1, \quad (4.7)$$

or

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{y} - \mathbf{Ax}\|_2^2 + \tau \|\mathbf{x}\|_1, \quad (4.8)$$

for nonnegative regularizer parameters λ and τ . Although very similar in nature, (4.7) and (4.8) show in general different statistical behaviors [4, 47]. Lagrange duality ensures that there exist λ and τ such that they both become equivalent to the constrained optimization (4.6). However, in practice, the challenge lies in tuning the regularizer parameters to achieve good estimation, with as little possible prior knowledge¹. Observe that letting $\lambda = \tau \rightarrow 0$ reduces (4.7) and (4.8) to the standard ℓ_1 -minimization [10, 17]

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t. } \mathbf{y} = \mathbf{Ax}.$$

The problem in (4.8) is arguably the most encountered in the literature and amongst (4.6) and (4.7), it is the one which is widely termed the LASSO estimator.

4.1.4 Generalized LASSO

The Generalized LASSO algorithm is a generalization of the standard LASSO introduced by Tibshirani, which can be used to enforce other types of structures apart from sparsity. In order to maintain a generic treatment of “structure,” we associate with it a convex function $f(\cdot)$. We will commonly refer to $f(\cdot)$ as the structure-inducing or structure-promoting function. In the case of sparse signals, f is chosen to be the ℓ_1 -norm. When the unknown signal is a low-rank matrix, f is chosen as the nuclear norm (sum of the singular values) and for a block-sparse signal the

¹ Assuming that the entries of the noise vector \mathbf{z} are i.i.d., it is well known that a sensible choice of τ in (4.8) must scale with the standard deviation σ of the noise components [6, 12, 42]. On the other hand, (4.7) eliminates the need to know or to pre-estimate σ [4].

associated structure-inducing function is the $\ell_{1,2}$ -norm. Recently, Chandrasekaran et al. [14] have proposed a principled way to convert notions of simplicity into such corresponding convex structure-promoting functions.

When a structure-inducing function f is available, we can try estimate \mathbf{x}_0 by solving generalized versions of (4.6), (4.7), and (4.8) where the ℓ_1 -norm is substituted by f . For instance, the generalized version of (4.8) solves for

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{y} - \mathbf{Ax}\|_2^2 + \tau f(\mathbf{x}).$$

4.1.5 The Squared-Error of the Generalized LASSO

The Generalized LASSO algorithm solves a convex optimization problem and is, thus, a computationally tractable way for estimating the desired signal \mathbf{x}_0 . When viewed as a natural extension of the OLS to the high-dimensional setting $m < n$, it is natural to ask whether we can give performance bounds for the squared error of the Generalized LASSO. While the LASSO method has been around for 20 years and has enjoyed great success in practice, there has been relatively little such analysis of its performance. In particular, most performance bounds derived in the literature are rather loose ([59] and the references therein). Here, we are interested in bounds that are sharp and simple, similar to those that characterize the OLS. In particular, under same assumptions on the distribution of the measurement matrix and the noise vector, we ask whether it is possible to derive bounds that resemble (4.4) and (4.5). It turns out that we can and this chapter is dedicated to providing a full performance analysis and computation of such bounds.

4.1.6 Organization

In Section 4.2 we review OLS. Section 4.3 formally introduces the generalized LASSO, and sets the goals of the chapter, while Section 4.4 reviews the relevant technical background. Sections 4.5, 4.6, and 4.7 are devoted to the analysis of the squared error of the generalized LASSO. We conclude in Section 4.8 with directions for future work.

4.2 Least Squares

We start by briefly reviewing the OLS equations and derive performance bounds under the generic assumption that the entries of \mathbf{A} are i.i.d. zero-mean normal with variance $1/m$. In Section 4.2.1 we examine the case where the entries of the noise

vector \mathbf{z} are also i.i.d. zero-mean normal with variance σ^2 . In Section 4.2.2 we compute the squared error of OLS for any *fixed* noise vector \mathbf{z} , while in Section 4.2.3 we perform an analysis of the worst-case error of OLS.

Recall that the OLS solves (4.1). It is clear that when $m < n$, (4.1) is ill posed. However, when $m > n$ and \mathbf{A} has i.i.d. normal entries, \mathbf{A} is full column rank with high probability. The solution of (4.1) is then unique and given by $\hat{\mathbf{x}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{y}$. Recalling that $\mathbf{y} = \mathbf{A} \mathbf{x}_0 + \mathbf{z}$, this gives a squared error-loss as in (4.3).

4.2.1 Gaussian Noise

For the purposes of this section, further assume that the entries of \mathbf{z} are i.i.d. zero-mean normal with variance σ^2 and independent of the entries of \mathbf{A} . In this case, the normalized mean-squared-error takes the form,

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2 &= \mathbb{E}[\mathbf{z}^T \mathbf{A} (\mathbf{A}^T \mathbf{A})^{-2} \mathbf{A}^T \mathbf{z}] \\ &= \sigma^2 \mathbb{E}[\text{trace}(\mathbf{A} (\mathbf{A}^T \mathbf{A})^{-2} \mathbf{A}^T)] \\ &= \sigma^2 \mathbb{E}[\text{trace}((\mathbf{A}^T \mathbf{A})^{-1})].\end{aligned}$$

$\mathbf{A}^T \mathbf{A}$ is a Wishart matrix and the distribution of its inverse is well studied. In particular, when $m > n + 1$, we have $\mathbb{E}[(\mathbf{A}^T \mathbf{A})^{-1}] = \frac{m}{m-n-1} \mathbf{I}_n$ [35]. Hence,

$$\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2 = m\sigma^2 \frac{n}{(m-1)-n}.$$

Noting that $\mathbb{E}\|\mathbf{z}\|_2^2 = m\sigma^2$ and letting m, n large enough we conclude with the stronger concentration result on the squared-error of OLS:

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2} \approx \frac{n}{(m-1)-n}. \quad (4.9)$$

4.2.2 Fixed Noise

Fix any noise vector \mathbf{z} , with the only assumption being that it is chosen independently of the measurement matrix \mathbf{A} . Denote the projection of \mathbf{z} onto the range space of \mathbf{A} by $\text{Proj}(\mathbf{z}, \text{Range}(\mathbf{A})) := \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{z}$ and the minimum singular value of \mathbf{A} by $\sigma_{\min}(\mathbf{A})$. Then,

$$\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2 \leq \frac{\|\mathbf{A}(\hat{\mathbf{x}} - \mathbf{x}_0)\|_2}{\sigma_{\min}(\mathbf{A})} = \frac{\|\text{Proj}(\mathbf{z}, \text{Range}(\mathbf{A}))\|_2}{\sigma_{\min}(\mathbf{A})}. \quad (4.10)$$

It is well known that, when \mathbf{A} has entries i.i.d. zero-mean normal with variance $1/m$, then $\sigma_{\min}(\mathbf{A}) \approx 1 - \sqrt{\frac{n}{m}}$, [63]. Also, since \mathbf{z} is independent of \mathbf{A} , and the range space of \mathbf{A} is uniformly random subspace of dimension n in \mathbb{R}^m , it can be shown that $\|\text{Proj}(\mathbf{z}, \text{Range}(\mathbf{A}))\|_2^2 \approx \frac{n}{m} \|\mathbf{z}\|_2^2$ (e.g., [9, p. 13]). With these, we conclude that with high probability on the draw of \mathbf{A} ,

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2} \lesssim \frac{n}{(\sqrt{m} - \sqrt{n})^2}. \quad (4.11)$$

4.2.3 Worst-Case Squared-Error

Next, assume no restriction at all on the noise vector \mathbf{z} . In particular, this includes the case of *adversarial* noise, i.e., noise that has information of the sensing matrix \mathbf{A} and can adapt itself accordingly. As expected, this can cause the reconstruction error to be, in general, significantly worse than the guarantees in (4.9) and (4.11). In more detail, we can write

$$\begin{aligned} \|\hat{\mathbf{x}} - \mathbf{x}_0\|_2 &\leq \frac{\|\mathbf{A}(\hat{\mathbf{x}} - \mathbf{x}_0)\|_2}{\sigma_{\min}(\mathbf{A})} = \frac{\|\mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{z}\|_2}{\sigma_{\min}(\mathbf{A})} \\ &\leq \|\mathbf{z}\|_2 \frac{\|\mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T\|_2}{\sigma_{\min}(\mathbf{A})} \leq \|\mathbf{z}\|_2 \sigma_{\min}^{-1}(\mathbf{A}), \end{aligned} \quad (4.12)$$

where $\|\mathbf{M}\|_2$ denotes the spectral norm of a matrix \mathbf{M} and we used the fact that the spectral norm of a symmetric projection matrix is upper bounded by 1. It is not hard to show that equality in (4.12) is achieved when \mathbf{z} is equal to the left singular value of \mathbf{A} corresponding to its minimum singular value. Using the fact that $\sigma_{\min}(\mathbf{A}) \approx 1 - \sqrt{\frac{n}{m}}$, we conclude that

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2} \lesssim \frac{m}{(\sqrt{n} - \sqrt{m})^2}. \quad (4.13)$$

4.3 The Generalized LASSO

In Section 4.3.1, we formally introduce the generalized LASSO algorithm. We do so by following a “compressed sensing” perspective. In particular, we view the LASSO as an algorithm performing *noisy compressed sensing*. This motivates the setup of our analysis, such as the assumption on the entries of the sensing matrix A being

i.i.d. normal, which is typical for yielding precise analytical results in the field of compressed sensing (e.g., [1, 10, 14, 17, 55]). Also, it helps appreciate the fact that our analysis of the error of the LASSO involves and builds upon notions of convex geometry that are also inherent in the classical analysis of noiseless compressed sensing.

Next, in Section 4.3.2 we restrict attention to the case where the unknown signal is either sparse or low-rank. For those, we present the formulae that characterize the LASSO error under different assumptions on the noise vector. We also contrast them to the ones corresponding to the OLS as they were discussed in Section 4.2. Generalizations of the formulae to arbitrary convex structure-inducing functions and the formal statement of the results follow in Sections 4.5–4.7 after introducing the necessary mathematical background in Section 4.4.

4.3.1 Noisy Compressed Sensing

The LASSO problem can be viewed as a “merger” of two closely related problems, the problems of noiseless compressed sensing (CS) and that of proximal denoising.

4.3.1.1 Noiseless compressed sensing

In the noiseless CS problem one wishes to recover $\mathbf{x}_0 \in \mathbb{R}^n$ from the random linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x}_0 \in \mathbb{R}^m$. A common approach is solving the following convex optimization problem

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \quad (4.14)$$

A critical performance criteria for the problem (4.14) concerns the minimum number of measurements needed to guarantee successful recovery of \mathbf{x}_0 [1, 14, 18, 19, 21, 55, 57]. Here, success means that \mathbf{x}_0 is the unique minimizer of (4.14), with high probability, over the realizations of the random matrix \mathbf{A} .

4.3.1.2 Proximal denoising

The proximal denoising problem tries to estimate \mathbf{x}_0 from noisy but uncompressed observations $\mathbf{y} = \mathbf{x}_0 + \mathbf{z}$, where the entries of \mathbf{z} are i.i.d. zero-mean Gaussian with variance σ^2 . In particular, it solves

$$\min_{\mathbf{x}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda \sigma f(\mathbf{x}) \right\}. \quad (4.15)$$

A closely related approach to estimate \mathbf{x}_0 , which requires prior knowledge $f(\mathbf{x}_0)$ about the signal of interest \mathbf{x}_0 , is solving the constrained denoising problem:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{x}\|_2^2 \text{ subject to } f(\mathbf{x}) \leq f(\mathbf{x}_0). \quad (4.16)$$

The natural question to be posed in both cases is how well one can estimate \mathbf{x}_0 via (4.15) (or (4.16)) [13, 15, 27, 43]. The minimizer $\hat{\mathbf{x}}$ of (4.15) (or (4.16)) is a function of the noise vector \mathbf{z} and the common measure of performance is the normalized mean-squared-error which is defined as $\frac{\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2}{\sigma^2}$.

4.3.1.3 The “merger” LASSO

The Generalized LASSO problem is naturally merging the problems of noiseless CS and proximal denoising. One assumes compressed and noisy measurements $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z} \in \mathbb{R}^m$ of an unknown signal $\mathbf{x}_0 \in \mathbb{R}^n$. The following three algorithms are all variants of the LASSO that try to estimate \mathbf{x}_0 .

* **C-LASSO**²:

$$\hat{\mathbf{x}}_c(\mathbf{A}, \mathbf{z}) = \arg \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \text{ subject to } f(\mathbf{x}) \leq f(\mathbf{x}_0). \quad (4.17)$$

* **ℓ_2 -LASSO**³:

$$\hat{\mathbf{x}}_{\ell_2}(\lambda, \mathbf{A}, \mathbf{z}) = \arg \min_{\mathbf{x}} \left\{ \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 + \frac{\lambda}{\sqrt{m}} f(\mathbf{x}) \right\}. \quad (4.18)$$

* **ℓ_2^2 -LASSO**:

$$\hat{\mathbf{x}}_{\ell_2^2}(\tau, \mathbf{A}, \mathbf{z}) = \arg \min_{\mathbf{x}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \frac{\sigma\tau}{\sqrt{m}} f(\mathbf{x}) \right\}. \quad (4.19)$$

The compressed nature of observations poses the question of finding the minimum number of measurements required to recover \mathbf{x}_0 *robustly*, that is with error proportional to the noise level. When recovery is robust, it is of importance to be able to explicitly characterize how good the estimate is. In this direction, a common measure of performance for the LASSO estimate $\hat{\mathbf{x}}$ is defined to be the **normalized squared error** (NSE):

²C-LASSO in (4.17) stands for “Constrained LASSO.” The algorithm assumes *a priori* knowledge of $f(\mathbf{x}_0)$.

³In the statistics literature the variant of the LASSO algorithm in (4.7) is mostly known as the “square-root LASSO” [4]. For the purposes of our presentation, we stick to the more compact term “ ℓ_2 -LASSO” [47].

$$\text{NSE} = \frac{\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2}.$$

Certainly, the NSE is not the sole measure of performance of the estimator and other measures can be of special interest for different kinds of structures of the unknown vector and for specific applications. As an example, if \mathbf{x}_0 is sparse, then the “support recovery criteria,” which measures how well the LASSO recovers the subset of nonzero indices of \mathbf{x}_0 might be of interest [24, 64, 66].

The results presented in this chapter are on the NSE of the generalized LASSO. When restricted to sparse signals, this has been the subject of analysis of lots of research papers (e.g., see [6, 11, 42] and the references therein). Yet, those bounds are in general order-wise and do not capture precisely the behavior of the NSE. In [3] and [56] manage to accurately characterize the NSE when \mathbf{z} is i.i.d. normal, \mathbf{x}_0 is sparse and f is the ℓ_1 -norm. In [3, 25], Bayati and Montanari are able to prove that the mean-squared-error of the LASSO problem is equivalent to the one achieved by a properly defined “Approximate Message Passing” (AMP) algorithm [38]. Following this connection and after evaluating the error of the AMP algorithm, they obtain an explicit expression for the mean squared error of the LASSO algorithm in an asymptotic setting. In [39], Maleki et al. propose Complex AMP, and characterize the performance of LASSO for sparse signals with complex entries. On the other hand, Stojnic’s approach [56] relies on results on Gaussian processes [33, 34] to derive sharp bounds for the *worst case NSE* of the ℓ_1 -constrained LASSO problem in (4.17). Subsequent work in [47] builds upon the analysis framework introduced in [56] and generalizes the results of the latter to the regularized LASSO and to arbitrary convex functions. In [46] and [60], sharp upper bounds on the NSE of the C-LASSO and the ℓ_2 -LASSO are derived, for noise vector provided that it is chosen independently of the sensing matrix. The outcome of those works is a general theory that yields precise bounds for the NSE of the LASSO algorithm for arbitrary convex regularizer $f(\cdot)$. The bounds are simple in nature and offer nice interpretations as generalizations of the classical OLS error bounds discussed in Section 4.2.

4.3.2 Motivating Examples

In Section 4.2 and in particular in equations (4.9)–(4.13), we reviewed classical bounds on the normalized square-error of the OLS, which corresponds to the LASSO in the trivial case $f(\cdot) = 0$. How do those results change when a nontrivial convex function $f(\cdot)$ is introduced? What is a precise and simple upper bound on the NSE of the LASSO when the unknown signal is sparse and $f(\cdot) = \|\cdot\|_1$? What if the unknown signal is low-rank and nuclear norm is chosen as the regularizer? Is it possible to generalize such bounds to arbitrary structures and corresponding convex regularizers? In the following paragraphs, we assume that the entries of the sensing matrix \mathbf{A} are i.i.d. zero-mean normal with variance $1/m$ and provide answers to those questions. The details and proofs of the statements made are deferred to Sections 4.5–4.7.

4.3.2.1 Sparse Signal Estimation

Assume $\mathbf{x}_0 \in \mathbb{R}^n$ has k nonzero entries. We estimate \mathbf{x}_0 via the LASSO with f being the ℓ_1 -norm. First, suppose that the noise vector has i.i.d. zero-mean normal entries with variance σ^2 . Then, the NSE of the C-LASSO admits the following sharp upper bound⁴, which is attained in the limit as the noise variance σ^2 goes to zero [47, 56]:

$$\frac{\|\hat{\mathbf{x}}_c - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2} \lesssim \frac{2k(\log \frac{n}{k} + 1)}{m - 2k(\log \frac{n}{k} + 1)}. \quad (4.20)$$

Compare this to the formula (4.9) for the OLS. (4.20) is obtained from (4.9) after simply replacing the ambient dimension n in the latter with $2k(\log \frac{n}{k} + 1)$. Also, while (4.9) requires $m > n$, (4.20) relaxes this requirement to $m > 2k(\log \frac{n}{k} + 1)$. This is to say that any number of measurements greater than $2k(\log \frac{n}{k} + 1) \ll n$ are sufficient to guarantee robust recovery. Note that this coincides with the classical phase-transition threshold in the noiseless compressed sensing [14, 55]. If instead of the C-LASSO, one uses the ℓ_2 -LASSO with $\lambda \geq \sqrt{2 \log \frac{n}{k}}$, then

$$\frac{\|\hat{\mathbf{x}}_{\ell_2} - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2} \lesssim \frac{(\lambda^2 + 3)k}{m - (\lambda^2 + 3)k}. \quad (4.21)$$

Again, observe how (4.21) is obtained from (4.9) after simply replacing the ambient dimension n with $(\lambda^2 + 3)k$. The role of the regularizer parameter λ is explicit in (4.21). Also, substituting $\lambda \approx \sqrt{2 \log \frac{n}{k}}$ in (4.21) (almost) recovers (4.20). This suggests that choosing this value of the regularizer parameter is optimal in that it results in the regularized LASSO (4.7) to perform as good as the constrained version (4.6). Note that this value for the optimal regularizer parameter only depends on the sparsity level k of the unknown signal \mathbf{x}_0 and *not* the unknown signal itself.

Next, consider the more general case in which the noise vector \mathbf{z} can be anything but drawn independently of the sensing matrix \mathbf{A} . If ones uses the C-LASSO to estimate \mathbf{x}_0 , then the estimation error is bounded as follows⁵ [46]:

$$\frac{\|\hat{\mathbf{x}}_c - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2} \lesssim \frac{2k(\log \frac{n}{k} + 1)}{(\sqrt{m} - \sqrt{2k(\log \frac{n}{k} + 1)})^2}. \quad (4.22)$$

Accordingly, the ℓ_2 -LASSO for $\lambda \geq \sqrt{2 \log \frac{n}{k}}$ gives [60]:

$$\frac{\|\hat{\mathbf{x}}_{\ell_2} - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2} \lesssim 2 \frac{(\lambda^2 + 3)k}{(\sqrt{m} - \sqrt{(\lambda^2 + 3)k})^2}. \quad (4.23)$$

⁴ The statements in this sections hold true with high probability in \mathbf{A}, \mathbf{z} and under mild assumptions. See Section 4.5 for the formal statement of the results.

⁵ The formula below is subject to some simplifications meant to highlight the essential structure. See Section 4.6 for the details.

Once more, (4.22) and (4.23) resemble the corresponding formula describing OLS in (4.11). The only difference is that the ambient dimension n is substituted with $2k(\log \frac{n}{k} + 1)$ and $(\lambda^2 + 3)k$, respectively⁶. As a last comment, observe that the bounds in (4.22) and (4.23) have squares in the denominators. This is in contrast to equations (4.20) and (4.21). A detailed comparison is included in Section 4.6.1.2.

4.3.2.2 Low-rank Matrix Estimation

Assume $\mathbf{X}_0 \in \mathbb{R}^{\sqrt{n} \times \sqrt{n}}$ is a rank- r matrix, and let $\mathbf{x}_0 = \text{vec}(\mathbf{X}_0) \in \mathbb{R}^n$ be the vectorization of \mathbf{X}_0 . We use the generalized LASSO with $f(\mathbf{x}) = \|\text{vec}^{-1}(\mathbf{x})\|_*$. The nuclear norm of a matrix (i.e., sum of singular values) is known to promote low-rank solutions [30, 50].

As previously, suppose first that \mathbf{z} has i.i.d. zero-mean normal entries with variance σ^2 . Then, the NSE of the C-LASSO and that of the ℓ_2 -LASSO for $\lambda \geq 2n^{1/4}$ are bounded as follows:

$$\frac{\|\hat{\mathbf{x}}_c - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2} \lesssim \frac{6\sqrt{nr}}{m - 6\sqrt{nr}}, \quad (4.24)$$

and

$$\frac{\|\hat{\mathbf{x}}_c - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2} \lesssim \frac{\lambda^2 r + 2\sqrt{n}(r+1)}{m - (\lambda^2 r + 2\sqrt{n}(r+1))}. \quad (4.25)$$

Just like in the estimation of sparse signals in Section 4.3.2.1, it is clear from the bounds above that they can be obtained from the OLS bound in (4.9) after only substituting the dimension of the ambient space n with $6\sqrt{nr}$ and $\lambda^2 r + 2\sqrt{n}(r+1)$, respectively. And again, $6\sqrt{nr}$ is exactly the phase transition threshold for the noiseless compressed sensing of low-rank matrices [1, 14, 44].

Moving to the case where \mathbf{z} is arbitrary but independent of \mathbf{A} , we find that

$$\frac{\|\hat{\mathbf{x}}_c - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2} \lesssim \frac{6\sqrt{nr}}{(\sqrt{m} - 6\sqrt{nr})^2}, \quad (4.26)$$

and

$$\frac{\|\hat{\mathbf{x}}_{\ell_2} - \mathbf{x}_0\|_2^2}{\|\mathbf{z}\|_2^2} \lesssim 2 \frac{\lambda^2 r + 2\sqrt{n}(r+1)}{(\sqrt{m} - \sqrt{\lambda^2 r + 2\sqrt{n}(r+1)})^2}. \quad (4.27)$$

⁶The factor of 2 in (4.23) is conjectured in [60] that is not essential and, only, appears as an artifact of the proof technique therein. See, also, Section 4.6.2.1.

4.3.2.3 General Structures

From the discussion in Sections 4.3.2.1 and 4.3.2.2, it is becoming clear that the error bounds for the OLS admit nice and simple generalizations to error bounds for the generalized LASSO. What changes in the formulae bounding the NSE of the OLS when considering the NSE of the LASSO, is only that the ambient dimension n is substituted by a specific parameter.

This parameter depends on the particular structure of the unknown signal, but not the signal itself. For example, in the sparse case, it depends only on the sparsity of \mathbf{x}_0 , not \mathbf{x}_0 itself, and in the low-rank case, it only depends on the rank of \mathbf{X}_0 , not \mathbf{X}_0 itself. Furthermore, it depends on the structure-inducing function $f(\cdot)$ that is being used. Finally, it is naturally dependent on whether the constrained or the regularized LASSO is being used. In the case of regularized LASSO, it also depends on the value λ of the regularizer parameter. Interestingly, the value of this parameter corresponding to the NSE of the constrained LASSO is exactly the phase-transition threshold of the corresponding noiseless CS problem.

Let us use the following notation, $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ to represent this parameter for the cases of the constrained and regularized LASSO, correspondingly. The notation used makes explicit the dependence of the parameter on the quantities just discussed. The formal definition of $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ is not at all involved. In fact, they both have a nice and insightful geometric interpretation and are not hard to compute for many practical examples of interest (see the Appendix). We will present all these in Section 4.4.

To conclude this section, we repeat once more: *the classical and well-known error analysis of the NSE of the OLS can be naturally extended to describe the NSE of the generalized LASSO*. In particular, when the entries of \mathbf{A} are i.i.d. normal, then an error bound on the NSE of the OLS translates to a bound on the NSE of the generalized (constrained or regularized) LASSO after (almost) only substituting the ambient dimension n by either $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ or $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$. $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ are *summary parameters* that capture the geometry of the LASSO problem. For instance, for $f(\mathbf{x}) = \ell_1$ and \mathbf{x}_0 a k -sparse vector, it will be seen that $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) = 2k \log \frac{n}{k}$. Hence, one obtains (4.20) and (4.22) when substituting the n with $2k \log \frac{n}{k}$ in (4.9) and (4.11), respectively.

4.4 Background

This section is meant to summarize the background required for the formal statement of the results on the NSE of the generalized LASSO. In Section 4.4.2 we revise fundamental notions of convex geometry. Next, in Section 4.4.3 we outline the main tools that underlie the analysis. These include a strong probabilistic comparison lemma on Gaussian processes proven by Gordon in [34], as well as some standard but powerful concentration inequalities on the Gaussian measure.

4.4.1 Notation

For the rest of the paper, let $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution of mean μ and variance σ^2 . Also, to simplify notation, let us write $\|\cdot\|$ instead of $\|\cdot\|_2$. For a vector $\mathbf{g} \in \mathbb{R}^m$ with independent $\mathcal{N}(0, 1)$ entries, we define $\gamma_m := \mathbb{E}[\|\mathbf{g}\|]$. It is well known ([34]) that $\gamma_m = \sqrt{2} \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})}$ and $\sqrt{m} \geq \gamma_m \geq \frac{m}{\sqrt{m+1}}$. Also, we reserve the variables \mathbf{h} and \mathbf{g} to denote i.i.d. Gaussian vectors in \mathbb{R}^n and \mathbb{R}^m , respectively. Finally, the Euclidean unit ball and unit sphere are, respectively, denoted as

$$\mathcal{B}^{n-1} := \{\mathbf{v} \in \mathbb{R}^n \mid \|\mathbf{v}\| \leq 1\} \quad \text{and} \quad \mathcal{S}^{n-1} := \{\mathbf{v} \in \mathbb{R}^n \mid \|\mathbf{v}\| = 1\}.$$

4.4.2 Convex Geometry

4.4.2.1 Subdifferential

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex function and $\mathbf{x}_0 \in \mathbb{R}^n$ be an arbitrary point that is *not* a minimizer of $f(\cdot)$. The subdifferential of $f(\cdot)$ at \mathbf{x}_0 is the set of vectors,

$$\partial f(\mathbf{x}_0) = \{\mathbf{s} \in \mathbb{R}^n \mid f(\mathbf{x}_0 + \mathbf{w}) \geq f(\mathbf{x}_0) + \mathbf{s}^T \mathbf{w}, \forall \mathbf{w} \in \mathbb{R}^n\}.$$

$\partial f(\mathbf{x}_0)$ is a convex and closed set [52]. For our purposes, further assume that it is nonempty and bounded [62]. It, also, does not contain the origin since we assumed that \mathbf{x}_0 is not a minimizer. For any number $\lambda \geq 0$, we denote the scaled (by λ) subdifferential as $\lambda \partial f(\mathbf{x}_0) = \{\lambda \mathbf{s} \mid \mathbf{s} \in \partial f(\mathbf{x}_0)\}$.

4.4.2.2 Convex Cones

A convex cone $\mathcal{K} \subset \mathbb{R}^n$ is a set that is convex and closed under multiplication by positive scalars. The polar cone \mathcal{K}° is defined as the closed convex cone

$$\mathcal{K}^\circ := \{\mathbf{p} \in \mathbb{R}^n \mid \mathbf{p}^T \mathbf{s} \leq 0 \quad \text{for all } \mathbf{s} \in \mathcal{K}\}.$$

For the conic hull of a set $\mathcal{C} \subset \mathbb{R}^n$, we write

$$\text{cone}(\mathcal{C}) := \{\mathbf{s} \mid \mathbf{s} \in \lambda \mathcal{C}, \text{ for some } \lambda \geq 0\}.$$

In particular, we denote the *convex hull of the subdifferential* as $\text{cone}(\partial f(\mathbf{x}_0))$.

Next, we introduce the notion of the tangent cone $\mathcal{T}_f(\mathbf{x}_0)$ of f at \mathbf{x}_0 . First, define the set of descent directions of f at \mathbf{x}_0 as $\mathcal{F}_f(\mathbf{x}_0) = \{\mathbf{v} \mid f(\mathbf{x}_0 + \mathbf{v}) \leq f(\mathbf{x}_0)\}$. The tangent cone $\mathcal{T}_f(\mathbf{x}_0)$ of f at \mathbf{x}_0 is defined as

$$\mathcal{T}_f(\mathbf{x}_0) := \text{Cl}(\text{cone}(\mathcal{F}_f(\mathbf{x}_0))),$$

where $\text{Cl}(\cdot)$ denotes the closure of a set. Under the assumption that \mathbf{x}_0 is *not* a minimizer of $f(\cdot)$, then the polar of the tangent cone (commonly termed the normal cone [5]) can be equivalently written [52, p. 222] as the conic hull of the subdifferential.

$$(\mathcal{T}_f(\mathbf{x}_0))^\circ = \text{cone}(\partial f(\mathbf{x}_0)). \quad (4.28)$$

4.4.2.3 Gaussian Squared Distance

Let $\mathcal{C} \subset \mathbb{R}^n$ be a closed and nonempty convex set. For any vector $\mathbf{v} \in \mathbb{R}^n$, we denote its (unique) projection onto \mathcal{C} as $\text{Proj}(\mathbf{v}, \mathcal{C})$, i.e.

$$\text{Proj}(\mathbf{v}, \mathcal{C}) := \underset{\mathbf{s} \in \mathcal{C}}{\text{argmin}} \|\mathbf{v} - \mathbf{s}\|.$$

The distance of \mathbf{v} to the set \mathcal{C} can then be written as

$$\text{dist}(\mathbf{v}, \mathcal{C}) := \|\mathbf{v} - \text{Proj}(\mathbf{v}, \mathcal{C})\|.$$

Definition 4.1 (Gaussian squared distance). Let $\mathbf{h} \in \mathbb{R}^n$ have i.i.d. $\mathcal{N}(0, 1)$ entries. The Gaussian squared distance of a set $\mathcal{C} \subset \mathbb{R}^n$ is defined as $\mathbf{D}(\mathcal{C}) := \mathbb{E}_h [\text{dist}^2(\mathbf{h}, \mathcal{C})]$.

Accordingly, define the Gaussian projection and correlation as

$$\mathbf{P}(\mathcal{C}) := \mathbb{E} [\|\text{Proj}(\mathbf{h}, \mathcal{C})\|^2] \quad \text{and} \quad \mathbf{C}(\mathcal{C}) := \mathbb{E} [(\mathbf{h} - \text{Proj}(\mathbf{h}, \mathcal{C}))^T \text{Proj}(\mathbf{h}, \mathcal{C})].$$

Of particular interest to us is the *Gaussian squared distance to the scaled subdifferential* and *to the cone of subdifferential* (Fig. 4.1)

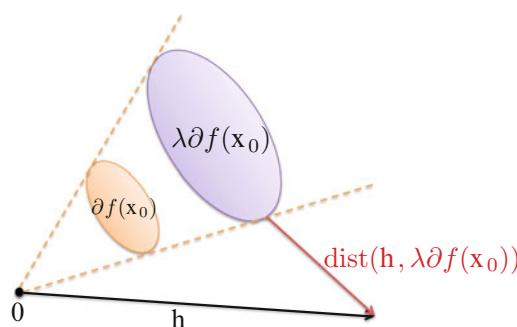


Fig. 4.1 Illustration of the distance of a vector to the scaled subdifferential $\lambda \partial f(\mathbf{x}_0)$.

$$\mathbf{D}(\lambda \partial f(\mathbf{x}_0)) = \mathbb{E} \left[\text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0)) \right],$$

$$\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) = \mathbb{E} \left[\text{dist}^2(\mathbf{h}, \text{cone}(\lambda \partial f(\mathbf{x}_0))) \right].$$

Following (4.28), observe that

$$\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) = \mathbf{D}((\mathcal{T}_f(\mathbf{x}_0))^\circ).$$

4.4.2.4 Gaussian Squared Distance to the Scaled Subdifferential

The Gaussian squared distance of the cone of subdifferential $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ appears as a fundamental quantity in the Noiseless CS problem. In particular, [14, 55] prove that $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ number of measurements suffice so that (4.14) uniquely recovers \mathbf{x}_0 . Later on, [1] and [57] showed⁷ that this number of measurements is also necessary. Hence, $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ precisely characterizes the phase-transition in the noiseless CS⁸.

The Gaussian squared distance of the scaled subdifferential $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ naturally arises in the analysis of the normalized mean-square error of the proximal denoising (4.15). Under the assumption of \mathbf{z} having i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, [43] shows that the normalized mean-square error of (4.15) admits a sharp upper bound (attained for $\sigma \rightarrow 0$) equal to $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$. For the constrained proximal denoiser in (4.16) the corresponding upper bound becomes $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$.

There is also a deep relation between the two quantities $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$, stronger than the obvious fact that

$$\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) \leq \min_{\lambda \geq 0} \mathbf{D}(\lambda \partial f(\mathbf{x}_0)).$$

In particular, it can be shown [1, 31, 43] that in high-dimensions

$$\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) \approx \min_{\lambda \geq 0} \mathbf{D}(\lambda \partial f(\mathbf{x}_0)). \quad (4.29)$$

Moreover, as the next lemma shows, the minimum of $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ in (4.29) is uniquely attained. The lemma also reveals an interesting relation between $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ and $\mathbf{C}(\lambda \partial f(\mathbf{x}_0))$. We will make use of this property in the analysis of the NSE of the ℓ_2 -LASSO in Section 4.6.

⁷The tools used in [1] and [57] differ. Amelunxen et al [1] use tools from conic integral geometry, while Stojnic [55] relies on a comparison lemma for Gaussian processes (see Lemma 3).

⁸In [14], $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ shows up indirectly via a closely related notion, that of the “Gaussian width” [34] of the restricted tangent cone $\mathcal{T}_f(\mathbf{x}_0) \cap \mathcal{S}^{n-1}$. In the terminology used in [1, 40], $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ corresponds to the “statistical dimension” of $\mathcal{T}_f(\mathbf{x}_0) = (\text{cone}(\partial f(\mathbf{x}_0)))^\circ$.

Lemma 1 ([1]). *Suppose $\partial f(\mathbf{x}_0)$ is nonempty and does not contain the origin. Then,*

1. $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ is a strictly convex function of $\lambda \geq 0$, and is differentiable for $\lambda > 0$.
2. $\frac{\partial \mathbf{D}(\lambda \partial f(\mathbf{x}_0))}{\partial \lambda} = -\frac{2}{\lambda} \mathbf{C}(\lambda \partial f(\mathbf{x}_0))$.

In the Appendix we show how to use the definitions and the properties just discussed to compute highly accurate and closed-form approximations for the quantities $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ and $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ for a number of practical convex regularizers and signal structures.

4.4.3 Probability Tools

4.4.3.1 Gordon's Lemma

Perhaps the most important technical ingredient underlying the analysis of the NSE of the generalized LASSO is the following lemma proved by Gordon in [34]. Gordon's Lemma establishes a very useful (probabilistic) inequality for Gaussian processes.

Lemma 2 (Gordon's Lemma [34]). *Let $\mathbf{G} \in \mathbb{R}^{m \times n}, g \in \mathbb{R}, \mathbf{g} \in \mathbb{R}^m, \mathbf{h} \in \mathbb{R}^n$ be independent of each other and have independent standard normal entries. Also, let $\mathcal{S} \subset \mathbb{R}^n$ be an arbitrary set and $\psi : \mathcal{S} \rightarrow \mathbb{R}$ be an arbitrary function. Then, for any $c \in \mathbb{R}$,*

$$\mathbb{P} \left(\min_{\mathbf{x} \in \mathcal{S}} \{ \|\mathbf{G}\mathbf{x}\| + \|\mathbf{x}\|g - \psi(\mathbf{x}) \} \geq c \right) \geq \mathbb{P} \left(\min_{\mathbf{x} \in \mathcal{S}} \{ \|\mathbf{x}\| \|\mathbf{g}\| - \mathbf{h}^T \mathbf{x} - \psi(\mathbf{x}) \} \geq c \right). \quad (4.30)$$

For the analysis of the LASSO, we use a slightly modified version of the original lemma (see [47, Lemma 5.1]).

Lemma 3 (Modified Gordon's Lemma). *$\mathbf{G} \in \mathbb{R}^{m \times n}, \mathbf{g} \in \mathbb{R}^m$ and $\mathbf{h} \in \mathbb{R}^n$ have i.i.d. $\mathcal{N}(0, 1)$ entries and be independent of each other. Also, let $\psi(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$, and, $\Phi_1 \subset \mathbb{R}^n$ and $\Phi_2 \subset \mathbb{R}^m$ such that either both Φ_1 and Φ_2 are compact or Φ_1 is arbitrary and Φ_2 is a scaled unit sphere. Then, for any $c \in \mathbb{R}$:*

$$\begin{aligned} & \mathbb{P} \left(\min_{\mathbf{x} \in \Phi_1} \max_{\mathbf{a} \in \Phi_2} \{ \mathbf{a}^T \mathbf{G}\mathbf{x} - \psi(\mathbf{x}, \mathbf{a}) \} \geq c \right) \geq \\ & 2\mathbb{P} \left(\min_{\mathbf{x} \in \Phi_1} \max_{\mathbf{a} \in \Phi_2} \{ \|\mathbf{x}\| \|\mathbf{g}^T \mathbf{a}\| - \|\mathbf{a}\| \|\mathbf{h}^T \mathbf{x}\| - \psi(\mathbf{x}, \mathbf{a}) \} \geq c \right) - 1. \end{aligned}$$

It is worth mentioning that the “escape through a mesh” lemma, which has been the backbone of the approach introduced by Stojnic [55] (and subsequently refined

in [14]) for computing an asymptotic upper bound to the minimum number of measurements required in the Noiseless CS problem, is a corollary of Lemma 2. The “escape through a mesh” lemma gives a bound on the “restricted minimum singular value” of an operator \mathbf{G} , which is defined as the minimum gain of \mathbf{G} restricted to a subset of the unit sphere. The concept is similar to the restricted isometry property and has been topic of several related works, [6, 14, 42, 49].

Let $\mathcal{C} \subset \mathbb{R}^n$ be a convex subset of the unit ℓ_2 -sphere \mathcal{S}^{n-1} . Then, the minimum singular value of $\mathbf{A} \in \mathbb{R}^{m \times n}$ restricted to \mathcal{C} is defined as

$$\sigma_{\min}(\mathbf{A}, \mathcal{C}) = \min_{\mathbf{v} \in \mathcal{C}} \|\mathbf{Av}\|.$$

Observe that, $\sigma_{\min}(\mathbf{A}, \mathcal{S}^{n-1})$ reduces to the standard definition of the minimum singular value of the matrix \mathbf{A} .

Lemma 4 (Restricted eigenvalue or Escape through a mesh). *Let $\mathbf{G} \in \mathbb{R}^{m \times n}$ have i.i.d. standard normal entries and $\mathcal{K} \subset \mathbb{R}^n$ be a convex cone. Assume $0 < t \leq \gamma_m - \sqrt{\mathbf{D}(\mathcal{K}^\circ)}$. Then,*

$$\mathbb{P} \left(\min_{\mathbf{v} \in \mathcal{K} \cap \mathcal{S}^{n-1}} \|\mathbf{Gv}\| \geq \gamma_m - \sqrt{\mathbf{D}(\mathcal{K}^\circ)} - t \right) \geq 1 - \exp(-\frac{t^2}{2}).$$

4.4.3.2 Gaussian Concentration of Lipschitz Functions

The upper bounds on the NSE of the LASSO, similar to the OLS case, are probabilistic in nature. In particular they hold with high probability over the realizations of \mathbf{A} (and occasionally of \mathbf{z}). Key to the derivation of such probabilistic statements is the concentration of measure phenomenon. The lemma below is a well-known and powerful result on the concentration of Lipschitz functions under the Gaussian measure. Recall that a function $\psi(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}$ is L-Lipschitz, if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $|\psi(\mathbf{x}) - \psi(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|$.

Lemma 5 (Lipschitz concentration, [37]). *Let $\mathbf{h} \in \mathbb{R}^n$ have i.i.d. $\mathcal{N}(0, 1)$ entries and $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$ be an L-Lipschitz function. Then, $\text{Var}[\psi(\mathbf{h})] \leq L^2$. Furthermore, for all $t > 0$, the events $\{\psi(\mathbf{h}) - \mathbb{E}[\psi(\mathbf{h})] \geq t\}$ and $\{\psi(\mathbf{h}) - \mathbb{E}[\psi(\mathbf{h})] \leq -t\}$ hold with probability no greater than $\exp(-t^2/(2L^2))$, each.*

4.5 The NSE of Generalized LASSO in Gaussian Noise

In this section we assume that the noise vector \mathbf{z} has entries distributed i.i.d. normal $\mathcal{N}(0, \sigma^2)$ and derive sharp upper bounds on the NSE of the generalized LASSO. We begin in Section 4.5.1 with describing the main steps of the required technical analysis. That is only a highlight of the key ideas and insights and we provide

specific references in [47] for the details⁹. Next, Sections 4.5.2, 4.5.3, and 4.5.4 are each devoted to upper-bounding the NSE of the C-LASSO, ℓ_2 -LASSO, and ℓ_2^2 -LASSO, respectively.

4.5.1 Technical Framework

Proving the sharp upper bounds for the NSE of the generalized LASSO requires several steps and can be challenging to include all the details in the limited space of a book chapter. However, the main strategy for the proof and the fundamental tools and ideas being used are neat and not hard to appreciate. This will be the purpose of this section. The framework to be presented was first introduced by Stojnic in [56] and subsequently simplified and extended in [47].

The bulk of the results to be presented in this section are for the C-LASSO and the ℓ_2 -LASSO. Based on this and a mapping between the ℓ_2 -LASSO and the ℓ_2^2 -LASSO, some conclusions can be drawn for the NSE of the ℓ_2^2 -LASSO, as well. For the purposes of exposition we use the ℓ_2 -LASSO. The analysis for the constrained version C-LASSO is to a large extent similar. In fact, [47] treats those two under a common framework.

4.5.1.1 First-Order Approximation

Recall the ℓ_2 -LASSO problem introduced in (4.18):

$$\hat{\mathbf{x}}_{\ell_2} = \arg \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{Ax}\| + \frac{\lambda}{\sqrt{m}} f(\mathbf{x}). \quad (4.31)$$

A key idea behind our approach is using the linearization of the convex structure inducing function $f(\cdot)$ around the vector of interest \mathbf{x}_0 [7, 52]. From convexity of f , for all $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{s} \in \partial f(\mathbf{x}_0)$, we have $f(\mathbf{x}) \geq f(\mathbf{x}_0) + \mathbf{s}^T(\mathbf{x} - \mathbf{x}_0)$. In particular,

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \sup_{\mathbf{s} \in \partial f(\mathbf{x}_0)} \mathbf{s}^T(\mathbf{x} - \mathbf{x}_0) =: \hat{f}(\mathbf{x}), \quad (4.32)$$

and approximate equality holds when $\|\mathbf{x} - \mathbf{x}_0\|$ is ‘‘small.’’ Recall that $\partial f(\mathbf{x}_0)$ denotes the subdifferential of $f(\cdot)$ at \mathbf{x}_0 and is always a compact and convex set [52]. We also assume that \mathbf{x}_0 is not a minimizer of $f(\cdot)$, hence, $\partial f(\mathbf{x}_0)$ does not contain the origin.

⁹When referring to [47] keep in mind the following: a) in [47] the entries of \mathbf{A} have variance 1 and not $1/m$ as here, b) [47] uses slightly different notation for $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ and $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ ($\mathbf{D}_f(\mathbf{x}_0, \lambda)$ and $\mathbf{D}_f(\mathbf{x}_0, \mathbf{R}^+)$, respectively).

We substitute $f(\cdot)$ in (4.31) by its first-order approximation $\hat{f}(\cdot)$, to get a corresponding “*Approximated LASSO*” problem. To write the approximated problem in an easy-to-work-with format, recall that $\mathbf{y} = \mathbf{Ax}_0 + \mathbf{z} = \mathbf{Ax}_0 + \sigma \mathbf{v}$, for $\mathbf{v} \sim \mathcal{N}(0, \mathbf{I}_m)$ and change the optimization variable from \mathbf{x} to $\mathbf{w} = \mathbf{x} - \mathbf{x}_0$:

$$\tilde{\mathbf{w}}_{\ell_2} = \arg \min_{\mathbf{w}} \left\{ \|\mathbf{Aw} - \sigma \mathbf{v}\| + \frac{1}{\sqrt{m}} \sup_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} \mathbf{s}^T \mathbf{w} \right\}. \quad (4.33)$$

We denote $\tilde{\mathbf{w}}_{\ell_2}$ the optimal solution of the approximated problem in (4.33) and $\hat{\mathbf{w}}_{\ell_2} = \hat{\mathbf{x}}_{\ell_2} - \mathbf{x}_0$ for the optimal solution of the original problem in (4.31)¹⁰. Also, denote the optimal cost achieved in (4.33) by $\tilde{\mathbf{w}}_{\ell_2}$, as $\tilde{\mathcal{F}}_{\ell_2}(\mathbf{A}, \mathbf{v})$. Finally, note that the approximated problem corresponding to C-LASSO can be written as in (4.33), with only $\lambda \partial f(\mathbf{x}_0)$ being substituted by $\text{cone}(\partial f(\mathbf{x}_0))$.

Taking advantage of the simple characterization of $\hat{f}(\cdot)$ via the subdifferential $\partial f(\mathbf{x}_0)$, we are able to *precisely* analyze the optimal cost and the normalized squared error of the resulting approximated problem. The approximation is tight when $\|\hat{\mathbf{x}}_{\ell_2} - \mathbf{x}_0\| \rightarrow 0$ and we later show that this is the case when the noise level $\sigma \rightarrow 0$. This fact allows us to translate the results obtained for the Approximated LASSO problem to corresponding *precise* results for the Original LASSO problem, in the *small noise variance regime*.

4.5.1.2 Importance of $\sigma \rightarrow 0$

Our focus is on the precise characterization of the NSE. While we show that the first order characteristics of the function, i.e. $\partial f(\mathbf{x}_0)$, suffice to provide sharp and closed-form bounds for small noise level σ , we believe that higher order terms are required for such precise results when σ is arbitrary. On the other hand, empirical observations suggest that the worst case NSE for the LASSO problem is achieved when $\sigma \rightarrow 0$. This statement admits a rigorous proof in the case of the C-LASSO. Interestingly, the same phenomena has been observed and proved to be true for related estimation problems. Examples include the proximal denoising problem (4.15) in [27, 28, 43] and, the LASSO problem with ℓ_1 penalization [2].

To recap, in what follows we derive formulae that sharply characterize the NSE of the generalized LASSO in the *small σ* regime. For the C-LASSO Section 10 in [47] proves that the corresponding such formula upper bounds the NSE when σ is arbitrary. Extended numerical simulations suggest that this is also the case for the regularized LASSO.

¹⁰We follow this convention throughout: use the symbol “~” over variables that are associated with the approximated problems. To distinguish, use the symbol “^” for the variables associated with the original problem .

4.5.1.3 Applying Gordon's Lemma

The (approximated) LASSO problem in (4.33) is simpler than the original one in (4.44), yet, still hard to directly analyze. We will apply Gordon's Lemma 3 to further simplify the problem. This trick is critical and repeatedly used in our analysis. First, write $\sqrt{m}\|\mathbf{Aw} - \sigma\mathbf{v}\| = \max_{\|\mathbf{a}\|=1} \mathbf{a}^T [\sqrt{m}\mathbf{A}, -\mathbf{v}] \begin{bmatrix} \mathbf{w} \\ \sigma\sqrt{m} \end{bmatrix}$ and choose the function $\psi(\cdot)$ in Lemma 3 to be $\sup_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} \mathbf{s}^T \mathbf{w}$. Then, the following result is a simple corollary of applying Gordon's Lemma to the LASSO objective. Recall that we write $\mathcal{F}_{\ell_2}(\mathbf{A}, \mathbf{v})$ for the optimal objective value of the approximated LASSO in (4.33).

Corollary 5.1 (Lower Key Optimization). *Let $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_m)$ and $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$ be independent of each other. Define the following optimization problem:*

$$\mathcal{L}(\mathbf{g}, \mathbf{h}) = \min_{\mathbf{w}} \left\{ \sqrt{\|\mathbf{w}\|_2^2 + m\sigma^2} \|\mathbf{g}\| - \mathbf{h}^T \mathbf{w} + \max_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} \mathbf{s}^T \mathbf{w} \right\}. \quad (4.34)$$

Then, for any $c \in \mathbb{R}$:

$$\mathbb{P}(\mathcal{F}_{\ell_2}(\mathbf{A}, \mathbf{v}) \geq c) \geq 2\mathbb{P}(\mathcal{L}(\mathbf{g}, \mathbf{h}) \geq c\sqrt{m}) - 1.$$

Corollary 5.1 establishes a probabilistic connection between the LASSO problem and the minimization (4.34). The advantage of the latter is that it is much easier to analyze. Intuitively, the main reason for that is that instead of an $m \times n$ matrix, (4.34) only involves two vectors of sizes $m \times 1$ and $n \times 1$. Even more, those vectors have independent standard normal entries and are independent of each other, which greatly facilitates probabilistic statements about the value of $\mathcal{L}(\mathbf{g}, \mathbf{h})$. Due to its central role in our analysis, we often refer to problem (4.34) as “key optimization” or “lower key optimization.” The term “lower” is attributed to the fact that analysis of (4.34) results in a probabilistic lower bound for the optimal cost of the LASSO problem.

4.5.1.4 Analyzing the Key Optimization

Deterministic Analysis: First, we perform the deterministic analysis of $\mathcal{L}(\mathbf{g}, \mathbf{h})$ for fixed $\mathbf{g} \in \mathbb{R}^m$ and $\mathbf{h} \in \mathbb{R}^n$. In particular, we reduce the optimization in (4.34) to a *scalar* optimization. To see this, perform the optimization over a fixed ℓ_2 -norm of \mathbf{w} to equivalently write

$$\mathcal{L}(\mathbf{g}, \mathbf{h}) = \min_{\alpha \geq 0} \left\{ \sqrt{\alpha^2 + m\sigma^2} \|\mathbf{g}\| - \max_{\|\mathbf{w}\|=\alpha} \min_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} (\mathbf{h} - \mathbf{s})^T \mathbf{w} \right\}.$$

The maximin problem that appears in the objective function of the optimization above has a simple solution. It can be shown [47, Lemma E.1] that if $\mathbf{h} \notin \lambda \partial f(\mathbf{x}_0)$:

$$\begin{aligned} \max_{\|\mathbf{w}\|=\alpha} \min_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} (\mathbf{h} - \mathbf{s})^T \mathbf{w} &= \min_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} \max_{\|\mathbf{w}\|=\alpha} (\mathbf{h} - \mathbf{s})^T \mathbf{w} \\ &= \alpha \min_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} \|\mathbf{h} - \mathbf{s}\|. \end{aligned} \quad (4.35)$$

This reduces (4.34) to a scalar optimization problem over α , for which one can compute the optimal value $\hat{\alpha}$ and the corresponding optimal cost. The result is summarized in Lemma 6 below.

Lemma 6 (Deterministic Result). *Let $\tilde{\mathbf{w}}(\mathbf{g}, \mathbf{h})$ be a minimizer of the problem in (4.34). If $\|\mathbf{g}\| > \text{dist}(\mathbf{h}, \lambda \partial f(\mathbf{x}_0)) > 0$, then*

$$\begin{aligned} a) \quad \|\tilde{\mathbf{w}}(\mathbf{g}, \mathbf{h})\|^2 &= m\sigma^2 \frac{\text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))}{\|\mathbf{g}\|^2 - \text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))}, \\ b) \quad \mathcal{L}(\mathbf{g}, \mathbf{h}) &= \sqrt{m\sigma^2 \sqrt{\|\mathbf{g}\|^2 - \text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))}}. \end{aligned}$$

Probabilistic Analysis: Of interest is making probabilistic statements about $\mathcal{L}(\mathbf{g}, \mathbf{h})$ and the norm of its minimizer $\|\tilde{\mathbf{w}}(\mathbf{g}, \mathbf{h})\|$. Lemma 6 provided closed-form deterministic solutions for both of them, which only involve the quantities $\|\mathbf{g}\|^2$ and $\text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))$. The ℓ_2 -norm and the distance function to a convex set are 1-Lipschitz functions. Application of Lemma 5 then shows that $\|\mathbf{g}\|^2$ and $\text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))$ concentrate nicely around their means $\mathbb{E}[\|\mathbf{g}\|^2] = m$ and $\mathbb{E}[\text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))] = \mathbf{D}(\lambda \partial f(\mathbf{x}_0))$, respectively. Combining this with Lemma 6, we conclude with Lemma 7 below.

Lemma 7 (Probabilistic Result). *Assume that $(1 - \varepsilon_L)m \geq \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) \geq \varepsilon_L m$ for some constant $\varepsilon_L > 0$. Define¹¹,*

$$\eta = \sqrt{m - \mathbf{D}(\lambda \partial f(\mathbf{x}_0))} \quad \text{and} \quad \gamma = \frac{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))}{m - \mathbf{D}(\lambda \partial f(\mathbf{x}_0))}.$$

Then, for any $\varepsilon > 0$, there exists a constant $c > 0$ such that, for sufficiently large m , with probability $1 - \exp(-cm)$,

$$|\mathcal{L}(\mathbf{g}, \mathbf{h}) - \sqrt{m\sigma}\eta| \leq \varepsilon\sqrt{m\sigma}\eta, \quad \text{and} \quad \left| \frac{\|\tilde{\mathbf{w}}(\mathbf{g}, \mathbf{h})\|^2}{m\sigma^2} - \gamma \right| \leq \varepsilon\gamma.$$

¹¹Observe that the dependence of η and γ on λ , m and $\partial f(\mathbf{x}_0)$, is implicit in this definition.

Remark. In Lemma 7, the condition “ $(1 - \varepsilon_L)m \geq \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) \geq \varepsilon_L m$ ” ensures that $\|\mathbf{g}\| > \text{dist}(\mathbf{h}, \lambda \partial f(\mathbf{x}_0)) > 0$ (cf. Lemma 6) with high probability over the realizations of \mathbf{g} and \mathbf{h} .

4.5.1.5 From the Key Optimization Back to the LASSO

Before proceeding, let us recap. Application of Gordon’s Lemma to the approximated LASSO problem in (4.33) introduced the simpler lower key optimization (4.34). Without much effort, we found in Lemma 7 that its cost $\mathcal{L}(\mathbf{g}, \mathbf{h})$ and the normalized squared norm of its minimizer $\frac{\|\tilde{\mathbf{w}}(\mathbf{g}, \mathbf{h})\|^2}{m\sigma^2}$ concentrate around $\sqrt{m}\sigma\eta$ and γ , respectively. This brings us to the following question:

- *To what extent do such results on $\mathcal{L}(\mathbf{g}, \mathbf{h})$ and $\tilde{\mathbf{w}}(\mathbf{g}, \mathbf{h})$ translate to useful conclusions about $\mathcal{F}_{\ell_2}(\mathbf{A}, \mathbf{v})$ and $\tilde{\mathbf{w}}_{\ell_2}(\mathbf{A}, \mathbf{v})$?*

Application of Gordon’s Lemma as performed in Corollary 5.1 when combined with Lemma 7 provides a first answer to this question: $\mathcal{F}_{\ell_2}(\mathbf{A}, \mathbf{v})$ is lower bounded by $\sigma\eta$ with overwhelming probability. Formally,

Lemma 8 (Lower Bound). *Assume $(1 - \varepsilon_L)m \geq \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) \geq \varepsilon_L m$ for some constant $\varepsilon_L > 0$ and m is sufficiently large. Then, for any $\varepsilon > 0$, there exists a constant $c > 0$ such that, with probability $1 - \exp(-cm)$,*

$$\mathcal{F}_{\ell_2}(\mathbf{A}, \mathbf{v}) \geq (1 - \varepsilon)\sigma\eta.$$

But is that all? It turns out that the connection between the LASSO problem and the simple optimization (4.34) is much *deeper* than Lemma 8 predicts. In short, under certain conditions on λ and m (similar in nature to those involved in the assumption of Lemma 8), we can prove the following being true:

- Similar to $\mathcal{L}(\mathbf{g}, \mathbf{h})$, the optimal cost $\mathcal{F}_{\ell_2}(\mathbf{A}, \mathbf{v})$ of the approximated ℓ_2 -LASSO concentrates around $\sigma\eta$.
- Similar to $\frac{\|\tilde{\mathbf{w}}(\mathbf{g}, \mathbf{h})\|^2}{m\sigma^2}$, the NSE of the approximated ℓ_2 -LASSO $\frac{\|\tilde{\mathbf{w}}_{\ell_2}(\mathbf{A}, \mathbf{v})\|^2}{m\sigma^2}$ concentrates around γ .

In some sense, $\mathcal{L}(\mathbf{g}, \mathbf{h})$ “predicts” $\mathcal{F}_{\ell_2}(\mathbf{A}, \mathbf{v})$ and $\|\tilde{\mathbf{w}}(\mathbf{g}, \mathbf{h})\|$ “predicts” $\|\tilde{\mathbf{w}}_{\ell_2}(\mathbf{A}, \mathbf{v})\|$. The main effort in [47, 56] is in proving those claims. The next section contains a synopsis of the main steps of the proof.

4.5.1.6 Synopsis of the Technical Framework

1. Apply Gordon’s Lemma to the *dual* of $\mathcal{F}_{\ell_2}(\mathbf{A}, \mathbf{v})$ to compute a *high-probability upper bound* for it. Following essentially the same ideas as in Section 4.5.1.4 it is shown that $\mathcal{F}_{\ell_2}(\mathbf{A}, \mathbf{v})$ is no larger than $\sigma\eta$ with high probability (see Lemmas 5.2 and 6.2 in [47]).

2. Assume $\|\tilde{\mathbf{w}}_{\ell_2}\|^2/(m\sigma^2)$ deviates from γ . Yet another application of Gordon's Lemma shows that such a deviation would result in a *significant increase* in the optimal cost, namely $\mathcal{F}_{\ell_2}(\mathbf{A}, \mathbf{v})$ would be significantly larger than $\sigma\eta$ (see Lemmas 5.2 and 6.3 in [47]).
3. Combining the previous steps shows that $\|\tilde{\mathbf{w}}_{\ell_2}\|^2/(m\sigma^2)$ concentrates with high probability around γ (see Lemma 6.4 in [47]).
4. The final step requires us to translate this bound on the NSE of the Approximated LASSO to a bound on the NSE of the original one. We choose σ small enough such that $\|\tilde{\mathbf{w}}_{\ell_2}\|$ is small and so $f(\mathbf{x}_0 + \tilde{\mathbf{w}}_{\ell_2}) \approx \hat{f}(\mathbf{x}_0 + \tilde{\mathbf{w}}_{\ell_2})$. Using this and combining the results of Steps 2 and 3 we show that $\|\hat{\mathbf{w}}_{\ell_2}\|^2/(m\sigma^2)$ concentrates with high probability around γ (see Section 9.1.2 in [47]).

4.5.2 C-LASSO

4.5.2.1 NSE

Theorem 1 ([47]). *Assume there exists a constant $\varepsilon_L > 0$ such that $(1 - \varepsilon_L)m \geq \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) \geq \varepsilon_L m$ and m is sufficiently large. For any $\varepsilon > 0$, there exists a constant $C = C(\varepsilon, \varepsilon_L)$ such that, with probability $1 - \exp(-Cm)$,*

$$\frac{\|\hat{\mathbf{x}}_c - \mathbf{x}_0\|^2}{m\sigma^2} \leq (1 + \varepsilon) \frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}, \quad (4.36)$$

Furthermore, there exists a deterministic number $\sigma_0 > 0$ (i.e., independent of \mathbf{A}, \mathbf{v}) such that, if $\sigma \leq \sigma_0$, with the same probability,

$$\left| \frac{\|\hat{\mathbf{x}}_c - \mathbf{x}_0\|^2}{m\sigma^2} \times \frac{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} - 1 \right| < \varepsilon. \quad (4.37)$$

Observe in Theorem 1 that as m approaches $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$, the NSE increases and when $m = \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$, $\text{NSE} = \infty$. This behavior is not surprising as when $m < \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$, one cannot even recover \mathbf{x}_0 from noiseless observations via (4.14), hence it is futile to expect noise robustness.

Example (sparse signals): Figure 4.2 illustrates Theorem 1 when \mathbf{x}_0 is a k -sparse vector and $f(\cdot)$ is the ℓ_1 norm. In this case, $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ is only a function of k and n and can be exactly calculated, [21] (also see Section ‘‘Sparse signals’’). The dark-blue region corresponds to the unstable region $m < \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$. The dashed gray line obeys $m = 1.4 \times \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and yields a constant (worst-case) NSE of 2.5 as sparsity varies. We note that for ℓ_1 minimization, the NSE formula was first proposed by Donoho et al. in [26].

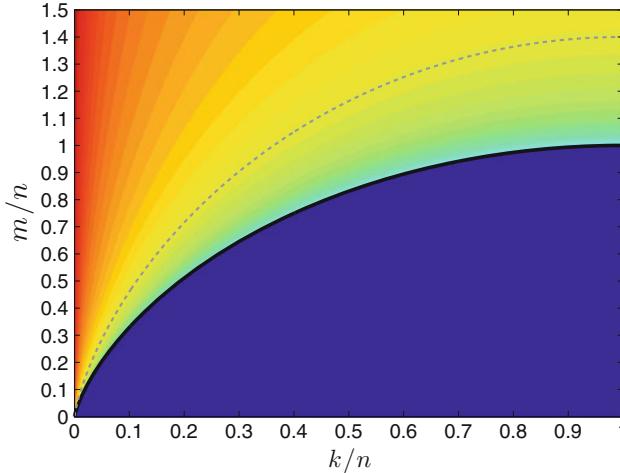


Fig. 4.2 NSE heatmap for ℓ_1 minimization based on Theorem 1. The x and y axes are the sparsity and measurements normalized by the ambient dimension. To obtain the figure, we plotted the heatmap of the function $-\log \frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}$ (clipped to ensure the values are between $[-10, 5]$).

4.5.2.2 Relation to Proximal Denoising

It is interesting to compare the NSE of the C-LASSO to the MSE risk of the constrained proximal denoiser in (4.16). Recall from Section 4.4.2.3 that the normalized MSE of (4.16) is upper bounded by $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ [13, 43]. Furthermore, this bound is attained asymptotically as $\sigma \rightarrow 0$. From Theorem 1 we find that the corresponding quantity $\|\hat{\mathbf{x}}_c - \mathbf{x}_0\|^2/\sigma^2$ is upper bounded by

$$\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) \frac{m}{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))},$$

and is again attained asymptotically as $\sigma \rightarrow 0$. We conclude that the NSE of the LASSO problem is amplified compared to the corresponding quantity of proximal denoising by a factor of $\frac{m}{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} > 1$. This factor can be interpreted as the penalty paid in the estimation error for observing noisy linear measurements of the unknown signal instead of just noisy measurements of the signal itself.

4.5.3 ℓ_2 -LASSO

Characterization of the NSE of the ℓ_2 -LASSO is more involved than that of the NSE of the C-LASSO. For this problem, choice of λ naturally plays a critical role. We state the main result in Section 4.5.3.1 and discuss it in detail in Section 4.5.3.2–4.5.3.4.

4.5.3.1 NSE

Definition 5.1 (\mathcal{R}_{ON}). Suppose $m > \min_{\lambda \geq 0} \mathbf{D}(\lambda \partial f(\mathbf{x}_0))$. Define \mathcal{R}_{ON} as follows,

$$\mathcal{R}_{\text{ON}} = \{\lambda > 0 \mid m - \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) > \max\{0, \mathbf{C}(\lambda \partial f(\mathbf{x}_0))\}\}.$$

Theorem 2 ([47]). Assume there exists a constant $\varepsilon_L > 0$ such that $(1 - \varepsilon_L)m \geq \max\{\mathbf{D}(\lambda \partial f(\mathbf{x}_0)), \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) + \mathbf{C}(\lambda \partial f(\mathbf{x}_0))\}$ and $\mathbf{D}(\lambda \partial f(\mathbf{x}_0)) \geq \varepsilon_L m$. Further, assume that m is sufficiently large. Then, for any $\varepsilon > 0$, there exist a constant $C = C(\varepsilon, \varepsilon_L)$ and a deterministic number $\sigma_0 > 0$ (i.e., independent of \mathbf{A}, \mathbf{v}) such that, whenever $\sigma \leq \sigma_0$, with probability $1 - \exp(-C \min\{m, \frac{m^2}{n}\})$,

$$\left| \frac{\|\hat{\mathbf{x}}_{\ell_2} - \mathbf{x}_0\|^2}{m\sigma^2} \times \frac{m - \mathbf{D}(\lambda \partial f(\mathbf{x}_0))}{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))} - 1 \right| < \varepsilon.$$

4.5.3.2 Regions Of Operation

First, we identify the regime in which the ℓ_2 -LASSO can robustly recover \mathbf{x}_0 . In this direction, the number of measurements should be large enough to guarantee at least noiseless recovery in (4.14), which is the case when $m > \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ [1, 14]. To translate this requirement in terms of $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$, recall (4.29) and Lemma 1, and define λ_{best} to be the *unique* minimizer of $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ over $\lambda \in \mathbb{R}^+$. We then write the regime of interest as $m > \mathbf{D}(\lambda_{\text{best}} \cdot \partial f(\mathbf{x}_0)) \approx \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$.

Next, we identify three important values of the penalty parameter λ , needed to describe the distinct regions of operation of the estimator.

1. λ_{best} : λ_{best} is optimal in the sense that the NSE is minimized for this particular choice of the penalty parameter (see Section 4.5.3.4). This also explains the term “best” we associate with it.
2. λ_{max} : Over $\lambda \geq \lambda_{\text{best}}$, the equation $m = \mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ has a unique solution. We denote this solution by λ_{max} . For values of λ larger than λ_{max} , we have $m \leq \mathbf{D}(\lambda \partial f(\mathbf{x}_0))$.
3. λ_{crit} : Over $0 \leq \lambda \leq \lambda_{\text{best}}$, if $m \leq n$, the equation $m - \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) = \mathbf{C}(\lambda \partial f(\mathbf{x}_0))$ has a unique solution which we denote λ_{crit} . Otherwise, it has no solution and $\lambda_{\text{crit}} := 0$.

Based on the above definitions, we recognize the three distinct regions of operation of the ℓ_2 -LASSO, as follows (Figs. 4.3 and 4.4),

1. $\mathcal{R}_{\text{ON}} = \{\lambda \in \mathbb{R}^+ \mid \lambda_{\text{crit}} < \lambda < \lambda_{\text{max}}\}$.
2. $\mathcal{R}_{\text{OFF}} = \{\lambda \in \mathbb{R}^+ \mid \lambda \leq \lambda_{\text{crit}}\}$.
3. $\mathcal{R}_{\infty} = \{\lambda \in \mathbb{R}^+ \mid \lambda \geq \lambda_{\text{max}}\}$.

See Figure 4.5 for an illustration of the definitions above and Section 8 in [47] for the detailed proofs of the statements.

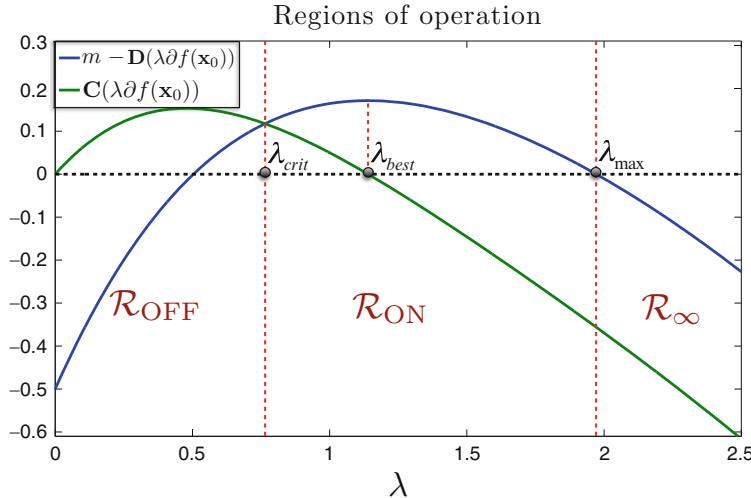


Fig. 4.3 Regions of operation of the ℓ_2 -LASSO.

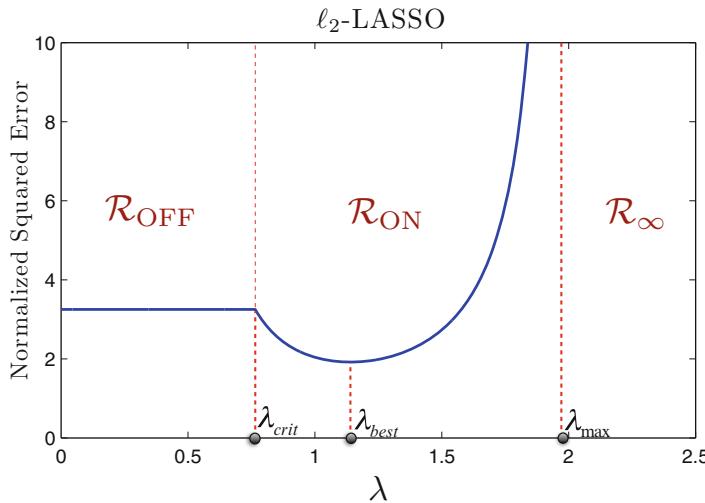


Fig. 4.4 We consider the ℓ_1 -penalized ℓ_2 -LASSO problem for a k sparse signal in \mathbb{R}^n . For $\frac{k}{n} = 0.1$ and $\frac{m}{n} = 0.5$, we have $\lambda_{\text{crit}} \approx 0.76$, $\lambda_{\text{best}} \approx 1.14$, $\lambda_{\text{max}} \approx 1.97$.

4.5.3.3 Characterizing the NSE in each Region

Theorem 2 upper bounds the NSE of the ℓ_2 -LASSO in \mathcal{R}_{ON} . Here, we also briefly discuss on some observations that can be made regarding \mathcal{R}_{OFF} and \mathcal{R}_∞ :

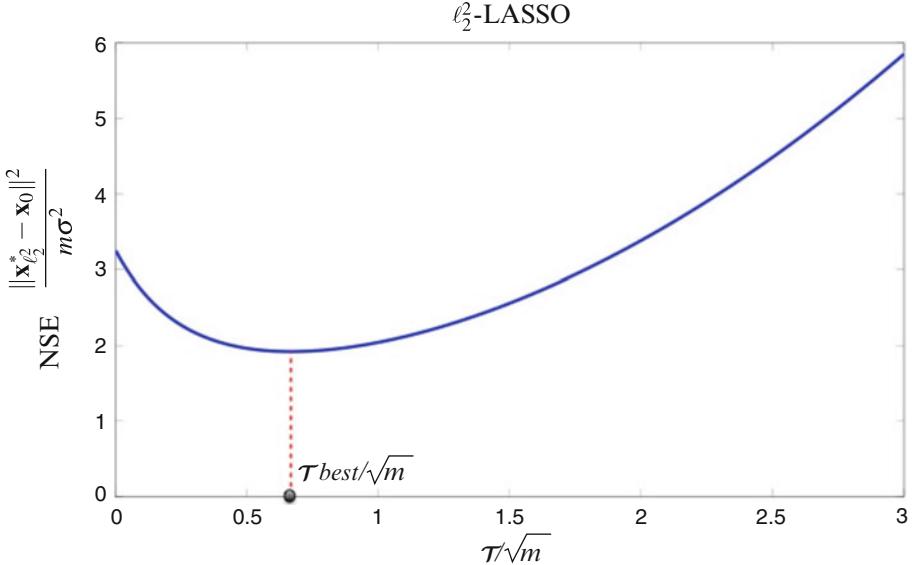


Fig. 4.5 Consider the exact same setup of Figure 4.4. We plot the NSE of the ℓ_2 -LASSO as a function of the regularizer parameter.

- \mathcal{R}_{OFF} : For $\lambda \in \mathcal{R}_{\text{OFF}}$, extended empirical observations suggest that the LASSO estimate $\hat{\mathbf{x}}_{\ell_2}$ satisfies $\mathbf{y} = \mathbf{A}\hat{\mathbf{x}}_{\ell_2}$ and the optimization (4.18) reduces to the standard ℓ_1 minimization (4.14) of noiseless CS. Indeed, Lemma 9.2 in [47] proves that this reduction is indeed true for sufficiently small values of λ . Proving the validity of the claim for the entire region \mathcal{R}_{OFF} would show that when $\sigma \rightarrow 0$, the NSE is $\mathbf{D}(\lambda_{\text{crit}} \cdot \partial f(\mathbf{x}_0)) / (m - \mathbf{D}(\lambda_{\text{crit}} \cdot \partial f(\mathbf{x}_0)))$, for all $\lambda \in \mathcal{R}_{\text{OFF}}$.
- \mathcal{R}_{ON} : Begin with observing that \mathcal{R}_{ON} is a nonempty and open interval. In particular, $\lambda_{\text{best}} \in \mathcal{R}_{\text{ON}}$ since $m > \mathbf{D}(\lambda_{\text{best}} \cdot \partial f(\mathbf{x}_0))$. Theorem 2 proves that for all $\lambda \in \mathcal{R}_{\text{ON}}$ and for σ sufficiently small,

$$\frac{\|\hat{\mathbf{x}}_{\ell_2} - \mathbf{x}_0\|}{m\sigma^2} \approx \frac{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))}{m - \mathbf{D}(\lambda \partial f(\mathbf{x}_0))}. \quad (4.38)$$

Also, empirical observations suggest that (4.38) holds for arbitrary σ when \approx is replaced with \lesssim . Finally, we should note that the NSE formula $\mathbf{D}(\lambda \partial f(\mathbf{x}_0)) / (m - \mathbf{D}(\lambda \partial f(\mathbf{x}_0)))$ is a convex function of λ over \mathcal{R}_{ON} .

- \mathcal{R}_{∞} : Empirically, we observe that the stable recovery of \mathbf{x}_0 is not possible for $\lambda \in \mathcal{R}_{\infty}$.

4.5.3.4 Optimal Tuning of the Penalty Parameter

It is not hard to see that the formula in (4.38) is strictly increasing in $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$. Thus, when $\sigma \rightarrow 0$, the NSE achieves its minimum value when the penalty parameter is set to λ_{best} . Recall from (4.29) that $\mathbf{D}(\lambda_{\text{best}} \cdot \partial f(\mathbf{x}_0)) \approx \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and compare the formulae in Theorems 1 and 2, to conclude that the C-LASSO and ℓ_2 -LASSO can be related by choosing $\lambda = \lambda_{\text{best}}$. In particular, we have

$$\frac{\|\hat{\mathbf{x}}_{\ell_2}(\lambda_{\text{best}}) - \mathbf{x}_0\|^2}{m\sigma^2} \approx \frac{\mathbf{D}(\lambda_{\text{best}} \cdot \partial f(\mathbf{x}_0))}{m - \mathbf{D}(\lambda_{\text{best}} \cdot \partial f(\mathbf{x}_0))} \approx \frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} \approx \frac{\|\hat{\mathbf{x}}_c - \mathbf{x}_0\|^2}{m\sigma^2}.$$

It is important to note that deriving λ_{best} does not require knowledge of any properties (e.g., variance) of the noise vector neither does it require knowledge of the unknown signal \mathbf{x}_0 itself. All it requires is knowledge of the particular structure of the unknown signal. For example, in the ℓ_1 -case, λ_{best} depends only on the sparsity of \mathbf{x}_0 , not \mathbf{x}_0 itself, and in the nuclear norm case, it only depends on the rank of \mathbf{x}_0 , not \mathbf{x}_0 itself.

4.5.4 ℓ_2^2 -LASSO

4.5.4.1 Connection to ℓ_2 -LASSO

We propose a mapping between the penalty parameters λ of the ℓ_2 -LASSO program (4.18) and τ of the ℓ_2^2 -LASSO program (4.19), for which the NSE of the two problems behaves the same. The mapping function is defined as follows.

Definition 5.2 (Mapping Function). For any $\lambda \in \mathcal{R}_{\text{ON}}$, define

$$\text{map}(\lambda) = \lambda \frac{m - \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) - \mathbf{C}(\lambda \partial f(\mathbf{x}_0))}{\sqrt{m - \mathbf{D}(\lambda \partial f(\mathbf{x}_0))}}.$$

Observe that $\text{map}(\lambda)$ is well defined over the region \mathcal{R}_{ON} , since $m > \mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ and $m - \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) > \mathbf{C}(\lambda \partial f(\mathbf{x}_0))$ for all $\lambda \in \mathcal{R}_{\text{ON}}$. It can be proven that $\text{map}(\cdot)$ defines a bijective mapping from \mathcal{R}_{ON} to \mathbb{R}^+ [47, Theorem 3.3].

Theorem 3 (Properties of $\text{map}(\cdot)$). Assume $m > \min_{\lambda \geq 0} \mathbf{D}(\lambda \partial f(\mathbf{x}_0))$. The function $\text{map}(\cdot) : \mathcal{R}_{\text{ON}} \rightarrow \mathbb{R}^+$ is strictly increasing and continuous. Thus, its inverse function $\text{map}^{-1}(\cdot) : \mathbb{R}^+ \rightarrow \mathcal{R}_{\text{ON}}$ is well defined.

Some other useful properties of the mapping function include the following:

- $\text{map}(\lambda_{\text{crit}}) = 0$,
- $\lim_{\lambda \rightarrow \lambda_{\text{max}}} \text{map}(\lambda) = \infty$,

4.5.4.2 Proposed Formula

The mapping function can potentially be used to translate results on the NSE of the ℓ_2 -LASSO over \mathcal{R}_{ON} to corresponding results on the ℓ_2^2 -LASSO for $\tau \in \mathbb{R}^+$. Assume $m > \mathbf{D}(\lambda_{\text{best}} \cdot \partial f(\mathbf{x}_0))$. It is conjectured in [47] that for any $\tau > 0$,

$$\frac{\mathbf{D}(\text{map}^{-1}(\tau) \cdot \partial f(\mathbf{x}_0))}{m - \mathbf{D}(\text{map}^{-1}(\tau) \cdot \partial f(\mathbf{x}_0))},$$

accurately characterizes the NSE $\|\hat{\mathbf{x}}_{\ell_2^2} - \mathbf{x}_0\|^2 / (m\sigma^2)$ for sufficiently small σ , and upper bounds it for arbitrary σ . Extended numerical simulations (see Section 13 in [47]) support the conjecture.

This formula would suggest a simple recipe for computing the optimal value of the penalty parameter, which we call τ_{best} . Recall that λ_{best} minimizes the error in the ℓ_2 -LASSO. Then, the proposed mapping between the two problems suggests that $\tau_{\text{best}} = \text{map}(\lambda_{\text{best}})$. To evaluate $\text{map}(\lambda_{\text{best}})$ we make use of Lemma 1 and the fact that $\frac{d\mathbf{D}(\lambda \partial f(\mathbf{x}_0))}{d\lambda} = -\frac{2}{\lambda} \mathbf{C}(\lambda \partial f(\mathbf{x}_0))$ for all $\lambda \geq 0$. Combine this with the fact that λ_{best} is the unique minimizer of $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$, to show that $\mathbf{C}(\lambda_{\text{best}} \cdot \partial f(\mathbf{x}_0)) = 0$, and to conclude with,

$$\tau_{\text{best}} = \lambda_{\text{best}} \sqrt{m - \mathbf{D}(\lambda_{\text{best}} \cdot \partial f(\mathbf{x}_0))} \approx \lambda_{\text{best}} \sqrt{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}.$$

4.6 The NSE of Generalized LASSO with Arbitrary Fixed Noise

Here, we relax the assumption of Section 4.5 that the entries of \mathbf{z} are i.i.d. normal. Instead, assume that the noise vector \mathbf{z} is arbitrary, but still independent of the sensing matrix \mathbf{A} . Under this assumption, we derive simple and non-asymptotic upper bounds on the NSE of the C-LASSO and of the ℓ_2 -LASSO. Those upper bounds can be interpreted as generalizations of the bound on the error of the OLS as was discussed in Section 4.2.2. Compared to the bounds of Section 4.5, the bounds derived here not only hold under more general assumption on the noise vector, but they are also non-asymptotic.

4.6.1 C-LASSO

Recall the generalized C-LASSO in (4.6). Section 4.6.1.1 introduces an upper bound on its NSE for arbitrary fixed noise vector that is independent of \mathbf{A} . In Section 4.6.1.2 we compare this bound to the result of Theorem 1, and Section 4.6.1.3 provides an overview of the proof technique.

4.6.1.1 NSE

Theorem 4 ([46]). Assume $m \geq 2$ and $0 < t \leq \sqrt{m-1} - \sqrt{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}$. Then, with probability, $1 - 6\exp(-t^2/26)$,

$$\frac{\|\hat{\mathbf{x}}_c - \mathbf{x}_0\|}{\|\mathbf{z}\|} \leq \frac{\sqrt{m}}{\sqrt{m-1}} \frac{\sqrt{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} + t}{\sqrt{m-1} - \sqrt{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} - t}.$$

4.6.1.2 Comparison to Theorem 1

It is interesting to see how the bound of Theorem 4 compares to the result of Theorem 1 in the case $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$. Of course, when this is the case the bound of Theorem 1 is tight and our intention is to see how loose is the bound of Theorem 4. Essentially¹², the only difference appears in the denominators of the two bounds; $\sqrt{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} \geq \sqrt{m} - \sqrt{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}$ for all regimes of $0 \leq \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) < m$. The contrast becomes significant when $m \approx \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$. In particular, setting $m = (1 + \varepsilon)^2 \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$, we have

$$\frac{\sqrt{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}}{\sqrt{m} - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} = \frac{\sqrt{2\varepsilon + \varepsilon^2}}{\varepsilon} = \sqrt{\frac{2}{\varepsilon} + 1}.$$

Thus, when ε is large, the bound of Theorem 4 is arbitrarily tight. On the other hand, when ε is small, it can be arbitrarily worse. Simulation results (see Figure 4.6) verify that the error bound of Theorem 4 becomes sharp as the number of measurements m increases. Besides, even if tighter, the bound of Theorem 1 requires stronger assumptions namely, an i.i.d. Gaussian noise vector \mathbf{z} and an asymptotic setting where m and $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ are large enough.

4.6.1.3 Proof Overview

We only provide an overview of the proof. The details can be found in [46]. We begin with introducing some useful notation. $\mathbf{A}\mathcal{T}_f(\mathbf{x}_0)$ will denote the cone obtained by multiplying elements of $\mathcal{T}_f(\mathbf{x}_0)$ by \mathbf{A} , i.e.,

$$\mathbf{A}\mathcal{T}_f(\mathbf{x}_0) = \{\mathbf{Av} \in \mathbb{R}^m \mid \mathbf{v} \in \mathcal{T}_f(\mathbf{x}_0)\}.$$

The lemma below derives a deterministic upper bound on the squared error of the C-LASSO. It is interesting to compare this to the corresponding bound (4.10) for the OLS. Recall the notions of “tangent cone” and “restricted minimum singular value” introduced in Section 4.4.

¹²Precisely: assuming $m \approx m - 1$ and ignoring the t 's in the bound of Theorem 4.

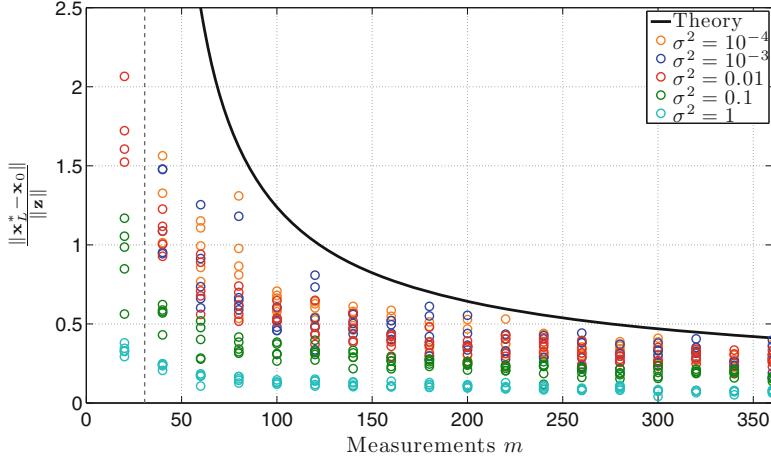


Fig. 4.6 NSE of the C-LASSO with ℓ_1 . The dimension of the \mathbf{x}_0 is $n = 500$ and its sparsity is $k = 5$. The number of measurements m varies from 0 to 360. We plot the empirical NSE assuming $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$ for several values of σ . The solid black line corresponds to the bound of Theorem 4. The dashed line corresponds to the phase transition line of noiseless CS $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$.

Lemma 9 (Deterministic error bound).

$$\|\hat{\mathbf{x}}_c - \mathbf{x}_0\| \leq \frac{\|\text{Proj}(\mathbf{z}, \mathbf{A}\mathcal{T}_f(\mathbf{x}_0))\|}{\sigma_{\min}(\mathbf{A}, \mathcal{T}_f(\mathbf{x}_0) \cap \mathcal{S}^{n-1})}.$$

Proof. From first-order optimality conditions (e.g., [52, p. 270–271]),

$$\langle \mathbf{A}^T(\mathbf{A}\hat{\mathbf{x}}_c - \mathbf{y}), \hat{\mathbf{x}}_c - \mathbf{x}_0 \rangle \leq 0.$$

Writing $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z}$, and rearranging terms, we find that

$$\begin{aligned} \|\mathbf{A}(\hat{\mathbf{x}}_c - \mathbf{x}_0)\| &\leq \left\langle \mathbf{z}, \frac{\mathbf{A}(\hat{\mathbf{x}}_c - \mathbf{x}_0)}{\|\mathbf{A}(\hat{\mathbf{x}}_c - \mathbf{x}_0)\|} \right\rangle \\ &\leq \sup_{\mathbf{v} \in \mathbf{A}\mathcal{T}_f(\mathbf{x}_0) \cap \mathcal{S}^{n-1}} \langle \mathbf{z}, \mathbf{v} \rangle \end{aligned} \tag{4.39}$$

$$\begin{aligned} &\leq \sup_{\mathbf{v} \in \mathbf{A}\mathcal{T}_f(\mathbf{x}_0) \cap \mathcal{B}^{n-1}} \langle \mathbf{z}, \mathbf{v} \rangle \\ &= \|\text{Proj}(\mathbf{z}, \mathbf{A}\mathcal{T}_f(\mathbf{x}_0))\|. \end{aligned} \tag{4.40}$$

(4.39) follows since $\hat{\mathbf{x}}_c - \mathbf{x}_0 \in \mathcal{T}_f(\mathbf{x}_0)$. For (4.40), we applied Moreau's decomposition Theorem [52, Theorem 31.5]. To conclude with the desired result it remains to invoke the definition of the restricted singular values $\sigma_{\min}(\mathbf{A}, \mathcal{T}_f(\mathbf{x}_0) \cap \mathcal{S}^{n-1})$.

To prove Theorem 4 we will translate the deterministic bound of Lemma 9 to a probabilistic one. For this, we need a high-probability lower bound for $\sigma_{\min}(\mathbf{A}, \mathcal{T}_f(\mathbf{x}_0) \cap \mathcal{S}^{n-1})$ and a high-probability upper bound for $\|\text{Proj}(\mathbf{z}, \mathbf{A}\mathcal{T}_f(\mathbf{x}_0))\|$. The former is given by a direct application of the “escape through a mesh” Lemma 4. The latter requires some more effort to derive. The result is summarized in Lemma 10 below and is the main technical contribution of [46]. The proof makes use of Gordon’s Lemma 3 and can be found in [46].

Lemma 10 (Restricted correlation). *Let $\mathcal{K} \in \mathbb{R}^n$ be a convex and closed cone, $\mathbf{G} \in \mathbb{R}^{m \times n}$ have independent standard normal entries, $m \geq 2$ and $\mathbf{z} \in \mathbb{R}^m$ be arbitrary and independent of \mathbf{G} . For any $t > 0$, pick $\alpha \geq \frac{\sqrt{\mathbf{D}(\mathcal{K}^\circ)} + t}{\gamma_{m-1}} \|\mathbf{z}\|$. Then,*

$$\sup_{\mathbf{v} \in \mathcal{K} \cap \mathcal{S}^{n-1}} \{\mathbf{z}^T \mathbf{Gv} - \alpha \|\mathbf{Gv}\|\} \leq 0, \quad (4.41)$$

with probability $1 - 5 \exp(-\frac{t^2}{26})$.

We may now complete the proof of Theorem 4.

Proof (of Theorem 4). Suppose $0 \leq t < \gamma_m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$. First apply Lemma 4 with $\mathbf{G} = \sqrt{m}\mathbf{A}$ and $\mathcal{K} = \mathcal{T}_f(\mathbf{x}_0)$. Then, with probability $1 - \exp(-\frac{t^2}{2})$,

$$\sigma_{\min}(\mathbf{A}, \mathcal{T}_f(\mathbf{x}_0) \cap \mathcal{S}^{n-1}) \geq \frac{\gamma_m - \sqrt{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} - t}{\sqrt{m}}. \quad (4.42)$$

Next, apply Lemma 10 with $\mathbf{G} = \sqrt{m}\mathbf{A}$ and $\mathcal{K} = \mathcal{T}_f(\mathbf{x}_0)$. With probability $1 - 5 \exp(-\frac{t^2}{26})$,

$$\begin{aligned} \|\text{Proj}(\mathbf{z}, \mathbf{A}\mathcal{T}_f(\mathbf{x}_0))\| &= \mathbf{z}^T \frac{\text{Proj}(\mathbf{z}, \mathbf{A}\mathcal{T}_f(\mathbf{x}_0))}{\|\text{Proj}(\mathbf{z}, \mathbf{A}\mathcal{T}_f(\mathbf{x}_0))\|} \leq \sup_{\mathbf{v} \in \mathcal{T}_f(\mathbf{x}_0) \cap \mathcal{S}^{n-1}} \frac{\mathbf{z}^T \mathbf{Av}}{\|\mathbf{Av}\|} \\ &\leq \frac{\sqrt{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} + t}{\gamma_{m-1}} \|\mathbf{z}\|. \end{aligned} \quad (4.43)$$

Theorem 4 now follows after substituting (4.42) and (4.43) in Lemma 9 and using the following: $\gamma_m \gamma_{m-1} = m - 1$ and $\gamma_{m-1} \leq m - 1$.

4.6.2 ℓ_2 -LASSO

Recall the generalized ℓ_2 -LASSO in (4.7). Section 4.6.2.1 derives an upper bound on its NSE for arbitrary fixed noise vector that is independent of \mathbf{A} . In Section 4.6.2.2 we compare this bound to the result of Theorem 2 and Section 4.6.2.3 provides an overview of the proof.

4.6.2.1 NSE

Theorem 5 ([60]). Assume $m \geq 2$. Fix the regularizer parameter in (4.7) to be $\lambda \geq 0$ and let $\hat{\mathbf{x}}_{\ell_2}$ be a minimizer of (4.7). Then, for any $0 < t \leq (\sqrt{m-1} - \sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))})$, with probability $1 - 5 \exp(-t^2/32)$,

$$\|\hat{\mathbf{x}} - \mathbf{x}_0\| \leq 2\|\mathbf{z}\| \frac{\sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))} + t}{\sqrt{m-1} - \sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))} - t}.$$

Theorem 5 provides a simple, general, non-asymptotic and (rather) sharp upper bound on the error of the regularized lasso estimator (4.7), which also takes into account the specific choice of the regularizer parameter $\lambda \geq 0$. It is non-asymptotic and is applicable in any regime of m , λ and $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$. Also, the constants involved in it are small making it rather tight¹³.

For the bound of Theorem 5 to be at all meaningful, we require $m > \min_{\lambda \geq 0} \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) = \mathbf{D}(\lambda_{\text{best}} \partial f(\mathbf{x}_0))$. Recall that this translates to the number of measurements being large enough to at least guarantee noiseless recovery. Also, similar to the discussion in Section 4.5.3.2 there exists¹⁴ a unique λ_{\max} satisfying $\lambda_{\max} > \lambda_{\text{best}}$ and $\sqrt{\mathbf{D}(\lambda_{\max} \partial f(\mathbf{x}_0))} = \sqrt{m-1}$, and, when $m \leq n$, there exists unique $\lambda_{\min} < \lambda_{\text{best}}$ satisfying $\sqrt{\mathbf{D}(\lambda_{\min} \partial f(\mathbf{x}_0))} = \sqrt{m-1}$. From this, it follows that $\sqrt{m-1} > \sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))}$ if and only if $\lambda \in (\lambda_{\min}, \lambda_{\max})$. This is a superset of \mathcal{R}_{ON} (recall the definition in Section 4.5.3.2) and is exactly the range of values of the regularizer parameter λ for which the bound of Theorem 5 is meaningful.

As a superset of \mathcal{R}_{ON} , $(\lambda_{\min}, \lambda_{\max})$ contains λ_{best} , for which, the bound of Theorem 5 achieves its minimum value since it is strictly increasing in $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$. Recall from Section 4.5.3.4 that deriving λ_{best} does not require knowledge of any properties (e.g. variance) of the noise vector neither does it require knowledge of the unknown signal \mathbf{x}_0 itself.

As a final remark, comparing Theorem 5 to Theorem 4 reveals the similar nature of the two results. Apart from a factor of 2, the upper bound on the error of the regularized lasso (4.7) for fixed λ is essentially the same as the upper bound on the error of the constrained lasso (4.6), with $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ replaced by $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$. This when combined with (4.29) suggests that setting $\lambda = \lambda_{\text{best}}$ in (4.7) achieves performance almost as good as that of (4.6).

¹³It is conjectured in [60] and supported by simulations (e.g., Figure 4.7) that the factor of 2 in Theorem 5 is an artifact of the proof technique and not essential.

¹⁴For proofs of those claims, see Section 8 and in particular Lemma 8.1 in [47].

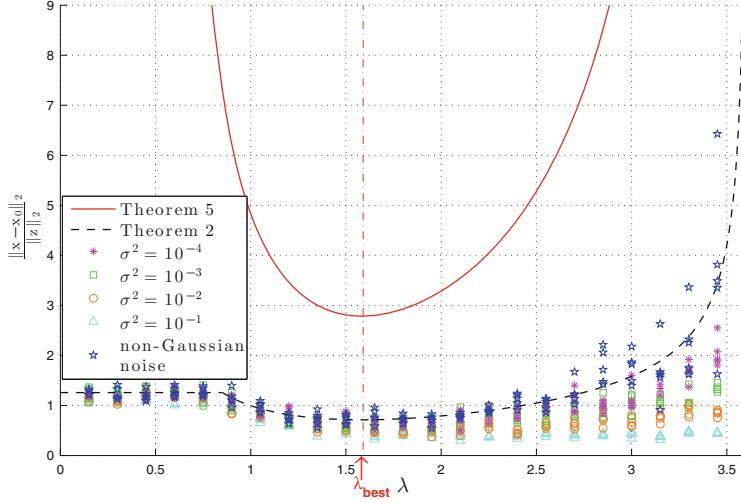


Fig. 4.7 Figure 4.7 illustrates the bound of Theorem 5, which is given in red for $n = 340$, $m = 140$, $k = 10$ and for \mathbf{A} having $\mathcal{N}(0, \frac{1}{m})$ entries. The upper bound of Theorem 2, which is asymptotic in m and only applies to i.i.d. Gaussian \mathbf{z} , is given in black. In our simulations, we assume \mathbf{x}_0 is a random unit norm vector over its support and consider both i.i.d. $\mathcal{N}(0, \sigma^2)$, as well as, non-Gaussian noise vectors \mathbf{z} . We have plotted the realizations of the normalized error for different values of λ and σ . As noted, the bound of Theorem 2 is occasionally violated since it requires very large m , as well as, i.i.d. Gaussian noise. On the other hand, the bound of Theorem 5 always holds.

4.6.2.2 Comparison to Theorem 2

To start with, Theorem 5 is advantageous to Theorem 2 in that it holds in more general setting than standard Gaussian noise and, also, characterizes a superset of \mathcal{R}_{ON} . Furthermore, it is non-asymptotic, while Theorem 2 requires $m, \mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ to be large enough. On the other side, when $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$, then, Theorem 2 offers clearly a tighter bound on the NSE. Yet, apart from a factor of 2, this bound only differs from the bound of Theorem 5 in the denominator, where instead of $\sqrt{m-1} - \sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))}$ we have the larger quantity $\sqrt{m - \mathbf{D}(\lambda \partial f(\mathbf{x}_0))}$. This difference becomes insignificant and indicates that our bound is rather tight when m is large. Finally, the bound of Theorem 2 is only conjectured in [47] to upper bound the estimation error for arbitrary values of the noise variance σ^2 . In contrast, Theorem 5 is a fully rigorous upper bound on the estimation error of (4.7).

4.6.2.3 Proof Overview

It is convenient to rewrite the generalized ℓ_2 -LASSO in terms of the error vector $\mathbf{w} = \mathbf{x} - \mathbf{x}_0$ as follows:

$$\min_{\mathbf{w}} \|\mathbf{Aw} - \mathbf{z}\| + \frac{\lambda}{\sqrt{m}} (f(\mathbf{x}_0 + \mathbf{w}) - f(\mathbf{x}_0)). \quad (4.44)$$

Denote the solution of (4.44) by $\hat{\mathbf{w}}$. Then, $\hat{\mathbf{w}} = \hat{\mathbf{x}} - \mathbf{x}_0$ and we want to bound $\|\hat{\mathbf{w}}\|$. To simplify notation, for the rest of the proof, we denote the value of that desired upper bound as

$$\ell(t) := 2\|\mathbf{z}\| \frac{\sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))} + t}{\sqrt{m-1} - \sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))} - t}. \quad (4.45)$$

It is easy to see that the optimal value of the minimization in (4.44) is no greater than $\|\mathbf{z}\|$. Observe that $\mathbf{w} = \mathbf{0}$ achieves this value. However, Lemma 11 below shows that if we constrain the minimization in (4.44) to be only over vectors \mathbf{w} whose norm is greater than $\ell(t)$, then the resulting optimal value is (with high probability on the measurement matrix \mathbf{A}) strictly greater than $\|\mathbf{z}\|$. Combining those facts yields the desired result, namely $\|\hat{\mathbf{w}}\| \leq \ell(t)$. The fundamental technical tool in the proof of Lemma 11 is (not surprisingly at this point) Gordon's Lemma 3.

Lemma 11. *Fix some $\lambda \geq 0$ and $0 < t \leq (\sqrt{m-1} - \sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))})$. Let $\ell(t)$ be defined as in (4.45). Then, with probability $1 - 5 \exp(-t^2/32)$, we have*

$$\min_{\|\mathbf{w}\| \geq \ell(t)} \{ \|\mathbf{Aw} - \mathbf{z}\| + \frac{\lambda}{\sqrt{m}} (f(\mathbf{x}_0 + \mathbf{w}) - f(\mathbf{x}_0)) \} > \|\mathbf{z}\|. \quad (4.46)$$

Proof. Fix λ and t , as in the statement of the lemma. From the convexity of $f(\cdot)$, $f(\mathbf{x}_0 + \mathbf{w}) - f(\mathbf{x}_0) \geq \max_{\mathbf{s} \in \partial f(\mathbf{x}_0)} \mathbf{s}^T \mathbf{w}$. Hence, it suffices to prove that w.h.p. over \mathbf{A} ,

$$\min_{\|\mathbf{w}\| \geq \ell(t)} \{ \sqrt{m} \|\mathbf{Aw} - \mathbf{z}\| + \max_{\mathbf{s} \in \partial f(\mathbf{x}_0)} \mathbf{s}^T \mathbf{w} \} > \sqrt{m} \|\mathbf{z}\|.$$

We begin with applying Gordon's Lemma 3 to the optimization problem in the expression above. Define $\bar{\mathbf{z}} = \sqrt{m}\mathbf{z}$, rewrite $\|\mathbf{Aw} - \mathbf{z}\|$ as $\max_{\|\mathbf{a}\|=1} \{ \mathbf{a}^T \mathbf{Aw} - \mathbf{a}^T \mathbf{z} \}$ and, then, apply Lemma 3 with $\mathbf{G} = \sqrt{m}\mathbf{A}$, $\mathcal{S} = \{ \mathbf{w} \mid \|\mathbf{w}\| \geq \ell(t) \}$ and $\psi(\mathbf{w}, \mathbf{a}) = -\mathbf{a}^T \bar{\mathbf{z}} + \max_{\mathbf{s} \in \partial f(\mathbf{x}_0)} \mathbf{s}^T \mathbf{w}$. This leads to the following statement:

$$\mathbb{P}((4.46) \text{ is true}) \geq 2 \cdot \mathbb{P}(\mathcal{L}(t; \mathbf{g}, \mathbf{h}) > \|\bar{\mathbf{z}}\|) - 1,$$

where $\mathcal{L}(t; \mathbf{g}, \mathbf{h})$ is defined as

$$\min_{\|\mathbf{w}\| \geq \ell(t)} \max_{\|\mathbf{a}\|=1} \{ (\|\mathbf{w}\| \mathbf{g} - \bar{\mathbf{z}})^T \mathbf{a} - \min_{\mathbf{s} \in \partial f(\mathbf{x}_0)} (\mathbf{h} - \mathbf{s})^T \mathbf{w} \}. \quad (4.47)$$

In the remaining, we analyze the simpler optimization problem defined in (4.47), and prove that $\mathcal{L}(t; \mathbf{g}, \mathbf{h}) > \|\bar{\mathbf{z}}\|$ holds with probability $1 - \frac{5}{2} \exp(-t^2/32)$. We begin with simplifying the expression for $\mathcal{L}(t; \mathbf{g}, \mathbf{h})$, as follows:

$$\begin{aligned}\mathcal{L}(t; \mathbf{g}, \mathbf{h}) &= \min_{\|\mathbf{w}\| \geq \ell(t)} \{ \|\|\mathbf{w}\| \mathbf{g} - \bar{\mathbf{z}}\| - \min_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} (\mathbf{h} - \mathbf{s})^T \mathbf{w} \} \\ &\geq \min_{\alpha \geq \ell(t)} \{ \|\alpha \mathbf{g} - \bar{\mathbf{z}}\| - \alpha \text{dist}(\mathbf{h}, \lambda \partial f(\mathbf{x}_0)) \} \\ &= \min_{\alpha \geq \ell(t)} \{ \sqrt{\alpha^2 \|\mathbf{g}\|^2 + \|\bar{\mathbf{z}}\|^2 - 2\alpha \mathbf{g}^T \bar{\mathbf{z}}} - \alpha \text{dist}(\mathbf{h}, \lambda \partial f(\mathbf{x}_0)) \}.\end{aligned}$$

The first equality above follows after performing the trivial maximization over \mathbf{a} in (4.47). For the inequality that follows, we apply the minimax inequality [52, Lemma 36.1] and use the second equality in (4.35). Next, we show that $\mathcal{L}(t; \mathbf{g}, \mathbf{h})$ is strictly greater than $\|\bar{\mathbf{z}}\|$ with the desired high probability over realizations of \mathbf{g} and \mathbf{h} . Consider the event \mathcal{E}_t of \mathbf{g} and \mathbf{h} satisfying all three conditions listed below,

$$1. \|\mathbf{g}\| \geq \gamma_m - t/4, \tag{4.48a}$$

$$2. \text{dist}(\mathbf{h}, \lambda \partial f(\mathbf{x}_0)) \leq \sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))} + t/4, \tag{4.48b}$$

$$3. \mathbf{g}^T \bar{\mathbf{z}} \leq (t/4) \|\bar{\mathbf{z}}\|. \tag{4.48c}$$

The conditions in (4.48) hold with high probability. In particular, the first two hold with probability no less than $1 - \exp(-t^2/32)$. This is because the ℓ_2 -norm and the distance function to a convex set are both 1-Lipschitz functions and, thus, Lemma 5 applies. The third condition holds with probability at least $1 - (1/2) \exp(-t^2/32)$, since $\mathbf{g}^T \bar{\mathbf{z}}$ is statistically identical to $\mathcal{N}(0, \|\bar{\mathbf{z}}\|^2)$. Union bounding yields

$$\mathbb{P}(\mathcal{E}_t) \geq 1 - (5/2) \exp(-t^2/32). \tag{4.49}$$

Furthermore, it can be shown (see Lemma 4.2 in [60]) that if \mathbf{g} and \mathbf{h} are such that \mathcal{E}_t is satisfied, then $\mathcal{L}(t; \mathbf{g}, \mathbf{h}) > \|\bar{\mathbf{z}}\|$. This, when combined with (4.49) shows that $\mathbb{P}(\mathcal{L}(t; \mathbf{g}, \mathbf{h}) > \|\bar{\mathbf{z}}\|) \geq 1 - (5/2) \exp(-t^2/32)$, completing the proof of Lemma 11.

4.7 The Worst-Case NSE of Generalized LASSO

Here, we assume no restriction at all on the distribution of the noise vector \mathbf{z} . In particular, this includes the case of *adversarial* noise, i.e., noise that has information on \mathbf{A} and can adapt itself accordingly. We compute the resulting worst-case NSE of the C-LASSO in the next section.

4.7.1 C-LASSO

4.7.1.1 NSE

Theorem 6. Assume $0 < t \leq \sqrt{m} - \sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))}$. Then, with probability $1 - \exp(-t^2/2)$,

$$\frac{\|\hat{\mathbf{x}}_c - \mathbf{x}_0\|}{\|\mathbf{z}\|} \leq \frac{\sqrt{m}}{\gamma_m - \sqrt{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} - t}.$$

Recall in the statement of Theorem 6 that $\gamma_m = \mathbb{E}[\|\mathbf{g}\|]$, with $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_m)$. For large m , $\gamma_m \approx \sqrt{m}$ and according to Theorem 6 the worst-case NSE of the C-LASSO can be as large as 1. Contrast this to Theorem 4 and the case where \mathbf{z} is not allowed to depend on \mathbf{A} . There, the NSE is approximately $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))/m$ for large m .

4.7.1.2 Proof

The proof of Theorem 6 follows easily from Lemma 9:

$$\|\hat{\mathbf{x}}_c - \mathbf{x}_0\| \leq \frac{\|\mathbf{A}(\hat{\mathbf{x}}_c - \mathbf{x}_0)\|}{\sigma_{\min}(\mathbf{A}, \mathcal{T}_f(\mathbf{x}_0) \cap \mathcal{S}^{n-1})} \leq \frac{\|\mathbf{z}\|}{\sigma_{\min}(\mathbf{A}, \mathcal{T}_f(\mathbf{x}_0) \cap \mathcal{S}^{n-1})}.$$

Apply (4.42) to the above, to conclude with the desired upper bound.

4.8 Conclusion

Precise undersampling theorems were first studied by Donoho and Tanner [19, 20]. They characterized the phase transitions for ℓ_1 minimization in the noiseless setup. Since then, there have been major improvements in several directions. The theory was initially limited to sparse signal recovery and noiseless setup. With recent advances, we can precisely analyze arbitrary convex functions under noise.

We have presented a detailed analysis of the normalized squared error of the generalized LASSO under the assumption that the entries of the sensing matrix being i.i.d. normal. The derived formulae are precise and simple. They only involve two geometric measures of signal complexity, the Gaussian squared distance to the scaled subdifferential and to the cone of subdifferential, which also appear in the most recent literature of noiseless compressed sensing [1, 14, 55, 57]. Moreover, they admit insightful interpretations when seen as generalizations of the corresponding formulae for the classical ordinary least-squares estimation.

There are several remaining directions for future work. Some of these involve improving and adding on the Theorems presented here. For instance, the multiplicative factor of 2 in Theorem 5 seems to be unnecessary, urging for an improvement.

Similar in nature are the problems of establishing the conjecture of [47] on the NSE of the ℓ_2^2 -LASSO (see Section 4.5.4.2) and studying the worst-case NSE of the ℓ_2 -LASSO. A potentially more challenging direction for future work involves studying different measurement ensembles and the universality phenomenon. It is now widely accepted that, whether the sensing matrix is i.i.d. Gaussian or i.i.d. Bernoulli, this does not change the phase transition and stability characteristics of the linear inverse problems [22, 23]. A provable extension of the results of this chapter to other measurement ensembles would, thus, be of great interest.

Appendix

The upper bounds on the NSE of the generalized LASSO presented in Sections 4.5–4.7 are in terms of the summary parameters $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ and $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$. While the bounds are simple, concise, and nicely resemble the corresponding ones in the case of OLS, it may appear to the reader that the formulae are rather abstract, because of the presence of $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$.

However, as discussed here, for a large number of widely used convex regularizers $f(\cdot)$, one can calculate (tight) upper bounds or even explicit formulae for these quantities. For example, for the estimation of a k -sparse signal \mathbf{x}_0 with $f(\cdot) = \|\cdot\|_1$, it has been shown that $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) \lesssim 2k(\log \frac{n}{k} + 1)$. Substituting this into Theorems 1 and 4 results in the “closed-form” upper bounds given in (4.20) and (4.22), i.e. ones expressed only in terms of m , n , and k . Analogous results have been derived [14, 31, 44, 54] for other well-known signal models as well, including low rankness and block-sparsity¹⁵. The first column of Table 4.1 summarizes some of the results for $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ found in the literature [14, 31]. The second column provides closed form results on $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ when λ is sufficiently large [47]. Note that, by setting λ to its lower bound in the second row, one approximately obtains the corresponding result in the first row. This should not be surprising due to (4.29). Also, this value of λ is a good proxy for the optimal regularizer λ_{best} of the ℓ_2 -LASSO as was discussed in Sections 4.5.3.4 and 4.6.2.1.

We refer the reader to [1, 14, 31, 47] for the details and state-of-the-art bounds on $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$. Identifying the subdifferential $\partial f(\mathbf{x}_0)$ and calculating $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ for all $\lambda \geq 0$ are the critical steps. Once those are available, computing $\min_{\lambda \geq 0} \mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ provides upper approximation formulae for $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$. This idea was first introduced by Stojnic [55] and was subsequently refined and generalized in [14]. Most recently [1, 31] proved (4.29),

¹⁵ We say $\mathbf{x}_0 \in \mathbb{R}^n$ is block-sparse if it can be grouped into t known blocks of size $b = n/t$ each so that only k of these t blocks are nonzero. To induce the structure, the standard approach is to use the $\ell_{1,2}$ norm which sums up the ℓ_2 norms of the blocks, [29, 48, 54, 58]. In particular, denoting the subvector corresponding to i 'th block of a vector \mathbf{x} by \mathbf{x}_i , the $\ell_{1,2}$ norm is defined as $\|\mathbf{x}\|_{1,2} = \sum_{i=1}^t \|\mathbf{x}_i\|_2$.

Table 4.1 Closed form upper bounds for $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$.

	$\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$	$\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$
<i>k</i> -sparse, $\mathbf{x}_0 \in \mathbb{R}^n$	$2k(\log \frac{n}{k} + 1)$	$(\lambda^2 + 3)k$ for $\lambda \geq \sqrt{2 \log \frac{n}{k}}$
<i>k</i> -block sparse, $\mathbf{x}_0 \in \mathbb{R}^{tb}$	$6\sqrt{nr}$	$\lambda^2 r + 2\sqrt{n}(r + 1)$ for $\lambda \geq 2n^{1/4}$
Rank r , $\mathbf{X}_0 \in \mathbb{R}^{\sqrt{n} \times \sqrt{n}}$	$4k(\log \frac{t}{k} + b)$	$(\lambda^2 + b + 2)k$ for $\lambda \geq \sqrt{b} + \sqrt{2 \log \frac{t}{k}}$

thus showing that the resulting approximation on $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ is in fact highly accurate. Section 4 of [1] is an excellent reference for further details and the notation used there is closer to ours.

We should emphasize that examples of regularizers are not limited to the ones discussed here and presented in Table 4.1. There are increasingly more signal classes that exhibit low-dimensionality and to which the theorems of Sections 4.5–4.7 would apply. Some of these are as follows.

- Non-negativity constraint: \mathbf{x}_0 has non-negative entries, [20].
- Low-rank plus sparse matrices: \mathbf{x}_0 can be represented as sum of a low-rank and a sparse matrix, [65].
- Signals with sparse gradient: Rather than \mathbf{x}_0 itself, its gradient $\mathbf{d}_{\mathbf{x}_0}(i) = \mathbf{x}_0(i) - \mathbf{x}_0(i-1)$ is sparse, [8].
- Low-rank tensors: \mathbf{x}_0 is a tensor and its unfoldings are low-rank matrices, [32, 36].
- Simultaneously sparse and low-rank matrices: For instance, $\mathbf{x}_0 = \mathbf{s}\mathbf{s}^T$ for a sparse vector \mathbf{s} , [45, 51].

Establishing new and tighter analytic bounds for $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ and $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ for more regularizers f is certainly an interesting direction for future research. In the case where such analytic bounds do not already exist in literature or are hard to derive, one can numerically estimate $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ and $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ once there is an available characterization of the subdifferential $\partial f(\mathbf{x}_0)$. Using the concentration property of $\text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))$ around $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$, when $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$, we can compute $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$, as follows:

1. draw a vector $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$,
2. return the solution of the convex program $\min_{\mathbf{s} \in \partial f(\mathbf{x}_0)} \|\mathbf{h} - \lambda \mathbf{s}\|^2$.

Computing $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ can be built on the same recipe by recognizing $\text{dist}^2(\mathbf{h}, \text{cone}(\partial f(\mathbf{x}_0)))$ as $\min_{\lambda \geq 0, \mathbf{s} \in \partial f(\mathbf{x}_0)} \|\mathbf{h} - \lambda \mathbf{s}\|^2$.

To sum up, any bound on $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ and $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ translates, through Theorems 1–6, into corresponding upper bounds on the NSE of the generalized LASSO. For purposes of illustration and completeness, we review next the details of computing $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ for the celebrated case where \mathbf{x}_0 is sparse and the ℓ_1 -norm is used as the regularizer.

Sparse signals

Suppose \mathbf{x}_0 is a k -sparse signal and $f(\cdot) = \|\cdot\|_1$. Denote by S the support set of \mathbf{x}_0 , and by S^c its complement. The subdifferential at \mathbf{x}_0 is [52],

$$\partial f(\mathbf{x}_0) = \{\mathbf{s} \in \mathbb{R}^n \mid \|\mathbf{s}\|_\infty \leq 1 \text{ and } \mathbf{s}_i = \text{sign}((\mathbf{x}_0)_i), \forall i \in S\}.$$

Let $\mathbf{h} \in \mathbb{R}^n$ have i.i.d. $\mathcal{N}(0, 1)$ entries and define

$$\text{shrink}(\chi, \lambda) = \begin{cases} \chi - \lambda & , \chi > \lambda, \\ 0 & , -\lambda \leq \chi \leq \lambda, \\ \chi + \lambda & , \chi < -\lambda. \end{cases}$$

Then, $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ is equal to ([1, 14])

$$\begin{aligned} \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) &= \mathbb{E}[\text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))] \\ &= \sum_{i \in S} \mathbb{E}[(\mathbf{h}_i - \lambda \text{sign}((\mathbf{x}_0)_i))^2] + \sum_{i \in S^c} \mathbb{E}[\text{shrink}^2(\mathbf{h}_i, \lambda)] = \\ &= k(1 + \lambda^2) + (n - k) \sqrt{\frac{2}{\pi}} \left[(1 + \lambda^2) \int_{\lambda}^{\infty} e^{-t^2/2} dt - \lambda \exp(-\lambda^2/2) \right]. \end{aligned} \tag{4.50}$$

Note that $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ depends only on n, λ and $k = |S|$, and *not* explicitly on S itself (which is not known). Substituting the expression in (4.50) in place of the $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ in Theorems 2 and 5 yields explicit expressions for the corresponding upper bounds in terms of n, m, k , and λ .

We can obtain an even simpler upper bound on $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ which does not involve error functions as we show below. Denote $Q(t) = \frac{1}{\sqrt{2\pi}} \int_t^{\infty} e^{-\tau^2/2} d\tau$ the complementary c.d.f. of a standard normal random variable. Then,

$$\begin{aligned} \frac{1}{2} \mathbb{E}[\text{shrink}^2(\mathbf{h}_i, \lambda)] &= \int_{\lambda}^{\infty} (t - \lambda)^2 d(-Q(t)) \\ &= -[(t - \lambda)^2 Q(t)]_{\lambda}^{\infty} + 2 \int_{\lambda}^{\infty} (t - \lambda) Q(t) dt \\ &\leq \int_{\lambda}^{\infty} (t - \lambda) e^{-t^2/2} dt \end{aligned} \tag{4.51}$$

$$\leq e^{-\lambda^2/2} - \frac{\lambda^2}{\lambda^2 + 1} e^{-\lambda^2/2} \tag{4.52}$$

$$= \frac{1}{\lambda^2 + 1} e^{-\lambda^2/2}.$$

(4.51) and (4.52) follow from standard upper and lower tail bounds on normal random variables, namely $\frac{1}{\sqrt{2\pi}} \frac{t}{t^2+1} e^{-t^2/2} \leq Q(t) \leq \frac{1}{2} e^{-t^2/2}$. From this, we find that

$$\mathbf{D}(\lambda \partial f(\mathbf{x}_0)) \leq k(1 + \lambda^2) + (n - k) \frac{2}{\lambda^2 + 1} e^{-\lambda^2/2}.$$

Letting $\lambda \geq \sqrt{2 \log(\frac{n}{k})}$ in the above expression recovers the corresponding entry in Table 4.1:

$$\mathbf{D}(\lambda \partial f(\mathbf{x}_0)) \leq (\lambda^2 + 3)k, \text{ when } \lambda \geq \sqrt{2 \log(\frac{n}{k})}. \quad (4.53)$$

Substituting (4.53) in Theorems 2 and 5 recovers the bounds in (4.21) and (4.23), respectively.

Setting $\lambda = \sqrt{2 \log(\frac{n}{k})}$ in (4.53) provides an approximation to $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$. In particular, $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) \leq 2k(\log(\frac{n}{k}) + 3/2)$. [14] obtains an even tighter bound $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) \leq 2k(\log(\frac{n}{k}) + 3/4)$ starting again from (4.50), but using different tail bounds for Gaussians. We refer the reader to Proposition 3.10 in [14] for the exact details.

Acknowledgements The authors gratefully acknowledge the anonymous reviewers for their attention and their helpful comments.

References

1. Amelunxen, D., Lotz, M., McCoy, M.B., Tropp, J.A.: Living on the edge: a geometric theory of phase transitions in convex optimization. arXiv preprint. arXiv:1303.6672 (2013)
2. Bayati, M., Montanari, A.: The dynamics of message passing on dense graphs, with applications to compressed sensing. IEEE Trans. Inf. Theory **57**(2), 764–785 (2011)
3. Bayati, M., Montanari, A.: The LASSO risk for gaussian matrices. IEEE Trans. Inf. Theory **58**(4), 1997–2017 (2012)
4. Belloni, A., Chernozhukov, V., Wang, L.: Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika **98**(4), 791–806 (2011)
5. Bertsekas, D., Nedic, A., Ozdaglar, A.: Convex Analysis and Optimization. Athena Scientific (2003)
6. Bickel, P.J., Ritov, Y., Tsybakov, A.B.: Simultaneous analysis of Lasso and Dantzig selector. Ann. Stat. **37**(4), 1705–1732 (2009)
7. Borwein, J.M., Lewis, A.S.: Convex Analysis and Nonlinear Optimization: Theory and Examples, vol. 3. Springer, New York (2010)
8. Cai, J.-F., Xu, W.: Guarantees of total variation minimization for signal recovery. arXiv preprint. arXiv:1301.6791 (2013)
9. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717–772 (2009)
10. Candès, E., Tao, T.: Near-optimal signal recovery from random projections: Universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
11. Candès, E., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n . Ann. Stat. **35**, 2313–2351 (2007)

12. Candes, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59**(8), 1207–1223 (2006)
13. Chandrasekaran, V., Jordan, M.I.: Computational and statistical tradeoffs via convex relaxation. *Proc. Natl. Acad. Sci.* **110**(13), E1181–E1190 (2013)
14. Chandrasekaran, V., Recht, B., Parrilo, P.A., Willsky, A.S.: The convex geometry of linear inverse problems. *Found. Comput. Math.* **12**(6), 805–849 (2012)
15. Donoho, D.L.: De-noising by soft-thresholding. *IEEE Trans. Inf. Theory* **41**(3), 613–627 (1995)
16. Donoho, D.L.: High-dimensional data analysis: the curses and blessings of dimensionality. *Aide-memoire of a lecture at “AMS Conference on Math Challenges of the 21st Century”*. Citeseer (2000)
17. Donoho, D.L.: Compressed sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006)
18. Donoho, D.L.: High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput. Geom.* **35**(4), 617–652 (2006)
19. Donoho, D.L., Tanner, J.: Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA* **102**(27), 9452–9457 (2005)
20. Donoho, D.L., Tanner, J.: Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci. USA* **102**(27), 9446–9451 (2005)
21. Donoho, D.L., Tanner, J.: Thresholds for the recovery of sparse solutions via ℓ_1 minimization. In: *The 40th Annual Conference on Information Sciences and Systems, 2006*, pp. 202–206. IEEE, New York (2006)
22. Donoho, D., Tanner, J.: Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. Roy. Soc. A Math. Phys. Eng. Sci.* **367**(1906), 4273–4293 (2009)
23. Donoho, D.L., Tanner, J.: Precise undersampling theorems. *Proc. IEEE* **98**(6), 913–924 (2010)
24. Donoho, D.L., Elad, M., Temlyakov, V.N.: Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inf. Theory* **52**(1), 6–18 (2006)
25. Donoho, D.L., Maleki, A., Montanari, A.: Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.* **106**(45), 18914–18919 (2009)
26. Donoho, D.L., Maleki, A., Montanari, A.: The noise-sensitivity phase transition in compressed sensing. *IEEE Trans. Inf. Theory* **57**(10), 6920–6941 (2011)
27. Donoho, D., Johnstone, I., Montanari, A.: Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE Trans. Inf. Theory* **59**(6), 3396–3433 (2013)
28. Donoho, D.L., Gavish, M., Montanari, A.: The phase transition of matrix recovery from Gaussian measurements matches the minimax mse of matrix denoising. *Proc. Natl. Acad. Sci. USA* **110**(21), 8405–8410 (2013)
29. Eldar, Y.C., Kuppinger, P., Bolcskei, H.: Block-sparse signals: uncertainty relations and efficient recovery. *IEEE Trans. Signal Process.* **58**(6), 3042–3054 (2010)
30. Fazel, M.: Matrix rank minimization with applications. Ph.D. thesis (2002)
31. Foygel, R., Mackey, L.: Corrupted sensing: novel guarantees for separating structured signals. arXiv preprint. arXiv:1305.2524 (2013)
32. Gandy, S., Recht, B., Yamada, I.: Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Prob.* **27**(2), 025010 (2011)
33. Gordon, Y.: Some inequalities for Gaussian processes and applications. *Isr. J. Math.* **50**(4), 265–289 (1985)
34. Gordon, Y.: *On Milman’s Inequality and Random Subspaces Which Escape Through a Mesh in \mathbb{R}^n* . Springer, New York (1988)
35. Härdle, W., Simar, L.: *Applied Multivariate Statistical Analysis*, vol. 2. Springer, Berlin (2007)
36. Kressner, D., Steinlechner, M., Vandereycken, B.: Low-rank tensor completion by Riemannian optimization. *BIT Numer. Math.* **54**(2):447–468 (2014)
37. Ledoux, M., Talagrand, M.: *Probability in Banach Saces: Isoperimetry and Processes*, vol. 23. Springer, Berlin (1991)

38. Maleki, M.A.: Approximate Message Passing Algorithms for Compressed Sensing. Stanford University, Stanford (2010)
39. Maleki, A., Anitori, L., Yang, Z., Baraniuk, R.G.: Asymptotic analysis of complex lasso via complex approximate message passing (camp). *IEEE Trans. Inf. Theory* **59**(7):4290–4308 (2013)
40. McCoy, M.B., Tropp, J.A.: From Steiner formulas for cones to concentration of intrinsic volumes. *Discrete Comput. Geom.* **51**(4), 926–963 (2014)
41. Merriman, M.: On the history of the method of least squares. *Analyst* **4**, 33–36 (1877)
42. Negahban, S.N., Ravikumar, P., Wainwright, M.J., Yu, B.: A unified framework for high-dimensional analysis of m -estimators with decomposable regularizers. *Stat. Sci.* **27**(4), 538–557 (2012)
43. Oymak, S., Hassibi, B.: Sharp MSE bounds for proximal denoising. arXiv preprint. arXiv:1305.2714 (2013)
44. Oymak, S., Mohan, K., Fazel, M., Hassibi, B.: A simplified approach to recovery conditions for low rank matrices. In: *IEEE International Symposium on Information Theory Proceedings (ISIT)*, 2011, pp. 2318–2322. IEEE, New York (2011)
45. Oymak, S., Jalali, A., Fazel, M., Eldar, Y.C., Hassibi, B.: Simultaneously structured models with application to sparse and low-rank matrices. arXiv preprint. arXiv:1212.3753 (2012)
46. Oymak, S., Thrampoulidis, C., Hassibi, B.: Simple bounds for noisy linear inverse problems with exact side information. arXiv preprint. arXiv:1312.0641 (2013)
47. Oymak, S., Thrampoulidis, C., Hassibi, B.: The squared-error of generalized LASSO: a precise analysis. arXiv preprint. arXiv:1311.0830 (2013)
48. Rao, N., Recht, B., Nowak, R.: Tight measurement bounds for exact recovery of structured sparse signals. arXiv preprint. arXiv:1106.4355 (2011)
49. Raskutti, G., Wainwright, M.J., Yu, B.: Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **99**, 2241–2259 (2010)
50. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52**(3), 471–501 (2010)
51. Richard, E., Savalle, P.-A., Vayatis, N.: Estimation of simultaneously sparse and low rank matrices. arXiv preprint. arXiv:1206.6474 (2012)
52. Rockafellar, R.T.: Convex Analysis, vol. 28. Princeton University Press, Princeton (1997)
53. Stigler, S.M.: Gauss and the invention of least squares. *Ann. Stat.* **9**, 465–474 (1981)
54. Stojnic, M.: Block-length dependent thresholds in block-sparse compressed sensing. arXiv preprint. arXiv:0907.3679 (2009)
55. Stojnic, M.: Various thresholds for ℓ_1 -optimization in compressed sensing. arXiv preprint. arXiv:0907.3666 (2009)
56. Stojnic, M.: A framework to characterize performance of LASSO algorithms. arXiv preprint. arXiv:1303.7291 (2013)
57. Stojnic, M.: A rigorous geometry-probability equivalence in characterization of ℓ_1 -optimization. arXiv preprint. arXiv:1303.7287 (2013)
58. Stojnic, M., Parvaresh, F., Hassibi, B.: On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Trans. Signal Process.* **57**(8), 3075–3085 (2009)
59. Taylor, J., et al.: The geometry of least squares in the 21st century. *Bernoulli* **19**(4), 1449–1464 (2013)
60. Thrampoulidis, C., Oymak, S., Hassibi, B.: Simple error bounds for regularized noisy linear inverse problems. In: *2014 IEEE International Symposium on Information Theory (ISIT)*, pp. 3007–3011. IEEE (2014)
61. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. Ser. B (Methodological)* **58**, 267–288 (1996)
62. Vandenberghe, L.: Subgradients. <http://www.seas.ucla.edu/~vandenbe/236C/lectures/subgradients.pdf> (2013)
63. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. arXiv preprint. arXiv:1011.3027 (2010)

64. Wainwright, M.J.: Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55**(5), 2183–2202 (2009)
65. Wright, J., Ganesh, A., Min, K., Ma, Y.: Compressive principal component pursuit. *Inf. Infer.* **2**(1), 32–68 (2013)
66. Zhao, P., Yu, B.: On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541–2563 (2006)

Chapter 5

The Quest for Optimal Sampling: Computationally Efficient, Structure-Exploiting Measurements for Compressed Sensing

Ben Adcock, Anders C. Hansen, and Bogdan Roman

Abstract An intriguing phenomenon in many instances of compressed sensing is that the reconstruction quality is governed not just by the overall sparsity of the object to recover, but also on its structure. This chapter is about understanding this phenomenon, and demonstrating how it can be fruitfully exploited by the design of suitable sampling strategies in order to outperform more standard compressed sensing techniques based on random matrices.

5.1 Introduction

Compressed sensing concerns the recovery of signals and images from a small collection of linear measurements. It is now a substantial area of research, accompanied by a mathematical theory that is rapidly reaching a mature state. Applications of compressed sensing can roughly be divided into two areas. First, *type I* problems, where the physical device imposes a particular type of measurements. This is the case in numerous real-world problems, including medical imaging (e.g., Magnetic Resonance Imaging (MRI) and Computerized Tomography (CT)), electron microscopy, seismic tomography, and radar. Second, *type II* problems, where the sensing mechanism allows substantial freedom to design the measurements so as to improve the reconstructed image or signal. Applications include compressive imaging and fluorescence microscopy.

B. Adcock (✉)

Department of Mathematics, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6
e-mail: ben_adcock@sfu.ca

A.C. Hansen • B. Roman

DAMTP, Centre for Mathematical Sciences, University of Cambridge,
Wilberforce Rd, Cambridge CB3 0WA, UK
e-mail: a.hansen@damtp.cam.ac.uk; abr28@cam.ac.uk

This chapter is devoted to the role of structured sparsity in both classes of problems. It is well known that the standard sparsifying transforms of compressed sensing, i.e. wavelets and their various generalizations, not only give sparse coefficients, but that there is a distinct structure to this sparsity: wavelet coefficients of natural signals and images are far more sparse at fine scales than at coarse scales. For type I problems, recent developments [2, 4] have shown that this structure plays a key role in the observed reconstruction quality. Moreover, the optimal subsampling strategy depends crucially on this structure. We recap this work in this chapter.

Since structure is vitally important in type I problems, it is natural to ask whether or not it can be exploited to gain improvements in type II problems. In this chapter we answer this question in the affirmative. We show that appropriately designed structured sensing matrices can successfully exploit structure. Doing so leads to substantial gains over classical approaches in type II problems, based on convex optimization with universal, random Gaussian or Bernoulli measurements, as well as more recent structure-exploiting algorithms—such as model-based compressed sensing [6], TurboAMP [53], and Bayesian compressed sensing [32, 33]—which incorporate structure using bespoke recovery algorithms. The matrices we propose, based on appropriately subsampled Fourier/Hadamard transforms, are also computationally efficient, and thus allow for fast reconstructions at high resolutions.

We also review the new theory of compressed sensing introduced in [2, 7] for such structured sensing matrices. The corresponding sensing matrices are highly non-universal and do not satisfy a meaningful Restricted Isometry Property (RIP). Yet, recovery is still possible, and is vastly superior to that obtained from standard RIP matrices. It transpires that the standard RIP is unsuitable if one seeks to exploit structured sparsity by designing appropriate measurements. Thus we consider an alternative that takes such structure into account, known as the *RIP in levels* [7].

5.2 Recovery of wavelet coefficients

In many applications of compressed sensing, we are faced with the problem of recovering an image or signal x , considered as a vector in \mathbb{C}^n or a function in $L^2(\mathbb{R}^d)$, that is sparse or compressible in an orthonormal basis of wavelets. If $\Phi \in \mathbb{C}^{n \times n}$ or $\Phi \in \mathcal{B}(L^2(\mathbb{R}^d), \ell_2(\mathbb{N}))$ (the set of bounded linear operators) is the corresponding sparsifying transformation, then we write $x = \Phi c$, where $c \in \mathbb{C}^n$ or $c \in \ell_2(\mathbb{N})$ is the corresponding sparse or compressible vector of coefficients. Given a sensing operator $A \in \mathbb{C}^{m \times n}$ or $A \in \mathcal{B}(L^2(\mathbb{R}^d), \mathbb{C}^m)$ and noisy measurements $y = Ax + e$ with $\|e\|_2 \leq \eta$, the usual approach is to solve the ℓ_1 -minimization problem:

$$\min_{z \in \mathbb{C}^n} \|\Phi^* z\|_1 \quad \text{s.t.} \quad \|y - Az\|_2 \leq \eta. \quad (5.1)$$

or

$$\inf_{z \in L^2(\mathbb{R}^d)} \|\Phi^* z\|_1 \quad \text{s.t.} \quad \|y - Az\|_2 \leq \eta. \quad (5.2)$$

Throughout this chapter, we shall denote a minimizer of (5.1) or (5.2) as \hat{x} . Note that (5.2) must be discretized in order to be solved numerically, and this can be done by restricting the minimization to be taken over a finite-dimensional space spanned by the first n wavelets, where n is taken sufficiently large [1].

As mentioned, compressed sensing problems arising in applications can be divided into two classes:

- I. *Imposed sensing operators.* The operator A is specified by the practical device and is therefore considered fixed. This is the case in MRI—where A arises by subsampling the Fourier transform [44, 45]—as well as other examples, including X-ray CT (see [17] and the references therein), radar [35], electron microscopy [9, 41], seismic tomography [43], and radio interferometry [60].
- II. *Designed sensing operators.* The sensing mechanism allows substantial freedom to design A so as to improve the compressed sensing reconstruction. Some applications belonging to this class are *compressive imaging*, e.g. the single-pixel camera [24] and the more recent lensless camera [37], and compressive fluorescence microscopy [54]. In these applications A is assumed to take binary values (typically $\{-1, 1\}$), yet, as we will see later, this is not a significant practical restriction.

As stated, the purpose of this chapter is to show that insight gained from understanding the application of compressed sensing to type I problems leads to more effective strategies for type II problems.

5.2.1 Universal sensing matrices

Let us consider type II problems. In finite dimensions, the traditional compressed sensing approach has been to construct matrices A possessing the following two properties. First, they should satisfy the *Restricted Isometry Property (RIP)*. Second, they should be *universal*. That is, if $\Phi \in \mathbb{C}^{n \times n}$ is an arbitrary isometry, then $A\Phi$ also satisfies the RIP of the same order as A . Subject to these conditions, a typical result in compressed sensing is as follows (see [27], for example): if A satisfies the RIP of order $2k$ with constant $\delta_{2k} < 4/\sqrt{41}$ then, for any $x \in \mathbb{C}^n$, we have

$$\|x - \hat{x}\|_2 \leq C \frac{\sigma_k(\Phi^* x)_1}{\sqrt{k}} + D\eta, \quad (5.3)$$

where \hat{x} is any minimizer of (5.1), C and D are positive constants depending only on δ_{2k} and, for $c \in \mathbb{C}^n$,

$$\sigma_k(c)_1 = \inf_{z \in \Sigma_k} \|c - z\|_1, \quad \Sigma_k = \{z \in \mathbb{C}^n : \|z\|_0 \leq k\}.$$

Hence, x is recovered exactly up to the noise level η and the error $\sigma_k(c)_1$ of the best approximation of $c = \Phi^*x$ with a k -sparse vector. Since A is universal, one has complete freedom to choose the sparsifying transformation Φ so as to minimize the term $\sigma_k(\Phi^*x)_1$ for the particular signal x under consideration.

Typical examples of universal sensing matrices A arise from random ensembles. In particular, Gaussian or Bernoulli random matrices (with the latter having the advantage of being binary) both have this property with high probability whenever m is proportional to k times by a log factor. Correspondingly, such matrices are often thought of as ‘optimal’ matrices for compressed sensing.

Remark 1. A significant drawback of random ensembles, however, is that they lead to dense and unstructured matrices. Storage and the lack of fast transforms render them impractical for all but small problem sizes. To overcome this, various structured random matrices have also been developed and studied (see, for example, [5, 28, 34, 36–38]). Often these admit fast, $\mathcal{O}(n \log n)$ transforms. However, the best known theoretical RIP guarantees are usually larger than for (sub)Gaussian random matrices [27].

5.2.2 Sparsity structure dependence and the flip test

Since it will become important later, we now describe a quick and simple test, which we call the *flip test*, to investigate the presence or absence of an RIP. Success of this test suggests the existence of an RIP and failure demonstrates its lack.

Let $A \in \mathbb{C}^{m \times n}$ be a sensing matrix, $x \in \mathbb{C}^n$ an image and $\Phi \in \mathbb{C}^{n \times n}$ a sparsifying transformation. Recall that sparsity of the vector $c = \Phi^*x$ is unaffected by permutations. Thus, let us define the flipped vector

$$P(c) = c' \in \mathbb{C}^n, \quad c'_i = c_{n+1-i}, \quad i = 1, \dots, n,$$

and using this, we construct the flipped image $x' = \Phi c'$. Note that, by construction, we have $\sigma_k(c)_1 = \sigma_k(c')_1$. Now suppose we perform the usual compressed sensing reconstruction (5.1) on both x and x' , giving approximations $\hat{x} \approx x$ and $\check{x} \approx x'$. We now wish to reverse the flipping operation. Thus, we compute $\check{x} = \Phi P(\Phi^* \check{x})$, which gives a second approximation to the original image x .

This test provides a simple way to investigate whether or not the RIP holds. To see why, suppose that A satisfies the RIP. Then by construction, we have that

$$\|x - \hat{x}\|_2, \|x - \check{x}\|_2 \leq C \frac{\sigma_k(\Phi^* x)_1}{\sqrt{k}} + D\eta.$$

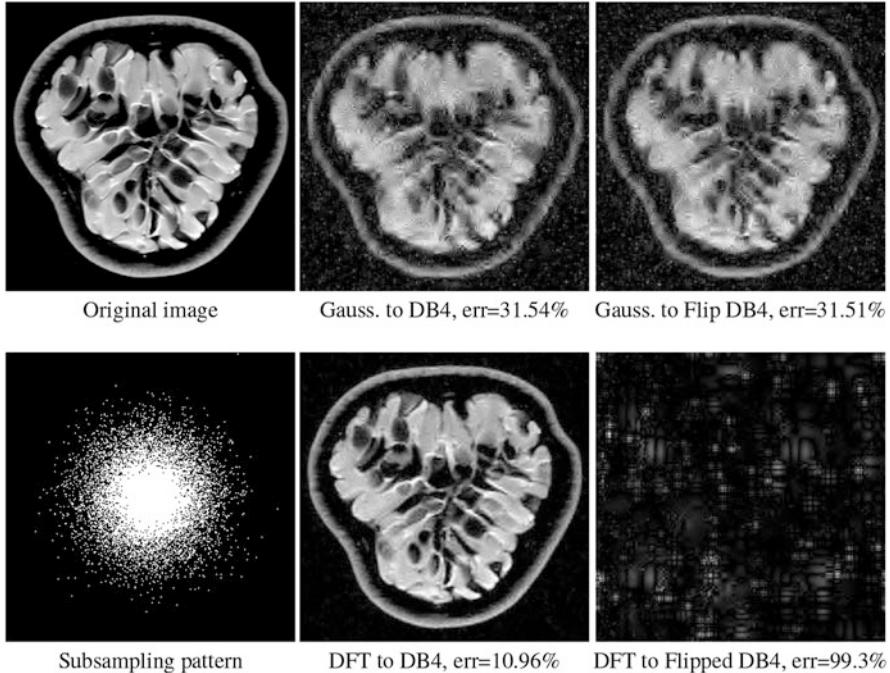


Fig. 5.1 Recovery of an MRI image of a passion fruit using (5.1) from $m = 8192$ samples at $n = 256 \times 256$ resolution (i.e., a 12.5 % subsampling rate) with Daubechies-4 wavelets as the sparsifying transform Φ . Top row: Flip test for Gaussian random measurements. Bottom row: Flip test for subsampled DFT measurements taken according to the subsampling pattern shown in the bottom left panel. The flip test suggests that the RIP holds for random sensing matrices (top row), but that there is no RIP for structured sensing matrices with structured sampling (bottom row).

Hence both \hat{x} and \check{x} should recover x equally well. In the top row of Figure 5.1 we present the result of the flip test for a Gaussian random matrix. As is evident, the reconstructions \hat{x} and \check{x} are comparable, thus indicating the RIP.

Having considered a type II problem setup, let us now examine the flip test for a type I problem. As discussed, in applications such as MRI, X-ray CT, radio interferometry, etc., the matrix A is imposed by the physical sensing device and arises from subsampling the rows of the DFT matrix $F \in \mathbb{C}^{n \times n}$.¹ Whilst one often has some freedom to choose which rows to sample (corresponding to selecting particular frequencies at which to take measurements), one cannot change the matrix F .

¹In actual fact, the sensing device takes measurements of the *continuous* Fourier transform of a function $x \in L^2(\mathbb{R}^d)$. As discussed in [1, 4], modelling continuous Fourier measurements as discrete Fourier measurements can lead to inferior reconstructions, as well as inverse crimes. To avoid this, one must consider an infinite-dimensional compressed sensing approach, as in (5.2). See [2, 4] for details, as well as [29] for implementation in MRI. For simplicity, we shall continue to work with the finite-dimensional model in the remainder of this chapter.

It is well known that in order to ensure a good reconstruction, one cannot subsample the DFT uniformly at random (recall that the sparsifying transform is a wavelet basis), but rather one must sample randomly according to an appropriate nonuniform density [2, 15, 45, 59]. See the bottom left panel of Figure 5.1 for an example of a typical density. As can be seen in the next panel, by doing so one achieves an excellent reconstruction.

Since an RIP has been recently shown for this matrix (in the case of the 2D Haar transform) [39] one may be tempted to think it is this property which underpins the high reconstruction quality seen in Figure 5.1 and witnessed empirically in many real-world CS MRI experiments. However, the result of the flip test in the bottom right panel clearly demonstrates that this cannot be the case. In particular, the ordering of the wavelet coefficients plays a crucial role in the reconstruction quality; a phenomenon which is impossible in RIP theory setting. Although [39] shows that the RIP can be guaranteed for this matrix, doing so requires an unrealistically large amount of samples. In particular, the results of the test performed imply that ensuring an RIP will necessarily require more measurements than suffice to recover the original, unflipped image to high accuracy. Later in this chapter, by introducing a new analytical framework that dispenses with the RIP and takes the structure of wavelet coefficients into account, we will explain precisely why such good recovery is possible when taking substantially fewer measurements than is needed to ensure an RIP.

Note that the flip test in Figure 5.1 also highlights another important phenomenon: namely, the effectiveness of the subsampling strategy depends on the sparsity structure of the image. In particular, two images with the same total sparsity (the original x and the flipped x') result in wildly different errors when the same sampling pattern is used. Thus also we conclude that there is no one optimal sampling strategy for all sparse vectors of wavelet coefficients.

Let us also note that the same conclusions of the flip test hold when (5.1) with wavelets is replaced by TV-norm minimization:

$$\min_{z \in \mathbb{C}^n} \|z\|_{TV} \quad \text{s.t.} \quad \|y - Az\|_2 \leq \eta. \quad (5.4)$$

Recall that $\|x\|_{TV} = \sum_{i,j} \|\nabla x(i, j)\|_2$, where we have $\nabla x(i, j) = \{D_1 x(i, j), D_2 x(i, j)\}$, $D_1 x(i, j) = x(i+1, j) - x(i, j)$, $D_2 x(i, j) = x(i, j+1) - x(i, j)$. In the experiment leading to Figure 5.2, we chose an image $x \in [0, 1]^{N \times N}$, and then built a different image x' from the gradient of x so that $\{\|\nabla x'(i, j)\|_2\}$ is a permutation of $\{\|\nabla x(i, j)\|_2\}$ for which $x' \in [0, 1]^{N \times N}$. Thus, the two images have the same “TV sparsity” and the same TV norm. In Figure 5.2 we demonstrate how the errors differ substantially for the two images when using the same sampling pattern. Note also how the improvement depends both on the TV sparsity structure and on the subsampling pattern. Analysis of this phenomenon is work in progress. In the remainder of this chapter we will focus on structured sparsity for wavelets.

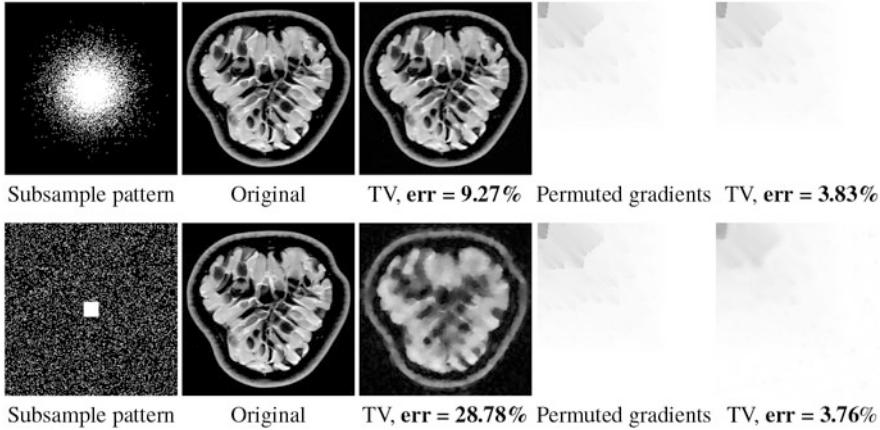


Fig. 5.2 TV recovery (5.4) from $m = 16384$ DFT samples at $n = 512 \times 512$ (6.25 % subsampling rate). The percentage error values listed are relative errors in the image domain. The *Permuted gradients* image was constructed as follows. The matrix of gradient vectors of the *Original image* was first computed, and then the ordering of the entries was permuted and signs of the vectors changed. The *Permuted gradients* image was then constructed with gradient equal to these new vectors, giving a new image whose TV norm and gradient sparsity is the same as that of the original image. The large error differences confirm that, much like the flip test for wavelet coefficients, the sparsity structure matters for TV reconstructions as well.

5.2.3 Structured sparsity

One of the foundational results of nonlinear approximation is that, for natural images and signals, the best k -term approximation error in a wavelet basis decays rapidly in k [21, 46]. In other words, wavelet coefficients are approximately k -sparse. However, wavelet coefficients possess far more structure than mere sparsity. Recall that a wavelet basis for $L^2(\mathbb{R}^d)$ is naturally partitioned into dyadic scales. Let $0 = M_0 < M_1 < \dots < \infty$ be such a partition, and note that $M_{l+1} - M_l = \mathcal{O}(2^l)$ in one dimension and $M_{l+1} - M_l = \mathcal{O}(4^l)$ in two dimensions. If $x = \Phi c$, let $c^{(l)} \in \mathbb{C}^{M_l - M_{l-1}}$ denote the wavelet coefficients of x at scale $l = 1, 2, \dots$, so that $c = (c^{(1)} | c^{(2)} | \dots)^\top$. Let $\varepsilon \in (0, 1]$ and define the global sparsity k and the sparsity at the l^{th} level k_l as follows:

$$k = k(\varepsilon) = \min \left\{ n : \left\| \sum_{i \in \mathcal{M}_n} c_i \varphi_i \right\| \geq \varepsilon \left\| \sum_{j=1}^{\infty} c_j \varphi_j \right\| \right\}, \quad (5.5)$$

$$k_l = k_l(\varepsilon) = |\mathcal{M}_{k(\varepsilon)} \cap \{M_{k-1} + 1, \dots, M_k\}|,$$

where \mathcal{M}_n is the set of indices of the largest n coefficients in absolute value and $|\cdot|$ is the set cardinality. Figure 5.3 reveals that there is a distinct sparsity structure, in particular, we have so-called *asymptotic sparsity*. That is

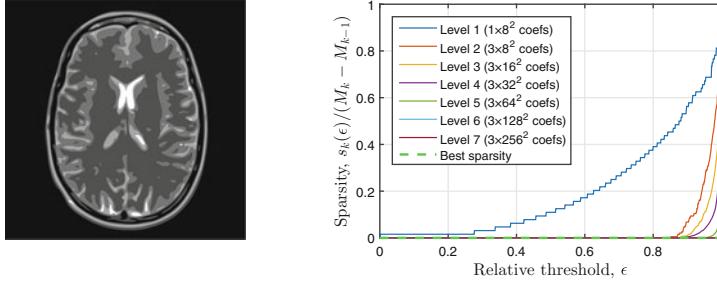


Fig. 5.3 Left: GLPU phantom [30]. Right: Relative sparsity of Daubechies-4 coefficients. Here the levels correspond to wavelet scales and $k_l(\epsilon)$ is given by (5.5). Each curve shows the relative sparsity at level l as a function of ϵ . The decreasing nature of the curves for increasing l confirms (5.6).

$$k_l/(M_l - M_{l-1}) \rightarrow 0, \quad l \rightarrow \infty. \quad (5.6)$$

Put simply, wavelet coefficients are much more sparse at fine scales than they are at coarse scales.

Note that this observation is by no means new: it is a simple consequence of the dyadic scaling of wavelets, and is a crucial step towards establishing the nonlinear approximation result mentioned above. However, given that wavelet coefficients always exhibit such structure, one may ask the following question. For type II problems, are the traditional sensing matrices of compressed sensing—which, as shown by the flip test (top row of Figure 5.1), recover all sparse vectors of coefficients equally well, regardless of ordering—optimal when wavelets are used as the sparsifying transform? It has been demonstrated in Figure 5.1 (bottom row) that structure plays a key role in type I compressed sensing problems. Leveraging this insight, in the next section we show that significant gains are possible for type II problems in terms of both the reconstruction quality and computational efficiency when the sensing matrix A is designed specifically to take advantage of such inherent structure.

Remark 2. Asymptotic sparsity is by no means limited to wavelets. A similar property holds for other types of -lets, including curvelets [12, 13], contourlets [22, 48], or shearlets [19, 20, 40] (see [4] for examples of Figure 5.3 based on these transforms). More generally, *any* sparsifying transform that arises (possibly via discretization) from a countable basis or frame will typically exhibit asymptotic sparsity.

Remark 3. Wavelet coefficients (and their various generalizations) actually possess far more structure than the asymptotic sparsity (5.6). Specifically, they tend to live on rooted, connected trees [18]. There are a number of existing algorithms which seek to exploit such structure within a compressed sensing context. We shall discuss these further in Section 5.5.

5.3 Efficient sensing matrices for structured sparsity

For simplicity, we now consider the finite-dimensional setting, although the arguments extend to the infinite-dimensional case [2]. Suppose that $\Phi \in \mathbb{C}^{n \times n}$ corresponds to a wavelet basis so that $c = \Phi^*x$ is not just k -sparse, but also asymptotically sparse with the sparsities k_1, \dots, k_r within the wavelet scales being known. As before, write $c^{(l)}$ for set of coefficients at scale l . Considering type II problems, we now seek a sensing matrix A with as few rows m as possible that exploits this local sparsity information.

5.3.1 Block-diagonality and structured Fourier/Hadamard sampling

For the l^{th} level, suppose that we assign a total of $m_l \in \mathbb{N}$ rows of A in order to recover the k_l nonzero coefficients of $c^{(l)}$. Note that $m = \sum_{l=1}^r m_l$. Consider the product $B = A\Phi$ of the sensing matrix A and the sparsifying transform Φ . Then there is a natural decomposition of B into blocks $\{B_{jl}\}_{j,l=1}^r$ of size $m_j \times (M_l - M_{l-1})$, where each block corresponds to the m_j measurements of the $M_l - M_{l-1}$ wavelet functions at the l^{th} scale.

Suppose it were possible to construct a sensing matrix A such that (i) B was block diagonal, i.e. $B_{jl} = 0$ for $j \neq l$, and (ii) the diagonal blocks B_{ll} satisfied an RIP of order $2k_l$ whenever m_l was proportional to k_l times by the usual log factor. In this case, one recovers the coefficients $c^{(l)}$ at scale l from near-optimal numbers of measurements using the usual reconstruction (5.1).

This approach, originally proposed by Donoho [23] and Tsaig & Donoho [55] under the name of ‘multiscale compressed sensing,’ allows for structure to be exploited within compressed sensing. Similar ideas were also pursued by Romberg [51] within the context of compressive imaging. Unfortunately, it is normally impossible to design an $m \times N$ matrix A such that $B = A\Phi$ is exactly block diagonal. Nevertheless, the notion of block-diagonality provides insight into better designs for A than purely random ensembles. To proceed, we relax the requirement of strict block-diagonality, and instead ask whether there exist practical sensing matrices A for which B is approximately block-diagonal whenever the sparsifying transform Φ corresponds to wavelets. Fortunately, the answer to this question is affirmative: as we shall explain next, and later confirm with a theorem, approximate block-diagonality can be ensured whenever A arises by appropriately subsampling the rows of the Fourier or Hadamard transform. Recalling that the former arises naturally in type I problems, this points towards the previously claimed conclusion that new insight brought about by studying imposed sensing matrices leads to better approaches for the type II problem. We note that some recent work has also considered nonuniform density sampling of the Hadamard transform [34] in compressive imaging, albeit from a sparsity (as opposed to structured sparsity) viewpoint and without analysis.

Let $F \in \mathbb{C}^{n \times n}$ be either the discrete Fourier or discrete Hadamard transform. Let $\Omega \subseteq \{1, \dots, n\}$ be an index set of size $|\Omega| = m$. We now consider choices for A of the form $A = P_\Omega F$, where $P_\Omega \in \mathbb{C}^{m \times n}$ is the restriction operator that selects rows of F whose indices lie in Ω . We seek an Ω that gives the desired block-diagonality. To do this, it is natural to divide up Ω itself into r disjoint blocks

$$\Omega = \Omega_1 \cup \dots \cup \Omega_r, \quad |\Omega_l| = m_l,$$

where the l^{th} block $\Omega_l \subseteq \{N_{l-1}, \dots, N_l\}$ corresponds to the m_l samples required to recover the k_l nonzero coefficients at scale l . Here the parameters $0 = N_0 < N_1 < \dots < N_r = n$ are appropriately chosen and delineate frequency bands from which the m_l samples are taken.

In Section 5.3.3, we explain why this choice of A works, and in particular, how to choose the sampling blocks Ω_l . In order to do this, it is first necessary to recall the notion of incoherent bases.

5.3.2 Incoherent bases and compressed sensing

Besides random ensembles, a common approach in compressed sensing is to design sensing matrices using orthonormal systems that are incoherent with the particular choice of sparsifying basis Φ [14, 15, 27]. Let $\Psi \in \mathbb{C}^{n \times n}$ be an orthonormal basis of \mathbb{C}^n . The (mutual) coherence of Φ and Ψ is the quantity

$$\mu = \mu(\Psi^* \Phi) = \max_{i,j=1,\dots,n} |(\Psi^* \Phi)_{i,j}|^2.$$

We say Ψ and Φ are *incoherent* if $\mu(\Psi, \Phi) \leq a/n$ for some $a \geq 1$ independent of n . Given such a Ψ , one constructs the sensing matrix $A = P_\Omega \Psi$, where $\Omega \subseteq \{1, \dots, N\}$, $|\Omega| = m$ is chosen uniformly at random. A standard result gives that a k -sparse signal x in the basis Φ is recovered exactly with probability at least $1 - p$, provided

$$m \geq Ck \log(1 + p^{-1}) \log(n),$$

for some universal constant $C > 0$ [27]. As an example, consider the Fourier basis $\Psi = F$. This is incoherent with the canonical basis $\Phi = I$ with optimally small constant $a = 1$. Fourier matrices subsampled uniformly at random are efficient sensing matrices for signals that are themselves sparse.

However, the Fourier matrix is not incoherent with a wavelet basis: $\mu(F, \Phi) = \mathcal{O}(1)$ as $n \rightarrow \infty$ for any orthonormal wavelet basis [2]. Nevertheless, Fourier samples taken within appropriate frequency bands (i.e., not uniformly at random) are *locally* incoherent with wavelets in the corresponding scales. This observation, which we demonstrate next, explains the success of the sensing matrix A constructed in the previous subsection for an appropriate choice of $\Omega_1, \dots, \Omega_r$.

5.3.3 Local incoherence and near block-diagonality of Fourier measurements with wavelets

For expository purposes, we consider the case of one-dimensional Haar wavelets. We note however that the arguments generalize to arbitrary compactly supported orthonormal wavelets, and to the infinite-dimensional setting where the unknown image x is a function. See Section 5.4.3.

Let $j = 0, \dots, r-1$ (for convenience we now index from 0 to $r-1$, as opposed to 1 to r) be the scale and $p = 0, \dots, 2^j - 1$ the translation. The Haar basis consists of the functions $\{\psi\} \cup \{\phi_{j,p}\}$, where $\psi \in \mathbb{C}^n$ is the normalized scaling function and $\phi_{j,p}$ are the scaled and shifted versions of the mother wavelet $\phi \in \mathbb{C}^n$. It is a straightforward, albeit tedious, exercise to show that

$$|\mathcal{F}\phi_{l,p}(\omega)| = 2^{l/2-r+1} \frac{|\sin(\pi\omega/2^{l+1})|^2}{|\sin(\pi\omega/2^r)|} \lesssim 2^{l/2} \frac{|\sin(\pi\omega/2^{l+1})|^2}{|\omega|}, \quad |\omega| < 2^r,$$

where \mathcal{F} denotes the DFT [3]. This suggests that the Fourier transform $\mathcal{F}\phi_{l,p}(\omega)$ is large when $\omega \approx 2^l$, yet smaller when $\omega \approx 2^j$ with $j \neq l$. Hence we should separate frequency space into bands of size roughly 2^j .

Let $F \in \mathbb{C}^{n \times n}$ be the DFT matrix with rows indexed from $-n/2 + 1$ to $n/2$. Following an approach of [15], we now divide these rows into the following disjoint frequency bands

$$W_0 = \{0, 1\}, \quad W_j = \{-2^j + 1, \dots, -2^{j-1}\} \cup \{2^{j-1} + 1, \dots, 2^j\}, \quad j = 0, \dots, r-1.$$

With this to hand, we now define Ω_j to be a subset of W_j of size $|\Omega_j| = m_j$ chosen uniformly at random. Thus, the overall sensing matrix $A = P_\Omega F$ takes measurements of the signal x by randomly drawing m_j samples of its Fourier transform within the frequency bands W_j .

Having specified Ω , let us note that the matrix $U = F\Phi$ naturally divides into blocks, with the rows corresponding to the frequency bands W_l and the columns corresponding to the wavelet scales. Write $U = \{U_{jl}\}_{j,l=0}^{r-1}$ where $U_{jl} \in \mathbb{C}^{2^j \times 2^l}$. For compressed sensing to succeed in this setting, we require two properties. First, the diagonal blocks U_{jj} should be incoherent, i.e. $\mu(U_{jj}) = \mathcal{O}(2^{-j})$. Second, the coherences $\mu(U_{jl})$ of the off-diagonal blocks U_{jl} should be appropriately small in comparison with $\mu(U_{jj})$. These two properties are demonstrated in the following lemma:

Lemma 1. *We have $\mu(U_{jj}) \lesssim 2^{-j}$ and, in general,*

$$\mu(U_{jl}) \lesssim \mu(U_{jj}) 2^{-|j-l|}, \quad j, l = 0, \dots, r-1$$

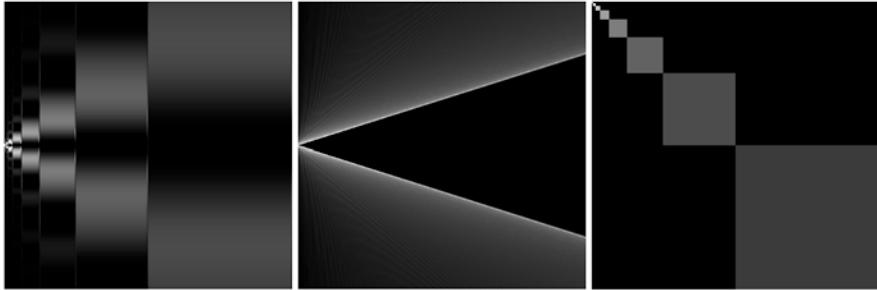


Fig. 5.4 The absolute values of the matrix $U = F\Phi$, where F is the discrete Fourier transform (left and middle) or the Hadamard transform (right) and the sparsifying transform Φ corresponds to Haar wavelets (left and right) or Legendre polynomials (middle). Light colors correspond to large values and dark colors to small values.

Hence U is approximately block diagonal, with exponential decay away from the diagonal blocks. Fourier measurements subsampled according to the above strategy are therefore ideally suited to recover structured sparse wavelet coefficients.²

The left panel of Figure 5.4 exhibits this decay by plotting the absolute values of the matrix U . In the right panel, we also show a similar result when the Fourier transform is replaced by the Hadamard transform. This is an important case, since the measurement matrix is binary. The middle panel of the figure shows the U matrix when Legendre polynomials are used as the sparsifying transform, as is sometimes the case for smooth signals. It demonstrates that diagonally-dominated coherence is not just a phenomenon associated with wavelets.

Having identified a measurement matrix to exploit structured sparsity, let us demonstrate its effectiveness. In Figure 5.5 we compare these measurements with the case of random Bernoulli measurements (this choice was made over random Gaussian measurements because of storage issues). Note that in this and later experiments we use the pattern proposed in [50] for the multilevel subsampling strategy, which has been shown empirically to deliver good all-round performance. As is evident, at all resolutions we see a significant advantage, since the former strategy exploits the structured sparsity. Note that for both approaches, the reconstruction quality is *resolution dependent*: the error decreases as the resolution increases, due to the increasing sparsity of wavelet coefficients at higher resolutions. However, because the Fourier/wavelets matrix U is *asymptotically incoherent* (see also Section 5.4.1), it exploits the inherent asymptotic sparsity structure (5.6) of the wavelet coefficients as the resolution increases, and thus gives successively greater improvements over random Bernoulli measurements.

²For brevity, we do not give the proof of this lemma or the later recovery results for Haar wavelets, Theorem 2. Details of the proof can be found in the short note [3].

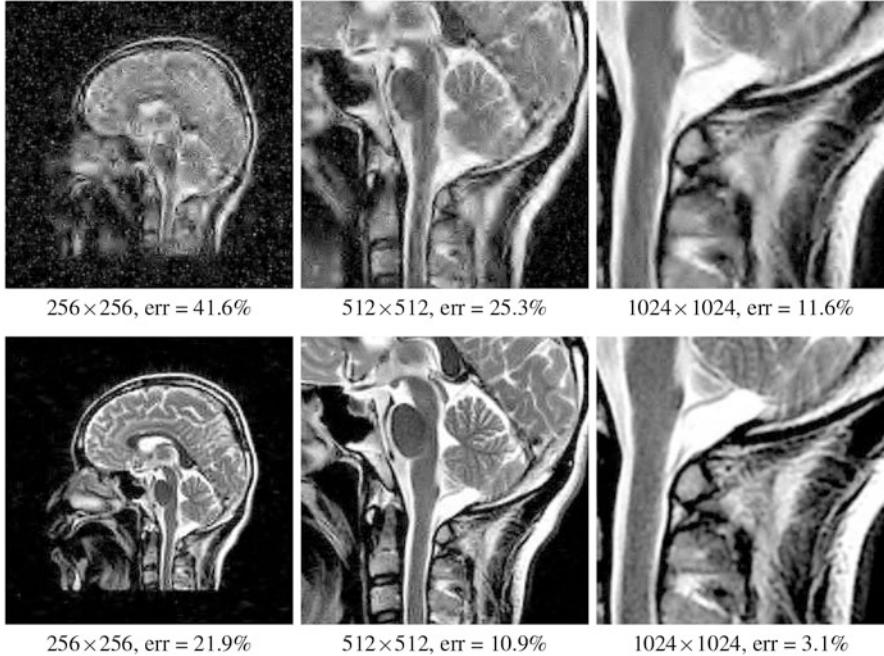


Fig. 5.5 Recovery from 12.5 % measurements using (5.1) with Daubechies-4 wavelets as the sparsifying transform. Top row: Random Bernoulli sensing matrix. Bottom row: Fourier sensing matrix with multilevel subsampling (see Definition 3). All images are 256×256 crops of the original full resolution versions in order to aid the visual comparison.

Remark 4. Besides improved reconstructions, an important feature of this approach is storage and computational time. Since F and Φ have fast, $\mathcal{O}(n \log n)$, transforms, the matrix $A = P_\Omega F \Phi$ does not need to be stored, and the reconstruction (5.1) can be performed efficiently with standard ℓ_1 solvers (we use SPGL1 [56, 57] throughout).

Recall that in type I problems such as MRI, we are constrained by the physics of the device to take Fourier measurements. A rather strange conclusion of Figure 5.5 is the following: compressed sensing actually works better for MRI with the intrinsic measurements, than if one were able to take optimal (in the sense of the standard sparsity-based theory) random (sub)Gaussian measurements. This has practical consequences. In MRI there is actually a little flexibility to design measurements, based on specifying appropriate pulses. By doing this, a number of approaches [31, 42, 49, 52, 61] have been proposed to make MRI measurements closer to uniformly incoherent with wavelets (i.e. similar to random Gaussians). On the other hand, Figure 5.5 suggests that one can obtain great results in practice by appropriately subsampling the unmodified Fourier operator.

Another aspect of certain type I problems such as MRI is that physical constraints dictate that sampling follow continuous trajectories in Fourier space, e.g. radial lines or spirals. The sampling strategy Ω proposed in this section is not of this form, yet it

allows for rigorous recovery guarantees. It is an interesting question for future work to provide such guarantees for realistic contours. For some initial, sparsity-based work in this direction see [8, 11].

5.4 A general framework for compressed sensing based on structured sparsity

Having argued for the particular case of Fourier samples with Haar wavelets, we now describe a general mathematical framework for structured sparsity. This is based on work in [2].

5.4.1 Concepts

We shall work in both the finite- and infinite-dimensional settings, where $U \in \mathbb{C}^{n \times n}$ or $U \in \mathcal{B}(\ell_2(\mathbb{N}))$, respectively. We assume throughout that U is an isometry. This occurs, for example, when $U = \Psi^* \Phi$ for an orthonormal basis Ψ and an orthonormal system Φ , as is the case for the example studied in the previous section: namely, Fourier sampling and a wavelet sparsifying basis. However, framework we present now is valid for an arbitrary isometry U , not just this particular example. We discuss this case further in Section 5.4.3.

We first require the following definitions. In the previous section it was suggested to divide both the sampling strategy and the sparse vector of coefficients into disjoint blocks. We now formalize these notions:

Definition 1 (Sparsity in levels). Let c be an element of either \mathbb{C}^N or $\ell_2(\mathbb{N})$. For $r \in \mathbb{N}$ let $\mathbf{M} = (M_1, \dots, M_r) \in \mathbb{N}^r$ with $1 \leq M_1 < \dots < M_r$ and $\mathbf{k} = (k_1, \dots, k_r) \in \mathbb{N}^r$, with $k_j \leq M_j - M_{j-1}$, $k = 1, \dots, r$, where $M_0 = 0$. We say that c is (\mathbf{k}, \mathbf{M}) -sparse if, for each $j = 1, \dots, r$, the set

$$\Delta_j := \text{supp}(c) \cap \{M_{j-1} + 1, \dots, M_j\},$$

satisfies $|\Delta_j| \leq k_j$. We denote the set of (\mathbf{k}, \mathbf{M}) -sparse vectors by $\Sigma_{\mathbf{k}, \mathbf{M}}$.

This definition allows for differing amounts of sparsity of the vector c in different levels. Note that the levels \mathbf{M} do not necessarily correspond to wavelet scales—for now, we consider a general setting. We also need a notion of best approximation:

Definition 2 ((\mathbf{k}, \mathbf{M})-term approximation). Let c be an element of either \mathbb{C}^N or $\ell_2(\mathbb{N})$. We say that c is (\mathbf{k}, \mathbf{M}) -compressible if $\sigma_{\mathbf{k}, \mathbf{M}}(c)$ is small, where

$$\sigma_{\mathbf{k}, \mathbf{M}}(c)_1 = \inf_{z \in \Sigma_{\mathbf{k}, \mathbf{M}}} \|c - z\|_1. \quad (5.7)$$

As we have already seen for wavelet coefficients, it is often the case that $k_j/(M_j - M_{j-1}) \rightarrow 0$ as $j \rightarrow \infty$. In this case, we say that c is *asymptotically sparse in levels*. However, we stress that this framework does not explicitly require such decay.

We now consider the level-based sampling strategy:

Definition 3 (Multilevel random sampling). Let $r \in \mathbb{N}$, $\mathbf{N} = (N_1, \dots, N_r) \in \mathbb{N}^r$ with $1 \leq N_1 < \dots < N_r$, $\mathbf{m} = (m_1, \dots, m_r) \in \mathbb{N}^r$, with $m_j \leq N_j - N_{j-1}$, $j = 1, \dots, r$, and suppose that

$$\Omega_j \subseteq \{N_{j-1} + 1, \dots, N_j\}, \quad |\Omega_j| = m_j, \quad j = 1, \dots, r,$$

are chosen uniformly at random, where $N_0 = 0$. We refer to the set

$$\Omega = \Omega_{\mathbf{N}, \mathbf{m}} = \Omega_1 \cup \dots \cup \Omega_r.$$

as an (\mathbf{N}, \mathbf{m}) -multilevel sampling scheme.

As discussed in Section 5.3.2, the (infinite) Fourier/wavelets matrix $U = F\Phi$ is globally coherent. However, as shown in Lemma 1, the coherence of its $(j, l)^{\text{th}}$ block is much smaller. We therefore require a notion of local coherence:

Definition 4 (Local coherence). Let U be an isometry of either \mathbb{C}^N or $\ell_2(\mathbb{N})$. If $\mathbf{N} = (N_1, \dots, N_r) \in \mathbb{N}^r$ and $\mathbf{M} = (M_1, \dots, M_r) \in \mathbb{N}^r$ with $1 \leq N_1 < \dots < N_r$ and $1 \leq M_1 < \dots < M_r$ the $(j, l)^{\text{th}}$ local coherence of U with respect to \mathbf{N} and \mathbf{M} is given by

$$\mu_{\mathbf{N}, \mathbf{M}}(j, l) = \sqrt{\mu(P_{N_j}^{N_{j-1}} U P_{M_l}^{M_{l-1}}) \mu(P_{N_j}^{N_{j-1}} U)}, \quad k, l = 1, \dots, r,$$

where $N_0 = M_0 = 0$ and P_b^a denotes the projection matrix corresponding to indices $\{a+1, \dots, b\}$. In the case where U is an operator on $\ell_2(\mathbb{N})$, we also define

$$\mu_{\mathbf{N}, \mathbf{M}}(j, \infty) = \sqrt{\mu(P_{N_j}^{N_{j-1}} U P_{M_{r-1}}^{\perp}) \mu(P_{N_j}^{N_{j-1}} U)}, \quad j = 1, \dots, r.$$

Note that the local coherence $\mu_{\mathbf{N}, \mathbf{M}}(j, l)$ is not just the coherence $\mu(P_{N_j}^{N_{j-1}} U P_{M_l}^{M_{l-1}})$ in the $(j, l)^{\text{th}}$ block. For technical reasons, one requires the product of this and the coherence $\mu(P_{N_j}^{N_{j-1}} U)$ in the whole j^{th} row block.

We remark also that our definition of local coherence (which applies to arbitrary isometries U , not just the previously discussed Fourier/Haar matrix) is different to other notions of local coherence in the literature. In [39], the authors define a local coherence of a matrix U in the j^{th} row (as opposed to row block) to be the maximum of its entries in that row. Using this, they prove recovery guarantees based on the RIP and the global sparsity k . Unfortunately, as demonstrated by the flip test, such notions of local coherence fail to capture the critical role of sparsity structure in the reconstruction, as well as the key dependence of the optimal subsampling strategy

on the structure. Conversely, our definition of local coherence takes into account the sparsity levels. As we will see in Section 5.4.2, it allows one to establish recovery guarantees that are consistent with the flip test and thus properly explain the key role played by structure.

Recall that in practice (see Figure 5.4), the local coherence often decays along the diagonal blocks and in the off-diagonal blocks. Loosely speaking, we say that the matrix U is *asymptotically incoherent* in this case.

In Section 5.3.3 we argued that the Fourier/wavelets matrix was nearly block-diagonal. In our theorems, we need to account for the off-diagonal terms. To do this in the general setup, we require a notion of a relative sparsity:

Definition 5 (Relative sparsity). Let U be an isometry of either \mathbb{C}^N or $\ell_2(\mathbb{N})$. For $\mathbf{N} = (N_1, \dots, N_r) \in \mathbb{N}^r$, $\mathbf{M} = (M_1, \dots, M_r) \in \mathbb{N}^r$ with $1 \leq N_1 < \dots < N_r$ and $1 \leq M_1 < \dots < M_r$, $\mathbf{k} = (k_1, \dots, k_r) \in \mathbb{N}^r$ and $j = 1, \dots, r$, the j^{th} relative sparsity is given by

$$K_j = K_j(\mathbf{N}, \mathbf{M}, \mathbf{k}) = \max_{z \in \Sigma_{\mathbf{k}, \mathbf{M}}, \|z\|_\infty \leq 1} \|P_{N_j}^{N_{j-1}} U z\|^2,$$

where $N_0 = M_0 = 0$.

The relative sparsities K_j take into account *interferences* between different sparsity level caused by the non-block diagonality of U .

5.4.2 Main theorem

Given the matrix/operator U and a multilevel sampling scheme Ω , we now consider the solution of the convex optimization problem

$$\min_{z \in \mathbb{C}^n} \|z\|_1 \quad \text{s.t.} \quad \|P_\Omega y - P_\Omega U z\|_2 \leq \eta, \quad (5.8)$$

where $y = Uc + e$, $\|e\|_2 \leq \eta$. Note that if $U = \Psi^* \Phi$, $x = \Phi c$ is the signal we wish to recover and \hat{c} is a minimizer of (5.8) then this gives the approximation $\hat{x} = \Phi \hat{c}$ to x .

Theorem 1. Let $U \in \mathbb{C}^{N \times N}$ be an isometry and $c \in \mathbb{C}^N$. Suppose that $\Omega = \Omega_{\mathbf{N}, \mathbf{m}}$ is a multilevel sampling scheme, where $\mathbf{N} = (N_1, \dots, N_r) \in \mathbb{N}^r$, $N_r = n$, and $\mathbf{m} = (m_1, \dots, m_r) \in \mathbb{N}^r$. Let $\varepsilon \in (0, e^{-1}]$ and suppose that (\mathbf{k}, \mathbf{M}) , where $\mathbf{M} = (M_1, \dots, M_r) \in \mathbb{N}^r$, $M_r = n$, and $\mathbf{k} = (k_1, \dots, k_r) \in \mathbb{N}^r$, are any pair such that the following holds:

(i) We have

$$1 \gtrsim \frac{N_j - N_{j-1}}{m_j} \log(\varepsilon^{-1}) \left(\sum_{l=1}^r \mu_{\mathbf{N}, \mathbf{M}}(j, l) k_l \right) \log(n), \quad j = 1, \dots, r. \quad (5.9)$$

(ii) We have $m_j \gtrsim \hat{m}_j \log(\varepsilon^{-1}) \log(n)$, where \hat{m}_j is such that

$$1 \gtrsim \sum_{j=1}^r \left(\frac{N_j - N_{j-1}}{\hat{m}_j} - 1 \right) \mu_{\mathbf{N}, \mathbf{M}}(j, l) \tilde{k}_j, \quad l = 1, \dots, r, \quad (5.10)$$

for all $\tilde{k}_1, \dots, \tilde{k}_r \in (0, \infty)$ satisfying

$$\tilde{k}_1 + \dots + \tilde{k}_r \leq k_1 + \dots + k_r, \quad \tilde{k}_j \leq K_j(\mathbf{N}, \mathbf{M}, \mathbf{k}).$$

Suppose that $\hat{c} \in \mathbb{C}^N$ is a minimizer of (5.8) with $\eta = \tilde{\eta}/\sqrt{E}$, where $E = \max_{j=1, \dots, r} \{(N_j - N_{j-1})/m_j\}$. Then, with probability exceeding $1 - k\varepsilon$, where $k = k_1 + \dots + k_r$, we have that

$$\|c - \hat{c}\|_2 \leq C \left(\tilde{\eta} \left(1 + D\sqrt{k} \right) + \sigma_{\mathbf{k}, \mathbf{M}}(c)_1 \right), \quad (5.11)$$

for some constant C , where $\sigma_{\mathbf{k}, \mathbf{M}}(c)_1$ is as in (5.7) and $D = 1 + \frac{\sqrt{\log_2(6\varepsilon^{-1})}}{\log_2(4En\sqrt{s})}$. If $m_j = N_j - N_{j-1}$, $j = 1, \dots, r$, then this holds with probability 1.

A similar theorem can be stated and proved in the infinite-dimensional setting [2]. For brevity, we shall not do this.

The key feature of Theorem 1 is that the bounds (5.9) and (5.10) involve only local quantities: namely, local sparsities k_j , local coherences $\mu(j, l)$, local measurements m_j , and relative sparsities K_j (note that the k_j 's in the theorem are arbitrary, and in particular, are not necessarily of the form (5.5)). This theorem directly generalizes a standard compressed sensing result for sampling with incoherent bases [1, 14] to the case of multiple levels. Specifically, this standard result is

$$\|c - \hat{c}\|_2 \leq C \left(\tilde{\eta} \left(1 + D\sqrt{k} \right) + \sigma_k(c)_1 \right),$$

where k is the global sparsity and $\sigma_k(c)_1$ is the best k -term approximation error, and this holds provided

$$m \gtrsim N\mu k \log(\varepsilon^{-1}) \log(n),$$

where $\mu = \mu(U)$ is the global coherence. Having said this, it is not immediately obvious how to understand these bounds of Theorem 1 in terms of how many measurements m_j are actually required in the j^{th} level. Whilst one may hope for simpler bounds, it can be shown that these estimates are in fact sharp for a large class of matrices U [2]. Moreover, in the important case Fourier/wavelets, one can analyze the local coherences $\mu(j, l)$ and relative sparsities K_j to get such explicit estimates. We consider this next.

5.4.3 The case of Fourier sampling with wavelets

Let us consider the example of Section 5.3.3, where the matrix U arises from the Fourier/Haar wavelet pair, the sampling levels correspond to the aforementioned frequency bands W_j and the sparsity levels are the Haar wavelet scales.

Theorem 2. ³ Let U and Ω be as in Section 5.3.3 (recall that we index over $j, l = 0, \dots, r-1$) and suppose that $x \in \mathbb{C}^n$. Let $\varepsilon \in (0, \varepsilon^{-1}]$ and suppose that

$$m_j \gtrsim \left(k_j + \sum_{\substack{l=0 \\ l \neq j}}^{r-1} 2^{-\frac{|j-l|}{2}} k_l \right) \log(\varepsilon^{-1}) \log(n), \quad j = 0, \dots, r-1. \quad (5.12)$$

Then, with probability exceeding $1 - k\varepsilon$, where $k = k_0 + \dots + k_{r-1}$, any minimizer \hat{x} of (5.1) satisfies

$$\|x - \hat{x}\|_2 \leq C \left(\eta \sqrt{D} (1 + E \sqrt{k}) + \sigma_{\mathbf{k}, \mathbf{M}}(\Phi^* x)_1 \right),$$

where $\sigma_{\mathbf{k}, \mathbf{M}}(\Phi^* x)_1$ is as in (5.7), $D = 1 + \frac{\sqrt{\log_2(6\varepsilon^{-1})}}{\log_2(4En\sqrt{k})}$ and $E = \max_{j=0, \dots, r-1} \{(N_j - N_{j-1})/m_j\}$. If $m_j = |W_j|$, $j = 0, \dots, r-1$, then this holds with probability 1.

The key part of this theorem is (5.12). Recall that if U were exactly block diagonal, then $m_j \gtrsim k_j$ would suffice (up to log factors). The estimate (5.12) asserts that we require only slightly more samples, and this is due to interferences from the other sparsity levels. However, as $|j - l|$ increases, the effect of these levels decreases exponentially. Thus, the number of measurements m_j required in the j^{th} frequency band is determined predominantly by the sparsities in the scales $l \approx j$. Note that $k_l \approx \mathcal{O}(k_j)$ when $l \approx j$ for typical signals and images, so the estimate (5.12) is typically on the order of k_j in practice.

The estimate (5.12) both agrees with the conclusion of the flip test in Figure 5.1 and explains the results seen. Flipping the wavelet coefficients changes the local sparsities k_1, \dots, k_r . Therefore to recover the flipped image to the same accuracy as the unflipped image, (5.12) asserts that one must change the local numbers of measurements m_j . But in Figure 5.1 the same sampling pattern was used in both cases, thereby leading to the worse reconstruction in the flipped case. Note that (5.12) also highlights why the optimal sampling pattern must depend on the image, and specifically, the local sparsities. In particular, there can be no optimal sampling strategy for all images.

³For a proof, we refer to [3].

Note that Theorem 2 is a simplified version, presented here for the purposes of elucidation, of a more general result found in [2] which applies to all compactly supported orthonormal wavelets in the infinite-dimensional setting.

5.5 Structured sampling and structured recovery

Structured sparsity within the context of compressed sensing has been considered in numerous previous works. See [6, 10, 16, 23, 25, 26, 32, 33, 47, 53, 55, 58] and the references therein. For the problem of reconstructing wavelet coefficients, most efforts have focused on their inherent tree structure (see Remark 3). Three well-known algorithmic approaches for doing this are model-based compressed sensing [6], TurboAMP [53], and Bayesian compressed sensing [32, 33]. All methods use Gaussian or Bernoulli random measurements, and seek to leverage the wavelet tree structure—the former deterministically, the latter two in a probabilistic manner—by appropriately designed recovery algorithms (based on modifications of existing iterative algorithms for compressed sensing). In other words, structure is incorporated solely in the recovery algorithm, and not in the measurements themselves.

In Figure 5.6 we compare these algorithms with the previously described method of multilevel Fourier sampling (similar results are also witnessed with the Hadamard matrix). Note that the latter, unlike other three methods, exploits structure by taking appropriate measurements, and uses an unmodified compressed sensing algorithm (ℓ^1 minimization). As is evident, this approach is able to better exploit the sparsity structure, leading to a significantly improved reconstruction. This experiment is representative of a large set tested. In all cases, we find that exploiting structure by sampling with asymptotically incoherent Fourier/Hadamard bases outperforms such approaches that seek to leverage structure in the recovery algorithm.

5.6 The Restricted Isometry Property in levels

The flip test demonstrates that the subsampled Fourier/wavelets matrix $P_\Omega U$ does not satisfy a meaningful RIP. However, due to the sparsity structure of the signal, we are still able to reconstruct, as was confirmed in Theorem 1. The RIP is therefore too crude a tool to explain the recoverability properties of structured sensing matrices. Having said that, the RIP is a useful for deriving uniform, as opposed to nonuniform, recovery results, and for analyzing other compressed sensing algorithms besides convex optimization, such as greedy methods [27]. This raises the question of whether there are alternatives which are satisfied by such matrices. One possibility which we now discuss is the RIP in levels. Throughout this section we shall work in the finite-dimensional setting. For proofs of the theorems, see [7].

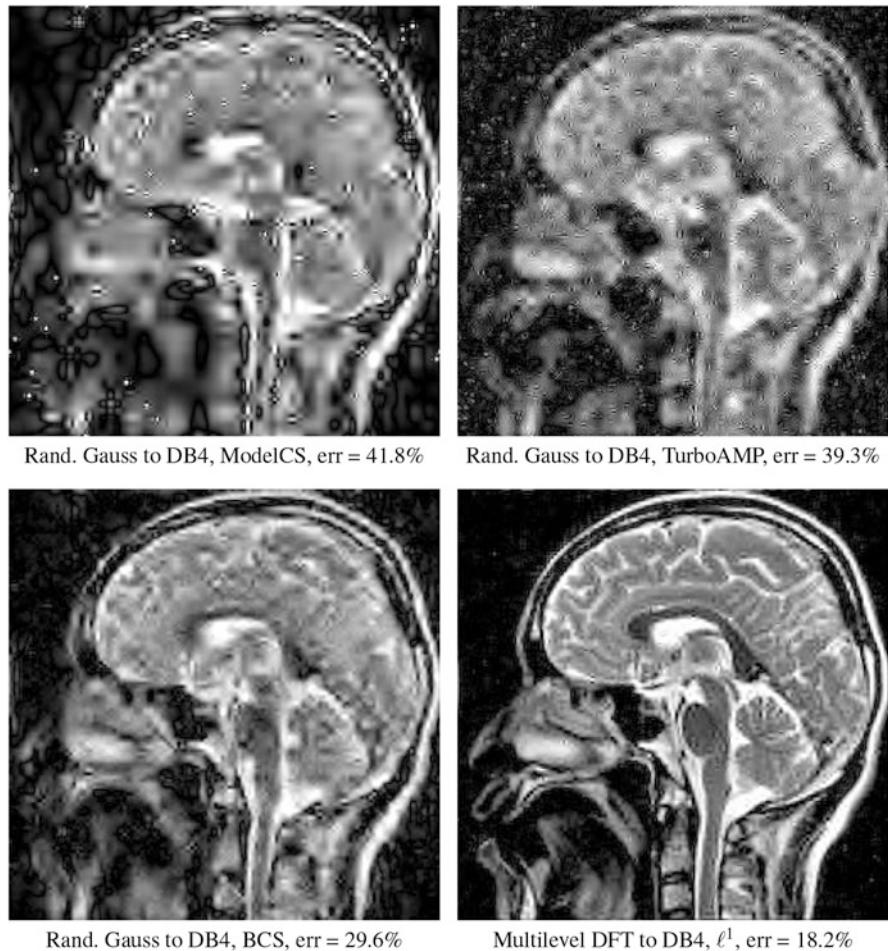


Fig. 5.6 Recovery from 12.5 % measurements at 256×256 . Comparison between random sampling with structured recovery and structured sampling with ℓ^1 -minimization recovery.

Definition 6. Given an r -level sparsity pattern (\mathbf{k}, \mathbf{M}) , where $M_r = n$, we say that the matrix $A \in \mathbb{C}^{m \times n}$ satisfies the *RIP in levels* (RIP_L) with RIP_L constant $\delta_{\mathbf{k}} \geq 0$ if for all x in $\Sigma_{\mathbf{k}, \mathbf{M}}$ we have

$$(1 - \delta_{\mathbf{k}}) \|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_{\mathbf{k}}) \|x\|_2^2.$$

The motivation for this definition is the following. Suppose that we repeat the flip test from Figure 5.1 except that instead of completely flipping the coefficients we only flip them within levels corresponding to the wavelet scales. We will refer to this as the *flip test in levels*. Note the difference between Figure 5.1 and Figure 5.7,

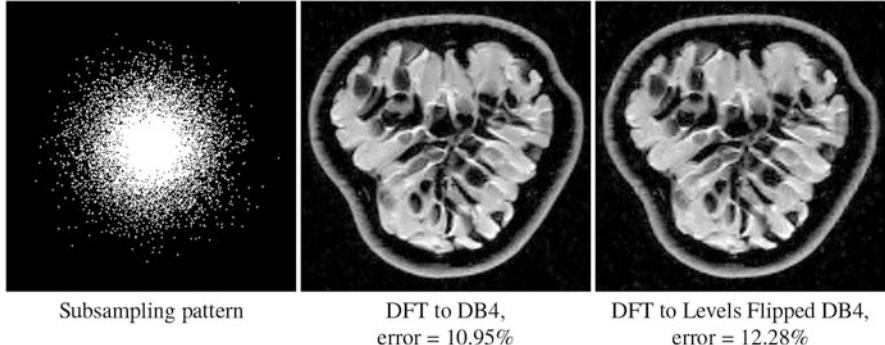


Fig. 5.7 The flip test in levels for the image considered in Figure 5.1.

where latter presents the flip test in levels: clearly, flipping within scales does not alter the reconstruction quality. In light of this experiment, we propose the above RIP in levels definition so as to respect the level structure.

We now consider the recovery properties of matrices satisfying the RIP in levels. For this, we define the *ratio constant* $\lambda_{\mathbf{k}, \mathbf{M}}$ of a sparsity pattern (\mathbf{k}, \mathbf{M}) to be $\lambda_{\mathbf{k}, \mathbf{M}} := \max_{j, l} k_j / k_l$. We assume throughout that $k_j \geq 1, \forall j$, so that $\eta_{\mathbf{k}, \mathbf{M}} < \infty$, and also that $M_r = n$. We now have the following:

Theorem 3. *Let (\mathbf{k}, \mathbf{M}) be a sparsity pattern with r levels and ratio constant $\lambda = \lambda_{\mathbf{k}, \mathbf{M}}$. Suppose that the matrix A has RIP_L constant $\delta_{2\mathbf{k}}$ satisfying*

$$\delta_{2\mathbf{k}} < \frac{1}{\sqrt{r(\sqrt{\lambda} + 1/4)^2 + 1}}. \quad (5.13)$$

Let $x \in \mathbb{C}^n$, $y \in \mathbb{C}^m$ be such that $\|Ux - y\|_2 \leq \eta$, and let \hat{x} be a minimizer of

$$\min_{z \in \mathbb{C}^n} \|z\|_1 \quad \text{s.t.} \quad \|y - Az\|_2 \leq \eta.$$

Then

$$\|x - \hat{x}\|_1 \leq C\sigma_{\mathbf{k}, \mathbf{M}}(x)_1 + D\sqrt{k}\eta, \quad (5.14)$$

where $k = k_1 + \dots + k_r$ and the constants C and D depend only on $\delta_{2\mathbf{k}}$.

This theorem is a generalization of a known result in standard (i.e., one-level) compressed sensing. Note that (5.13) reduces to the well-known estimate $\delta_{2k} \leq 4/\sqrt{41}$ [27] when $r = 1$. On the other hand, in the multiple level case the reader may be concerned that the bound ceases to be useful, since the right-hand side of (5.13) deteriorates with both the number of levels r and the sparsity ratio λ . As we show in the following two theorems, the dependence on r and λ in (5.13) is sharp:

Theorem 4. Fix $a \in \mathbb{N}$. There exists a matrix A with two levels and a sparsity pattern (\mathbf{k}, \mathbf{M}) such that the RIP_L constant $\delta_{a\mathbf{k}}$ and ratio constant $\lambda = \lambda_{\mathbf{k}, \mathbf{M}}$ satisfy

$$\delta_{a\mathbf{k}} \leq 1/|f(\lambda)|, \quad (5.15)$$

where $f(\lambda) = o(\sqrt{\lambda})$, but there is an $x \in \Sigma_{\mathbf{k}, \mathbf{M}}$ such that x is not the minimizer of

$$\min_{z \in \mathbb{C}^n} \|z\|_1 \quad \text{s.t.} \quad Az = Ax.$$

Roughly speaking, Theorem 4 says that if we fix the number of levels and try to replace (5.13) with a condition of the form

$$\delta_{2\mathbf{k}} < \frac{1}{C\sqrt{r}} \lambda^{-\frac{\alpha}{2}}$$

for some constant C and some $\alpha < 1$ then the conclusion of Theorem 3 ceases to hold. In particular, the requirement on $\delta_{2\mathbf{k}}$ cannot be independent of λ . The parameter a in the statement of Theorem 4 also means that we cannot simply fix the issue by changing $\delta_{2\mathbf{k}}$ to $\delta_{3\mathbf{k}}$, or any further multiple of \mathbf{k} .

Similarly, we also have a theorem that shows that the dependence on the number of levels r cannot be ignored.

Theorem 5. Fix $a \in \mathbb{N}$. There exists a matrix A and an r -level sparsity pattern (\mathbf{k}, \mathbf{M}) with ratio constant $\lambda_{\mathbf{k}, \mathbf{M}} = 1$ such that the RIP_L constant $\delta_{a\mathbf{k}}$ satisfies

$$\delta_{a\mathbf{k}} \leq 1/|f(r)|,$$

where $f(r) = o(\sqrt{r})$, but there is an $x \in \Sigma_{\mathbf{k}, \mathbf{M}}$ such that x is not the minimizer of

$$\min_{z \in \mathbb{C}^n} \|z\|_1 \quad \text{s.t.} \quad Az = Ax.$$

These two theorems suggest that, at the level of generality of the RIP_L , one must accept a bound that deteriorates with the number of levels and ratio constant. This begs the question: What is the effect of such deterioration? To understand this, consider the case studied earlier, where the r levels correspond to wavelet scales. For typical images, the ratio constant λ grows only very mildly with n , where $n = 2^r$ is the dimension. Conversely, the number of levels is equal to $\log_2(n)$. This suggests that estimates for the Fourier/wavelet matrix that ensure an RIP in levels (thus guaranteeing uniform recovery) will involve at least several additional factors of $\log(n)$ beyond what is sufficient for nonuniform recovery (see Theorem 2). Proving such estimates for the Fourier/wavelets matrix is work in progress.

Acknowledgements The authors thank Andy Ellison from Boston University Medical School for kindly providing the MRI fruit image, and General Electric Healthcare for kindly providing the brain MRI image. BA acknowledges support from the NSF DMS grant 1318894.

ACH acknowledges support from a Royal Society University Research Fellowship. ACH and BR acknowledge the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/L003457/1.

References

1. Adcock, B., Hansen, A.C.: Generalized sampling and infinite-dimensional compressed sensing. Technical report NA2011/02, DAMTP, University of Cambridge (2011)
2. Adcock, B., Hansen, A.C., Poon, C., Roman, B.: Breaking the coherence barrier: a new theory for compressed sensing. arXiv:1302.0561 (2014)
3. Adcock, B., Hansen, A.C., Roman, B.: A note on compressed sensing of structured sparse wavelet coefficients from subsampled Fourier measurements. arXiv:1403.6541 (2014)
4. Adcock, B., Hansen, A.C., Roman, B., Teschke, G.: Generalized sampling: stable reconstructions, inverse problems and compressed sensing over the continuum. *Adv. Imaging Electron Phys.* **182**, 187–279 (2014)
5. Ailon, N., Liberty, E.: An almost optimal unrestricted fast Johnson–Lindenstrauss transform. In: Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA) (2011)
6. Baraniuk, R.G., Cevher, V., Duarte, M.F., Hedge, C.: Model-based compressive sensing. *IEEE Trans. Inf. Theory* **56**(4), 1982–2001 (2010)
7. Bastounis, A., Hansen, A.C.: On the absence of the RIP in real-world applications of compressed sensing and the RIP in levels. arXiv:1411.4449 (2014)
8. Bigot, J., Boyer, C., Weiss, P.: An analysis of block sampling strategies in compressed sensing. arXiv:1305.4446 (2013)
9. Binev, P., Dahmen, W., DeVore, R.A., Lamby, P., Savu, D., Sharpley, R.: Compressed sensing and electron microscopy. In: Vogt, T., Dahmen, W., Binev, P. (eds.) *Modeling Nanoscale Imaging in Electron Microscopy, Nanostructure Science and Technology*, pp. 73–126. Springer, New York (2012)
10. Bourrier, A., Davies, M.E., Peleg, T., Pérez, P., Gribonval, R.: Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. arXiv:1311.6239 (2013)
11. Boyer, C., Weiss, P., Bigot, J.: An algorithm for variable density sampling with block-constrained acquisition. *SIAM J. Imag. Sci.* **7**(2), 1080–1107 (2014)
12. Candès, E., Donoho, D.L.: Recovering edges in ill-posed inverse problems: optimality of curvelet frames. *Ann. Stat.* **30**(3), 784–842 (2002)
13. Candès, E.J., Donoho, D.L.: New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Commun. Pure Appl. Math.* **57**(2), 219–266 (2004)
14. Candès, E.J., Plan, Y.: A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inf. Theory* **57**(11), 7235–7254 (2011)
15. Candès, E.J., Romberg, J.: Sparsity and incoherence in compressive sampling. *Inverse Prob.* **23**(3), 969–985 (2007)
16. Carson, W.R., Chen, M., Rodrigues, M.R.D., Calderbank, R., Carin, L.: Communications-inspired projection design with application to compressive sensing. *SIAM J. Imag. Sci.* **5**(4), 1185–1212 (2012)
17. Chartrand, R., Sidky, Y., Pan, X.: Nonconvex compressive sensing for X-ray CT: an algorithm comparison. In: Asilomar Conference on Signals, Systems, and Computers (2013)
18. Crouse, M.S., Nowak, R.D., Baraniuk, R.G.: Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Process.* **46**, 886–902 (1998)
19. Dahlke, S., Kutyniok, G., Steidl, G., Teschke, G.: Shearlet coorbit spaces and associated Banach frames. *Appl. Comput. Harmon. Anal.* **27**(2), 195–214 (2009)

20. Dahlke, S., Kutyniok, G., Maass, P., Sagiv, C., Stark, H.-G., Teschke, G.: The uncertainty principle associated with the continuous shearlet transform. *Int. J. Wavelets Multiresolution Inf. Process.* **6**(2), 157–181 (2008)
21. DeVore, R.A.: Nonlinear approximation. *Acta Numer.* **7**, 51–150 (1998)
22. Do M.N., Vetterli, M.: The contourlet transform: An efficient directional multiresolution image representation. *IEEE Trans. Image Process.* **14**(12), 2091–2106 (2005)
23. Donoho, D.L.: Compressed sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006)
24. Duarte, M.F., Davenport, M.A., Takhar, D., Laska, J., Kelly, K., Baraniuk, R.G.: Single-pixel imaging via compressive sampling. *IEEE Signal Process. Mag.* **25**(2), 83–91 (2008)
25. Duarte M.F., Eldar, Y.C.: Structured compressed sensing: from theory to applications. *IEEE Trans. Signal Process.* **59**(9), 4053–4085 (2011)
26. Eldar, Y.C.: Robust recovery of signals from a structured union of subspaces. *IEEE Trans. Inf. Theory* **55**(11), 5302–5316 (2009)
27. Foucart, S., Rauhut, H.: *A Mathematical Introduction to Compressive Sensing*. Birkhauser, New York (2013)
28. Gan, L., Do, T.T., Tran, T.D.: Fast compressive imaging using scrambled Hadamard ensemble. *Proc. EUSIPCO* (2008)
29. Guerquin-Kern, M., Häberlin, M., Pruessmann, K.P., Unser, M.: A fast wavelet-based reconstruction method for magnetic resonance imaging. *IEEE Trans. Med. Imaging* **30**(9), 1649–1660 (2011)
30. Guerquin-Kern, M., Lejeune, L., Pruessmann, K.P., Unser, M.: Realistic analytical phantoms for parallel magnetic resonance imaging. *IEEE Trans. Med. Imaging* **31**(3), 626–636 (2012)
31. Haldar, J., Hernando, D., Liang, Z.: Compressed-sensing MRI with random encoding. *IEEE Trans. Med. Imaging* **30**(4), 893–903 (2011)
32. He, L., Carin, L.: Exploiting structure in wavelet-based Bayesian compressive sensing. *IEEE Trans. Signal Process.* **57**(9), 3488–3497 (2009)
33. He, L., Chen, H., Carin, L.: Tree-structured compressive sensing with variational Bayesian analysis. *IEEE Signal Process. Lett.* **17**(3), 233–236 (2010)
34. Herman, M.A.: Compressive sensing with partial-complete, multiscale Hadamard waveforms. In: *Imaging and Applied Optics*, p. CM4C.3. Optical Society of America (2013) <https://www.osapublishing.org/viewmedia.cfm?uri=COSI-2013-CM4C.3&seq=0>
35. Herman, M.A., Strohmer, T.: High-resolution radar via compressed sensing. *IEEE Trans. Signal Process.* **57**(6), 2275–2284 (2009)
36. Hinrichs, A., Vybird, J.: Johnson–Lindenstrauss lemma for circulant matrices. *Random Struct. Algorithm* **39**(3), 391–398 (2011)
37. Huang, G., Jiang, H., Matthews, K., Wilford, P.: Lensless imaging by compressive sensing. *arXiv:1305.7181* (2013)
38. Krahmer, F., Ward, R.: New and improved Johnson–Lindenstrauss embeddings via the restricted isometry property. *SIAM J. Math. Anal.* **43**(3), 1269–1281 (2011)
39. Krahmer, F., Ward, R.: Stable and robust recovery from variable density frequency samples. *IEEE Trans. Image Proc.* **23**(2), 612–622 (2014)
40. Kutyniok, G., Lemvig, J., Lim, W.-Q.: Compactly supported shearlets. In: Neamtu M., Schumaker, L. (eds.) *Approximation Theory XIII: San Antonio 2010*. Springer Proceedings in Mathematics, vol. 13, pp. 163–186. Springer, New York (2012)
41. Leary, R., Saghi, Z., Midgley, P.A., Holland, D.J.: Compressed sensing electron tomography. *Ultramicroscopy* **131**(0), 70–91 (2013)
42. Liang, D., Xu, G., Wang, H., King, K.F., Xu, D., Ying, L.: Toeplitz random encoding MR imaging using compressed sensing. In: *Proceedings/IEEE International Symposium on Biomedical Imaging*, pp. 270–273, June 2009
43. Lin, T., Herrman, F.J.: Compressed wavefield extrapolation. *Geophysics* **72**(5), SM77–SM93 (2007)
44. Lustig, M., Donoho, D.L., Pauly, J.M.: Sparse MRI: the application of compressed sensing for rapid MRI imaging. *Magn. Reson. Imaging* **58**(6), 1182–1195 (2007)

45. Lustig, M., Donoho, D.L., Santos, J.M., Pauly, J.M.: Compressed sensing MRI. *IEEE Signal Process. Mag.* **25**(2), 72–82 (2008)
46. Mallat, S.G.: *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd edn. Academic, London (2009)
47. Mishali, M., Eldar, Y.C., Elron, A.J.: Xampling: Signal acquisition and processing in union of subspaces. *IEEE Trans. Signal Process.* **59**(10), 4719–4734 (2011)
48. Po, D.D.-Y., Do, M.N.: Directional multiscale modeling of images using the contourlet transform. *IEEE Trans. Image Process.* **15**(6), 1610–1620 (2006)
49. Puy, G., Marques, J.P., Gruetter, R., Thiran, J., Van De Ville, D., Vanderghenst, P., Wiaux, Y.: Spread spectrum magnetic resonance imaging. *IEEE Trans. Med. Imaging* **31**(3), 586–598 (2012)
50. Roman, B., Adcock, B., Hansen, A.C.: On asymptotic structure in compressed sensing. *arXiv:1406.4178* (2014)
51. Romberg, J.: Imaging via compressive sampling. *IEEE Signal Process. Mag.* **25**(2), 14–20 (2008)
52. Sebert, F., Xou, Y.M., Ying, L.: Compressed sensing MRI with random B1 field. *Proc. Int. Soc. Magn. Reson. Med.* **16**, 3151 (2008)
53. Som, S., Schniter, P.: Compressive imaging using approximate message passing and a markov-tree prior. *IEEE Trans. Signal Process.* **60**(7), 3439–3448 (2012)
54. Studer, V., Bobin, J., Chahid, M., Moussavi, H., Candès, E., Dahan, M.: Compressive fluorescence microscopy for biological and hyperspectral imaging. *Natl. Acad. Sci. USA* **109**(26), 1679–1687 (2011)
55. Tsaig, Y., Donoho, D.L.: Extensions of compressed sensing. *Signal Process.* **86**(3), 549–571 (2006)
56. van den Berg, E., Friedlander, M.P.: SPGL1: a solver for large-scale sparse reconstruction. <http://www.cs.ubc.ca/labs/scl/spgl1> (2007)
57. van den Berg, E., Friedlander, M.P.: Probing the pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comput.* **31**(2), 890–912 (2008)
58. Wang, L., Carlson, D., Rodrigues, M.R.D., Wilcox, D., Calderbank, R., Carin, L.: Designed measurements for vector count data. *Adv. Neural Inf. Process. Syst.*, pp. 1142–1150 (2013)
59. Wang, Z., Arce, G.R.: Variable density compressed image sampling. *IEEE Trans. Image Process.* **19**(1), 264–270 (2010)
60. Wiaux, Y., Jacques, L., Puy, G., Scaife, A.M.M., Vanderghenst, P.: Compressed sensing imaging techniques for radio interferometry. *Mon. Not. R. Astron. Soc.* **395**(3), 1733–1742 (2009)
61. Wong, E.C.: Efficient randomly encoded data acquisition for compressed sensing. In: *Proceedings on International Society for Magnetic Resonance in Medicine*, p. 4893 (2010)

Chapter 6

Compressive Sensing in Acoustic Imaging

Nancy Bertin, Laurent Daudet, Valentin Emiya, and Rémi Gribonval

Abstract Acoustic sensing is at the heart of many applications, ranging from underwater sonar and nondestructive testing to the analysis of noise and their sources, medical imaging, and musical recording. This chapter discusses a palette of acoustic imaging scenarios where sparse regularization can be leveraged to design compressive acoustic imaging techniques. Nearfield acoustic holography (NAH) serves as a guideline to describe the general approach. By coupling the physics of vibrations and that of wave propagation in the air, NAH can be expressed as an inverse problem with a sparsity prior and addressed through sparse regularization. In turn, this can be coupled with ideas from compressive sensing to design semi-random microphone antennas, leading to improved hardware simplicity, but also to new challenges in terms of sensitivity to a precise calibration of the hardware and software scalability. Beyond NAH, this chapter shows how compressive sensing is being applied to other acoustic scenarios such as active sonar, sampling of the plenacoustic function, medical ultrasound imaging, localization of directive sources, and interpolation of plate vibration response.

N. Bertin (✉)

IRISA - CNRS UMR 6074, PANAMA team (Inria & CNRS), Campus de Beaulieu,
F-35042 Rennes Cedex, France
e-mail: nancy.bertin@irisa.fr

L. Daudet

Paris Diderot University, Institut Langevin, ESPCI ParisTech, CNRS
UMR 7587, F-75005 Paris, France
e-mail: laurent.daudet@espci.fr

V. Emiya

Aix-Marseille Université, CNRS, LIF UMR 7279, F-13288 Marseille Cedex 9, France
e-mail: valentin.emyia@lif.univ-mrs.fr

R. Gribonval

Inria, PANAMA team (Inria & CNRS), Campus de Beaulieu, F-35042 Rennes Cedex, France
e-mail: remi.gribonval@inria.fr

6.1 Introduction

Acoustic sensing is at the heart of many applications, ranging from underwater sonar and nondestructive testing to the analysis of noise and their sources, medical imaging, and musical recording. With the emergence of compressive sensing and its successes from magnetic resonance medical imaging and optics to astrophysics, one can naturally envision new acoustic sensing techniques where the nature of the sensors is revisited jointly with the models and techniques used to extract acoustic information from raw recordings, under the auspices of sparsity.

Acoustic imaging is indeed a domain that combines a number of features calling for sparse regularization and compressive sensing:

- **high dimensionality:** acoustic data such as pressure fields are high-dimensional spatio-temporal objects whose complete acquisition could generate huge volumes of data and require high throughput interfaces;
- **structure and correlation:** sensor arrays such as acoustic antennas tend to capture correlated information;
- **linear sensors:** the behavior of most acoustic sensors such as microphones or hydrophones is well-approximated in standard regimes as being linear;
- **linear equations:** similarly, in standard regimes the wave equation as well as related PDEs that drive the physics of acoustic phenomena can be considered as linear, so the observed phenomena depend linearly on the generating sources, whether in active or passive scenarios.

These features call for the expression of acoustic imaging in the context of linear inverse problems and dimensionality reduction, where low-dimensional models can be leveraged to acquire high-dimensional objects through few, non-adaptive, linear measurements. However, the deployment of sparse regularization and compressive sensing tools in acoustic imaging raises a number of questions:

- Where does sparsity or low-dimensionality emerge from? In other words, in what domain can we expect the considered objects to be sparse?
- Can we drive the design of sensor arrays using the sparsity assumption?
- What are the practical gains in exploiting sparsity?
- Can we go as far as compressive sensing, *i.e.*, can we leverage sparsity to voluntarily reduce the number of array elements while preserving imaging quality?

This chapter discusses a palette of acoustic imaging scenarios where recent advances and challenges in compressive acoustic imaging are highlighted.

Nearfield acoustic holography (NAH) serves as a guideline to describe the general approach. This technique, which is used to study vibrating structures producing noise in the car industry, in the aircraft industry and the railway industry (for acoustic comfort in or outside the vehicles), or in the naval industry (acoustic signature of ships), consists in imaging a vibrating structure by “listening” to the sound produced using a large set of microphones. By coupling the physics of vibrations and that of wave propagation in the air, NAH can be expressed as an

inverse problem with a sparsity prior and addressed through sparse regularization. In turn, this can be coupled with ideas from compressive sensing to design new semi-random microphone antennas, the goal being primarily to sub-sample in space rather than in time, since time-sampling is mostly “free” in such an acoustic context. This leads not only to substantial practical benefits in terms of hardware simplicity, but also to new challenges in terms of sensitivity to a precise calibration of the hardware. Following the general framework described in the context of NAH in Section 6.2, we further discuss a number of acoustic scenarios:

- a) Active sonar for underwater and air ultrasound imaging (Section 6.3.1);
- b) Sampling of the plenacoustic function (Section 6.3.2);
- c) Medical ultrasound imaging (Section 6.3.3);
- d) Localization of directive sources (Section 6.4.1.1);
- e) Interpolation of plate vibration response (Section 6.4.1.1);

6.2 Compressive Nearfield Acoustic Holography

NAH is traditionally expressed as a linear inverse problem where the goal is to estimate the vibration of the structure given the pressure field recorded in a plane at a short distance from the structure. Acoustic images are usually obtained with Tikhonov regularization techniques.

Traditional NAH typically suffers from hardware complexity (size of the microphone antenna) and the large duration of the acquisition, which are both necessary to obtain high quality images. It is however possible to circumvent these issues by adopting a compressive sensing approach to NAH. This can be achieved by coupling the choice of models (or dictionaries) adapted to vibrating plates with the design of a new shape for the microphone antenna (semi-random geometry). The reconstruction of acoustic images exploits sparse regularization techniques, for example relying on convex optimization.

Numerical and experimental results demonstrate practical benefits both in acquisition time and in hardware simplicity, while preserving a good reconstruction quality. They also highlight some important practical issues such as a sensitivity of the imaging process to the precise modeling of the hardware implementation.

6.2.1 Standard Nearfield Acoustic Holography (NAH)

The direct problem of NAH is the expression of the pressure field $p(\mathbf{r}, t)$ measured at a distance z_0 above the vibrating plate, located in the (Oxy) plane, as a function of the vibration of the plate, here described by its normal velocity field $u(\mathbf{r}, t)$.

For a fixed eigenfrequency ω , the discrete formulation goes as

$$p = F^{-1}GFu = Au \quad (6.1)$$

where:

- u denotes the vector of source normal velocities to be identified, discretized on a rectangular regular grid,
- p is the vector of measured pressures, also discretized in the hologram plane,
- F is the square 2-D spatial DFT operator,
- G is a known propagation operator, derived from the Green's function of free-field acoustic propagation,
- $A = F^{-1}GF$ is the measurement matrix gathering all the linear operators.

Assuming square matrices, a naive inversion of Equation (6.1) yields

$$u = A^{-1}p \quad (6.2)$$

However, the operator A is badly ill-conditioned, as G expresses the propagation of so-called evanescent waves, whose amplitudes are exponentially decaying with distance. The computation of the sources using this equation is therefore very unstable, and thus requires regularization. In its most standard form called Tikhonov regularization [22], this is done by adding an extra ℓ_2 -norm penalty term and, generally, involves the solution of the following minimization problem:

$$\hat{u} = \min_u \|p - Au\|_2^2 + \lambda \|Lu\|_2^2 \quad (6.3)$$

where L is the so-called Tikhonov matrix and λ the regularization parameter. Denoting $R_\lambda = (A^T A + \lambda L^T L)^{-1} A^T A$, the result of the Tikhonov regularization can be expressed in closed form as:

$$\hat{u} = R_\lambda A^{-1} p \quad (6.4)$$

It should be noted that in this analysis, it is implicitly assumed that the pressure field is completely known in the hologram plane at $z = z_0$, in order to “retro-propagate” the acoustic field. For high frequencies (small wavelengths) and relatively large plates, the corresponding regular spatial samples at Nyquist rates may involve several hundreds of sampling points. In practice, microphone arrays with significantly more than 100 microphones are costly and one has to repeat the experiment in different positions of the array in order to get a sufficiently fine sampling of the measurements. In a typical experiment, a 120-microphone array¹ was positioned in 16 different positions (4 positions in each x and y direction), leading to 1920

¹Experiments performed by François Ollivier and Antoine Peillot at Institut Jean Le Rond d'Alembert, UPMC Univ. Paris 6

sampling locations and a lengthy measurement process. Furthermore, Tikhonov regularization here amounts to a low-pass filtering (in spatial frequencies) of the vibration wave field u , and this leads to non-negligible artifacts in the estimated field \hat{u} especially at low frequencies, and near the plate boundaries.

6.2.2 Sparse modeling

In the standard approach described above, only weak assumptions are made on the field under study, namely on its spatial frequency bandwidth. In this section, we show that more precise models actually lead to a significant reduction in the number of measurements necessary for its sampling. The models here are based on the sparsity of the wave field in an appropriate dictionary Ψ : in matrix form one has $u = \Psi c$, where c is a sparse (or compressible) vector.

Theoretical results [25] indicate that linear combinations of plane waves provide good approximations to solutions of the Helmholtz equation on any star-shaped plate (in particular, all convex plates are star-shaped), under any type of boundary conditions. These results have been recently extended to the solutions of the Kirchhoff–Love equation for vibrations of thin isotropic homogeneous plates [7]. Mathematically, the velocity of the plate u can be approximated by a sum of a sparse number of plane waves (evanescent waves are here neglected):

$$u(\mathbf{r}) \approx \left(\sum_j c_j e^{i\mathbf{k}_j \cdot \mathbf{r}} \right) \mathbf{1}_{\mathcal{S}}(\mathbf{r}) \quad (6.5)$$

where $\mathbf{1}_{\mathcal{S}}(\mathbf{r})$ is the indicator function that restricts the plane waves to the domain \mathcal{S} of the plate, the vectors \mathbf{k}_j are the wavevectors of the plane waves, $\mathbf{r} = (x, y)$, and the c_j are the corresponding coefficients. To build the dictionary Ψ , we generate plane waves with wavevectors \mathbf{k}_j regularly sampling the 2D Fourier plane and restrict them to the domain of the plate \mathcal{S} . This is actually equivalent to restricting to \mathcal{S} the basis vectors of the discrete Fourier transform on a larger rectangular domain containing \mathcal{S} . This is illustrated in Figure 6.1.

6.2.3 Sparse regularization for inverse problems

With this approximation framework, the sparsity of the coefficient vector c is now used to regularize the NAH inverse problem, which can be recast as follows: for a given set of pressure measurements p , find the sparsest set of coefficients c leading to a reconstructed wavefield consistent with the measurements:

$$\operatorname{argmin}_c \|c\|_0 \quad \text{s.t.} \quad p = A\Psi c, \quad (6.6)$$

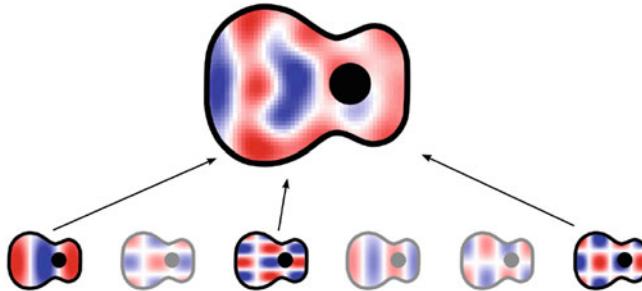


Fig. 6.1 At a given frequency, the complex vibration pattern of the plate (here, a guitar soundboard) can be approximated as a sparse sum of plane waves *Figure courtesy of F. Ollivier, UPMC*

This can be solved (approximately) using greedy algorithms, or, alternatively, through (noisy) ℓ_1 relaxation, such as a basis pursuit denoising (BPDN) [10] framework:

$$\underset{c}{\operatorname{argmin}} \|p - A\Psi c\|_2^2 + \lambda \|c\|_1 \quad (6.7)$$

with an appropriate choice of λ . Comparing Equations (6.7) and (6.3), one can see that the main difference lies in the choice of the norm: the ℓ_2 -norm of the Tikhonov regularization spreads the energy of the solution on all decomposition coefficients c , while the ℓ_1 -norm approach of BPDN promotes sparsity. In addition, sparse regularization gives an extra degree of freedom with the choice of the dictionary Ψ .

6.2.4 Sensor design for compressive acquisition

Interestingly, in the sparse modeling framework, we have dropped the need to completely sample the hologram plane: one only needs the pressure measurements to have sufficient diversity to capture the different degrees of freedom involved in the measurement. Bearing in mind that reducing the number of point samples has important practical consequences, both in terms of hardware cost (less microphones and analog-to-digital converters) and acquisition time, one may then ask—in the spirit of compressive sensing—the following two questions:

- How many point measurements are really necessary?
- Can we design better sensing matrices, *i.e.*, in practice, find a better positioning of the microphones, in order to even further reduce their number?

In the case of a signal sparse in the spatial Fourier basis, it has been shown that few point measurements in the spatial domain are sufficient to recover exactly the signal [30] and that the reconstruction is robust to noise. In acoustic experiments,

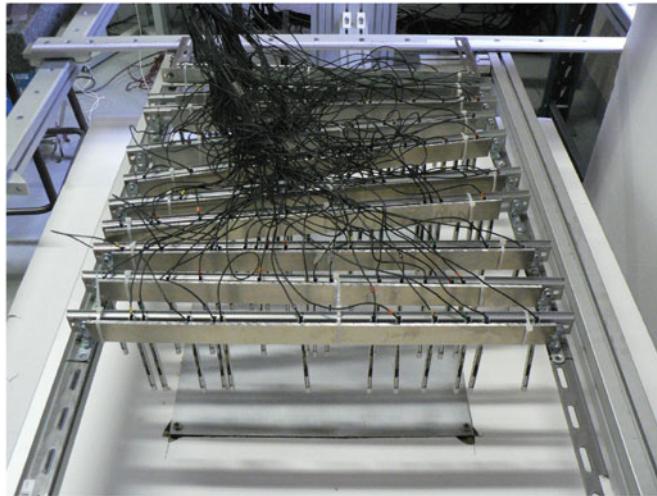


Fig. 6.2 Picture of the random array for compressive NAH. The microphones are placed at the tip of the small vertical rods, randomly placed along the ten horizontal bars. Part of the rectangular plate under study can be seen below the array. *Picture courtesy of F. Ollivier, UPMC*

the measurements are not strictly point measurements, not only because of the finite size of the microphones membrane, but also due to the acoustic propagation: each microphone gathers information about the whole vibration, although with a higher weight for the sources nearby. The theory suggests that an array with randomly placed sensors is a good choice of measurement scheme: in conjunction with sparse reconstruction principles, random microphone arrays perform better than regular arrays, as the measurement subspace becomes less coherent with the sparse signal subset (and therefore each measurement / microphone carries more global information about the whole experiment).

However, uniformly distributed random arrays are difficult to build for practical reasons (microphone mounts); therefore, there is the additional constraint to use an array that can be built using several (here, 10) straight bars, each of them holding 12 microphones. Extensive numerical simulations have shown that a good array design is obtained by bars that are tilted with respect to the axis, with small random angles, and microphones placed randomly with a uniform distribution along each bar. This is illustrated in Figure 6.2.

6.2.5 Practical benefits and sensitivity to hardware implementation

Figure 6.3 shows some results of the wave field reconstruction on two different plates—using the standard NAH technique and the compressive measurements presented above—at different frequencies and number of measurements. It can be

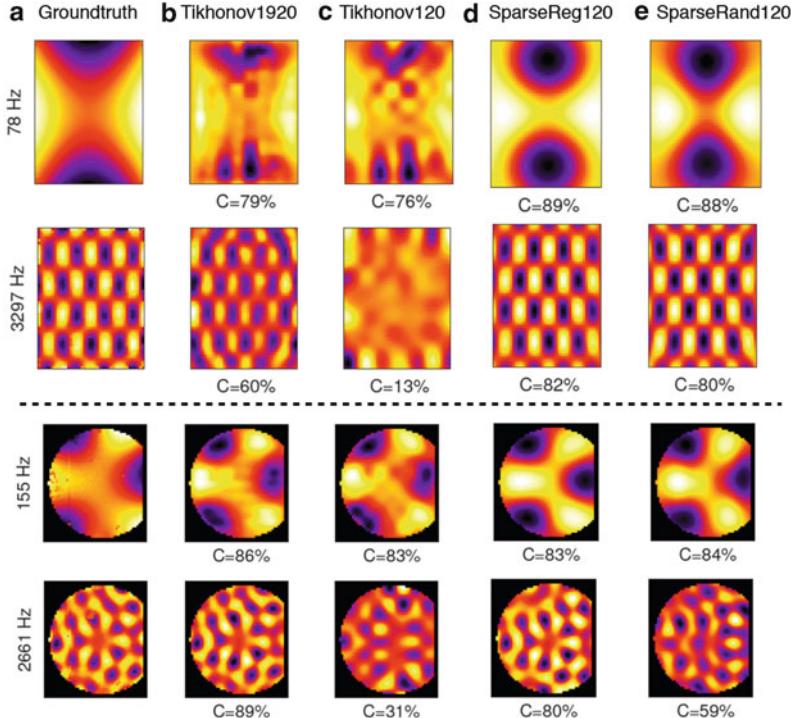


Fig. 6.3 Comparative results of NAH. Columns (b) and (c) are for standard NAH, at 1920 and 120 microphones, respectively. Columns (d) and (e) are for the sparse modeling framework, with 120 measurements, in a regular and random configuration, respectively. The number C represents the normalized cross-correlation with the reference measurements (column (a), obtained with laser velocimetry). *Figure adapted from [9]*

shown that, *with a single snapshot*, the random antenna (120 microphones) has similar performance than the dense grid of 1920 measurements (120 microphones, 16 snapshots) required for Tikhonov inversion. For some plate geometries, the number of microphones can be even further reduced, down to about 40 in the case of a rectangular plate. Furthermore, the low-frequency artifacts that were observed are no more present: the dictionary can natively model the discontinuities of the wave field at the plate boundaries.

However, using such a random array raises its share of difficulties too. One of them is that, in order to construct the propagation operator G , one has to know exactly the position of each sensor [8]—in practice this involves significantly more care than in the regular array case. Secondly, compressive sensing is much more sensitive than ℓ_2 -based methods to sensor calibration issues: it can be shown that even what would be considered a benign mismatch in sensor gain (typically few dB of error in gain) can severely impede the sparse reconstruction [5, 17].

To summarize what we have learnt from this study on NAH, applying compressive sensing to a real-world inverse problem involves a number of key components:

- *sparsity*, that here emerges from the physics of the problem. Taking sparsity into account in the inverse problem formulation may already produce some significant performance gains, at the cost of increased computation;
- *measurements*, that can be designed to optimally leverage on sparsity, although there are usually strong physical constraints in these measurements—in particular, completely random measurements as often used in theoretical studies of compressive sensing are almost never met in practice;
- *pitfalls*, that are often encountered, not only in terms of computational complexity, but also the sensitivity to the knowledge of the whole measurement system.

6.3 Acoustic imaging scenarios

Compressive NAH illustrates the potential of acoustic compressive sensing, as well as its main challenges. As we will now illustrate on a number of other scenarios, many acoustic imaging problems can be seen as linear inverse problems. While these are traditionally addressed with linear techniques (such as beamforming and matched filtering), it is often possible to exploit simple physical models to identify a domain where the considered objects—which seem intrinsically high-dimensional—are in fact rather sparse or low-dimensional, in the sense that they can be described with few parameters. State-of-the-art sparse regularization algorithms can therefore be leveraged to replace standard linear inversion. They can lead to improved resolution with the same acquisition hardware, but sometimes raise issues in terms of the computational resources they demand.

In all the scenarios considered below, it is even possible to go one step further than sparse regularization by designing new “pseudo-random” acquisition hardware, in the spirit of compressive sensing. For example, new semi-random acoustic antennas can be proposed and manufactured for both air acoustic imaging (vibrating plates, room acoustics) and sonar. By combining these antennas with the proposed sparse models, acoustic imaging techniques can be designed that improve the image quality and/or reduce the acquisition time and the number of sensors. This raises two main challenges. First, it shifts the complexity from hardware to software since the numerical reconstruction algorithms can be particularly expensive when the objects to reconstruct are high-dimensional. Second, pseudo-random sensor arrays come with a price: the precise calibration of the sensors’ response and position has been identified as a key problem to which sparse reconstruction algorithms are particularly sensitive, which opens new research perspectives around blind calibration or “autofocus” techniques.

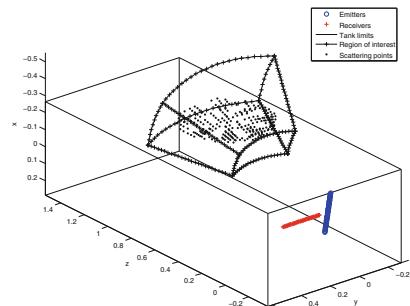
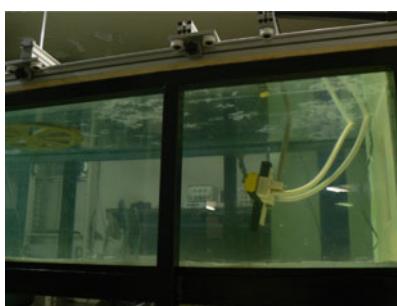


Fig. 6.4 Active sonar: 3D imaging of a spatially sparse scene composed of a few scattering objects in air or water. A wheel-shaped target is placed in a tank filled with water and imaged by means of two perpendicular arrays of emitting and receiving hydrophones (left). After discretization, it is modeled as a set of point scatterers included in a limited region of interest so as to make the sparse recovery tractable (right). *Picture courtesy of J. Marchal, UPMC*

6.3.1 Active sonar (scenario a)

Active sonar is an air or underwater ultrasound imaging scenario in which the image of the scattering objects is obtained by modeling the echoes. An emitting antenna (appearing in yellow on the left of Figure 6.4) first generates a signal sequence called a *ping* which is backscattered by the objects, and the back-propagated signal is then recorded by a receiving antenna. Several pings with different emitted signals are generated one after the other. The set of recordings is finally processed to obtain an image of the scene.

The design of an active sonar system depends on a number of features including the number of transducers in the emission and reception antennas, the directivity of the transducers, the geometry of the antennas, the emission sequences, and the imaging algorithm.

6.3.1.1 Problem formulation

The region of interest in the 3D space is discretized into voxels indexed by k . The sample $m_{npj} \in \mathbb{C}$ recorded at (discrete) time n , ping p and receiver j is stored in a multidimensional array $m \triangleq [m_{npj}]_{npj}$ which is modeled as the sum of the contribution of all the scattered signals:

$$m = \sum_k c_k \psi_k \quad (6.8)$$

where $c_k \in \mathbb{C}$ is the unknown omnidirectional scattering coefficient at voxel k and $\psi_k \triangleq [\psi_k(n, p, j)]_{n, p, j}$ is the known 3D array that models the synthesis of the recorded signal when the emission is scattered at voxel k . More precisely, we have

$$\psi_k(n, p, j) = \sum_i e_{pi}(n/f_s - \tau_{ik} - \tau_{kj})$$

where e_{pi} is the emission signal at ping p and emitter i , f_s is the sampling frequency, and τ_{ik} and τ_{kj} are the propagation delays from emitter i to voxel k and from voxel k to receiver j , respectively.

The objective is to estimate c_k for each k from the recording m and the known atoms ψ_k , which only depend on the design of the sonar device and on known physical constants such as the speed of sound.

6.3.1.2 From beamforming to sparse approaches

Beamforming—or matched filtering—is a well-established principle to address active sonar imaging by estimating c_k as a linear combination of the recorded data m . It can be written as

$$\hat{c}_k = \frac{\langle m, \psi_k \rangle}{\|\psi_k\|_F^2}. \quad (6.9)$$

This technique is called beamforming when the emission antenna and the emission signals e_{pi} are designed such that the resulting emission is focusing on a controlled area at each given ping p . The estimate \hat{c}_k also results from the formation of a beam towards a controlled area at the receiver-array level so that the intersection of the emission and reception beams is related to the direction of voxel k . A typical setting consists of linear, orthogonal emission and reception antennas that form orthogonal planar beams, so that the imaging is the result of the concatenation of 2D slices obtained at each ping.

Non-linear estimation using sparsity. A sparse assumption naturally comes from the idea that the scattering objects are supported by very few voxels, *i.e.*, $c_k = 0$ for most indices k . With a slight abuse of notation, we obtain a basic sparse estimation problem

$$\operatorname{argmin}_c \|c\|_0 \quad \text{s.t.} \quad m = \Psi c. \quad (6.10)$$

where c is a vector composed of scattering coefficients c_k for all voxels k , m is the vectorized version of the recorded data, and Ψ is the known dictionary matrix, in which column k is the vectorized version of ψ_k .

Compared to beamforming, sparse approaches provide a non-linear sparse estimate by (approximately) solving (6.10). Pings are not processed independently but simultaneously, resulting in a true 3D approach instead of a combination of

2D-slice processings: this may improve the accuracy at the price of a higher computational burden. Computational issues are closely related to the size of the 3D space discretization into voxels and are a challenge in many similar contexts. First investigations with a real sonar device have been proposed in [33].

6.3.1.3 Open questions on sparse model design and related algorithms

Investigations of sparse approaches for active sonar are still in their early developmental stage. Only the setting with synthetic data [6] and preliminary considerations on models with real data [33] are available today. A number of open questions should be addressed in order to enhance the accuracy of the results and the computational complexity of the algorithms.

How to design a good dictionary? Designing a dictionary for active sonar mainly consists in choosing the number of transducers, the geometry of the antennas, the number of pings, and the emission sequences. How these parameters relate to the imaging quality is still unclear. Such knowledge would provide cues to reduce the number of sensors and the acquisition time.

Is the omnidirectional scattering model improvable? Instead of modeling the scattering in each voxel k by a scalar c_k that assumes an omnidirectional scattering, one may propose new scattering models. For instance, one may extend the concept of a scattering coefficient c_k to a vector or a matrix \mathbf{C}_k which can model directional scattering at voxel k . Such investigations are physically motivated by arguments including near field issues or blind adaptation to imperfect calibration and lead to models with structured sparsity such as joint sparse models [12], harmonic and molecular sparse models [11, 16], or a combination of them.

Is 3D imaging tractable? Discretizing a region of interest in 3D space results in high-dimensional models. In such a context, standard estimation strategies such as convex minimization or greedy algorithms are computationally demanding [33], even for small regions of interests. A major challenge is to provide new, possibly approximate algorithms to estimate the sparse representation within a reasonable computation time under realistic imaging conditions.

6.3.2 Sampling the plenacoustic function (scenario *b*)

Compressive sensing principles can similarly be applied to sample the so-called *plenacoustic function* \mathcal{P} that gathers the set of all impulse responses between any source and receiver position (\mathbf{r}_s and \mathbf{r}_p , respectively) within a given room [1]: $\mathcal{P}(\mathbf{r}_s, \mathbf{r}_p, t)$, which of course depends on the room geometry and mechanical properties of boundary materials. In a linear setting—a reasonable assumption at audible sound levels—this function \mathcal{P} completely characterizes the acoustics of the room. Quoting the authors of [1], one may ask: “*How many microphones do we need to place in the room in order to completely reconstruct the sound field*

at any position in the room?" By a crude computation, sampling \mathcal{P} in the whole audible range (with frequencies up to 20 kHz) seems hopeless, as one would have to be able to move source and receiver (microphone) on a 3D grid with a step size of less than 1 cm (half of the smallest wavelength), leading to more than 1 million sensor positions (or microphones) per cubic meter. However, the propagative nature of acoustic waves introduces some strong constraints on \mathcal{P} ; it is, for instance, well known that the acoustic field within a bounded domain \mathcal{D} is entirely determined by the pressure field and its normal derivative at the boundary $\partial\mathcal{D}$. This is known as Kirchhoff's integral theorem and derives from Green's identity. To take advantage of these constraints within a compressive sensing framework, one must find sparse models for the acoustic field p itself. Let us assume that the source is fixed in a given room; the goal is then to estimate the acoustic pressure impulse response $p(\mathbf{r}, t)$ within a whole spatial domain \mathcal{D} , where $\mathbf{r} \in \mathcal{D}$ is the position of the receiver – by the reciprocity principle, this is equivalent to fixing the receiver and moving the source. Sparsity arises from two physically motivated, and dual, assumptions:

- **A time viewpoint:** for any $\mathbf{r} = \mathbf{r}_0$ fixed, $p(\mathbf{r}_0, t)$ is sparse in the beginning of the impulse responses, *i.e.* at $t < t_{mix}$. Indeed, the beginning of the impulse responses is characterized by a set of isolated pulses, corresponding first to the direct sound (direct wave propagation between source and receiver) and then to the so-called *early echoes* of the impulse bouncing on the walls (first-order reflexions bouncing on one wall and then higher order reflexions). After t_{mix} called mixing time, the density of echoes and their dispersion makes them impossible to isolate, the impulse response being then better characterized by a stochastic model. To take into account the spatial variations (as a function of \mathbf{r}), this sparsity is exploited in the framework of an *image-source model*: first-order reflexions may be modeled as impulses coming in direct path from a set of virtual sources located symmetrically from the (real) source with respect to the walls, assumed planar. Similarly, higher-order reflexions are caused by higher-order symmetries of these virtual sources with respect to the walls—or their spatial extension in the virtual space. Noting that the position of the virtual sources only depend on the position of the real source and the geometry of the room, the model for $p(\mathbf{r}, t)$ is now written as

$$p(\mathbf{r}, t) = \sum_{k=0}^K c_k \frac{\delta(t - \|\mathbf{s}_k - \mathbf{r}\|)}{4\pi \|\mathbf{s}_k - \mathbf{r}\|} \quad (6.11)$$

for $\mathbf{r} \in \mathcal{D}$, $t < t_{mix}$, where K is the number of real and virtual sources in the ball of radius κt_{mix} around the real source, κ is the sound velocity in air (typically $\kappa = 340 \text{ m.s}^{-1}$), \mathbf{s}_k is the position of virtual source k , and c_k is the corresponding intensity—taking into account some possible attenuation at the reflexion. The denominator simply expresses the free-field geometrical attenuation, where the energy of the impulse gets evenly spread on a sphere of growing area during the propagation. In short, for $t < t_{mix}$, the plenacoustic function

(for a fixed source at \mathbf{r}_0) is entirely determined by a linear combination of impulses from a *sparse* set of virtual sources within a finite spherical domain.

Given some measurements (pressure signals at a number of microphones), the model is estimated by looking for a sparse number of (real and virtual) sources, whose combination optimally models the observed data. Considering the size of the problem, greedy searches are often used for this task.

Note that this assumption of sparsity in the time domain of the impulse responses can also be exploited for the simultaneous measurement of the room impulse responses from different source locations [2], as it can be shown that this problem is equivalent to the estimation of the mixing filters in the context of convolutive source separation.

- **A frequency viewpoint:** At low frequencies, below the so-called Schroeder frequency f_{Sch} , a Fourier transform of the impulse responses shows isolated peaks that correspond to the modal response of the room. Above f_{Sch} , again the modal density gets too high to be able to isolate peaks with a clear physical meaning. There is therefore sparsity in the frequency domain below f_{Sch} , but the modes themselves have a very specific spatial distribution. Actually, for a given modal frequency f_0 , and sufficiently far from the walls to neglect evanescent waves, the mode only contains the wavelength $\lambda = \kappa/f_0$. In other words, the modes are entirely described by (infinitely many) plane waves $e^{i(\mathbf{k}\cdot\mathbf{r}-2\pi f_0 t)}$, with the wave vector \mathbf{k} of fixed modulus $\|\mathbf{k}\| = 2\pi f_0/\kappa$. Now, the (discretized) full spatio-temporal model is written as

$$p(\mathbf{r}, t) = \sum_{r=1}^R \sum_{p=1}^P c_{r,p} e^{i(\mathbf{k}_{r,p}\cdot\mathbf{r}-2\pi f_r t)} \quad (6.12)$$

with $\|\mathbf{k}_{r,p}\| = 2\pi f_r/\kappa$, R is the number of modes (sparse in frequency, *i.e.* only few modal f_r are significant), P is the number of plane waves used to discretize the 3D sphere of all possible directions, and the $c_{r,p}$ are the corresponding coefficients.

Given some measurements, this model is estimated in two steps: first, the sparse set of modal frequencies is estimated, jointly across microphone signals. Then, at a given modal frequency f_r , the $c_{r,p}$ coefficients are estimated by least-squares projections on the discrete set of plane waves $e^{i(\mathbf{k}\cdot\mathbf{r}-2\pi f_0 t)}$, computed at the sensor positions.

The above-described model has been tested in real experimental conditions [23, 24], with 120 microphones distributed within a $2 \times 2 \times 2$ m volume (with an approximately uniform distribution, as displayed in Figure 6.5), in a large room with strong reverberation. In a leave-one-out setting (model built using 119 microphone signals, tested on the remaining one), the above-described model leads to accurate interpolation of the impulse responses within the whole volume (see Figure 6.6 for an example), but with a precision significantly decreased near the edges of the volume. Actually, it was observed that this method gave poor results for extrapolation of the plenacoustic function outside the volume of interest where



Fig. 6.5 Microphone array used to sample the plenacoustic function. (a) Picture of the 120-microphone array in the room. The black omnidirectional source can be seen on the left. (b) Geometry of the microphone array. Blue dots indicate microphone capsules. From [24]

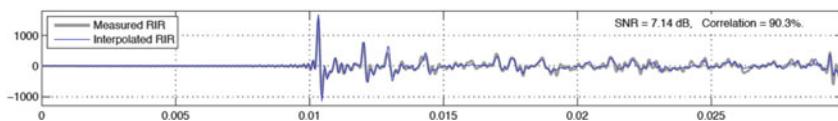


Fig. 6.6 Recorded impulse response at one microphone location (gray line), and interpolated impulse response (blue line) at the same location, using the time-sparse model. From [23]

measurements are performed. It should be noted that a similar CS technique [34] has been used to interpolate the sound field for the reconstruction of a spatially encoded echoic sound field, as an alternative to the widely used Higher Order Ambisonics techniques.

NB: The method presented here only performed compressive-sensing-based interpolation on the part of the plenacoustic function where some suitable sparse model could be established: at the beginning of the response and at low frequencies. While these are here treated independently, their joint processing could likely be beneficial and will be the focus of future work, see also Chapter 5.

6.3.3 *Medical ultrasound (scenario c)*

Ultrasonography is a widespread medical imaging method relying on the propagation of acoustic waves in the body. A wave emitted by a probe of piezoelectric transducers travels through the soft tissues and partially back-propagates to the probe again, revealing the presence of scatterers, similarly to the aforementioned scenario of underwater imaging. Classical ultrasonography uses beamforming both at emission and reception to scan slices of the targeted region of the body, providing two-dimensional images. In this imaging mode called “B-mode,” many emissions are performed successively, each time with a different beam orientation. For this

reason, medical ultrasonography is known to be a very user-dependent technique, relying on the ability of the practitioner to precisely handle the probe and to mentally figure out a 3D scene of the structures visualized in 2D. A natively 3D acquisition would lessen the dependency on the echographist movements. However, it would imply the use of a matrix array of transducers, meaning a high number of elements, which cannot be activated all at the same time due to technical limitations.

The context of ultrasonography sounds like a favorable context for the deployment of compressive sensing strategies, with a spectrum of expected improvements unsurprisingly ranging from a reduction of the data flow and a decrease of the sampling rate to better image resolution. Attempts to show the feasibility of compressive sensing in this context are however very recent and still far from technological applicability [20].

6.3.3.1 Sparse modeling of the ultrasound imaging process

As in other scenarios, the first step to compressive sensing is to identify the sources of sparsity of the problem. As ultrasonography is concerned, several routes have been taken, placing the sparsity hypothesis at different stages of the ultrasonography workflow.

A first approach is to assume a sparse distribution of scatterers in the body [32, 35]. By considering that echogeneity of a very large proportion of the insonified region is zero, the raw signals m received at each probe transducers after emission of a plane wave are modeled as:

$$m = Ac, \text{ with } c \text{ sparse.} \quad (6.13)$$

where the matrix A embeds the emitted signals as well as propagation effects, and c is a supposedly sparse scattering distribution (or “diffusion map”). The dimension of c is the number of voxels in the discretized space, and its estimation directly gives the reconstructed image. This is very reminiscent of the aforementioned underwater acoustics scenario. Such a sparsity hypothesis turns out to be efficient on simulated data which comply with the model but suffers from its bad representation of speckle patterns, which convey important information for practitioners.

To circumvent this issue, other authors rather suppose the sparsity of the raw received signals in some appropriate basis:

$$m = \Psi c \text{ with } c \text{ sparse.} \quad (6.14)$$

Fourier, wavelets and wave-atom bases have been used, with evidence for outperformance of the latter on signals obtained by classical pre-beamformed pulse-echo emissions [21]. After estimation of the support and coefficient \hat{c} , images are obtained by post-beamforming on $\hat{m} = \Psi \hat{c}$.

A third hypothesis places the sparsity assumption later in the process. Images y obtained from m by conventional post-beamforming are supposed to possess a sparse 2D Fourier transform [29]:

$$y = F^{-1}c \text{ with } c \text{ sparse,} \quad (6.15)$$

and sparse recovery is performed in the spatial frequency domain. This third approach is designed to ease the choice of a sensing matrix that would be compatible with instrumentation constraints while remaining incoherent with the sparsity basis.

6.3.3.2 Doppler and duplex modes

Medical ultrasound scanners can also be used for visualizing the blood flow dynamically. This so-called Doppler-mode repeatedly measures the flow velocity at a given position to recover its distribution over time, by pulsing multiple times in the same directions and exploiting the Doppler effect. Doppler-mode ultrasonography can be acquired alone but is also often acquired together with the B-mode. This duplex imaging implies alternating emission modes and is traditionally obtained by halving the time devoted to each mode. Compressive sensing strategies such as random mode alternation have been shown to be efficient [31]. Doppler signals are approximately sparse in the Fourier domain, which is a favorable situation here for compressive sensing (incoherence between sensing basis and sparsity basis, technical ease of random subsampling). The “savings” from compressive sensing of Doppler signals can then be reinvested in B-mode.

6.3.3.3 Subsampling and compressive sensing strategies

Most of these early work in compressive ultrasonography validate the chosen sparse model by random removal of a certain amount of samples among all acquired, or random linear combinations of the acquired signals. Though it is a valuable step to show the feasibility of compressive sensing in this context, the shift from sparse modeling and simulated subsampling to actual compressive sensing devices remains to be done. Indeed, spatially uniform random acquisition of the raw signals is technically as costly as acquiring them all. Different subsampling masks and their practical feasibility are discussed in [29] and in particular, the possibility to completely disconnect some elements of the probe, which would reduce acquisition time and data flow (especially for 3D imaging with matrix arrays).

Reduction of the sampling rate by exploiting the specificities of the used ultrasound signals (narrowbandness and finite rate of innovation) has also been proposed through the Xampling scheme [35]. Among proposed compressive sensing strategies, this is probably the closest to hardware feasibility, but remains far from an actual 3D imaging.

6.4 Beyond sparsity

Just as with other imaging modalities, acoustic imaging can benefit from low-dimensional models that go beyond traditional dictionary-based sparse models.

6.4.1 Structured sparsity

6.4.1.1 Localization of directive sources (scenario d)

The sparse models described above can be adapted to the joint localization and characterization of sources, in terms of their directivity patterns. Indeed, in practice every acoustic source has a radiation pattern that is non-uniform in direction, and furthermore this directivity pattern is frequency-dependent. Measuring these 3D directivity patterns usually requires a lengthy measurement protocol, with a dense array of microphones at a fixed distance of the source (usually 1m or 2m). The radiation pattern is usually described as its expansion on spherical harmonics of increasing order: monopolar, dipolar, quadrupolar, etc. Restricting the expansion to a finite order L ($L = 0$ for monopolar only, $L = 1$ for monopolar and dipolar, etc.), one can write the sound field in polar coordinates (r, θ, φ) at wavenumber k for a source located at the origin:

$$p(kr, \theta, \varphi) = \sum_{l=0}^L \sum_{q=-l}^l c_l^q(k) h_l(kr) Y_l^q(\theta, \varphi) \quad (6.16)$$

where the Y_l^q are the spherical harmonic of degree l and order q , h_l the propagative Hankel functions of order l , and c_l^q the corresponding coefficients.

This can be used to build a *group-sparse* model for the field produced by a sparse number of sources with non-uniform radiation: the sparsity in space restricts the number of active locations; for a given active location, all the coefficients of the corresponding spherical harmonic decomposition are non-zero. From a number of pressure measurements at different locations, the inverse problem amounts to finding both source location and directivity pattern. Using the group-sparse model, this can, for instance, be solved using an ℓ_1/ℓ_2 type of penalty on the set of activity coefficients (reorganized in column form for each location), or group-OMP.

The experimental results raise the interesting issue of the sampling step for the spatial locations. If the actual sources are on sampling points, or very close, the model successfully identifies the radiation pattern at least up to order $L = 2$ (quadrupolar). However, a source located between sampling points will appear as a linear combination of two or more sources with complex radiation coefficients: for instance, in the simplest case a dipolar source may appear as a combination of 2 neighboring monopoles in opposite phase, but more complex combinations also

arise, where the solutions eventually cannot easily be given a physical interpretation. Future work would therefore have to investigate sparse optimization on continuous parameter space [13].

6.4.1.2 Interpolation of plate vibration responses (scenario e)

In the NAH case described in section 6.2, the different plane waves, spatially restricted to the domain of the plate, could be selected independently. Actually, similarly to the plenacoustic case above, there are some further constraints that can be enforced for further modeling: the set of selected wave vectors must be of fixed modulus $\|\mathbf{k}\|$. However, in plates we may not know in advance the dispersion relation, linking the temporal frequency f to the spatial wavelength λ , or equivalently to the wavenumber $\|\mathbf{k}\|$. Hence, the problem may be recast as finding the best value of $\|\mathbf{k}\|$, such that the linear combination of plane waves *constrained to* a given wavenumber $\|\mathbf{k}\|$ maximally fits the observed data. In [7], this principle was employed for the interpolation of impulse responses in a plate, from a set of point-like measurements obtained by laser velocimetry, randomly chosen on the plate. Results show that accurate interpolation on the whole plate was possible with a number of points significantly below the spatial Nyquist range, together with an estimation of the dispersion relation. Interestingly, similar results also held with a *regular* sampling of the measurement points: constraining the wave vectors to lie on a circle allows us to undo the effect of spatial aliasing.

6.4.2 Cosparsity

Some of the aforementioned sparse modeling of acoustic fields, as a pre-requisite to the deployment of a compressive sensing strategy, rely on the ability to build a dictionary of solutions (or approximate solutions) of the wave equation ruling the propagation of the target acoustic field, such as the spherical waves in Eq. (6.11) or plane waves in Eq. (6.5) and Eq. (6.12). In most cases, exact closed-form expressions of these *Green's functions* do not exist, and computationally costly numerical methods have to be used. Moreover, the resulting dictionary Ψ is usually dense and its size grows polynomially with dimensions, which will cause tractability issues when used in solving the corresponding optimization problem in real scale conditions.

An idea to circumvent these issues would be to find an alternative model which would not rely on “solving” the wave equation, but rather on the wave equation itself. The so-called *cosparseness modeling* described in this section (see also Chapter 11) offers such an alternative. It also happens to have the potential to reduce the computational burden.

Let us first recall the wave equation obeyed by the (continuous) sound pressure field $p(\mathbf{r}, t)$ at position \mathbf{r} and time t :

$$\Delta p(\mathbf{r}, t) - \frac{1}{\kappa^2} \frac{\partial^2 p(\mathbf{r}, t)}{\partial t^2} = \begin{cases} 0, & \text{if no source at location } \mathbf{r} \\ f(\mathbf{r}, t), & \text{if source at location } \mathbf{r} \end{cases} \quad (6.17)$$

where Δ is the spatial Laplacian operator and the constant κ is the sound propagation speed in the medium. This can be concisely written as $\square p(\mathbf{r}, t) = f(\mathbf{r}, t)$ where \square denotes the linear D'Alembertian wave operator. Discretizing the signal in time and space, as well as the operator \square which becomes a matrix denoted Ω (augmented with the initial and boundary conditions, so as to define a determined system of linear equations), and finally assuming that the number of sound sources is small compared to the size of the spatial domain, reconstruction of the pressure field p from measurements y can be expressed as the following optimization problem:

$$\min_p \|\Omega p\|_0 \text{ s.t. } \|y - Ap\|_2 \leq \epsilon \quad (6.18)$$

The measurement matrix A is obtained by selecting, in the identity matrix, the rows corresponding to the microphone locations. This formulation can be directly compared to the counterpart sparse optimization, which consisted in minimizing the number of non-zeros in the expansion coefficients c such that $p = \Psi c$. Here, they have been replaced by the sparse product $z = \Omega p$. It is easy to see that in this special case, with $\Psi = \Omega^{-1}$, both problems are equivalent². However, while the dictionary Ψ is dense, the operator Ω obtained by a first-order finite-difference-method (FDM) discretization is extremely sparse: Ω has exactly 7 non-zero coefficients per row, no matter the global dimension of the problem, and is thus easy and cheap to compute, store, and apply in a matrix product.

Most of the well-known sparse recovery algorithms can be adapted to fit the cosparse recovery problem: greedy schemes [15, 27], convex relaxation [18, 26]. Thanks to the strong structural properties of sparsity and shift-invariance of Ω , they can be implemented in very efficient ways.

Solving the cosparse problem from a set of incomplete measurements $y = Ap$ produces different levels of outputs, that can be used in different applicative scenarios:

- **Source localization:** determining the *support* (locations of nonzero entries in the product Ωp) gives straightforwardly an estimation of the source locations.
- **Source identification:** once \hat{p} is determined, the product $\Omega \hat{p}$ gives estimations of the source signals $f(\mathbf{r}, t)$ at their estimated locations.
- **Field reconstruction:** \hat{p} is itself an estimation of the sound pressure field in the whole domain \mathcal{D} and at all instants.

²This is no longer true as soon as Ψ and Ω are not square and invertible matrices [14].

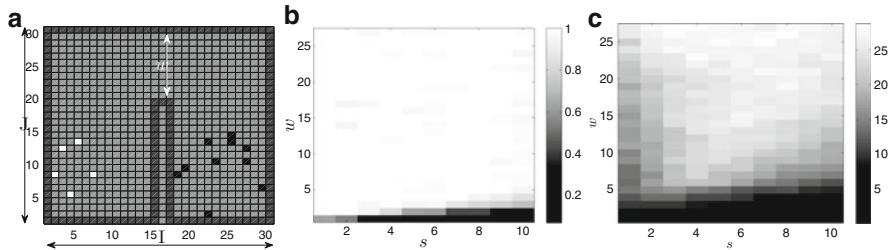


Fig. 6.7 A simulated compressive source localization and identification scenario. (a) 10 microphones (in black) and s sound sources (in white) are placed opposite sides of an inner wall in a shoebox room. A “door” of w pixels is open in the wall. (b) Varying the size of the inner wall and the number of sources, the empirical probability of correct source localization is very high in many situations. (c) The sound pressure field p is estimated with a relatively satisfying signal-to-noise ratio ($\text{SNR}_p = 20\log_{10} \|p\|_2 / \|p - \hat{p}\|_2$) in these conditions. This experiment scales up in 3D, while the equivalent sparse synthesis solver becomes intractable. From [19]

Localization has been the main target of early work in this domain [18, 28]. An illustration of such a scenario, taken from [19], is given in Figure 6.7. Cosparse modeling has thus been proven efficient in situations where typical geometric approaches fail and where the now traditional sparse synthesis approach fails to scale up computationally speaking.

As for source and field reconstruction, more investigations remain to be done on this emerging approach. Some other potentialities of the cosparse modeling can be envisioned. The capability to, at least partially, *learn* an operator Ω that would be known almost everywhere (but not at all boundaries for instance), or known only up to a physical parameter (such as the sound velocity κ), also contributes to make this modeling particularly attractive.

6.5 Conclusion / discussion

Acoustics offers a large playground where compressive sensing can efficiently address many imaging tasks. A non-exhaustive sample of such scenarios has been introduced in this chapter. We hope these are representative enough to enable the reader to connect the acoustic world to the main principles of compressive sensing, and to make his/her mind on how advances in compressive sensing may disseminate in acoustics.

Some good news for applying the theory of compressive sensing to acoustics is that, in many scenarios, a sparsity assumption naturally emerges from physics: a sparse distribution may be assumed for objects in space, for plane waves in a domain of interest, for early echoes in time, or peaks in modal responses. A cosparse modeling also happens to relate to the wave equation.

Another noticeable specificity of acoustic signal processing is that conventional acquisition devices (point microphones, sensing in the time domain) provide measures that are “naturally” incoherent with the sparsity basis of acoustic waves (Fourier basis), leading to a favorable wedding between acoustic applications and compressive sensing theory. In a way, many traditional sound acquisition and processing tasks, such as underdetermined sound source separation or other common settings with few microphones can be seen as ancestors of compressive sensing, even if not explicitly stated as such.

These two elements give the most encouraging signs towards the actual development of compressive sensing in its now well-established meaning. Thus, one can now hope to use compressive sensing in acoustics with several goals in mind:

- reducing the cost of hardware by using less sensors;
- reducing the data complexity, including acquisition time, data flow, and storage;
- improving the accuracy of the results by opening the door to super resolution.

In practice, the specific constraints of each application – *e.g.*, real-time processing—often tell which of those promises from the theory of compressive sensing can be achieved.

Now that it is possible to handle high-dimensional objects and to image 3D-regions, important questions remain open about the actual devices to be designed.

First, the theory of compressive sensing generally assumes that the sensing device is perfectly known. In practice, some parameters may vary and have to be estimated. They include the calibration of sensor gain, phase, and positions for instance. This is even more challenging for large arrays with many cheap sensors that suffer from a large variability in sensor characteristics. Preliminary studies have shown that sparse regularization can help to adapt the sensing matrix in the case of unknown gains and phase [3–5].

The design of new sparse models is another challenge, *e.g.*, to model the directivity of sensors or of scattering material using structured sparsity, or by studying how it can relate to the speckle distribution. Compressive acoustic imaging also calls for new views on the interplay between discrete and continuous signal processing, especially to handle the challenges of 3D imaging of large regions of interest.

Eventually, the last major missing step towards implementation of real-life compressive sensing acoustic device is now, in many scenarios, the possibility to build or adapt hardware to compressive sensing requirements, such as randomness in a subsampling scheme or incoherence with the sparsity basis, while actually getting some gain compared to conventional state of the art. Feasibility of compressive sensing is often shown by simulating random subsampling, keeping a certain percentage of all collected samples and exhibiting satisfying signal reconstruction from those samples. Hardware which performs such random subsampling can be simply as costly and complicated to build as conventional hardware: acquiring more samples than needed to finally drop unneeded samples is obviously a suboptimal strategy to reduce acquisition time and cost. The shift from theory and proofs of

concept to actual devices with practical gains, shown here in the case of nearfield acoustic holography, is now one of the next main challenges in many other acoustic compressive sensing scenarios.

Acknowledgements The authors wish to warmly thank François Ollivier, Jacques Marchal, and Srdjan Kitic for the figures, as well as Gilles Chardon, Rémi Mignot, Antoine Peillot, and the colleagues from the ECHANGE and PLEASE project whose contributions have been essential in the work described in this chapter. This work was supported in part by French National Research, ECHANGE project (ANR-08-EMER-006 ECHANGE) and by the European Research Council, PLEASE project (ERC-StG-2011-277906).

References

1. Ajdler, T., Vetterli, M.: Acoustic based rendering by interpolation of the plenacoustic function. In: SPIE/IS &T Visual Communications and Image Processing Conference, pp. 1337–1346. EPFL (2003)
2. Benichoux, A., Vincent, E., Gribonval, R.: A compressed sensing approach to the simultaneous recording of multiple room impulse responses. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics 2011, p. 1, New Paltz, NY, USA (2011)
3. Bilen, C., Puy, G., Gribonval, R., Daudet, L.: Blind Phase Calibration in Sparse Recovery. In: EUSIPCO - 21st European Signal Processing Conference-2013, IEEE, Marrakech, Maroc (2013)
4. Bilen, C., Puy G., Gribonval, R., Daudet, L.: Blind Sensor Calibration in Sparse Recovery Using Convex Optimization. In: SAMPTA - 10th International Conference on Sampling Theory and Applications - 2013, Bremen, Germany (2013)
5. Bilen, C., Puy, G., Gribonval, R., Daudet, L.: Convex optimization approaches for blind sensor calibration using sparsity. *IEEE Trans. Signal Process.* **99**(1), (2014)
6. Boufounos, P.: Compressive sensing for over-the-air ultrasound. In: Proceedings of ICASSP, pp. 5972–5975, Prague, Czech Republic, (2011)
7. Chardon, G., Leblanc, A., Daudet L.: Plate impulse response spatial interpolation with sub-nyquist sampling. *J. Sound Vib.* **330**(23), 5678–5689 (2011)
8. Chardon, G., Bertin, N., Daudet, L.: Multiplexage spatial aléatoire pour l'échantillonnage compressif - application à l'holographie acoustique. In: XXIIIe Colloque GRETSI, Bordeaux, France (2011)
9. Chardon, G., Daudet, L., Peillot, A., Ollivier, F., Bertin, N., Gribonval, R.: Nearfield acoustic holography using sparsity and compressive sampling principles. *J. Acoust. Soc. Am.* **132**(3), 1521–1534 (2012) Code & data for reproducing the main figures of this paper are available at <http://echange.inria.fr/nah>.
10. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic Decomposition by Basis Pursuit. *SIAM J. Sci. Comput.* **20**(1), 33–61 (1999)
11. Daudet, L.: Sparse and structured decompositions of signals with the molecular matching pursuit. *IEEE Trans Audio Speech Lang. Process.* **14**(5), 1808–1816 (2006)
12. Duarte, M.F., Sarvotham, S., Baron, D., Wakin, M.B., Baraniuk, R.G.: Distributed compressed sensing of jointly sparse signals. In: Proceedings of Asilomar Conference Signals, Systems and Computers, pp. 1537–1541 (2005)
13. Ekanadham, C., Trachina, D., Simoncelli, E.P.: Recovery of sparse translation-invariant signals with continuous basis pursuit. *IEEE Trans. Signal Process.* **59**(10), 4735–4744 (2011)
14. Elad, M., Milanfar, P., Rubinstein, R.: Analysis versus synthesis in signal priors. *Inverse Prob.* **23**, 947–968 (2007)

15. Giryes, R., Nam, S., Elad, M., Gribonval, R., Davies, M.E.: Greedy-Like algorithms for the cosparse analysis model. arXiv preprint arXiv:1207.2456 (2013)
16. Gribonval, R., Bacry, E.: Harmonic decomposition of audio signals with matching pursuit. *IEEE Trans. Signal Process.* **51**(1), 101–111 (2003)
17. Gribonval, R., Chardon, G., Daudet, L.: Blind Calibration For Compressed Sensing By Convex Optimization. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, Kyoto, Japan (2012)
18. Kitić, S., Bertin, N., Gribonval, R.: A review of cosparse signal recovery methods applied to sound source localization. In: Le XXIVe colloque Gretsi, Brest, France (2013)
19. Kitic, S., Bertin, N., Gribonval, R.: Hearing behind walls: localizing sources in the room next door with cosparsity. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy (2014)
20. Liebgott, H., Basarab, A., Kouamé, D., Bernard, O., Friboulet, D.: Compressive Sensing in Medical Ultrasound. pp. 1–6. IEEE, Dresden (Germany) (2012)
21. Liebgott, H., Prost, R., Friboulet, D.: Pre-beamformed RF signal reconstruction in medical ultrasound using compressive sensing. *Ultrasonics* **53**(2), 525–533 (2013)
22. Maynard, J.D., Williams, E.G., Lee, Y.: Nearfield acoustic holography: I, theory of generalized holography and the development of NAH. *J. Acoust. Soc. Am.* **78**(4), 1395–1413 (1985)
23. Mignot, R., Chardon, G., Daudet, L.: Low frequency interpolation of room impulse responses using compressed sensing. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(1), 205–216 (2014)
24. Mignot, R., Daudet, L., F. Ollivier. Room reverberation reconstruction: Interpolation of the early part using compressed sensing. *IEEE Trans. Audio, Speech, Lang. Process.* **21**(11), 2301–2312 (2013)
25. Moiola, A., Hiptmair, R., Perugia, I.: Plane wave approximation of homogeneous helmholtz solutions. *Z. Angew. Math. Phys. (ZAMP)* **62**, 809–837 (2011). 10.1007/s00033-011-0147-y
26. Nam, S., Davies, M.E., Elad, M., Gribonval, R.: The cosparse analysis model and algorithms. *Appl. Comput. Harmon. Anal.* **34**(1), 30–56 (2013)
27. Nam, S., Davies, M.E., Elad, M., Gribonval, R.: Recovery of cosparse signals with greedy analysis pursuit in the presence of noise. In: 2011 4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pp. 361–364. IEEE, New York (2011)
28. Nam, S., Gribonval, R.: Physics-driven structured cosparse modeling for source localization. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5397–5400. IEEE, New York (2012)
29. Quinsac, C., Basarab, A., Kouamé, D.: Frequency domain compressive sampling for ultrasound imaging. *Adv. Acoust. Vib. Adv. Acoust. Sens. Imag. Signal Process.* **12** 1–16 (2012)
30. Rauhut, H.: Random sampling of sparse trigonometric polynomials. *Appl. Comput. Harmon. Anal.* **22**(1), 16–42 (2007)
31. Richy, J., Liebgott, H., Prost, R., Friboulet, D.: Blood velocity estimation using compressed sensing. In: IEEE International Ultrasonics Symposium, pp. 1427–1430. Orlando (2011)
32. Schifflner, M.F., Schmitz, G.: Fast pulse-echo ultrasound imaging employing compressive sensing. In: Ultrasonics Symposium (IUS), 2011 IEEE International, pp. 688–691. IEEE, New York (2011)
33. Stefanakis, N., Marchal, J., Emiya, V., Bertin, N., Gribonval, R., Cervenka, P.: Sparse underwater acoustic imaging: a case study. In: Proceeding of International Conference Acoustics, Speech, and Signal Processing. Kyoto, Japan (2012)
34. Wabnitz, A., Epain, N., van Schaik, A., Jin, C.: Time domain reconstruction of spatial sound fields using compressed sensing. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 465–468. IEEE, New York (2011)
35. Wagner, N., Eldar, Y.C., Feuer, A., Danin, G., Friedman, Z.: Xampling in ultrasound imaging. SPIE Medical Imaging Conference, 7968 (2011)

Chapter 7

Quantization and Compressive Sensing

Petros T. Boufounos, Laurent Jacques, Felix Krahmer, and Rayan Saab

Abstract Quantization is an essential step in digitizing signals, and, therefore, an indispensable component of any modern acquisition system. This chapter explores the interaction of quantization and compressive sensing and examines practical quantization strategies for compressive acquisition systems. Specifically, we first provide a brief overview of quantization and examine fundamental performance bounds applicable to any quantization approach. Next, we consider several forms of scalar quantizers, namely uniform, non-uniform, and 1-bit. We provide performance bounds and fundamental analysis, as well as practical quantizer designs and reconstruction algorithms that account for quantization. Furthermore, we provide an overview of Sigma-Delta ($\Sigma\Delta$) quantization in the compressed sensing context, and also discuss implementation issues, recovery algorithms, and performance bounds. As we demonstrate, proper accounting for quantization and careful quantizer design has significant impact in the performance of a compressive acquisition system.

P.T. Boufounos (✉)

Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA 02139, USA
e-mail: petrosb@merl.com

L. Jacques

ISPGGroup, ICTEAM/ELEN, Université catholique de Louvain, Place du Levant 2,
PO box L5.04.04, B1348 Louvain-la-Neuve, Belgium
e-mail: laurent.jacques@uclouvain.be

F. Krahmer

Technische Universität München, Boltzmannstr. 3, 85748 Garching bei München
e-mail: felix.krahmer@tum.de

R. Saab

University of California, San Diego, 9500 Gilman Drive #0112, La Jolla, CA 92093-0112, USA
e-mail: rsaab@ucsd.edu

7.1 Introduction

In order to store and manipulate signals using modern devices, it is necessary to digitize them. This involves two steps: sampling (or measurement) and quantization. The compressed sensing theory and practice described in the remainder of this book provides a novel understanding of the measurement process, enabling new technology and approaches to reduce the sampling burden. This chapter explores the very interesting interaction of compressed sensing with quantization.

Sampling maps a signal to a set of coefficients, typically using linear measurements. This map can often be designed to be lossless, i.e., to perfectly represent all signals in a certain class, as well as robust to noise and signal modeling errors. The Nyquist theorem, as well as more recent compressive sampling theorems are examples of such sampling approaches [9, 25, 75].

The guarantees in sampling theorems are typically stated in terms of the critical measurement rate, i.e., the number of measurements necessary to perfectly represent signals in a given class. Oversampling, compared to that minimum rate, typically provides robustness to errors in the representation, noise in the acquisition, and mismatches in signal models. The latter is especially important in compressive sensing systems as they provide perfect reconstruction guarantees for exactly sparse signals; in practice, the acquired signal is almost never exactly sparse.

Quantization, on the other hand, is the process of mapping the representation coefficients—which potentially belong to an uncountably infinite set—to elements in a finite set, and representing them using a finite number of bits. Due to the many-to-one nature of such a map, the quantized representation is in general lossy, i.e., distorts the representation and, therefore, the signal. This distortion occurs even if the measurement process is lossless.

The interaction of quantization with sampling introduces interesting trade-offs in the acquisition process. A system designed to sample signals at (or slightly above) the critical rate may be less robust to errors introduced by quantization. Consequently, it requires a sophisticated quantizer design that ensures very small quantization errors. On the other hand, a simpler quantizer architecture (e.g., with fewer bits per measurement) could introduce significant error to the representation and require some oversampling to compensate. Practical systems designs navigate this trade-off, for example, according to the complexity of the corresponding hardware.

Compressive acquisition systems amplify the importance of the trade-off between quantizer complexity and oversampling. The sampling rate is significantly reduced in such systems, at the expense of increased sensitivity to noise and signal model mismatch. Thus, loss of information due to quantization can be detrimental, especially when not properly handled. One may revert to oversampling here as well, however the incoherent and often randomized nature of compressive measurements poses challenges. Thus, powerful oversampling based quantization approaches, such as Sigma-Delta quantization can be applied, but only after careful consideration.

Nevertheless, the sparse signal models and the computational methods developed for compressed sensing can alleviate a number of performance bottlenecks due to quantization in conventional systems. Using computational approaches originating in frame theory and oversampling, it is possible to significantly reduce the distortion due to quantization, to significantly improve the performance due to saturation, and to enable reconstruction from measurements quantized as coarsely as 1 bit. The theory and practice for such methods are described in Sec. 7.3.

It might seem counterintuitive that compressed sensing attempts to remove sampling redundancy, yet successful reconstruction approaches employ tools developed for oversampled representations. In fact there is a strong connection between compressed sensing and oversampling, which we explore in various points in this chapter. Furthermore, with sufficient care, this connection can be exposed and exploited to implement Sigma-Delta quantization in CS-based acquisition systems, and significantly improve performance over scalar quantization. The details are discussed in Sec. 7.4.

The next section presents general principles of quantization, including a brief background on vector, scalar, and Sigma-Delta quantization for general acquisition systems. It is not an exhaustive survey of the topic. For this we refer the reader to [32, 39, 77]. Instead, it serves to establish notation and as quick reference for the subsequent discussion. Sec. 7.3 and Sec. 7.4 examine the interaction of compressive sensing and quantization in significant detail. Sec. 7.5 concludes with some discussion of the literature, promising directions and open problems.

Notation: In addition to the notational conventions defined in Chapter 1, this chapter also uses the following general notations. The logarithm in base $a > 0$ is noted \log_a and whenever the base is not specified, \log refers to the natural logarithm. Note that in some cases, such as asymptotic results, the logarithm base is not important. This chapter also uses the following non-asymptotic orderings: For two functions f and g , we write $f \lesssim g$ if there exists a constant $C > 0$ independent of the function arguments such that $f \leq Cg$, with a similar definition for $f \gtrsim g$. Moreover, $f \asymp g$ if we have both $f \lesssim g$ and $f \gtrsim g$. Occasionally, we also rely on the well-established big- O and big- Ω asymptotic notation to concisely explain asymptotic behavior when necessary. More specific notation is defined at first occurrence.

7.2 Fundamentals of Quantization

For the purposes of this section, a quantizer operates on signals x , viewed as vectors in a bounded set $V \subset \mathbb{R}^n$. The goal of a quantizer $Q(\cdot)$ is to represent those signals as accurately as possible using a rate of R bits, i.e., using a quantization point $q = Q(x)$ chosen from a set of 2^R possible ones often referred to as codebook. Of course, when V contains an infinite number of signals, signals will be distorted through this representation.

In this section, we first define common quantization performance metrics and determine fundamental bounds on the performance of a quantizer. Then, in preparation for the next sections, we examine common approaches to quantization, namely scalar and Sigma-Delta quantization, which are very useful in compressive sensing applications.

7.2.1 Quantization Performance Bounds

To measure the accuracy of the quantizer we consider the distortion, i.e., the ℓ_2 distance of a quantization point from its original signal $\|x - Q(x)\|_2$. The overall performance of the quantizer is typically evaluated either using the average distortion over all the signals—often computed using a probability measure on the signal space V —or using the worst-case distortion over all signals in V . In this chapter, in the spirit of most of the compressed sensing literature, we quantify the performance of the quantizer using the worst-case distortion on any signal, i.e.,

$$\varepsilon = \sup_{x \in V} \|x - Q(x)\|_2. \quad (7.1)$$

This choice enables very strong guarantees, irrespective of the accuracy of any probabilistic assumption on the signal space.

A lower bound on the distortion of any quantizer can be derived by constructing a covering of the set V . A covering of radius r is a set of points q such that each element in V has distance at most r from its closest point in the covering. If we can construct a covering using P points, then we can also define a quantizer that uses $R = \lceil \log_2 P \rceil$ bits and has worst-case distortion $\varepsilon = r$ as each signal is quantized to the closest point in the covering.

To determine a lower bound for the number of points in such a covering, we consider balls of radius r centered at q , defined as

$$\mathcal{B}_r(q) = \{x \in \mathbb{R}^n \mid \|q - x\|_2 \leq r\}. \quad (7.2)$$

Since each signal in V is at most r away from some point in the covering, if we place a ball of radius r at the center of each point of the covering, then the union of those balls covers V . Thus, the total volume of the balls should be at least as large as the volume of the set, denoted $\text{vol}(V)$. Since the volume of a ball of radius r in n dimensions is $\text{vol}(\mathcal{B}_r(q)) = r^n \pi^{n/2} / \Gamma(1 + n/2)$, where $\Gamma(\cdot)$ is the Gamma function, the best possible error given the rate R can be derived using

$$\text{vol}(V) \leq \frac{\pi^{n/2}}{\Gamma(1 + \frac{n}{2})} 2^R r^n \Rightarrow r \gtrsim 2^{-\frac{R}{n}}. \quad (7.3)$$

In other words, the worst-case error associated with an optimal quantizer can, at best, decay exponentially as the bit rate increases. Moreover, the decay rate depends on the ambient dimension of the signal. In short,

$$\varepsilon \gtrsim 2^{-\frac{R}{n}}. \quad (7.4)$$

The smallest achievable worst-case distortion for a set is also known as the $(R+1)$ -dyadic entropy number of the set, whereas the number of bits necessary to achieve a covering with worst-case distortion equal to ε is known as the Kolmogorov ε -entropy or metric entropy of the set.

For the models commonly assumed in compressive sensing, these quantities are not straightforward to calculate and depend on the sparsity model assumed. For example, compressible signals are commonly modeled as being drawn from a unit ℓ_p ball, where $0 < p < 1$ (cf. Chapter 1 for a discussion on compressibility). In this case, the worst-case distortion is bounded by

$$\varepsilon \gtrsim \begin{cases} 1 & \text{if } 1 \leq R \leq \log_2 n \\ \left(\frac{1}{R} \log_2 \left(\frac{n}{R} + 1\right)\right)^{\frac{1}{p} - \frac{1}{2}} & \text{if } \log_2 n \leq R \leq n \\ 2^{-\frac{R}{n}} n^{\frac{1}{2} - \frac{1}{p}} & \text{if } R \geq n, \end{cases} \quad (7.5)$$

where the constant implicit in our nation is independent of R and n [24, 36, 64, 88].

In the case of exactly k -sparse signals, the volume of the union of subspaces they occupy has measure zero in the n -dimensional ambient space. However, by considering the $\binom{n}{k}$ k -dimensional subspaces and coverings of their unit balls, a lower bound on the error can be derived [18], namely

$$\varepsilon \gtrsim \frac{2^{-\frac{R}{k}} n}{k}. \quad (7.6)$$

Note that this lower bound can be achieved in principle using standard transform coding (TC), i.e., by first representing the signal using its sparsity basis, using $\log_2 \binom{n}{k} \lesssim k \log_2 (n/k)$ bits to represent the support of the non-zero coefficients and using the remaining bits to represent the signal in the k -dimensional subspace at its Kolmogorov entropy

$$\varepsilon_{\text{TC}} \lesssim 2^{-\frac{R - k \log_2 (n/k)}{k}} = \frac{2^{-\frac{R}{k}} n}{k}. \quad (7.7)$$

Unfortunately, compressive sensing systems do not have direct access to the sparse vectors. They can only access the measurements, $y = Ax$, which must be quantized upon acquisition—in practice using analog circuitry. Thus, transform coding is not possible. Instead, we must devise simple quantization algorithms that act directly on the measurements in such a way that permits accurate reconstruction.

7.2.2 *Scalar Quantization*

The simplest approach to quantization is known as *scalar quantization* and often referred to as *pulse code modulation* (PCM), or *memoryless scalar quantization* (MSQ). Scalar quantization directly quantizes each measurement of the signal, without taking other measurements into account. In other words a 1-dimensional, i.e., scalar, quantizer is applied separately to each measurement of the signal.

7.2.2.1 Measurement and Scalar Quantization

A scalar quantizer can be defined using a set of levels, $\mathcal{Q} = \{l_i \in \mathbb{R} : l_j < l_{j+1}\}$, comprising the quantization codebook, and a set of thresholds $\mathcal{T} = \{t_i \in \overline{\mathbb{R}} : t_j < t_{j+1}\}$, implicitly defining the quantization intervals $\mathcal{C}_j = [t_j, t_{j+1})$. Assuming no measurement noise, the quantizer is applied element-wise to the measurement coefficients, $y = Ax$, to produce the quantized measurements $q = Q(y)$, $q_i = Q(y_i)$. Using a rate of B bits per coefficient, i.e., $R = mB$ total bits, the quantizer represents $L = 2^B$ total levels per coefficient. A scalar value y_i quantizes to the quantization level corresponding to the quantization interval in which the coefficient lies.

$$Q(y_i) = l_j \Leftrightarrow y_i \in \mathcal{C}_j. \quad (7.8)$$

A scalar quantizer is designed by specifying the quantization levels and the corresponding thresholds. Given a source signal with measurements modeled as a continuous random variable X , a (*distortion*) *optimal* scalar quantizer minimizes the error

$$\mathbb{E}|X - Q(X)|^2. \quad (7.9)$$

Such an optimal quantizer necessarily satisfies the Lloyd-Max conditions [71, 74]

$$l_j = \mathbb{E}\{X | X \in \mathcal{C}_j\}, \quad t_j = \frac{1}{2}(l_j + l_{j+1}), \quad (7.10)$$

which define a fixed point equation for levels and thresholds and the corresponding fixed-point iteration—known as the Lloyd–Max algorithm—to compute them.

Alternatively, a simpler design approach is the uniform scalar quantizer, which often performs almost as well as an optimal scalar quantizer design. It is significantly less complex and can be shown to approach optimality as the bit-rate increases [39]. The thresholds of a uniform scalar quantizer are defined to be equi-spaced, i.e., $t_{j+1} - t_j = \Delta$, where Δ is referred to as the quantization bin width or *resolution*. The levels are typically set to the mid-point $l_j = \frac{1}{2}(t_j + t_{j+1})$ of the quantization bin \mathcal{C}_j . Thus, the quantization error introduced to each coefficient is bounded by $\Delta/2$. A uniform quantizer defines a uniform grid in the m -dimensional measurement space, as shown in Fig. 7.1.

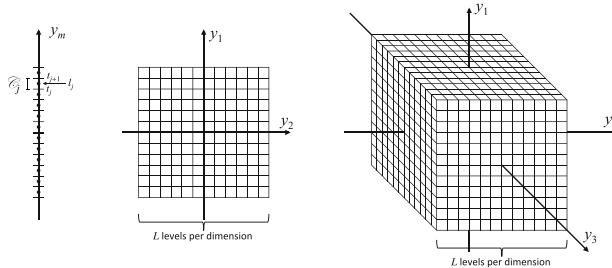


Fig. 7.1 A finite uniform scalar quantizer and the uniform grid it generates in 2 and 3 dimensions

In practical systems, the scalar quantizer has finite range, i.e., it saturates if the signal exceeds a saturation level S . In particular, a uniform finite-range scalar quantizer using B bits per coefficient has quantization interval $\Delta = S2^{-B+1}$. If a coefficient exceeds S , the quantizer maps the coefficient to the largest quantization level, i.e., it saturates. Depending on the magnitude of the coefficient, this may introduce significant error. However, it is often convenient in theoretical analysis to assume an infinite quantizer that does not saturate. This assumption is often justified, as S in practice is set large enough to avoid saturation given a signal class. As described in Sec. 7.3.3, this is often suboptimal in compressive sensing applications.

Compared to classical systems, optimal scalar quantizer designs for compressive sensing measurements require extra care. An optimal design with respect to the measurement error is not necessarily optimal for the signal, due to the non-linear reconstruction inherent in compressed sensing. While specific designs have been derived for very specific reconstruction and probabilistic signal models, e.g., [56, 89], a general optimal design remains an open problem. Thus the literature has focused mostly, but not exclusively, on uniform scalar quantizers.

7.2.2.2 Scalar Quantization and Oversampling

When a signal is oversampled, a scalar quantizer makes suboptimal use of the bit-rate. The k -dimensional signal space mapped through the measurement operator to an m -dimensional measurement space, where $m > k$, spans, at most, a k -dimensional subspace of \mathbb{R}^m , as shown in Fig. 7.2. As evident from the figure, this subspace intersects only a few of the available quantization cells and, therefore, does not use the available bits effectively. For an L -level quantizer, the number of quantization cells intersected $I_{k,m,L}$ is bounded by [14, 38, 90]

$$I_{k,m,L} \lesssim \left(\frac{Lm}{k} \right)^k \quad (7.11)$$

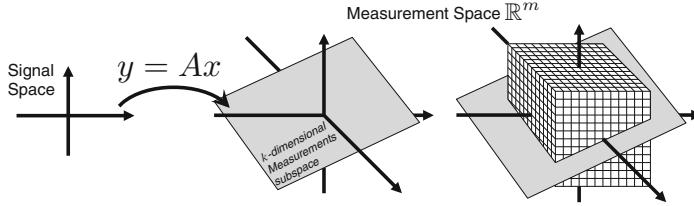


Fig. 7.2 A k -dimensional space measured using m measurements spans a k -dimensional subspace of \mathbb{R}^m and intersects only a few of the L^m available quantization cells

Using a simple covering argument as in Sec. 7.2.1, it is thus possible to derive a lower bound on the error performance as a function of the number of measurements m

$$\varepsilon \gtrsim \left(\frac{2^{-B}k}{m} \right) \quad (7.12)$$

The bounds hold for any scalar quantizer design, not just uniform ones.

Linear reconstruction, i.e., reconstruction using a linear operator acting on the scalar quantized measurements, does not achieve the bound (7.12) [38, 91]. The quantization error using linear reconstruction can only decay as fast as

$$\varepsilon \gtrsim \frac{2^{-B}k}{\sqrt{m}}. \quad (7.13)$$

Instead, *consistent* reconstruction achieves the optimal bound in a number of cases. Consistent reconstruction treats the quantization regions as reconstruction constraints and ensures that the reconstructed signal \hat{x} quantizes to the same quantization points when measured using the same system. Thus in the oversampled setting where A is an $m \times k$ matrix with $m > k$, and where $q = Q(Ax)$ one solves the problem:

$$\text{find any } \hat{x} \text{ s.t. } q = Q(A\hat{x}). \quad (7.14)$$

If the measurement operator A is a tight frame formed by an oversampled discrete Fourier transform (DFT), the root mean square error (RMSE) of such a reconstruction (with respect to a random signal model) decays as $O(1/m)$ [38, 91], i.e., as (7.12). In the case of random frames with frame vectors drawn independently from a Gaussian distribution [50] or from a suitable distribution on the $(m-1)$ -sphere [84], the reconstruction method in (7.14) also displays RMSE and worst case reconstruction error decreasing as $O(1/m)$ and $O((\log m)/m)$, respectively.

The constraints imposed by consistent reconstruction are convex and can be imposed on any convex optimization algorithm. This makes them particularly suitable for a number of reconstruction algorithms already used in compressive sensing systems, as we explore in Sec. 7.3.

The bounds (7.12) and (7.13)—which can be achieved with proper design of the measurement process and the reconstruction algorithm—demonstrate that the most efficient use of the rate $R = mB$ is in refining each measurement using more bits per measurement, B , rather than in increasing the number of measurements, m . They suggest that in terms of error performance, by doubling the oversampling it is possible to save 0.5 bits per coefficient if linear reconstruction is used and 1 bit per coefficient if consistent reconstruction is used. This means that a doubling of the rate by doubling the oversampling factor is equivalent to a linear increase in the rate by $m/2$ or m through an increase in B , for linear and consistent reconstruction, respectively. So in principle, if rate-efficiency is the objective, the acquisition system should only use a sufficient number of measurements to reconstruct the signal and no more. All the rate should be devoted to refining the quantizer. However, these bounds ignore the practical advantages in oversampling a signal, such as robustness to erasures, robustness to measurement noise, and implementation complexity of high-rate scalar quantizers. Thus in practice, oversampling is often preferred, despite the rate-inefficiency. Techniques such as Sigma-Delta quantization, which we discuss in Sec. 7.2.3, have been developed to improve some of the trade-offs and are often used in conjunction with oversampling.

7.2.2.3 Performance Bounds on Sparse Signals

Scalar quantization in compressive sensing exhibits similar bounds as scalar quantization of oversampled signals. Signals that are k -sparse in \mathbb{R}^n belong to a union of k -dimensional subspaces. When measured using m linear measurements, they occupy a union of k -dimensional subspaces of \mathbb{R}^m , $\binom{n}{k}$ of them. Using the same counting argument as above, it is evident that the number of quantization cells intersected, out of the L^m possible ones, is at most

$$\binom{n}{k} I_{k,m,L} \gtrsim \left(\frac{Lmn}{k^2} \right)^k \quad (7.15)$$

The resulting error bound is

$$\varepsilon \gtrsim \frac{2^{-B} k}{m} \quad (7.16)$$

$$\gtrsim \frac{2^{-\frac{R}{m}} k}{m}, \quad (7.17)$$

which decays slower than (7.6) as the rate increases keeping the number of measurements m constant. Furthermore, as the rate increases with the number measurements m , keeping B , the number of bits per measurement constant, the behavior is similar to quantization of oversampled frames: the error can only decay linearly with m .

These bounds are not surprising, considering the similarities of oversampling and compressive sensing of sparse signals. It should, therefore, be expected that more sophisticated techniques, such as Sigma-Delta ($\Sigma\Delta$) quantization should improve performance, as they do in oversampled frames. However, their application is not as straightforward. The next section provides an overview of $\Sigma\Delta$ quantization and Sec. 7.4 discusses in detail how it can be applied to compressive sensing.

7.2.3 Sigma-Delta Quantization

An alternative approach to the scalar quantization techniques detailed in the previous section is feedback quantization. The underlying idea is that the fundamental limits for the reconstruction accuracy discussed above can be overcome if each quantization step takes into account errors made in previous steps. The most common feedback quantization scheme is $\Sigma\Delta$ quantization, originally introduced for bandlimited signals in [47] (cf. [46]). A simple $\Sigma\Delta$ scheme, illustrated in Figure 7.3, shows this feedback structure.

A motivation in $\Sigma\Delta$ quantization is that, in some applications, reducing circuit complexity is desirable, even at the expense of a higher sampling rate. Indeed, $\Sigma\Delta$ designs drastically reduce the required bit depth per sample while allowing for accurate signal reconstruction using simple circuits. In fact, since its introduction, $\Sigma\Delta$ quantization has seen widespread use (see, e.g., [77] and the references therein) in applications ranging from audio coding to wireless communication.

Nevertheless, a mathematical analysis of $\Sigma\Delta$ quantization in its full generality has been challenging. A preliminary analysis of simple $\Sigma\Delta$ schemes for restricted input classes (including constant input and sinusoidal input) was presented in [40] and follow-up works. However, most of these results were limited to linear, or at best low-order polynomial error decay in the oversampling rate. This type of error decay is sub-optimal (albeit better than scalar quantization), and rather far from the optimal exponential error decay. Specifically, a major difficulty that prevented

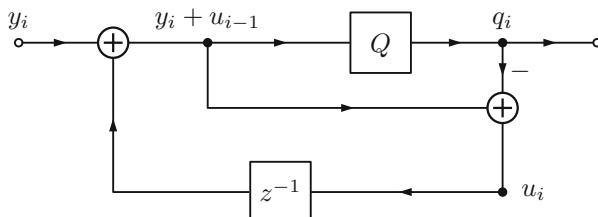


Fig. 7.3 A block diagram of a simple 1st order $\Sigma\Delta$ scheme: The input y_i is added to the state variable u_{i-1} (initialized as $u_0 = 0$) and the sum is scalar quantized. Subsequently, the state variable is updated as the difference between the scalar quantizer's input and its output. More complex designs, featuring higher order $\Sigma\Delta$ quantization with more feedback loops are possible. We discuss such designs in more detail in Section 7.4

a more comprehensive treatment was understanding the instabilities caused by the positive feedback inherent to the $\Sigma\Delta$ circuit designs. For example, depending on the design of the $\Sigma\Delta$ scheme, the state variables could grow without bound. A crucial idea to prevent such phenomena for arbitrary *band-limited* inputs was developed in [32]; their analysis led, for the first-time, to super-polynomial bounds for the error decay. To date, the best known error bounds decay exponentially in the oversampling rate [35, 41]. While this is near-optimal (optimal up to constants in the exponent), it has been shown that with a fixed bit budget per sample, the achievable rate-distortion relationship is strictly worse than for scalar quantization of Nyquist rate samples [63]. That said, increasing the bit budget per sample entails more expensive and complex circuitry, which grows increasingly costly with every added bit (in fact, the best current quantizers provide a resolution of about 20 bits per sample). Thus, for quantizing bandlimited functions, if one wishes to improve the performance or reduce the cost, one must revert to oversampling-based methods such as $\Sigma\Delta$ quantization.

The accuracy gain of $\Sigma\Delta$ quantization is most prominent when a significant oversampling rate and, therefore, a high redundancy of samples is inherent or desired. Such redundant representations can also be encountered in a finite-dimensional discrete context. Namely, this corresponds to a finite frame expansion in the sense of (1.32). This observation served as a motivation to devise $\Sigma\Delta$ schemes for finite-frame expansions, and the first such construction was provided in [7]. In contrast to oversampled representations of bandlimited signals, which directly correspond to a temporal ordering, finite frames generally do not have an inherent order, nor are the frame vectors necessarily close enough to each other to allow for partial error compensation. Due to this difficulty, the first works on $\Sigma\Delta$ quantization for finite frame expansions focus on frames with special smoothness properties. Namely, they assume that the frame $\Phi = \{\phi_j\}_{j=1}^N$ has a well-controlled *frame variation*

$$v_\Phi := \sum_{j=1}^{N-1} \|\phi_{j+1} - \phi_j\|_2.$$

The constructions in [7] coupled with (linear) reconstruction via the canonical dual frame (that is, the Moore-Penrose pseudo-inverse of the matrix that generates the redundant representation) was shown to yield an error decay on the order of $v_\Phi N^{-1}$, i.e., linear error decay whenever the frame variation is bounded by a constant. By using more sophisticated $\Sigma\Delta$ schemes these results were later improved to higher order polynomial error decay [6, 12, 13] in the number of measurements, thereby beating the bound (7.12) associated with scalar quantization. Again, these constructions require certain smoothness conditions on the frame and employ the canonical dual frame for recovery. In a slightly different approach, the design of the feedback and the ordering of the frame vectors has been considered as part of the quantizer design [14, 20].

A new take on the frame quantization problem was initiated in [10, 65] where the authors realized that reconstruction accuracy can be substantially improved by

employing an appropriate alternative dual frame (i.e., a different left-inverse) for recovery. At the core of this approach is still a smoothness argument, but this time for the dual frame. Given a frame, an appropriate dual frame, the so-called Sobolev dual, can be obtained by solving a least-squares problem over the space of all duals [10]. Again, this yields polynomial error decay, albeit now in more general settings. Moreover, by optimizing over such constructions, root-exponential error decay can be achieved [60].

While the definition of the Sobolev dual does not require any smoothness of the frame, the concrete examples discussed in the aforementioned works still exclusively focused on smooth frames. Similar results on recovery guarantees for frames without smoothness properties were first obtained for frames consisting of independent standard Gaussian vectors [42] and subsequently generalized to vectors with independent subgaussian entries [61].

The underlying constructions also form the basis for the $\Sigma\Delta$ quantization schemes for compressed sensing measurements. Details on such schemes are given in Sec. 7.4. The insight behind the schemes is that the number of measurements taken in compressed sensing is typically larger than the support size by at least a logarithmic factor in the dimension, and there is an interest in choosing it even larger than that, as this induces additional stability and robustness. Thus, once the support of the signal has been identified and only the associated signal coefficients need to be determined, one is dealing with a redundant representation. The goal is now to employ frame quantization schemes to exploit this redundancy.

For typical compressed sensing matrices, any k columns indeed form a frame; this follows, for example, from the restricted isometry property. However, as the support of the signal is not known when quantizing the measurements, it is crucial that $\Sigma\Delta$ quantization is universal. That is, it must not require knowledge regarding which of a given collection of frames (namely, those forming the rows of an $m \times k$ submatrix of A) has been used for encoding. The reconstruction from the resulting digital encodings then typically proceeds in two steps. First the support is identified using standard compressed sensing recovery techniques, just treating the quantization error as noise. In a second step, only the restriction of the measurement matrix to the identified support columns is considered. For the frame consisting of the rows of this matrix, one then applies frame quantization reconstruction techniques. Recovery guarantees for such an approach have been proven for Gaussian measurements [42] and measurements with independent subgaussian entries [61]. It is of great importance that the dual frame used for recovery is chosen properly (e.g., the Sobolev dual), as it follows from the RIP that the frames never have a small frame variation. Here again the recovery error bounds decay polynomially in the number of measurements and beat the analogous bounds for scalar quantization.

Preliminary steps towards a unified approach to support and signal recovery have been considered in [29]. The reconstruction techniques studied in this work, however, intrinsically rely on certain non-convex optimization problems, for which no efficient solution methods are known. Thus the quest remains open for an integrated approach to reconstruction from $\Sigma\Delta$ -quantized compressed sensing measurements that combines numerical tractability and guaranteed recovery.

7.3 Scalar Quantization and Compressive Sensing

The interplay of scalar quantization and compressed sensing has been widely explored in the literature. In addition to the lower bounds discussed in Section 7.2.2.3, there is significant interest in providing practical quantization schemes and reconstruction algorithms with strong performance guarantees.

This part explores these results. Our development considers the following quantized compressed sensing (QCS) model:

$$q = Q(y) = Q(Ax), \quad (7.18)$$

where $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$. The sensing matrix can be, for instance, a random Gaussian sensing matrix A such that $a_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$. Note that the scaling of the entries of the sensing matrix should be independent of m . This allows us to fix the design of the scalar quantizer Q since the dynamic range of the components of Ax is then independent of the number of measurements. This has no consequence on some of the common requirements the sensing matrix must satisfy, such as the Restricted Isometry Property (see Chap. 1), as soon as an appropriate rescaling of A is applied. For instance, if A has RIP of order $2k$ and if $A \rightarrow \lambda A$ for some $\lambda > 0$, then A/λ has RIP of the same order and the error bound (1.20) in the stability Theorem 1.6 remains unchanged [52].

The first two parts, Sec. 7.3.1 and Sec. 7.3.2, focus on the *high resolution assumption* (HRA) that simplifies the QCS model. Under HRA, the quantization bin widths— Δ or the distance between two consecutive thresholds—are small with respect to the dynamic range of the unquantized input. This allows us to model the quantization distortion $Q(Ax) - Ax$ as uniform white noise [39]. Determining bounds on its power and moments can better constrain signal reconstruction methods, such as the basis pursuit denoise (BPDN) program [26, 28], which is commonly used for reconstructing signals whose CS measurements are corrupted by Gaussian noise. However, the price to pay is an oversampling in CS measurements.

Sec. 7.3.3 considers scalar quantizers with *saturation*. Saturation induces information loss in the measurements exceeding the saturation level. However, democracy—a key property of compressive sensing measurements that makes every measurement equally informative—provides robustness against such corruption.

In Sec. 7.3.4, very low-resolution quantization is studied through 1-bit compressed sensing. In this case, the HRA cannot be assumed anymore—the quantization bins are the two semi-infinite halves of the real line—and the analysis of the QCS model relies on high dimensional geometric arguments.

Finally, Sec. 7.3.5 studies how noise, either on the signal or on the measurements, can impact the QCS model (7.18), the reconstruction error and the quantizer trade-offs. In particular, at constant bit budget $R = mB$, the total noise power determines the optimal trade-off between quantizer precision and number of measurements.

7.3.1 Uniform Scalar Quantization

First we consider the QCS model (7.18) using a uniform quantizer with resolution Δ and a set of levels \mathcal{Q} ,

$$q = Q(y) = Q(Ax) \in \mathcal{Q}^m,$$

measuring a signal $x \in \mathbb{R}^n$ using a sensing matrix $A \in \mathbb{R}^{m \times n}$. For simplicity, we assume henceforth that x is sparse in the canonical basis, i.e., $\Psi = I$.

We consider a quantizer Q that has uniform quantization regions, i.e., $t_{j+1} - t_j = \Delta$ for all j , and, setting $t_j = j\Delta$, quantization levels $l_j = \frac{t_j + t_{j+1}}{2} = (j + \frac{1}{2})\Delta$ in \mathcal{Q} .

By definition, the signal x satisfies the following *quantization consistency* constraint (QC_u)

$$\|q - Ax\|_\infty \leq \Delta/2. \quad (\text{QC}_u)$$

From this fact, we can also deduce that

$$\|Ax - q\|_2 \leq \sqrt{m} \|Ax - q\|_\infty \leq \sqrt{m}\Delta/2.$$

This shows that the QCS model can be assimilated to a noisy CS model

$$q = Q(Ax) = Ax + \xi, \quad (7.19)$$

with a “noise” $\xi = Q(Ax) - Ax$ of bounded ℓ_2 -norm, i.e., $\|\xi\|_2 \leq \sqrt{m}\Delta/2$.

The quantization noise power can be further reduced using the high resolution assumption. Under this assumption, the coefficients of y may lie anywhere in the quantization region determined by the coefficients of q and it is natural to model the quantization distortion ξ as a uniform white noise, i.e.,

$$\xi_i \sim_{\text{iid}} \mathcal{U}([- \Delta/2, \Delta/2]).$$

Under this model, a simple use of the Chernoff–Hoeffding bound [45] provides, with high probability

$$\|\xi\|_2^2 \leq \varepsilon_2^2 := \frac{\Delta^2}{12}m + \zeta \frac{\Delta^2}{6\sqrt{5}}m^{1/2},$$

for a small constant $\zeta > 0$.

The first approach in modeling and understanding QCS exploited this bound and the development of noise-robust CS approaches to impose a *distortion consistency constraint* (DC_u) [25]

$$\|q - Ax'\|_2 \leq \varepsilon_2, \quad (\text{DC}_u)$$

on any candidate signal x' estimating x . This was indeed a natural constraint to consider since most noise-robust compressed sensing reconstruction methods can incorporate a bounded ℓ_2 -norm distortion on the measurements. For instance, the **BPDN** program can find a solution \hat{x} of

$$\hat{x} = \arg \min_z \|z\|_1 \text{ s.t. } \|q - Az\|_2 \leq \varepsilon_2. \quad (\text{BPDN})$$

Then, if the sensing matrix $A' = A/\sqrt{m}$ satisfies the RIP with constant $\delta \leq 1/\sqrt{2}$ on $2k$ sparse signals, it is known [22] that

$$\|x - \hat{x}\|_2 \lesssim \frac{1}{\sqrt{m}} \varepsilon_2 + \frac{1}{\sqrt{k}} \sigma_k(x)_1 \asymp \Delta + \frac{1}{\sqrt{k}} \sigma_k(x)_1,$$

where $\sigma_k(x)_1$ is the best k -term approximation defined in (1.2).

This approach has two drawbacks. First, there is no guarantee that the solution \hat{x} satisfies the QC_u constraint above, i.e., $\|q - A\hat{x}\|_\infty \not\leq \Delta/2$. This shows that some sensing information has been lost in the reconstruction. Moreover, as described in Sec. 7.2.2.2, the consistency of the solution helps in reaching the lower bound [38, 50, 84]

$$(\mathbb{E}\|x - \hat{x}\|^2)^{1/2} \gtrsim \frac{k}{m} \Delta$$

in the oversampled setting. Second, from a maximum a posteriori standpoint, since every constrained optimisation corresponds to an unconstrained Lagrangian formulation, imposing a small ℓ_2 -norm on the residual $q - A\hat{x}$ can be viewed as enforcing a Gaussian distribution on ξ , which is not the uniform one expected from the HRA.

To circumvent these two limitations, [52] studied the Basis Pursuit DeQuantizer (**BDPQ**) program

$$\hat{x}_p = \arg \min_z \|z\|_1 \text{ s.t. } \|q - Az\|_p \leq \varepsilon_p, \quad (\text{BDPQ}_p)$$

where ε_p must be carefully selected in order for x to be a feasible point of this new ℓ_p -constraint. If $\varepsilon_p \rightarrow \Delta$ as $p \rightarrow \infty$, the **BDPQ** _{p} solution \hat{x}_p tends to be consistent with the quantized measurements. But what is the price to pay, e.g., in terms of number of measurements, for being allowed to increase p beyond 2?

To answer this, we need a variant of the restricted isometry property.

Definition 1. Given two normed spaces $\mathcal{X} = (\mathbb{R}^m, \|\cdot\|_{\mathcal{X}})$ and $\mathcal{Y} = (\mathbb{R}^n, \|\cdot\|_{\mathcal{Y}})$ (with $m < n$), a matrix $A \in \mathbb{R}^{m \times n}$ has the Restricted Isometry Property from \mathcal{X} to \mathcal{Y} at order $k \in \mathbb{N}$, radius $0 \leq \delta < 1$ and for a normalization $\mu > 0$, if for all $x \in \Sigma_k := \{u \in \mathbb{R}^N : \|u\|_0 \leq k\}$,

$$(1 - \delta)^{1/\kappa} \|x\|_{\mathcal{Y}} \leq \frac{1}{\mu} \|Ax\|_{\mathcal{X}} \leq (1 + \delta)^{1/\kappa} \|x\|_{\mathcal{Y}}, \quad (7.20)$$

the exponent κ depending on the spaces \mathcal{X} and \mathcal{Y} . To lighten notation, we write that A is $\text{RIP}_{\mathcal{X}, \mathcal{Y}}(k, \delta, \mu)$.

In this general definition, the common RIP is equivalent to $\text{RIP}_{\ell_2^m, \ell_2^n}(k, \delta, 1)$ with $\kappa = 2$ (see Chap. 1, Eq. (1.8)). Moreover, the $\text{RIP}_{p, k, \delta'}$ defined in [8] is equivalent to the $\text{RIP}_{\ell_p^m, \ell_p^n}(k, \delta, \mu)$ with $\kappa = 1$, $\delta' = 2\delta/(1-\delta)$ and $\mu = 1/(1-\delta)$. Finally, the Restricted p -Isometry Property proposed in [27] is also equivalent to the $\text{RIP}_{\ell_p^m, \ell_2^n}(k, \delta, 1)$ with $\kappa = p$.

To characterize the stability of BPDQ we consider the space $\mathcal{X} = \ell_p^m := (\mathbb{R}^m, \|\cdot\|_p)$ and $\mathcal{Y} = \ell_2^n := (\mathbb{R}^n, \|\cdot\|_2)$ with $\kappa = 1$, and we write RIP_p as a shorthand for $\text{RIP}_{\ell_p^m, \ell_2^n}$. At first sight, it could seem unnatural to define an embedding of $\mathcal{X} = \ell_p^m$ in $\mathcal{Y} = \ell_2^m$ for $p \neq 2$, those spaces being not isometrically isomorphic to each other for $m = n$. However, the RIP_p rather sustains the possibility of an isometry between $\mathcal{X} \cap A\Sigma_k$ and $\mathcal{Y} \cap \Sigma_k$. We will see in Proposition 1 that the existence of such a relation comes with an exponential growth of m as p increases, a phenomenon that can be related to Dvoretzky's theorem when specialized to those Banach spaces [69].

From this new characterization, one can prove the following result.

Theorem 1 ([52, 53]). *Let $k \geq 0$, $2 \leq p < \infty$ and $A \in \mathbb{R}^{m \times n}$ be a $\text{RIP}_p(s, \delta_s, \mu_p)$ matrix for $s \in \{k, 2k, 3k\}$ and some normalization constant $\mu_p > 0$. If*

$$\delta_{2k} + \sqrt{(1 + \delta_k)(\delta_{2k} + \delta_{3k})(p - 1)} < 1/3, \quad (7.21)$$

then, for any signal $x \in \mathbb{R}^n$ observed according to the noisy sensing model $y = Ax + n$ with $\|n\|_p \leq \varepsilon_p$, the unique solution \hat{x}_p obeys

$$\|x^* - x\| \leq 4 \frac{1}{\sqrt{k}} \sigma_k(x)_1 + 8 \varepsilon_p / \mu_p, \quad (7.22)$$

where, again, $\sigma_k(x)_1$ denotes the best k -term approximation.

This theorem follows by generalizing the fundamental result of Candès in [26] to the particular geometry of Banach spaces ℓ_p^m . It shows that, if A is RIP_p with particular requirement on the RIP_p constant, the BPDQ_p program is stable under both measurement noise corruption and departure from the strict sparsity model, as measured by e_0 . In particular, under the same conditions, given a measurement noise ξ and some upper bounds ε_p on its ℓ_p -norm, (7.22) provides the freedom to find the value of p that minimizes ε_p / μ_p .

This is exactly how QCS signal recovery works. Following Theorem 1 and its stability result (7.22), we jointly determine a RIP_p sensing matrix with known value μ_p and a tight error bound ε_p on the ℓ_p norm of the residual $q - Ax$ under HRA. The existence of a RIP_p matrix is guaranteed by the following result [52, 53].

Proposition 1 (RIP_p Matrix Existence). *Let a random Gaussian sensing matrix $A \in \mathbb{R}^{m \times n}$ be such that $a_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$, $p \geq 1$ and $0 \leq \eta < 1$. Then, A is $\text{RIP}_p(k, \delta_k, \mu_p)$ with probability higher than $1 - \eta$ when we have jointly $m \geq 2^{p+1}$ and*

$$m \geq m_0^{\max(p/2, 1)} \quad \text{with } m_0 = O(\delta_k^{-2} (k \log(\frac{n}{k}) + k \log(\delta_k^{-1}) + \log \frac{2}{\eta})). \quad (7.23)$$

Moreover, $\mu_p = \Theta(m^{1/p} \sqrt{p+1})$.

There is thus an exponential price to pay for a matrix A to be RIP_p as p increases: roughly speaking, for $p \geq 2$, we need $m \geq m_0^{p/2} = O(k^{p/2} \log^{p/2}(n/k))$ measurements for satisfying this property with non-zero probability.

To estimate a tight value of ε_p in the case of quantization noise—since, under HRA $\xi_j \sim_{\text{iid}} \mathcal{U}(-\Delta/2, \Delta/2)$ —we can show that

$$\|\xi\|_p \leq \varepsilon_p := \frac{\Delta}{2(p+1)^{1/p}} (m + \zeta(p+1) \sqrt{m})^{\frac{1}{p}}, \quad (7.24)$$

with probability higher than $1 - e^{-2\zeta^2}$. Actually, for $\zeta = 2$, x is a feasible solution of the BPDQ_p fidelity constraint with a probability exceeding $1 - e^{-8} > 1 - 3.4 \times 10^{-4}$.

Finally, combining the estimation ε_p with the bound on μ_p , we find, under the conditions of Proposition 1,

$$\frac{\varepsilon_p}{\mu_p} \lesssim \frac{\Delta}{\sqrt{p+1}}. \quad (7.25)$$

This shows that, in the high *oversampled sensing scenario* driven by (7.23), and provided the RIP_p constants $\{\delta_k, \delta_{2k}, \delta_{3k}\}$ satisfy (7.21), the part of the reconstruction error due to quantization noise behaves as $O(\Delta/\sqrt{p+1})$. This is also the error we get if x is exactly k -sparse since then e_0 vanishes in (7.22).

If we solve for p , we can see that the error decays as $O(\Delta/\sqrt{\log m})$ as m increases. There is possibly some room for improvements since, as explained in Sec. 7.2.2.2, the lower bound on reconstruction of sparse signal is $\Omega(\Delta/m)$. Beyond scalar quantization schemes, Sec. 7.4 will also show that much better theoretical error reduction can be expected using $\Sigma\Delta$ quantization.

Interestingly, we can, however, observe a numerical gain in using BPDQ_p for increasing values of p when the signal x is observed by the model (7.19) and when m increases beyond the minimal value m_0 needed for stabilizing **BPQN** (i.e., BPDQ_2).

This gain is depicted in Fig. 7.4. The plots on the left correspond to the reconstruction quality, i.e., the value $\text{SNR} = 20 \log(\|x\|/\|x - \hat{x}_p\|)$ expressed in dB, reached by BPDQ_p for different values of p and m/k . The original signal x

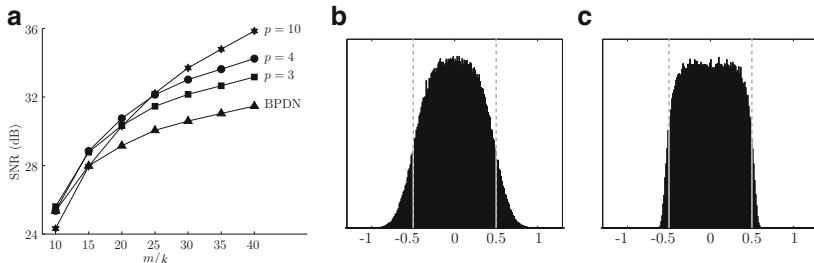


Fig. 7.4 (a) Quality of BPDQ_p for different m/k and p . (b) and (c): Histograms of $\Delta^{-1}(A\hat{x} - q)_i$ for $p = 2$ and for $p = 10$, respectively

has dimension $n = 1024$ and is k -sparse in the canonical basis, with support of size $k = 16$ uniformly random and normally distributed non-zero coefficients. Each point of each curve represents average quality over 500 trials. For each sparse signal x , m quantized measurements were recorded using (7.19) with a random Gaussian sensing matrix A and $\Delta = \|Ax\|_\infty/40$. The reconstruction was done by solving BPDQ_p with the Douglas–Rachford algorithms [52], an efficient convex optimization method solving constrained programs, such as BPDQ ,¹ using simpler *proximal operators* [30]. Fig. 7.4a shows that higher oversampling ratio m/k allows the use of higher p with significant gain in the reconstruction quality. However, if m/k is low, i.e., close to $m/k = 10$, the best quality is still reached by BPDN. The quantization consistency of the reconstruction, i.e., the original motivation for introducing the BPDQ_p program, can also be tested. This is shown in Fig. 7.4b and Fig. 7.4c where the histograms of the components of $\Delta^{-1}(A\hat{x}_p - q)$ are represented for $p = 2$ and $p = 10$ at $m/k = 40$. This histogram for $p = 10$ is indeed closer to a uniform distribution over $[-1/2, 1/2]$, while the one at $p = 2$ is mainly Gaussian.

7.3.2 Non-Uniform Scalar Quantization

If the distribution of the measurements is known, quantization distortion can be decreased by adopting a non-uniform scalar quantizer. For instance, when A is a random Gaussian matrix viewing the signal as fixed and the matrix as randomly drawn, the distribution of the components of $y = Ax$ is also Gaussian with a variance proportional to the signal energy $\|x\|_2^2$ (and similarly, for other matrix constructions, such as ones drawn with random sub-Gaussian entries). Assuming the acquired signal energy can be fixed, e.g., using some automatic gain control, the known distribution of the measurements can be exploited in the design of the quantizer, thanks, for example, to the Lloyd–Max algorithm mentioned in Sec. 7.2.2 [71]. In particular, the quantization thresholds and levels are then optimally adjusted to this distribution.

This section shows that the formalism developed in Sec. 7.3.1 can indeed be adapted to non-uniform scalar quantizer. To understand this adaptation, we exploit a common tool in quantization theory [39]: any non-uniform quantizer can be factored as the composition of a “compression” of the real line over $[0, 1]$ followed by a uniform quantization of the result that is finally re-expanded on \mathbb{R} . Mathematically,

$$Q = \mathcal{G}^{-1} \circ Q_\Delta \circ \mathcal{G}, \quad (7.26)$$

where $\mathcal{G} : \mathbb{R} \rightarrow [0, 1]$ is the *compressor* and $\mathcal{G}^{-1} : [0, 1] \rightarrow \mathbb{R}$ is the *expander*, giving the name *comparator* as a portemanteau.

¹The code of BPDQ is freely available at <http://wiki.epfl.ch/bpdq>.

In particular, under HRA, the compressor \mathcal{G} of a distortion optimal quantizer, i.e., one that minimizes $\mathbb{E}|X - Q(X)|^2$ for a source modeled as a random variable X with pdf φ , must satisfy

$$\frac{d}{d\lambda} \mathcal{G}(\lambda) = \left(\int \varphi^{1/3}(t) dt \right)^{-1} \varphi^{1/3}(\lambda),$$

and if Q is an optimal B -bit quantizer (e.g., obtained by Lloyd-Max method) then $\Delta = 2^{-B}$ in (7.26). In this case, the Panter and Dite formula estimates the quantizer distortion as [79]

$$\mathbb{E}|X - Q(X)|^2 \simeq_B \frac{2^{-2B}}{12} \|\varphi\|_{1/3} =: \sigma_{\text{PD}}^2,$$

with L_s -norm $\|\varphi\|_s = (\int |\varphi^s(t)| dt)^{1/s}$ and where “ \simeq_B ” means that the relation tends to an equality when B is large. The rest of this section assumes that the expected distribution is Gaussian, i.e., if $\varphi \sim \mathcal{N}(0, \sigma_0^2)$ and $\|\varphi\|_{1/3} = \frac{1}{2}\sqrt{3}\pi\sigma_0^2$, as it comes by seeing the signal fixed (with known energy) and the Gaussian matrix random in CS.

Compander theory generalizes quantization consistency in the “compressed” domain, i.e.,

$$|\mathcal{G}(\lambda) - \mathcal{G}(Q(\lambda))| \leq \Delta/2 = 2^{-B-1}.$$

Therefore, for the right compressor \mathcal{G} , in the noiseless QCS model (7.18), the signal x provides consistency constraints to be imposed on any reconstruction candidate x' :

$$\|\mathcal{G}(Ax') - \mathcal{G}(q)\|_\infty \leq \Delta/2 = 2^{-B-1}. \quad (\text{QC})$$

This generalizes the uniform quantization consistency (QC_u) introduced in Sec. 7.3.1.

The compander formalism is leveraged in [53], to generalize the approach described in Sec. 7.3.1 to non-uniform quantization. In particular, a new set of parametric constraints are introduced, the p -Distortion Consistency (or D_pC) for $p \geq 2$. These have for limit cases the QC above and the *distortion consistency* constraint (DC) arising from Panter and Dite formula, namely, the constraint imposing any reconstruction candidate x' to satisfy [31]

$$\|Ax' - q\|_2^2 \leq \varepsilon_{\text{PD}}^2 := m\sigma_{\text{PD}}^2, \quad (\text{DC})$$

with DC asymptotically satisfied by x when both B and m are large.

The D_pC constraint corresponds to imposing that a candidate signal x' satisfies

$$\|Ax' - Q_p[q]\|_{p,w} = \|Ax' - Q_p[Ax]\|_{p,w} \leq \varepsilon_{p,w}, \quad (\text{D}_p\text{C})$$

where $\|v\|_{p,w} = \|\text{diag}(w)v\|_p$ is the weighted ℓ_p -norm of $v \in \mathbb{R}^m$ with weights $w \in \mathbb{R}_+^m$, denoting by $\text{diag}(w)$ the diagonal matrix having w on its diagonal. The mapping $Q_p : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is a post-quantization modification of q characterized componentwise hereafter and such that $Q_p[q] = Q_p[Ax]$.

Under HRA, a careful design of Q_p , w and the bounds $\varepsilon_{p,w}$ ensures that D₂C amounts to imposing DC on x' and, that as $p \rightarrow +\infty$, D_pC tends to QC [53]. Briefly, if q_i falls in the quantization bin \mathcal{C}_j , $Q_p(q_i)$ is defined as the minimizer of

$$\min_{\lambda \in \mathcal{C}_j} \int_{\mathcal{C}_j} |t - \lambda|^p \varphi(t) dt.$$

Actually, $Q_2(q_i) = q_i$ by equivalence with (7.10), and $\lim_{p \rightarrow \infty} Q_p(q_i) = \frac{1}{2}(t_j + t_{j+1})$. The weights are defined by the quantizer compressor \mathcal{G} with $w_i(p) = \frac{d}{d\lambda} \mathcal{G}(Q_p[q_i])^{\frac{p-2}{p}}$. Moreover, under HRA and asymptotically in m , an optimal bound ε_p reads $\varepsilon_{p,w}^p = m^{\frac{2-B_p}{(p+1)2^p}} \|\varphi\|_{1/3}$. For $p = 2$, $\varepsilon_{2,w} = \varepsilon_{\text{PD}}$ matches the distortion power estimated by the Panter and Dite formula, while for $p \rightarrow +\infty$, $\varepsilon_{p,w} \rightarrow \frac{1}{2}2^{-B}$, i.e., half the size of the uniform quantization bins in the domain compressed by \mathcal{G} .

Similarly to Sec. 7.3.1, using (D_pC) as a fidelity constraint in the signal reconstruction leads to the definition of a Generalized Basis Pursuit DeNoise program:

$$\hat{x}_{p,w} = \arg \min_{z \in \mathbb{R}^n} \|z\|_1 \text{ s.t. } \|Q_p(q) - Az\|_{p,w} \leq \varepsilon_{p,w}. \quad (\text{GBPDN}(\ell_{p,w}))$$

Ideally, we would like to directly set $p = \infty$ in order to enforce consistency of $\hat{x}_{p,w}$ with q . However, as studied in [53], it is not certain that this limit case minimizes the reconstruction error $\|x - \hat{x}_{p,w}\|$ as a function of p , given a certain number of measurements m .

Actually, the stability of GBPDN can be established from the one of BPDQ (Sec. 7.3.1) if we impose A to satisfy the more general RIP _{$\ell_{p,w}^m, \ell_2^n$} , as formally defined in (7.20). Indeed, for any weighting vector w , we have always $\|Q_p(q) - Az\|_{p,w} = \|q' - A'z\|_p$ with $q' = \text{diag}(w)Q_p(q)$ and $A' = \text{diag}(w)A$. Therefore, we know from Theorem 1 that if A' is RIP _{p} , or equivalently if A is RIP _{$\ell_{p,w}^m, \ell_2^n$} , with the additional condition (7.21) on its RIP constants at different sparsity levels, then the solution of GBPDN($\ell_{p,w}$) will be stable in the sense of (7.22), i.e.,

$$\|\hat{x}_{p,w} - x\| \lesssim \frac{\varepsilon_{p,w}}{\mu_{p,w}} + \frac{\sigma_k(x)_1}{\sqrt{k}}.$$

Compared to the unit weights case (as involved by the RIP _{p}), a random Gaussian matrix A with $a_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$ satisfies the RIP _{$\ell_{p,w}^m, \ell_2^n$} ($k, \delta_k, \mu_{p,w}$) with high probability provided that m grows like $O((\theta_p \delta_k^{-2} (k \log(n/k))^{p/2}))$. The ratio $\theta_p := \|w\|_\infty / (m^{-1/p} \|w\|_p)$ depends on the *conditioning* of w . It is equal to 1 for constant weights (recovering (7.23)), while it increases with the dynamic range of w . For the weight $w(p)$ defined previously and with a Gaussian optimal quantizer, $\theta_p^{p/2} \simeq_{m,B} \sqrt{p+1}$ asymptotically in m and B .

As for the uniform case, a strong (polynomial) oversampling in m is thus required for satisfying the RIP $\text{RIP}_{p,w}^{\ell_p^m, \ell_2^n}$ at $p > 2$ compared to the minimal number of measurements needed at $p = 2$. However, an asymptotic analysis of $\varepsilon_{p,w}/\mu_{p,w}$ shows that the GBP DN reconstruction error due to quantization for a Gaussian sensing matrix behaves like [53]

$$\|\hat{x}_{p,w} - x\| \lesssim \frac{2^{-B}}{\sqrt{p+1}} + \frac{\sigma_k(x)_1}{\sqrt{k}},$$

This error decay is thus similar to the one found in (7.25) for uniform QCS with now a direct interpretation in terms of the quantizer bit-depth B .

Efficient convex optimization methods, like those relying on proximal algorithms [30], can also be used to numerically solve GBP DN. In [53], numerical simulations show that the reconstruction qualities reached in the reconstruction of sparse signals from their non-uniformly quantized measurements behave similarly, with respect to p and m , to those observed in Sec. 7.3.1 for the uniformly quantized CS setting.

We should also remark that beyond QCS, the stability of GBP DN (when A is $\text{RIP}_{p,w}$) can also be used for reconstructing signals acquired under a (heteroscedastic) noisy sensing model $y = Ax + \xi$ where $\xi \in \mathbb{R}^m$ is an additive generalized Gaussian noise with bounded $\ell_{p,w}$ -norm for some specific weight $w \in \mathbb{R}_+^m$ [53, 92].

7.3.3 Finite-Range Scalar Quantizer Design

So far we have only considered a scalar quantizer model without saturation. Practical scalar quantizers have a finite range, which implies a saturation level $\pm S$ and, using B bits per coefficient, a quantization interval equal to

$$\Delta = S2^{-B+1}. \quad (7.27)$$

In order to determine the optimal saturation rate, the system designed needs to balance the loss of information due to saturation, as S decreases, with the increased quantization error due to an increasing quantization interval in (7.27), as S increases. In classical systems, this balance requires setting the quantization level relatively close to the signal amplitude to avoid saturation. On the other hand, in compressive sensing systems, the incoherence of the measurements with the sparsity basis of the signal makes them more robust to loss of information and enables higher saturation levels with smaller quantization intervals.

A key property of compressive measurements, which provides the robustness to loss of information, is *democracy*. Intuitively, each measurement contributes an equal amount of information to the reconstruction. If the signal is slightly oversampled, relative to the rate required for CS reconstruction, then any subset with enough measurements should be sufficient to recover the signal. The notion of democracy was first introduced in [23, 43] in the context of information carried in each bit of the representation; the definition below strengthens the concept and formulates it in the context of compressive sensing [33, 66].

Definition 2. Let $A \in \mathbb{R}^{m \times n}$, and let $\tilde{m} \leq m$ be given. We say that A is (\tilde{m}, k, δ_k) -democratic if, for all row index sets Γ such that $|\Gamma| \geq \tilde{m}$, any matrix $\tilde{A} = ((A^T)_{\Gamma})^T$, i.e., comprised of a Γ -subset of the rows of A , satisfies the RIP of order k with constant δ_k .

This definition takes an adversarial view of democracy: a matrix A is democratic if an adversary can pick any $d = m - \tilde{m}$ rows to remove from A , and the remaining matrix still satisfies the RIP. This is a much stronger guarantee than just randomly selecting a subset of the rows to be removed. Such a guarantee is important in the case of saturation robustness because the saturated measurements are the largest ones in magnitude, i.e., potentially the ones most aligned with the measured signal and, presumably, the ones that capture a significant amount of information. Still, despite this strict requirement, randomly generated matrices can be democratic if they have a sufficient number of rows.

Theorem 2 ([33]). Let $A \in \mathbb{R}^{m \times n}$ with elements a_{ij} drawn according to $\mathcal{N}(0, \frac{1}{m})$ and let $\tilde{m} \leq m$, $k < \tilde{m}$, and $\delta \in (0, 1)$ be given. Define $d = m - \tilde{m}$. If

$$m = C_1(k + d) \log \left(\frac{n + m}{k + d} \right), \quad (7.28)$$

then with probability exceeding $1 - 3e^{-C_2 m}$ we have that A is $(\tilde{m}, k, \delta/(1 - \delta))$ -democratic, where C_1 is arbitrary and $C_2 = (\delta/8)^2 - \log(42e/\delta)/C_1$.

The practical implication of democratic measurements is that information loss due to saturated measurements can be tolerated.

Saturated measurements are straightforward to detect, since they quantize to the highest or the lowest level of the quantizer. The simplest approach is to treat saturated measurements as corrupted, and reject them from the reconstruction, together with the corresponding rows of A . As long as the number of saturated measurements is not that large, the RIP still holds and reconstruction is possible using any sparse reconstruction algorithm.

However, saturated measurements do contain the information that the measurement is large. In the context of consistent reconstruction, they can be used as constraints in the reconstruction process. If a measurement i is positively saturated, then we know that $(Ax)_i \geq S - \Delta$. Similarly, if it is negatively saturated, $(Ax)_i \leq -S + \Delta$. These constraints can be imposed on any reconstruction algorithm to improve performance [66].

Fig. 7.5 demonstrates the effect of each approach. As demonstrated in the plots, rejecting saturated measurements or treating them as consistency constraints significantly outperforms just ignoring saturation. Furthermore, if saturation is properly taken into account, a distortion optimal finite-range scalar quantizer should be designed with significant saturation rate, often more than 20%. While the figures suggest that saturation rejection and saturation consistency have very similar performance, careful examination demonstrates, as expected, that consistency provides more robustness in a larger range of saturation rates and conditions. A more careful

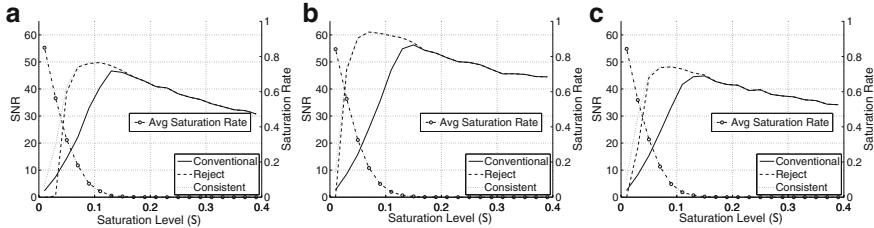


Fig. 7.5 Saturation performance using ℓ_1 minimization for (a) exactly $k = 20$ -sparse signals and compressible signals in weak ℓ_p for (b) $p = 0.4$ and (c) $p = 0.8$, with $n = 1024$, $m = 384$, and $B = 4$. The reconstruction SNR as a function of the saturation level is measured on the left y-axis assuming (solid line) conventional reconstruction, i.e., ignoring saturation, (dotted line) enforcing saturation consistency, and (dashed line) rejecting saturated measurements. The dashed-circled line, measured on the right y-axis, plots the average saturation rate given the saturation level (a) $k = 20$ (b) $x \in w\ell_{0.4}$ (c) $x \in w\ell_{0.8}$

study and detailed discussion can be found in [66]. Furthermore, further gains in the bit-rate can be achieved by coding for the location of the saturated measurements and transmitting those separately [58].

7.3.4 1-Bit Compressive Sensing

The simplest scalar quantizer design to implement in hardware is a 1-bit quantizer, which only computes the sign of its input. Its simplicity makes it quite appealing for compressive sensing systems.

The sensing model of 1-bit CS, first introduced in [19], is very similar to the standard scalar quantization model

$$q = \text{sign}(Ax), \quad (7.29)$$

where $\text{sign}(x_i)$ is a scalar function applied element-wise to its input and equals 1 if $x_i \geq 0$ and -1 otherwise.

One of the challenges of this model is that it is invariant under changes of the signal amplitude since $\text{sign}(cx) = \text{sign}(x)$ for any positive c . For that reason, enforcing consistency is not straightforward. A signal can be scaled arbitrarily and still be consistent with the measurements. Thus, a magnitude constraint is typically necessary. Of course, the signal can only be recovered within a positive scalar factor.

Similarly to multi-bit scalar quantization models, the literature in this area focuses on deriving lower bounds for the achievable performance, reconstruction guarantees, as well as practical algorithms to invert this problem.

7.3.4.1 Theoretical Performance Bounds

A lower bound on the achievable performance can be derived using a similar analysis as in Sec. 7.2.2.3. The main difference is that the quantization cells are now orthants in the m -dimensional space, shown in Fig. 7.6a, corresponding to each measured sign pattern. Each subspace of the $\binom{n}{k}$ possible ones intersects very few of those orthants, as shown in Fig. 7.6b, i.e., uses very few quantization points. In total, at most $I \leq 2^k \binom{n}{k} \binom{m}{k}$ quantization cells are intersected by the union of all subspaces [54].

Since the signal amplitude cannot be recovered, the lower bound is derived on k -dimensional spheres and coverings using spherical caps instead of balls. The derivation ensures that the spherical caps have radius sufficiently large to cover

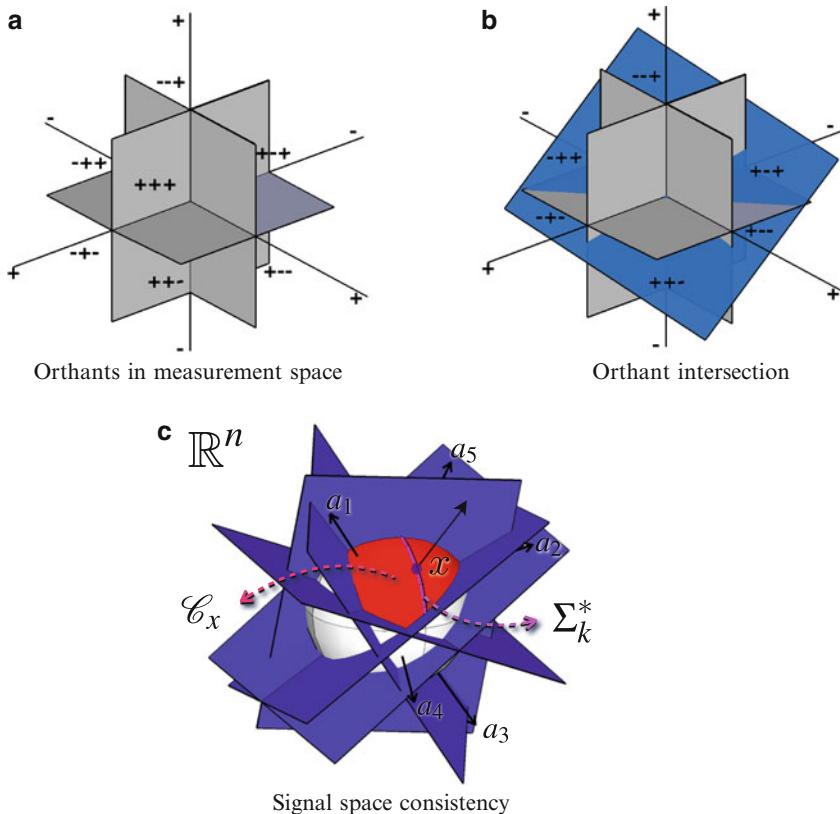


Fig. 7.6 Behavior of 1-bit measurements. (a) The measurement space, \mathbb{R}^m is separated to high-dimensional orthant, according to the sign of each orthant. (b) Signals in a k -dimensional space ($k < m$) will only map to a k dimensional subspace of \mathbb{R}^m and intersect only a few orthants of the measurement space. (c) The same behavior in the signal space. Each measurement vector defines its orthogonal hyperplane. The measurement sign identifies which side of the hyperplane the signal lies on; all signals in the shaded region have consistent measurements. Newer measurements provide less and less information; the chance of intersecting the consistency region decreases

the $\binom{n}{k}$ spheres. Despite the similarity to the argument in Sec. 7.2.2.3, this case requires a little bit more care in the derivation; details can be found in [54]. Still, the result is very similar in nature. Defining $\Sigma_k^* := \{x \in \Sigma_k, \|x\|_2 = 1\}$, we have:

Theorem 3 ([54]). *Given $x \in \Sigma_k^*$, any estimation $\hat{x} \in \Sigma_k^*$ of x obtained from $q = \text{sign}(Ax)$ has a reconstruction error of at least*

$$\|\hat{x} - x\| \gtrsim \frac{k}{m + k^{3/2}},$$

which is on the order of $\frac{k}{m}$ as m increases.

If the sensing matrix A is Gaussian, i.e., if $a_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$, any k -sparse signal that has consistent measurements will not be very far from the signal producing the measurements, assuming a sufficient number of them. This guarantee approaches the lower bound of Theorem 3 within a logarithmic factor.

Theorem 4 ([54]). *Fix $0 \leq \eta \leq 1$ and $\varepsilon_o > 0$. If the number of measurements is*

$$m \geq \frac{2}{\varepsilon_o} \left(2k \log(n) + 4k \log\left(\frac{17}{\varepsilon_o}\right) + \log\frac{1}{\eta} \right), \quad (7.30)$$

then for all $x, x' \in \Sigma_k^*$ we have that

$$\|x - x'\|_2 > \varepsilon_o \Rightarrow \text{sign}(Ax) \neq \text{sign}(Ax'), \quad (7.31)$$

with probability higher than $1 - \eta$. Equivalently, if m and k are given, solving for ε_0 above leads to

$$\|x - x'\|_2 \lesssim \frac{k}{m} \log \frac{mn}{k}, \quad (7.32)$$

with the same probability.

Fig. 7.6c provides further intuition on these bounds by illustrating how 1-bit measurements operate in the signal space. Specifically, each measurement corresponds to a hyperplane in the signal space, orthogonal to the measurement vector. The measurement sign determines on which side of the hyperplane the signal lies. Furthermore, the signal is sparse, i.e., lies in Σ_k^* . A consistent sparse reconstruction algorithm can produce any sparse signal in the indicated shaded region.

A new measurement provides new information about the signal only if the corresponding hyperplane intersects the region of consistent signals and, therefore, shrinks it. However, as more measurements are obtained and the consistency region shrinks, newer measurements have lower probability of intersecting that region and providing new information, leading to the $1/m$ decay of the error.

Consistency can be quantified using the normalized hamming distance between measurements

$$d_H(q, q') = \frac{1}{m} \sum_i q_i \oplus q'_i,$$

where \oplus denotes the exclusive-OR operator. It is, thus, possible to show that if x and x' above differ by no more than s bits in their 1-bit measurements, i.e., if $d_H(\text{sign}(Ax), \text{sign}(Ax')) \leq s/m$, then, with $m \gtrsim \frac{1}{\epsilon_0} k \log \max(m, n)$ and with high probability [51],

$$\|x - x'\|_2 \leq \frac{k+s}{k} \epsilon_0.$$

A bound similar to (7.32) exists for sign measurements of non-sparse signals in the context of quantization using frame permutations [76]. In particular, reconstruction from sign measurements of signals exhibits (almost surely) an asymptotic error decay rate arbitrarily close to $O(1/m)$. However, in contrast to Theorem 4 this result holds only for a fixed signal and not uniformly for all signals of interest.

Note that these results focus on matrices generated using the normal distribution. It has been shown that matrices generated from certain distributions do not perform well in this setting, even though they can be used in standard compressive sensing [82]. For instance, consider a random Bernoulli matrix A such that $a_{ij} = 1$ or -1 with equal probability. In this case, the two distinct sparse vectors $(1, 0, \dots, 0)^T$ and $(1, \lambda, 0, \dots, 0)^T$ with $0 \leq \lambda < 1$ are λ apart and they generate the same quantization vector $q = \text{sign}(A_1)$, where A_1 is the first column of A . It is not possible, therefore, to distinguish those two vectors from their 1-bit observations by increasing m and guarantee that the reconstruction error will decay as measurements increase. This counterexample, however, is exceptional in the sense that such failures can only happen if the signal can have a very large entry. Under mild flatness assumptions on the ℓ_∞ -norm of the signal, arbitrary subgaussian measurements can be utilized [1].

These results establish lower and upper bounds on distances between two sparse signals that have (almost) consistent 1-bit measurements. It is also possible to provide an embedding guarantee similar to the RIP [25]. Since the measurement does not preserve the signal magnitude, we should not expect distances of signals to be preserved. However, the measurements do preserve angles between signals. Defining $d_S(u, v) = \frac{1}{\pi} \arccos(u^T v)$, $u, v \in S^{n-1}$, we have:

Theorem 5 (Binary ϵ -Stable Embedding (B ϵ SE) [54]). *Let $A \in \mathbb{R}^{m \times n}$ be a random Gaussian matrix such that $a_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$. Fix $0 \leq \eta \leq 1$ and $\epsilon > 0$. If the number of measurements satisfies*

$$m \geq \frac{2}{\epsilon^2} \left(k \log(n) + 2k \log\left(\frac{35}{\epsilon}\right) + \log\left(\frac{2}{\eta}\right) \right), \quad (7.33)$$

then with probability exceeding $1 - \eta$

$$d_S(x, x') - \epsilon \leq d_H(\text{sign}(Ax), \text{sign}(Ax')) \leq d_S(x, x') + \epsilon, \quad (7.34)$$

for all $x, x' \in \Sigma_k^$.*

In other words, up to an additive distortion that decays as $\epsilon \lesssim (\frac{k}{m} \log \frac{mn}{k})^{1/2}$, the Hamming distance between $\text{sign}(Ax)$ and $\text{sign}(Ax')$ tends to concentrate around the angular distance between x and x' . Notice that, in contrast to the RIP, a vanishing

distance between the quantized measurements of two signals does not imply they are equal, i.e., we observe a (restricted) *quasi-isometry* between Σ_k^* and $\text{sign}(A\Sigma_k^*)$ instead of the common RIP [49]. This comes from the additive nature of the distortion in (7.34) and is a direct effect of the inherent ambiguity due to quantization.

This embedding result has been extended to signals belonging to convex sets $\mathcal{K} \subset \mathbb{R}^n$ provided that their Gaussian mean width

$$w(\mathcal{K}) = \mathbb{E} \sup\{u^T g : u \in \mathcal{K} - \mathcal{K}\}, \quad g \sim \mathcal{N}(0, \mathbf{I}_{n \times n}), \quad (7.35)$$

with $\mathcal{K} - \mathcal{K} := \{v - v' : v, v' \in \mathcal{K}\}$, can be computed [80–82]. In particular, if

$$m \geq C\epsilon^{-6}w^2(\mathcal{K})$$

for some constant $C > 0$, then (7.34) holds with high probability for any $x, x' \in \mathcal{K} \cap S^{n-1}$. In particular, for

$$\mathcal{K} = K_{n,k} := \{u \in \mathbb{R}^n : \|u\|_1 \leq k^{1/2}, \|u\|_2 \leq 1\},$$

since $w^2(K_{n,k}) = O(k \log n / k)$ [81], an embedding exists between the set of compressible vectors modeled by $K_{n,k}$ and $\{-1, +1\}^m$ provided that $m \geq C\epsilon^{-6}k \log n / k$.

Note that generalizations of these embeddings to non-linear functions other than the sign operator, or to stochastic processes whose expectation is characterizable by such functions, are also possible [81].

7.3.4.2 Reconstruction from 1-Bit Measurements

The original efforts in reconstructing from 1-bit measurements enforced $\|x\|_2 = 1$ as a reconstruction constraint, formulating the non-convex ℓ_1 minimization problem

$$\hat{x} = \arg \min_x \|x\|_1, \text{ s.t. } q = \text{sign}(Ax), \|x\|_2 = 1. \quad (7.36)$$

Even though the problem is not convex, a number of algorithms have been shown experimentally to converge to the solution [19, 67]. More recently, a number of greedy algorithmic alternatives have also been proposed [3, 15, 54].

Most of these algorithms attempt to enforce consistency by introducing a one-sided penalty for sign violations

$$J(Az, q) = \|(q \circ Az)_-\|_q, \quad (7.37)$$

where \circ is the element-wise product between vectors, $(y_i)_- = y_i$ if y_i is negative and 0 otherwise, also applied element-wise, and the ℓ_q norm is typically the ℓ_1 or the ℓ_2 norm. Typically, a descent step is performed using the gradient of (7.37),

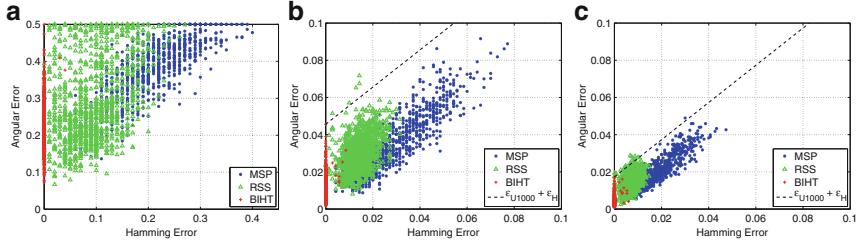


Fig. 7.7 Angular reconstruction error $\varepsilon_S = d_S(x, \hat{x})$ vs. consistency error $\varepsilon_H = d_H(\text{sign}(A\hat{x}), q)$ for different greedy reconstructions (MSP, RSS, and BIHT). BIHT returns a consistent solution in most trials. When A is a BeSE, (7.34) predicts that the angular error ε_S is bounded by the hamming error ε_H (and conversely) in addition to an offset ε . This phenomenon is confirmed by an experimental linear trend (in dashed) between the two errors that improves when m/n increases [54]. **(a)** $m/n = 0.1$ **(b)** $m/n = 0.7$ **(c)** $m/n = 1.5$

followed by a support identification and sparsity enforcement step. Often, care is taken in selecting the descent step, especially considering the signal is on the unit ℓ_2 sphere [67]. Assuming a certain noise level, a maximum likelihood formulation can also be used to remove the norm constraint [3].

For example, the Binary IHT (BIHT), a variation of the popular Iterative Hard Thresholding (IHT) [11], uses the one-sided ℓ_1 norm in (7.37) and follows its subgradient $\frac{1}{2}A^T(q - \text{sign}(Az))$. The algorithm is defined by the iteration

$$z^{n+1} = \mathcal{H}_k\left(z^n + \frac{1}{2}A^T(y - \text{sign}(Az^n))\right), \quad z^0 = 0, \quad (7.38)$$

where $\mathcal{H}_k(\cdot)$ is a hard threshold, keeping the largest k coefficients of its input and setting the remaining ones to zero.

The BIHT does not have convergence or reconstruction guarantees to a consistent output. Still, as shown in Fig. 7.7, it works surprisingly well compared to other greedy approaches. Moreover, variations exist to make it more robust to potential binary errors in the knowledge of q [54] or to extend it to multi-bit scalar quantization [51].

The first iteration of BIHT is a simple truncated back-projection, $\hat{x}_0 = \mathcal{H}_k(A^T q)$ whose distance to x is known to decay asymptotically as $\sqrt{k/m}$ for a Gaussian matrix A [3, 51]. Furthermore, \hat{x}_0 matches the solution of the (feasible) problem

$$\arg \max_z q^T A z \text{ s.t. } z \in \Sigma_k^*,$$

where maximizing $q^T A z$ also promotes the 1-bit consistency of z with q .

This optimization can be generalized to any convex sets $\mathcal{K} \subset \mathbb{R}^n$ where x can lie, such as the set $\mathcal{K} = K_{n,k}$ of compressible signals with Gaussian width $w(K_{n,k}) \asymp k \log n / k$ [81]. If $m \geq C\epsilon^{-2}w(\mathcal{K})^2$ for some $C > 0$, and a fixed x is sensed using (7.29) with a Gaussian sensing matrix A , then the solution to

$$\hat{x} = \arg \max_z q^T A z \text{ s.t. } z \in \mathcal{K},$$

satisfies $\|\hat{x} - x\|^2 = O(\varepsilon)$ with high probability. Interestingly, under certain conditions, this holds also for sensing models other than (7.29), where the sign operator is replaced, for instance, by the logistic function [81].

What makes it difficult to provide reconstruction error estimates for algorithms motivated by the problem (7.36) is the non-convex constraint $\|x\|_2 = 1$, whose convex relaxation allows for the zero solution and is hence meaningless. To overcome this obstacle, it has been proposed in [80, 81] to impose a norm constraint to prevent trivial solutions on the measurements rather than the signal. This results in a different problem, which allows for a meaningful convex relaxation. Namely, since $q = \text{sign}(Ax)$, it follows that at the solution $q^T(Ax) = \|Ax\|_1$. Thus, by constraining this norm, the following convex problem can be formulated:

$$\hat{x} = \arg \min_x \|x\|_1, \text{ s.t. } q = \text{sign}(Ax), q^T Ax = 1 \quad (7.39)$$

As shown in [80], the problem in (7.39) does allow for reconstruction guarantees: If $m \sim \varepsilon^{-5} k \log(n/k)$, the solution \hat{x} recovered from quantized Gaussian measurements of a sparse signal x is such that $d_S(x, \hat{x}) \leq \varepsilon$ with high probability. This holds uniformly for all signals $x \in \mathbb{R}^n$. Under flatness assumptions on the signal, recovery guarantees can also be proved for arbitrary subgaussian measurements [1].

7.3.5 Noise, Quantization, and Trade-offs

The sections above were focused on noiseless QCS models. These models only consider the statistical or the geometrical properties of quantization of CS measurements under high or low resolution modes. However, any signal acquisition system is subject to noise corruption before quantization, either on the measurement process or on the signal itself. Such noise can be incorporated in a more general model

$$q = Q(A(x + \xi_x) + \xi_s), \quad (7.40)$$

where $\xi_x \in \mathbb{R}^n$ and $\xi_s \in \mathbb{R}^m$ corrupt the signal and the sensing, respectively, before quantization. Examining the impact of such noise in signal recovery leads to new interesting questions.

In [93] two efficient reconstruction methods are developed for sparse or compressible signals sensed according (7.40) under sensing noise only, i.e., $\xi_x = 0$. The two approaches are mainly numerical: one relies on a maximum likelihood formulation, built on the quantization model and on a known Gaussian noise distribution, the other follows a least square principle. The two resulting methods are both regularized by an ℓ_1 -norm accounting for sparse signal prior. A provably convergent procedure inherited from a fixed point continuation method is used for reconstructing the signal in the two possible frameworks. With their approach, the combined effects of noise and coarse quantization can be jointly handled. Reasonable reconstruction results are achieved even using 1 or 2 bits per measurement.

The case $\xi_x \neq 0, \xi_s = 0$ boils down to an interaction of the well-understood phenomenon of *noise folding* in CS [34] and quantization [68]. Noise-folding in unquantized CS says that under a weak assumption of orthogonality between the rows of A , the variance of the component $A\xi_x$ undergoes a multiplication by n/m compared to the variance σ_ξ^2 of ξ_x . This impacts directly the reconstruction error of signals. The corresponding MSE is then n/m times higher than the noise power, or equivalently, the SNR loses 3 dB each time m is divided by 2 [34].

An extension of this result to noisy QCS has been provided in [68], assuming the sensing matrix A is RIP of order k and constant δ . In this case, if ξ_x is standard normally distributed and if the quantizer has resolution B , then, under a random signal model where the signal support T is chosen uniformly at random in $\{1, \dots, n\}$ and the amplitudes of the non-zero coefficients are standard normally distributed,

$$(1 - \delta)\mathbb{E}\|x - \hat{x}\|^2 = 2^{-2B+1} \frac{k}{m} \mathbb{E}\|x\|^2 + 2(2^{-2B} + 1) \frac{n}{m} \mathbb{E}\|\xi_x|_T\|^2 + km\kappa, \quad (7.41)$$

where $\hat{x} = (A_T^\dagger q)_T$ is the oracle-assisted reconstruction of x knowing the support T of x for each of its realization, and

$$\kappa = \max_{i \neq j} |\mathbb{E}Q(a_i^T(x + \xi_x))Q(a_j^T(x + \xi_x))|,$$

measures the worst correlation between distinct quantized measurements.

In (7.41), the first term accounts for the quantization error of the signal itself, while the second term represents both the error due to folded signal noise as well as the quantization of that noise.

Finally, the third term reflects a distortion due to correlation between quantized measurement. It is expected to be negligible in CS scenarios, especially when B increases or if a *dithering* is added to Q [39].

Numerical study of (7.41) shows that, at constant rate $R = mB$, a trade-off can be expected between a *measurement compression* (MC) regime, where m is small (but still high enough to guarantee A to be RIP) and B is high, and a *quantization compression* (QC) regime, where m is high compared to the standard CS setting but B is small. Interestingly, the optimal bit-depth B , minimizing the expected reconstruction error, depends on the input SNR: $\text{ISNR} = 20 \log_{10} \|x\|/\|\xi_x\|$. This is illustrated in Fig. 7.8 where the evolution of (7.41) (discarding the effect of the third term) is plotted for four different noise scenarios. The optimal bit depth decays smoothly with the ISNR, suggesting that the QC regime is preferable at low ISNR while MC is clearly better at high ISNR. The general behavior of Fig. 7.8 is also confirmed on Monte Carlo error estimation of the oracle-assisted reconstruction defined above [68].

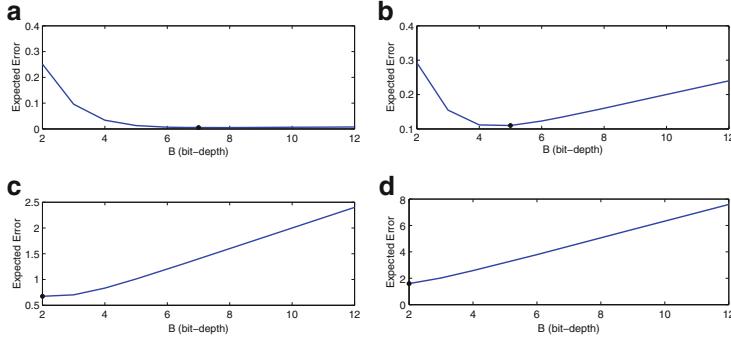


Fig. 7.8 Upper bound on the oracle-assisted reconstruction error as a function of bit-depth B and ISNR at constant rate $R = mB$ [68]. The black dots denote the minimum point on each curve (a) ISNR = 35dB, optimal bit-depth = 7 (b) ISNR = 20dB, optimal bit-depth = 5 (c) ISNR = 10dB, optimal bit-depth = 2 (d) ISNR = 5dB, optimal bit-depth = 2

7.4 Sigma-Delta Quantization for Compressive Sensing

As mentioned in the introduction, $\Sigma\Delta$ quantization for compressed sensing fundamentally builds on corresponding schemes for finite frames. Thus before presenting an analysis specific to compressed sensing, we first discuss the finite frame case.

7.4.1 $\Sigma\Delta$ Quantization for Frames

Let $\Phi \in \mathbb{R}^{n \times N}$ with columns $\{\phi_j\}_{j=1}^N$ be a frame in the sense of (1.32) and consider the frame expansion

$$c = \Phi^T x$$

of a signal $x \in \mathbb{R}^n$. The goal is now to quantize c as a whole such that the quantized representation q allows for approximate recovery of x . $\Sigma\Delta$ quantization schemes obtain such a q using a recursive procedure, which we will now explain in detail.

At the core of the schemes is a uniform scalar quantizer Q , which maps a real number to the closest point in a codebook of the form

$$\mathcal{Q} = \{(\pm j - 1/2)\Delta, j \in \{1, \dots, L\}\}. \quad (7.42)$$

A $\Sigma\Delta$ scheme applies such a quantizer sequentially to the entries of c , taking in each quantization step the errors made in r previous steps into account. The complexity parameter r is referred to as the order of the $\Sigma\Delta$ scheme; it quantifies the trade-off between required storage and achievable accuracy.

A first order $\Sigma\Delta$ quantization scheme, the simplest such algorithm, hence retains the error only for one step. In the following formalization associated with the so-called *greedy* first order $\Sigma\Delta$ scheme, the error parameter appears as the state variable u_i ; it measures the total accumulated error up to step i . The quantized frame coefficient vector $q \in \mathcal{Q}^N$ is computed by running the iteration

$$\begin{aligned} q_i &= Q(u_{i-1} + c_i) \\ u_i &= u_{i-1} + c_i - q_i. \end{aligned} \quad (7.43)$$

As initialization, one typically uses $u_0 = 0$. In matrix-vector notation, the above recurrence relation reads

$$Du = c - q. \quad (7.44)$$

Here $D \in \mathbb{R}^{N \times N}$ is the finite difference matrix with entries given in terms of the Kronecker delta by $D_{ij} = \delta_{i,j} - \delta_{i+1,j}$, that is,

$$D = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & & 0 \\ 0 & -1 & 1 & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}. \quad (7.45)$$

The scheme is explicitly designed such that each q_j partly cancels the error made up to q_{j-1} . When the signal is approximated as $\tilde{\Phi}q$ using a dual frame $\tilde{\Phi} \in \mathbb{R}^{n \times N}$ with columns $\{\tilde{\phi}_j\}_{j=1}^N$, this entails that one seeks to compensate an error in the direction of a dual frame vector $\tilde{\phi}_{j-1}$ using a distortion in the direction of the next dual frame vector $\tilde{\phi}_j$. This serves as a motivation to choose a smoothly varying dual frame, i.e., with subsequent dual frame vectors close to each other.

Bounding the reconstruction error using (7.44) in terms of the operator norm $\|A\|_{2 \rightarrow 2} := \sup_{\|x\|_2 \leq 1} \|Ax\|_2$, one obtains

$$\|x - \tilde{\Phi}q\|_2 = \|\tilde{\Phi}(c - q)\|_2 = \|\tilde{\Phi}Du\|_2 \leq \|\tilde{\Phi}D\|_{2 \rightarrow 2} \|u\|_2.$$

The smoothness intuition is reflected in the fact that the columns of $\tilde{\Phi}D$ are given by $\tilde{\phi}_j - \tilde{\phi}_{j-1}$. Thus more precisely, finding a smooth dual frame $\tilde{\Phi}$ amounts to minimizing $\|\tilde{\Phi}D\|_{2 \rightarrow 2}$.

If one is willing to store more than one previous value of the state variable, that is, to consider a higher order $\Sigma\Delta$ scheme, it is possible to profit from higher order smoothness of the dual frame. Such a generalization of (7.43) is the greedy r -th order $\Sigma\Delta$ scheme, which is associated with the recurrence relation

$$D^r u = c - q. \quad (7.46)$$

Here, the iteration to compute the quantized coefficients is explicitly given by

$$q_i = Q \left(\sum_{j=1}^r (-1)^{j-1} \binom{r}{j} u_{i-j} + c_i \right)$$

$$u_i = \sum_{j=1}^r (-1)^{j-1} \binom{r}{j} u_{i-j} + c_i - q_i. \quad (7.47)$$

As before, one initializes $u_i = 0$, $i \leq 0$. The reconstruction error is now bounded by

$$\|x - \tilde{\Phi}q\|_2 = \|\tilde{\Phi}(c - q)\|_2 = \|\tilde{\Phi}D^r u\|_2 \leq \|\tilde{\Phi}D^r\|_{2 \rightarrow 2} \|u\|_2. \quad (7.48)$$

Examining (7.48), it is advantageous to choose a dual frame that minimizes $\|\tilde{\Phi}D^r\|_{2 \rightarrow 2}$, and a $\Sigma\Delta$ scheme that yields a state-variable sequence with well-bounded $\|u\|_2$. This motivates the following definitions.

Definition 3. Let $\Phi \in \mathbb{R}^{n \times N}$ be a frame and r be a positive integer. Then the r -th order Sobolev dual of Φ is given by

$$\tilde{\Phi}^{(r)} := \arg \min \|\tilde{\Phi}D^r\|_{2 \rightarrow 2} = (D^{-r}\Phi)^\dagger D^{-r}, \quad (7.49)$$

where the minimum is taken over all dual frames of Φ .

Definition 4. A $\Sigma\Delta$ scheme with a codebook \mathcal{Q} is *stable* if there exist constants C_1 and C_2 such that whenever $\|c\|_\infty \leq C_1$ we have $\|u\|_\infty \leq C_2$.

In general, designing and proving the stability of $\Sigma\Delta$ quantization schemes of arbitrary order can be quite difficult if the number of elements in the associated codebook is held fixed. This challenge is especially difficult in the case of 1-bit quantizers and overcoming it is the core of the contributions of [32, 35, 41], where stable $\Sigma\Delta$ quantization schemes of arbitrary order are designed. On the other hand, if the number of elements in the codebook (7.42) is allowed to increase with order, then even the simple greedy $\Sigma\Delta$ schemes (7.47) are stable, as the following proposition shows (see, e.g., [13]).

Proposition 2. *The greedy r -th order $\Sigma\Delta$ scheme (7.47) associated with the $2L$ -level scalar quantizer (7.42) is stable, with $\|u\|_\infty \leq \Delta/2$, whenever $\|c\|_\infty \leq \Delta(L - 2^{r-1} + 2^{-1})$.*

Proof. The proof is by induction. We begin by rewriting (7.47) in terms of auxiliary state variables $u_i^{(j)}$, $j = 1, \dots, r$ and $u_i^{(0)} = c_i - q_i$ as

$$q_i = Q \left(\sum_{j=1}^r u_{i-1}^{(j)} + c_i \right)$$

$$u_i^{(j)} = u_{i-1}^{(j)} + u_i^{(j-1)}, \quad j = 1, \dots, r \quad (7.50)$$

with $u_0^{(j)} = 0$ for $j = 1, \dots, r$. Note that with this notation $u_i^{(r)} = u_i$. Now suppose that $|u_{i-1}^{(j)}| \leq 2^{r-j}\Delta/2$ for all $j \in \{1, \dots, r\}$, then $|\sum_{j=1}^r u_{i-1}^{(j)}| \leq (2^r - 1)\Delta/2$. Since

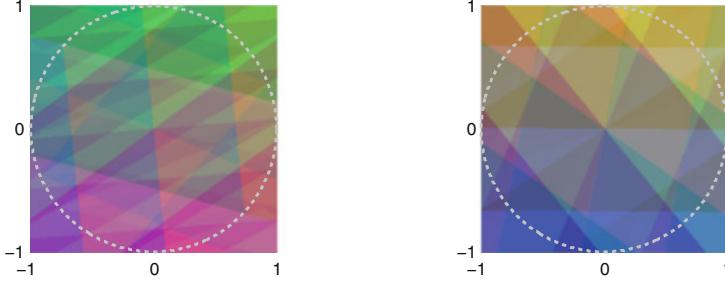


Fig. 7.9 The first order $\Sigma\Delta$ (left) and scalar quantization (right) cells associated with 2-bit quantization of $\Phi^T x$ where x is in the unit ball of \mathbb{R}^2 and Φ is a 2×15 Gaussian random matrix

by the $\Sigma\Delta$ iterations we have $u_i^{(j)} = \sum_{k=1}^j u_{i-1}^{(k)} + c_i - q_i$ we deduce that

$$|u_i^{(r)}| = \left| \sum_{k=1}^r u_{i-1}^{(k)} + c_i - Q \left(\sum_{k=1}^r u_{i-1}^{(k)} + c_i \right) \right| \leq \Delta/2$$

provided $\|c\|_\infty \leq \Delta(L - 2^{r-1} + 1/2)$. Moreover, by (7.50), $|u_i^{(j)}| \leq 2^{r-j}\Delta/2$.

Working with stable r -th order $\Sigma\Delta$ schemes and frames with smoothness properties, and employing the Sobolev dual for reconstruction, it was shown in [10] that the reconstruction error satisfies $\|x - \tilde{\Phi}q\|_2 \leq C_{r,\Phi}N^{-r}$, where the constant $C_{r,\Phi}$ depends only on the quantization scheme and the frame. Such results for $\Sigma\Delta$ -quantization show that its error decay rate breaks the theoretical $\sim 1/N$ lower bound of scalar quantization described in the introduction. Fig. 7.9 helps to illustrate why such a result is possible. It shows the quantization cells associated with two-bit quantization of $\Phi^T x$, where x in the unit ball of \mathbb{R}^2 , using both first order $\Sigma\Delta$ quantization and scalar quantization. For vectors belonging to a given cell, the worst case error achieved by an optimal decoder is proportional to the diameter of the cell. The figure shows that the cells resulting from $\Sigma\Delta$ quantization are smaller than those resulting from scalar quantization, indicating the potential for a smaller reconstruction error. For a detailed overview of $\Sigma\Delta$ quantization of frame expansions, see, e.g., [83].

The existing recovery algorithms for $\Sigma\Delta$ quantized compressed sensing measurements rely on a two-stage algorithm. In the first stage, the signal support is recovered and in the second stage, the signal coefficients are estimated using the Sobolev dual of the frame associated with the recovered support.

7.4.2 Finding the Signal Support

Let $q \in \mathcal{Q}^m$ be the r -th order $\Sigma\Delta$ quantization of the compressed sensing measurements $y = Ax \in \mathbb{R}^m$ associated with the sparse vector $x \in \Sigma_k$ and the measurement matrix $A \in \mathbb{R}^{m \times n}$. In order to preserve the codebook definition (7.42), we assume as in Sec. 7.3 that the scaling of the entries of A is independent of m .

The goal of the first stage of the reconstruction algorithm is to recover $T := \text{supp}(x)$. To that end, following [42] we will use a (standard) compressed sensing decoder $\mathcal{D} : \mathbb{R}^m \rightarrow \mathbb{R}^N$ that has uniform robustness guarantees for matrices with an appropriate RIP constant. For such a decoder and an arbitrary scalar κ

$$x \in \Sigma_k \text{ and } \gamma \in \mathbb{R}^m : \|\gamma\|_2 \leq \kappa\sqrt{m} \implies \|\mathcal{D}(Ax + \gamma) - x\|_2 \leq C\kappa. \quad (7.51)$$

For example, if $\mathcal{D}(Ax + \gamma)$ is the output of an ℓ_1 -minimization algorithm such as Basis Pursuit DeNoise (BPDN), it satisfies (7.51) with constant $C := C(\delta, k)$ when the matrix A (more precisely A/\sqrt{m}) satisfies an appropriate restricted isometry property [26]. As the next proposition shows, robust decoders allow recovering the support of a sparse vector when its smallest non-zero entry is above the error level.

Proposition 3. *Let \mathcal{D} be a compressed sensing decoder satisfying (7.51) and let $x \in \Sigma_k$ with $T := \text{supp}(x)$. Define $\hat{x} := \mathcal{D}(Ax + \gamma)$. If $\min_{i \in T} |x_i| > 2C\kappa$ then the largest k coefficients of \hat{x} are supported on T .*

Proof. First, note that for all $i \in T$, (7.51) yields $|\hat{x}_i - x_i| \leq C\kappa$. Since $\min_{i \in T} |x_i| > 2C\kappa$, the reverse triangle inequality gives $|\hat{x}_i| > C\kappa$ for all i in T . On the other hand, (7.51) also ensures that $|\hat{x}_i| \leq C\kappa$ for all $i \in T^c$.

A sharper version of this argument appears in [42] but Proposition 3 is sufficient for our purposes. In particular, consider an r th order greedy $\Sigma\Delta$ quantization associated with a codebook \mathcal{Q} having $2L$ elements. Applying such a scheme to Ax yields a quantized vector q satisfying $\|q - Ax\|_2 \leq \frac{\Delta}{2} 2^r \sqrt{m}$ provided

$$L > \|Ax\|_\infty / \Delta + 2^{r-1} - 1/2. \quad (7.52)$$

Thus assuming that A/\sqrt{m} has appropriate RIP constants, Proposition 3 shows that using a decoder satisfying (7.51), the support T of $x \in \Sigma_k \subset \mathbb{R}^n$ can be accurately recovered provided $|z_i| > 2^r C \Delta$ for all $i \in T$. What remains is to choose the number of levels L in the codebook to satisfy (7.52); this in turn requires an estimate of $\|Ax\|_\infty$.

To that end, we now consider subgaussian measurement matrices, i.e., matrices whose entries are subgaussian random variables as defined below.

Definition 5. Let ξ be a Gaussian random variable drawn according to $\mathcal{N}(0, \sigma^2)$. If a random variable η satisfies $P(|\eta| > t) \leq eP(|\xi| > t)$ for all t , then we say η is subgaussian with parameter $\sigma > 0$.

Examples of subgaussian random variables include Gaussian, Bernoulli, and bounded random variables, as well as their linear combinations. For matrices populated with such subgaussian entries, the following proposition from [61] gives a bound on $\|Ax\|_\infty$ when the non-zero entries of x are restricted to a fixed support T so that $Ax = \Phi^T x_T$ for a frame Φ associated with the support.

Proposition 4. *Let $\hat{\Phi}$ be a $k \times m$ subgaussian matrix with mean zero, unit variance, and parameter σ , where $k < m$. Let $\Phi = \frac{1}{\sqrt{m}}\hat{\Phi}$ and fix $\alpha \in (0, 1)$. Then, with probability at least $1 - e^{-\frac{1}{4}m^{1-\alpha}k^\alpha}$, we have for all $m > C\frac{1}{1-\alpha}k$ and $x \in \mathbb{R}^k$*

$$\|\Phi^T x\|_\infty \leq e^{1/2} \left(\frac{m}{k}\right)^{-\frac{\alpha}{2}} \|x\|_2. \quad (7.53)$$

Here C is a constant that may depend on σ , but is independent of k and α .

Taking a union bound over all the $\binom{n}{k}$ submatrices of A of size $m \times k$ yields an identical uniform bound on $\|Ax\|_\infty$, which holds for sparse vectors x with high probability, provided $m > Ck(\log n)^{\frac{1}{1-\alpha}}$.

Thus an r -th order greedy $\Sigma\Delta$ scheme with sufficiently many quantization levels allows the recovery of a sparse signal's support from its compressed sensing measurements. Equipped with this knowledge, we can estimate the signal coefficients using the Sobolev dual of the frame associated with the recovered support.

7.4.3 Recovering the Signal Coefficients

We continue to consider Gaussian or subgaussian measurement matrices, now assuming that the support T of the signal x has been identified. Our goal is to approximate the coefficients x_i , $i \in T$. With high probability, the matrix A/\sqrt{m} has the restricted isometry property of order $2k$ and level $\delta_{2k} \leq 1/\sqrt{2}$ provided one takes at least on the order of $k \log(n/k)$ measurements. Then the matrix A_T/\sqrt{m} restricted to the columns indexed by T is close to an isometry and its rows hence form a frame. Consequently, the measurement vector is the associated frame expansion of x_T , and q is the corresponding $\Sigma\Delta$ frame quantization.

As shown in Sec. 7.4.1, it is advantageous to reconstruct x from the r -th order $\Sigma\Delta$ quantization q of the measurement vector Ax using the Sobolev dual $\tilde{A}_T^{(r)}$ of A_T , see (7.48) and (7.49). A possible bound for the reconstruction error is then proportional to $\|\tilde{A}_T^{(r)} D^r\|_{2 \rightarrow 2}$. Thus to show a uniform recovery guarantee, one needs a bound for this quantity which is uniform over all potential support sets T . In the initial work [42], dealing with Gaussian compressed sensing matrices, the approach to proving such a bound consisted of explicitly controlling the lowest singular value of $D^{-r} A_T$. Their approach utilized the unitary invariance of the Gaussian measure to identify the distribution of the singular values of the random matrix $D^{-r} A_T$ with those of $S_{D^{-r}} \Psi$, where $S_{D^{-r}}$ is a diagonal matrix whose entries are the singular

values of D^{-r} , and Ψ is a Gaussian matrix. This, coupled with bounds on the singular values of D^{-r} , allowed [42] to derive bounds that held with probability high enough to survive a union bound over all $\binom{n}{k}$ Gaussian submatrices of A . In [61], this approach was extended to subgaussian matrices. Herein, to prove such a bound on $\|\tilde{A}_T^{(r)} D^r\|_{2 \rightarrow 2}$, we follow the simpler, RIP-based approach presented in [37].

To that end, let $E = U_E S_E V_E^T$ be the singular value decomposition (SVD) of any matrix E (for some orthogonal matrices U_E and V_E) where the matrix S_E is diagonal with (ordered) diagonal entries $\sigma_j(E)$. We denote also $\sigma_{\min}(E) := \sigma_1(E)$ the smallest singular value of E . Then the following proposition (see, e.g., [42]) holds.

Proposition 5. *There are positive constants $C_1(r)$ and $C_2(r)$, independent of m , such that*

$$C_1(r) \left(\frac{m}{j}\right)^r \leq \sigma_j(D^{-r}) \leq C_2(r) \left(\frac{m}{j}\right)^r, \quad j = 1, \dots, m. \quad (7.54)$$

Denote by P_ℓ the $\ell \times m$ matrix that maps a vector to its first ℓ components. Moreover, denote by $\tilde{\Sigma}_k(A, \mathcal{D}) \subset \Sigma_k$ the set of k -sparse signals x whose support can be recovered from q with the decoder \mathcal{D} as in Proposition 3. The following theorem describes the reconstruction performance.

Theorem 6 ([37]). *Let $A \in \mathbb{R}^{m \times n}$ be a matrix such that for a fixed $\ell \leq m$, the $\ell \times n$ matrix $\frac{1}{\sqrt{\ell}} P_\ell V_{D^{-r}}^T A$ has restricted isometry constant $\delta_k \leq \delta$. Then the following holds uniformly for all $x \in \tilde{\Sigma}_k(A, \mathcal{D})$.*

If x has support T , q is the r -th order $\Sigma\Delta$ quantization of Ax , and $\hat{x} := \tilde{A}_T^{(r)} q$, then

$$\|x - \hat{x}\|_2 \leq \frac{\Delta}{C(r)\sqrt{1-\delta}} \left(\frac{m}{\ell}\right)^{-r+\frac{1}{2}},$$

where $C(r) > 0$ is a constant depending only on r and Δ is the quantization step size.

Proof. As the SVD of D^{-r} provides $D^{-r} = U_{D^{-r}} S_{D^{-r}} V_{D^{-r}}^T$, the smallest singular value of $D^{-r} A_T$ satisfies

$$\begin{aligned} \sigma_{\min}(D^{-r} A_T) &= \sigma_{\min}(S_{D^{-r}} V_{D^{-r}}^T A_T) \\ &\geq \sigma_{\min}(P_\ell S_{D^{-r}} V_{D^{-r}}^T A_T) \\ &= \sigma_{\min}((P_\ell S_{D^{-r}} P_\ell^T)(P_\ell V_{D^{-r}}^T A_T)) \\ &\geq \sigma_\ell(D^{-r}) \sigma_{\min}(P_\ell V_{D^{-r}}^T A_T), \end{aligned}$$

To bound $\sigma_{\min}(P_\ell V_{D^{-r}}^T A_T)$ uniformly over all support sets T of size k we simply note that if $\frac{1}{\sqrt{\ell}} P_\ell V_{D^{-r}}^T \Phi$ has restricted isometry constant $\delta_k \leq \delta$ then $\sigma_{\min}(P_\ell V_{D^{-r}}^T A_T)$ is uniformly bounded from below by

$$\sqrt{\ell} \sqrt{1 - \delta}. \quad (7.55)$$

The theorem follows by applying (7.48), (7.54), (7.55) as

$$\frac{1}{\sigma_{\min}(D^{-r}A_T)} \|u\|_2 \leq \frac{\Delta}{C(r)\sqrt{(1-\delta)}} \left(\frac{m}{\ell}\right)^{-r+\frac{1}{2}} \quad (7.56)$$

□

The above theorem can be applied almost directly to Gaussian compressed sensing matrices. If A is a Gaussian matrix with independent zero mean and unit variance entries, then by rotation invariance so is the matrix $P_\ell V_{D-r}^T A$. Regarding the choice of ℓ , note from Theorem 6 that the smaller ℓ is, the better the bound. On the other hand, ℓ has to be large enough for $\frac{1}{\sqrt{\ell}}(P_\ell V_{D-r}^T \Phi)$ to have restricted isometry constant $\delta_k \leq \delta$. This prompts the choice $\ell \asymp k \log n$, as then $\frac{1}{\sqrt{\ell}}(P_\ell V_{D-r}^T \Phi)$ has the restricted isometry constant $\delta_k < \delta$ with high probability, as discussed in Chapter 1. In particular, if

$$m \gtrsim k(\log n)^{\frac{1}{1-\alpha}}, \quad \alpha \in (0, 1)$$

and

$$\ell \asymp k \log n$$

then

$$\frac{m}{\ell} \asymp \frac{m}{k \log n} = \left(\frac{m}{k}\right)^\alpha \cdot \left(\frac{m}{k(\log n)^{\frac{1}{1-\alpha}}}\right)^{1-\alpha} \gtrsim \left(\frac{m}{k}\right)^\alpha$$

Applying Theorem 6 directly, we obtain

$$\|x - \hat{x}\|_2 \lesssim \Delta \left(\frac{m}{k}\right)^{-\alpha(r-\frac{1}{2})}.$$

This essentially recovers the result in [42] and a similar, albeit more technical argument for subgaussian matrices, using either bounds on tail probabilities for quadratic forms [44, 87] or bounds for suprema of chaos processes [59] recovers the analogous result in [61].

To illustrate the advantage of using $\Sigma\Delta$ schemes for quantizing compressed sensing measurements we conduct a numerical experiment with k -sparse signals in \mathbb{R}^n , as we vary the number of measurements m . We fix $k = 10$, $n = 1,000$, and the quantization step-size $\Delta = 0.01$. We draw $m \times n$ Gaussian matrices A for $m \in \{100, 200, 400, 800\}$ and quantize the measurements Ax using scalar quantization and r th order $\Sigma\Delta$ schemes with $r = 1, 2, 3$. We then use the two-stage reconstruction method described herein to obtain an approximation \hat{x} of x using its quantized measurements. Repeating this experiment 30 times, we compute the average of the reconstruction error $\|x - \hat{x}\|_2$ for each of the quantization methods and plot them against the oversampling ratio m/k in Fig. 7.10.

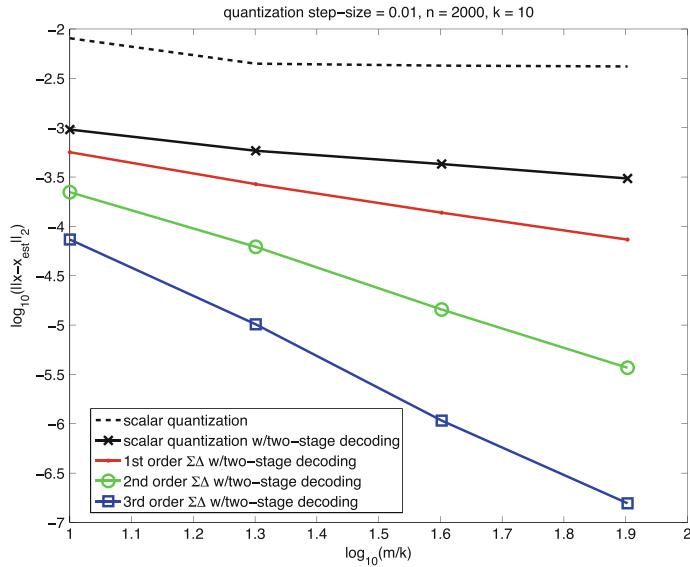


Fig. 7.10 Average errors over 30 experiments. The figure shows the reconstruction errors resulting from scalar quantization, using ℓ_1 -minimization for reconstruction (dashed line). Also corresponding to scalar quantization, the figure shows the errors resulting from reconstructing via the two-stage algorithm described herein (solid black line), using the canonical dual of the frame corresponding to the recovered support in the second stage. It also shows the reconstruction errors resulting from 1st, 2nd, and 3rd order $\Sigma\Delta$ quantization, respectively. These errors decay as $(m/k)^{-r}$ for $r = 1, 2, 3$ respectively, slightly outperforming the theoretical predictions presented here

In summary, using Gaussian and subgaussian compressed sensing matrices recovery of sparse signals from their $\Sigma\Delta$ quantized measurements is possible. More importantly, the reconstruction error decays polynomially in the number of measurements and thus outperforms the (at best) linear error decay that can be achieved with scalar quantization. This improvement comes at the cost of introducing memory elements, and feedback, into the quantization procedure.

7.5 Discussion and Conclusion

Quantization is an essential component of any acquisition system, and, therefore, an important part of compressive sensing theory and practice. While significant work has been done in understanding the interaction of quantization and compressive sensing, there are several open problems and questions.

One of the most interesting open problems is the interaction of quantization with noise. While the discussion and references in Sec. 7.3.5 provide some initial results and theoretical analysis, a comprehensive understanding is still missing.

An understanding of the optimal bit allocation and the optimal quantizer design, uniform or non-uniform scalar, or $\Sigma\Delta$, given the noise level, as well as the robustness of the reconstruction to noise and quantization is still elusive.

While $\Sigma\Delta$ can be used to improve the rate efficiency of compressive sensing, compared to scalar quantization, the performance is still not comparable to the state of the art in conventional $\Sigma\Delta$ methods. For example, conventional $\Sigma\Delta$ quantization of band-limited functions can achieve error that decays exponentially as the sampling rate increases, not currently possible with existing compressive sensing $\Sigma\Delta$. Furthermore, the analysis in Sec. 7.4 does not hold for 1-bit quantization, often desirable in practical systems due to its simplicity. Such an extension has significant practical importance.

Even with $\Sigma\Delta$ approaches, the rate efficiency of compressive sensing systems is not ideal. As evident from the fundamental bounds in Sec. 7.2, compressive sensing is not rate-efficient compared to classical methods such as transform coding. In others word while compressive sensing is very promising in building sensing systems because it can significantly reduce the number of measurements and the sampling burden, it is not a good data compression approach if the measurements have already been obtained and the achievable bit-rate is important. That said, due to the intimate connection between frame quantization and quantization for compressed sensing, promising results in the finite frames context, e.g., [48] can inform future developments in compressed sensing.

The potential encoding simplicity of a compressive sensing system is very appealing. Acquiring generalized linear measurements and quantizing them can be less complex than typical transform-coding approaches and much more attractive in low-power and computationally restricted sensing applications. The complexity is shifted to the reconstruction, which, in many applications, can bear significantly more computational complexity. Nevertheless, the rate inefficiency of compressive sensing can be a barrier in such applications.

A number of promising approaches have been proposed to overcome this barrier using modifications of the quantizer that produce non-contiguous quantization regions [16, 17, 57, 78]. Initial theoretical analysis and experimental results are promising. However, our understanding is still limited. One of the drawbacks of such approaches is that the reconstruction is no longer convex and, therefore, not as simple to provide guarantees for.

Alternatively, recent work on adaptive quantization strategies has shown that error decay exponential in the bit-rate can be achieved, even using a 1-bit quantizer, at the cost of adaptivity in the measurements and—in contrast with the methods presented in this chapter—significant computation at the encoder. Specifically, [5] shows that adaptively choosing the threshold of a 1-bit quantizer allows the error to decay exponentially with the number of measurements. The cost is that the thresholds are updated by solving an ℓ_1 minimization problem, or running an iterative hard thresholding scheme. It is thus interesting to quantify the trade-off between computational complexity at the quantizer, and achievable reconstruction accuracy.

Another important aspect is that while the best recovery guarantees in compressed sensing are obtained for Gaussian and subgaussian measurement matrices,

which are also mainly considered in this article, applications usually require structured matrices, such as subsampled Fourier matrices, e.g., as a model for subsampled MRI measurements [72], or subsampled convolution, e.g., as a model for coded aperture imaging [73]. In both cases, when the subsampling is randomized, near-optimal recovery guarantees are known for unquantized compressed sensing [59, 86]. Combined with quantization, however, hardly anything is known for such matrices. Such results would be of great importance to move the approaches discussed in this survey closer to the application scenarios.

Quantization is also important when considering randomized embeddings, an area of research intimately related to compressive sensing [4, 62]. Embeddings are transformations that preserve the geometry of the space they operate on; reconstruction of the embedded signal is not necessarily the goal. They have been proven quite useful, for example, in signal-based retrieval applications, such as augmented reality, biometric authentication, and visual search [21, 70, 85].

These applications require storage or transmission of the embedded signals, and, therefore, quantizer design is very important in controlling the rate used by the embedding. Indeed, significant analysis has been performed for embeddings followed by conventional scalar quantization, some of it in the context of quantized compressive sensing [54, 81, 82] or in the study of quantized extensions to the Johnson Lindenstrauss Lemma [49, 55, 70, 85]. Furthermore, since reconstruction is not an objective anymore, non-contiguous quantization is more suitable, leading to very interesting quantized embedding designs and significant rate reduction [21]. In this context, quantization can also provide significant computation savings in the retrieval, leading to Locality Sensitive Hashing (LSH) and similar methods [2].

Acknowledgements Petros T. Boufounos is exclusively supported by Mitsubishi Electric Research Laboratories. Laurent Jacques is a Research Associate funded by the Belgian F.R.S.-FNRS. Felix Krahmer and Rayan Saab acknowledge support by the German Science Foundation (DFG) in the context of the Emmy-Noether Junior Research Group KR 4512/1-1 “RaSenQuaSI.”

References

1. Ai, A., Lapanowski, A., Plan, Y., Vershynin, R.: One-bit compressed sensing with non-gaussian measurements. *Linear Algebra Appl.* **441**, 222–239 (2014)
2. Andoni A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM* **51**(1), 117–122 (2008)
3. Bahmani, S., Boufounos, P.T., Raj, B.: Robust 1-bit compressive sensing via Gradient Support Pursuit. *arXiv preprint arXiv:1304.6627* (2013)
4. Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices. *Constr. Approx.* **28**(3), 253–263 (2008)
5. Baraniuk, R., Foucart, S., Needell, D., Plan, Y., Wootters, M.: Exponential decay of reconstruction error from binary measurements of sparse signals. *arXiv preprint arXiv:1407.8246* (2014)
6. Benedetto, J.J., Powell, A.M., Yılmaz, Ö.: Second-order Sigma-Delta ($\Sigma \Delta$) quantization of finite frame expansions. *Appl. Comput. Harmon. Anal.* **20**(1), 126–148 (2006)

7. Benedetto, J.J., Powell, A.M., Yilmaz, Ö.: Sigma-Delta ($\Sigma \Delta$) quantization and finite frames. *IEEE Trans. Inform. Theory* **52**(5), 1990–2005 (2006)
8. Berinde, R., Gilbert, A.C., Indyk, P., Karloff, H., Strauss, M.J.: Combining geometry and combinatorics: a unified approach to sparse signal recovery. In: Proceedings 46th Annual Allerton Conference Communication Control Computing, pp. 798–805. IEEE, New York (2008)
9. Blu, T., Dragotti, P.-L., Vetterli, M., Marziliano, P., Coulot, L.: Sparse sampling of signal innovations. *IEEE Signal Process. Mag.* **25**(2), 31–40 (2008)
10. Blum, J., Lammers, M., Powell, A.M., Yilmaz, Ö.: Sobolev duals in frame theory and Sigma-Delta quantization. *J. Fourier Anal. Appl.* **16**(3), 365–381 (2010)
11. Blumensath, T., Davies, M.: Iterative hard thresholding for compressive sensing. *Appl. Comput. Harmon. Anal.* **27**(3), 265–274, (2009)
12. Bodmann, B.G., Paulsen, V.I.: Frame paths and error bounds for Sigma-Delta quantization. *Appl. Comput. Harmon. Anal.* **22**(2), 176–197 (2007)
13. Bodmann, B.G., Paulsen, V.I., Abdulbaki, S.A.: Smooth frame-path termination for higher order Sigma-Delta quantization. *J. Fourier Anal. Appl.* **13**(3), 285–307 (2007)
14. Boufounos, P.T.: Quantization and Erasures in Frame Representations. D.Sc. Thesis, MIT EECS, Cambridge, MA (2006)
15. Boufounos, P.T.: Greedy sparse signal reconstruction from sign measurements. In: Proceeding of Asilomar Conference on Signals Systems and Computing. Asilomar, California (2009)
16. Boufounos, P.T.: Hierarchical distributed scalar quantization. In: Proceedings of International Conference Sampling Theory and Applications (SampTA), pp. 2–6. Singapore (2011)
17. Boufounos, P.T.: Universal rate-efficient scalar quantization. *IEEE Trans. Inform. Theory* **58**(3), 1861–1872 (2012)
18. Boufounos, P.T., Baraniuk, R.G.: Quantization of sparse representations. In: Rice University ECE Department Technical Report 0701. Summary appears in Proceeding Data Compression Conference (DCC), pp. 27–29. Snowbird, UT (2007)
19. Boufounos P.T., Baraniuk, R.G.: 1-bit compressive sensing. In: Proceedings of Conference Informatin Science and Systems (CISS), pp. 19–21. IEEE Princeton, NJ (2008)
20. Boufounos, P.T., Oppenheim, A.V.: Quantization noise shaping on arbitrary frame expansions. *IEEE EURASIP J. Adv. Signal Process.* Article ID:053807 (2006)
21. Boufounos, P.T., Rane, S.: Efficient coding of signal distances using universal quantized embeddings. In: Proceedings Data Compression Conference (DCC), pp. 20–22. IEEE, Snowbird, UT (2013)
22. Cai, T.T., Zhang, A.: Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE Trans. Inform. Theory* **60**(1), 122–132 (2014)
23. Calderbank, A., Daubechies, I.: The pros and cons of democracy. *IEEE Trans. Inform. Theory* **48**(6), 1721–1725 (2002)
24. Candès, E., Romberg, J.: Encoding the ℓ_p ball from limited measurements. In: Proceeding Data Compression Conference (DCC), pp. 28–30. IEEE, Snowbird, UT (2006)
25. Candès, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59**(8), 1207–1223 (2006)
26. Candès, E.J.: The restricted isometry property and its implications for compressed sensing. *C. R. Acad. Sci., Ser. I* **346**, 589–592 (2008)
27. Chartrand, R., Staneva, V.: Restricted isometry properties and nonconvex compressive sensing. *Inverse Prob.* **24**(3), 1–14 (2008)
28. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic Decomposition by Basis Pursuit. *SIAM J. Sci. Comput.* **20**(1), 33–61 (1998)
29. Chou, E.: Non-convex decoding for sigma delta quantized compressed sensing. In: Proceeding International Conference Sampling Theory and Applications (SampTA 2013), pp. 101–104. Bremen, Germany (2013)
30. Combettes, P.L., Pesquet, J.-C.: Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212. Springer, New York (2011)

31. Dai, W., Pham, H.V., Milenkovic, O.: Distortion-Rate Functions for Quantized Compressive Sensing. Technical Report arXiv:0901.0749 (2009)
32. Daubechies, I., DeVore, R.: Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order. *Ann. Math.* **679**–710 (2003)
33. Davenport, M.A., Laska, J.N., Boufounos, P.T., Baraniuk, R.G.: A simple proof that random matrices are democratic. Technical report, Rice University ECE Department Technical Report TREE-0906, Houston, TX (2009)
34. Davenport, M.A., Laska, J.N., Treichler, J., Baraniuk, R.G.: The pros and cons of compressive sensing for wideband signal acquisition: Noise folding versus dynamic range. *IEEE Trans. Signal Process.* **60**(9), 4628–4642 (2012)
35. Deift, P., Güntürk, C.S., Krahmer, F.: An optimal family of exponentially accurate one-bit sigma-delta quantization schemes. *Commun. Pure Appl. Math.* **64**(7):883–919 (2011)
36. Edmunds, D.E., Triebel, H.: Function Spaces, Entropy Numbers, Differential Operators. Cambridge University Press, Cambridge (1996)
37. Feng, J., Krahmer, F.: An RIP approach to Sigma-Delta quantization for compressed sensing. *IEEE Signal Process. Lett.* **21**(11), 1351–1355 (2014)
38. Goyal, V.K., Vetterli, M., Thao, N.T.: Quantized overcomplete expansions in \mathbb{R}^N : Analysis, synthesis, and algorithms. *IEEE Trans. Inform. Theory* **44**(1), 16–31 (1998)
39. Gray, R.M., Neuhoff, D.L.: Quantization. *IEEE Trans. Inform. Theory* **44**(6), 2325–2383 (1998)
40. Gray, R.M.: Oversampled sigma-delta modulation. *IEEE Trans. Comm.* **35**(5), 481–489 (1987)
41. Güntürk, C.S.: One-bit sigma-delta quantization with exponential accuracy. *Commun. Pure Appl. Math.* **56**(11), 1608–1630 (2003)
42. Güntürk, C.S., Lammers, M., Powell, A.M., Saab, R., Yilmaz, Ö.: Sobolev duals for random frames and $\Sigma \Delta$ quantization of compressed sensing measurements. *Found. Comput. Math.* **13**(1), 1–36 (2013)
43. Güntürk, S.: Harmonic analysis of two problems in signal compression. PhD thesis, Program in Applied and Computation Mathematics, Princeton University, Princeton, NJ (2000)
44. Hanson, D.L., Wright, F.T.: A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Stat.* **42**(3), 1079–1083 (1971)
45. Hoeffding, W.: Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.* **58**(301), 13–30 (1963)
46. Inose, H., Yasuda, Y.: A unity bit coding method by negative feedback. *Proc. IEEE* **51**(11), 1524–1535 (1963)
47. Inose, H., Yasuda, Y., Murakami, J.: A telemetering system by code modulation – Δ – Σ modulation. *IRE Trans. Space El. Tel. SET* **8**(3), 204–209 (1962)
48. Iwen, M., Saab, R.: Near-optimal encoding for sigma-delta quantization of finite frame expansions. *J. Fourier Anal. Appl.* **19**(6), 1255–1273 (2013)
49. Jacques, L.: A quantized Johnson Lindenstrauss lemma: The finding of buffon’s needle. arXiv preprint arXiv:1309.1507 (2013)
50. Jacques, L.: Error decay of (almost) consistent signal estimations from quantized random gaussian projections. arXiv preprint arXiv:1406.0022 (2014)
51. Jacques, L., Degraux, K., De Vleeschouwer, C.: Quantized iterative hard thresholding: Bridging 1-bit and high-resolution quantized compressed sensing. In: Proceedings of International Conference Sampling Theory and Applications (SampTA 2013), arXiv:1305.1786, pp. 105–108. Bremen, Germany (2013)
52. Jacques, L., Hammond, D.K., Fadili, M.J.: Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine. *IEEE Trans. Inform. Theory* **57**(1), 559–571 (2011)
53. Jacques, L., Hammond, D.K., Fadili, M.J.: Stabilizing nonuniformly quantized compressed sensing with scalar companders. *IEEE Trans. Inform. Theory* **5**(12), 7969–7984 (2013)
54. Jacques, L., Laska, J.N., Boufounos, P.T., Baraniuk, R.G.: Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Trans. Inform. Theory* **59**(4), 2082–2102 (2013)

55. Johnson, W.B., Lindenstrauss, J.: Extensions of lipschitz mappings into a hilbert space. *Contemp. Math.* **26**(189–206), 1 (1984)
56. Kamilov, U., Goyal, V.K., Rangan, S.: Optimal quantization for compressive sensing under message passing reconstruction. In: Proceeding of IEEE International Symposium on Information Theory (ISIT), pp. 459–463 (2011)
57. Kamilov, U.S., Goyal, V.K., Rangan, S.: Message-passing de-quantization with applications to compressed sensing. *IEEE Trans. Signal Process.* **60**(12), 6270–6281 (2012)
58. Kostina, V., Duarte, M.F., Jafarpour, S., Calderbank, R.: The value of redundant measurement in compressed sensing. In: Proceeding of International Conference Acoustics, Speech and Signal Processing (ICASSP), pp. 3656–3659 (2011)
59. Krahmer, F., Mendelson, S., Rauhut, H.: Suprema of chaos processes and the restricted isometry property. *Commun. Pure Appl. Math.* **67**(11), 1877–1904 (2014)
60. Krahmer, F., Saab, R., Ward, R.: Root-exponential accuracy for coarse quantization of finite frame expansions. *IEEE Trans. Inform. Theory* **58**(2), 1069–1079 (2012)
61. Krahmer, F., Saab, R., Yilmaz, Ö.: Sigma-delta quantization of sub-gaussian frame expansions and its application to compressed sensing. *Inform. Inference* **3**(1), 40–58 (2014)
62. Krahmer, F., Ward, R.: New and improved johnson-lindenstrauss embeddings via the restricted isometry property. *SIAM J. Math. Anal.* **43**(3), 1269–1281 (2011)
63. Krahmer, F., Ward, R.: Lower bounds for the error decay incurred by coarse quantization schemes. *Appl. Comput. Harmonic Anal.* **32**(1), 131–138 (2012)
64. Kühn, T.: A lower estimate for entropy numbers. *J. Approx. Theory* **110**(1), 120–124 (2001)
65. Lammers, M., Powell, A.M., Yilmaz, Ö.: Alternative dual frames for digital-to-analog conversion in sigma–delta quantization. *Adv. Comput. Math.* **32**(1), 73–102 (2010)
66. Laska, J., Boufounos, P., Davenport, M., Baraniuk, R.: Democracy in action: Quantization, saturation, and compressive sensing. *Appl. Comput. Harmon. Anal.* **31**(3), 429–443 (2011)
67. Laska, J., Wen, Z., Yin, W., Baraniuk, R.: Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements. *IEEE Trans. Signal Process.* **59**(11), 5289–5301 (2010)
68. Laska, J.N., Baraniuk, R.G.: Regime change: Bit-depth versus measurement-rate in compressive sensing. *IEEE Trans. Signal Process.* **60**(7), 3496–3505 (2012)
69. Ledoux, M.: The Concentration of Measure Phenomenon. American Mathematical Society, Providence, RI (2005)
70. Li, M., Rane, S., Boufounos, P.T.: Quantized embeddings of scale-invariant image features for mobile augmented reality. In: Proceeding of IEEE Internatioal Workshop on Multimedia Signal Processing (MMSP), pp. 17–19. Banff, Canada (2012)
71. Lloyd, S.: Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28**(2), 129–137 (1982)
72. Lustig, M., Donoho, D., Pauly, J.M.: Sparse MRI: The application of compressed sensing for rapid MRI imaging. *Magn. Reson. Med.* **58**(6), 1182–1195 (2007)
73. Marcia, R.F., Willett, R.M.: Compressive coded aperture superresolution image reconstruction. In: Proceeding of International Conference Acoustics, Speech and Signal Processing (ICASSP), pp. 833–836. IEEE, New York (2008)
74. Max, J.: Quantizing for minimum distortion. *IEEE Trans. Inform. Theory* **6**(1), 7–12 (1960)
75. Mishali, M., Eldar, Y.C.: Sub-Nyquist sampling. *IEEE Signal Proc. Mag.* **28**(6), 98–124 (2011)
76. Nguyen, H.Q., Goyal, V.K., Varshney, L.R.: Frame permutation quantization. *Appl. Comput. Harmon. Anal.* (2010)
77. Norsworthy, S.R., Schreier, R., Temes, G.C. et al.: Delta-Sigma Data Converters: Theory, Design, and Simulation, vol. 97. IEEE press, New York (1996)
78. Pai, R.J.: Nonadaptive lossy encoding of sparse signals. M.eng. thesis, MIT EECS, Cambridge, MA (2006)
79. Panter, P.F., Dite, W.: Quantization distortion in pulse-count modulation with nonuniform spacing of levels. *Proc. IRE* **39**(1), 44–48 (1951)
80. Plan, Y., Vershynin, R.: One-bit compressed sensing by linear programming. *Commun. Pure Appl. Math.* **66**(8), 1275–1297 (2013)
81. Plan, Y., Vershynin, R.: Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Trans. Inform. Theory* **59**(1), 482–494 (2013)

82. Plan, Y., Vershynin, R.: Dimension reduction by random hyperplane tessellations. *Discret Comput. Geom.* **51**(2), 438–461 (2014)
83. Powell, A.M., Saab, R., Yilmaz, Ö.: Quantization and finite frames. In: *Finite Frames*, pp. 267–302. Springer, New York (2013)
84. Powell, A.M., Whitehouse, J.T.: Error bounds for consistent reconstruction: Random polytopes and coverage processes. *Found. Comput. Math.* (2013). doi:[10.1007/s10208-015-9251-2](https://doi.org/10.1007/s10208-015-9251-2). arXiv preprint arXiv:1405.7094
85. Rane, S., Boufounos, P.T., Vetro, A.: Quantized embeddings: An efficient and universal nearest neighbor method for cloud-based image retrieval. In: *Proceedings of SPIE Applications of Digital Image Processing XXXVI*, (2013) 885609
86. Rudelson, M., Vershynin, R.: On sparse reconstruction from Fourier and Gaussian measurements. *Commun. Pure Appl. Math.* **61**, 1025–1045 (2008)
87. Rudelson, M., Vershynin, R.: Hanson-wright inequality and sub-gaussian concentration. *Electron. Comm. Probab.* **18**, 1–9 (2013)
88. Schütt, C.: Entropy numbers of diagonal operators between symmetric Banach spaces. *J. Approx. Theory* **40**(2), 121–128 (1984)
89. Sun, J.Z., Goyal, V.K.: Optimal quantization of random measurements in compressed sensing. In: *Proceedings IEEE International Symposium on Information Theory (ISIT)*, pp. 6–10 (2009)
90. Thao, N.T., Vetterli, M.: Lower bound on the mean-squared error in oversampled quantization of periodic signals using vector quantization analysis. *IEEE Trans. Inform. Theory*, **42**(2), 469–479 (1996)
91. Thao, N.T., Vetterli, M.: Reduction of the MSE in R-times oversampled A/D conversion $O(1/R)$ to $O(1/R^2)$. *IEEE Trans. Signal Process.* **42**(1), 200–203 (1994)
92. Varanasi, M.K., Aazhang, B.: Parametric generalized Gaussian density estimation. *J. Acoust. Soc. Am.* **86**, 1404–1415 (1989)
93. Zymnis, A., Boyd, S., Candes, E.: Compressed sensing with quantized measurements. *IEEE Signal Proc. Lett.* **17**(2), 149–152 (2010)

Chapter 8

Compressive Gaussian Mixture Estimation

Anthony Bourrier, Rémi Gribonval, and Patrick Pérez

Abstract When performing a learning task on voluminous data, memory and computational time can become prohibitive. In this chapter, we propose a framework aimed at estimating the parameters of a density mixture on training data in a compressive manner by computing a low-dimensional *sketch* of the data. The sketch represents empirical moments of the underlying probability distribution. Instantiating the framework on the case where the densities are isotropic Gaussians, we derive a reconstruction algorithm by analogy with compressed sensing. We experimentally show that it is possible to precisely estimate the mixture parameters provided that the sketch is large enough, while consuming less memory in the case of numerous data. The considered framework also provides a privacy-preserving data analysis tool, since the sketch does not disclose information about individual datum it is based on.

8.1 A general idea of compressive learning

8.1.1 Conceptual compressive learning outline

When considering a learning problem, one is interested in performing a task on a certain type of data. A learning procedure will usually consist in trying to fit an underlying model to the type of data one is interested in by picking a model in a parametrized set $M = \{M_\theta : \theta \in \Theta\}$, where Θ is a parameter set which is used to

A. Bourrier
Inria Rennes-Bretagne Atlantique, Rennes, France

Technicolor, Cesson-Sévigné, France
e-mail: anthony.bourrier@gmail.com

R. Gribonval (✉)
Inria Rennes-Bretagne Atlantique, Rennes, France
e-mail: remi.gribonval@inria.fr

P. Pérez
Technicolor, Cesson-Sévigné, France
e-mail: patrick.perez@technicolor.com

index the models in M . In order to achieve this model fitting, a learning procedure will consist in finding a parameter θ^* so that the model M_{θ^*} is adequate to a training set $\mathcal{X} = \{x_1, \dots, x_L\}$ containing some data one is interested in. The computational cost of estimating such a parameter θ^* will depend on the size of \mathcal{X} , on the size of the models in M , and on the algorithm used for the estimation.

When considering numerous and/or high-dimensional training data, or even refined models, it is necessary to come up with approximate learning schemes which will allow one to learn parameters from data with a fair precision while requiring less memory and/or computational time. A possible conceptual compressive framework to perform such an estimation is outlined in Figure 8.1, which represents two main ways of compressing learning data in order to apply a learning algorithm to data of reduced size. The top scheme represents a case where each vector of \mathcal{X} is compressed individually: this is performed, for instance, in [7] and in Chapter “Compressive Classification: Where Wireless Communications Meets Machine Learning.” A widely used technique to instantiate this first scheme is the Principal Component Analysis (PCA). The bottom scheme, which is the scheme we will be interested in throughout this chapter, represents the case where \mathcal{X} will be compressed into a single representation usually called *sketch*, of size m which should not depend on the number L of elements in the database but rather on the complexity of the model one wants to fit (represented by the parameter set Θ) and on the task one aims at achieving after the learning. This second scheme has been instantiated in a simple estimation problem in [17], which will be discussed in the next section.

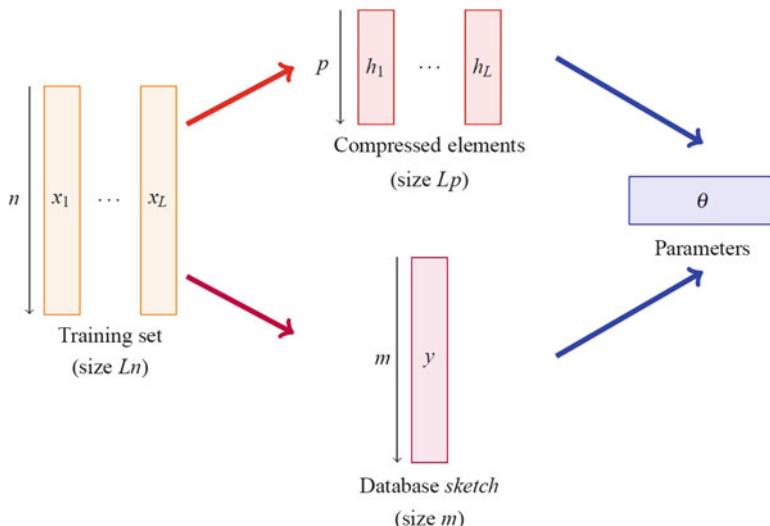


Fig. 8.1 Compressive learning outline. The learning data $\mathcal{X} = \{x_1, \dots, x_L\}$ is compressed into a smaller representation, which can either consist in reducing the dimensions of each individual entry x_r or in computing a more global compressed representation of the data, called *sketch*. Parameters θ are then inferred from such a compressed representation by an algorithm adapted to the representation.

8.1.2 Density mixture estimation

In this chapter, we will focus on a classical unsupervised learning problem: density mixture estimation.

Let's model the problem as follows: assume the data set \mathcal{X} comprises vectors in \mathbb{R}^n , and consider a set P of parametrized probability densities defined on \mathbb{R}^n , that is

$$P = \left\{ p_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^+ \left| \int_{\mathbb{R}^n} p_\theta \, d\ell = 1, \theta \in \Theta \right. \right\}, \quad (8.1)$$

where ℓ is the Lebesgue measure on \mathbb{R}^n .

The goal of density mixture estimation is to find a density function p on \mathbb{R}^n which satisfies two prerequisites:

- p must be written as a linear combination, or *mixture*, of functions in P , that is $p = \sum_{s=1}^k c_s p_{\theta_s}$, where the quantities c_s are positive *weights* satisfying $\sum_{s=1}^k c_s = 1$ and θ_s are the *parameters* of the mixture. Let's notice the similarity of this condition with a usual sparsity condition over a dictionary.
- \mathcal{X} can “reasonably well” be considered as a set of vectors drawn *i.i.d.* with respect to a probability law of density p . This condition is typically enforced by optimizing a certain quantity representing the consistency between the data \mathcal{X} and the probability p , such as the likelihood of the parameters $\mathbb{P}(\mathcal{X} | (c_s, \theta_s)_{s=1}^k)$, which represents the probability of the drawing of the data in \mathcal{X} knowing the probability density p . Alternatively, this condition can be formulated by supposing the data \mathcal{X} is drawn *i.i.d.* from a probability distribution of density f , and that one searches for the best approximation p of f , p being an element of the set of sparse linear combinations of functions in P .

Since p must be written as a linear combination of a few functions of P , one can consider it as a “sparse vector” over the family P . In that sense, several works have considered the density mixture estimation problem in a linear inverse problem fashion, and will be discussed in the next section. Drawing our inspiration from these works, our goal in this chapter will be to propose an instantiation of the compressive scheme represented in the bottom part of Figure 8.1, applying it to a density mixture estimation problem. The method we propose can be interpreted as compressive sensing with respect to a signal model¹ beyond the standard sparse model. Such an extension of compressed sensing to other or more general models than sparse finite-dimensional vectors has been studied, for instance, for tensors in Chapter “Two algorithms for compressed sensing of sparse tensors” or for the cosparse model in Chapter “Cosparcity in Compressed Sensing.” Fundamental performance limits of decoders for generalized models are also studied in [4].

¹Namely, a model of probability measures.

In the next section, on top of discussing the works relative to density mixture estimation in a linear inverse problem fashion, we will also mention other works which rely on computing a *sketch* of some data.

8.2 Related work: density mixture estimation and sketching

In this section, we present different sets of works which motivate our contribution: on the one hand, works on density mixture estimation expressed as a linear inverse problem; on the other hand, works aimed at performing learning tasks on compressed data. In particular, some of these latter works focus on analyzing data streams, thanks to global compressed representations called *sketches*. Our approach will combine both aspects.

8.2.1 Density mixture estimation as a linear inverse problem

To express the density mixture estimation problem in a linear inverse fashion, the works [2, 6] consider P as a finite set $\{p_1, \dots, p_M\}$, where all the densities p_s and the density f belong to the Hilbert space $L^2(\mathbb{R}^n)$. Assume the vectors \mathcal{X}_i are drawn *i.i.d.* from a probability distribution of density f .

In this case, the density mixture estimation consists in finding a sparse vector $c \in \mathbb{R}^M$, such that $f \approx f_c$, where $f_c = \sum_{s=1}^M c_s p_s$. This objective can be specified in different ways in a linear inverse problem fashion. They all share the following observation: if \mathbb{E} is the expectation with respect to the density f , then the scalar product between f and p_s , defined by

$$\langle f, p_s \rangle = \int_{\mathbb{R}^n} p_s f \, d\ell = \mathbb{E}[p_s(X)], \quad (8.2)$$

can be approximated by an empirical mean obtained from the data \mathcal{X} . This empirical estimator is defined as

$$\hat{\mathbb{E}}[p_s(X)] = \frac{1}{L} \sum_{r=1}^L p_s(x_r), \quad (8.3)$$

where $\hat{\mathbb{E}}$ is the expectation with respect to the empirical probability distribution of the data \mathcal{X} .

In [6], the authors aim at minimizing over c the quantity

$$\|f - f_c\|_2^2 + J(c), \quad (8.4)$$

where J is a penalty function aimed at promoting sparsity, and which is a weighted version of the ℓ_1 -norm depending on the family P .

The first term can be developed using the scalar product, and can be replaced without changing the solution to problem (8.4) by the term

$$-2\mathbb{E}[f_c(X)] + \|f_c\|_2^2. \quad (8.5)$$

From there, the authors simply replace the theoretical expectation \mathbb{E} in (8.5) by its empirical counterpart $\hat{\mathbb{E}}$. This defines the “SPADES” estimator as

$$\hat{c} = \underset{c \in \mathbb{R}^M}{\operatorname{argmin}} -\frac{2}{L} \sum_{r=1}^L f_c(x_r) + \|f_c\|_2^2 + J(c). \quad (8.6)$$

They further propose inequalities for the estimation error under different assumptions on the family P and the choice of the penalty term J .

In [2], the authors apply the Dantzig selector [8] to the problem by expressing it as

$$\underset{c \in \mathbb{R}^M}{\operatorname{argmin}} \|c\|_1, \text{ s.t. } |(Gc)_s - \hat{\mathbb{E}}[p_s(X)]| \leq \eta_s \text{ for all } 1 \leq s \leq M. \quad (8.7)$$

In this formulation, the matrix G is the Gram matrix of the family P , for which the (s_1, s_2) entry is equal to $\langle p_{s_1}, p_{s_2} \rangle$. The quantity η_s is an adaptive threshold relative to each density p_s . The objective aims at finding a sparse solution via ℓ_1 relaxation, while the constraints ensure that the theoretical correlation $\langle p_s, f_c \rangle = (Gc)_s$ is close to the empirical estimator of $\langle p_s, f \rangle$ defined as $\hat{\mathbb{E}}[p_s(X)]$.

These two methods have been successfully applied to density mixture estimation for several models in dimension $n = 1, 2$ in [6] and $n = 1$ in [2]. However, they suffer from two major drawbacks if one aims at applying them to higher-dimensional models. Let's give somewhat intuitive insights into the reasons for these drawbacks:

- *Finiteness of the density model:* The family P is assumed to be finite. However, several density models typically used in estimation are infinite, such as Gaussian Mixture Models (GMMs), which consider Gaussian densities indexed on a certain number of parameters depending on n . If one aims, for instance, at applying these methods to GMM estimation, one needs to discretize the continuous model of Gaussian densities. Since a number of $(\frac{R}{h})^n$ is required to sample a cube of side R with a mesh h in each direction, the number of centers in a discrete representation will grow exponentially fast with n , which will not be viable computationally.
- *Incoherence of the density model:* Both methods rely on the fact that P is a family of *incoherent* densities, that is the quantities

$$\frac{\langle p_{s_1}, p_{s_2} \rangle}{\|p_{s_1}\|_2 \|p_{s_2}\|_2} \quad (8.8)$$

are not too close to 1 if $s_1 \neq s_2$. This incoherence necessity prevents from decomposing f on a refined model containing many similar densities.

Therefore, if these approaches are viable and theoretically sound for density mixture estimation for small dimension ($n = 1, 2, 3$), they may not be applied in all generality for moderate dimensions (say, even $n = 5, 10$). For such dimensions, it would be more interesting to consider a continuous model of densities p_θ indexed by a continuous parameter θ , such as a vector of a certain space \mathbb{R}^d . However, such a continuous model cannot be treated with the aforementioned methods.

Keeping these limitations in mind, let's present the other set of inspiring contributions for our work, which can be viewed as compressive learning instances using sketches.

8.2.2 *Learning with data stream sketches*

8.2.2.1 Data stream sketches

Data streams consist in a flow of data $(x_r)_{r \geq 1}$. In some models, the data stream is important in itself; in others, the data x_r act simply as modification step of an underlying item x , which is the object of interest (for instance, x could be a vector and the elements x_r could encode modifications of this vector such as adding or subtracting a unit in a particular entry).

Standard statistical problems involving data streams include the search for frequent items among the x_r , usually called *heavy hitters* [10, 13], or more generally estimation of quantiles of the content of the stream. When aiming at obtaining such information at a given time without having to store all the data flow up to this point, it is necessary to maintain a compressed representation of the data stream which will allow one to perform the desired estimations. Such compressed representations can be deterministically built, but are sometimes built similarly to hashing functions [14, 15]. In this case, they are called *sketches*.

A sketch y is usually updated each time a new element (x_r) is streamed. Examples of usual sketches include the Count Sketch [9] and the Count-Min Sketch [11]. They both rely on updating y thanks to randomly chosen hashing functions. They are mainly used to estimate the heavy hitters.

8.2.2.2 Compressed histogram estimation

Interestingly, [17] proposes a sketching procedure which can be linked to compressed sensing and to density mixture estimation. In this work, the authors consider a data stream $(x_r)_{r \geq 1}$, where each x_r is an n -dimensional vector taken from a finite set $A \subset \mathbb{R}^n$. The goal is to obtain, at a given point r_0 , a histogram H_{r_0} approximating the distribution of vectors x_r for $r \leq r_0$.

In order to avoid storing and updating a complete histogram of the data stream as it flows, the authors propose instead to build and update a sketch of such a histogram. This sketch is obtained by considering a low-dimensional projection of a histogram H by a randomly built linear operator A designed to approximately

preserve distances between histograms, still in a way similar to [1, 16]. The sketch can be updated at each time r by considering x_r as a histogram H_r which is null everywhere except in the bin corresponding to x_r , where it is 1.

The benefit of this framework is the reduction of memory costs, since the whole histogram need not be updated, while still allowing one to compute at any time a good approximation of the histogram. However, the main drawback is the complexity of the recovery procedure, which is exponential in the dimension n of the data. This prevents from applying this method to even moderate dimensions (say, $n = 10$).

8.3 Compressive estimation framework

We will now present the proposed compressive estimation method [5]. Let's recall that we want to consider the following problem in a compressive way: let $\mathcal{X} = \{x_r\}_{r=1}^L$ be vectors in \mathbb{R}^n , drawn *i.i.d.* from a certain probability distribution of density $p \in L^1(\mathbb{R}^n)$. In the rest of this chapter, the notation $\Sigma_k(P)$ will denote the positive linear combinations of k densities in P , that is

$$\Sigma_k(P) = \left\{ \sum_{s=1}^k \lambda_s p_{\theta_s} : \lambda_s \in \mathbb{R}_+, \theta_s \in \Theta \right\}. \quad (8.9)$$

Let's note that $P \subset \Sigma_k(P)$. Our goal is to find a good estimate of p in the set $\Sigma_k(P)$.

To simplify the problem, we will assume that p is an exact k -sparse mixture of densities taken in P , defined in (8.1), that is $p = \sum_{s=1}^k c_s p_{\theta_s}$, with $c_s \geq 0$, $\sum_{s=1}^k c_s = 1$ and $\theta_s \in \Theta$. In this case, the most natural way to approximate p with a density of Σ_k is to try and estimate $c = (c_1, \dots, c_k)$ and $\theta_1, \dots, \theta_k$ from \mathcal{X} . We further want to perform this estimation in a compressive fashion. By analogy with compressed sensing, let's derive a conceptual method.

8.3.1 The compressive operator

The unknown “signal” p is a sparse linear combination of the p_{θ} . One would like to reconstruct p from a “compressive” measure Ap , where A is a linear operator which transforms a density function into a finite-dimensional representation, so that one is able to manipulate it. To reconstruct p from Ap , one will then look for an element q of $\Sigma_k(P)$ satisfying $Aq \approx Ap$. The main raised issue for now is to design an adequate linear measurement operator A . This operator should satisfy the following requirements:

- *Estimation of Ap :* The empirical representation of the density p is the discrete density associated with the collection of vectors \mathcal{X} . Since p is unknown, one cannot compute directly Ap , and so the operator A must be such that Ap can be estimated through this empirical distribution.

- *Computation of AP*: The reconstruction algorithm will aim at finding densities of Σ_k which will have an image by A similar to the empirical value of Ap computed from \mathcal{X} . To reconstruct the density, one should therefore be able to effectively compute the value of Af for any $f \in \Sigma_k(P)$ (or $f \in P$, which is equivalent).

Assume the operator A transforms a function into a compressed representation of dimension m . A can be seen as the concatenation of m linear forms A_1, \dots, A_m . Since we made the simplifying assumption that p belonged to $\Sigma_k \subset \langle P \rangle$, one only needs to define the linear forms A_j on the complex span of P , denoted $\langle P \rangle$. Considering the complex span instead of the real span will allow us to simplify the expressions of linear forms based on Fourier transforms, which we consider in the following.

These linear forms must satisfy the two above conditions. Simple linear forms on $\langle P \rangle$ can be defined as

$$A_g : f \mapsto \int_{\mathbb{R}^n} fg \, d\ell, \quad (8.10)$$

where g is a bounded measurable function on \mathbb{R}^n . If f is a probability density on \mathbb{R}^n , such linear forms *generalized moments* of f . In particular, the required conditions are easily interpreted for A_g :

- *Estimation of $A_g p$* : Since p is a probability density on \mathbb{R}^n ,

$$A_g p = \int_{\mathbb{R}^n} gp \, d\ell = \mathbb{E}[g(X)], \quad (8.11)$$

where $\mathbb{E}[\cdot]$ is the expectation taken with respect to the probability law of density p . Therefore, the value of $A_g p$ can be approximated by computing the empirical estimate

$$\hat{A}_g(X) = \frac{1}{L} \sum_{r=1}^L g(x_r) = \hat{\mathbb{E}}[g(X)], \quad (8.12)$$

where $\hat{\mathbb{E}}[\cdot]$ is the expectation with respect to the empirical distribution with L equal masses $\frac{1}{L}$ at each vector of $\mathcal{X} = \{x_1, \dots, x_L\}$. For such a choice of linear forms, it is therefore possible to estimate $A_g p$ for virtually any g .

Moreover, concentration of measure results such as Hoeffding's inequality or McDiarmid's inequality provide confidence intervals on the estimation error. In particular, if the function g takes value in a ball of radius R , then with probability greater than $1 - \delta$ on the drawing of the data vectors, one has

$$\left| \frac{1}{L} \sum_{r=1}^L g(x_r) - \int_{\mathbb{R}^n} gp \, d\ell \right| \leq \varepsilon, \quad (8.13)$$

provided

$$L \geq \frac{R}{\sqrt{2}\varepsilon} \sqrt{\ln\left(\frac{2}{\delta}\right)}. \quad (8.14)$$

- *Computation of AP*: The functions g should be chosen so that the value of $A_g p_\theta$ is computable in closed form.

A compression scheme analog to compressed sensing will consist in considering a family G of functions so that $g.p_\theta$ is integrable for any $\theta \in \Theta$ and $g \in G$. Having defined a probability distribution on the family G , one will be able to randomly choose a compressive operator A by drawing *i.i.d.* m functions $g_1, \dots, g_m \in G$ and defining $A = (A_1, \dots, A_m)$.

8.3.2 Proposed instantiation: isotropic Gaussians

Let's now propose a particular instantiation of this framework which we will use in the rest of this chapter. Given $\sigma > 0$, let's define the family P as

$$P_\sigma = \left\{ p_\mu : x \mapsto \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left(-\frac{\|x - \mu\|_2^2}{2\sigma^2}\right), \mu \in \mathbb{R}^n \right\}. \quad (8.15)$$

This family contains all isotropic Gaussians of \mathbb{R}^n of variance σ^2 , indexed by their n -dimensional mean μ , which uniquely characterizes them. As before, we define $\Sigma_k(P_\sigma)$ as the linear combinations of at most k functions of P_σ and we adopt the simplifying notation $\Sigma_k(P_\sigma) = \Sigma_k$.

Natural linear forms associated with this type of functions are Fourier measurements. Each Fourier measurement can be indexed by a frequency vector $\omega \in \mathbb{R}^n$ and the corresponding function g can be defined as $g_\omega(x) = \exp(-i\langle x, \omega \rangle)$. The corresponding linear form, denoted as A_ω , is therefore:

$$A_\omega : q \mapsto \int_{\mathbb{R}^n} q(x) e^{-i\langle x, \omega \rangle} dx. \quad (8.16)$$

Given data $X = \{x_1, \dots, x_L\}$, the empirical counterpart of this linear form is

$$\hat{A}_\omega(\mathcal{X}) = \frac{1}{L} \sum_{r=1}^L \exp(-i\langle x_r, \omega \rangle). \quad (8.17)$$

A compressive operator can therefore be defined by choosing m frequencies $\omega_1, \dots, \omega_m$ and posing $A = (A_{\omega_1}, \dots, A_{\omega_m})$. The frequencies ω_j will typically be randomly drawn, so that for a probability density $f \in \langle P_\sigma \rangle$, Af can be interpreted as the sampling of the characteristic function of f at m random frequencies. We specify the random model for choosing the frequencies in section 8.5.

The value of $A_\omega p_\mu$ is explicitly computable: one has

$$A_\omega p_\mu = \exp\left(-\frac{\sigma^2}{2}\|\omega\|_2^2\right) \exp(-i\langle\omega, \mu\rangle). \quad (8.18)$$

8.3.3 Recovery Problem formulation

Given the data \mathcal{X} , we will denote \hat{y} the empirical sketch of \mathcal{X} , which is the m -dimensional vector $\hat{y} = (\hat{A}_{\omega_1}, \dots, \hat{A}_{\omega_m})$. One can express the recovery of the initial density p from \hat{y} as the following minimization problem:

$$\hat{p} = \underset{q \in \Sigma_k}{\operatorname{argmin}} \frac{1}{2} \|\hat{y} - Aq\|_2^2. \quad (8.19)$$

Despite being nonconvex, this kind of formulation of the problem is addressed in a regular compressed sensing setting by greedy algorithms. In the next section, we will derive from such a standard method an algorithm aimed at solving (8.19).

8.4 Compressive reconstruction algorithm

To address the estimation problem (8.19), we propose an algorithm analogous to Iterative Hard Thresholding (IHT) [3].

8.4.1 Reminder of Iterative Hard Thresholding

Consider a k -sparse signal x of dimension n and a measurement matrix A of size $m \times n$ (with $m < n$). Denoting $y = Ax$ the measurement of x , IHT considers the minimization

$$\underset{z \in \Sigma_k}{\operatorname{argmin}} \|y - Az\|_2^2, \quad (8.20)$$

where Σ_k is now the set of k -sparse vectors in \mathbb{R}^n . At each iteration, IHT updates an estimate \hat{x} of x , decreasing the objective function $F : \hat{x} \mapsto \frac{1}{2} \|y - A\hat{x}\|_2^2$ while ensuring the k -sparsity of \hat{x} . The quantity $r = y - A\hat{x}$ is named the *residual*. The update step is performed in two steps:

1. The n -dimensional gradient of F , noted ∇F , is computed.
2. The update is given by $\hat{x} \leftarrow H_k(\hat{x} - \lambda \nabla F)$, where λ is a descent step and H_k is a hard thresholding operator which keeps only the k entries of the vector with largest module and sets the others to 0.

8.4.2 Proposed continuous case algorithm

We also adopt an iterative greedy method to perform the reconstruction of p . We therefore iteratively update an estimate \hat{p} , which is parametrized by a vector $\hat{a} \in \mathbb{R}^k$ of positive weights and by the support $\hat{\Gamma} = \{\hat{\mu}_1, \dots, \hat{\mu}_k\} \subset \mathbb{R}^n$ corresponding to the means of the current estimated Gaussians. The current residual is defined by $\hat{r} = \hat{y} - A\hat{p}$. In our case, the function F takes \hat{p} as an argument and is defined as

$$F(\hat{p}) = \frac{1}{2} \|\hat{y} - A\hat{p}\|_2^2. \quad (8.21)$$

There are some differences between our density estimation problem and the problem addressed by IHT which require modifications of the procedure. They are explained in the following sections. The algorithm is then summarized.

8.4.2.1 The “gradient”: continuous version

In IHT, the signal one wants to reconstruct is assumed to be sparse in a finite basis of vectors. The gradient computed in the first step is a finite-dimensional vector, and each entry corresponds to the infinitesimal shift of the objective function F when a certain entry of the vector is shifted.

In our case, the density is assumed to be sparse in the “infinite basis” P_σ , which is parametrized by \mathbb{R}^n . The “canonical” directions in which \hat{p} can be shifted are therefore also parametrized by \mathbb{R}^n , and the “gradient” is the collection of these possible shifts, denoted by $\nabla_\mu F$ for all $\mu \in \mathbb{R}^n$ and defined as follows:

$$\nabla_\mu F = \left(\frac{\partial}{\partial t} \frac{1}{2} \|\hat{y} - A(\hat{p} + tp_\mu)\|_2^2 \right)_{t=0} = -\langle Ap_\mu, \hat{r} \rangle. \quad (8.22)$$

Again, this quantity represents the local variation of the objective function (8.19) when the density p_μ is added to the current estimate. Since we cannot compute these values for every $\mu \in \mathbb{R}^n$, we must only choose a finite number of μ for which we will compute $\nabla_\mu F$.

Since we aim at decreasing F , these directions should be chosen so that $\nabla_\mu(\hat{p})$ is negatively minimal, so that p_μ is a seemingly good candidate to be added to the current estimate \hat{p} . Therefore, we seek instead a certain number M of local minima of $\mu \mapsto \nabla_\mu \hat{p}$, where M will typically be chosen as $\mathcal{O}(k)$. These local minima parametrize elements of P which are the best correlated elements to the residual \hat{r} . They are the best directions in which to “move” locally the estimate \hat{p} in order to decrease the objective function F . These local minima are searched for by a randomly initialized minimization algorithm. When they are found, they are added to the current support $\hat{\Gamma}$, increasing its size up to $K = M + k$ elements.

8.4.2.2 Hard Thresholding

The second step in IHT consists in choosing a descent step used to shift the current estimate in the direction of the gradient, and enforcing sparsity through hard thresholding. In our case, we have at this point an updated collection of candidate means $\hat{\Gamma}$, and we want to keep only k of these means.

In order to do this, we aim at decomposing the empirical sketch \hat{y} as a positive linear combination of vectors in Ap_v , with $v \in \hat{\Gamma}$. We therefore project the sketch \hat{y} on the cone generated by the sketches of the functions $\{p_v : v \in \hat{\Gamma}\}$. With $\hat{\Gamma} = \{v_1, \dots, v_K\}$, this cone is defined as

$$C(A\hat{\Gamma}) = \left\{ \sum_{j=1}^K \lambda_j Ap_{v_j} : \lambda_j \geq 0 \right\}. \quad (8.23)$$

The aforementioned projection of \hat{y} on $C(A\hat{\Gamma})$ is expressed as the following minimization problem:

$$\underset{\beta \in \mathbb{R}_+^K}{\operatorname{argmin}} \|\hat{y} - N\beta\|_2, \quad (8.24)$$

where N is the concatenation of the sketches of the functions parametrized by $\hat{\Gamma}$, that is

$$N = [Ap_{v_1} \dots Ap_{v_K}]. \quad (8.25)$$

The hard thresholding step is then performed by keeping the k largest coefficients and the k corresponding parameters of $\hat{\Gamma}$ found in (8.24). Note that in the framework we consider (isotropic Gaussians with Fourier measurements), the quantity $\|Ap_v\|_2$ does not depend on v , that is the sketches all have the same energy. In the case where they do not have the same energy, one should keep the k coefficients such that $\|\beta_j Ap_{v_j}\|_2$ is maximal.

8.4.2.3 Gradient descent step

In IHT, an iteration stops when hard thresholding is performed. In our case, we can still perform an additional step, which consists in decreasing further the objective function F .

At this point in the iteration, \hat{p} is defined as

$$\sum_{s=1}^k \hat{c}_s p_{\hat{\mu}_s}, \quad (8.26)$$

where the parameters \hat{c}_s and $\hat{\mu}_s$ hopefully estimate the real parameters c_s and μ_s of p .

Since the family P_σ is extremely coherent, the local minima we found in the previous steps may be shifted from the true mean vectors because of the imprecision induced by the other components of the mixture. This imprecision on the μ_s obviously also implies imprecision on the coefficients c_s . However, there may exist a better estimate for p in the vicinity of \hat{p} .

To find it, we simply consider G as a slightly different version of F : G represents the same cost function, but takes the parameters $\hat{c}_1, \dots, \hat{c}_k$ and μ_1, \dots, μ_k as arguments. Therefore, it is defined as:

$$\begin{aligned} G : \mathbb{R}^k \times (\mathbb{R}^n)^k &\rightarrow \mathbb{R} \\ (c, \mu_1, \dots, \mu_k) &\mapsto \frac{1}{2} \|\hat{y} - [A \mu_1 \dots A \mu_k] c\|_2^2. \end{aligned} \quad (8.27)$$

Initializing the parameters to the current estimators \hat{c} and $\hat{\mu}_1, \dots, \hat{\mu}_k$, we can apply a gradient descent algorithm on G to find, in the vicinity of \hat{p} , a better estimate in the sense that it has a smaller image under F .

8.4.2.4 Algorithmic scheme

The overall procedure is summarized below. It consists of three steps per iteration:

1. M local minima of $\mu \mapsto \nabla \hat{p}(\mu)$ are sought with a gradient descent algorithm with random initialization and are added to the current support \hat{I} .
2. The sketch \hat{y} is projected on $A \hat{I}$ with a positivity constraint on the coefficients. Only the k highest coefficients and the corresponding vectors in the support are kept.
3. A gradient descent algorithm is applied to further decrease the objective function with respect to the weights and support vectors.

The global iteration is repeated until a convergence criterion is achieved.

8.4.3 Memory usage

Let's now estimate the order of magnitude of the memory required by the compressive algorithm to estimate p from \hat{y} . Let's consider that n , k , and m are much larger than 1. If we assume that optimization algorithms only use first-order quantities, their memory costs are dominated by $\mathcal{O}(kn)$. The computation of the cost function F has cost $\mathcal{O}(km)$. The storage of the operator A (via the frequencies ω_j) requires $\mathcal{O}(mn)$.

Therefore, the total memory usage is $\mathcal{O}((k+n)m + kn)$ and does not depend on the number L of vectors. In comparison, the memory requirement of a standard Expectation-Maximization (EM) algorithm is $\mathcal{O}(L(k+n))$ to store both the vectors and their probabilities to belong to each current component of the mixture. The compressed algorithm allows memory savings as soon as $m + \frac{kn}{k+n} \lesssim L$. Since $kn \lesssim m$, this condition is nearly equivalent to $m \lesssim L$.

This suggests that one will be able to make memory savings if the number of vectors in the training set \mathcal{X} is larger than the size of the sketch required to perform the reconstruction.

8.4.4 Computational complexity

Computational complexity is the main drawback of the compressed procedure, since this procedure relies on several optimization steps, which can involve many variables for large k and n .

More precisely, the computational bottleneck is the last step of the iteration where a gradient descent is performed. This optimization procedure involves $k(n + 1)$ variables and the cost for computing the function at a certain point is $\mathcal{O}(mk)$. Therefore, since a first-order optimization algorithm requires the computation of the gradient for each variable, the complexity of a simple gradient descent implementation is $\mathcal{O}(mk^2(n + 1))$. Moreover, the cost of computing the sketch from the training data is $\mathcal{O}(mnL)$, and must be taken into account unless the data is streamed so that the sketch can be computed “on the fly” before the estimation procedure *per se*. The overall cost is virtually $\mathcal{O}(mnL + mk^2(n + 1)n_{comp})$, where n_{comp} is the number of iterations performed in the compressed estimation procedure.

This must be compared to a standard EM algorithm, which has complexity $\mathcal{O}(knLn_{EM})$, where n_{EM} is again the number of iterations performed in the EM algorithm. In the case the sketch is not computed on the fly, there will be a gain in complexity in the compressed case only if $m \ll kn_{EM}$. This was not the case in our experiments (described in the next section), since the EM algorithm converged quickly enough so that $m \sim kn_{EM}$. We will further discuss the computational outlooks in Section 8.6.

8.5 Experiments

8.5.1 Experimental setup

To evaluate the behavior of the compressive reconstruction algorithm, we conducted experiments on vectors drawn from a mixture of k isotropic Gaussians with identity covariance matrices ($\sigma = 1$). In each case, we drew weights $\{c_s\}$ uniformly on the simplex² so that they were positive and sum up to 1. We chose the Gaussian means μ_j by drawing random vectors, each entry being drawn from a probability law of density $\mathcal{N}(0, 1)$.

²We also performed experiments where all the weights were equal to $\frac{1}{k}$ and this didn’t alter the conclusions drawn from the experiments.

The experiments were performed in the following way: after the choice of the probability distribution p , we drew L random vectors from this probability distribution and computed the empirical sketch of the distribution in one pass on the data. The training samples were then discarded from hard memory. We chose the sketching operator A randomly, following the scheme described in Section 8.5.2. We then applied the reconstruction algorithm to the empirical sketch \hat{y} to get an approximated mixture \hat{p} . The random initialization for the reconstruction algorithm is detailed in Section 8.5.3. Experimental results in dimension $n = 10$ are given in Section 8.5.4.

To evaluate the quality of the estimation, we relied on two usual discrepancy measures between probability density functions. They were used to quantify the difference between the true mixture p and the estimated mixture \hat{p} . The two considered quantities are defined by integrals, which in our case could not be computed explicitly. Therefore, we approximated the integrals by empirical means: we drew $N = 10^5$ points $(y_i)_{i=1}^N$ *i.i.d.* from p and computed the empirical estimates described below. The two chosen measures were:

- *Kullback–Leibler (KL) divergence:* A symmetrized version of KL divergence can be defined as

$$D_{KL}(p, \hat{p}) = \int_{\mathbb{R}^n} \left[\ln \left(\frac{p(x)}{\hat{p}(x)} \right) p(x) + \ln \left(\frac{\hat{p}(x)}{p(x)} \right) \hat{p}(x) \right] dx. \quad (8.28)$$

The empirical estimate we considered is defined as

$$\hat{D}_{KL}(p, \hat{p}) = \frac{1}{N} \sum_{r=1}^N \left[\ln \left(\frac{p(y_r)}{\hat{p}(y_r)} \right) + \frac{\hat{p}(y_r)}{p(y_r)} \ln \left(\frac{\hat{p}(y_r)}{p(y_r)} \right) \right]. \quad (8.29)$$

The KL divergence ranges from 0 to $+\infty$, lower values meaning closer distributions.

- *Hellinger distance:* The Hellinger distance can be defined as

$$D_H(p, \hat{p}) = 1 - \int_{\mathbb{R}^n} \sqrt{p(x)\hat{p}(x)} dx. \quad (8.30)$$

The empirical estimate we considered is defined as

$$\hat{D}_H(p, \hat{p}) = 1 - \frac{1}{N} \sum_{r=1}^N \sqrt{\frac{\hat{p}(y_r)}{p(y_r)}}. \quad (8.31)$$

The Hellinger distance ranges from 0 to 1. Here again, lower values mean closer distributions.

8.5.2 Choice of the frequencies

Let's now describe the heuristic we considered to randomly choose the compressive operator A . Let's recall that $p \in \Sigma_k$, so that $p = \sum_{s=1}^k c_s p_{\mu_s}$, with the c_s positive which sum to 1. Denoting by $\mathcal{F}(f).\omega$ the Fourier transform of a function f taken at frequency ω , we have

$$\begin{aligned} |\mathcal{F}(p).\omega| &= \left| \sum_{s=1}^k c_s p_{\mu_s} \right| \leq \sum_{s=1}^k c_s |\mathcal{F}(p_{\mu_s}).\omega| \\ &= \sum_{s=1}^k c_s \exp\left(-\frac{\sigma^2}{2} \|\omega\|_2^2\right) = \exp\left(-\frac{\sigma^2}{2} \|\omega\|_2^2\right). \end{aligned} \quad (8.32)$$

This upper bound on the value of $\mathcal{F}(p)$ gives a hint on the way to design the random choice of frequencies. Indeed, we want to sample frequencies which are likely to be “energetic,” so that $|\mathcal{F}(p).\omega|$ is not “too low.” The frequencies were therefore chosen as $\omega = ru$, where r is a real number drawn with respect to a centered Gaussian of variance $1/\sigma^2$ (corresponding to the upper bound found in (8.32)) and u is a direction uniformly drawn on the ℓ^2 sphere.

8.5.3 Heuristic for random initialization

The search for local minima in the first step of the algorithm was initialized randomly by exploiting a measure performed during the construction of the sketch: during the single pass on the data, the norms of the vectors x_r are computed and the maximum of the norms, $R = \max_{x \in \mathcal{X}} \|x\|_2$, is computed. These calculations have a negligible impact on the computation time of the sketch, and on its size (it only adds one component which “completes” the sketch). The knowledge of R allows us to delimit a ball in which the centers of the Gaussians are very probably contained.

We performed the random initialization by drawing a direction uniformly on the unit sphere and multiplying this unit vector by a scalar uniformly drawn in $[0; R]$.

8.5.4 Results

Figure 8.2 visually illustrates the behavior of the algorithm on a simple mixture of 4 Gaussians in dimension 2. $L = 10^3$ points were drawn from this mixture and used to compute an $m = 30$ -dimensional sketch. As shown in the figure, the mixture parameters are precisely estimated without referring to the initial data. The symmetric KL divergence and Hellinger distance are, respectively, 0.026 and 0.003.

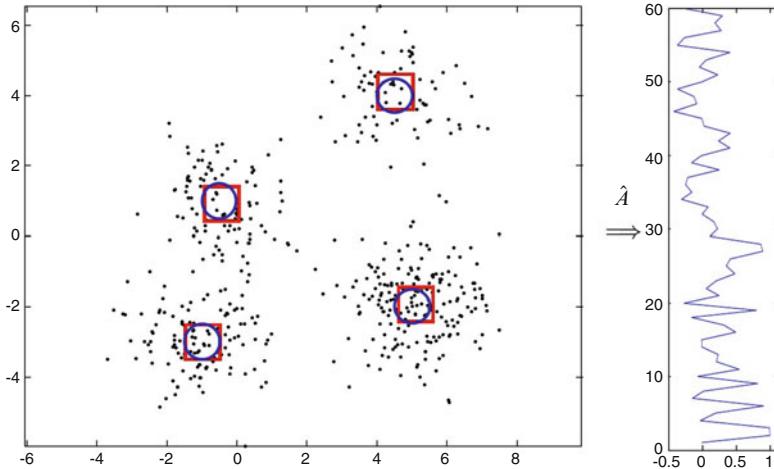


Fig. 8.2 Real and reconstructed centroids (respectively represented as circles and squares) of 4 Gaussians in dimension 2 from 10^3 points drawn from the mixture. To estimate the 12 real parameters of the mixture, the data was compressed to a complex-valued sketch of dimension 30, represented to the right as a 60-dimensional real signal.

Figure 8.3 illustrates the reconstruction quality of our algorithm in dimension 10 for different values of mixture components k and sketch sizes m in terms of Hellinger distance. For each sketch size m ranging from 200 to 2000 with step size 200, k was chosen to range from $m/200$ to $m/10$ with step $m/200$. For each choice of parameters, 10 experiments were performed and the depicted value is the Hellinger distance such that 80% of the experiments lead to a smaller Hellinger distance. We can observe a gradually increasing measure of the Hellinger distance as the number of mixture components rises. For the considered parameters range, choosing $m = 10kn$, *i.e.*, choosing m so that it contains 10 times more values than the number of parameters to estimate, leads to a Hellinger distance smaller than 0.03 for 80% of the cases.

Table 8.1 compares our algorithm with a standard EM algorithm [12] in the case where $n = 20$, $k = 10$, $m = 1000$ for values of dataset size L ranging from 10^3 to 10^5 . For each case, we can see that the precision of the estimation increases with the number of samples. In the compressed case, this can be explained by the fact that the components of the sketch are better estimated with more points. We notice that the memory used for EM is proportional to the number L of samples in the dataset, while the memory required by the compressed algorithm does not depend on this parameter, which leads to a substantial improvement in memory usage for $L \geq 10^4$. Even with this reduced memory cost, the compressed algorithm is able to provide a precision comparable to the precision of the EM algorithm.

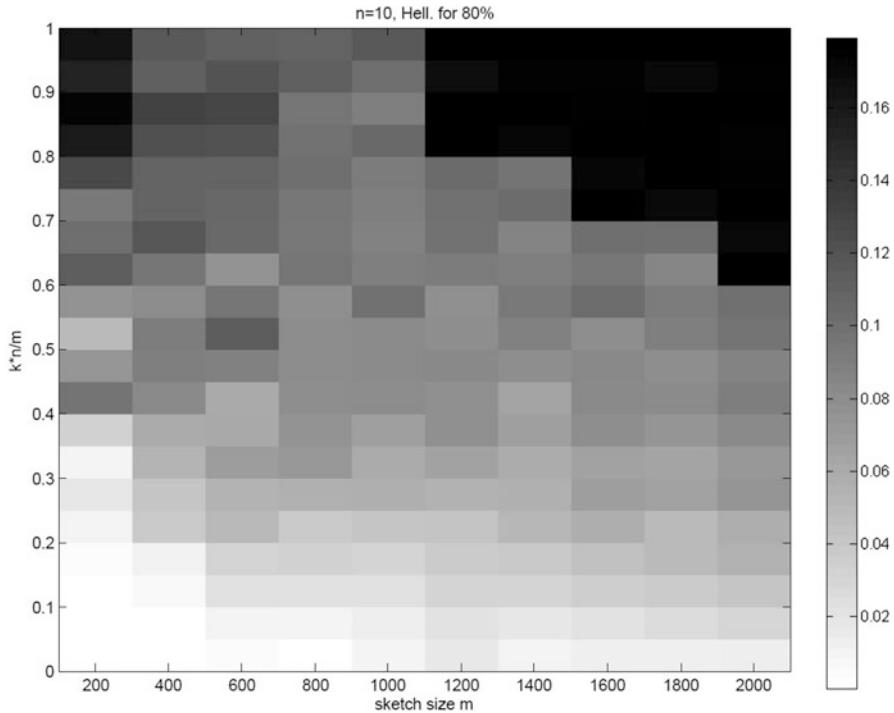


Fig. 8.3 Quality of reconstruction in dimension $n = 10$, with $N = 10^4$ points, measured as a Hellinger distance. Each square corresponds to 10 experiments, and the depicted values are the values of the Hellinger distance under which 80% of performed experiments are placed. The black area at the top right corresponds to values which have not been computed.

Table 8.1 Comparison between our compressed estimation algorithm and an EM algorithm in terms of precision of the estimation and memory usage (in megabytes). Experiments were performed with $n = 20$, $k = 10$, $m = 1000$. In each cell, the value is a median on 10 experiments with the standard deviation for the precision measures.

L	Compressed			L	EM		
	KL div.	Hell.	Mem.		KL div.	Hell.	Mem.
10^3	0.68 ± 0.28	0.06 ± 0.01	0.6	10^3	0.68 ± 0.44	0.07 ± 0.03	0.24
10^4	0.24 ± 0.31	0.02 ± 0.02	0.6	10^4	0.19 ± 0.21	0.01 ± 0.02	2.4
10^5	0.13 ± 0.15	0.01 ± 0.02	0.6	10^5	0.13 ± 0.21	0.01 ± 0.02	24

8.6 Conclusion and outlooks

In this chapter, we first proposed a review of techniques reminiscent of inverse problems and compressed sensing applied to certain learning tasks, mainly to density estimation. Our contribution consisted in a framework for density mixture estimation, which was instantiated to isotropic Gaussians and random Fourier

sampling. We could derive an algorithm which experimentally shows good reconstruction properties with respect to a standard estimation algorithm, while requiring an amount of memory independent of the number of training vectors for the reconstruction. Moreover, the reconstruction method only requires access to the *sketch* computed from the data, thus preserving privacy in the learning step. Different outlooks can be foreseen:

- Density mixture estimation using isotropic Gaussians can be seen as a clustering problem. It would be particularly interesting to extend the experimental results to more general families of Gaussians, for instance with diagonal covariance matrices, to allow for variations in the form of the “clusters.” Considering larger families of densities would probably require finer choices for the sketching operator A , in order to be able to separate the compressed representations of all vectors in the enriched family.
- Even if such a compressive framework has the potential to reduce memory requirements by computing the sketch on the fly with streamed data, the computational complexity of the density reconstruction is still large, especially due to the last stage of the algorithm. Algorithmic savings could be made by finding a faster alternative to this step. The cost of computing the sketch may also be reduced by finding better sketching operators or procedures.
- Another important research direction is the study of the well-posedness of such a reconstruction problem. Producing conclusions require a study of the action of A on Σ_k when the frequencies are chosen at random with a Gaussian distribution such as the distribution chosen in our experiments.
- This work shows that a simple learning problem can be cast in a compressive fashion by only retaining a fixed-size representation of the training data. This representation is not dependent on the number of data vectors. Such methods, if they are found to be applicable to more difficult learning problems, would form a theory of compressive statistical learning, potentially leading to memory and/or computationally cheap algorithms to perform learning tasks.

Acknowledgements This work was supported in part by the European Research Council, PLEASE project (ERC-StG-2011-277906).

References

1. Achlioptas, D.: Database-friendly random projections. In: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 274–281 (2001)
2. Bertin, K., Pennec, E.L., Rivoirard, V.: Adaptive Dantzig density estimation. Annales de l’Institut Henri Poincaré (B) Probabilités et Statistiques **47**(1), 43–74 (2011)
3. Blumensath, T., Davies, M.E.: Iterative hard thresholding for compressed sensing. Appl. Comput. Harmon. Anal. **27**(3), 265–274 (2009)
4. Bourrier, A., Davies, M.E., Peleg, T., Pérez, P., Gribonval, R.: Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. IEEE Trans. Inf. Theory **60**, 7928–7946 (2013)

5. Bourrier, A., Gribonval, R., Pérez, P.: Compressive Gaussian mixture estimation. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (2013)
6. Bunea, F., Tsybakov, A.B., Wegkamp, M., Barbu, A.: Spades and mixture models. *Ann. Stat.* **38**(4), 2525–2558 (2010)
7. Calderbank, R., Schapire, R., Jafarpour, S.: Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Preprint (2009)
8. Candès, E.J., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n . *Ann. Stat.* **35**(6), 2313–2351 (2007)
9. Charikar, M., Chen, K., Farach-Colton, M.: Finding frequent items in data streams. In: ICALP, pp. 693–703 (2002)
10. Cormode, G., Hadjieleftheriou, M.: Methods for finding frequent items in data streams. *VLDB J.* **19**(1), 3–20 (2010)
11. Cormode, G., Muthukrishnan, S.: An improved data stream summary: the count-min sketch and its applications. In: LATIN, pp. 29–38 (2004)
12. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc.* **39**(1), 1–38 (1977)
13. Gilbert, A.C., Strauss, M.J., Tropp, J.A., Vershynin, R.: One sketch for all: fast algorithms for compressed sensing. In: STOC, pp. 237–246 (2007)
14. Gionis, A., Indyk, P., Motwani, R.: Similarity search in high dimensions via hashing. In: VLDB, pp. 518–529 (1999)
15. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: STOC, pp. 604–613 (1998)
16. Johnson, W., Lindenstrauss, J.: Extensions of Lipschitz mappings into a Hilbert space. In: Conference in Modern Analysis and Probability (New Haven, Conn., 1982). Contemporary Mathematics, vol. 26, pp. 189–206. American Mathematical Society, Providence (1984)
17. Thaper, N., Guha, S., Indyk, P., Koudas, N.: Dynamic multidimensional histograms. In: ACM SIGMOD International Conference on Management of Data (2002)

Chapter 9

Two Algorithms for Compressed Sensing of Sparse Tensors

Shmuel Friedland, Qun Li, Dan Schonfeld, and Edgar A. Bernal

Abstract Compressed sensing (CS) exploits the sparsity of a signal in order to integrate acquisition and compression. CS theory enables exact reconstruction of a sparse signal from relatively few linear measurements via a suitable nonlinear minimization process. Conventional CS theory relies on vectorial data representation, which results in good compression ratios at the expense of increased computational complexity. In applications involving color images, video sequences, and multi-sensor networks, the data is intrinsically of high order, and thus more suitably represented in tensorial form. Standard applications of CS to higher-order data typically involve representation of the data as long vectors that are in turn measured using large sampling matrices, thus imposing a huge computational and memory burden. In this chapter, we introduce Generalized Tensor Compressed Sensing (GTCS)—a unified framework for compressed sensing of higher-order tensors which preserves the intrinsic structure of tensorial data with reduced computational complexity at reconstruction. We demonstrate that GTCS offers an efficient means for representation of multidimensional data by providing simultaneous acquisition and compression from all tensor modes. In addition, we propound two reconstruction procedures, a serial method (GTCS-S) and a parallelizable method (GTCS-P), both capable of recovering a tensor based on noiseless or noisy observations. We then compare the performance of the proposed methods with Kronecker compressed sensing (KCS) and multi-way compressed sensing (MWCS). We demonstrate experimentally that GTCS outperforms KCS and MWCS

S. Friedland (✉)

Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7045, USA

e-mail: friedlan@uic.edu

This work was supported by NSF grant DMS-1216393.

Q. Li • E.A. Bernal

PARC, A Xerox Company, 800 Phillips Road, Webster, NY 14580, USA

e-mail: Qun.Li@parc.com; Edgar.Bernal@parc.com

D. Schonfeld

Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL 60607, USA

e-mail: daans@uic.edu

in terms of both reconstruction accuracy (within a range of compression ratios) and processing speed. The major disadvantage of our methods (and of MWCS as well) is that the achieved compression ratios may be worse than those offered by KCS.

9.1 Introduction

Compressed sensing [2, 7] is a framework for reconstructing signals that have sparse representations. A vector $x \in \mathbb{R}^N$ is called *k-sparse* if x has at most k nonzero entries. The sampling scheme can be modelled by a linear operation. Assuming the number of measurements m satisfies $m < N$, and $A \in \mathbb{R}^{m \times N}$ is the matrix used for sampling, then the encoded information is $y \in \mathbb{R}^m$, where $y = Ax$. The decoder knows A and recovers y by finding a solution $\hat{z} \in \mathbb{R}^N$ satisfying

$$\hat{z} = \arg \min_z \|z\|_1 \quad \text{s.t.} \quad y = Az. \quad (9.1)$$

Since $\|\cdot\|$ is a convex function and the set of all z satisfying $y = Az$ is convex, minimizing Eq. (9.1) is polynomial in N . Each *k*-sparse solution can be recovered uniquely if A satisfies the null space property (NSP) of order k , denoted as NSP_k [5]. Given $A \in \mathbb{R}^{m \times N}$ which satisfies the NSP_k property, a *k*-sparse signal $x \in \mathbb{R}^N$ and samples $y = Ax$, recovery of x from y is achieved by finding the z that minimizes Eq. (9.1). One way to generate such A is by sampling its entries using numbers generated from a Gaussian or a Bernoulli distribution. This matrix generation process guarantees that there exists a universal constant c such that if

$$m \geq 2ck \ln \frac{N}{k}, \quad (9.2)$$

then the recovery of x using Eq. (9.1) is successful with probability greater than $1 - \exp(-\frac{m}{2c})$ [14].

The objective of this document is to consider the case where the *k*-sparse vector x is represented as a *k*-sparse tensor $\mathcal{X} = [x_{i_1, i_2, \dots, i_d}] \in \mathbb{R}^{N_1 \times N_2 \times \dots \times N_d}$. Specifically, in the sampling phase, we construct a set of measurement matrices $\{U_1, U_2, \dots, U_d\}$ for all tensor modes, where $U_i \in \mathbb{R}^{m_i \times N_i}$ for $i = 1, 2, \dots, d$, and sample \mathcal{X} to obtain $\mathcal{Y} = \mathcal{X} \times_1 U_1 \times_2 U_2 \times \dots \times_d U_d \in \mathbb{R}^{m_1 \times m_2 \times \dots \times m_d}$ (see Sec. 9.3.1 for a detailed description of tensor mode product notation). Note that our sampling method is mathematically equivalent to that proposed in [10], where A is expressed as a Kronecker product $A := U_1 \otimes U_2 \otimes \dots \otimes U_d$, which requires m to satisfy

$$m \geq 2ck \left(-\ln k + \sum_{i=1}^d \ln N_i \right). \quad (9.3)$$

We show that if each U_i satisfies the NSP_k property, then we can recover \mathcal{X} uniquely from \mathcal{Y} by solving a sequence of ℓ_1 minimization problems, each

similar to the expression in Eq. (9.1). This approach is advantageous relative to vectorization-based compressed sensing methods such as that from [10] because the corresponding recovery problems are in terms of U_i 's instead of A , which results in greatly reduced complexity. If the entries of U_i are sampled from Gaussian or Bernoulli distributions, the following set of conditions needs to be satisfied:

$$m_i \geq 2ck \ln \frac{N_i}{k}, \quad i = 1, \dots, d. \quad (9.4)$$

Observe that the dimensionality of the original signal \mathcal{X} , namely $N = N_1 \cdot \dots \cdot N_d$, is compressed to $m = m_1 \cdot \dots \cdot m_d$. Hence, the number of measurements required by our method must satisfy

$$m \geq (2ck)^d \prod_{i=1}^d \ln \frac{N_i}{k}, \quad (9.5)$$

which indicates a worse compression ratio than that from Eq. (9.3). This is consistent with the observations from [11] (see Fig. 4(a) in [11]). We first discuss our method for matrices, i.e., $d = 2$, and then for tensors, i.e., $d \geq 3$.

9.2 Compressed Sensing of Matrices

9.2.1 Vector and Matrix Notation

Column vectors are denoted by italic letters as $x = (x_1, \dots, x_N)^T \in \mathbb{R}^N$. Norms used for vectors include

$$\|x\|_2 := \sqrt{\sum_{i=1}^N x_i^2}, \quad \|x\|_1 := \sum_{i=1}^N |x_i|.$$

Let $[N]$ denote the set $\{1, 2, \dots, N\}$, where N is a positive integer. Let $S \subset [N]$. We use the following notation: $|S|$ is the cardinality of set S , $S^c := [N] \setminus S$, and $\|x_S\|_1 := \sum_{i \in S} |x_i|$.

Matrices are denoted by capital italic letters as $A = [a_{ij}] \in \mathbb{R}^{m \times N}$. The transposes of x and A are denoted by x^T and A^T , respectively. Norms of matrices used include the Frobenius norm $\|A\|_F := \sqrt{\text{tr}(AA^T)}$, and the spectral norm $\|A\|_2 := \max_{\|x\|_2=1} \|Ax\|_2$. Let $R(X)$ denote the column space of X . The singular value decomposition (SVD) [12] of A with rank $(A) = r$ is:

$$A = \sum_{i=1}^r (\sqrt{\sigma_i} u_i)(\sqrt{\sigma_i} v_i)^T, \quad u_i^T u_j = v_i^T v_j = \delta_{ij}, \quad i, j \in [r]. \quad (9.6)$$

Here, $\sigma_1(A) = \sigma_1 \geq \dots \geq \sigma_r(A) = \sigma_r > 0$ are all positive singular values of A . u_i and v_i are the left and the right singular vectors of A corresponding to σ_i . Recall that

$$Av_i = \sigma_i u_i, \quad A^T v_i = \sigma_i v_i, \quad i \in [r], \quad \|A\|_2 = \sigma_1(A), \quad \|A\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2(A)}.$$

For $k < r$, let

$$A_k := \sum_{i=1}^k (\sqrt{\sigma_i} u_i) (\sqrt{\sigma_i} v_i)^T.$$

For $k \geq r$, we have $A_k := A$. Then A_k is a solution to the following minimization problems:

$$\begin{aligned} \min_{B \in \mathbb{R}^{m \times N}, \text{rank}(B) \leq k} \|A - B\|_F &= \|A - A_k\|_F = \sqrt{\sum_{i=k+1}^r \sigma_i^2(A)}, \\ \min_{B \in \mathbb{R}^{m \times N}, \text{rank}(B) \leq k} \|A - B\|_2 &= \|A - A_k\|_2 = \sigma_{k+1}(A). \end{aligned}$$

We call A_k the best rank- k approximation to A . Note that A_k is unique if and only if $\sigma_j(A) > \sigma_{j+1}(A)$ for $j \in [k-1]$.

$A \in \mathbb{R}^{m \times N}$ satisfies the *null space property of order k* , abbreviated as NSP_k property, if the following condition holds: let $Aw = 0, w \neq 0$; then for each $S \subset [N]$ satisfying $|S| = k$, the inequality $\|w_S\|_1 < \|w_{S^c}\|_1$ is satisfied.

Let $\Sigma_{k,N} \subset \mathbb{R}^N$ denote all vectors in \mathbb{R}^N which have at most k nonzero entries. The fundamental lemma of noiseless recovery in compressed sensing that has been introduced in Chapter 1 is:

Lemma 1. *Suppose that $A \in \mathbb{R}^{m \times N}$ satisfies the NSP_k property. Assume that $x \in \Sigma_{k,N}$ and let $y = Ax$. Then for each $z \in \mathbb{R}^N$ satisfying $Az = y$, $\|z\|_1 \geq \|x\|_1$. Equality holds if and only if $z = x$. That is, $x = \arg \min_z \|z\|_1$ s.t. $y = Az$. The complexity of this minimization problem is $O(N^3)$ [8, 9].*

9.2.2 Noiseless Recovery

9.2.2.1 Compressed Sensing of Matrices - Serial Recovery (CSM-S)

The serial recovery method for compressed sensing of matrices in the noiseless case is described by the following theorem.

Theorem 1 (CSM-S). *Let $X = [x_{ij}] \in \mathbb{R}^{N_1 \times N_2}$ be k -sparse. Let $U_i \in \mathbb{R}^{m_i \times N_i}$ and assume that U_i satisfies the NSP_k property for $i \in [2]$. Define*

$$Y = [y_{pq}] = U_1 X U_2^T \in \mathbb{R}^{m_1 \times m_2}. \quad (9.7)$$

Then X can be recovered uniquely as follows. Let $y_1, \dots, y_{m_2} \in \mathbb{R}^{m_1}$ be the columns of Y . Let $\hat{z}_i \in \mathbb{R}^{N_1}$ be a solution of

$$\hat{z}_i = \arg \min_{z_i} \|z_i\|_1 \quad \text{s.t.} \quad y_i = U_1 z_i, \quad i \in [m_2]. \quad (9.8)$$

Then each \hat{z}_i is unique and k -sparse. Let $Z \in \mathbb{R}^{N_1 \times m_2}$ be the matrix whose columns are $\hat{z}_1, \dots, \hat{z}_{m_2}$. Let $w_1^T, \dots, w_{N_1}^T$ be the rows of Z . Then $v_j \in \mathbb{R}^{N_2}$, whose transpose is the j -th row of X , is the solution of

$$\hat{v}_j = \arg \min_{v_j} \|v_j\|_1 \quad \text{s.t.} \quad w_j = U_2 v_j, \quad j \in [N_1]. \quad (9.9)$$

Proof. Let Z be the matrix whose columns are $\hat{z}_1, \dots, \hat{z}_{m_2}$. Then Z can be written as $Z = XU_2^T \in \mathbb{R}^{N_1 \times m_2}$. Note that \hat{z}_i is a linear combination of the columns of X . \hat{z}_i has at most k nonzero coordinates, because the total number of nonzero elements in X is k . Since $Y = U_1 Z$, it follows that $y_i = U_1 \hat{z}_i$. Also, since U_1 satisfies the NSP_k property, we arrive at Eq. (9.8). Observe that $Z^T = U_2 X^T$; hence, $w_j = U_2 \hat{v}_j$. Since X is k -sparse, then each \hat{v}_j is k -sparse. The assumption that U_2 satisfies the NSP_k property implies Eq. (9.9). \square

If the entries of U_1 and U_2 are drawn from random distributions as described above, then the set of conditions from Eq. (9.4) needs to be met as well. Note that although Theorem 1 requires both U_1 and U_2 to satisfy the NSP_k property, such constraints can be relaxed if each row of X is k' -sparse, where $k' < k$. In this case, it follows from the proof of Theorem 1 that X can be recovered as long as U_1 and U_2 satisfy the NSP_k and the $\text{NSP}_{k'}$ properties, respectively.

9.2.2.2 Compressed Sensing of Matrices - Parallelizable Recovery (CSM-P)

The parallelizable recovery method for compressed sensing of matrices in the noiseless case is described by the following theorem.

Theorem 2 (CSM-P). Let $X = [x_{ij}] \in \mathbb{R}^{N_1 \times N_2}$ be k -sparse. Let $U_i \in \mathbb{R}^{m_i \times N_i}$ and assume that U_i satisfies the NSP_k property for $i \in [2]$. If Y is given by Eq. (9.7), then X can be recovered approximately as follows. Consider a rank decomposition (e.g., SVD) of Y such that

$$Y = \sum_{i=1}^K b_i^{(1)} (b_i^{(2)})^T, \quad (9.10)$$

where $K = \text{rank}(Y)$. Let $\hat{w}_i^{(j)} \in \mathbb{R}^{N_j}$ be a solution of

$$\hat{w}_i^{(j)} = \arg \min_{w_i^{(j)}} \|w_i^{(j)}\|_1 \quad \text{s.t.} \quad b_i^{(j)} = U_j w_i^{(j)}, \quad i \in [K], j \in [2].$$

Then each $\hat{w}_i^{(j)}$ is unique and k -sparse, and

$$X = \sum_{i=1}^K \hat{w}_i^{(1)} (\hat{w}_i^{(2)})^T. \quad (9.11)$$

Proof. First observe that $R(Y) \subset U_1 R(X)$ and $R(Y^T) \subset U_2 R(X^T)$. Since Eq. (9.10) is a rank decomposition of Y , it follows that $b_i^{(1)} \in U_1 R(X)$ and $b_i^{(2)} \in U_2 R(X^T)$. Hence $\hat{w}_i^{(1)} \in R(X)$, $\hat{w}_i^{(2)} \in R(X^T)$ are unique and k -sparse. Let $\hat{X} := \sum_{i=1}^K \hat{w}_i^{(1)} (\hat{w}_i^{(2)})^T$. Assume to the contrary that $X - \hat{X} \neq 0$. Clearly $R(X - \hat{X}) \subset R(X), R(X^T - \hat{X}^T) \subset R(X^T)$. Let $X - \hat{X} = \sum_{i=1}^J u_i^{(1)} (u_i^{(2)})^T$ be a rank decomposition of $X - \hat{X}$. Hence $u_1^{(1)}, \dots, u_J^{(1)} \in R(X)$ and $u_1^{(2)}, \dots, u_J^{(2)} \in R(X^T)$ are two sets of J linearly independent vectors. Since each vector either in $R(X)$ or in $R(X^T)$ is k -sparse, and U_1, U_2 satisfy the NSP_k property, it follows that $U_1 u_1^{(j)}, \dots, U_1 u_J^{(j)}$ are linearly independent for $j \in [2]$ (see Appendix for proof). Hence the matrix $Z := \sum_{i=1}^J (U_1 u_i^{(1)}) (U_2 u_i^{(2)})^T$ has rank J . In particular, $Z \neq 0$. On the other hand, $Z = U_1 (X - \hat{X}) U_2^T = Y - Y = 0$, which contradicts the previous statement. So $X = \hat{X}$. \square

The above recovery procedure consists of two stages, namely, the decomposition stage and the reconstruction stage, where the latter can be implemented in parallel for each matrix mode. Note that the above theorem is equivalent to multi-way compressed sensing for matrices (MWCS) introduced in [15].

9.2.2.3 Simulation Results

We demonstrate experimentally the performance of CSM (denoted by GTCS in general in the following simulation discussions and results) methods on the reconstruction of sparse images and video sequences. As demonstrated in [10], KCS outperforms several other methods including independent measurements and partitioned measurements in terms of reconstruction accuracy in tasks related to compression of multidimensional signals. A more recently proposed method is MWCS, which stands out for its reconstruction efficiency. For the above reasons, we compare our methods with both KCS and MWCS. Our experiments use the ℓ_1 -minimization solvers from [1]. We set the same threshold to determine the termination of the ℓ_1 -minimization process in all subsequent experiments. All simulations are executed on a desktop with a 2.4 GHz Intel Core i7 CPU and 16GB RAM.

The original grayscale image (see Fig. 9.1) is of size 128×128 pixels ($N = 16384$). We use the discrete cosine transform (DCT) as the sparsifying transform, and zero-out the coefficients outside the 16×16 sub-matrix in the upper left corner of the transformed image. We refer to the inverse DCT of the resulting sparse set of transform coefficients as the target image. Let m denote the number of measurements along both matrix modes; we generate the measurement matrices with



Fig. 9.1 The original grayscale image.

entries drawn from a Gaussian distribution with mean 0 and standard deviation $\sqrt{\frac{1}{m}}$. For simplicity, we set the number of measurements for two modes to be equal; that is, the randomly constructed Gaussian matrix U is of size $m \times 128$ for each mode. Therefore, the KCS measurement matrix $U \otimes U$ is of size $m^2 \times 16384$, and the total number of measurements is m^2 . We refer to $\frac{m^2}{N}$ as the normalized number of measurements. For GTCS, both the serial recovery method GTCS-S and the parallelizable recovery method GTCS-P are implemented. In the matrix case, for a given choice of rank decomposition method, GTCS-P and MWCS are equivalent; in this case, we use SVD as the rank decomposition approach. Although the reconstruction stage of GTCS-P is parallelizable, we recover each vector in series. Consequently, we note that the reported performance data for GTCS-P can be improved upon. We examine the performance of the above methods by varying the normalized number of measurements from 0.1 to 0.6 in steps of 0.1. Reconstruction performance for the different methods is compared in terms of reconstruction accuracy and computational complexity. Reconstruction accuracy is measured via the peak signal to noise ratio (PSNR) between the recovered and the target image (both in the spatial domain), whereas computational complexity is measured in terms of the reconstruction time (see Fig. 9.2).

9.2.3 Recovery of Data in the Presence of Noise

Consider the case where the observation is noisy. For a given integer k , a matrix $A \in \mathbb{R}^{m \times N}$ satisfies the restricted isometry property (RIP $_k$) [4] if

$$(1 - \delta_k) \|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k) \|x\|_2^2$$

for all $x \in \Sigma_{k,N}$ and for some $\delta_k \in (0, 1)$.

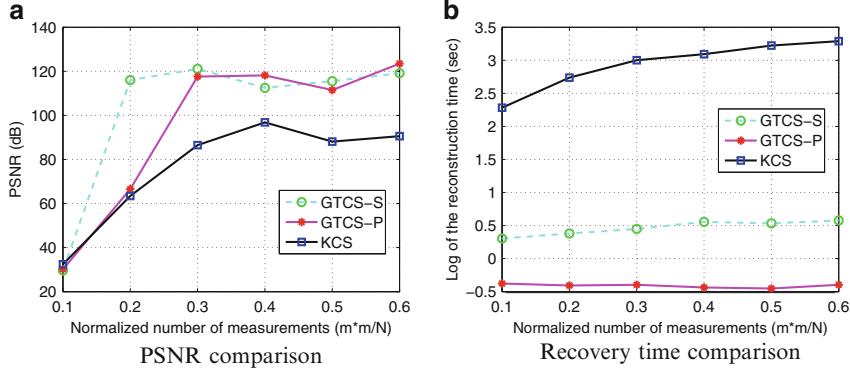


Fig. 9.2 Performance comparison among the tested methods in terms of PSNR and reconstruction time in the scenario of noiseless recovery of a sparse image.

It was shown in [3] that the reconstruction in the presence of noise is achieved by solving

$$\hat{x} = \arg \min_z \|z\|_1, \quad \text{s.t.} \quad \|Az - y\|_2 \leq \varepsilon, \quad (9.12)$$

which has complexity $O(N^3)$.

Lemma 2. Assume that $A \in \mathbb{R}^{m \times N}$ satisfies the RIP_{2k} property for some $\delta_{2k} \in (0, \sqrt{2} - 1)$. Let $x \in \Sigma_{k,N}$, $y = Ax + e$, where e denotes the noise vector, and $\|e\|_2 \leq \varepsilon$ for some real nonnegative number ε . Then

$$\|\hat{x} - x\|_2 \leq C_2 \varepsilon, \quad \text{where } C_2 = \frac{4\sqrt{1 + \delta_{2k}}}{1 - (1 + \sqrt{2})\delta_{2k}}. \quad (9.13)$$

9.2.3.1 Compressed Sensing of Matrices - Serial Recovery (CSM-S) in the Presence of Noise

The serial recovery method for compressed sensing of matrices in the presence of noise is described by the following theorem.

Theorem 3 (CSM-S in the presence of noise). Let $X = [x_{ij}] \in \mathbb{R}^{N_1 \times N_2}$ be k -sparse. Let $U_i \in \mathbb{R}^{m_i \times N_i}$ and assume that U_i satisfies the RIP_{2k} property for some $\delta_{2k} \in (0, \sqrt{2} - 1)$, $i \in [2]$. Define

$$Y = [y_{pq}] = U_1 X U_2^T + E, \quad Y \in \mathbb{R}^{m_1 \times m_2}, \quad (9.14)$$

where E denotes the noise matrix, and $\|E\|_F \leq \varepsilon$ for some real nonnegative number ε . Then X can be recovered approximately as follows. Let $c_1(Y), \dots, c_{m_2}(Y) \in \mathbb{R}^{m_1}$ denote the columns of Y . Let $\hat{z}_i \in \mathbb{R}^{N_1}$ be a solution of

$$\hat{z}_i = \arg \min_{z_i} \|z_i\|_1 \quad \text{s.t.} \quad \|c_i(Y) - U_1 z_i\|_2 \leq \varepsilon, \quad i \in [m_2]. \quad (9.15)$$

Let $Z \in \mathbb{R}^{N_1 \times m_2}$ be the matrix whose columns are $\hat{z}_1, \dots, \hat{z}_{m_2}$. According to Eq. (9.13), $\|c_i(Z) - c_i(XU_2^T)\|_2 = \|\hat{z}_i - c_i(XU_2^T)\|_2 \leq C_2\epsilon$, hence $\|Z - XU_2^T\|_F \leq \sqrt{m_2}C_2\epsilon$. Let $c_1(Z^T), \dots, c_{N_1}(Z^T)$ be the rows of Z . Then $u_j \in \mathbb{R}^{N_2}$, the j -th row of X , is the solution of

$$\hat{u}_j = \arg \min_{u_j} \|u_j\|_1 \quad \text{s.t.} \quad \|c_j(Z^T) - U_2 u_j\|_2 \leq \sqrt{m_2}C_2\epsilon, \quad j \in [N_1]. \quad (9.16)$$

Denote by \hat{X} the recovered matrix, then according to Eq. (9.13),

$$\|\hat{X} - X\|_F \leq \sqrt{m_2 N_1} C_2^2 \epsilon. \quad (9.17)$$

Proof. The proof of the theorem follows from Lemma 2. \square

The upper bound in Eq. (9.17) can be tightened by assuming that the entries of E adhere to a specific type of distribution. Let $E = [e_1, \dots, e_{m_2}]$. Suppose that each entry of E is an independent random variable with a given distribution having zero mean. Then we can assume that $\|e_j\|_2 \leq \frac{\epsilon}{\sqrt{m_2}}$, which implies that $\|E\|_F \leq \epsilon$.

Each z_i can be recovered by finding a solution to

$$\hat{z}_i = \arg \min_{z_i} \|z_i\|_1 \quad \text{s.t.} \quad \|c_i(Y) - U_1 z_i\|_2 \leq \frac{\epsilon}{\sqrt{m_2}}, \quad i \in [m_2]. \quad (9.18)$$

Let $Z = [\hat{z}_1 \dots \hat{z}_{m_2}] \in \mathbb{R}^{N_1 \times m_2}$. According to Eq. (9.13), $\|c_i(Z) - c_i(XU_2^T)\|_2 = \|\hat{z}_i - c_i(XU_2^T)\|_2 \leq C_2 \frac{\epsilon}{\sqrt{m_2}}$; therefore, $\|Z - XU_2^T\|_F \leq C_2\epsilon$.

Let $E_1 := Z - XU_2^T$ be the error matrix, and assume that the entries of E_1 adhere to the same distribution as the entries of E . Hence, $\|c_i(Z^T) - c_i(U_2 X^T)\|_2 \leq \frac{C_2 \epsilon}{\sqrt{N_1}}$.

\hat{X} can be reconstructed by recovering each row of X :

$$\hat{u}_j = \arg \min_{u_j} \|u_j\|_1 \quad \text{s.t.} \quad \|c_j(Z^T) - U_2 u_j\|_2 \leq \frac{C_2 \epsilon}{\sqrt{N_1}}, \quad j \in [N_1]. \quad (9.19)$$

Consequently, $\|\hat{u}_j - c_j(X^T)\|_2 \leq \frac{C_2^2 \epsilon}{\sqrt{N_1}}$, and the recovery error is bounded as follows:

$$\|\hat{X} - X\|_F \leq C_2^2 \epsilon. \quad (9.20)$$

When Y is not full-rank, the above procedure is equivalent to the following alternative. Let Y_k be a best rank- k approximation of Y :

$$Y_k = \sum_{i=1}^k (\sqrt{\tilde{\sigma}_i} \tilde{u}_i) (\sqrt{\tilde{\sigma}_i} \tilde{v}_i)^T. \quad (9.21)$$

Here, $\tilde{\sigma}_i$ is the i -th singular value of Y , and \tilde{u}_i, \tilde{v}_i are the corresponding left and right singular vectors of Y for $i \in [k]$, assume that $k \leq \min(m_1, m_2)$. Since X is assumed to be k -sparse, then $\text{rank}(X) \leq k$. Hence the ranks of XU_2 and $U_1 XU_2^T$ are less than or equal to k . In this case, recovering X amounts to following the procedure described above with Y_k and Z_k taking the place of Y and Z , respectively.

9.2.3.2 Compressed Sensing of Matrices - Parallelizable Recovery (CSM-P) in the Presence of Noise

The parallelizable recovery method for compressed sensing of matrices in the presence of noise is described by the following theorem.

Theorem 4 (CSM-P in the presence of noise). *Let $X = [x_{ij}] \in \mathbb{R}^{N_1 \times N_2}$ be k -sparse. Let $U_i \in \mathbb{R}^{m_i \times N_i}$ and assume that U_i satisfies the RIP_{2k} property for some $\delta_{2k} \in (0, \sqrt{2} - 1)$, $i \in [2]$. Let Y be as defined in Eq. (9.14). Then X can be recovered approximately as follows. Let Y_k be a best rank- k' approximation of Y as in Eq. (9.21), where k' is the minimum of k and the number of singular values of Y greater than $\frac{\varepsilon}{\sqrt{k}}$. Then $\hat{X} = \sum_{i=1}^{k'} \frac{1}{\sigma_i} \hat{x}_i \hat{y}_i^T$ and*

$$\|X - \hat{X}\|_F \leq C_2^2 \varepsilon, \quad (9.22)$$

where

$$\begin{aligned} \hat{x}_i &= \arg \min_{x_i} \|x_i\|_1 \quad \text{s.t.} \quad \|\tilde{\sigma}_i \tilde{u}_i - U_1 x_i\|_2 \leq \frac{\varepsilon}{\sqrt{2k}}, \\ \hat{y}_i &= \arg \min_{y_i} \|y_i\|_1 \quad \text{s.t.} \quad \|\tilde{\sigma}_i \tilde{v}_i - U_2 y_i\|_2 \leq \frac{\varepsilon}{\sqrt{2k}}, \\ &\quad i \in [k]. \end{aligned} \quad (9.23)$$

Proof. Assume that $k < \min(m_1, m_2)$, otherwise $Y_k = Y$. Since $\text{rank}(U_1 X U_2) \leq k$, $Y_k = U_1 X U_2 + E_k$. Let

$$U_1 X U_2^T = \sum_{i=1}^k (\sqrt{\sigma_i} u_i) (\sqrt{\sigma_i} v_i)^T \quad (9.24)$$

be the SVD of $U_1 X U_2^T$. Then $\|u_i\| = \|\tilde{u}_i\| = \|v_i\| = \|\tilde{v}_i\| = 1$ for $i \in [k]$.

Assuming

$$e_i := \sqrt{\tilde{\sigma}_i} \tilde{u}_i - \sqrt{\sigma_i} u_i, \quad f_i := \sqrt{\tilde{\sigma}_i} \tilde{v}_i - \sqrt{\sigma_i} v_i, \quad i \in [k], \quad (9.25)$$

then the entries of e_i and f_i are independent Gaussian variables with zero mean and standard deviation $\frac{\varepsilon}{\sqrt{2\sigma_i m_1 k}}$ and $\frac{\varepsilon}{\sqrt{2\sigma_i m_2 k}}$, respectively, for $i \in [k]$. When $\varepsilon^2 \ll \varepsilon$,

$$E_k \approx \sum_{i=1}^k e_i (\sqrt{\sigma_i} v_i^T) + \sum_{i=1}^k (\sqrt{\sigma_i} u_i) f_i^T. \quad (9.26)$$

In this scenario,

$$\|\sqrt{\sigma_i} u_i - \sqrt{\tilde{\sigma}_i} \tilde{u}_i\| \leq \frac{\varepsilon}{\sqrt{2k\sigma_i}}, \quad \|\sqrt{\sigma_i} v_i - \sqrt{\tilde{\sigma}_i} \tilde{v}_i\| \leq \frac{\varepsilon}{\sqrt{2k\sigma_i}}. \quad (9.27)$$

Note that

$$\sum_{i=1}^{\min(m_1, m_2)} (\sigma_i - \sigma(Y_k))^2 \leq \text{tr}(EE^T) \leq \varepsilon^2, \quad \sum_{i=1}^k (\sigma_i - \tilde{\sigma}_i)^2 \leq \text{tr}(E_k E_k^T) \leq \varepsilon^2. \quad (9.28)$$

Given the way k' is defined, it can be interpreted as the numerical rank of Y . Consequently, Y can be well represented by its best rank k' approximation. Thus

$$U_1 X U_2^T \approx \sum_{i=1}^{k'} (\sqrt{\sigma_i} u_i) (\sqrt{\sigma_i} v_i^T), \quad Y_{k'} = \sum_{i=1}^{k'} (\sqrt{\tilde{\sigma}_i} \tilde{u}_i) (\sqrt{\tilde{\sigma}_i} \tilde{v}_i^T), \quad i \in [k']. \quad (9.29)$$

Assuming $\sigma_i \approx \tilde{\sigma}_i$ for $i \in [k']$, we conclude that

$$\|\tilde{\sigma}_i \tilde{u}_i - \sigma_i u_i\| \leq \frac{\varepsilon}{\sqrt{2k}}, \quad \|\tilde{\sigma}_i \tilde{v}_i - \sigma_i v_i\| \leq \frac{\varepsilon}{\sqrt{2k}}. \quad (9.30)$$

A compressed sensing framework can be used to solve the following set of minimization problems, for $i \in [k']$:

$$\hat{x}_i = \arg \min_{x_i} \|x_i\|_1 \quad \text{s.t.} \quad \|\tilde{\sigma}_i \tilde{u}_i - U_1 x_i\|_2 \leq \frac{\varepsilon}{\sqrt{2k}}, \quad (9.31)$$

$$\hat{y}_i = \arg \min_{y_i} \|y_i\|_1 \quad \text{s.t.} \quad \|\tilde{\sigma}_i \tilde{v}_i - U_2 y_i\|_2 \leq \frac{\varepsilon}{\sqrt{2k}}. \quad (9.32)$$

The error bound from Eq. (9.22) follows. \square

9.2.3.3 Simulation Results

In this section, we use the same target image and experimental settings used in Section 9.2.2.3. We simulate the noisy recovery scenario by modifying the observation with additive, zero-mean Gaussian noise having standard deviation values ranging from 1 to 10 in steps of 1, and attempt to recover the target image via solving a set of minimization problems as in Eq. (9.12). As before, reconstruction performance is measured in terms of PSNR between the recovered and the target image, and in terms of reconstruction time, as illustrated in Figs. 9.3 and 9.4.

9.3 Compressed Sensing of Tensors

9.3.1 A Brief Introduction to Tensors

A tensor is a multidimensional array. The order of a tensor is the number of modes. For instance, tensor $\mathcal{X} \in \mathbb{R}^{N_1 \times \dots \times N_d}$ has order d and the dimension of its i -th mode (denoted mode i) is N_i .

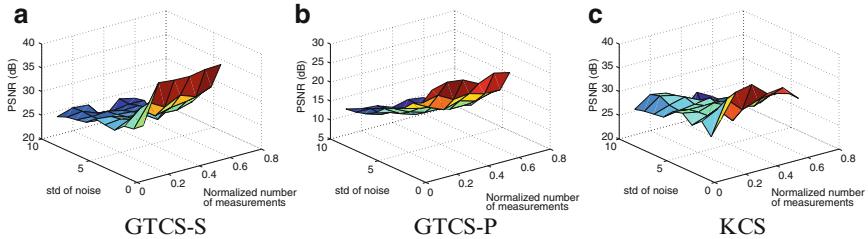


Fig. 9.3 PSNR between target and recovered image for the tested methods in the noisy recovery scenario.

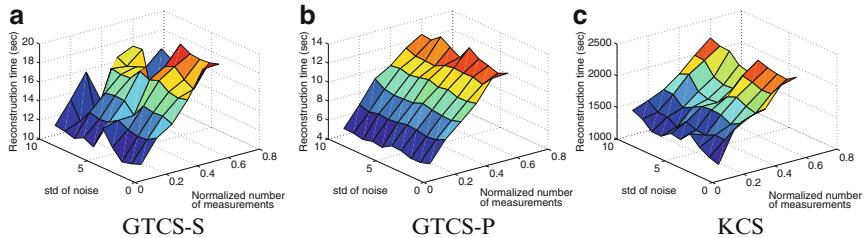


Fig. 9.4 Execution time for the tested methods in the noisy recovery scenario.

Definition 1 (Kronecker Product). The Kronecker product between matrices $A \in \mathbb{R}^{I \times J}$ and $B \in \mathbb{R}^{K \times L}$ is denoted by $A \otimes B$. The result is the matrix of dimensions $(I \cdot K) \times (J \cdot L)$ defined by

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1J}B \\ a_{21}B & a_{22}B & \cdots & a_{2J}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}B & a_{I2}B & \cdots & a_{IJ}B \end{pmatrix}.$$

Definition 2 (Outer Product and Tensor Product). The operator \circ denotes the tensor product between two vectors. In linear algebra, the outer product typically refers to the tensor product between two vectors, that is, $u \circ v = uv^T$. In this chapter, the terms outer product and tensor product are equivalent. The Kronecker product and the tensor product between two vectors are related by $u \circ v = u \otimes v^T$.

Definition 3 (Mode- i Product). The mode- i product of a tensor $\mathcal{X} = [x_{\alpha_1, \dots, \alpha_d}] \in \mathbb{R}^{N_1 \times \dots \times N_d}$ and a matrix $U = [u_{j, \alpha_i}] \in \mathbb{R}^{J \times N_i}$ is denoted by $\mathcal{X} \times_i U$ and is of size $N_1 \times \dots \times N_{i-1} \times J \times N_{i+1} \times \dots \times N_d$. Element-wise, the mode- i product can be written as $(\mathcal{X} \times_i U)_{\alpha_1, \dots, \alpha_{i-1}, j, \alpha_{i+1}, \dots, \alpha_d} = \sum_{\alpha_i=1}^{N_i} x_{\alpha_1, \dots, \alpha_d} u_{j, \alpha_i}$.

Definition 4 (Mode- i Fiber and Mode- i Unfolding). The mode- i fiber of tensor $\mathcal{X} = [x_{\alpha_1, \dots, \alpha_d}] \in \mathbb{R}^{N_1 \times \dots \times N_d}$ is the set of vectors obtained by fixing every index

but α_i . The mode- i unfolding $X_{(i)}$ of \mathcal{X} is the $N_i \times (N_1 \cdot \dots \cdot N_{i-1} \cdot N_{i+1} \cdot \dots \cdot N_d)$ matrix whose columns are the mode- i fibers of \mathcal{X} . $\mathcal{Y} = \mathcal{X} \times_1 U_1 \times \dots \times_d U_d$ is equivalent to $Y_{(i)} = U_i X_{(i)} (U_d \otimes \dots \otimes U_{i+1} \otimes U_{i-1} \otimes \dots \otimes U_1)^T$.

Definition 5 (Core Tucker Decomposition [16]). Let $\mathcal{X} \in \mathbb{R}^{N_1 \times \dots \times N_d}$ be a tensor with mode- i unfolding $X_{(i)} \in \mathbb{R}^{N_i \times (N_1 \cdot \dots \cdot N_{i-1} \cdot N_{i+1} \cdot \dots \cdot N_d)}$ such that $\text{rank}(X_{(i)}) = r_i$. Let $R_i(\mathcal{X}) \subset \mathbb{R}^{N_i}$ denote the column space of $X_{(i)}$, and $c_{1,i}, \dots, c_{r_i,i}$ be a basis in $R_i(\mathcal{X})$. Then \mathcal{X} is an element of the subspace $\mathbf{V}(\mathcal{X}) := R_1(\mathcal{X}) \circ \dots \circ R_d(\mathcal{X}) \subset \mathbb{R}^{N_1 \times \dots \times N_d}$. Clearly, vectors $c_{i_1,1} \circ \dots \circ c_{i_d,d}$, where $i_j \in [r_j]$ and $j \in [d]$, form a basis of \mathbf{V} . The core Tucker decomposition of \mathcal{X} is

$$\mathcal{X} = \sum_{i_j \in [r_j], j \in [d]} \xi_{i_1, \dots, i_d} c_{i_1,1} \circ \dots \circ c_{i_d,d} \quad (9.33)$$

for some decomposition coefficients ξ_{i_1, \dots, i_d} , $i_j \in [r_j]$ and $j \in [d]$.

A special case of the core Tucker decomposition is the higher-order singular value decomposition (HOSVD). Any tensor $\mathcal{X} \in \mathbb{R}^{N_1 \times \dots \times N_d}$ can be written as

$$\mathcal{X} = \mathcal{S} \times_1 U_1 \times \dots \times_d U_d, \quad (9.34)$$

where $U_i = [u_1, \dots, u_{N_i}]$ is an orthonormal matrix for $i \in [d]$, and $\mathcal{S} = \mathcal{X} \times_1 U_1^T \times \dots \times_d U_d^T$ is called the core tensor. For a more in-depth discussion on HOSVD, including the set of properties the core tensor is required to satisfy, please refer to [6].

\mathcal{X} can also be expressed in terms of weaker decompositions of the form

$$\mathcal{X} = \sum_{i=1}^K a_i^{(1)} \circ \dots \circ a_i^{(d)}, \quad a_i^{(j)} \in R_j(\mathcal{X}), j \in [d]. \quad (9.35)$$

For instance, first decompose $X_{(1)}$ as $X_{(1)} = \sum_{j=1}^{r_1} c_{j,1} g_{j,1}^T$ (e.g., via SVD); then each $g_{j,1}$ can be viewed as a tensor of order $d-1 \in R_2(\mathcal{X}) \circ \dots \circ R_d(\mathcal{X}) \subset \mathbb{R}^{N_2 \times \dots \times N_d}$. Secondly, unfold each $g_{j,1}$ in mode 2 to obtain $g_{j,1(2)}$ and decompose $g_{j,1(2)} = \sum_{l=1}^{r_2} d_{l,2,j} f_{l,2,j}^T$, $d_{l,2,j} \in R_2(\mathcal{X})$, $f_{l,2,j} \in R_3(\mathcal{X}) \circ \dots \circ R_d(\mathcal{X})$. We simply say a vector belongs to a tensor product of several vector spaces when we mean its corresponding tensorial representation belongs to that space. By successively unfolding and decomposing each remaining tensor mode, a decomposition of the form in Eq. (9.35) is obtained. Note that if \mathcal{X} is k -sparse, then each vector in $R_i(\mathcal{X})$ is k -sparse and $r_i \leq k$ for $i \in [d]$. Hence, $K \leq k^{d-1}$.

Definition 6 (CANDECOMP/PARAFAC Decomposition [13]). For a tensor $\mathcal{X} \in \mathbb{R}^{N_1 \times \dots \times N_d}$, the CANDECOMP/PARAFAC (CP) decomposition is defined as $\mathcal{X} \approx [\lambda; A^{(1)}, \dots, A^{(d)}] \equiv \sum_{r=1}^R \lambda_r a_r^{(1)} \circ \dots \circ a_r^{(d)}$, where $\lambda = [\lambda_1, \dots, \lambda_R]^T \in \mathbb{R}^R$ and $A^{(i)} = [a_1^{(i)}, \dots, a_R^{(i)}] \in \mathbb{R}^{N_i \times R}$ for $i \in [d]$.

9.3.2 Noiseless Recovery

9.3.2.1 Generalized Tensor Compressed Sensing—Serial Recovery (GTCS-S)

The serial recovery method for compressed sensing of tensors in the noiseless case is described by the following theorem.

Theorem 5. *Let $\mathcal{X} = [x_{i_1, \dots, i_d}] \in \mathbb{R}^{N_1 \times \dots \times N_d}$ be k -sparse. Let $U_i \in \mathbb{R}^{m_i \times N_i}$ and assume that U_i satisfies the NSP_k property for $i \in [d]$. Define*

$$\mathcal{Y} = [y_{j_1, \dots, j_d}] = \mathcal{X} \times_1 U_1 \times \dots \times_d U_d \in \mathbb{R}^{m_1 \times \dots \times m_d}. \quad (9.36)$$

Then \mathcal{X} can be recovered uniquely as follows. Unfold \mathcal{Y} in mode 1,

$$Y_{(1)} = U_1 X_{(1)} [\otimes_{k=2}^d U_k]^T \in \mathbb{R}^{m_1 \times (m_2 \dots m_d)}.$$

Let $y_1, \dots, y_{m_2 \dots m_d}$ be the columns of $Y_{(1)}$. Then $y_i = U_1 z_i$, where each $z_i \in \mathbb{R}^{N_1}$ is k -sparse. Recover each z_i using Eq. (9.1). Let $\mathcal{Z} = \mathcal{X} \times_2 U_2 \times \dots \times_d U_d \in \mathbb{R}^{N_1 \times m_2 \times \dots \times m_d}$, and let $z_1, \dots, z_{m_2 \dots m_d}$ denote its mode-1 fibers. Unfold \mathcal{Z} in mode 2,

$$Z_{(2)} = U_2 X_{(2)} [\otimes_{k=3}^d U_k \otimes I]^T \in \mathbb{R}^{m_2 \times (N_1 \cdot m_3 \dots m_d)}.$$

Let $w_1, \dots, w_{N_1 \cdot m_3 \dots m_d}$ be the columns of $Z_{(2)}$. Then $w_j = U_2 v_j$, where each $v_j \in \mathbb{R}^{N_2}$ is k -sparse. Recover each v_j using Eq. (9.1). \mathcal{X} can be reconstructed by successively applying the above procedure to tensor modes $3, \dots, d$.

Proof. The proof of this theorem is a straightforward generalization of that of Theorem 1. \square

Note that although Theorem 5 requires U_i to satisfy the NSP_k property for $i \in [d]$, such constraints can be relaxed if each mode- i fiber of $\mathcal{X} \times_{i+1} U_{i+1} \times \dots \times_d U_d$ is k_i -sparse for $i \in [d-1]$, and each mode- d fiber of \mathcal{X} is k_d -sparse, where $k_i \leq k$, for $i \in [d]$. In this case, it follows from the proof of Theorem 5 that X can be recovered as long as U_i satisfies the NSP_{k_i} property, for $i \in [d]$.

9.3.2.2 Generalized Tensor Compressed Sensing—Parallelizable Recovery (GTCS-P)

The parallelizable recovery method for compressed sensing of tensors in the noiseless case is described by the following theorem.

Theorem 6 (GTCS-P). *Let $\mathcal{X} = [x_{i_1, \dots, i_d}] \in \mathbb{R}^{N_1 \times \dots \times N_d}$ be k -sparse. Let $U_i \in \mathbb{R}^{m_i \times N_i}$ and assume that U_i satisfies the NSP_k property for $i \in [d]$. If \mathcal{Y} is given by Eq. (9.36), then \mathcal{X} can be recovered uniquely as follows. Consider a decomposition of \mathcal{Y} such that*

$$\mathcal{Y} = \sum_{i=1}^K b_i^{(1)} \circ \dots \circ b_i^{(d)}, \quad b_i^{(j)} \in R_j(\mathcal{Y}) \subseteq U_j R_j(\mathcal{X}), j \in [d]. \quad (9.37)$$

Let $\hat{w}_i^{(j)} \in R_j(\mathcal{X}) \subset \mathbb{R}^{N_j}$ be a solution of

$$\hat{w}_i^{(j)} = \arg \min_{w_i^{(j)}} \|w_i^{(j)}\|_1 \quad \text{s.t.} \quad b_i^{(j)} = U_j w_i^{(j)}, \quad i \in [K], j \in [d]. \quad (9.38)$$

Thus each $\hat{w}_i^{(j)}$ is unique and k -sparse. Then,

$$\mathcal{X} = \sum_{i=1}^K w_i^{(1)} \circ \dots \circ w_i^{(d)}, \quad w_i^{(j)} \in R_j(\mathcal{X}), j \in [d]. \quad (9.39)$$

Proof. Since \mathcal{X} is k -sparse, each vector in $R_j(\mathcal{X})$ is k -sparse. If each U_j satisfies the NSP_k property, then $w_i^{(j)} \in R_j(\mathcal{X})$ is unique and k -sparse. Define \mathcal{Z} as

$$\mathcal{Z} = \sum_{i=1}^K w_i^{(1)} \circ \dots \circ w_i^{(d)}, \quad w_i^{(j)} \in R_j(\mathcal{X}), j \in [d]. \quad (9.40)$$

Then

$$(\mathcal{X} - \mathcal{Z}) \times_1 U_1 \times \dots \times_d U_d = 0. \quad (9.41)$$

To show $\mathcal{Z} = \mathcal{X}$, assume a slightly more general scenario, where each $R_j(\mathcal{X}) \subseteq \mathbf{V}_j \subset \mathbb{R}^{N_j}$, such that each nonzero vector in \mathbf{V}_j is k -sparse. Then $R_j(\mathcal{Y}) \subseteq U_j R_j(\mathcal{X}) \subseteq U_j \mathbf{V}_j$ for $j \in [d]$. Assume to the contrary that $\mathcal{X} \neq \mathcal{Z}$. This hypothesis can be disproven via induction on mode m as follows.

Suppose

$$(\mathcal{X} - \mathcal{Z}) \times_m U_m \times \dots \times_d U_d = 0. \quad (9.42)$$

Unfold \mathcal{X} and \mathcal{Z} in mode m , then the column (row) spaces of $X_{(m)}$ and $Z_{(m)}$ are contained in \mathbf{V}_m ($\hat{\mathbf{V}}_m := \mathbf{V}_1 \circ \dots \circ \mathbf{V}_{m-1} \circ \mathbf{V}_{m+1} \circ \dots \circ \mathbf{V}_d$). Since $\mathcal{X} \neq \mathcal{Z}$, $X_{(m)} - Z_{(m)} \neq 0$. Then $X_{(m)} - Z_{(m)} = \sum_{i=1}^p u_i v_i^T$, where $\text{rank}(X_{(m)} - Z_{(m)}) = p$, and $u_1, \dots, u_p \in \mathbf{V}_m, v_1, \dots, v_p \in \hat{\mathbf{V}}_m$ are two sets of linearly independent vectors.

Since $(\mathcal{X} - \mathcal{Z}) \times_m U_m \times \dots \times_d U_d = 0$,

$$\begin{aligned} 0 &= U_m (X_{(m)} - Z_{(m)}) (U_d \otimes \dots \otimes U_{m+1} \otimes I)^T \\ &= U_m (X_{(m)} - Z_{(m)}) \hat{U}_m^T \\ &= \sum_{i=1}^p (U_m u_i) (\hat{U}_m v_i)^T. \end{aligned}$$

Since $U_m u_1, \dots, U_m u_p$ are linearly independent (see Appendix for proof), it follows that $\hat{U}_m v_i = 0$ for $i \in [p]$. Therefore,

$$(X_{(m)} - Z_{(m)}) \hat{U}_m^T = \left(\sum_{i=1}^p u_i v_i^T \right) \hat{U}_m^T = \sum_{i=1}^p u_i (\hat{U}_m v_i)^T = 0,$$

which is equivalent to (in tensor form, after folding)

$$\begin{aligned} (\mathcal{X} - \mathcal{Z}) &\times_m I_m \times_{m+1} U_{m+1} \times \dots \times_d U_d \\ &= (\mathcal{X} - \mathcal{Z}) \times_{m+1} U_{m+1} \times \dots \times_d U_d = 0, \end{aligned} \quad (9.43)$$

where I_m is the $N_m \times N_m$ identity matrix. Note that Eq. (9.42) leads to Eq. (9.43) upon replacing U_m with I_m . Similarly, when $m = 1$, U_1 can be replaced with I_1 in Eq. (9.41). By successively replacing U_m with I_m for $2 \leq m \leq d$,

$$\begin{aligned} &(\mathcal{X} - \mathcal{Z}) \times_1 U_1 \times \dots \times_d U_d \\ &= (\mathcal{X} - \mathcal{Z}) \times_1 I_1 \times \dots \times_d I_d \\ &= \mathcal{X} - \mathcal{Z} = 0, \end{aligned}$$

which contradicts the assumption that $\mathcal{X} \neq \mathcal{Z}$. Thus, $\mathcal{X} = \mathcal{Z}$. This completes the proof. \square

Note that although Theorem 6 requires U_i to satisfy the NSP_k property for $i \in [d]$, such constraints can be relaxed if all vectors $\in R_i(\mathcal{X})$ are k_i -sparse. In this case, it follows from the proof of Theorem 6 that X can be recovered as long as U_i satisfies the NSP_{k_i} for $i \in [d]$.

As in the matrix case, the reconstruction stage of the recovery process can be implemented in parallel for each tensor mode.

Note additionally that Theorem 6 does not require tensor rank decomposition, which is an NP-hard problem. Weaker decompositions such as the one described by Eq. 9.35 can be utilized.

The above described procedure allows exact recovery. In some cases, recovery of a rank- R approximation of \mathcal{X} , $\hat{\mathcal{X}} = \sum_{r=1}^R w_r^{(1)} \circ \dots \circ w_r^{(d)}$, suffices. In such scenarios, \mathcal{Y} in Eq. (9.37) can be replaced by its rank- R approximation, namely, $\hat{\mathcal{Y}} = \sum_{r=1}^R b_r^{(1)} \circ \dots \circ b_r^{(d)}$ (obtained, e.g., by CP decomposition).

9.3.2.3 Simulation Results

Examples of data that is amenable to tensorial representation include color and multi-spectral images and video. We use a 24-frame, 24×24 pixel grayscale video to test the performance of our algorithm (see Fig. 9.5). In other words, the video data is represented as a $24 \times 24 \times 24$ tensor ($N = 13824$). We use the three-dimensional



Fig. 9.5 The original 24 video frames.

DCT as the sparsifying transform, and zero-out coefficients outside the $6 \times 6 \times 6$ cube located on the front upper left corner of the transformed tensor. As in the image case, let m denote the number of measurements along each tensor mode; we generate the measurement matrices with entries drawn from a Gaussian distribution with mean 0 and standard deviation $\sqrt{\frac{1}{m}}$. For simplicity, we set the number of measurements for each tensor mode to be equal; that is, the randomly constructed Gaussian matrix U is of size $m \times 24$ for each mode. Therefore, the KCS measurement matrix $U \otimes U \otimes U$ is of size $m^3 \times 13824$, and the total number of measurements is m^3 . We refer to $\frac{m^3}{N}$ as the normalized number of measurements. For GTCS-P, we employ the weaker form of the core Tucker decomposition as described in Section 9.3.1. Although the reconstruction stage of GTCS-P is parallelizable, we recover each vector in series. We examine the performance of KCS and GTCS-P by varying the normalized number of measurements from 0.1 to 0.6 in steps of 0.1. Reconstruction accuracy is measured in terms of the average PSNR across all frames between the recovered and the target video, whereas computational complexity is measured in terms of the log of the reconstruction time (see Fig. 9.6).

Note that in the tensor case, due to the serial nature of GTCS-S, the reconstruction error propagates through the different stages of the recovery process. Since exact reconstruction is rarely achieved in practice, the equality constraint in the ℓ_1 -minimization process described by Eq. (9.1) becomes increasingly difficult to satisfy for the later stages of the reconstruction process. In this case, a relaxed recovery procedure as described in Eq. (9.12) can be employed. Since the relaxed constraint from Eq. (9.12) results in what effectively amounts to recovery in the presence of noise, we do not compare the performance of GTCS-S with that of the other two methods.

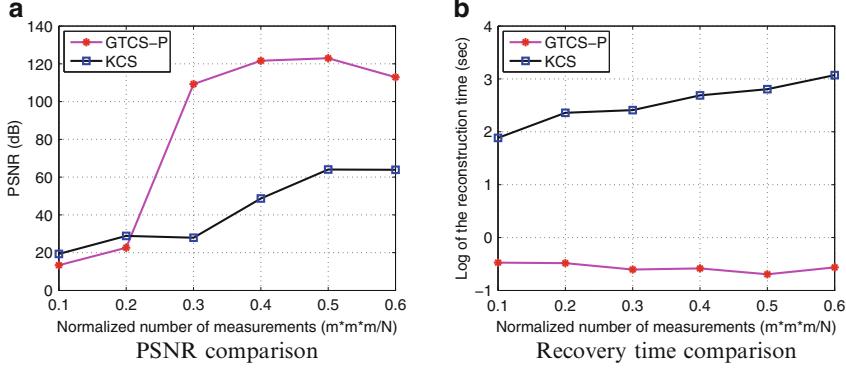


Fig. 9.6 Performance comparison among the tested methods in terms of PSNR and reconstruction time in the scenario of noiseless recovery of the sparse video.

9.3.3 Recovery in the Presence of Noise

9.3.3.1 Generalized Tensor Compressed Sensing—Serial Recovery (GTCS-S) in the Presence of Noise

Let $\mathcal{X} = [x_{i_1, \dots, i_d}] \in \mathbb{R}^{N_1 \times \dots \times N_d}$ be k -sparse. Let $U_i \in \mathbb{R}^{m_i \times N_i}$ and assume that U_i satisfies the NSP _{k} property for $i \in [d]$. Define

$$\mathcal{Y} = [y_{j_1, \dots, j_d}] = \mathcal{X} \times_1 U_1 \times \dots \times_d U_d + \mathcal{E} \in \mathbb{R}^{m_1 \times \dots \times m_d}, \quad (9.44)$$

where \mathcal{E} is the noise tensor and $\|\mathcal{E}\|_F \leq \varepsilon$ for some real nonnegative number ε . Although the norm of the noise tensor is not equal across different stages of GTCS-S, it is assumed that at any given stage, the entries of the error tensor are independent and identically distributed. The upper bound of the reconstruction error for GTCS-S recovery in the presence of noise is derived next by induction on mode k .

When $k = 1$, unfold \mathcal{Y} in mode 1 to obtain matrix $Y_{(1)} \in \mathbb{R}^{m_1 \times (m_2 \cdot \dots \cdot m_d)}$. Recover each $z_i^{(1)}$ by

$$\hat{z}_i^{(1)} = \arg \min_{z_i^{(1)}} \|z_i^{(1)}\|_1 \quad \text{s.t.} \quad \|c_i(Y_{(1)}) - U_1 z_i^{(1)}\|_2 \leq \frac{\varepsilon}{\sqrt{m_2 \cdot \dots \cdot m_d}}. \quad (9.45)$$

Let $\hat{Z}^{(1)} = [\hat{z}_1^{(1)} \dots \hat{z}_{m_2 \cdot \dots \cdot m_d}^{(1)}] \in \mathbb{R}^{N_1 \times (m_2 \cdot \dots \cdot m_d)}$. According to Eq. (9.13), $\|\hat{z}_i^{(1)} - c_i(X_{(1)}[\otimes_{k=d}^2 U_k]^T)\|_2 \leq C_2 \frac{\varepsilon}{\sqrt{m_2 \cdot \dots \cdot m_d}}$, and $\|\hat{Z}^{(1)} - X_{(1)}[\otimes_{k=d}^2 U_k]^T\|_F \leq C_2 \varepsilon$. In tensor form, after folding, this is equivalent to $\|\hat{\mathcal{Z}}^{(1)} - \mathcal{X} \times_2 U_2 \times \dots \times_d U_d\|_F \leq C_2 \varepsilon$.

Assume when $k = n$, $\|\hat{\mathcal{Z}}^{(n)} - \mathcal{X} \times_{n+1} U_{n+1} \times \dots \times_d U_d\|_F \leq C_2^n \varepsilon$ holds. For $k = n+1$, unfold $\hat{\mathcal{Z}}^{(n)}$ in mode $n+1$ to obtain $\hat{Z}_{(n+1)}^{(n)} \in \mathbb{R}^{m_{n+1} \times (N_1 \cdot \dots \cdot N_n \cdot m_{n+2} \cdot \dots \cdot m_d)}$, and recover each $z_i^{(n+1)}$ by

$$\begin{aligned} \hat{z}_i^{(n+1)} &= \arg \min_{z_i^{(n+1)}} \|z_i^{(n+1)}\|_1 \quad \text{s.t.} \\ \|c_i(\hat{Z}_{(n+1)}^{(n)}) - U_{n+1} z_i^{(n+1)}\|_2 &\leq C_2 \frac{\epsilon}{\sqrt{N_1 \cdot \dots \cdot N_n \cdot m_{n+2} \cdot \dots \cdot m_d}}. \end{aligned} \quad (9.46)$$

Let $\hat{Z}^{(n+1)} = [\hat{z}_1^{(n+1)} \dots \hat{z}_{N_1 \cdot \dots \cdot N_n \cdot m_{n+2} \cdot \dots \cdot m_d}^{(n+1)}] \in \mathbb{R}^{N_{n+1} \times (N_1 \cdot \dots \cdot N_n \cdot m_{n+2} \cdot \dots \cdot m_d)}$. Then $\|\hat{z}_i^{(n+1)} - c_i(X_{(n+1)} [\otimes_{k=d}^{n+2} U_k]^T)\|_2 \leq C_2^{n+1} \frac{\epsilon}{\sqrt{N_1 \cdot \dots \cdot N_n \cdot m_{n+2} \cdot \dots \cdot m_d}}$, and $\|\hat{Z}^{(n+1)} - X_{(n+1)} [\otimes_{k=d}^{n+2} U_k]^T\|_F \leq C_2^{n+1} \epsilon$. Folding back to tensor form, $\|\hat{\mathcal{X}}^{(n+1)} - \mathcal{X} \times_{n+2} U_{n+2} \times \dots \times_d U_d\|_F \leq C_2^{n+1} \epsilon$.

When $k = d$, $\|\hat{\mathcal{X}}^{(d)} - \mathcal{X}\|_F \leq C_2^d \epsilon$ by induction on mode k .

9.3.3.2 Generalized Tensor Compressed Sensing—Parallelizable Recovery (GTCS-P) in the Presence of Noise

Let $\mathcal{X} = [x_{i_1, \dots, i_d}] \in \mathbb{R}^{N_1 \times \dots \times N_d}$ be k -sparse. Let $U_i \in \mathbb{R}^{m_i \times N_i}$ and assume that U_i satisfies the NSP _{k} property for $i \in [d]$. Let \mathcal{Y} be defined as in Eq. (9.44). GTCS-P recovery in the presence of noise operates as in the noiseless recovery case described in Section 9.3.2.2, except that $\hat{w}_i^{(j)}$ is recovered via

$$\hat{w}_i^{(j)} = \arg \min_{w_i^{(j)}} \|w_i^{(j)}\|_1 \quad \text{s.t.} \quad \|U_j w_i^{(j)} - b_i^{(j)}\|_2 \leq \frac{\epsilon}{2k}, \quad i \in [K], j \in [d]. \quad (9.47)$$

It follows from the proof of Theorem 4 that the recovery error of GTCS-P in the presence of noise between the original tensor \mathcal{X} and the recovered tensor $\hat{\mathcal{X}}$ is bounded as follows:

$$\|\hat{\mathcal{X}} - \mathcal{X}\|_F \leq C_2^d \epsilon.$$

9.3.3.3 Simulation Results

In this section, we use the same target video and experimental settings used in Section 9.3.2.3. We simulate the noisy recovery scenario by modifying the observation tensor with additive, zero-mean Gaussian noise having standard deviation values ranging from 1 to 10 in steps of 1, and attempt to recover the target video via solving a set of minimization problems as in Eq. (9.12). As before, reconstruction performance is measured in terms of the average PSNR across all frames between the recovered and the target video, and in terms of log of reconstruction time, as illustrated in Figs. 9.7 and 9.8. Note that the illustrated results correspond to the performance of the methods for a given choice of upper bound on the l_2 norm in Eq. (9.12); the PSNR numbers can be further improved by tightening this bound.

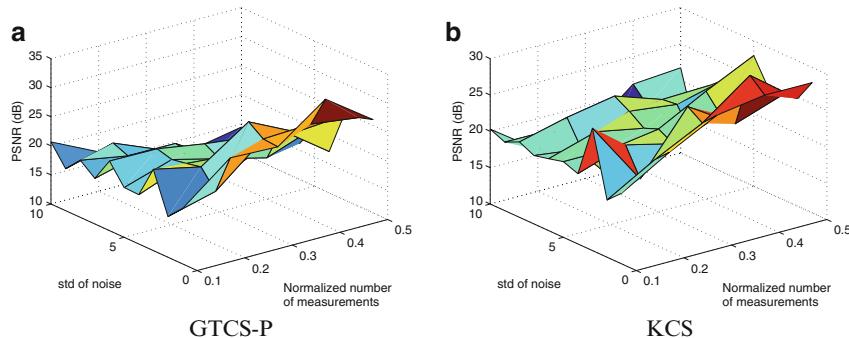


Fig. 9.7 PSNR for the tested methods in the scenario of recovering the sparse video in the presence of noise.

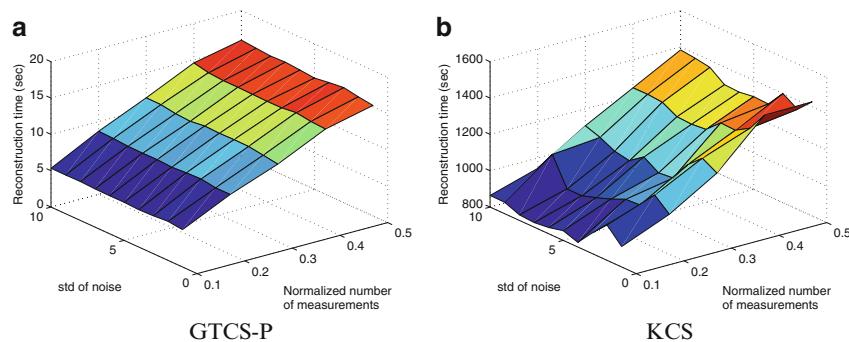


Fig. 9.8 Execution time for the tested methods in the scenario of recovering the sparse video in the presence of noise.

9.3.4 Tensor Compressibility

Let $\mathcal{X} = [x_{i_1, \dots, i_d}] \in \mathbb{R}^{N_1 \times \dots \times N_d}$. Assume the entries of the measurement matrix are drawn from a Gaussian or Bernoulli distribution as described above. For a given level of reconstruction accuracy, the number of measurements for \mathcal{X} required by GTCS should satisfy

$$m \geq 2^d c^d \prod_{i \in [d]} \ln \frac{N_i}{k}. \quad (9.48)$$

Suppose that $N_1 = \dots = N_d = N^{\frac{1}{d}}$. Then

$$m \geq 2^d c^d \left(\ln \frac{N^{\frac{1}{d}}}{k} \right)^d = 2^d c^d \left(\frac{1}{d} \ln N - \ln k \right)^d. \quad (9.49)$$

On the other hand, the number of measurements required by KCS should satisfy

$$m \geq 2c \ln \frac{N}{k}. \quad (9.50)$$

Note that the lower bound in Eq. (9.50) is indicative of a better compression ratio relative to that in Eq. (9.49). In fact, this phenomenon has been observed in simulations (see Ref. [11]), which indicate that KCS reconstructs the data with better compression ratios than GTCS.

9.4 Conclusion

In applications involving color images, video sequences, and multi-sensor networks, the data is intrinsically of high-order, and thus more suitably represented in tensorial form. Standard applications of CS to higher-order data typically involve representation of the data as long vectors that are in turn measured using large sampling matrices, thus imposing a huge computational and memory burden. As a result, extensions of CS theory to multidimensional signals have become an emerging topic. Existing methods include Kronecker compressed sensing (KCS) for sparse tensors and multi-way compressed sensing (MWCS) for sparse and low-rank tensors. KCS utilizes Kronecker product matrices as the sparsifying bases and to represent the measurement protocols used in distributed settings. However, due to the requirement to vectorize multidimensional signals, the recovery procedure is rather time consuming and not applicable in practice. Although MWCS achieves more efficient reconstruction by fitting a low-rank model in the compressed domain, followed by per-mode decompression, its performance relies highly on the quality of the tensor rank estimation results, the estimation being an NP-hard problem. We introduced the Generalized Tensor Compressed Sensing (GTCS)—a unified framework for compressed sensing of higher-order tensors which preserves the intrinsic structure of tensorial data with reduced computational complexity at reconstruction. We demonstrated that GTCS offers an efficient means for representation of multidimensional data by providing simultaneous acquisition and compression from all tensor modes. We introduced two reconstruction procedures, a serial method (GTCS-S) and a parallelizable method (GTCS-P), both capable of recovering a tensor based on noiseless and noisy observations, and compared the performance of the proposed methods with KCS and MWCS. As shown, GTCS outperforms KCS and MWCS in terms of both reconstruction accuracy (within a range of compression ratios) and processing speed. The major disadvantage of our methods (and of MWCS as well) is that the achieved compression ratios may be worse than those offered by KCS. GTCS is advantageous relative to vectorization-based compressed sensing methods such as KCS because the corresponding recovery problems are in terms of a multiple small measurement matrices U_i 's, instead of a single, large measurement matrix A , which results in greatly reduced complexity.

In addition, GTCS-P does not rely on tensor rank estimation, which considerably reduces the computational complexity while improving the reconstruction accuracy in comparison with other tensorial decomposition-based method such as MWCS.

Appendix

Let $X = [x_{ij}] \in \mathbb{R}^{N_1 \times N_2}$ be k -sparse. Let $U_i \in \mathbb{R}^{m_i \times N_i}$, and assume that U_i satisfies the NSP_k property for $i \in [2]$. Define Y as

$$Y = [y_{pq}] = U_1 X U_2^T \in \mathbb{R}^{m_1 \times m_2}. \quad (9.51)$$

Given a rank decomposition of X , $X = \sum_{i=1}^r z_i u_i^T$, where $\text{rank}(X) = r$, Y can be expressed as

$$Y = \sum_{i=1}^r (U_1 z_i)(U_2 u_i)^T, \quad (9.52)$$

which is also a rank- r decomposition of Y , where $U_1 z_1, \dots, U_1 z_r$ and $U_2 u_1, \dots, U_2 u_r$ are two sets of linearly independent vectors.

Proof. Since X is k -sparse, $\text{rank}(Y) \leq \text{rank}(X) \leq k$. Furthermore, both $R(X)$, the column space of X , and $R(X^T)$ are vector subspaces whose elements are k -sparse. Note that $z_i \in R(X), u_i \in R(X^T)$. Since U_1 and U_2 satisfy the NSP_k property, then $\dim(U_1 R(X)) = \dim(U_2 R(X^T)) = \text{rank}(X)$. Hence the decomposition of Y in Eq. (9.52) is a rank- r decomposition of Y , which implies that $U_1 z_1, \dots, U_1 z_r$ and $U_2 u_1, \dots, U_2 u_r$ are two sets of linearly independent vectors. This completes the proof. \square

References

1. Candes, E.J., Romberg, J.K.: The l_1 magic toolbox, available online: <http://www.l1-magic.org> (2006)
2. Candes, E.J., Romberg, J.K., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**(2), 489–509 (2006)
3. Candes, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59**, 1207–1223 (2006)
4. Candes, E.J.: The restricted isometry property and its implications for compressed sensing. *C. R. Math.* **346**(9–10), 589–592 (2008)
5. Cohen, A., Dahmen, W., DeVore, R.: Compressed sensing and best k -term approximation. *J. Am. Math. Soc.* **22**(1), 211–231 (2009)
6. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **21**, 1253–1278 (2000)
7. Donoho, D.L.: Compressed sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006)

8. Donoho, D.L., Tsaig, Y.: Extensions of compressed sensing. *Signal Process.* **86**(3), 533–548 (2006)
9. Donoho, D.L., Tsaig, Y.: Fast solution of l_1 -norm minimization problems when the solution may be sparse. *IEEE Trans. Inf. Theory* **54**(11), 4789–4812 (2008)
10. Duarte, M.F., Baraniuk, R.G.: Kronecker compressive sensing. *IEEE Trans. Image Process.* **21**(2), 494–504 (2012)
11. Friedland, S., Qun, L., Schonfeld, D.: Compressive sensing of sparse tensors. *IEEE Trans. Image Process.* **23**(10), 4438–4447 (2014)
12. Golub, G.H., Charles, F.V.L.: *Matrix Computations*, 4th edn. Johns Hopkins University Press, Baltimore (2013)
13. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. *SIAM Rev.* **51**(3), 455–500 (2009)
14. Rauhut, H.: On the impossibility of uniform sparse reconstruction using greedy methods. *Sampl. Theory Signal Image Process* (2008)
15. Sidiropoulos, N.D., Kyriolidis, A.: Multi-way compressed sensing for sparse low-rank tensors. *IEEE Signal Process. Lett.* **19**(11), 757–760 (2012)
16. Tucker, L.R.: The extension of factor analysis to three-dimensional matrices. In: *Contributions to Mathematical Psychology*. Holt, Rinehart and Winston, New York (1964)

Chapter 10

Sparse Model Uncertainties in Compressed Sensing with Application to Convolutions and Sporadic Communication

Peter Jung and Philipp Walk

Abstract The success of the compressed sensing paradigm has shown that a substantial reduction in sampling and storage complexity can be achieved in certain linear and non-adaptive estimation problems. It is therefore an advisable strategy for noncoherent information retrieval in, for example, sporadic blind and semi-blind communication and sampling problems. But, the conventional model is not practical here since the compressible signals have to be estimated from samples taken solely on the output of an un-calibrated system which is unknown during measurement but often compressible. Conventionally, one has either to operate at suboptimal sampling rates or the recovery performance substantially suffers from the dominance of model mismatch. In this work we discuss such type of estimation problems and we focus on *bilinear inverse problems*. We link this problem to the recovery of low-rank and sparse matrices and establish stable low-dimensional embeddings of the uncalibrated receive signals whereby addressing also efficient communication-oriented methods like *universal* random demodulation. Exemplarily, we investigate in more detail sparse convolutions serving as a basic communication channel model. In using some recent results from additive combinatorics we show that such type of signals can be efficiently low-rate sampled by semi-blind methods. Finally, we present a further application of these results in the field of phase retrieval from intensity Fourier measurements.

10.1 Introduction

Noncoherent compressed reception of information is a promising approach to cope with several future challenges in sporadic communication where short compressible messages have to be communicated in an unsynchronized manner over unknown,

P. Jung (✉)

Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany
e-mail: peter.jung@tu-berlin.de

P. Walk

Technische Universität München, Arcistrasse 21, 80333 München, Germany
e-mail: philipp.walk@tum.de

but compressible, dispersive channels. To enable such new communication concepts efficiently, it is therefore necessary to investigate blind and semi-blind sampling strategies which explicitly account for the low-dimensional structure of the signals. Since the compressed sensing paradigm provides a substantial reduction in sampling and storage complexity it is therefore also an advisable strategy for noncoherent information retrieval. However, in this and many related applications the conventional linear estimation model is a quite strong assumption since here the compressible signals of interest are not accessible in the usual way. Instead they have to be estimated from sampling data taken solely on the output of an additional linear system which is itself unknown during the measurement but often compressible. Thus, in the standard scheme one has either to operate at suboptimal rates or the overall estimation performance substantially suffers from the dominance of model mismatch. It is therefore important to evaluate the additional amount of sampling which is necessary to cope in a stable way with such model uncertainties. The output signals to be sampled do not constitute anymore a fixed finite union of low-dimensional canonical subspaces but a more complicated but still compressible set. In this chapter we focus on *bilinear models* and we discuss conditions which ensure additive complexity in input signals and model uncertainty. We motivate the relevance of this topic for sporadic communication in future cellular wireless networks and its random access strategies. In this setting the main dispersion is caused by convolutions of s -sparse channel impulse responses x with f -sparse user signals y . The convolution $x * y$ in dimension n can be recovered by conventional compressed sensing methods from $\mathcal{O}(sf \log n)$ incoherent samples whereby only $s + f$ “active” components contribute. However, we will show that for fixed x ordinary (non-circular) convolutions are invertible in y (and vice-versa) in a uniformly stable manner and can be compressed into $2^{2(s+f-2)\log(s+f-2)}$ dimensions *independent* of n . This demonstrates the possibility of low-rate sampling strategies in the order $\mathcal{O}((s + f) \log n)$ in our setting. Although efficient recovery algorithms operating at this rate are still unknown we show that sampling itself can be achieved efficiently with a considerable derandomized and universal approach, with a *random demodulator*. This proceeding contribution contains material from the joint work of the authors, presented in two talks at the CSA13 workshop, i.e., “*Low-Complexity Model Uncertainties in Compressed Sensing with Application to Sporadic Communication*” by Peter Jung and “*Stable Embedding of Sparse Convolutions*” by Philipp Walk.

Outline of the Work: First, we state in Section 10.1 the bilinear sampling problem and we discuss the relevance of this topic for sporadic communication in future cellular wireless networks. In Section 10.2 we will present a general framework for a stable random low-dimensional embedding of the signal manifolds beyond the standard linear vector model. We discuss structured measurements in this context and propose a *universal random demodulator* having an efficient implementation. At the end of this section we summarize in Theorem 1 that additive scaling in sampling complexity can be achieved for certain bilinear inverse problems, once a particular stability condition is fulfilled independent of the ambient dimension.

In Section 10.3 we will discuss such a condition for sparse convolutions in more detail and we show in Theorem 2 by arguments from additive combinatorics that an ambient dimension will not occur in this case. Finally, we show a further application for quadratic problems and draw the link to our work [40] on complex phase retrieval from intensity measurements of symmetrized Fourier measurements and presenting this result in Theorem 3.

10.1.1 Problem Statement

The standard linear model in compressed sensing is that the noisy observations $b = \Phi\Psi y + e$ are obtained from a *known* model under the additional assumption that y is essentially concentrated on a few components in a fixed or a few components in a fixed basis. Let us assume for the following exposition, that $y \in \Sigma_f$ is an f -sparse vector. Φ is the, possibly random, measurement matrix and Ψ denotes the dictionary (not necessarily a basis) in which the object can be sparsely described, both have to be known for decoding. For our purpose it is not important to understand Ψ as a property of the data. Instead, Ψ can also be understood as a part of the measurement process, i.e., viewing $\Phi\Psi$ as the overall measurement matrix. Solving for a sparse parameter vector y in this case can be done with a substantially reduced number of incoherent measurements. However what happens if Ψ (or Φ) is *not perfectly known*, i.e., depends on some unknown parameters resulting in an overall uncertainty in the estimation model? To our knowledge, this is one of the most sensible points for the application of compressed sensing to practical problems.

Model Uncertainties: Additive uncertainties in the overall measurement process have been investigated, for example, by [23]. An extension of this work with explicit distinction between errors in Φ and Ψ , suitable for redundant dictionaries, has been undertaken in [2]. Another situation, referring more to the multiplicative case, is the basis mismatch as has been studied, for example, by [12]. The strategy in the previous work was to estimate the degradation of the recovery performance in terms of the perturbation. However, if the unknown uncertainty is itself compressible in some sense one might treat it as a further unknown variable to be estimated from the same (blind) or prior (semi-blind, without calibrating the sampling device) observations as well. For example, can one handle the case where Ψ is known to have a compressible representation $\Psi = \sum_j x_j \Psi_j$ such that, for example, the coefficient vector $x \in \Sigma_s$ is s -sparse:

$$b = \Phi\left(\sum_j x_j \Psi_j\right)y + e =: \Phi B(x, y) + e \quad (10.1)$$

In principle, the original goal is here to estimate the sparse signal vector y from b under the condition that x is sparse as well. In this setting it would only be necessary to infer on the support of x . On the other hand, in many applications more precise

knowledge on the model parameters x is desirable as well and the task is then to recover the pair (x, y) up to indissoluble ambiguities.

Sampling Methods for Sporadic Communication: Our motivation for investigating this problem are universal sampling methods, which may become relevant for sporadic communication scenarios, in particular in wireless cellular networks. Whereby voice telephone calls and human generated data traffic were the main drivers for 2/3/4G networks (GSM, UMTS and LTE) this is expected to change dramatically in the future. Actually, 5G will bring again a completely new innovation cycle with many completely new and challenging applications (see, for example, [43] and the references therein). The *Internet of Things* will connect billions of smart devices for monitoring, controlling and assistance in, for example, the tele-medicine area, smart homes and smart factory, etc. In fact, this will change the internet from a human-to-human interface towards a more general machine-to-machine platform. However, machine-to-machine traffic is completely sporadic in nature which cannot be handled efficiently by only providing sufficient bandwidth.

A rather unexplored field here is the *instantaneous* joint estimation of user activity, channel coefficients, and data messages. As indicated, e.g., in [14] such approaches are necessary for one-stage random access protocols and therefore key enablers for machine-type communication within the vision of the “Internet of Things.” For a brief exposition, let us focus only on the estimation of a single channel vector x and data vector y from a single or only few observation cycles b . This vector b represents the samples taken by the receiver on elements $B(x, y)$ from a bilinear set under sparsity or more general compressibility constraints. A typical (circular) channel model for B is obtained in (10.1) with unitary representations of the finite Weyl–Heisenberg group on, e.g., \mathbb{C}^n :

$$(\Psi_j)_{kl} = e^{i2\pi \frac{j_1 l}{n}} \delta_{k, l \ominus j_2} \quad \text{for } j = (j_1, j_2) \in \{0, \dots, n-1\}^2. \quad (10.2)$$

These n^2 unitary operators fulfill the Weyl commutation rule and cyclically (\ominus denotes subtraction modulo n) shift the data signal y by j_2 and its discrete Fourier transform by j_1 . Even more they form an operator (orthonormal) basis with respect to the Hilbert–Schmidt inner product, i.e., *every* channel (linear mapping on y) can be represented as $\sum_j x_j \Psi_j$ (spreading representation of a “discrete pseudo-differential operator”). Since B is dispersive in the standard and the Fourier basis such channels are called doubly-dispersive and in most of the applications here the spreading function x is sparse (or compressible). Furthermore, at moderate mobility between transmitter and receiver x is essentially supported only on $j \in \{0\} \times \{0, \dots, n-1\}$, hence, the dominating single-dispersive effect is the *sparse circular convolution*, see Section 10.3.

For a *universal* dimensioning of, for example, a future random access channel architecture, where “universal” means that sampling strategies Φ should be independent of the particular low-dimensional structure of x and y , it is important to know how many samples m have to be taken in an efficient manner for stable distinguishing:

- (i) $B(x, y)$ from $B(x, y')$ and from $B(x', y)$ by universal measurements not depending on the low-dimensional structure and on x or y (*semi-blind methods*)
- (ii) completely different elements $B(x, y)$ and $B(x', y')$ by universal measurements not depending on the low-dimensional structure (*blind methods*)

In the view of the expected system parameters like (n, s, f) there will be substantial difference in whether a *multiplicative* $m = \mathcal{O}(sf \log n)$ or *additive* $m = \mathcal{O}((s+f)\log n)$ scaling can be achieved. Even more, we argue that achieving additive scaling without further compactness assumptions is closely related to the question whether (x, y) can be deconvolved up to indissoluble ambiguities at all from $B(x, y)$ which is in some cases also known as blind deconvolution. That this is indeed possible in a suitable random setting has already been demonstrated for the non-sparse and sufficiently oversampled case in [1].

10.1.2 Bilinear Inverse Problems with Sparsity Priors

Here we consider the model (10.1) from a compressive viewpoint which means that, due to the low complexity of signal sets, the measurement matrix $\Phi \in \mathbb{C}^{m \times n}$ corresponds to undersampling $m \ll n$. We use the well-known approach of lifting the bilinear map $B : \mathbb{C}^{n_1} \times \mathbb{C}^{n_2} \rightarrow \mathbb{C}^n$ to a linear map $B : \mathbb{C}^{n_1 \times n_2} \rightarrow \mathbb{C}^n$. Hereby, we can understand $x \otimes y = xy^T = x\bar{y}^*$ as a complex rank-one $n_1 \times n_2$ -matrix or as an $n_1 \cdot n_2$ -dimensional complex vector $\text{vec}(x \otimes y)$. As long as there arises no confusion we will always use the same symbol B , i.e., the structured signal z to be sampled in a compressive manner can be written in several ways:

$$z = B(x, y) = B(x \otimes y) = B(\text{vec}(x \otimes y)). \quad (10.3)$$

A step-by-step approach would be (i) estimating z from m noisy observations $b = \Phi z + e$ and (ii) deconvolving $(\lambda x, y/\lambda)$ from that estimate up to a scaling $\lambda \neq 0$ (due to the bilinearity) and, depending on B , further ambiguities. The second step requires injectivity of B on the desired subset in $\mathbb{C}^{n_1 \cdot n_2}$ and full inversion requires obviously $n_1 \cdot n_2 \leq n$. Both steps usually fall into the category of inverse problems and here we will consider the case with sparsity priors on x and y . For $x \in \Sigma_s$ and $y \in \Sigma_f$ the vector $\text{vec}(x \otimes y)$ is sf -sparse in $n_1 \cdot n_2$ dimensions, i.e. $z = B(x, y)$ is the image of the sf -sparse vector $\text{vec}(x \otimes y) \in \Sigma_{sf}$ under B . For recovering z (step (i)) via convex methods one could use the framework in [10]:

$$\min \|B^* z\|_1 \quad \text{s.t.} \quad \|\Phi z - b\|_2 \leq \varepsilon. \quad (10.4)$$

This (analysis-sparsity) approach recovers z successfully (or within the usually desired error scaling) if Φ acts almost-isometrically on $B(\Sigma_{sf})$ (called D-RIP in [10]). For example, if Φ obeys a certain concentration bound (like i.i.d. Gaussian matrices), $m = \mathcal{O}(sf \log(n_1 \cdot n_2 / (sf)))$ and B is a partial isometry (BB^* is a scaled

identity, i.e. the columns of B form a tight frame) (10.4) succeeds with exponential probability. Once B would be injective on Σ_{sf} it is in principle possible to extract $\text{vec}(x \otimes y)$ from z (step (ii)). In this case one could also consider directly the (synthesis-sparsity) approach for recovery, i.e., minimizing, for example, ℓ_1 -norm over the vectors $u \in \mathbb{C}^{n_1 \cdot n_2}$:

$$\min \|u\|_1 \quad \text{s.t.} \quad \|\Phi B(u) - b\|_2 \leq \varepsilon. \quad (10.5)$$

This one-step approach in turn depends on the precise mapping properties of B (in particular its anisotropy) and a detailed characterization could be done in terms of the RE-condition in [6]. For B being unitary (10.5) agrees with (10.4). Another approach in estimating the RIP-properties of the composed matrix ΦB for random measurements Φ having the concentration property was given in [31] yielding successful recovery in the regime $m = \mathcal{O}(sf \log(n_1 \cdot n_2 / (sf)))$.

But the set Σ_{sf} is considerably larger than $\text{vec}(M_{s,f})$ where $M_{s,f}$ denotes the rank-one tensor products, i.e.:

$$M_{s,f} := \{x \otimes y : x \in \Sigma_s \text{ and } y \in \Sigma_f\} \quad (10.6)$$

are the sparse rank-one $n_1 \times n_2$ matrices with not more than s non-zero rows and f non-zero columns. All the previous considerations make only use of the vector-properties of $x \otimes y$ and, hence, result in a multiplicative sf -scaling. Although the approach [20] termed *blind compressed sensing* gives structural insights into the rank-sparsity relation it also results in the sf -regime. Since the non-zero entries in $\text{vec}(M_{s,f})$ occur in s equal-sized blocks, each having at most f non-zero values, one might extend the vector-concept to block-sparse vectors. However, to fully exploit the correlation properties in the non-zero coefficients one has to consider the original estimation problem as a low-rank matrix recovery problem with sparsity constraints as already investigated in [13] in the view of noiseless identifiability. Unfortunately, already without sparsity this setting does not fall directly into the usual isotropic low-rank matrix recovery setting since the matrix picture is somehow “hidden behind” B resembling the anisotropy of ΦB . Whereby the anisotropic vector case has been extensively investigated in vector-RIP context by [6, 34] or in the RIP’less context by [27] (noiseless non-uniform recovery) very little is known here in the matrix case.

We already mentioned that restricting the problem (10.1) solely to the diagonal $B(x, x)$, which impose a quadratic inverse problem, resembles closely the phase-retrieval problem. We will make this precise for the non-compressive case in Section 10.3. Unfortunately this does not extend to the compressive case. Finally, we mention that our framework applies also to a certain extent to the multi-linear setting with minor modification, i.e., for higher order tensors. Such constructions occur not only in image and video processing but also in the communication context. For example, a separable spreading is characterized by $x_{(j_1, j_2)} = x_{j_1}^{(1)} \cdot x_{j_2}^{(2)}$ with (10.2) and is widely used as a simplified channel model yielding a 3rd order inverse problem in (10.1).

10.2 Stable Low-Dimensional Embedding

In this section we will establish a generic approach for stable embedding a non-compact and non-linear set V of a finite-dimensional normed space into a lower-dimensional space. Hereby $V \subseteq Z - Z$ will usually represent (not necessarily all) differences (chords/secants) between elements from another set Z . In this case, stable lower-dimensional embedding essentially establishes the existence of a “shorter description” of elements in Z whereby decoding can be arbitrarily complex. Once Z obeys a suitable convex surrogate function $f : Z \rightarrow \mathbb{R}_+$ the geometric convex approach [11] is applicable and via Gordon’s “escape through a mesh” theorem the Gaussian width of the descent cones of f provide estimates on the sampling complexity. But, e.g., for $Z = M_{s,f}$ this seems not be the case and conventional multi-objective convex programs (f is the infimal convolution of multiple surrogates like ℓ_1 and nuclear norm) are limited to the Pareto boundary [30] formed by the single objectives and have a considerable gap to the optimal sampling complexity.

10.2.1 Structure-aware Approximation

The first step is to establish an approximation statement in terms of an intrinsic distance description for a given subset V of a finite-dimensional space with a given norm $\|\cdot\|$. The problem arises since we need to quantify certain differences $v - r$ of two elements $v, r \in V$ whereby potentially $v - r \notin V$. Clearly $v - r \in \text{span}(V)$, but in this way we will potentially lose the low-complexity structure of V .

Intrinsic Distance: We consider an intrinsic notion of a *distance function* $d(v, r)$. For example, if V is path-connected, one might take the length of a smooth path $\gamma : [0, 1] \rightarrow V$ from $v = \gamma(0)$ to $r = \gamma(1)$ with $\gamma([0, 1]) \subset V$ and use $\|v - r\| \leq \int_0^1 \|\dot{\gamma}(t)\| dt$. However, we will not exploit Riemannian metrics here as has been done, for example, in [4] for *compact* manifolds. Instead, since a *rectifiable* path can be approached by finite partition sums, we consider a construction called *projective norm* in the case of tensors, see here [15, p.7] or [35, Ch.2]. More precisely, for differences $w = v - r \in V - V$ we define:

$$\|w\|_\pi := \inf \left\{ \sum_{i=1}^k \|v_i\| : w = \sum_{i=1}^k v_i \text{ with } v_i \in V \right\}, \quad (10.7)$$

which is the infimum over all finite representations of w by elements in V , whereby for any $w \in V$ one has $\|w\|_\pi = \|w\|$. More generally (beyond difference sets), if there is no decomposition for w , then $\|w\|_\pi$ is set to ∞ . In the examples later on we will always have that V is a central-symmetric linear cone, i.e., $V = \xi V$ for all $0 \neq \xi \in \mathbb{R}$. In this case V is generated by a central-symmetric atomic subset of the unit sphere and $\|w\|_\pi$ is then called an *atomic norm*, see here for example [11]. But,

instead of approaching the optimum in (10.7), we will later consider a particularly chosen decomposition $v - r = \sum_i v_i$ depending from the application (we specify the relevant cases later in Section 10.2.3). Once, for $v - r$ a decomposition $\{v_i\}$ in V has been specified, $d(v, r)$ is defined (and lower bounded) as:

$$d(v, r) := \sum_i \|v_i\| \geq \|v - r\|_\pi \geq \|v - r\|. \quad (10.8)$$

However, then $d(v, r)$ is not necessarily a metric on V . There is a useful condition: if $v - r$ has a k -term decomposition $v - r = \sum_{i=1}^k v_i$ in V and there is μ such that $\sum_{i=1}^k \|v_i\|^2 \leq \mu \|\sum_{i=1}^k v_i\|^2$ —it would follow from Cauchy–Schwartz inequality that:

$$\|v - r\| \leq d(v, r) = \sum_{i=1}^k \|v_i\| \leq (k \sum_{i=1}^k \|v_i\|^2)^{\frac{1}{2}} \leq \sqrt{k\mu} \|v - r\|. \quad (10.9)$$

Thus, within $\sqrt{k\mu}$ the norms $\|\cdot\|_\pi$ and $\|\cdot\|$ are then equivalent on V and for Euclidean norms we have $\mu = 1$ for orthogonal decompositions. A worst-case estimate is obviously $k = \dim(V)$ meaning that V contains a frame for its span with lower frame-bound $1/\mu > 0$ which, however, could be arbitrary small depending on the anisotropic structure of V . Instead, we shall therefore consider $V = B(U)$ as the image under a given mapping B of another “nicer” set U which a-priori has this property.

The Essential Approximation Step: Here we will now give a short but generalized form of the essential step in [5]. Variants of it can be found in almost all RIP-proofs based on nets. However, here we focus on the important property, that the (linear) compression mapping $\Phi : V \rightarrow W$ should be solely applied on elements of V .

Lemma 1. *Let $\delta, \varepsilon \in (0, 1)$ and $\Phi : V \rightarrow W$ be a linear map between subsets V and W of finite normed spaces, each with its norm $\|\cdot\|$. Assume that for each $v \in V$ there exists $r = r(v) \in V$ such that*

- (i) *a decomposition $\{v_i\}_{i=1}^k \subset V$ exists for $v - r = \sum_i v_i$ with $d(v, r) := \sum_i \|v_i\| \leq \varepsilon \|v\|$*
- (ii) *and $\|\Phi v\| - \|r\| \leq \frac{\delta}{2} \|r\|$.*

Then it holds for $\varepsilon < \delta/7$:

$$\|\Phi v\| - \|v\| \leq \delta \|v\| \quad \text{for all } v \in V. \quad (10.10)$$

If $\|v\| = \|r(v)\|$ for all $v \in V$, then (10.10) holds also for $\varepsilon < \delta/4$.

Let us make here the following remark: Lemma 1 neither requires that V is symmetric ($V = -V$) nor is a linear cone ($V = \xi V$ for all $\xi > 0$). However, if the lemma holds for a given V and approximation strategy $v \rightarrow r(v)$ then, it also holds

for ξV with $\xi \in \mathbb{C}$ and $r(\xi \cdot) := \xi r(\cdot)$ ¹. It holds therefore also for $\bigcup_{\xi \in \mathbb{C}} \xi V$ whereby the converse is wrong.

Proof. We set $a = 0$ if we already know that $\|v\| = \|r\|$ and $a = 1$ else. Using triangle inequalities we get for $v, r \in V$ with the decomposition $v - r = \sum_{i=1}^k v_i$ given in (i):

$$\begin{aligned} \|\Phi v\| - \|v\| &= \|\Phi v\| - \|\Phi r\| + \|\Phi r\| - \|v\| \\ &\leq \|\Phi v\| - \|\Phi r\| + a\|v\| - \|r\| + \|\Phi r\| - \|r\| \\ &\leq \|\Phi(v - r)\| + a\|v - r\| + \|\Phi r\| - \|r\| \\ &\stackrel{(ii)}{\leq} \sum_i \|\Phi v_i\| + a \cdot d(v, r) + \frac{\delta}{2} \|r\| \end{aligned} \quad (10.11)$$

where in the last step we also used the property of d given in (10.8). Since $\|v\| - \|r\| \leq a\|v - r\| \leq a \cdot d(v, r)$ we have $\|r\| \leq \|v\| + a \cdot d(v, r)$ and therefore:

$$\begin{aligned} \|\Phi v\| - \|v\| &\leq \sum_i \|\Phi v_i\| + a\left(1 + \frac{\delta}{2}\right) \cdot d(v, r) + \frac{\delta}{2} \|v\| \\ &\leq \sum_i \|\Phi v_i\| + \left(a\left(1 + \frac{\delta}{2}\right)\varepsilon + \frac{\delta}{2}\right) \|v\| \end{aligned} \quad (10.12)$$

where the last line follows from $d(v, r) \leq \varepsilon\|v\|$ given in the assumption (i) of the lemma. Note that, if we can ensure $\|r\| = \|v\|$, then $a = 0$. We now follow the same strategy as in [5] and define the constant:

$$A := \sup_{0 \neq v \in V} \frac{\|\Phi v\| - \|v\|}{\|v\|}. \quad (10.13)$$

implying that for any $\varepsilon' > 0$ there is $v^* \in V$ with $(A - \varepsilon')\|v^*\| \leq \|\Phi v^*\| - \|v^*\|$. From the prerequisite (i) of the lemma there also exists $r^* = r(v^*) \in V$ with $d(v^*, r^*) := \sum_i \|v_i^*\| \leq \varepsilon\|v^*\|$ for a decomposition $v^* - r^* = \sum_i v_i^*$. We have then from (10.13) that $\sum_i \|\Phi v_i^*\| \leq (1+A)d(v^*, r^*) \leq (1+A)\varepsilon\|v^*\|$ and using (10.12) for $v = v^*$ gives:

$$(A - \varepsilon')\|v^*\| \leq \|\Phi v^*\| - \|v^*\| \leq \left((1+A)\varepsilon + a\left(1 + \frac{\delta}{2}\right)\varepsilon + \frac{\delta}{2}\right) \|v^*\|. \quad (10.14)$$

Solving for A gives:

¹To see this, let $v, r(v) \in V$ with a decomposition $v - r(v) = \sum_i v_i$. Then $r(\xi v) = \xi r(v) \in \xi V$ and $\xi v - r(\xi v) = \xi(v - r(v)) = \sum_i \xi v_i$ with $\xi v_i \in \xi V$.

$$A \leq \frac{\varepsilon + a(1 + \frac{\delta}{2})\varepsilon + \frac{\delta}{2} + \varepsilon'}{1 - \varepsilon} \stackrel{(!)}{\leq} \delta \Leftrightarrow \varepsilon \leq \frac{\delta - 2\varepsilon'}{2 + a(2 + \delta) + 2\delta} \Leftrightarrow \varepsilon < \frac{\delta}{4 + 3a},$$

(10.15)

since for each fixed $\delta < 1$ there exists a sufficiently small $\varepsilon' > 0$ such that (10.15) holds. Recall, in general, $a = 1$ but if we are able to choose $\|r\| = \|v\|$ we have $a = 0$. \square

Summarizing, the approximation strategy of [5] applies in a quite generalized context. For a given V one has (i) to find a suitable² $d(v, r)$ and (ii) find covering number estimates for V in terms of d which are better than those of the ambient space. However, the second step seems notoriously difficult and we approach this by sticking to a particular parametrization of the set V .

10.2.2 Bi-Lipschitz Mappings and the RNMP

Here, we consider now the non-linear set V as the image $V = B(U)$ of a (parameter) set U of a normed space under a linear map $B : U \rightarrow V$, i.e., B is always a surjection. The domain U can be, for example, subsets of vectors or matrices equipped with the norm of the ambient space (usually some Euclidean norm). We shall approximate each element $v \in V$ by another element $r = r(v) \in V$ but taking care of the case $v - r \notin V$. To this end we will perform the approximation in the domain U of B and translate this afterwards to its range V . Thus, we will need the following two properties (A_σ) and $(B_{\alpha, \beta})$:

(A_σ) : A set U has the property (A_σ) for $\sigma > 0$ if it is the finite union $U = \bigcup_{l=1}^L U_l$ of L subsets of a normed space and for each $u, \rho \in U$ with $u, \rho \in U_l$ for some $l = 1 \dots L$ there exists $\{u_i\}_{i=1}^k \subset U_l$ yielding a k -term decomposition $u - \rho = \sum_{i=1}^k u_i$ with:

$$\sum_{i=1}^k \|u_i\| \leq \sigma \cdot \|\sum_{i=1}^k u_i\|. \quad (10.16)$$

For example, if U is a subspace, then $u - \rho \in U$ for each $u, \rho \in U$. In this case, the “ $k = 1$ ”-decomposition $u_1 = u - \rho$ is valid giving $\sigma = 1$. However, if U is a union of L subspaces U_l then u and ρ usually have to be in the *same* subspace for $\sigma = 1$. On the other hand, if U is some subset equipped with an Euclidean norm and $u - \rho \notin U$ but is guaranteed to have an *orthogonal* k -term decomposition in U , then $\sigma = \sqrt{k}$, see (10.9) for $U = V$. For example, let U be the matrices of maximal rank κ equipped with the Frobenius (Hilbert–Schmidt) norm. In this case it might happen that $u - \rho \notin U$ but the singular value decomposition provides an orthogonal (in the

²A suitable *decomposition strategy* for $v - r = \sum_i v_i$ with all $v_i \in V$ has to be found. Then $d(v, r) := \sum_i \|v_i\|$ defines an intrinsic distance function. We will give examples in Section 10.2.3.

Hilbert–Schmidt inner product) “ $k = 2$ ”-decomposition in U for any $u, \rho \in U$, i.e., $\sigma = \sqrt{2}$. However, if U is the union of L matrix subsets U_l of maximal rank κ (like sparse low-rank matrices), then u and ρ have usually to be from the *same* subset for (10.16) to hold with $\sigma = \sqrt{2}$.

To switch now between domain and range of B we will also need the property:

$(B_{\alpha,\beta})$: A map $B : U \rightarrow V$ has the property $(B_{\alpha,\beta})$ if there is $0 < \alpha \leq \beta < \infty$ such that it holds:

$$\alpha \|u\| \leq \|B(u)\| \leq \beta \|u\| \quad \text{for all } u \in U \quad (10.17)$$

In [39] the authors have considered condition (ii) for $U = \{x \otimes y : x \in X, y \in Y\}$ where X and Y are two given cones of an Euclidean space under the name *restricted norm multiplicativity property* (RNMP) since in this case $\|x \otimes y\| = \|x\| \|y\|$. We will further discuss such models for U in (10.26) and (10.27) below and in much more detail for convolutions in Section 10.3. On the other hand, for difference sets $U = M - M$ and linear mappings B this is the *bi-Lipschitz condition* of B on M . We have the following lemma:

Lemma 2. *Let $\hat{\varepsilon} > 0$ and $B : U \rightarrow V$ be a linear map having property $(B_{\alpha,\beta})$. If $u, \rho \in U$ fulfill $\|u - \rho\| \leq \hat{\varepsilon} \|u\|$ and there exists a decomposition $\{u_i\} \subset U$ with $u - \rho = \sum_i u_i$ such that (10.16) holds for some $\sigma > 0$, then it holds:*

$$\|v - r\| \leq d(v, r) \leq \frac{\beta \sigma}{\alpha} \hat{\varepsilon} \|v\|. \quad (10.18)$$

where $v := B(u)$, $r := B(\rho)$ and $d(v, r) := \sum_i \|B(u_i)\|$.

Proof. The assertion follows directly from:

$$\begin{aligned} \|v - r\| &= \|B(u) - B(\rho)\| = \|B(u - \rho)\| = \left\| \sum_i B(u_i) \right\| \leq \sum_i \|B(u_i)\| = d(v, r) \\ &\stackrel{(10.17)}{\leq} \beta \sum_i \|u_i\| \stackrel{(10.16)}{\leq} \beta \sigma \|u - \rho\| \leq \beta \sigma \hat{\varepsilon} \|u\| \stackrel{(10.17)}{\leq} \frac{\beta \sigma}{\alpha} \hat{\varepsilon} \|v\| \quad \square \end{aligned} \quad (10.19)$$

In the next section we will use this lemma to translate the accuracy in approximating u by some $\rho = \rho(u)$ from domain U of B to its image V . Note that linearity of B is used only in the first step of (10.19) whereby extension is possible once there holds $\|B(u) - B(\rho)\| \leq c \cdot \sum_{i=1}^k \|B(u_i)\|$ uniformly for every $u \in U$ and $\rho = \rho(u)$. However, we will not further argue on this here.

10.2.3 Covering and Entropy

The remaining task is now to specify for given $\hat{\varepsilon} > 0$ an approximation strategy $u \rightarrow \rho(u)$ such that Lemma 2 can be applied to all $u \in U$, i.e., for each $u \in U$ there is $\rho = \rho(u) \in U$ with $\|u - \rho\| \leq \hat{\varepsilon}\|u\|$ and each $u - \rho$ has a finite decomposition in U . From this equation it is clear that we have to consider $\hat{\varepsilon}$ -coverings³ for the set $U' := \{u/\|u\| : 0 \neq u \in U\}$ and to estimate its *covering number*:

$$N_{\hat{\varepsilon}}(U') := \min\{|R| : R \text{ is an } \hat{\varepsilon}\text{-covering for } U'\} \quad (10.20)$$

Its logarithm $H_{\hat{\varepsilon}}(U') = \log N_{\hat{\varepsilon}}(U')$ is called the (metric) $\hat{\varepsilon}$ -entropy of U' . Due to pre-compactness of U' as a subset of the unit ball these quantities are always finite. Furthermore we will abbreviate now $V' := \{v/\|v\| : 0 \neq v \in V\}$. Let us restate Lemma 2 in this context:

Corollary 1. *Let $B : U \rightarrow V$ be linear with property $(B_{\alpha,\beta})$, U be a linear cone with property (A_σ) and $\hat{\varepsilon} > 0$. Then each $\hat{\varepsilon}$ -covering for U' induces an $\varepsilon = \frac{\beta\sigma\hat{\varepsilon}}{\alpha}$ -covering for V' for the norm in V as well as for an intrinsic distance and:*

$$H_\varepsilon(V') \leq H_{\alpha\varepsilon/(\beta\sigma)}(U') \quad (10.21)$$

holds.

The property (A_σ) always induces an intrinsic distance on V as will be seen in the proof below.

Proof. Let be $R \subset U'$ an $\hat{\varepsilon}$ -covering for U' , i.e., for each $u \in U'$ there exists $\rho(u) \in R$ such that $\|u - \rho(u)\| \leq \hat{\varepsilon}$. Since U is a linear cone, i.e., $\xi U = U$ for $\xi > 0$, it follows for all $0 \neq u \in U$ that $\|u - \rho(u)\| \leq \hat{\varepsilon}\|u\|$ holds with $\rho(u) := \rho(u/\|u\|)\|u\| \in U$.

Property (A_σ) asserts now that there always exists a decomposition $\{u_i\} \subset U$ for $u - \rho(u) = \sum_i u_i$ in U satisfying (10.16). For each $v := B(u)$ set $r = r(v) := B(\rho(u))$ and therefore $r - v = \sum_i B(u_i)$ has an intrinsic decomposition in V . Define $d(v, r) := \sum_i \|B(u_i)\|$. From Lemma 2 it follows that:

$$\|v - r\| \leq d(v, r) \leq \frac{\beta\sigma\hat{\varepsilon}}{\alpha}\|v\|.$$

Indeed, for each $v \in V'$ this means $\|v - r\| \leq d(v, r) \leq \beta\sigma\hat{\varepsilon}/\alpha$ which yields (10.21) and shows that v and r are also close in the intrinsic distance induced by (A_σ) . \square

We will now give a short overview on some cases for U which have property (A_σ) , their entropy bounds and the corresponding values for σ in (10.16). All examples are central-symmetric linear cones, i.e., $U = \xi U$ for all $0 \neq \xi \in \mathbb{R}$. Hence, Corollary 1 will translate this via $\hat{\varepsilon} = \alpha\varepsilon/(\beta\sigma)$ to an entropy estimate for V once B has

³ R is an $\hat{\varepsilon}$ -net for U' if for each $u \in U'$ exists $\rho = \rho(u) \in R$ with $\|u - \rho\| \leq \hat{\varepsilon}$, i.e., the union of these $\hat{\varepsilon}$ -balls centered at ρ cover U' .

property $(B_{\alpha,\beta})$. If we assume that $U \subseteq \bigcup_{l=1}^L U_l$ we have $N_{\hat{\epsilon}}(U') \leq \sum_{l=1}^L N_{\hat{\epsilon}}(U'_l)$ and if furthermore all $U'_l := \{u/\|u\| : 0 \neq u \in U_l\}$ have the same covering number as U'_l , we get therefore:

$$H_{\hat{\epsilon}}(U') \leq H_{\hat{\epsilon}}(U'_l) + \log L. \quad (10.22)$$

Of most interest here is the dependency on the ambient dimension of U . If there is sufficient compressibility, the ambient dimension will *explicitly* occur only in L whereby $H_{\hat{\epsilon}}(U'_l)$ could depend on it, for fixed $\epsilon > 0$, only through $\hat{\epsilon} = \alpha\epsilon/(\beta\sigma)$. This is indeed the case for sparse vectors and matrices as it will be shown now.

Finite Union of Subspaces: If each U_l is contained in a subspace of real dimension d , then one can choose for any $\hat{\epsilon} > 0$ and each $l = 1 \dots L$ an $\hat{\epsilon}$ -net for the unit ball \tilde{U}'_l in $\tilde{U}_l := \text{span}(U_l)$ and one has the well-known estimate $H_{\hat{\epsilon}}(U'_l) \leq H_{\hat{\epsilon}}(\tilde{U}'_l) \leq d \log(3/\hat{\epsilon})$ being valid for any norm not only for the Euclidean norm [38, Sec. 2.2]. Even more, any smooth manifold of real dimension d behaves in this way for $\hat{\epsilon} \rightarrow 0$. The union of these L nets is an $\hat{\epsilon}$ -net for U' . Thus, if U is therefore contained in a union of L subspaces of the same dimension d , we have from (10.22):

$$H_{\hat{\epsilon}}(U') \leq d \log(3/\hat{\epsilon}) + \log L \quad (10.23)$$

In particular, in a subspace we have $\sigma = 1$ in (10.16) as already explained after (10.16). Furthermore, in the sparse vector case, $U = \Sigma_{2k}$ is the union of $L := \binom{n}{d} \leq \binom{en}{d}^d$ different $d = 2k$ -dimensional subspaces and we have in this case $H_{\hat{\epsilon}}(U') \leq d \log(3/\hat{\epsilon}) + d \log(en/d)$.

Low-rank Matrices: Consider differences of rank- κ matrices M , i.e., $U = M - M$ are $n \times n$ matrices of rank at most 2κ with the Euclidean (Frobenius) norm $\|u\|^2 := \langle u, u \rangle$ defined by the Hilbert–Schmidt inner product. From [9, Lemma 3.1] it follows:

$$H_{\hat{\epsilon}}(U') \leq (2n+1)2\kappa \log(9/\hat{\epsilon}). \quad (10.24)$$

A matrix $u - \rho$ for *any* $u, \rho \in U$ has rank at most 4κ and can be decomposed as $u - \rho = u_1 + u_2$ for $u_1, u_2 \in U$ with $\langle u_1, u_2 \rangle = 0$, i.e. it fulfills (10.16) for $k = 2$ and $\sigma \leq \sqrt{2}$. Hence, U has property (A_σ) for $\sigma = \sqrt{2}$.

Low-rank and Sparse Matrices: Here we consider the union $U = M_{s,f}^\kappa - M_{s,f}^\kappa$ of $L = \binom{n}{2s} \binom{n}{2f}$ different sets of differences of rank- κ matrices $M_{s,f}^\kappa$ (equipped with the Frobenius norm) as defined in (10.6) and it follows from (10.22) and (10.24) that:

$$H_{\hat{\epsilon}}(U') \leq (2s+2f+1)2\kappa \log(9/\hat{\epsilon}) + 2(s+f) \log \frac{en}{2 \min(s,f)}. \quad (10.25)$$

The *bilinear and sparse model* is here the special case for $\kappa = 1$ ($M_{s,f} = M_{s,f}^1$ in (10.6)) and, once $\hat{\epsilon}$ does not depend on n , entropy scales at most as $\mathcal{O}((s+f)\log n)$ for sufficiently large n . Again, U has here the property (A_σ) for $\sigma = \sqrt{2}$.

Sparse Bilinear Case with one Known Input: Lemma 1 and Lemma 2 do not require that V is a *full* difference set. Here, we essentially consider the set:

$$V = \bigcup_{x \in \Sigma_s} (B(x \otimes \Sigma_f) - B(x \otimes \Sigma_f)) = B(M_{s,2f}). \quad (10.26)$$

This case will be relevant when we, universally, have to sample and store measurements in a repetitive blind manner whereby we will have knowledge about one of the components during decoding, i.e. this comprises a *universal sampling method*. Thus, once (10.17) holds for this rank-one set U with (α, β) being independent of the ambient dimension its entropy bound scales additive in s and f , i.e., $\mathcal{O}((s+f) \log n)$ according to (10.25) instead of $\mathcal{O}(s \cdot f \log n)$. In our first covering estimate on this set in [39] we have established this scaling for cones directly, not using [9, Lemma 3.1].

The Quadratic and Symmetric Case: Here, we consider again differences of the form $V = Z - Z$ for $Z = \bigcup_{x \in \Sigma_s} B(x \otimes x)$. If B is symmetric, the binomial formula asserts that:

$$V = \bigcup_{x,y \in \Sigma_s} B((x+y) \otimes (x-y)) = B(M_{2s,2s}) \quad (10.27)$$

This model is important for sparse convolutions and sparse phase retrieval as discussed in Section 10.3. Once again, if (10.17) holds for $U = M_{2s,2s}$ independent of the ambient dimension, entropy scales linearly in the sparsity s , i.e., $\mathcal{O}(s \log n)$ as follows from (10.25) and *not* as $\mathcal{O}(s^2 \log n)$.

10.2.4 Random Sampling Methods

Based on the properties (A_σ) condition $\|\Phi v\| - \|v\| \leq \delta \|v\|$ should hold simultaneously for all $v \in V = B(U)$ with high probability. For difference sets $V = Z - Z$ (meaning that $U = M - M$ for another set M since B is linear) this condition provides a stable embedding of Z in W and, by (10.17), *it always implies stable embedding M in W —but in the anisotropic situation*. An estimate for the RIP-like constant $\hat{\delta}$ of the composed map $\Phi B : U \rightarrow W$ follows with $\alpha = (1 - \eta)\xi$ and $\beta = (1 + \eta)\xi$ as:

$$\begin{aligned} \|\Phi B(u) - \xi \|u\| \| &\leq \|\Phi B(u)\| - \|B(u)\| + \|B(u)\| - \xi \|u\| \\ &\leq \delta \|B(u)\| + \eta \xi \|u\| \leq ((1 + \eta)\delta \xi + \eta \xi) \|u\| \\ &= \xi ((1 + \eta)\delta + \eta) \|u\| = \xi (\delta + \eta(\delta + 1)) \|u\| =: \xi \hat{\delta} \|u\| \end{aligned} \quad (10.28)$$

The term $\eta(\delta + 1)$ reflects the degree of anisotropy caused by B . A similar relation for the usual definition of the RIP-property has been obtained, for example, in [31]. Although we do not discuss efficient recovery here, recall that, for example, [8] states that for $\hat{\delta} < 1/3$ certain convex recovery methods (ℓ_1 -minimization for sparse vectors and nuclear norm minimization for low rank matrices when $\|\cdot\|$ are Euclidean norms) are successful, implying $\eta < 1/3$.

Random Model with Generic Concentration: As shown already in the sparse vector case in [5, Thm 5.2] we have in this generalized setting a similar statement:

Lemma 3. *Let $\Phi : V \rightarrow W$ be a random linear map which obeys for $\delta \in (0, 1), \gamma > 0$ the uniform bound $\Pr(\{\|\Phi r\| - \|r\| \leq \frac{\delta}{2}\|r\|\}) \geq 1 - e^{-\gamma}$ for each $r \in V$. Let $B : U \rightarrow V$ linear with property $(B_{\alpha, \beta})$ where U is a linear cone having property (A_σ) . Then:*

$$\Pr(\{\forall v \in V : \|\Phi v\| - \|v\| \leq \delta\|v\|\}) \geq 1 - e^{-(\gamma - H_{\hat{\delta}}(U'))} \quad (10.29)$$

where $\hat{\delta} < \frac{\alpha}{7\beta\sigma}\delta$.

Proof. From (10.29) it follows that it is sufficient to consider the set $V' = \{v/\|v\| : 0 \neq v \in V\}$. From Corollary 1 we have for this set a covering ε -net R with respect to an intrinsic distance of cardinality $|R| \leq e^{H_\varepsilon(V')} \leq e^{H_{\hat{\delta}}(U')}$ with $\hat{\delta} = \alpha\varepsilon/(\beta\sigma)$. Taking the union bound over R asserts therefore that $\|\Phi r\| - \|r\| \leq \frac{\delta}{2}\|r\|$ with probability $\geq 1 - e^{-(\gamma - H_{\hat{\delta}}(U'))}$ for all $r \in R$ and the same Φ . From Lemma 1, if $\varepsilon = \frac{\beta\sigma}{\alpha}\hat{\delta} < \delta/7$ there holds $\|\Phi v\| - \|v\| \leq \delta\|v\|$ for all $v \in V$ and the same Φ simultaneously with probability exceeding $1 - e^{-(\gamma - H_{\hat{\delta}}(U'))}$. \square

This lemma shows that the concentration exponent γ must be in the order of the entropy $H_{\hat{\delta}}(U')$ to ensure embedding with sufficiently high probability. By construction such a random embedding is a *universal sampling method* where the success probability in (10.29) depends solely on the entropy and not on the particular “orientation” of U' which has several practical-relevant advantages as discussed already in the introduction.

Randomizing Fixed RIP Matrices: We extent the statement of Lemma 3 to include randomized classical RIP matrices, i.e., Φ is (k, δ_k) -RIP if $\|\Phi v\|_2^2 - \|v\|_2^2 \leq \delta_k\|v\|_2^2$ for each k -sparse vector v . The motivation behind is the use of structured or deterministic measurements with possibly fast and efficient transform implementation. Such measurements usually *fail to be universal* and do not have concentration properties. However, the important result of [26] states that this can be achieved by a moderate amount of randomization. Randomization can, for example, be done with a multiplier D_ξ performing point-wise multiplication with a vector ξ having i.i.d. ± 1 components, see here also [25] for more general ξ . We consider now $V \subseteq \mathbb{C}^n$ and ℓ_2 -norms.

Lemma 4. *Let $B : U \rightarrow V$ and U as in Lemma 3 and the random matrix D_ξ is distributed as given above. Let $\delta, \rho > 0$ and Φ be (k, δ_k) -RIP with $\delta_k \leq \delta/8$ and $k \geq 40(\rho + H_{\hat{\epsilon}}(U') + 3 \log(2))$. Then*

$$\Pr(\{\forall v \in V : \|\Phi D_\xi v\|_2 - \|v\|_2 \leq \delta \|v\|_2\}) \geq 1 - e^{-\rho} \quad (10.30)$$

where $\hat{\epsilon} < \frac{\alpha}{7\beta\sigma} \delta$.

Proof. For a given (k, δ_k) -RIP matrix Φ with $k \geq 40(\rho + p + \log(4))$ and $\delta_k \leq \frac{\delta}{8}$ it follows from [26]: ΦD_ξ is with probability $\geq 1 - e^{-\rho}$ a $\frac{\delta}{2}$ -Johnson–Lindenstrauss–embedding for any point cloud of cardinality e^p . Now, from Corollary 1, there exists an ε -net R for V' of cardinality $|R| \leq e^{H_\varepsilon}$ where $H_\varepsilon = H_\varepsilon(V') \leq H_{\hat{\epsilon}}(U')$ with $\hat{\epsilon} = \alpha\varepsilon/(\beta\sigma)$. When adding the zero-element to the point cloud it has cardinality:

$$|R| \leq e^{H_\varepsilon} + 1 = e^{H_\varepsilon}(1 + e^{-H_\varepsilon}) \leq 2e^{H_\varepsilon} = e^{H_\varepsilon + \log(2)} \quad (10.31)$$

Therefore, set $p = H_\varepsilon + \log(2)$ (or the next integer). From [26] it follows then that for each $k \geq 40(\rho + H_\varepsilon + \log(2) + \log(4)) = 40(\rho + H_\varepsilon + 3 \log(2))$ the point cloud R is mapped almost-isometrically (including norms since 0 is included), i.e., with probability $\geq 1 - e^{-\rho}$ we have $|\|\Phi D_\xi r\|_2^2 - \|r\|_2^2| \leq \frac{\delta}{2} \|r\|_2^2$ for all $r \in R$ which implies:

$$\Pr(\{\forall r \in R : \|\Phi D_\xi r\|_2 - \|r\|_2 \leq \frac{\delta}{2} \|r\|_2\}) \geq 1 - e^{-\rho}. \quad (10.32)$$

We will choose $\varepsilon = \frac{\beta\sigma}{\alpha} \hat{\epsilon} < \frac{\delta}{7}$. Then, since R is an ε -net for V' and U has property (A_σ) inducing an intrinsic decomposition and distance, it follows from Lemma 2 that:

$$\Pr(\{\forall v \in V : \|\Phi D_\xi v\|_2 - \|v\|_2 \leq \delta \|v\|_2\}) \geq 1 - e^{-\rho} \quad \square \quad (10.33)$$

Randomizing Random RIP Matrices: We extent Lemma 4 to random structured RIP models which itself are in many cases not universal and can therefore without further randomization not be used directly in the generalized framework. Assume an “ (M, p) RIP model,” meaning that the $m \times n$ random matrix Φ is (k, δ_k) -RIP with probability $\geq 1 - e^{-\gamma}$ and $\delta_k \leq \delta$ if $m \geq c\delta^{-2}k^p M(n, k, \gamma)$ for a constant $c > 0$. Define for a given U :

$$k_{\hat{\epsilon}}(\rho) := 40(\rho + H_{\hat{\epsilon}}(U') + 3 \log(2)) \quad (10.34)$$

We have the following lemma:

Lemma 5. *Let $\delta > 0$ and D_ξ , $B : U \rightarrow V$ and U as in Lemma 4. Let Φ be an $m \times n$ random (M, p) -RIP model (independent of D_ξ) and $k_{\hat{\epsilon}}(\rho)$ as given above for $\hat{\epsilon} < \frac{\alpha}{7\beta\sigma} \delta$. Then ΦD_ξ is universal in the sense that:*

$$\Pr(\{\forall v \in V : \|\Phi D_\xi v\|_2 - \|v\|_2 \leq \delta \|v\|_2\}) \geq 1 - (e^{-\rho} + e^{-\gamma}) \quad (10.35)$$

if $m \geq 64c\delta^{-2}k_{\hat{\epsilon}}(\rho)^p M(n, k_{\hat{\epsilon}}(\rho), \gamma)$.

Proof. The proof follows directly from Lemma 4. Define $\delta' = \delta/8$. Then the model assumptions assert that for $m \geq c\delta'^{-2}k_{\hat{\epsilon}}(\rho)^p M(n, k_{\hat{\epsilon}}(\rho), \gamma)$ the matrix Φ has $(k_{\hat{\epsilon}}(\rho), \delta_k)$ -RIP with $\delta_k \leq \delta' = \delta/8$ and probability $\geq 1 - e^{-\gamma}$. Thus, by Lemma 4 for any $\rho > 0$ the claim follows. \square

The best (ρ, γ) -combination for a fixed probability bound $\geq 1 - e^{-\lambda}$ can be estimated by minimizing $k_{\hat{\epsilon}}(\rho)^p M(n, k_{\hat{\epsilon}}(\rho), \gamma)$. We will sketch this for random *partial circulant matrices* $P_{\Omega} \hat{D}_{\eta}$. Let $F = (e^{i2\pi kl/n})_{k,l=0}^{n-1}$ be the $n \times n$ -matrix of the (non-unitary) discrete Fourier transform. Then, $\hat{D}_{\eta} := F^{-1} D_{\eta} F$ is an $n \times n$ circulant matrix with $\hat{\eta} := F\eta$ on its first row (Fourier multiplier η) and the $m \times n$ matrix $P_{\Omega} := \frac{1}{|\Omega|} \mathbf{1}_{\Omega}$ is the normalized projection onto coordinates in the set $\Omega \subset [1, \dots, n]$ of size $m = |\Omega|$. Random convolutions for compressed sensing and universal demodulation concepts are already proposed in [33]. In [37] a related approach has been called *random demodulator* and is used for sampling frequency-sparse signals via convolutions on the Fourier side (being not suitable for sporadic communication tasks). Measurement matrices $P_{\Omega} \hat{D}_{\eta}$ are systematically investigated in [32] showing that (k, δ_k) -RIP properties hold in the regime $m = \mathcal{O}((k \log n)^{\frac{3}{2}})$. Finally, linear scaling in k (and this will be necessary for the overall additivity statement in the bilinear setting) has been achieved in [26]. But $P_{\Omega} \hat{D}_{\eta}$ is *not universal* meaning that the signal has to be k -sparse in the canonical basis.

Therefore, we propose the *universal random demodulator* $P_{\Omega} \hat{D}_{\eta} D_{\xi}$ which still has an efficient FFT-based implementation but is independent of the sparsity domain. Such random matrices work again in our framework:

Lemma 6. *Let be D_{ξ} , $B : U \rightarrow V$ and U as in Lemma 4. Let $\Phi = P_{\Omega} \hat{D}_{\eta} D_{\xi}$ be an $m \times n$ partial random circulant matrix with η being a vector with independent zero-mean, unit-variance subgaussian entries and:*

$$m \geq \tilde{c} \delta^{-2} (\lambda + h_{\hat{\epsilon}}) \max((\log(\lambda + h_{\hat{\epsilon}}) \log(n))^2, \lambda + \log(2)) \quad (10.36)$$

where $h_{\hat{\epsilon}} = H_{\hat{\epsilon}}(U) + 4 \log(2)$ and \tilde{c} is a universal constant. If $\hat{\epsilon} < \frac{\alpha}{7\beta\sigma} \delta$ the LHS of statement (10.35) holds with probability $\geq 1 - e^{-\lambda}$.

Proof. From [25, Theorem 4.1] we have that:

$$M(n, k, \gamma) = \max((\log k \cdot \log n)^2, \gamma) \quad (10.37)$$

and $p = 1$ in Lemma 5. We choose $\rho = \gamma =: \lambda + \log(2)$ (being suboptimal). \square

Since this choice ρ and γ is not necessarily optimal the logarithmic order in n might be improved. However, for fixed λ and sufficiently small $\hat{\epsilon}$ we have $m = \mathcal{O}(h_{\hat{\epsilon}}(\log h_{\hat{\epsilon}} \cdot \log n)^2)$ which is sufficient to preserve, for example, additive scaling (up to logarithms and large n) for the bilinear sparse models once $\hat{\epsilon}$ does not depend on n and where $h_{\hat{\epsilon}} = \mathcal{O}((s + f) \log n)$.

Stable Embedding of Bilinear Signal Sets: Finally, we come back now to the application for bilinear inverse problems with sparsity priors as discussed in the introduction. From the communication theoretic and signal processing point of view we will consider the problems (i) and (ii) on page 287 and we give the results for both cases in one theorem. Although we will summarize this for generic random measurements due to concentration as in Lemma 3, it follows from Lemma 6 that *the scaling even remains valid in a considerable de-randomized setting*. The assertion (i) in the next theorem was already given in [39]. Recall that $M_{s,f} \subseteq \mathbb{C}^{n \times n}$ are the (s,f) —sparse rank—one matrices as defined in (10.6).

Theorem 1. Set (i) $U = M_{s,f}$ and $\kappa = 1$ or (ii) $U = M_{s,f} - M_{s,f}$ and $\kappa = 2$ equipped with the Frobenius norm. Let be $B : U \rightarrow V \subseteq \mathbb{C}^n$ linear with property $(B_{\alpha,\beta})$ and $\|\cdot\|$ be a norm in V . If α, β do not depend on n , $\Phi \in \mathbb{C}^{m \times n}$ obeys $\Pr(\{\|\Phi r\| - \|r\| \leq \frac{\delta}{2} \|r\|\}) \geq 1 - e^{-c\delta^2 m}$ for each $r \in V$ and $m \geq c''\delta^{-2}(s+f)\log(n/(\kappa\min(s,f)))$ it follows that:

$$\Pr(\{\forall v \in V : \|\Phi v\| - \|v\| \leq \delta \|v\|\}) \geq 1 - e^{-c'm} \quad (10.38)$$

where $c', c'' > 0$ (only depending on δ).

Proof. In both cases U has property (A_σ) with $\sigma = \sqrt{2}$. Fix exemplary $\hat{\varepsilon} := \frac{\alpha}{8\beta\sigma}\delta < \frac{\alpha}{7\beta\sigma}\delta$ for Lemma 3. From (10.25) we have in both cases (i) and (ii):

$$\begin{aligned} H_{\hat{\varepsilon}}(U) &\leq \left(s+f+\frac{1}{2}\right)4\kappa \log \frac{9}{\hat{\varepsilon}} + \kappa(s+f) \log \frac{n}{\kappa\min(s,f)} \\ &= \left(s+f+\frac{1}{2}\right)4\kappa \log \frac{8 \cdot 9\sigma\beta}{\alpha\delta} + \kappa(s+f) \log \frac{n}{\kappa\min(s,f)} =: h_\delta \end{aligned} \quad (10.39)$$

where (α, β) are the bounds for B in (10.17) and independent of n . Let $\gamma = c\delta^2 m$ for some $c > 0$. We have from Lemma 3:

$$\Pr(\{\forall v \in B(U) : \|\Phi v\| - \|v\| \leq \delta \|v\|\}) \geq 1 - e^{-(c\delta^2 m - h_\delta)}. \quad (10.40)$$

To achieve exponential probability of the form $\geq 1 - \exp(-c'm)$ we have to ensure a constant $c' > 0$ such that $c\delta^2 m - h_\delta \geq c'm$. In other words $\delta^2(c - \frac{\delta^{-2}h_\delta}{m}) \geq c' > 0$ meaning that there must be a constant c'' such that the number of measurements fulfills $m \geq c''\delta^{-2}(s+f)\log(n/(\kappa\min(s,f)))$. \square

Final Remarks on Recovery: In this section we have solely discussed embeddings. Hence, it is not at all clear that one can achieve recovery in the $s+f$ -regime even at moderate complexity. A negative result has been shown here already in [30] for multi-objective convex programs which are restricted to the Pareto boundary caused by the individual objectives. On the other hand, greedy algorithms or

alternating minimization algorithms like the “sparse power factorization” method [28] seem to be capable to operate in the desired regime once the algorithm is optimally initialized.

10.3 Sparse Convolutions and Stability

In this section, we will consider the central condition (10.17) for the special case where the bilinear mapping B refers to convolutions representing, for example, basic single-dispersive communication channels. Let us start with the case where $B(x, y) = x \circledast y$ is given as the *circular* convolution in \mathbb{C}^n . Denote with $k \ominus i$ the difference $k - i$ modulo n . Then this bilinear mapping is defined as:

$$(x \circledast y)_k = \sum_{i=0}^{n-1} x_i y_{k \ominus i} \quad \text{for all } k \in \{0, \dots, n-1\}. \quad (10.41)$$

Our analysis was originally motivated by the work in [22] where the authors considered circular convolutions with $x \in \Sigma_s$ and $y \in \Sigma_f$. We will show that under certain conditions circular convolutions fulfill property (10.17) for $U = \{x \otimes y : x \in X, y \in Y\}$ and suitable sets $X, Y \subset \mathbb{C}^{2n-1}$. In this case (10.17) reads as:

$$\alpha \|x\| \|y\| \leq \|x \circledast y\| \leq \beta \|x\| \|y\| \quad \text{for all } (x, y) \in X \times Y, \quad (10.42)$$

where from now on $\|x\| := \|x\|_2$ will always denote the ℓ_2 -norm. According to [39] we call this condition *restricted norm multiplicativity property* (RNMP). As already pointed out in the previous section, this condition ensures compression for the models (10.26) and (10.27) as summarized in Theorem 1. In fact, (10.42) follows as a special case of sparse convolutions, if one restricts the support of x and y to the first n entries. In this case circular convolution (10.41) equals (ordinary) convolution which is defined on \mathbb{Z} element-wise for absolute-summable $x, y \in \ell_1(\mathbb{Z})$ by

$$(x * y)_k = \sum_{i \in \mathbb{Z}} x_i y_{k-i} \quad \text{for all } k \in \mathbb{Z}. \quad (10.43)$$

Obviously, the famous Young inequality states for $1/p + 1/q - 1/r = 1$ and $1 \leq p, q, r \leq \infty$ that:

$$\|x * y\|_r \leq \|x\|_p \|y\|_q \quad (10.44)$$

and implies sub-multiplicativity of convolutions in ℓ_1 but a reverse inequality was only known for positive signals. However, the same is true when considering *s-sparse sequences* $\Sigma_s = \Sigma_s(\mathbb{Z})$ as we will show in Theorem 2. Moreover, the lower bound α in (10.42) depends solely on the sparsity levels of the signals and not on the support location.

Let us define $[n] := \{0, \dots, n-1\}$ and the set of subsets with cardinality s by $[n]_s := \{T \subset [n] \mid |T| = s\}$. For any $T \in [n]_s$ the matrix B_T denotes the $s \times s$ principal submatrix of B with rows and columns in T . Further, we denote by B_t an $n \times n$ –*Hermitian Toeplitz matrix* generated by $t \in \Sigma_s^n$ with symbol given for $\omega \in [0, 2\pi)$ by

$$b(t, \omega) = \sum_{k=-n+1}^{n-1} b_k(t) e^{ik\omega}, \quad (10.45)$$

which for $b_k(t) := (t * \bar{t^-})_k$ defines a *positive trigonometric polynomial of order not larger than n* by the FEJÉR-RIESZ factorization. Note, b_k are the samples of the auto-correlation of t which can be written as the convolution of $t = \{t_k\}_{k \in \mathbb{Z}}$ with the complex-conjugation of the time reversal t^- , given component-wise by $t_k^- = t_{-k}$. We will define the minimum of all k -restricted determinants of B by:

$$D_{n,k} := \min\{|\det(B_t)| : t \in \Sigma_k^n, \|t\| = 1\} \quad (10.46)$$

which exists by compactness arguments.

10.3.1 The RNMP for Sparse Convolutions

The following theorem is a generalization of a result in [42], (i) in the sense of the extension to infinite sequences on \mathbb{Z} (ii) extension to the complex case, which actually only replaces SZEGÖ factorization with FEJÉR-RIESZ factorization in the proof and (iii) with a precise determination of the dimension parameter n^4 .

Theorem 2. *For $s, f \in \mathbb{N}$ exist constants $0 < \alpha(s, f) \leq \beta(s, f) < \infty$ such that for all $x \in \Sigma_s$ and $y \in \Sigma_f$ it holds:*

$$\alpha(s, f) \|x\| \|y\| \leq \|x * y\| \leq \beta(s, f) \|x\| \|y\|, \quad (10.47)$$

where $\beta^2(s, f) = \min\{s, f\}$. Moreover, the lower bound only depends on the sparsity levels s and f of the sequences and can be lower bounded by

$$\alpha^2(s, f) \geq \frac{1}{\sqrt{n \cdot \min(s, f)^{n-1}}} \cdot D_{n, \min(s, f)}, \quad (10.48)$$

with $n = \lfloor 2^{2(s+f-2) \log(s+f-2)} \rfloor$. This bound is decreasing in s and f . For $\beta(s, f) = 1$ it follows that $\alpha(s, f) = 1$.

⁴Actually, the estimate of the dimension $n = \bar{n}$ of the constant $\alpha_{\bar{n}}$ in [42] was quite too optimistic.

The main assertion of the theorem is: The smallest ℓ^2 -norm over all convolutions of s - and f -sparse normalized sequences can be determined solely in terms of s and f , where we used the fact that the sparse convolution can be represented by sparse vectors in $n = \lfloor 2^{2(s+f-2)\log(s+f-2)} \rfloor$ dimensions, due to an additive combinatoric result. An analytic lower bound for α , which decays exponentially in the sparsity, has been found very recently in [41]. Although $D_{n,\min\{s,f\}}$ is decreasing in n (since we extend the minimum to a larger set by increasing n) nothing seems to be known on the precise scaling in n . Nevertheless, since n depends solely on s and f it is sufficient to ensure that $D_{n,\min\{s,f\}}$ is non-zero.

Proof. The upper bound is trivial and follows, for example, from the Young inequality (10.44) for $r = q = 2$ and $p = 1$ and with the Cauchy–Schwartz inequality, i.e., in the case $s \leq f$ this yields:

$$\|x * y\|_2 \leq \|x\|_1 \|y\|_2 \leq \sqrt{s} \|x\|_2 \|y\|_2. \quad (10.49)$$

For $x = 0$ or $y = 0$ the inequality is trivial as well, hence we assume that x and y are non-zero. We consider therefore the following problem:

$$\inf_{\substack{(x,y) \in (\Sigma_s, \Sigma_f) \\ x \neq 0 \neq y}} \frac{\|x * y\|}{\|x\| \|y\|} = \inf_{\substack{(x,y) \in (\Sigma_s, \Sigma_f) \\ \|x\| = \|y\| = 1}} \|x * y\|. \quad (10.50)$$

Such *bi-quadratic optimization problems* are known to be NP-hard in general [29]. According to (10.43) the squared norm can be written as:

$$\|x * y\|^2 = \sum_{k \in \mathbb{Z}} \left| \sum_{i \in \mathbb{Z}} x_i y_{k-i} \right|^2. \quad (10.51)$$

Take sets $I, J \subset \mathbb{Z}$ such that $\text{supp}(x) \subseteq I$ and $\text{supp}(y) \subseteq J$ with $|I| = s, |J| = f$ and let $I = \{i_0, \dots, i_{s-1}\}$ and $J = \{j_0, \dots, j_{f-1}\}$ (ordered sets). Thus, we represent x and y by complex vectors $u \in \mathbb{C}^s$ and $v \in \mathbb{C}^f$ component-wise, i.e., for all $i, j \in \mathbb{Z}$:

$$x_i = \sum_{\theta=0}^{s-1} u_{\theta} \delta_{i,i_{\theta}} \quad \text{and} \quad y_j = \sum_{\gamma=0}^{f-1} v_{\gamma} \delta_{j,j_{\gamma}}. \quad (10.52)$$

Inserting this representation in (10.51) yields:

$$\|x * y\|^2 = \sum_{k \in \mathbb{Z}} \left| \sum_{i \in \mathbb{Z}} \left(\sum_{\theta=0}^{s-1} u_{\theta} \delta_{i,i_{\theta}} \right) \left(\sum_{\gamma=0}^{f-1} v_{\gamma} \delta_{k-i,j_{\gamma}} \right) \right|^2 \quad (10.53)$$

$$= \sum_{k \in \mathbb{Z}} \left| \sum_{\theta=0}^{s-1} \sum_{\gamma=0}^{f-1} \sum_{i \in \mathbb{Z}} u_{\theta} \delta_{i,i_{\theta}} v_{\gamma} \delta_{k-i,j_{\gamma}+i} \right|^2. \quad (10.54)$$

Since the inner i -sum is over \mathbb{Z} , we can shift I by i_0 if we set $i \rightarrow i + i_0$ (note that $x \neq 0$), without changing the value of the sum:

$$= \sum_{k \in \mathbb{Z}} \left| \sum_{\theta} \sum_{\gamma} \sum_{i \in \mathbb{Z}} u_{\theta} \delta_{i+i_0, i_{\theta}} v_{\gamma} \delta_{k, j_{\gamma} + i + i_0} \right|^2. \quad (10.55)$$

By the same argument we can shift J by j_0 by setting $k \rightarrow k + i_0 + j_0$ and get:

$$= \sum_{k \in \mathbb{Z}} \left| \sum_{\theta} \sum_{\gamma} \sum_{i \in \mathbb{Z}} u_{\theta} \delta_{i, i_{\theta} - i_0} v_{\gamma} \delta_{k, j_{\gamma} - j_0 + i} \right|^2. \quad (10.56)$$

Therefore we always can assume that the supports $I, J \subset \mathbb{Z}$ fulfill $i_0 = j_0 = 0$ in (10.50). From (10.54) we get:

$$= \sum_{k \in \mathbb{Z}} \left| \sum_{\theta} \sum_{\gamma} u_{\theta} v_{\gamma} \delta_{k, j_{\gamma} + i_{\theta}} \right|^2 \quad (10.57)$$

$$= \sum_{k \in \mathbb{Z}} \sum_{\theta, \theta'} \sum_{\gamma, \gamma'} u_{\theta} \overline{u_{\theta'}} v_{\gamma} \overline{v_{\gamma'}} \delta_{k, j_{\gamma} + i_{\theta}} \delta_{k, j_{\gamma'} + i_{\theta'}} \quad (10.58)$$

$$= \sum_{\theta, \theta'} \sum_{\gamma, \gamma'} u_{\theta} \overline{u_{\theta'}} v_{\gamma} \overline{v_{\gamma'}} \delta_{i_{\theta} + j_{\gamma}, j_{\gamma'} + i_{\theta'}}. \quad (10.59)$$

The interesting question is now: *What is the smallest dimension n to represent this fourth order tensor $\delta_{i_{\theta} + j_{\gamma}, i_{\theta'} + j_{\gamma'}}$, i.e. representing the additive structure?* Let us consider an “index remapping” $\phi : A \rightarrow \mathbb{Z}$ of the indices $A \subset \mathbb{Z}$. Such a map ϕ which preserves additive structure:

$$a_1 + a_2 = a'_1 + a'_2 \Rightarrow \phi(a_1) + \phi(a_2) = \phi(a'_1) + \phi(a'_2) \quad (10.60)$$

for all $a_1, a_2, a'_1, a'_2 \in A$ is called a *Freiman homomorphism* on A of order 2 and a *Freiman isomorphism* if:

$$a_1 + a_2 = a'_1 + a'_2 \Leftrightarrow \phi(a_1) + \phi(a_2) = \phi(a'_1) + \phi(a'_2) \quad (10.61)$$

for all $a_1, a_2, a'_1, a'_2 \in A$, see, e.g., [21, 36]. For $A := I \cup J$ the property (10.61) gives exactly our desired indices $\phi(I)$ and $\phi(J)$ and we have to determine $n = n(s, f)$ such that $\phi(A) \subset [n]$. The minimization problem reduces then to an n -dimensional problem. Indeed, this was a conjecture in [24] and very recently proved in [21, Theorem 20.10] for sets with Freiman dimension $d = 1$. Fortunately, he could prove a more general compression argument for arbitrary sum sets in a torsion-free abelian group G having a finite Freiman dimension d . We will state here a restricted version of his result for the 1-dimensional group $G = (\mathbb{Z}, +)$ and $A_1 = A_2 = A$:

Lemma 7. *Let $A \subset \mathbb{Z}$ be a set containing zero with $m := |A| < \infty$ and Freiman dimension $d = \dim^+(A + A)$. Then there exists a Freiman isomorphism $\phi : A \rightarrow \mathbb{Z}$ of order 2 such that*

$$\text{diam}(\phi(A)) \leq d!^2 \left(\frac{3}{2}\right)^{d-1} 2^{m-2} + \frac{3^{d-1} - 1}{2}. \quad (10.62)$$

We here use the definition of a Freiman isomorphism according to ([21], p.299) which is a more generalized version as in [36]. In fact, $\phi : A \rightarrow \mathbb{Z}$ can be easily extended to $\phi' : A + A \rightarrow \mathbb{Z}$ by setting $\phi'(a_1 + a_2) = \phi(a_1) + \phi(a_2)$. Then Gryniewicz defines the map ϕ' to be a Freiman homomorphism, if $\phi'(a_1 + a_2) = \phi'(a_1) + \phi'(a_2)$ for all $a_1, a_2 \in A$. If ϕ' is also injective, then it holds

$$\phi'(a_1) + \phi'(a_2) = \phi'(a'_1) + \phi'(a'_2) \Leftrightarrow a_1 + a_2 = a'_1 + a'_2. \quad (10.63)$$

Since $0 \in A$ we have for every $a \in A$ that $\phi'(a + 0) = \phi(a) + \phi(0)$ and therefore (10.63) is equivalent to our definition (10.61). Furthermore, we have $\text{diam}(\phi'(A)) = \text{diam}(\phi(A) + \phi(0)) = \text{diam}(\phi(A)) = \max \phi(A) - \min \phi(A)$.

We continue with the proof of the theorem by taking $A = I \cup J$. Recall that there always exists sets $I, J \subset G$ with $|I| = s$ and $|J| = f$ containing the support of x resp. y . Since $0 \in I \cap J$ we always have $m = |A| \leq s + f - 1$. Unfortunately, the Freiman dimension can be much larger than the linear dimension of the ambient group \mathbb{Z} . But we can bound d for any $A \subset \mathbb{Z}$ by a result⁵ of Tao and Vu in [36, Corollary 5.42] by

$$\min\{|A + A|, |A - A|\} \leq \frac{|A|^2}{2} - \frac{|A|}{2} + 1 \leq (d + 1)|A| - \frac{d(d + 1)}{2} \quad (10.64)$$

where the smallest possible d is given by $d = |A| - 2$. Hence we can assume $d \leq m - 2$ in (10.62). By using the bound $\log(d!) \leq ((d + 1) \ln(d + 1) - d) / \ln 2$, we get the following upper bound:

$$\text{diam}(\phi(A)) < d!^2 \left(\frac{3}{2}\right)^{m-3} \cdot 2^{m-2} + \frac{3^{m-3}}{2} = (2(d!)^2 + 2^{-1})3^{m-3} \quad (10.65)$$

$$< (2^{[2(m-1)\log(m-1)\ln 2 - 2(m-2)]/\ln 2 + 1} + 2^{-1})3^{m-3} \quad (10.66)$$

using $3 < 2^2$ and $2/\ln 2 > 2$ we get

$$< 2^{2(m-1)\log(m-1) - 2(m-2) + 1 + 2(m-3)} + 2^{2(m-3)-1} \quad (10.67)$$

$$= 2^{2(m-1)\log(m-1)-1} + 2^{2(m-3)-1} \quad (10.68)$$

⁵Note that the Freiman dimension of order 2 in [36] is defined by $\dim(A) := \dim^+(A + A) - 1 = d - 1$.

$$< \lfloor 2^{2(m-1)\log(m-1)-1} + 2^{2(m-1)-1} \rfloor - 1 \quad (10.69)$$

$$< \lfloor 2^{2(m-1)[\log(m-1)-1]} + 2^{2(m-1)\log(m-1)-1} \rfloor - 1 \quad (10.70)$$

$$= \lfloor 2^{2(s+f-2)\log(s+f-2)} \rfloor - 1. \quad (10.71)$$

We translate ϕ by $a^* := \min \phi(A)$, i.e. $\phi' = \phi - a^*$ still satisfying (10.61). Abbreviate $\tilde{I} = \phi'(I)$ and $\tilde{J} = \phi'(J)$. From (10.62) we have with $n = \lfloor 2^{2(s+f-2)\log(s+f-2)} \rfloor$:

$$0 \in \tilde{I} \cup \tilde{J} \subset \{0, 1, 2, \dots, n-1\} = [n]. \quad (10.72)$$

and by (10.60) for all $\theta, \theta' \in [s]$ and $\gamma, \gamma' \in [f]$ we have the identity

$$\delta_{i_\theta + j_\gamma, i_{\theta'} + j_{\gamma'}} = \delta_{\tilde{i}_\theta + \tilde{j}_\gamma, \tilde{i}_{\theta'} + \tilde{j}_{\gamma'}}. \quad (10.73)$$

Although a Freiman isomorphism does not necessarily preserve the index order, this is not important for the norm in (10.59). We define the embedding of u, v into \mathbb{C}^n by setting for all $i, j \in [n]$:

$$\tilde{x}_i = \sum_{\theta=0}^{s-1} u_\theta \delta_{i, \tilde{i}_\theta} \quad \text{and} \quad \tilde{y}_j = \sum_{\gamma=0}^{f-1} v_\gamma \delta_{j, \tilde{j}_\gamma}. \quad (10.74)$$

Let us further set $\tilde{x}_i = \tilde{y}_i = 0$ for $i \in \mathbb{Z} \setminus [n]$. Then we get from (10.59):

$$\|x * y\|^2 = \sum_{\theta, \theta'} \sum_{\gamma, \gamma'} u_\theta \overline{u_{\theta'}} v_\gamma \overline{v_{\gamma'}} \delta_{i_\theta + j_\gamma, i_{\theta'} + j_{\gamma'}} \quad (10.75)$$

$$(10.73) \rightarrow = \sum_{\theta, \theta'} \sum_{\gamma, \gamma'} u_\theta \overline{u_{\theta'}} v_\gamma \overline{v_{\gamma'}} \delta_{\tilde{i}_\theta + \tilde{j}_\gamma, \tilde{i}_{\theta'} + \tilde{j}_{\gamma'}}. \quad (10.76)$$

Going analog backwards as in (10.59) to (10.53) we get

$$= \sum_{k \in \mathbb{Z}} \left| \sum_{i \in \mathbb{Z}} \left(\sum_{\theta=0}^{s-1} u_\theta \delta_{i, \tilde{i}_\theta} \right) \left(\sum_{\gamma=0}^{f-1} v_\gamma \delta_{k-i, \tilde{j}_\gamma} \right) \right|^2 \quad (10.77)$$

$$(10.74) \rightarrow = \sum_{k \in \mathbb{Z}} \left| \sum_{i \in \mathbb{Z}} \tilde{x}_i \tilde{y}_{k-i} \right|^2 = \|\tilde{x} * \tilde{y}\|^2. \quad (10.78)$$

Furthermore, we can rewrite the Norm by using the support properties of \tilde{x}, \tilde{y} as

$$\|\tilde{x} * \tilde{y}\|^2 = \sum_{i, i'=0}^{n-1} \tilde{x}_i \overline{\tilde{x}_{i'}} \sum_{k \in \mathbb{Z}} \tilde{y}_{k-i} \overline{\tilde{y}_{k-i'}} \quad (10.79)$$

and substituting by $k' = k - i$ we get

$$= \sum_{i,i'=0}^{n-1} \tilde{x}_i \overline{\tilde{x}_{i'}} \sum_{k'=\max\{0,i-i'\}}^{\min\{n-1,n-1-(i-i')\}} \tilde{y}_{k'} \overline{\tilde{y}_{k'+(i-i')}} = \langle \tilde{x}, B_{\tilde{y}} \tilde{x} \rangle, \quad (10.80)$$

where $B_{\tilde{y}}$ is an $n \times n$ Hermitian Toeplitz matrix with first row $(B_{\tilde{y}})_{0,k} = \sum_{j=0}^{n-k} \tilde{y}_j \tilde{y}_{j+k} =: b_k(\tilde{y}) = (\tilde{y} * \tilde{y}^-)_k$ resp. first column $(B_{\tilde{y}})_{k,0} =: b_{-k}(\tilde{y})$ for $k \in [n]$. Its symbol $b(\tilde{y}, \omega)$ is given by (10.45) and since $b_0 = \|\tilde{y}\| = 1$ it is for each $\tilde{y} \in \mathbb{C}^n$ a normalized trigonometric polynomial of order $n-1$. Minimizing the inner product in (10.80) over $\tilde{x} \in \Sigma_s^n$ with $\|\tilde{x}\| = 1$ includes all possible \tilde{I} and therefore establishes a *lower bound* (see the remarks after the proof). However, this then means to minimize the minimal eigenvalue λ_{\min} over all $s \times s$ principal submatrices of $B_{\tilde{y}}$:

$$\lambda_{\min}(B_{\tilde{y}}, s) := \min_{\tilde{x} \in \Sigma_s^n, \|\tilde{x}\| = 1} \langle \tilde{x}, B_{\tilde{y}} \tilde{x} \rangle \geq \lambda_{\min}(B_{\tilde{y}}) \quad (10.81)$$

whereby $\lambda_{\min}(B_{\tilde{y}}, s)$ is sometimes called the s -restricted eigenvalue (singular value) of $B_{\tilde{y}}$, see [34] or [27]. First, we show now that $\lambda_{\min}(B_{\tilde{y}}) > 0$. By the well-known Fejér-Riesz factorization, see, e.g., [16, Thm.3], the symbol of $B_{\tilde{y}}$ is *non-negative*⁶ for every $\tilde{y} \in \mathbb{C}^n$. By [7, (10.2)] it follows therefore that *strictly* $\lambda_{\min}(B_{\tilde{y}}) > 0$. Obviously, then also the determinant is non-zero. Hence $B_{\tilde{y}}$ is invertible and with $\lambda_{\min}(B_{\tilde{y}}) = 1/\|B_{\tilde{y}}^{-1}\|$ we can estimate the smallest eigenvalue (singular value) by the determinant ([7], Thm. 4.2):

$$\lambda_{\min}(B_{\tilde{y}}) \geq |\det(B_{\tilde{y}})| \frac{1}{\sqrt{n} (\sum_k |b_k(\tilde{y})|^2)^{(n-1)/2}} \quad (10.82)$$

whereby from $\|\tilde{y}\| = 1$ and the upper bound of the theorem or directly (10.49) it follows also that $\sum_k |b_k(\tilde{y})|^2 = \|\tilde{y} * \tilde{y}^-\|^2 \leq f$ if $\tilde{y} \in \Sigma_f^n$. Since the determinant is a continuous function in \tilde{y} over a compact set, the non-zero minimum is attained. Minimizing (10.82) over all sparse vectors \tilde{y} with smallest sparsity yields

$$\min_{\substack{\tilde{y} \in \Sigma_{\min\{s,f\}}^n \\ \|\tilde{y}\|=1}} \lambda_{\min}(B_{\tilde{y}}) \geq \sqrt{\frac{1}{nf^{n-1}}} \cdot \underbrace{\min_{t \in \Sigma_{\min\{s,f\}}^n, \|t\|=1} |\det(B_t)|}_{=D_{n,\min\{s,f\}}} > 0 \quad (10.83)$$

which shows the claim of the theorem. \square

⁶Note, there exist $\tilde{y} \in \mathbb{C}^n$ with $\|\tilde{y}\| = 1$ and $b(\tilde{y}, \omega) = 0$ for some $\omega \in [0, 2\pi)$. That is the reason why things are more complicated here. Moreover, we want to find a universal lower bound over all \tilde{y} , which is equivalent to a universal lower bound over all non-negative trigonometric polynomials of order $n-1$.

It is important to add here that the compression via the Freiman isomorphism $\phi : I \cup J \rightarrow [n]$ is obviously not global and depends on the support sets I and J . From numerical point of view one might therefore proceed only with the first assertion in (10.81) and evaluate the particular intermediate steps:

$$\begin{aligned}
\inf_{\substack{(x,y) \in (\Sigma_s, \Sigma_f) \\ \|x\| = \|y\| = 1}} \|x * y\|^2 &= \min_{\substack{(\tilde{x}, \tilde{y}) \in (\Sigma_s^n, \Sigma_f^n) \\ \|\tilde{x}\| = \|\tilde{y}\| = 1}} \|\tilde{x} * \tilde{y}\|^2 \\
&= \min \left\{ \min_{\substack{\tilde{I} \in [n]_s \\ \|\tilde{y}\| = 1}} \min_{\tilde{y} \in \Sigma_f^n} \lambda_{\min}(B_{\tilde{I}, \tilde{y}}), \min_{\substack{\tilde{J} \in [n]_f \\ \|\tilde{x}\| = 1}} \min_{\tilde{x} \in \Sigma_s^n} \lambda_{\min}(B_{\tilde{J}, \tilde{x}}) \right\} \\
&\geq \min_{T \in [n]_{\max\{s, f\}}} \min_{t \in \Sigma_{\min\{s, f\}}^n, \|t\| = 1} \lambda_{\min}(B_{T, t}) \\
&\geq \min_{\substack{t \in \Sigma_{\min\{s, f\}}^n \\ \|t\| = 1}} \lambda_{\min}(B_t) \geq \min_{\substack{t \in \mathbb{C}^n \\ \|t\| = 1}} \lambda_{\min}(B_t).
\end{aligned} \tag{10.84}$$

The first equality holds, since any support configuration in $\Sigma_s^n \times \Sigma_f^n$ is also realized by sequences in $\Sigma_s \times \Sigma_f$. The bounds in (10.85) can be used for numerical computation attempts.

Let us now summarize the implications for the RNMP of zero-padded sparse circular convolutions as defined in (10.41). Therefore we denote the zero-padded elements by $\Sigma_s^{n,n-1} := \{x \in \mathbb{C}^{2n-1} \mid \text{supp}(x) \in [n]_s\}$, for which the circular convolution (10.41) equals the ordinary convolution (10.43) restricted to $[2n-1]$. Hence, the bounds in Theorem 2 will be valid also in (10.42) for $X = \Sigma_s^{n,n-1}$ and $Y = \Sigma_f^{n,n-1}$.

Corollary 2. *For $s, f \leq n$ and all $(x, y) \in \Sigma_s^{n,n-1} \times \Sigma_f^{n,n-1}$ it holds:*

$$\alpha(s, f, n) \|x\| \|y\| \leq \|x \circledast y\| \leq \beta(s, f) \|x\| \|y\|. \tag{10.85}$$

Moreover, we have $\beta^2(s, f) = \min\{s, f\}$ and with $\tilde{n} = \min\{n, 2^{2(s+f-2)\log(s+f-2)}\}$:

$$\alpha^2(s, f, \tilde{n}) \geq \frac{1}{\sqrt{\tilde{n} \cdot \min(s, f)^{\tilde{n}-1}}} \cdot D_{\tilde{n}, \min(s, f)}, \tag{10.86}$$

which is a decreasing sequence in s and f . For $\beta(s, f) = 1$ we get equality with $\alpha(s, f) = 1$.

Proof. Since $x \in \Sigma_s^{n,n-1}$ and $y \in \Sigma_f^{n,n-1}$ we have

$$\|x \circledast y\|_{\ell^2([2n-1])} = \|x * y\|_{\ell^2([-n, n])}. \tag{10.87}$$

Hence, x, y can be embedded in Σ_s resp. Σ_f without changing the norms. If $n \geq \lfloor 2^{2(s+f-2)\log(s+f-2)} \rfloor =: \tilde{n}$, then we can find a Freiman isomorphism which expresses the convolution by vectors $\tilde{x}, \tilde{y} \in \mathbb{C}^{\tilde{n}}$. If $n \leq \tilde{n}$, there is no need to compress the convolution and we can set easily $\tilde{n} = n$. Hence, all involved Hermitian Toeplitz matrices B_t in (10.81) are $\tilde{n} \times \tilde{n}$ matrices and we just have to replace n by \tilde{n} in (10.48).

10.3.2 Implications for Phase Retrieval

In this section we will discuss an interesting application of the RNMP result in Theorem 2 and in particular we will exploit here the version presented in Corollary 2. We start with a bilinear map $B(x, y)$ which is *symmetric*, i.e., $B(x, y) = B(y, x)$ and let us denote its diagonal part by $A(x) := B(x, x)$. Already in (10.27) we mentioned *quadratic inverse problems* where $x \in \Sigma_s$ and there we argued that, due the binomial-type formula:

$$A(x_1) - A(x_2) = B(x_1 - x_2, x_1 + x_2) \quad (10.88)$$

different x_1 and x_2 can be (stable) distinguished modulo global sign on the basis of $A(x_1)$ and $A(x_2)$ whenever $B(x_1 - x_2, x_1 + x_2)$ is well-separated from zero. In the sparse case $x_1, x_2 \in \Sigma_s$ this assertion is precisely given by property (10.17) when lifting B to a linear map operating on the set $U = M_{2s, 2s}$ of rank-one matrices with at most $2s$ non-zero rows and columns (see again (10.27)). In such rank-one cases we call this as the RNMP condition and for sparse convolutions (being symmetric) we have shown in the previous Section 10.3.1 that this condition is fulfilled independent of the ambient dimension. As shown in Corollary 2 this statement translates to zero-padded circular convolutions. Hence, combining (10.88) with Corollary 2 and Theorem 1 asserts that each zero-padded s -sparse x can be stable recovered modulo global sign from $\mathcal{O}(s \log n)$ randomized samples of its circular auto-convolution (which itself is at most s^2 -sparse).

However, here we discuss now another important application for the *phase retrieval problem* and these implications will be presented also in [40]. The relation to the quadratic problems above is as follows: Let us define from the (symmetric) circular convolution \circledast the (sesquilinear) *circular correlation*:

$$x \circledast y := x \circledast \Gamma \bar{y} = F^*(Fx \odot \bar{Fy}) \quad (10.89)$$

where $(u \odot v)_k := u_k v_k$ denotes the Hadamard (point-wise) product, $(F)_{k,l} = n^{-\frac{1}{2}} e^{i2\pi kl/n}$ is the unitary Fourier matrix (here on \mathbb{C}^n), and $\Gamma := F^2 = F^{*2}$ is the time reversal (an involution). Whenever dimension is important we will indicate this by $F = F_n$ and $\Gamma = \Gamma_n$. Therefore, Fourier measurements on the circular auto-correlation $x \circledast x$ are intensity measurements on the Fourier transform of x :

$$F(x \circledast x) = |Fx|^2. \quad (10.90)$$

Recovering x from such intensity measurements is known as a phase retrieval problem, see, e.g., [3] and the references therein, which is without further support restrictions on x not possible [19]. Unfortunately, since the circular correlation in (10.89) is sesquilinear and not symmetric (10.88) does not hold in general. However, it will hold for structures which are consistent with a real-linear algebra, i.e. (10.88) symmetric for vectors with the property $x = \Gamma \bar{x}$ (if and only if and the same also for y). Hence, to enforce this symmetry and to apply our result, we perform a *symmetrization*. Let us consider two cases separately. First, assume that $x_0 = \bar{x}_0$ and define $\mathcal{S}: \mathbb{C}^n \rightarrow \mathbb{C}^{2n-1}$:

$$\mathcal{S}(x) := \underbrace{(x_0, x_1, \dots, x_{n-1})}_{=:x}^T, \quad \underbrace{(\bar{x}_{n-1}, \dots, \bar{x}_1)}_{=:x_-^\circ}^T. \quad (10.91)$$

Now, for $x_0 = \bar{x}_0$ the symmetry condition $\mathcal{S}(x) = \Gamma \overline{\mathcal{S}(x)}$ is fulfilled (note that here $\Gamma = \Gamma_{2n-1}$):

$$\mathcal{S}(x) = \left(\frac{x}{x_-^\circ} \right) = \Gamma \left(\frac{\bar{x}}{x_-^\circ} \right) = \Gamma \overline{\left(\frac{x}{x_-^\circ} \right)} = \Gamma \overline{\mathcal{S}(x)}. \quad (10.92)$$

Thus, for $x, y \in \mathbb{C}_0^n := \{x \in \mathbb{C}^n : x_0 = \bar{x}_0\}$, circular correlation of (conjugate) symmetrized vectors is symmetric and agrees with the circular convolution. Let us stress the fact that the symmetrization map is linear only for *real* vectors x since complex conjugation is involved. On the other hand, \mathcal{S} can obviously be written as a linear map on vectors like $(\text{Re}(x), \text{Im}(x))$ or (x, \bar{x}) .

Applying Corollary 2 to the *zero-padded symmetrization* (first zero padding $n \rightarrow 2n-1$, then symmetrization $2n-1 \rightarrow 4n-3$) $\mathcal{S}(x)$ for $x \in \Sigma_{0,n}^{n,n-1} := \Sigma_n^{n,n-1} \cap \mathbb{C}_0^{2n-1}$ we get the following stability result.

Theorem 3. *Let $n \in \mathbb{N}$, then $4n-3$ absolute-square Fourier measurements of zero-padded symmetrized vectors in \mathbb{C}^{4n-3} are stable up to a global sign for $x \in \Sigma_{0,n}^{n,n-1}$, i.e., for all $x_1, x_2 \in \Sigma_{0,n}^{n,n-1}$ it holds*

$$\| |F\mathcal{S}(x_1)|^2 - |F\mathcal{S}(x_2)|^2 \| \geq c \|\mathcal{S}(x_1 - x_2)\| \|\mathcal{S}(x_1 + x_2)\| \quad (10.93)$$

with $c = c(n) = \alpha(n, n, 4n-3) / \sqrt{4n-3} > 0$ and $F = F_{4n-3}$.

Remark 1. Note that we have

$$2\|x\|^2 \geq \|\mathcal{S}(x)\|^2 = \|x\|^2 + \|x_-^\circ\|^2 \geq \|x\|^2. \quad (10.94)$$

Thus, $\mathcal{S}(x) = 0$ if and only if $x = 0$ and the stability in distinguishing x_1 and x_2 up to a global sign follows from the RHS of (10.93) and reads explicitly as:

$$\| |F\mathcal{S}(x_1)|^2 - |F\mathcal{S}(x_2)|^2 \| \geq c \|x_1 - x_2\| \|x_1 + x_2\|. \quad (10.95)$$

Unfortunately, s -sparsity of x does not help in this context to reduce the number of measurements, but at least can enhance the stability bound α to $\alpha(2s, 2s, 4n-3)$.

Proof. For zero-padded symmetrized vectors, auto-convolution agrees with auto-correlation and we get from (10.91) for $x \in \Sigma_{0,n}^{n,n-1}$:

$$F(A(x)) = F(\mathcal{S}(x) \circledast \mathcal{S}(x)) = \sqrt{4n-3} |F\mathcal{S}(x)|^2. \quad (10.96)$$

Putting things together we get for every $x \in \Sigma_{0,n}^{n,n-1}$:

$$\begin{aligned} \left\| |F\mathcal{S}(x_1)|^2 - |F\mathcal{S}(x_2)|^2 \right\| &= (4n-3)^{-1/2} \|F(A(x_1) - A(x_2))\| \\ F \text{ is unitary} \rightarrow &= (4n-3)^{-1/2} \|A(x_1) - A(x_2)\| \\ &\stackrel{(10.88)}{=} (4n-3)^{-1/2} \|\mathcal{S}(x_1 - x_2) \circledast \mathcal{S}(x_1 + x_2)\| \\ &\geq \frac{\alpha(n, n, 4n-3)}{\sqrt{4n-3}} \|\mathcal{S}(x_1 - x_2)\| \cdot \|\mathcal{S}(x_1 + x_2)\|. \end{aligned} \quad (10.97)$$

In the last step we use that Corollary 2 applies whenever the non-zero entries are contained in a cyclic block of length $2n-1$.

In the *real case* (10.93) is equivalent to a *stable linear embedding* in \mathbb{R}^{4n-3} up to a global sign (see here also [18] where the ℓ_1 -norm is used on the left side) and therefore this is an *explicit phase retrieval statement* for *real* signals. Recently, stable recovery also in the complex case up to a global phase from the same number of subgaussian measurements has been achieved in [17] using lifting as in (10.27). Both results hold with exponential high probability whereby our result is deterministic. Even more, the greedy algorithm in [17, Thm.3.1] applies in our setting once the signals obey sufficient decay in magnitude. But, since \mathcal{S} is *not complex-linear* Theorem 3 cannot directly be compared with the usual complex phase retrieval results. On the other hand, our approach indeed (almost) distinguishes complex phases by the Fourier measurements since symmetrization provides injectivity here up to a global sign. To get rid of the odd definition \mathbb{C}_0^n one can symmetrize (and zero padding) $x \in \mathbb{C}^n$ also by:

$$\mathcal{S}'(x) := (\underbrace{0, \dots, 0}_n, x_0, \dots, x_{n-1}, \bar{x}_{n-1}, \dots, \bar{x}_0, \underbrace{0, \dots, 0}_{n-1})^T \in \mathbb{C}^{4n-1} \quad (10.98)$$

again satisfying $\mathcal{S}'(x) = \Gamma_{4n-1} \overline{\mathcal{S}'(x)}$ at the price of two further dimensions.

Corollary 3. *Let $n \in \mathbb{N}$, then $4n-1$ absolute-square Fourier measurements of zero-padded and symmetrized vectors given by (10.98) are stable up to a global sign for $x \in \mathbb{C}^n$, i.e., for all $x_1, x_2 \in \mathbb{C}^n$ it holds*

$$\left\| |F\mathcal{S}'(x_1)|^2 - |F\mathcal{S}'(x_2)|^2 \right\| \geq 2c \|x_1 - x_2\| \|x_1 + x_2\| \quad (10.99)$$

with $c = c(n) = \alpha(n, n, 4n-1) / \sqrt{4n-1} > 0$ and $F = F_{4n-1}$.

The proof of it is along the same steps as in Theorem 3. The direct extension to sparse signals as in [39] seems to be difficult since randomly chosen Fourier samples do not provide a sufficient measure of concentration property without further randomization.

Acknowledgements The authors would like to thank the anonymous reviewers for their detailed and valuable comments. We also thank Holger Boche, David Gross, Richard Kueng, and Götz Pfander for their support and many helpful discussions. This work was supported by the *Deutsche Forschungsgemeinschaft (DFG)* under grant JU 2795/2-1.

References

1. Ahmed, A., Recht, B., Romberg, J.: Blind deconvolution using convex programming. arXiv:1211.5608v1 (2012)
2. Aldroubi, A., Chen, X., Powell, A.: Perturbations of measurement matrices and dictionaries in compressed sensing. *Appl. Comput. Harmon. Anal.* **33**(2), 282–291 (2012)
3. Bandeira A, Cahill J, Mixon D, Nelson A. Saving phase: Injectivity and stability for phase retrieval. *Appl. Comput. Harmon. Anal.* **37**(1):106–125 (2014)
4. Baraniuk, R., Wakin, M.: Random projections of smooth manifolds. *Found. Comput. Math.* **9**(1), 51–77 (2009)
5. Baraniuk, R., Davenport, M., DeVore, R.A., Wakin, M.: A simple proof of the restricted isometry property for random matrices. *Constr. Approx.* **28**(3), 253–263 (2008)
6. Bickel, P.J., Ritov, Y., Tsybakov, A.B.: Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.* **37**(4), 1705–1732 (2009)
7. Böttcher, A., Grudsky, S.M.: *Spectral Properties of Banded Toeplitz Matrices*. SIAM, Philadelphia (2005)
8. Cai, T.T., Zhang, A.: Sharp RIP bound for sparse signal and lowrank matrix recovery. *Appl. Comput. Harmon. Anal.* **35**(1), 74–93 (2013)
9. Candes, E., Plan, Y.: Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theory* **54**, 1–30 (2011)
10. Candes, E., Eldar, Y., Needell, D., Randall, P.: Compressed sensing with coherent and redundant dictionaries. *Appl. Comput. Harmon. Anal.* **31**(1), 59–73 (2010)
11. Chandrasekaran, V., Recht, B.: The convex geometry of linear inverse problems. *Found. Comput. Math.* **12**, 805–849 (2012)
12. Chi, Y., Scharf, L.: Sensitivity to basis mismatch in compressed sensing. *IEEE Trans. Signal Process.* **59**(5), 2182–2195 (2011)
13. Choudhary, S., Mitra, U.: Identifiability Scaling Laws in Bilinear Inverse Problems, pp. 1–32. arXiv:1402.2637v1 (2014)
14. Dhillon HS, Huang H, Viswanathan H, Valenzuela RA, Jul IT. Fundamentals of Throughput Maximization with Random Arrivals for M2M Communications. *Trans. Commun.* **62**(11):4094–4109 (2014)
15. Diestel, J., Grothendieck, A., Fourie, J., Swart, J.: *The Metric Theory of Tensor Products: Grothendieck's R'esum'e Revisited* (2008)
16. Dimitrov, D.K.: *Approximation Theory: A Volume Dedicated to Blagovest Sendov*. Marin Drinov Academic, Sofia (2004)
17. Ehler M, Fornasier M, Sigl J. Quasi-linear compressed sensing. *SIAM Multiscale Model. Simul.* **12**(2):725–754 (2014)
18. Eldar, Y., Mendelson, S.: Phase retrieval: stability and recovery guarantees. *Appl. Comput. Harmon. Anal.* **36**(3), 473–494 (2014)

19. Fienup, J.R.: Reconstruction of a complex-valued object from the modulus of its Fourier transform using a support constraint. *J. Opt. Soc. Am. A* **4**, 118–123 (1987)
20. Gleichman, S., Eldar, Y.: Blind compressed sensing. *IEEE Trans. Inf. Theory* **57**(10), 6958–6975 (2011)
21. Grynkiewicz, D.J.: Structural Additive Theory. *Developments in Mathematics*, vol. 30. Springer, New York (2013)
22. Hegde, C., Baraniuk, R.G.: Sampling and recovery of pulse streams. *IEEE Trans. Signal Process.* **59**.4, 1505–1517 (2011)
23. Herman, M., Strohmer, T.: General deviants: An analysis of perturbations in compressed sensing. *IEEE J. Sel. Top. Sign. Process.* **4**(2), 342–349 (2010)
24. Konyagin, S., Lev, V.: Combinatorics and linear algebra of Freiman’s Isomorphism. *Mathematika* **47**, 39–51 (2000)
25. Krahmer, F., Mendelson, S., Rauhut, H.: Suprema of chaos processes and the restricted isometry property. *arXiv:1207.0235v3* (2012)
26. Krahmer, F., Ward, R.: New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. *arXiv:1009.0744v4* (2011)
27. Kueng R, Gross D. RIPless compressed sensing from anisotropic measurements. *Linear Algebra Appl.* **44**:110–123 (2014)
28. Lee, K., Wu, Y., Bresler, Y.: Near optimal compressed sensing of sparse rank-one matrices via sparse power factorization. *arXiv:1312.0525v1* (2013)
29. Ling, C., Nie, J., Qi, L., Ye, Y.: Biquadratic optimization over unit spheres and semidefinite programming relaxations. *SIAM J. Optim.* **20**, 1286–1310 (2009)
30. Oymak, S., Jalali, A., Fazel, M., Eldar, Y., Hassibi, B.: Simultaneously structured models with application to sparse and low-rank matrices. *arXiv:1212.3753v2* (2012)
31. Rauhut, H.: Compressed sensing and redundant dictionaries. *IEEE Trans. Inf. Theory* **54**(5), 2210–2219 (2008)
32. Rauhut, H., Romberg, J., Tropp, J.A.: Restricted isometries for partial random circulant matrices. *Appl. Comput. Harm. Anal.* **32**(2), 242–254 (2012)
33. Romberg, J.: Compressive sensing by random convolution. *SIAM J. Imaging Sci.* **2**(4), 1098 (2009)
34. Rudelson, M., Zhou, S.: Reconstruction from anisotropic random measurements. Tech. rep. University of Michigan, Department of Statistics, Technical Report 522 (2011)
35. Ryan, R.: Introduction to Tensor Products of Banach Spaces, p. 239. Springer, New York (2002)
36. Tao, T., Vu, V.: Additive Combinatorics. Cambridge University Press, Cambridge (2006)
37. Tropp, J., Laska, J.: Beyond Nyquist: Efficient sampling of sparse bandlimited signals. *Trans. Inf. Theory* **56**(1), 520–544 (2010)
38. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. In: Eldar, Y., Kutyniok, G. (eds.) *Compressed Sensing, Theory and Applications*, Chap. 5. Cambridge University Press, Cambridge (2012)
39. Walk, P., Jung, P.: Compressed sensing on the image of bilinear maps. In: *IEEE International Symposium on Information Theory*, pp. 1291–1295 (2012)
40. Walk, P., Jung, P.: Stable recovery from the magnitude of symmetrized Fourier measurements. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2014)
41. Walk P, Jung P, Pfander G.E. On the Stability of Sparse Convolutions. preprint available at <http://arxiv.org/abs/1409.6874>
42. Walk, P., Jung, P.: On a reverse ℓ^2 -inequality for sparse circular convolutions. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4638–4642 (2013)
43. Wunder, G., Jung, P., Kasparick, M., Wild, T., Schaich, F., Chen, Y., Gaspar, I., Michailow, N., Festag, A., Fettweis, G., Cassiau, N., Ktenas, D., Dryjanski, M., Pietrzyk, S., Eged, B., Vago, P.: 5GNOW: non-orthogonal, asynchronous waveforms for future applications. *IEEE Commun. Mag.* **52**(2), 97–105 (2014)

Chapter 11

Cosparsity in Compressed Sensing

Maryia Kabanava and Holger Rauhut

Abstract Analysis ℓ_1 -recovery is a strategy of acquiring a signal, that is sparse in some transform domain, from incomplete observations. In this chapter we give an overview of the analysis sparsity model and present theoretical conditions that guarantee successful nonuniform and uniform recovery of signals from noisy measurements. We derive a bound on the number of Gaussian and subgaussian measurements by examining the provided theoretical guarantees under the additional assumption that the transform domain is generated by a frame, which means that there are just few nonzero inner products of a signal of interest with frame elements.

11.1 Introduction

As already outlined in this book (see in particular Chapter 1), compressed sensing aims at acquiring signals from undersampled and possibly corrupted measurements. In mathematical terms, the available data about a signal $x \in \mathbb{R}^n$ is given by a set of measurements

$$y = Ax + e, \quad (11.1)$$

where $A \in \mathbb{R}^{m \times n}$ with $m \ll n$ is the sensing matrix and $e \in \mathbb{R}^m$ represents a noise vector. Since this system is undetermined it is hopeless to recover x from y without additional information. The key idea is to take into account our prior knowledge about the structure of x .

The standard assumption in compressed sensing is that the signal is sparse in some orthonormal basis (as outlined in Chapter 1), which means that it can be represented as a linear combination of only few basis elements. This setting corresponds to the *synthesis* sparsity model. This chapter is concerned with a more general sparsity model, where one assumes that the signal is sparse after a possibly redundant transform, see [7, 17, 34, 37, 38, 53] for initial papers

M. Kabanava • H. Rauhut (✉)

RWTH Aachen University, Lehrstuhl C für Mathematik (Analysis),
Templergraben 55, 52062 Aachen, Germany

e-mail: kabanava@mathc.rwth-aachen.de; rauhut@mathc.rwth-aachen.de

on this subject. This analysis sparsity model—also called cosparsity model—leads to more flexibility in the modeling of sparse signals. Many sparse recovery methods can be adapted to this setting including convex relaxation leading to analysis ℓ_1 -minimization (see below) and greedy-like methods [22, 38]. Relevant analysis operators can be generated by the discrete Fourier transform, wavelet [35, 46, 50], curvelet [6], or Gabor transforms [25]. The popular method of total variation minimization [5, 11, 39, 49] corresponds to analysis with respect to a difference operator.

This chapter aims at giving an overview on the analysis sparsity model and its use in compressed sensing. We will present in particular recovery guarantees for ℓ_1 -minimization including versions of the null space property and the restricted isometry property. Moreover, we give estimates on the number of measurements required for (approximate) recovery of signals using random Gaussian and subgaussian measurements. Parts of these results (or their proofs) are new and have not appeared elsewhere in the literature yet.

We emphasize that recovery guarantees for subgaussian measurements in the analysis sparsity framework should only be seen as the starting point of a theory for cosparse recovery. Practically relevant measurement scenarios rather use structured random matrices [43] such as random partial Fourier [8, 41, 48] or partial random circulant matrices [31, 42, 44, 45]. Nevertheless, theoretical results for subgaussian measurement matrices are important because they provide optimal guarantees and give benchmarks for other measurement matrices. We leave it as an interesting open problem for future research to study in detail the recovery of cosparse signals from structured random measurements. For the special case of total variation minimization, a few results for random Fourier measurements are already available in [32, 39].

Notation: We use Ω_Λ to refer to a submatrix of Ω with the rows indexed by Λ (we should emphasize that our notation differs from the general notation of the book, where this is rather the submatrix corresponding to the columns of Ω); α_Λ stands for the vector whose entries indexed by Λ coincide with the entries of α and the rest are filled by zeros. On some occasions with a slight abuse of notation we refer to α_Λ as an element of $\mathbb{R}^{|\Lambda|}$. We use $[p]$ to denote the set of all natural numbers not exceeding p , i.e., $[p] = \{1, 2, \dots, p\}$. The sign of a real number $r \neq 0$ is $\text{sgn}(r) = \frac{r}{|r|}$. For a vector $\alpha \in \mathbb{R}^p$ we define its sign vector $\text{sgn}(\alpha) \in \mathbb{R}^p$ by

$$(\text{sgn}(\alpha))_i = \begin{cases} \frac{\alpha_i}{|\alpha_i|}, & \text{for all } i \text{ such that } \alpha_i \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The operator norm of a matrix A is given by $\|A\|_{2 \rightarrow 2} := \sup_{\|x\|_2 \leq 1} \|Ax\|_2$; A^T is the transpose of A . The orthogonal complement of a subspace $S \subset \mathbb{R}^p$ is denoted by S^\perp . The orthogonal projection on S is performed by the operator \mathcal{P}_S . Notation B_2^p stands for the unit ball with respect to the ℓ_2 -norm and S^{n-1} is the unit sphere in \mathbb{R}^n .

11.2 Analysis vs. synthesis sparsity

We recall from Chapter 1 that a vector $x \in \mathbb{R}^n$ is k -sparse, if $\|x\|_0 = |\{\ell : x_\ell \neq 0\}| \leq k$. For signals arising in applications it is more common to have sparse expansion in some basis or dictionary rather than being sparse themselves. This means that for a matrix $D \in \mathbb{R}^{n \times q}$, $q \geq n$, whose columns form a so-called dictionary of \mathbb{R}^n (a spanning set), x can be represented as

$$x = D\alpha, \quad \alpha \in \mathbb{R}^q,$$

where α is k -sparse. It is common to choose an orthogonal matrix $D \in \mathbb{R}^{n \times n}$, so that we have sparsity with respect to an orthonormal basis. However, also a redundant frame may be used. Here, we “synthesize” x from a few columns d_j of D , which is the reason why this is also called the synthesis sparsity model. The set of all k -sparse signals can be described as

$$\bigcup_{T \subset [n]: |T|=k} V_T = \bigcup_{T \subset [n]: |T|=k} \text{span}\{d_j : j \in T\}, \quad (11.2)$$

i.e., it is a union of k -dimensional subspaces which are generated by k columns of D .

The analysis sparsity model assumes that Ωx is (approximately) sparse, where $\Omega \in \mathbb{R}^{p \times n}$ is a so-called analysis operator. Denoting the rows of Ω by $\omega_j \in \mathbb{R}^n$, $j = 1, \dots, p$, the entries of Ωx are given as $\langle \omega_j, x \rangle$, $j = 1, \dots, p$, i.e., we analyze x by taking inner products with the ω_j . If Ωx is k -sparse, then x is called ℓ -cosparse, where the number $\ell := p - k$ is referred to as *cosparsity* of x . The index set of the zero entries of Ωx is called the *cosupport* of x . The motivation to work with the cosupport rather than the support in the context of analysis sparsity is that it is the location of the zero-elements which define a corresponding subspace. In fact, if Λ is the cosupport of x , then

$$\langle \omega_j, x \rangle = 0, \quad \text{for all } j \in \Lambda.$$

Hence, the set of ℓ -cosparse signals can be written as

$$\bigcup_{\Lambda \subset [p]: \#\Lambda = \ell} W_\Lambda, \quad (11.3)$$

where W_Λ denotes the orthogonal complement of the linear span of $\{\omega_j : j \in \Lambda\}$. In this sense, the analysis sparsity model falls into the larger class of union of subspaces model [2].

When $\Omega \in \mathbb{R}^{n \times n}$ is an orthogonal matrix, then the analysis sparsity model coincides with the synthesis sparsity model. More precisely, if $x = \Omega^T \alpha$ for an k -sparse $\alpha \in \mathbb{R}^n$ (so taking $D = \Omega^T$ as the basis), then $\Omega x = \alpha$ is k -sparse as well, meaning that x is $(n - k)$ -cosparse. Taking $\Omega \in \mathbb{R}^{p \times n}$ with $p > n$, the analysis

sparsity model is more general and offers more flexibility than synthesis sparsity. The following considerations on dimensionality and number of subspaces in (11.2) and (11.3) illustrate this point, see also [38, Section 2.3].

Let us first consider the generic case that both the rows $\omega_j \in \mathbb{R}^n$, $j = 1, \dots, p$, of the analysis operator as well as the columns $d_j \in \mathbb{R}^n$, $j = 1, \dots, q$, of the dictionary are in general linear position, i.e., any collection of at most n of these vectors are linearly independent. Then the following table comparing the s -sparse model (11.2) and the ℓ -cosparse model (11.3) applies

model	subspaces	no. subspaces	subspace dim.
synthesis	$V_T := \text{span}\{d_j, j \in T\}$	$\binom{q}{k}$	k
analysis	$W_\Lambda := \text{span}\{\omega_j, j \in \Lambda\}^\perp$	$\binom{p}{\ell}$	$n - \ell$

As suggested in [38] one way of comparing the two models is to consider an ℓ -cosparse analysis model and an $(n - \ell)$ -sparse synthesis model so that the corresponding subspaces have the same dimensions. If, for instance, $\ell = n - 1$ so that the dimension is $n - \ell = 1$, there are $\binom{q}{1} = q$ subspaces in the synthesis model,

while there are $\binom{p}{n-1}$ subspaces in the analysis sparsity model. Typically p is somewhat larger than n , and if q is not extremely large, the number of subspaces of dimension 1 is much larger for the analysis sparsity model than for the synthesis sparsity model. Or in other words, if one is looking for a synthesis sparsity model having the same one-dimensional subspaces as in a given analysis sparsity model then one needs $q = \binom{p}{n-1}$ many dictionary elements—usually way too many to be handled efficiently. More generally speaking, the analysis sparsity model contains many more low-dimensional subspaces than the synthesis sparsity model, but the situation reverses for high-dimensional subspaces [38].

As another difference to the synthesis sparsity model where any value of the sparsity k between 1 and n can occur, the dimension of W_Λ is restricted to a value between 0 and n and in the case of “generic” analysis operator (i.e., any set of at most n rows of Ω is linearly independent) the cosparsity is restricted to values between $p - n$ and p so that the sparsity of Ωx must be at least $p - n$ for a non-trivial vector x .

Sometimes it is desired not to have too many low-dimensional subspaces in the model and then it is beneficial if there are linear dependencies among the rows of the analysis operator Ω . In this case, the above table does no longer apply and the number of subspace may be significantly smaller. A particular situation where this happens is connected to the popular method of total variation. The

analysis operator is a two-dimensional difference operator. It collects all vertical and horizontal derivatives of an image $X \in \mathbb{R}^{n \times n}$ into a single vector. If we concatenate the columns of X into the vector $x \in \mathbb{R}^{n^2}$, then this operator is given by the matrix $\Omega \in \mathbb{R}^{2n(n-1) \times n^2}$ with a lot of linear dependencies, which acts by the rule

$$\Omega x = (X_{21} - X_{11}, \dots, X_{nn} - X_{n-1n}, X_{12} - X_{11}, \dots, X_{nn} - X_{nn-1})^T.$$

Another important case that we will consider in more detail in this chapter appears when the rows ω_j of $\Omega \in \mathbb{R}^{p \times n}$ form a frame [10, 14, 16, 25], i.e., if there exist constants $0 < a \leq b < \infty$ such that

$$a\|x\|_2^2 \leq \|\Omega x\|_2^2 = \sum_{j=1}^p |\langle \omega_j, x \rangle|^2 \leq b\|x\|_2^2 \quad (11.4)$$

Clearly, in our finite dimensional case such constants always exist if the ω_j span \mathbb{R}^n . For simplicity, we will often refer to Ω itself as a frame. If a lower and an upper frame bounds are equal, then Ω is called a tight frame. Frames are more general than orthonormal bases and allow for stable expansions. They are useful, for instance, when orthonormal bases with certain properties do not exist (see, e.g., the Balian–Low theorem in [1]). Moreover, their redundancy can be useful for tasks like error corrections in transmission of information, etc.

In case that Ω is a frame, then a signal x is uniquely determined by its frame coefficients Ωx . To reconstruct x from Ωx we make use of the canonical dual frame. Its elements are given by the columns of the matrix $\Omega^\dagger = (\Omega^T \Omega)^{-1} \Omega^T$ and for any x we have $x = \Omega^\dagger(\Omega x)$. A lower and an upper frame bounds of the canonical dual frame are b^{-1} and a^{-1} , respectively.

Particular frames of importance include Gabor frames [25], wavelet frames [35, 46, 50], shearlet [26], and curvelet frames [6], etc.

Instead of using a predefined analysis operator we can design it by learning. Given a set of training signals $\{u_i\}_{i=1}^d$, $u_i \in \mathbb{R}^n$, the goal is to find a matrix $\Omega \in \mathbb{R}^{p \times n}$, which provides the highest cosparsity for each u_i . We refer to [13, 28, 47] for further details on this subject.

In practice, signals are usually not exactly sparse or cosparse. In order to measure the error of approximation we recall that the error of k -term approximation of $x \in \mathbb{R}^n$ in ℓ_1 is defined as

$$\sigma_k(x)_1 := \inf_{z: \|z\|_0 \leq k} \|x - z\|_1.$$

In the cosparse case we use the quantity $\sigma_k(\Omega x)_1$ as a measure how close x is to being $(p - k)$ -cosparse. We remark that although for generic analysis operators $\Omega \in \mathbb{R}^{p \times n}$, the vector Ωx has at least $p - n$ nonzero entries (unless x is trivial), the approximation error $\sigma_k(\Omega x)_1$ may nevertheless become small for values of $k < p - n$.

11.3 Recovery of cosparse signals

We now turn to the compressed sensing problem of recovering an (approximately) cosparse vector $x \in \mathbb{R}^n$ from underdetermined linear measurements

$$y = Ax,$$

where $A \in \mathbb{R}^{m \times n}$ is a measurement matrix with $m < n$. Let $\Omega \in \mathbb{R}^{p \times n}$ be the analysis operator generating the analysis cosparsity model.

In analogy with the standard sparsity case (synthesis sparsity model) outlined in Chapter 1, Section 1.3, one might start with the ℓ_0 -minimization problem

$$\min_{z \in \mathbb{R}^n} \|\Omega z\|_0 \text{ subject to } Az = y. \quad (11.5)$$

However, this combinatorial optimization problem is again NP-hard in general. As an alternative, we may use its ℓ_1 -relaxation

$$\min_{z \in \mathbb{R}^n} \|\Omega z\|_1 \text{ subject to } Az = y \quad (11.6)$$

or in the noisy case

$$\min_{z \in \mathbb{R}^n} \|\Omega z\|_1 \text{ subject to } \|Az - y\|_2 \leq \varepsilon. \quad (11.7)$$

Alternative approaches include greedy-type algorithms such as Greedy Analysis Pursuit (GAP) [37, 38], thresholding-based methods [22, 40], or reweighted ℓ_1 -minimization [9].

We start by presenting conditions under which the solution of (11.6) coincides with the solution of (11.5), so that the original cosparse vector is recovered. We discuss versions of the null space property and the restricted isometry property. When the measurement matrix $A \in \mathbb{R}^{m \times n}$ is taken at random (as usual in compressed sensing), then an analysis of these concepts leads to so-called uniform recovery bounds stating that using a random draw of the measurement matrix one can with high probability recover all k -sparse vectors under a certain lower bound on the number of measurements. In contrast, nonuniform recovery results state that a given (fixed) cosparse vector can be recovered from a random draw of the measurement matrix with a certain probability. Those guarantees can be derived with conditions that depend both on the matrix A and the vector x to be recovered. We will state such conditions later on in this section.

11.3.1 Analysis null space property

As in the standard synthesis sparsity case, the null space property of the measurement matrix A characterizes recovery via analysis ℓ_1 -minimization. In analogy to the null space property described in Section 1.3.2 of the introductory chapter, we

say that given an analysis operator $\Omega \in \mathbb{R}^{p \times n}$, a measurement matrix $A \in \mathbb{R}^{m \times n}$ satisfies the Ω -null space property of order k if, for all subsets $\Lambda \subset [p]$ of cardinality $|\Lambda| \geq p - k$, it holds

$$\|\Omega_{\Lambda^c} v\|_1 < \|\Omega_\Lambda v\|_1 \quad \text{for all } v \in \ker A \setminus \{0\}.$$

Analogously to Theorem 2 of Chapter 1 it can be shown that every $(p - k)$ -cosparse vector can be recovered exactly via analysis ℓ_1 -minimization (11.6). In order to guarantee stable and robust recovery we use the following version of the null space property extending the corresponding notions from the standard synthesis sparsity case [20, Chapter 4].

Definition 1. A matrix $A \in \mathbb{R}^{m \times n}$ is said to satisfy the robust ℓ_2 -stable Ω -null space property of order k with constant $0 < \rho < 1$ and $\tau > 0$, if for any set $\Lambda \subset [p]$ with $|\Lambda| \geq p - k$ it holds

$$\|\Omega_{\Lambda^c} v\|_2 < \frac{\rho}{\sqrt{k}} \|\Omega_\Lambda v\|_1 + \tau \|Av\|_2 \quad \text{for all } v \in \mathbb{R}^n. \quad (11.8)$$

The following theorem has been shown in [29], similarly to [20, Theorem 4.22]

Theorem 1. Let $\Omega \in \mathbb{R}^{p \times n}$ be a frame with a lower frame bound $a > 0$. Let $A \in \mathbb{R}^{m \times n}$ satisfy the robust ℓ_2 -stable Ω -null space property of order k with constants $0 < \rho < 1$ and $\tau > 0$. Then for any $x \in \mathbb{R}^n$ the solution \hat{x} of (11.7) with $y = Ax + e$, $\|e\|_2 \leq \varepsilon$, approximates the vector x with ℓ_2 -error

$$\|x - \hat{x}\|_2 \leq \frac{2(1 + \rho)^2}{\sqrt{a}(1 - \rho)} \frac{\sigma_k(\Omega x)_1}{\sqrt{k}} + \frac{2\tau(3 + \rho)}{\sqrt{a}(1 - \rho)} \varepsilon. \quad (11.9)$$

We will analyze the stable null space property for Gaussian random matrices directly in Section 11.4.4.

11.3.2 Restricted isometry property

It is a by now classical approach to analyze sparse recovery algorithms via the restricted isometry property. A version for the analysis sparsity called the D-RIP was introduced in [7]. (The D corresponds to the dictionary, so in our notation Ω^T -RIP would be actually more appropriate.)

Definition 2. A measurement matrix $A \in \mathbb{R}^{m \times n}$ satisfies the restricted isometry property adapted to $\Omega \in \mathbb{R}^{p \times n}$ (the D-RIP) with constant δ_k if

$$(1 - \delta_k) \|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k) \|x\|_2^2 \quad (11.10)$$

holds for all $x = \Omega^T \alpha$ with $\|\alpha\|_0 \leq k$.

If $\Omega = \text{Id}$, we get the standard RIP-property.

In the important case that the rows of Ω form a frame, see (11.4), we have the following recovery result generalizing the one from the synthesis sparsity case. For the case of tight frames, it has been shown in [7], while the general case can be found in [19, Proposition 9].

Theorem 2. *Let $\Omega \in \mathbb{R}^{p \times n}$ be a frame with frame bounds $a, b > 0$ and $(\Omega^\dagger)^T$ its canonical dual frame. Suppose that the measurement matrix $A \in \mathbb{R}^{m \times n}$ obeys the restricted isometry property with respect to $(\Omega^\dagger)^T$ with constant $\delta_{2k} < \sqrt{a/b}/9$. Let $y = Ax + e$ with $\|e\|_2 \leq \varepsilon$. Then the solution \hat{x} of (11.6) satisfies*

$$\sqrt{a}\|x - \hat{x}\|_2 \leq c_0 \frac{\sigma_k(\Omega x)_1}{\sqrt{k}} + c_1 \varepsilon \quad (11.11)$$

for constants c_0, c_1 that depend only on δ_{2k} .

For the case that Ω is a difference operator corresponding to total variation minimization, recovery guarantees have been provided in [39].

11.3.3 Recovery conditions via tangent cones

The null space property and the restricted isometry property of A guarantee that all (co-)sparse vectors can be recovered via analysis ℓ_1 -minimization from measurements obtained by applying A . It is also useful to have recovery conditions that not only depend on A but also on the vector to be recovered. In fact, such conditions are at the basis for the nonuniform recovery guarantees stated below.

This section follows the approach of [12] that works with tangent cones of $\|\Omega \cdot\|_1$ at the vector to be recovered. For fixed $x \in \mathbb{R}^n$ we define the convex cone

$$T(x) = \text{cone}\{z - x : z \in \mathbb{R}^n, \|\Omega z\|_1 \leq \|\Omega x\|_1\}, \quad (11.12)$$

where the notation ‘‘cone’’ stands for the conic hull of the indicated set. The set $T(x)$ consists of the directions from x , which do not increase the value of $\|\Omega x\|_1$. The following result describes a geometric property, which guarantees exact recovery, see Figure 11.1. It was proved in [29] and is analogous to Proposition 2.1 in [12], see also [20, Theorem 4.35].

Theorem 3. *Let $A \in \mathbb{R}^{m \times n}$. A vector $x \in \mathbb{R}^n$ is the unique minimizer of $\|\Omega z\|_1$ subject to $Az = Ax$ if and only if $\ker A \cap T(x) = \{0\}$.*

Proof. For convenience we prove that the condition $\ker A \cap T(x) = \{0\}$ implies recovery. For the other direction, we refer to [29].

Suppose there is $z \in \mathbb{R}^n$ such that $Az = Ax$ and $\|\Omega z\|_1 \leq \|\Omega x\|_1$. Then $z - x \in T(x)$ and $z - x \in \ker A$. Since $\ker A \cap T(x) = \{0\}$, we conclude that $z - x = 0$, so that x is the unique minimizer. \square

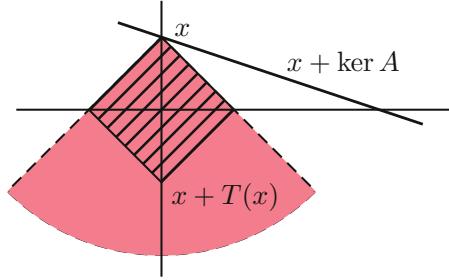


Fig. 11.1 Geometry of successful recovery. The dashed region corresponds to the set $\{z : \|\Omega z\|_1 \leq \|\Omega x\|_1\}$.

When the measurements are noisy, we use the following criteria for robust recovery [29].

Theorem 4. *Let $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ and $y = Ax + e$ with $\|e\|_2 \leq \varepsilon$. If*

$$\inf_{\substack{v \in T(x) \\ \|v\|_2=1}} \|Av\|_2 \geq \tau \quad (11.13)$$

for some $\tau > 0$, then a solution \hat{x} of the ℓ_1 -analysis minimization problem (11.7) satisfies

$$\|x - \hat{x}\|_2 \leq \frac{2\varepsilon}{\tau}.$$

Proof. Since \hat{x} is a minimizer of (11.7), we have $\|\Omega \hat{x}\|_1 \leq \|\Omega x\|_1$ and $\hat{x} - x \in T(x)$. Our assumption (11.13) implies

$$\|A(\hat{x} - x)\|_2 \geq \tau \|\hat{x} - x\|_2. \quad (11.14)$$

On the other hand, an upper bound for $\|A\hat{x} - Ax\|_2$ is given by

$$\|A\hat{x} - Ax\|_2 \leq \|A\hat{x} - y\| + \|Ax - y\|_2 \leq 2\varepsilon. \quad (11.15)$$

Combining (11.14) and (11.15) we get the desired estimate. \square

11.3.4 Dual certificates

Another common approach for recovery conditions is based on duality. For the standard synthesis sparsity model corresponding results have been obtained, for instance, in [21, 51], see also [20, Theorem 4.26–4.33]. In fact, the first contribution to compressed sensing by Candès et al. [8] is based on such an approach.

Apparently, Haltmeier [27] first addressed the problem of robust recovery of a signal by analysis ℓ_1 -minimization, when the analysis operator is given by a frame. His result has been generalized in [18, 52], which for our particular case is stated in the following theorem. In order to formulate it, we recall that the subdifferential of a convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined as

$$\partial f(x) = \{v \in \mathbb{R}^n : f(z) \geq f(x) + \langle z - x, v \rangle \text{ for all } z \in \mathbb{R}^n\}.$$

The subdifferential of the ℓ_1 -norm is given as $\partial \|\cdot\|(x) = \{(v_j)_j : v_j \in \partial |\cdot|(x_j)\}$, where the subdifferential of the absolute value function is

$$\partial |\cdot|(u) = \begin{cases} [-1, 1] & \text{if } u = 0, \\ \text{sgn}(u) & \text{if } u \neq 0. \end{cases}$$

Theorem 5. *Let $x \in \mathbb{R}^n$ be cosparse with cosupport Λ and let noisy measurements $y = Ax + e$ be given with $\|e\|_2 \leq \varepsilon$. Assume the following:*

1. *There exists $\eta \in \mathbb{R}^m$ (the dual vector) and $\alpha \in \partial \|\cdot\|_1(\Omega x)$ such that*

$$A^* \eta = \Omega^* \alpha, \text{ with } \|\alpha_\Lambda\|_\infty \leq \kappa < 1. \quad (11.16)$$

2. *The sensing matrix A is injective on $W_\Lambda = \ker \Omega_\Lambda$ implying that there exists $C_A > 0$ such that*

$$\|Ax\|_2 \geq C_A \|x\|_2, \text{ for any } x \in W_\Lambda. \quad (11.17)$$

Then any solution \hat{x} of the analysis ℓ_1 -minimization problem (11.7) approximates x with ℓ_2 -error

$$\|x - \hat{x}\|_2 \leq C\varepsilon,$$

where C depends on the same quantities as the constant τ in (11.18) below.

In the inverse problems community (11.16) is also referred to as source (range) condition, while (11.17) is called injectivity condition and they are commonly used to provide error estimates for the sparsity promoting regularizations [4, 24]. In [4] the error is measured by the Bregman distance. The work of Grasmair [24] provides a more general result, where the error is measured in terms of the regularization functional. The first results concerning the ℓ_2 -error estimates (as stated above) are given in [27], where the analysis operator is assumed to be a frame. The result [18] extends to general analysis operators and decomposable norms.

We close this section by showing that the conditions in Theorem 5 are stronger than the tangent cone condition of Theorem 4.

Theorem 6. *Let $A \in \mathbb{R}^{m \times n}$ and $x \in \mathbb{R}^n$ with cosupport Λ satisfy (11.16) and (11.17) with vectors $\alpha \in \partial \|\cdot\|_1(\Omega x)$, $\|\alpha_\Lambda\|_\infty \leq \kappa$, and $\eta \in \mathbb{R}^m$. Then*

$$\inf_{\substack{v \in T(x) \\ \|v\|_2=1}} \|Av\|_2 \geq \tau$$

with

$$\tau = \left(C_A^{-1} + \frac{(C_A + \|A\|_{2 \rightarrow 2}) \|\eta\|_2}{C_{\Omega, A} C_A (1 - \kappa)} \right)^{-1}, \quad (11.18)$$

where $C_{\Omega, A}$ depends only on Ω and Λ .

The proof follows the same lines as the proof of the main result in [18]. We start with the following lemma.

Lemma 1. *Let $A \in \mathbb{R}^{m \times n}$ and $x \in \mathbb{R}^n$ with cosupport Λ satisfy (11.16) and (11.17) with vectors $\alpha \in \partial \|\cdot\|_1(\Omega x)$, $\|\alpha_\Lambda\|_\infty \leq \kappa$, and $\eta \in \mathbb{R}^m$. Then for any $v \in T(x)$*

$$\|\Omega_\Lambda v\|_1 \leq \frac{\|\eta\|_2}{1 - \kappa} \|Av\|_2.$$

Proof. For $v \in T(x)$, there exists $\beta_\Lambda \in \mathbb{R}^p$ (that is, β_Λ is zero on Λ^c) such that $\|\beta_\Lambda\|_\infty \leq 1$ and $\|(\Omega v)_\Lambda\|_1 = \langle (\Omega v)_\Lambda, \beta_\Lambda \rangle$. The subdifferential $\partial \|\cdot\|_1(\Omega x)$ of the ℓ_1 -norm at the point Ωx consists of all vectors α , such that any $\beta \in \mathbb{R}^p$ satisfies

$$\|\beta\|_1 \geq \|\Omega x\|_1 + \langle \alpha, \beta - \Omega x \rangle.$$

Since $x \in \mathbb{R}^n$ is cosparse with the cosupport Λ , we can explicitly write

$$\partial \|\cdot\|_1(\Omega x) = \{\alpha \in \mathbb{R}^p : \alpha_{\Lambda^c} = \text{sgn}(\Omega x), \|\alpha_\Lambda\|_\infty \leq 1\}.$$

Taking into account that the vector β_Λ has zero-entries on the index set Λ^c and $\|\beta_\Lambda\|_\infty \leq 1$, it follows that $\text{sgn}(\Omega x) + \beta_\Lambda \in \partial \|\cdot\|_1(\Omega x)$. Every $v \in T(x)$ is represented as

$$v = \sum_j t_j v_j, \quad v_j = z_j - x, \quad \|\Omega z_j\|_1 \leq \|\Omega x\|_1, \quad t_j \geq 0.$$

Let $\alpha \in \partial \|\cdot\|_1(\Omega x)$, so that $\alpha_{\Lambda^c} = \text{sgn}(\Omega x)$. The definition of the subdifferential implies then that

$$\begin{aligned} 0 &\geq \|\Omega z_j\|_1 - \|\Omega x\|_1 \geq \langle \text{sgn}(\Omega x) + \beta_\Lambda, \Omega(z_j - x) \rangle \\ &= \langle \text{sgn}(\Omega x) + \beta_\Lambda - \alpha, \Omega v_j \rangle + \langle \alpha, \Omega v_j \rangle \geq \langle \beta_\Lambda - \alpha_\Lambda, \Omega_\Lambda v_j \rangle + \langle \alpha, \Omega v_j \rangle. \end{aligned}$$

Multiplying by $t_j \geq 0$ and summing up over all j gives

$$0 \geq \langle \beta_\Lambda - \alpha_\Lambda, \Omega_\Lambda (\sum_j t_j v_j) \rangle + \langle \alpha, \Omega (\sum_j t_j v_j) \rangle = \langle \beta_\Lambda - \alpha_\Lambda, (\Omega v)_\Lambda \rangle + \langle \alpha, \Omega v \rangle.$$

Due to the choice of β_Λ and duality of the ℓ_1 -norm and ℓ_∞ norm we obtain

$$0 \geq \|\Omega_\Lambda v\|_1 - \|\alpha_\Lambda\|_\infty \|\Omega_\Lambda v\|_1 + \langle \alpha, \Omega v \rangle,$$

which together with (11.16) gives

$$\begin{aligned}\|\Omega_A v\|_1 &\leq -\frac{\langle \alpha, \Omega v \rangle}{1 - \|\alpha\|_\infty} = -\frac{\langle \Omega^* \alpha, v \rangle}{1 - \|\alpha\|_\infty} = -\frac{\langle A^* \eta, v \rangle}{1 - \|\alpha\|_\infty} \\ &= -\frac{\langle \eta, Av \rangle}{1 - \|\alpha\|_\infty} \leq \frac{\|\eta\|_2}{1 - \kappa} \|Av\|_2.\end{aligned}$$

This concludes the proof. \square

Proof (of Theorem 6). The idea is to split $v \in T(x)$ into its projection onto the subspace $W_A = \ker \Omega_A$ and its complement W_A^\perp . Since we are in finite dimensions, it follows that $\|\Omega_A w\|_2 \geq C_{\Omega,A} \|w\|_2$ for all $w \in W_A^\perp$ for some constant $C_{\Omega,A}$. Taking into account (11.17) we obtain

$$\begin{aligned}\|v\|_2 &\leq \|\mathcal{P}_{W_A} v\|_2 + \|\mathcal{P}_{W_A^\perp} v\|_2 \leq C_A^{-1} \|A \mathcal{P}_{W_A} v\|_2 + \|\mathcal{P}_{W_A^\perp} v\|_2 \\ &= C_A^{-1} \|A(v - \mathcal{P}_{W_A^\perp} v)\|_2 + \|\mathcal{P}_{W_A^\perp} v\|_2 \\ &\leq C_A^{-1} \|Av\|_2 + C_A^{-1} \|A \mathcal{P}_{W_A^\perp} v\|_2 + \|\mathcal{P}_{W_A^\perp} v\|_2 \\ &\leq C_A^{-1} \|Av\|_2 + (1 + C_A^{-1} \|A\|_{2 \rightarrow 2}) \|\mathcal{P}_{W_A^\perp} v\|_2 \\ &\leq C_A^{-1} \|Av\|_2 + (1 + C_A^{-1} \|A\|_{2 \rightarrow 2}) C_{\Omega,A}^{-1} \|\Omega_A \mathcal{P}_{W_A^\perp} v\|_2\end{aligned}$$

Since $\Omega_A \mathcal{P}_{W_A} v = 0$ and $\Omega_A \mathcal{P}_{W_A^\perp} v = \Omega_A(v - \mathcal{P}_{W_A^\perp} v) = \Omega_A v$ by definition of W_A , the estimate above can be continued to obtain

$$\|v\|_2 \leq C_A^{-1} \|Av\|_2 + \frac{C_A + \|A\|_{2 \rightarrow 2}}{C_{\Omega,A} C_A} \|\Omega_A v\|_2 \leq C_A^{-1} \|Av\|_2 + \frac{C_A + \|A\|_{2 \rightarrow 2}}{C_{\Omega,A} C_A} \|\Omega_A v\|_1.$$

As a final step we apply Lemma 1. \square

11.4 Recovery from random measurements

A main task in compressed sensing is to obtain bounds for the minimal number of linear measurements required to recover a (co-)sparse vector via certain recovery methods, say analysis ℓ_1 -minimization. It is up till now open to rigorously prove such guarantees—and in particular, verify the conditions of the previous sections—for deterministic sensing matrix constructions in the optimal parameter regime, see, for instance, [20, Chapter 6.1] for a discussion. Therefore, we pass to random matrices.

A matrix that is populated with independent standard normal distributed entries (see also Chapter 1) is called a *Gaussian matrix*. We will also consider subgaussian matrices. To this end, we introduce the ψ_2 -norm of a random variable X which is defined as

$$\|X\|_{\psi_2} := \inf \left\{ c > 0 : \mathbb{E} \exp \left(|X|^2 / c^2 \right) \leq 2 \right\}.$$

A random variable X is called *subgaussian*, if $\|X\|_{\psi_2} < \infty$. Boundedness of the ψ_2 -norm of a random variable X is equivalent to the fact that its tail satisfies $\mathbb{P}(|X| > t) \leq 2e^{-ct^2}$ and its moments $(\mathbb{E}|X|^p)^{1/p}$, $p \geq 1$, grow like \sqrt{p} . Standard examples of subgaussian random variables are Gaussian, Bernoulli, and bounded random variables. A matrix with independent mean-zero and variance one subgaussian entries is called a *subgaussian matrix*.

Definition 3. A random vector X in \mathbb{R}^n is called *isotropic*, if $\mathbb{E}|\langle X, x \rangle|^2 = \|x\|_2^2$ for every $x \in \mathbb{R}^n$. A random vector $X \in \mathbb{R}^n$ is *subgaussian*, if $\langle X, x \rangle$ are subgaussian random variables for all $x \in \mathbb{R}^n$. The ψ_2 norm of X is defined as

$$\|X\|_{\psi_2} = \sup_{u \in S^{n-1}} \|\langle X, u \rangle\|_{\psi_2}.$$

A random matrix whose rows are independent, isotropic and subgaussian is called an isotropic subgaussian ensemble. Subgaussian random matrices are examples of such matrices.

We will give an overview on recovery results for analysis ℓ_1 -minimization where the analysis operator is a frame and the measurement matrix is Gaussian or more generally an isotropic subgaussian ensemble. We cover both uniform recovery, which is studied via the Ω -null space property or via the D-RIP, and nonuniform recovery bounds, where especially for the Gaussian case, it is possible to derive explicit bounds with small constants. For some of these results, new proofs are provided. For the important special case of total variation minimization where cosparsity is with respect to a one- or two-dimensional difference operator, some recent recovery results can be found in [30], which also contains an alternative approach for Gaussian matrices and cosparsity with respect to frames.

11.4.1 Uniform recovery via the restricted isometry property

We recall from Chapter 1 (Theorem 5) that a rescaled Gaussian random matrix $A \in \mathbb{R}^{m \times n}$ satisfies the standard restricted isometry property, i.e. $\delta_k \leq \delta$ with probability at least $1 - \theta$ provided that $m \geq C\delta^{-2}(k \ln(eN/k) + \ln(2\theta^{-1}))$. A similar estimate, derived in [3, 7], holds for the D-RIP in (2).

Theorem 7. Let $A \in \mathbb{R}^{m \times n}$ be a draw of a Gaussian random matrix and let $\Omega \in \mathbb{R}^{p \times n}$ be an analysis operator. If

$$m \geq C\delta^{-2}(k \ln(ep/k) + \ln(2\theta^{-1})), \quad (11.19)$$

then with probability at least $1 - \theta$ the matrix $\frac{1}{\sqrt{m}}A$ satisfies the restricted isometry property adapted to Ω with constant $\delta_k \leq \delta$.

Proof. This is a generalization of the standard restricted isometry property of Gaussian matrices. The proof relies on the concentration of measure phenomenon formulated in Theorem 4 of Chapter 1 and the covering argument presented in the same chapter in Lemma 3. The only difference in comparison with the proof of Theorem 5 in Chapter 1 occurs in the step of taking a union bound with respect to all k -dimensional subspaces. There are $\binom{p}{k} \leq \left(\frac{ep}{k}\right)^k$ subspaces, which is reflected in the term $k \ln(ep/k)$. \square

The above result extends to isotropic subgaussian random matrices.

Theorem 8 (Corollary 3.1 of [15]). Let $A \in \mathbb{R}^{m \times n}$ be a draw of an isotropic subgaussian ensemble and let $\Omega \in \mathbb{R}^{p \times n}$ be an analysis operator. If

$$m \geq C\delta^{-2}(k \ln(ep/k) + \ln(2\theta^{-1}))$$

then with probability at least $1 - \theta$ the matrix $\frac{1}{\sqrt{m}}A$ satisfies the restricted isometry property adapted to Ω with constant $\delta_k \leq \delta$.

An extension of the above bound to Weibull matrices has been shown in [19].

Applying the above results for the canonical dual frame $(\Omega^\dagger)^T$ of a frame Ω in combination with Theorem 2 shows that the analysis ℓ_1 -program

$$\min \|\Omega z\|_1 \quad \text{subject to } Az = Ax$$

with a random draw of a (sub-)gaussian matrix $A \in \mathbb{R}^{m \times n}$ recovers every ℓ -cosparse vector x with $\ell = p - k$ exactly with high probability provided

$$m \geq \frac{Cb}{a} k \ln(ep/k). \quad (11.20)$$

The difference to the standard synthesis sparsity case is merely the appearance of p instead of n inside the logarithmic factor as well as the ratio b/a of the frame bounds. Clearly, this ratio is one for a tight frame.

We will return to uniform recovery with Gaussian measurements in Subsection 11.4.4, where we will study the Ω -null space property directly which allows to give an explicit and small constant in the bound (11.19) on the number of measurements. The approach relies on techniques that are introduced in the next section concerning nonuniform recovery. (These methods are easier to apply in the nonuniform setting which is the reason why we postpone an analysis of the Ω -null space property to later.)

11.4.2 Nonuniform recovery from Gaussian measurements

We now turn nonuniform results for recovery of cosparse signals with respect to a frame being the analysis operator and using Gaussian measurement matrices, which state that a given (fixed) cosparse vector can be recovered with high probability under a certain bound on the number of measurements. We will not qualitatively improve over (11.20), but we will obtain a very good constant, which is in fact optimal in a certain “asymptotic” sense. The main result stated next appeared in [29], but we give a slightly different proof here, which allows us later to extend this approach also to the subgaussian case.

Theorem 9. *Let $\Omega \in \mathbb{R}^{p \times n}$ be a frame with frame bounds $a, b > 0$ and $x \in \mathbb{R}^n$ be an ℓ -cosparse vector and $k = p - \ell$. Let $A \in \mathbb{R}^{m \times n}$ be a Gaussian random matrix and let noisy measurements $y = Ax + e$ be taken with $\|e\|_2 \leq \varepsilon$. If for $0 < \theta < 1$ and some $\tau > 0$*

$$\frac{m^2}{m+1} \geq \frac{2bk}{a} \left(\sqrt{\ln \frac{ep}{k}} + \sqrt{\frac{a \ln(\theta^{-1})}{bk}} + \tau \sqrt{\frac{a}{2bk}} \right)^2, \quad (11.21)$$

then with probability at least $1 - \theta$, every minimizer \hat{x} of (11.7) satisfies

$$\|x - \hat{x}\|_2 \leq \frac{2\varepsilon}{\tau}.$$

Setting $\varepsilon = 0$ yields exact recovery via (11.6). Roughly speaking, i.e., for rather large k, m, p the bound (11.21) reads

$$m > 2 \frac{b}{a} k \ln(ep/k)$$

for having recovery with “high probability”.

Our proof relies on the recovery condition of Theorem 4 based on tangent cones as well as on convex geometry and Gordon’s escape through a mesh theorem [23] and is inspired by [12], see also [20, Chapter 9].

According to (11.13) the successful recovery of a signal is achieved, when the minimal gain of the measurement matrix over the tangent cone is greater than some positive constant. For Gaussian matrices the probability of this event is estimated by Gordon’s escape through a mesh theorem [23], [20, Theorem 9.21]. To present it formally we introduce some notation. For a set $T \subset \mathbb{R}^n$ we define its Gaussian width by

$$\ell(T) := \mathbb{E} \sup_{x \in T} \langle x, g \rangle, \quad (11.22)$$

where $g \in \mathbb{R}^n$ is a standard Gaussian random vector. Due to the rotation invariance (11.22) can be written as

$$\ell(T) = \mathbb{E} \|g\|_2 \cdot \mathbb{E} \sup_{x \in T} \langle x, u \rangle,$$

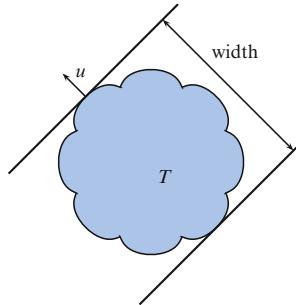


Fig. 11.2 The mean width of a set T in the direction u . When T is symmetric, $2\sup_{x \in T} \langle x, u \rangle = \sup_{x \in T} \langle x, u \rangle - \inf_{z \in T} \langle z, u \rangle$, which corresponds to the smallest distance between two hyperplanes orthogonal to the direction u , such that T is contained between them.

where u is uniformly distributed on S^{n-1} . We recall [20, Theorem 8.1] that

$$E_n := \mathbb{E} \|g\|_2 = \sqrt{2} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)}$$

satisfies

$$\frac{n}{\sqrt{n+1}} \leq E_n \leq \sqrt{n},$$

so that up to some factor of order \sqrt{n} the Gaussian width is basically equivalent to the mean width of a set, see Figure 11.2.

In order to gain more intuition about the Gaussian width we remark that a d -dimensional subspace $U \subset \mathbb{R}^n$ intersected with the sphere S^{n-1} satisfies $\ell(U \cap S^{n-1}) \sim \sqrt{d}$ so that for a subset T of the sphere the quantity $\ell(T)^2$ can somehow be interpreted as its dimension (although this interpretation should be handled with care.)

Next we state the version of Gordon's escape through a mesh theorem from [20, Theorem 9.21].

Theorem 10 (Gordon's escape through a mesh). *Let $A \in m \times n$ be a Gaussian random matrix and T be a subset of the unit sphere S^{n-1} . Then, for $t > 0$, it holds*

$$\mathbb{P} \left(\inf_{x \in T} \|Ax\|_2 > E_m - \ell(T) - t \right) \geq 1 - e^{-\frac{t^2}{2}}. \quad (11.23)$$

Recall the set

$$T(x) = \text{cone}\{z - x : z \in \mathbb{R}^n, \|\Omega z\|_1 \leq \|\Omega x\|_1\},$$

from (11.12) and set $T := T(x) \cap S^{n-1}$. To provide a bound on the number of Gaussian measurements, we compute the Gaussian width of T . We start with establishing a connection between $\ell(T)$ and $\ell(\Omega(T))$, where $\Omega(T)$ is the set obtained by applying operator Ω to each element of a set T .

Theorem 11. *Let $\Omega \in \mathbb{R}^{p \times n}$ be a frame with a lower frame bound $a > 0$ and $T \subset \mathbb{R}^n$. Then*

$$\ell(T) \leq a^{-1/2} \ell(\Omega(T)).$$

Before giving the proof of Theorem 11, we recall Slepian's inequality, see [33] or [20, Chapter 8.7] for details. For a random variable X we define $\|X\|_2 = (\mathbb{E}|X|^2)^{1/2}$.

Theorem 12. *Let $(X_t)_{t \in T}$ and $(Y_t)_{t \in T}$ be Gaussian centered processes. If for all $s, t \in T$*

$$\|X_t - X_s\|_2 \leq \|Y_t - Y_s\|_2,$$

then

$$\mathbb{E} \sup_{t \in T} X_t \leq \mathbb{E} \sup_{t \in T} Y_t.$$

Proof (of Theorem 11). By the definition of the Gaussian width

$$\begin{aligned} \ell(T) &= \mathbb{E} \sup_{v \in T} \langle g, v \rangle = \mathbb{E} \sup_{v \in T} \langle g, \Omega^\dagger \Omega v \rangle \\ &= \mathbb{E} \sup_{w \in \Omega(T)} \langle (\Omega^\dagger)^T g, w \rangle. \end{aligned} \tag{11.24}$$

We consider two Gaussian processes,

$$X_w = \langle (\Omega^\dagger)^T g, w \rangle \text{ and } Y_w = \|(\Omega^\dagger)^T\|_{2 \rightarrow 2} \langle h, w \rangle,$$

where $w \in \Omega(T)$ and $g \in \mathbb{R}^n$, $h \in \mathbb{R}^p$ are both standard Gaussian vectors. For any standard Gaussian vector h it holds $\mathbb{E}|\langle h, w \rangle|^2 = \|w\|_2^2$. So if $w, w' \in \Omega(T)$, then

$$\|X_w - X_{w'}\|_2 \leq \|(\Omega^\dagger)^T\|_{2 \rightarrow 2} \|w - w'\|_2 = \|Y_w - Y_{w'}\|_2.$$

It follows from Theorem 12 that

$$\mathbb{E} \sup_{w \in \Omega(T)} \langle (\Omega^\dagger)^T g, w \rangle \leq \|(\Omega^\dagger)^T\|_{2 \rightarrow 2} \mathbb{E} \sup_{w \in \Omega(T)} \langle h, w \rangle. \tag{11.25}$$

An upper bound of the canonical dual frame is $(\Omega^\dagger)^T$ is a^{-1} . Hence, $\|(\Omega^\dagger)^T\|_{2 \rightarrow 2} \leq a^{-1/2}$. Together with (11.24) and (11.25) this gives

$$\ell(T) \leq a^{-1/2} \mathbb{E} \sup_{w \in \Omega(T)} \langle h, w \rangle = a^{-1/2} \ell(\Omega(T)).$$

□

The next theorem from [29, Section 2.2] provides a good bound on $\ell(\Omega(T))$.

Theorem 13. *Let $\Omega \in \mathbb{R}^{p \times n}$ be a frame with an upper frame bound $b > 0$ and $x \in \mathbb{R}^n$ be ℓ -cosparse with $\ell = p - k$. For $T := T(x) \cap S^{n-1}$, it holds*

$$\ell(\Omega(T))^2 \leq 2bk \ln \left(\frac{ep}{k} \right). \quad (11.26)$$

Proof. Since Ω is a frame with an upper frame constant b , we have

$$\Omega(T) \subset \Omega(T(x)) \cap \Omega(S^{n-1}) \subset K(\Omega x) \cap \left(\sqrt{b} B_2^p \right),$$

where

$$K(\Omega x) = \text{cone} \{ y - \Omega x : y \in \mathbb{R}^p, \|y\|_1 \leq \|\Omega x\|_1 \}.$$

The supremum over a larger set can only increase, hence

$$\ell(\Omega(T)) \leq \sqrt{b} \ell(K(\Omega x) \cap B_2^p). \quad (11.27)$$

An upper bound for the Gaussian width $\ell(K(\Omega x) \cap B_2^p)$ can be given in terms of the polar cone $\mathcal{N}(\Omega x) = K(\Omega x)^\circ$ defined by

$$\mathcal{N}(\Omega x) = \{ z \in \mathbb{R}^p : \langle z, y - \Omega x \rangle \leq 0 \text{ for all } y \in \mathbb{R}^p \text{ such that } \|y\|_1 \leq \|\Omega x\|_1 \}.$$

By duality of convex programming, see [12] or [20, Proposition 9.22], we have

$$\ell(K(\Omega x) \cap B_2^p) \leq \mathbb{E} \min_{z \in \mathcal{N}(\Omega x)} \|g - z\|_2$$

and by Hölder's inequality

$$\ell(K(\Omega x) \cap B_2^p)^2 \leq \left(\mathbb{E} \min_{z \in \mathcal{N}(\Omega x)} \|g - z\|_2 \right)^2 \leq \mathbb{E} \min_{z \in \mathcal{N}(\Omega x)} \|g - z\|_2^2. \quad (11.28)$$

Let Λ be the cosupport of x . Then one can verify that

$$\mathcal{N}(\Omega x) = \bigcup_{t \geq 0} \{ z \in \mathbb{R}^p : z_i = t \text{sgn}(\Omega x)_i, i \in \Lambda^c, |z_i| \leq t, i \in \Lambda \}. \quad (11.29)$$

To proceed, we fix t , minimize $\|g - z\|_2^2$ over all possible entries z_j , take the expectation of the obtained expression, and finally optimize over t . Taking into account (11.29), we have

$$\begin{aligned} \min_{z \in \mathcal{N}(\Omega x)} \|g - z\|_2^2 &= \min_{\substack{t \geq 0 \\ |z_i| \leq t, i \in \Lambda^c}} \sum_{i \in \Lambda^c} (g_i - t \operatorname{sgn}(\Omega x)_i)^2 + \sum_{i \in \Lambda} (g_i - z_i)^2 \\ &= \min_{t \geq 0} \sum_{i \in \Lambda^c} (g_i - t \operatorname{sgn}(\Omega x)_i)^2 + \sum_{i \in \Lambda} S_t(g_i)^2, \end{aligned}$$

where S_t is the soft-thresholding operator given by

$$S_t(x) = \begin{cases} x + t & \text{if } x < -t, \\ 0 & \text{if } -t \leq x \leq t, \\ x - t & \text{if } x > t. \end{cases}$$

Taking expectation we arrive at

$$\begin{aligned} \mathbb{E} \min_{z \in \mathcal{N}(\Omega x)} \|g - z\|_2^2 &\leq \mathbb{E} \left[\sum_{i \in \Lambda^c} (g_i - t \operatorname{sgn}(\Omega x)_i)^2 \right] + \mathbb{E} \left[\sum_{i \in \Lambda} S_t(g_i)^2 \right] \\ &= k(1 + t^2) + (p - k) \mathbb{E} S_t(g)^2, \end{aligned} \quad (11.30)$$

where g is a univariate standard Gaussian random variable. The expectation of $S_t(g)^2$ is estimated by integration,

$$\begin{aligned} \mathbb{E} S_t(g)^2 &= \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{-t} (x+t)^2 e^{-\frac{x^2}{2}} dx + \int_t^{\infty} (x-t)^2 e^{-\frac{x^2}{2}} dx \right] \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x^2 e^{-\frac{(x+t)^2}{2}} dx = \frac{2e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \int_0^{\infty} x^2 e^{-\frac{x^2}{2}} e^{-xt} dx \\ &\leq e^{-\frac{t^2}{2}} \sqrt{\frac{2}{\pi}} \int_0^{\infty} x^2 e^{-\frac{x^2}{2}} dx = e^{-\frac{t^2}{2}}. \end{aligned} \quad (11.31)$$

Substituting the estimate (11.31) into (11.30) gives

$$\mathbb{E} \min_{z \in \mathcal{N}(\Omega x)} \|g - z\|_2^2 \leq k(1 + t^2) + (p - k)e^{-\frac{t^2}{2}}.$$

Setting $t = \sqrt{2 \ln(p/k)}$ finally leads to

$$\ell(K(\Omega x) \cap B_2^p)^2 \leq k(1 + 2 \ln(p/k)) + k = 2k \ln(ep/k). \quad (11.32)$$

By combining inequalities (11.27) and (11.32) we obtain

$$\ell(\Omega(T))^2 \leq 2bk \ln \frac{ep}{k}. \quad \square$$

By Theorem 11 and Theorem 13 we obtain

$$\ell(T) \leq \sqrt{\frac{2bk}{a} \ln \frac{ep}{k}}.$$

At this point we are ready to prove our main result (Theorem 9) concerning the number of Gaussian measurements that guarantee the robust recovery of a fixed cosparse vector.

Set $t = \sqrt{2 \ln(\theta^{-1})}$. The choice of m in (11.21) guarantees that

$$E_m - \ell(T) - t \geq \frac{m}{\sqrt{m+1}} - \sqrt{\frac{bk}{a} \ln \frac{ep}{k}} - \sqrt{2 \ln(\theta^{-1})} \geq \tau.$$

Theorem 10 yields

$$\mathbb{P} \left(\inf_{x \in T} \|Ax\|_2 \geq \tau \right) \geq \mathbb{P} \left(\inf_{x \in T} \|Ax\|_2 \geq E_m - \ell(T) - t \right) \geq 1 - \theta.$$

This completes the proof.

11.4.3 Nonuniform recovery from subgaussian measurements

Based on the estimates of the previous section in combination with Theorem 14 due to Mendelson et al. [36] (see below), we may extend the nonuniform recovery result also to subgaussian matrices. We remark, however, that due to an unspecified constant in (11.33) this technique does not necessarily improve over the uniform estimate (11.20) derived via the restricted isometry property. Nevertheless, we feel that this proof method is interesting for theoretical reasons.

Theorem 14 (Corollary 2.7 in [36]). *Let $T \subset S^{n-1}$, $X_i \in \mathbb{R}^n$, $i = 1, \dots, m$ be independent isotropic subgaussian random vectors with $\|X_i\|_{\psi_2} \leq \alpha$ and $0 < \delta < 1$. Define $A \in \mathbb{R}^{m \times n}$ as $A = (X_1, \dots, X_m)^T$. If*

$$m \geq \frac{c_1 \alpha^4}{\delta^2} \ell(T)^2, \quad (11.33)$$

then with probability at least $1 - \exp(-c_2 \delta^2 m / \alpha^4)$ for all $x \in T$ it holds

$$1 - \delta \leq \frac{\|Ax\|_2^2}{m} \leq 1 + \delta, \quad (11.34)$$

where c_1, c_2 are absolute constants.

Theorem 15. Let $\Omega \in \mathbb{R}^{p \times n}$ be a frame with frame bounds $a, b > 0$, $x \in \mathbb{R}^n$ be an ℓ -cosparse vector and $k = p - l$. Let $X_i \in \mathbb{R}^n$, $i = 1, \dots, m$, be independent isotropic subgaussian random vectors with $\|X_i\|_{\psi_2} \leq \alpha$ and $0 < \delta < 1$. Define $A \in \mathbb{R}^{m \times n}$ as $A = (X_1, \dots, X_m)^T$. If

$$m \geq \frac{c_1 b \alpha^4}{a \delta^2} s \ln \frac{ep}{k},$$

then with probability at least $1 - \exp(-c_2 \delta^2 m / \alpha^4)$ every minimizer \hat{x} of (11.7) approximates x with the following ℓ_2 -error

$$\|x - \hat{x}\|_2 \leq \frac{2\epsilon}{\sqrt{1 - \delta}}.$$

Proof. Inserting the estimate of Gaussian width (11.26) and (11.38) in (11.33) provides the above bound on the number of subgaussian measurements that guarantees successful nonuniform recovery via (11.6). \square

11.4.4 Uniform recovery via the Ω -null space property

Let us now consider the stable Ω -null space property introduced in (11.8), which implies (uniform) recovery of all cosparse vectors via analysis ℓ_1 -minimization, see Theorem 1. Since the restricted isometry property adapted to $(\Omega^\dagger)^T$ implies the Ω -null space property, see Theorem 2, we already know that Gaussian (and subgaussian) measurement matrices satisfy the Ω -null space property with high probability under Condition (11.19). The constant C in (11.19), however, is unspecified and inspecting the proof would reveal a rather large value. We follow now a different path based on convex geometry and Gordon's escape through a mesh theorem that leads to a direct estimate for the Ω -null space property and yields an explicit and small constant. This approach is inspired by [12] and has been applied in [20, Chapter 9.4] for the synthesis sparsity model for the first time. The next theorem was shown in [29] using additionally some ideas of [48].

Theorem 16. Let $A \in \mathbb{R}^{m \times n}$ be a Gaussian random matrix, $0 < \rho < 1$, $0 < \theta < 1$ and $\tau > 0$. If

$$\frac{m^2}{m+1} \geq \frac{2bk(2+\rho^{-1})^2}{a} \left(\sqrt{\ln \frac{ep}{k}} + \frac{1}{\sqrt{2}} + \frac{1}{2+\rho^{-1}} \sqrt{\frac{a \ln(\theta^{-1})}{bk}} + \frac{1}{\tau(2+\rho^{-1})} \sqrt{\frac{a}{2bk}} \right)^2,$$

then with probability at least $1 - \theta$ for every vector $x \in \mathbb{R}^n$ and perturbed measurements $y = Ax + e$ with $\|e\|_2 \leq \varepsilon$ a minimizer \hat{x} of (11.7) approximates x with ℓ_2 -error

$$\|x - \hat{x}\|_2 \leq \frac{2(1+\rho)^2}{\sqrt{a}(1-\rho)} \frac{\sigma_k(\Omega x)_1}{\sqrt{k}} + \frac{2\tau\sqrt{b}(3+\rho)}{\sqrt{a}(1-\rho)} \varepsilon.$$

Proof (Sketch). We verify that A satisfies the ℓ_2 -stable Ω -null space property (11.8). To this end we introduce the set

$$W_{\rho,k} := \left\{ w \in \mathbb{R}^n : \|\Omega_{\Lambda^c} w\|_2 \geq \rho/\sqrt{k} \|\Omega_\Lambda w\|_1 \text{ for some } \Lambda \subset [p], |\Lambda| = p - k \right\}.$$

If

$$\inf \left\{ \|Aw\|_2 : w \in W_{\rho,k} \cap S^{n-1} \right\} > \frac{1}{\tau}, \quad (11.35)$$

then for any $w \in \mathbb{R}^n$ such that $\|Aw\|_2 \leq \frac{1}{\tau} \|w\|_2$ and any set $\Lambda \subset [p]$ with $|\Lambda| \geq p - k$ it holds

$$\|\Omega_{\Lambda^c} w\|_2 < \frac{\rho}{\sqrt{k}} \|\Omega_\Lambda w\|_1.$$

For the remaining vectors $w \in \mathbb{R}^n$, we have $\|Aw\|_2 > \frac{1}{\tau} \|w\|_2$, which together with the fact that Ω is a frame with upper frame bound b leads to

$$\|\Omega_{\Lambda^c} w\|_2 \leq \|\Omega w\|_2 \leq \sqrt{b} \|w\|_2 < \tau \sqrt{b} \|Aw\|_2.$$

Thus, for any $w \in \mathbb{R}^n$,

$$\|\Omega_{\Lambda^c} w\|_2 < \frac{\rho}{\sqrt{k}} \|\Omega_\Lambda w\|_1 + \tau \sqrt{b} \|Aw\|_2,$$

which according to Theorem 1 guarantees stable and robust recovery.

To show (11.35), we have to study the Gaussian width of the set $W_{\rho,k} \cap S^{n-1}$. According to Theorem 11

$$\ell(W_{\rho,k} \cap S^{n-1}) \leq a^{-1/2} \ell(\Omega(W_{\rho,k} \cap S^{n-1})). \quad (11.36)$$

Since Ω is a frame with an upper frame bound b , we have

$$\Omega(W_{\rho,k} \cap S^{n-1}) \subset \Omega(W_{\rho,k}) \cap \left(\sqrt{b}B_2^p\right) \subset T_{\rho,k} \cap \left(\sqrt{b}B_2^p\right) = \sqrt{b}(T_{\rho,k} \cap B_2^p), \quad (11.37)$$

with

$$T_{\rho,k} = \left\{ u \in \mathbb{R}^p : \|u_S\|_2 \geq \rho/\sqrt{k}\|u_{S^c}\|_1 \text{ for some } S \subset [p], |S| = k \right\}.$$

Inspired by [48] it was shown in [29, Section 3.2] that

$$\ell(\Omega(W_{\rho,k} \cap S^{n-1})) \leq \sqrt{b}(2 + \rho^{-1}) \left(\sqrt{2k \ln \frac{ep}{k}} + \sqrt{k} \right).$$

Together with (11.36) this gives

$$\ell(W_{\rho,k} \cap S^{n-1}) \leq \sqrt{\frac{b}{a}}(2 + \rho^{-1}) \left(\sqrt{2k \ln \frac{ep}{k}} + \sqrt{k} \right). \quad (11.38)$$

An application of Theorem 10 and inequality (11.38) complete the proof. \square

Roughly speaking, with high probability every ℓ -cosparse vector can be recovered via analysis ℓ_1 -minimization using a single random draw of a Gaussian matrix if

$$m > 18(b/a)k \ln(ep/k). \quad (11.39)$$

Moreover, the recovery is stable under passing to approximately cosparse vectors when adding slightly more measurements.

With the approach outlined in Section 11.4.3 it is possible to give a bound on the number of subgaussian measurements. We have to combine Theorem 14 with an estimate (11.38) of the Gaussian width of the set $W_{\rho,k} \cap S^{n-1}$.

Acknowledgements M. Kabanava and H. Rauhut acknowledge support by the European Research Council through the grant StG 258926.

References

1. Benedetto, J.J., Heil, C., Walnut, D.F.: Differentiation and the Balian–Low theorem. *J. Fourier Anal. Appl.* **1**(4), 355–402 (1994)
2. Blumensath, T.: Sampling and reconstructing signals from a union of linear subspaces. *IEEE Trans. Inf. Theory* **57**(7), 4660–4671 (2011)
3. Blumensath, T., Davies, M.E.: Sampling theorems for signals from the union of finite-dimensional linear subspaces. *IEEE Trans. Inf. Theory* **55**(4), 1872–1882 (2009)

4. Burger, M., Osher, S.: Convergence rates of convex variational regularization. *Inverse Prob.* **20**(5), 1411–1421 (2004)
5. Cai, J.-F., Xu, W.: Guarantees of total variation minimization for signal recovery. In: Proceedings of 51st Annual Allerton Conference on Communication, Control, and Computing, 1266–1271 (2013)
6. Candès, E.J., Donoho, D.L.: New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Commun. Pure Appl. Math.* **57**(2), 219–266 (2004)
7. Candès, E.J., Eldar, Y.C., Needell, D., Randall, P.: Compressed sensing with coherent and redundant dictionaries. *Appl. Comput. Harmon. Anal.* **31**(1), 59–73 (2011)
8. Candès, E.J., Tao, J.T., Romberg, J.K.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**(2), 489–509 (2006)
9. Carrillo, R.E., McEwen, J.D., Van De Ville, D., Thiran, J.-Ph., Wiaux, Yv.: Sparsity averaging for compressive imaging. *IEEE Signal Process Lett.* **20**(6), 591–594 (2013)
10. Casazza, P.G., Kutyniok, G. (eds.): *Finite frames. Theory and applications. Applied and Numerical Harmonic Analysis*. Birkhäuser/Springer, New York (2013)
11. Chan, T., Shen, J.: *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods*. SIAM, Philadelphia (2005)
12. Chandrasekaran, V., Recht, B., Parrilo, P.A., Willsky, A.S.: The convex geometry of linear inverse problems. *Found. Comput. Math.* **12**(6), 805–849 (2012)
13. Chen, Y., Pock, Th., Bischof, H.: Learning ℓ_1 -based analysis and synthesis sparsity priors using bi-level optimization. Preprint arXiv:1401.4105 (2014)
14. Christensen, O.: *An Introduction to Frames and Riesz Bases. Applied and Numerical Harmonic Analysis*. Birkhäuser, Boston (2003)
15. Davenport, M., Wakin, M.: Analysis of orthogonal matching pursuit using the restricted isometry property. *IEEE Trans. Inf. Theory* **56**(9), 4395–4401 (2010)
16. Duffin, R.J., Schaeffer, A.C.: A class of nonharmonic Fourier series. *Trans. Am. Math. Soc.* **72**(2), 341–366 (1952)
17. Elad, M., Milanfar, P., Rubinstein, R.: Analysis versus synthesis in signal priors. *Inverse Prob.* **23**(3), 947–968 (2007)
18. Fadili, J., Peyré, G., Vaiter, S., Deledalle, C.-A., Salmon, J.: Stable recovery with analysis decomposable priors. In: 10th international conference on Sampling Theory and Applications (SampTA 2013), pp. 113–116, Bremen, Germany (2013)
19. Foucart, S.: Stability and robustness of ℓ_1 -minimizations with Weibull matrices and redundant dictionaries. *Linear Algebra Appl.* **441**, 4–21 (2014)
20. Foucart, S., Rauhut, H.: *A Mathematical Introduction to Compressive Sensing. Applied and Numerical Harmonic Analysis*. Birkhäuser, Boston (2013)
21. Fuchs, J.J.: On sparse representations in arbitrary redundant bases. *IEEE Trans. Inf. Theory* **50**(6):1341–1344 (2004)
22. Giryes, R., Nam, S., Elad, M., Gribonval, R., Davies, M.E.: Greedy-Like algorithms for the cosparse analysis model. *Linear Algebra Appl.* **441**, 22–60 (2014)
23. Gordon, Y.: On Milman’s inequality and random subspaces which escape through a mesh in \mathbb{R}^n . In: *Geometric Aspects of Functional Analysis (1986/87)* **1317** of Lecture Notes in Mathematics, pp. 84–106, Springer, Berlin (1988)
24. Grasmair, M.: Linear convergence rates for Tikhonov regularization with positively homogeneous functionals. *Inverse Prob.* **27**(7), 075014 (2011)
25. Gröchenig, K.: *Foundations of time-frequency analysis. Appl. Numer. Harmon. Anal.* Birkhäuser, Boston (2001)
26. Guo, K., Kutyniok, G., Labate, D.: Sparse multidimensional representations using anisotropic dilation and shear operators In: Chen, G., Lai, M. (eds.), *Wavelets and Splines: Athens 2005, Proceedings of the International Conference on the Interactions Between Wavelets and Splines*, Athens, GA, May 16–19 (2005)
27. Haltmeier, M.: Stable signal reconstruction via ℓ^1 -minimization in redundant, non-tight frames. *IEEE Trans. Signal Process.* **61**(2), 420–426 (2013)

28. Hawe, S., Kleinsteuber, M., Diepold, Kl.: Analysis operator learning and its application to image reconstruction. *IEEE Trans. Image Process.* **22**(6), 2138–2150 (2013)
29. Kabanova, M., Rauhut, H.: Analysis ℓ_1 -recovery with frames and Gaussian measurements. Preprint arXiv:1306.1356 (2013)
30. Kabanova, M., Rauhut, H., Zhang, H.: Robust analysis ℓ_1 -recovery from Gaussian measurements and total variation minimization. Preprint arXiv:1407.7402 (2014)
31. Krahmer, F., Mendelson, S., Rauhut, H.: Suprema of chaos processes and the restricted isometry property. *Commun. Pure Appl. Math.* **67**(11), 1877–1904 (2014)
32. Krahmer, F., Ward, R.: Stable and robust sampling strategies for compressive imaging. *IEEE Trans. Image Process.* **23**(2), 612–622 (2014)
33. Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Springer, Berlin, Heidelberg, New York (1991)
34. Liu, Y., Mi, T., Li, Sh.: Compressed sensing with general frames via optimal-dual-based ℓ_1 -analysis. *IEEE Trans. Inf. Theory* **58**(7), 4201–4214 (2012)
35. Mallat, S.: A Wavelet Tour of Signal Processing: The Sparse Way. Academic, San Diego (2008)
36. Mendelson, S., Pajor, A., Tomczak-Jaegermann, N.: Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.* **17**(4), 1248–1282 (2007)
37. Nam, S., Davies, M.E., Elad, M., Gribonval, R.: Cosparse analysis modeling – uniqueness and algorithms. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2011)
38. Nam, S., Davies, M.E., Elad, M., Gribonval, R.: The cosparse analysis model and algorithms. *Appl. Comput. Harmon. Anal.* **34**(1), 30–56 (2013)
39. Needell, D., Ward, R.: Stable image reconstruction using total variation minimization. *SIAM J. Imag. Sci.* **6**(2), 1035–1058 (2013)
40. Peleg, T., Elad, M.: Performance guarantees of the thresholding algorithm for the cosparse analysis model. *IEEE Trans. Inf. Theory* **59**(3), 1832–1845 (2013)
41. Rauhut, H.: Random sampling of sparse trigonometric polynomials. *Appl. Comput. Harmon. Anal.* **22**(1), 16–42 (2007)
42. Rauhut, H.: Circulant and Toeplitz matrices in compressed sensing. In: Proceedings on SPARS'09, Saint-Malo, France (2009)
43. Rauhut, H.: Compressive sensing and structured random matrices. In: Fornasier M. (ed.) Theoretical Foundations and Numerical Methods for Sparse Recovery, Radon Series on Computational and Applied Mathematics, vol. 9, pp. 1–92. deGruyter (2010)
44. Rauhut, H., Romberg, J.K., Tropp, J.A.: Restricted isometries for partial random circulant matrices. *Appl. Comput. Harmon. Anal.* **32**(2), 242–254 (2012)
45. Romberg, J.K.: Compressive sensing by random convolution. *SIAM J. Imag. Sci.* **2**(4), 1098–1128 (2009)
46. Ron, A., Shen, Z.: Affine systems in $L_2(\mathbb{R}^d)$: the analysis of the analysis operator. *J. Funct. Anal.* **148**(2), 408–447 (1997)
47. Rubinstein, R., Peleg, T., Elad, M.: Analysis K-SVD: a dictionary-learning algorithm for the analysis sparse model. *IEEE Trans. Signal Process.* **61**(3), 661–677 (2013)
48. Rudelson, M., Vershynin, R.: On sparse reconstruction from Fourier and Gaussian measurements. *Commun. Pure Appl. Math.* **61**(8), 1025–1045 (2008)
49. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Phys. D* **60**, 259–268 (1992)
50. Selesnick, I., Figueiredo, M.: Signal restoration with overcomplete wavelet transforms: comparison of analysis and synthesis priors. *Proc. SPIE* **7446**, 74460D (2009)
51. Tropp, J.A.: Recovery of short, complex linear combinations via ℓ_1 minimization. *IEEE Trans. Inf. Theory* **51**(4), 1568–1570 (2005)
52. Vaite, S., Golbabaei, M., Fadili, J.M., Peyré, G.: Model selection with low complexity priors. Preprint arXiv:1307.2342 (2013)
53. Vaite, S., Peyré, G., Dossal, Ch., Fadili, J.: Robust sparse analysis regularization. *IEEE Trans. Inf. Theory* **59**(4), 2001–2016 (2013)

Chapter 12

Structured Sparsity: Discrete and Convex Approaches

Anastasios Kyrillidis, Luca Baldassarre, Marwa El Halabi,
Quoc Tran-Dinh, and Volkan Cevher

Abstract During the past decades, sparsity has been shown to be of significant importance in fields such as compression, signal sampling and analysis, machine learning, and optimization. In fact, most natural data can be *sparsely* represented, i.e., a small set of coefficients is sufficient to describe the data using an appropriate basis. Sparsity is also used to enhance interpretability in real-life applications, where the relevant information therein typically resides in a low dimensional space.

However, the true underlying structure of many signal processing and machine learning problems is often more sophisticated than sparsity alone. In practice, what makes applications differ is the existence of sparsity patterns among coefficients. In order to better understand the impact of such *structured sparsity patterns*, in this chapter we review some realistic sparsity models and unify their convex and non-convex treatments. We start with the general group sparse model and then elaborate on two important special cases: the dispersive and hierarchical models. We also consider more general structures as defined by set functions and present their convex proxies. Further, we discuss efficient optimization solutions for structured sparsity problems and illustrate structured sparsity in action via three applications in image processing, neuronal signal processing, and confocal imaging.

12.1 Introduction

Information in many natural and man-made signals can be exactly represented or well approximated by a sparse set of nonzero coefficients in an appropriate basis [82]. This fact has found many applications in practice: Compressive sensing (CS) [23, 36] exploits sparsity to recover signals from their compressive samples through dimensionality-reducing, non-adaptive sensing mechanisms. From a different perspective, sparsity is used to enhance *interpretability* in machine learning [48, 53] and statistics [115] applications. For example, while the ambient dimension in feature selection applications might be huge in modern data analysis,

A. Kyrillidis • L. Baldassarre (✉) • M. El Halabi • Q. Tran-Dinh • V. Cevher
EPFL, Lausanne, Switzerland
e-mail: anastasios.kyrillidis@epfl.ch; luca.baldassarre@epfl.ch; marwa.elhalabi@epfl.ch;
quoc.trandinh@epfl.ch; volkan.cevher@epfl.ch

often a sparse set of features is sufficient to explain well the observations. Such observations have led to several new theoretical and algorithmic developments in different communities, including theoretical computer science [47], neuronal imaging [52, 68], bioinformatics [100, 105, 114, 134]. In all these disciplines, given a data set to analyze, one is usually interested in the *simplest* model that well explains the observations.

12.1.1 Why structured sparsity?

While many of the optimization solutions proposed nowadays result in *sparse* model selections, in many cases they do not capture the true underlying structure to best explain the observations [8]. In fact, this *uninformed* selection process has been the center of attention in sparse approximation theory [75, 131], because it not only prevents interpretability of results in many problems, but also fails to exploit key prior information that could radically improve estimation performance.

During the last decade, researchers have extended the simple sparsity idea to more realistic *structured* sparsity models which describe the interdependency between the nonzero coefficients. On the one hand, the use of these models yields solutions that are easier to interpret, since they can be decomposed into meaningful “parts,” such as predefined groups of coefficients or adaptive clusters that depend on the data. On the other hand, structured sparsity models lead to better recovery performance in terms of reducing the number of samples required for stable recovery under perturbations [8, 15, 40, 104]. To highlight the importance of this property, in the case of Magnetic Resonance Imaging (MRI), reducing the total number of measurements is highly desirable for both capturing functional activities within small time periods and rendering the whole procedure less time-consuming [81].

In recent years, we have witnessed many approaches that guide the selection process: (overlapping) group Lasso, fused Lasso, greedy approaches for signal approximation under tree-structure assumptions, just to name a few; see [42, 67, 116, 130]. To showcase the potential of such approaches, consider the problem of recovering an image of n pixels from a *limited* set of m linear measurements using the tree-structured group sparse model [6], which is well suited for describing natural images. This model uses a tree to impose dependencies over the coefficients of the signal to be estimated and will be described in detail in Section 12.5. Briefly, the coefficients are arranged on a tree and the discrete structure imposes that if a node is selected, then all its ancestors must be selected as well. Therefore, for a desired number of nonzero coefficients, or *sparsity*, K , the possible valid choices are much more limited, see Fig. 12.1.

To demonstrate these ideas visually, we use a stylized CS example. Figure 12.2 shows the promise of such model, compared to the simple sparsity one, in the compressed sensing setting:

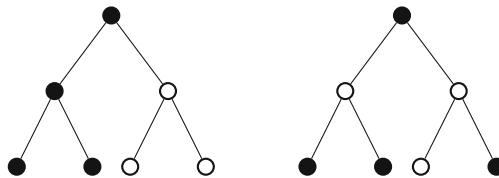


Fig. 12.1 Tree constraint. Each node represent a coefficient. (Left) A valid selection of four nodes. (Right) An *invalid* selection of four nodes.

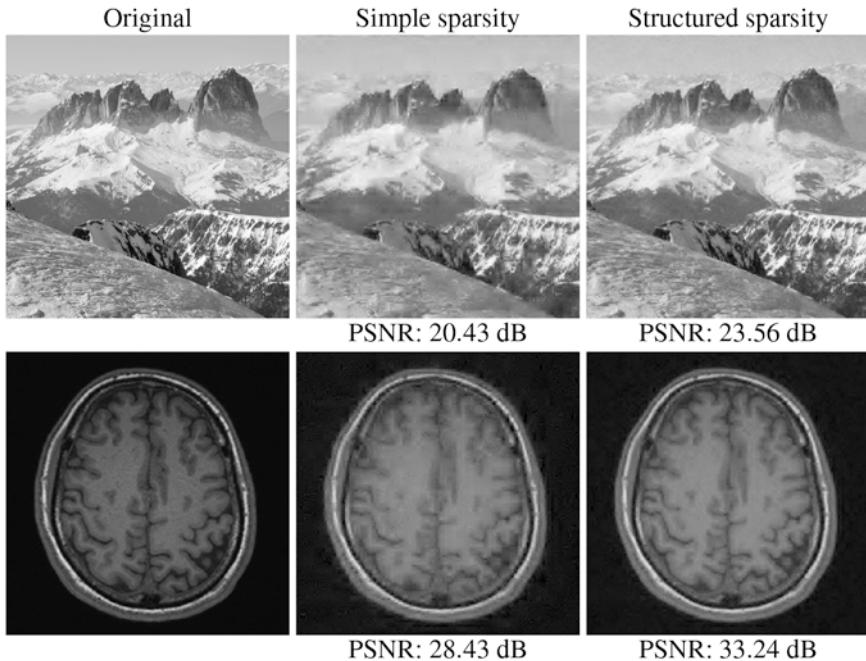


Fig. 12.2 Empirical performance of simple and structured sparsity recovery of natural images. In all cases, the number of measurements m is approximately 5% (top row) and 10% (bottom row) of the actual image dimensions n . **Left panel:** Original images of dimension: (Top row) 2048×2048 , (Bottom row) 512×512 . **Middle panel:** Conventional recovery using simple sparsity model. **Right panel:** Tree-structured sparse recovery; see Section 12.5.

$$y = Awc.$$

Here, y denotes the limited set of measurements, $x = Wc$ is the image representation with coefficients c under a 2D Wavelet basis W and A is the linear, dimensionality-reducing measurement matrix.

12.1.2 Chapter overview

In order to better understand the impact of structured sparsity, in this chapter we analyze the connections between the discrete models and their convex relaxations, highlighting their relative advantages. After setting notations and introducing some preliminaries in Section 12.2, we start with the general group sparse model in Section 12.3 and then elaborate on two important special cases: the dispersive and the hierarchical models, Sections 12.4 and 12.5, respectively. For each of these, we present the models in their discrete nature, discuss how to solve the ensuing discrete problems, and then describe convex alternatives. In Section 12.6, we consider more general structures as defined by set functions and present a recipe to obtain convex relaxations from discrete models. Further, we discuss efficient optimization solutions for structured sparsity problems in Section 12.7 and illustrate structured sparsity in action via three applications in Section 12.8. Section 12.9 gathers some concluding remarks and open questions.

12.2 Preliminaries

We use superscripts such as w^i to denote the estimate at the i th iteration of an algorithm. The sign of a scalar $\alpha \in \mathbb{R}$ is given by $\text{sign}(\alpha)$. Given a set $S \subseteq N := \{1, \dots, n\}$, the complement S^c is defined with respect to N , and its cardinality as $|S|$. The support set of w is $\text{supp}(w) = \{i \in N : w_i \neq 0\}$. Given a vector $w \in \mathbb{R}^n$, w_S is the projection (in \mathbb{R}^n) of w onto S , i.e. $(w_S)_{S^c} = 0$, whereas $w|_S \in \mathbb{R}^{|S|}$ is w limited to the entries in S . We define $\Sigma_k := \{w : w \in \mathbb{R}^n, |\text{supp}(w)| \leq k\}$ as the set of all k -sparse vectors in n -dimensions; with a slight abuse of further denote the set of k -sparse supports in N . We sometimes write $x \in \Sigma_k$ to mean with a slight abuse of notation, we use Σ_k to further denote $\text{supp}(x) \in \Sigma_k$.

We use \mathbb{B}^n to represent the space of n -dimensional binary vectors and define $\iota : \mathbb{R}^n \rightarrow \mathbb{B}^n$ to be the indicator function of the nonzero components of a vector in \mathbb{R}^n , i.e., $\iota(x)_i = 1$ if $x_i \neq 0$ and $\iota(x)_i = 0$, otherwise. We let $\mathbb{1}_n$ be the n -dimensional vector of all ones, $\mathbb{1}_{n,S}$ the n -dimensional vector of all ones projected onto S and I_n the $n \times n$ identity matrix; we often use I when the dimension is clear from the context.

Norms We define the ℓ_p^n -norm in n -dimensions as:

$$\|x\|_p = \begin{cases} (\sum_{i=1}^n |x_i|^p)^{1/p} & \text{if } p \in (0, \infty), \\ \max_i |x_i| & \text{if } p = \infty. \end{cases}$$

The ℓ_0 pseudo-norm is defined as $\|x\|_0 := |\text{supp}(x)|$.

Projections operations Assume $M_k \subset \Sigma_k$. Given M_k with sparsity level k and an anchor point $x \in \mathbb{R}^n$, a key problem in our subsequent discussions is the following projection problem:

$$\mathcal{P}_{M_k}(x) \in \operatorname{argmin}_{w \in \mathbb{R}^n} \left\{ \|w - x\|_2^2 \mid \operatorname{supp}(w) \in M_k \right\}. \quad (12.1)$$

Proximity operations Consider a given function $g(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}$. We define the proximity operator of $g(\cdot)$ as [30, eq. (2.13)]

$$\operatorname{prox}_\lambda^g(x) := \operatorname{argmin}_{w \in \mathbb{R}^n} \left\{ \frac{1}{2} \|w - x\|_2^2 + \lambda \cdot g(w) \right\}, \quad (12.2)$$

where $\lambda > 0$ is a constant weight.

Optimization preliminaries

Definition 1 (Convexity of Functions). A differentiable function $f \in \mathbb{R}^n$ is called convex on its domain if for any $v, w \in \mathbb{R}^n$:

$$f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle,$$

where $\nabla f(w)$ is the gradient of f evaluated at the point w .

Existing algorithmic solutions invariably rely on two structural assumptions on the objective function that particularly stand out among many others: the *Lipschitz continuous gradient* assumption, the *strong convexity* condition.

Definition 2 (Lipschitz Gradient Continuity). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a differentiable function. Then, f is a Lipschitz gradient function if and only if for any $v, w \in \operatorname{dom}(f)$:

$$\|\nabla f(v) - \nabla f(w)\|_2 \leq L\|v - w\|_2.$$

Here, $L > 0$ is known as the Lipschitz constant.

A useful property of Lipschitz gradient functions is given in the following Lemma.

Lemma 1. *Let $f \in \mathbb{R}^n$ be a L -Lipschitz gradient convex function. Then, we have*

$$|f(v) - f(w) - \langle \nabla f(w), v - w \rangle| \leq \frac{L}{2} \|v - w\|_2^2, \quad \forall w, v \in \mathbb{R}^n.$$

Definition 3 (Strong Convexity). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a twice differentiable convex function. Then, f is μ -strongly convex if and only if

$$\mu I \preceq \nabla^2 f(x), \quad \forall x \in \mathbb{R}^n,$$

for some global constant $\mu > 0$.

A useful property of μ -strongly convex functions is given in the next lemma.

Lemma 2. *A twice differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is μ -strongly convex if there exists a constant $\mu > 0$ such that for any $w, v \in \mathbb{R}^n$, we have*

$$f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\mu}{2} \|v - w\|_2^2$$

12.3 Sparse group models

We start our discussion with the general *group sparse* models because they can be seen as encompassing more specific models, such as the ones that we describe in the next sections. The group sparse models are based on the identification of groups of variables that should either be selected or discarded (i.e., set to zero) together [9, 60, 67, 100, 103, 104]. Such groupings of variables naturally emerge in many applications such as computer vision [8, 25], bioinformatics [105, 134], gene expression data [100, 114], and neuroimaging [52, 68]. When analyzing gene expression data for discovering disease-specific biomarkers, the groups might represent all those genes that participate in the same cellular process. These groups are also called genetic pathways. Identifying which pathways are implied on the onset of a disease can allow molecular biologists to focus their investigations on a more limited number of genes [114], speeding up the diagnosis and the development of therapies.

A group sparsity model—denoted by \mathfrak{G} —consists of a collection of groups of variables that could overlap arbitrarily; that is, $\mathfrak{G} = \{G_1, \dots, G_M\}$ where each G_j is a subset of the index set $N := \{1, \dots, n\}$.

A group structure \mathfrak{G} can be represented by a bipartite graph, with the n variable nodes on one side and the M group nodes on the other. A variable node i is connected by an edge to a group node j if $i \in G_j$. The bi-adjacency matrix $B \in \mathbb{B}^{n \times M}$ of the bipartite graph compactly encodes the group structure,

$$B_{ij} = \begin{cases} 1, & \text{if } i \in G_j \\ 0, & \text{otherwise.} \end{cases}$$

Figure 12.3 shows an example.

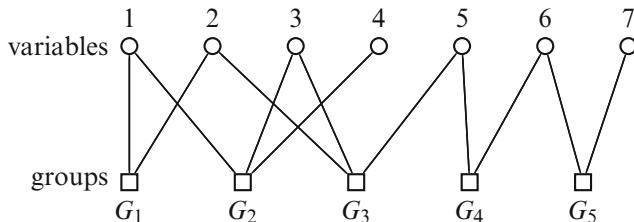


Fig. 12.3 The group structure \mathfrak{G}^1 defined by the groups $G_1 = \{1, 2\}$, $G_2 = \{1, 2, 3\}$, $G_3 = \{2, 3, 4\}$, $G_4 = \{4, 5, 6\}$, and $G_5 = \{6, 7\}$.

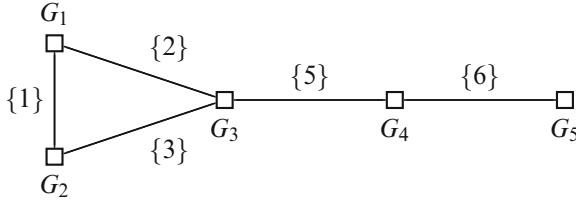


Fig. 12.4 Cyclic intersection graph induced by the group structure \mathfrak{G}^1 described in Figure 12.3, where on each edge we report the elements of the intersection.

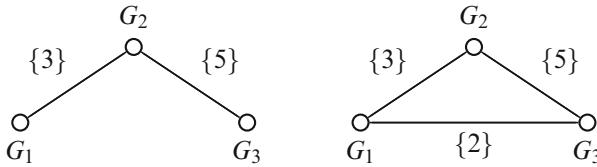


Fig. 12.5 Consider $G_1 = \{1, 2, 3\}$, $G_2 = \{3, 4, 5\}$, $G_3 = \{5, 6, 7\}$, whose intersection group is reported in the left plot. If G_3 included an element from G_1 , for example $\{2\}$, the induced intersection group would contain a cycle, as shown in the right plot.

The group structure can also be represented by a *intersection group* $(\mathcal{V}, \mathcal{E})$ where the nodes \mathcal{V} are the groups $G \in \mathfrak{G}$ and the edge set \mathcal{E} contains e_{ij} if $G_i \cap G_j \neq \emptyset$: if two groups overlap they are connected by an edge. A *cycle* in a graph is a sequence of connected nodes v_1, v_2, \dots, v_n , such that $v_1 = v_n$. For an illustrative example, see Figure 12.4.

Of particular interest are *acyclic* group structures whose intersection graph is acyclic, i.e., either a tree or a forest, see Figure 12.5. Since dynamic programs can be run efficiently on forests, this group structure leads to tractable projections [5].

Given a user-defined group budget $g \in \mathbb{Z}_+$, we define the group model $M_g := \{\bigcup_{G_\ell \in T} G_\ell, T \subseteq \mathfrak{G}, |T| \leq g\}$, containing all sets of indices that are the union of at most g groups from the collection \mathfrak{G} . Then, the projection problem (12.1) becomes

$$\hat{x} =: \mathcal{P}_{M_g}(x) \in \operatorname{argmin}_{z \in \mathbb{R}^n} \{ \|x - z\|_2^2 \mid \operatorname{supp}(z) \in M_g \}. \quad (12.3)$$

Moreover, one might be only interested in identifying the *group support* of the approximation \hat{x} , that is the G groups that constitute its support. We call this the group sparse *model selection* problem.

12.3.1 The discrete model

According to (12.3), we search for $\hat{x} \in \mathbb{R}^n$ such that $\|\hat{x} - x\|_2^2$ is minimized, while \hat{x} does not exceed a given group budget g . A useful notion in the group sparse model is that of the **group ℓ_0 -“norm”**

$$\|w\|_{\mathfrak{G},0} := \min_{\omega \in \mathbb{B}^M} \left\{ \sum_{j=1}^M \omega_j \mid B\omega \geq \iota(w) \right\}, \quad (12.4)$$

where, B denotes the adjacency matrix as defined in the previous subsection, the binary vector ω indicates which groups are selected and the constraint $B\omega \geq \iota(w)$ imposes that, for every nonzero component of w , at least one group that contains it is selected. Given the above definitions, the group-based signal approximation problem (12.3) can be reformulated as

$$\hat{x} \in \operatorname{argmin}_{w \in \mathbb{R}^n} \left\{ \|w - x\|_2^2 \mid \|w\|_{\mathfrak{G},0} \leq g \right\}. \quad (12.5)$$

One can easily observe that, in the case where we already know the active groups in the approximation \hat{x} , we can obtain \hat{x} as $\hat{x}_T = x_T$ and $\hat{x}_{T^c} = 0$, where $T = \bigcup_{G \in S^g(\hat{x})} G$, with $S^g(\hat{x})$ denoting the group support of \hat{x} and $T^c = N \setminus T$. That is, if we know the group support of the solution, the entries' values are naturally given by the anchor point x . The authors in [5] prove that the group support can be obtained by solving the discrete model selection problem, according to the next lemma.

Lemma 3 ([5]). *Given $x \in \mathbb{R}^n$ and a group structure \mathfrak{G} , the group support of the solution \hat{x} to (12.5)—denoted by $S^g(\hat{x}) = \{G_j \in \mathfrak{G} : \omega_j^g = 1\}$ —is given by the solution (ω^g, h^g) of the following binary maximization problem:*

$$\max_{\omega \in \mathbb{B}^M, h \in \mathbb{B}^n} \left\{ \sum_{i=1}^n h_i x_i^2 : B\omega \geq h, \sum_{j=1}^M \omega_j \leq g \right\}. \quad (12.6)$$

In (12.6), h denotes the selected variables in x , while ω is the group support indicator vector for x . Thus, the constraint $B\omega \geq h$ guarantees that, for every selected variable, there is at least one group that covers it.

The problem in (12.6) can produce all the instances of the weighted maximum coverage problem (WMC) [92], where the weights for each element are given by x_i^2 ($1 \leq i \leq n$) and the index sets are given by the groups $G_j \in \mathfrak{G}$ ($1 \leq j \leq M$). Since WMC is in general NP-hard [92] finding the groups that cover the solution \hat{x} is in general NP-hard.

However, it is possible to approximate the solution of (12.6) using the greedy WMC algorithm [93]. This algorithm iteratively selects the group that contains new variables with maximum combined weight until g groups have been selected. The

work in [5] presents a polynomial time algorithm for solving (12.6) for acyclic groups structures, using dynamic programming. The algorithm gradually explores the intersection group, which in this case is a tree or a forest of trees, from the leaves towards the root, storing the best solutions found so far and dynamically updating them until the entire tree is examined.

12.3.2 Convex approaches

The literature in compressive sensing and machine learning with group sparsity has mainly focused on leveraging the group structures for lowering the number of samples required for stable recovery of signals [8, 15, 40, 61, 64, 100, 104, 113].

For the special case of the block-sparsity model, i.e. non-overlapping groups, the problem of finding the group support, i.e. the model selection problem, is computationally tractable and offers a well-understood theory [113]. The first convex relaxations for group-sparse approximation [130] considered only non-overlapping groups: the authors proposed the **Group LARS** (Least Angle RegreSsion) algorithm to solve this problem, a natural extension of simple sparsity LARS algorithm [38]. Using the same algorithmic principles, its extension to overlapping groups [133] has the disadvantage of selecting supports defined as the complement of a union of groups, even though it is possible to engineer the groups in order to favor certain sparsity patterns over others [67]. Eldar *et al.* [40] consider the union of subspaces framework and reformulate the model selection problem as a block-sparse model selection one by duplicating the variables that belong to the overlaps between the groups, which is the optimization approach proposed also in [64]. Moreover, [40] considers a model-based pursuit approach [29] as potential solver for this problem, based on a predefined model M_k . For these cases, one uses the group lasso norm

$$\sum_{G \in \mathfrak{G}} \|x|_G\|_p, \quad (12.7)$$

where $x|_G$ is the restriction of x to only the components indexed by G and $1 \leq p \leq \infty$.

In addition, convex proxies to the group ℓ_0 -norm (12.4) have been proposed (e.g., [64]) for finding group-sparse approximations of signals. Given a group structure \mathfrak{G} , an example generalization is

$$\|x\|_{\mathfrak{G},\{1,p\}} := \inf_{\substack{v_1, \dots, v_M \in \mathbb{R}^n \\ \forall i, \text{supp}(v_i) = G_i}} \left\{ \sum_{i=1}^M d_i \|v_i\|_p \mid \sum_{i=1}^M v_i = x \right\}, \quad (12.8)$$

where $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ is the ℓ_p -norm, and d_j are positive weights that can be designed to favor certain groups over others [100]. This norm can be seen as a weighted instance of the atomic norm described in [28, 104], based on the

seminal work of [29] on the sparse synthesis model. There, the authors leverage convex optimization for signal recovery, but not for model selection: Let $\mathcal{A} := \{a_1, a_2, \dots \mid a_i \in \mathbb{R}^n, \forall i\}$ be the *atomic set* of *group sparse* signals x that can be synthesized as a k -sparse combination of atoms a_i in \mathcal{A} , i.e.,

$$x = \sum_{i=1}^k c_i a_i, \quad c_i \geq 0, a_i \in \mathcal{A} \text{ and } \|c\|_0 \leq k. \quad (12.9)$$

Here, $a_i \in \mathbb{R}^n$, $\text{supp}(a_i) = \mathfrak{G}_i$ and $\|a_i\|_2 = 1$. We can then define the atomic norm

$$\|x\|_{\mathcal{A}} = \inf \left\{ \sum_{i=1}^{|\mathcal{A}|} c_i \mid x = \sum_{i=1}^{|\mathcal{A}|} c_i a_i, c_i \geq 0, \forall a_i \in \mathcal{A} \right\}. \quad (12.10)$$

Lemma 4 ([28, 104]). *If in (12.8) the weights are all equal to 1 ($d_i = 1, \forall i$), we have*

$$\|x\|_{\mathcal{A}} = \|x\|_{\mathfrak{G}, \{1, p\}}.$$

The group-norm (12.8) can also be viewed as the tightest convex relaxation of a particular set function related to the *weighted set-cover* (see Section 12.6 and [99]).

One can in general use (12.8) to find a group-sparse approximation under the chosen group norm

$$\hat{x} \in \operatorname{argmin}_{w \in \mathbb{R}^N} \left\{ \|w - x\|_2^2 : \|w\|_{\mathfrak{G}, \{1, p\}} \leq \lambda \right\}, \quad (12.11)$$

where $\lambda > 0$ controls the trade-off between approximation accuracy and group-sparsity. However, solving (12.11) does not necessarily yield a group support for \hat{x} : even though we can recover one through the decomposition $\{v_j\}$ used to compute $\|\hat{x}\|_{\mathfrak{G}, \{1, p\}}$, it may not be unique and when it is unique it may not capture the minimal group-cover of x [100].

The regularized version of problem (12.11) is equivalent to the proximity operator of $\|x\|_{\mathfrak{G}, \{1, p\}}$. Recently, [88, 124] proposed an efficient algorithm for this proximity operator in large scale settings with extended overlap among groups. In this case, the proximity operator involves: (i) an active set preprocessing step [128] that restricts the prox operations on a subset of the model—i.e., “active” groups and, (ii) a dual optimization step based on Bertsekas’ projected Newton method [12]; however, its convergence requires the strong regularity of the Hessian of the objective near the optimal solution. The authors in [1] propose a fixed-point method to compute the proximity operator for a wide range of group sparse model variants, given the model-structured mapping of the fixed-point Picard iterations is non-expansive. [129] develops an efficient primal–dual criterion for the overlapping group sparse proximity operator: the solution of its dual formulation is used to compute the duality gap and tweak the accuracy of the solution and, thus, the convergence of the algorithm.

12.3.3 Extensions

In many applications, such as multinomial classification [126] and genome-wide association studies [134], it is desirable to find solutions that are both group sparse and sparse in the usual sense (see [112] for an extension of the group lasso). A classic illustrative example is the sparse group lasso approach [42], a regularization method that combines the lasso [115] and the group lasso [86].

From a discrete perspective, the original problem (12.6) can be generalized by introducing a sparsity constraint k and allowing to individually select variables within a group. The generalized integer problem then becomes

$$\max_{\omega \in \mathbb{B}^M, h \in \mathbb{B}^n} \left\{ \sum_{i=1}^n h_i x_i^2 \mid B\omega \geq h, \sum_{i=1}^n h_i \leq k, \sum_{j=1}^M \omega_j \leq g \right\}. \quad (12.12)$$

While this problem is also NP-hard in general', the dynamic program that solves (12.6) can be adapted to solve the general problem in polynomial time for acyclic group structures.

Proposition 1 ([5]). *Given a acyclic group structure \mathfrak{G} , there exists a dynamic programming algorithm that solves (12.12) with complexity linear in n and k .*

From a convex point of view, Simon et al. [112] are the first to propose the sparse group model regularization in the context of linear regression. Given a group model \mathfrak{G} and constants $\lambda_1, \lambda_2 > 0$, they consider the problem

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) + \lambda_1 \sum_{G \in \mathfrak{G}} \|x|_G\|_2 + \lambda_2 \|x\|_1 \right\}.$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a data model fidelity term, e.g., $f(x) = \|y - Bx\|_2^2$; the selection of f depends on the problem at hand. For this purpose, a generalized gradient descent algorithm is proposed. Vincent and Hansen [126] provide an efficient and robust sparse group lasso algorithm; in this case, a penalized quadratic approximation of the loss function is optimized via a Newton-type algorithm in a coordinate descent framework [129].

12.4 Sparse dispersive models

To describe the *dispersive* structure, we motivate our discussion with an application from neurobiology. Living beings function via transmission of electrical signals between electrically excitable neuronal brain cells. Such chemical “information” causes a swift change in the electrical potential of a possibly discharged neuron cell, which results in its electrical excitation. Currently, we are far from understanding the grid of neurons in its *entirety*: large-scale brain models are difficult to handle while complex neuronal signal models lead to non-interpretable results.

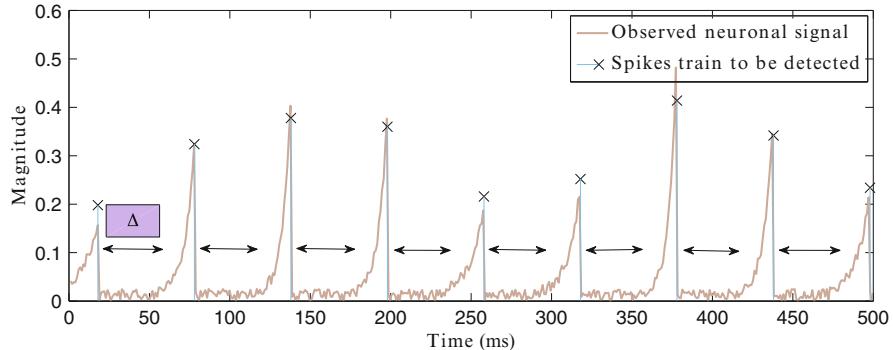


Fig. 12.6 Neuronal spike train example.

Inspired by the statistical analysis in [46], the authors in [57] consider a simple one-dimensional model, where the neuronal signal behaves as a train of spike signals with some *refractoriness* period $\Delta > 0$ [9]: there is a minimum nonzero time period Δ where a neuron remains inactive between two consecutive electrical excitations. In statistical terms, neuronal signals are defined by an inter-spike interval distribution that characterizes the probability that a new spike is generated as a function of the inter-arrival time. Figure 12.6 illustrates how a collection of noisy neuronal spike signals with $\Delta > 0$ might appear in practice.

12.4.1 The discrete model

We provide next a formal definition of the sparse dispersive model:

Definition 4 (Dispersive Model). The dispersive model D_k in n dimensions with sparsity level k and refractory parameter $\Delta \in \mathbb{Z}_+$ is given by

$$D_k = \{S_q \mid \forall q, S_q \subseteq N, |S_q| \leq k \text{ and } |i - j| > \Delta, \forall i, j \in S_q, i \neq j\}, \quad (12.13)$$

i.e., D_k is a collection of index subsets S_q in N with number of elements no greater than k and with distance between the indices in S_q greater than the interval Δ .

We note that if there are no constraints on the interval of consecutive spikes, the dispersive model naturally boils down to the simple sparsity model Σ_k .

Given the definition above, the projection (12.1) is

$$\mathcal{P}_{D_k}(x) \in \operatorname{argmin}_{w \in \mathbb{R}^n} \{ \|w - x\|_2^2 \mid \operatorname{supp}(w) \in D_k \}. \quad (12.14)$$

Let $\omega \in \mathbb{B}^n$ be a *support indicator* binary vector, i.e., ω represents the support set of a sparse vector x such that $\text{supp}(\omega) = \text{supp}(x)$. Moreover, let $D \in \mathbb{B}^{(n-\Delta+1) \times n}$ such that:

$$D = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 1 & 1 & 0 & \cdots & 0 \\ & & & \ddots & & & & & \\ 0 & \cdots & 0 & 0 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix}_{(n-\Delta+1) \times n} \quad (12.15)$$

Here, per row, there are Δ consecutive ones that denote the time interval between two potential consecutive spikes. Finally, let $b \in \mathbb{R}^{n-\Delta+2}$ such that $b := [k \ 1 \ 1 \ \cdots \ 1 \ 1]^T$.

According to [57], the following linear support constraints encode the definition of the dispersive model D_k :

$$B\omega := \begin{bmatrix} \mathbf{1} \\ D \end{bmatrix} \omega \leq b. \quad (12.16)$$

One can observe that $D_k \equiv \{\bigcup_{\omega \in \mathfrak{Z}} \text{supp}(\omega) \mid \mathfrak{Z} := \{\omega \in \mathbb{B}^n : B\omega \leq b\}\}$. To this end, (12.14) becomes:

$$\mathcal{P}_{D_k}(x) \in \operatorname{argmin}_{w \in \mathbb{R}^n} \left\{ \|w - x\|_2^2 \mid B \cdot \mathbf{1}(w) \leq b \right\}. \quad (12.17)$$

A key observation is given in the following lemma.

Lemma 5 ([57]). *Given the problem setting above, (12.17) has solution $\mathcal{P}_{D_k}(x)$ such that $S := \text{supp}(\mathcal{P}_{D_k}(x))$ and $(\mathcal{P}_{D_k}(x))_S = x_S$, where*

$$S \in \text{supp} \left(\operatorname{argmax}_{\omega \in \mathbb{B}^n: B\omega \leq b} \{c^T \omega\} \right), \quad \text{and } c := [x_1^2 \ x_2^2 \ \cdots \ x_n^2]^T, \quad (12.18)$$

i.e., we target to capture most of the signal's x energy, given structure D_k . To solve (12.18), the binary integer program (12.18) is identical to the solution of the linear program, obtained by relaxing the integer constraints into continuous constraints.

Lemma 5 indicates that (12.17) can be efficiently performed using linear programming tools [20]. Once (12.17) is relaxed to a convex problem, decades of knowledge on convex analysis and optimization can be leveraged. Interior point methods find a solution with fixed precision in polynomial time, but their complexity might be prohibitive even for moderate-size problems.

12.4.2 Convex approaches

The constraint matrix D describes a *collection of groups*, where each group is assumed to have at most one nonzero entry to model the refractoriness property.¹ Moreover, these groups are *overlapping* which aggrandizes the “clash” between neighboring groups: a nonzero entry in a group discourages every other overlapping group from having a distinct nonzero entry.

In mathematical terms, each row i of D defines a group G_i such that $G_i = \text{supp}(d_i) \subseteq N$ where d_i denotes the i th row of D , $\forall i \in \{1, \dots, M := n - \Delta + 1\}$:

$$D = \begin{bmatrix} & & \Delta & & \\ & \boxed{G_1} & & & \\ & \boxed{G_2} & & & \\ & & \ddots & \ddots & \\ & & & & \boxed{G_M} \end{bmatrix}$$

Given such group structure, the dispersive model is characterized both by *inter-group* and *intra-group* properties:

- *Intra-group sparsity*: we desire $\|D\omega\|_\infty \leq 1$, i.e., per refractoriness period of length Δ , we require only one “active” spike.
- *Inter-group exclusion*: due to the refractoriness property, the activation of a group implies the deactivation of its closely neighboring groups.

While the sparsity level within a group can be easily “convexified” using standard ℓ_1 -norm regularization, the dispersive model further introduces the notion of *inter-group exclusion*, which is highly *combinatorial*. However, one can relax it by introducing *competitions* among variables in overlapping groups: variables that have a “large” neighbor should be penalized more than variables with “smaller” neighbors.

In this premise and based on [135], we identify the following family of norms²:

¹Other convex structured models that can be described as the composition of a simple function over a linear transformation D can be found in [1].

²The proposed norm originates from the composite absolute penalties (CAP) convex norm, proposed in [133], according to which

$$g(x) = \sum_{G_i} \left(\sum_{j \in G_i} |x_j|^\gamma \right)^p, \quad (12.19)$$

for various values of γ and p . Observe that this model also includes the famous group sparse model where $g(x) = \sum_{G_i} \|x_{G_i}\|_2$, described in Section 12.3, for $p = 1/2$ and $\gamma = 2$.

$$\Omega_{\text{exclusive}}(x) = \sum_{G_i} \left(\sum_{j \in G_i} |x_j| \right)^p, \quad p = 2, 3, \dots, \quad (12.20)$$

as convex regularizers that imitate the dispersive model. In (12.20), $(\sum_{j \in G_i} |x_j|) := \|x_{G_i}\|_1$ promotes sparsity within each group G_i , while the outer sum over groups $\sum_{G_i} \|x_{G_i}\|_1^p$ imposes sparsity over the number of groups that are activated. Observe that for $p = 1$, (12.20) becomes the standard ℓ_1 -norm over N . Notice that the definition of the overlapping groups (instead of non-overlapping) is a key property for capturing the discrete structure: variables belonging to overlapping groups are weighted differently when considered parts of different groups. This leads to variable “suppression” (i.e., thresholding) of elements, depending on the “weight” of their neighborhood within the groups they belong to.

12.5 Hierarchical sparse models

Hierarchical structures are found in many signals and applications. For example, the wavelet coefficients of images are naturally organized on regular quad-trees³ to reflect their multi-scale structure, c.f. Figure 12.7 and [6–8, 31, 55, 61, 82, 110, 133]; gene networks are described by a hierarchical structure that can be leveraged for multi-task regression [71]; hierarchies of latent variables are typically used for deep learning [11].

In essence, a hierarchical structure defines an ordering of importance among the elements (either individual variables or groups of them) of a signal with the rule



Fig. 12.7 Wavelet coefficients naturally cluster along a rooted connected subtree of a regular tree and tend to decay towards the leaves. (Left) Example of wavelet tree for a 32×32 image. The root of the tree is the top-left corner, and there are three regular subtrees related to horizontal, vertical, and diagonal details. Each node is connected to four children representing detail at a finer scale. (Centre) Grayscale 512×512 image. (Right) Wavelet coefficients for the image at center. Best viewed in color, dark blue represents values closer to zero.

³A regular quad-tree is a finite tree whose nodes have exactly four children, leaves excluded.

that an element can be selected only after its ancestors. Such structured models result into more robust solutions and allow recovery with far fewer samples. In compressive sensing, assuming that the signal possesses a hierarchical structure with sparsity k leads to improved sample complexity bounds of the order of $O(k)$ for dense measurement matrices [8], compared to the bound of $O(k \log(n/k))$ for standard sparsity. Also in the case of sparse measurement matrices, e.g. expanders, hierarchical structures yield improved sample complexity bounds [4, 62].

12.5.1 The discrete model

The discrete model underlying the hierarchical structure is given by the next definition.

Definition 5 (Hierarchical Model). Let T denote an arbitrary *tree* or *forest* representation over the variables in a set N . We define a k rooted connected (RC) subtree S with respect to T as a collection of k variables in N such that $v \in S$ implies $\mathcal{A}(v) \in S$, where $\mathcal{A}(v)$ is the set that contains all the ancestors of the node v .

The hierarchical model of budget k , T_k , is the set of all k rooted-connected subtrees of T .

An example of rooted connected subtree over $|N| = 9$ variables is given in Figure 12.8.

Given a tree T , the rooted connected approximation is the solution of the following discrete problem

$$\mathcal{P}_{T_k}(x) \in \operatorname{argmin}_{z \in \mathbb{R}^n} \left\{ \|x - z\|_2^2 \mid \operatorname{supp}(z) \in T_k \right\}, \quad (12.21)$$

which can be reformulated as follows:

$$\hat{h} \in \operatorname{argmax}_{h \in \mathbb{B}^n} \left\{ \sum_{i=1}^N h_i x_i^2 : h \in T_k \right\}, \quad (12.22)$$

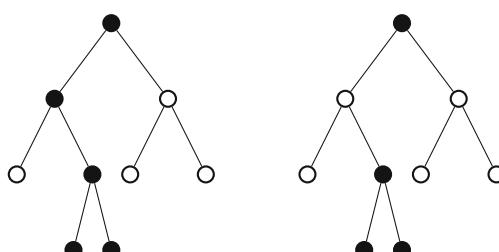


Fig. 12.8 Hierarchical constraints. Each node represents a variable. (Left) A valid selection of nodes. (Right) An *invalid* selection of nodes.

where h is a binary vector with k nonzero components that indicates which components of x are selected. Given a solution \hat{h} of the above problem, a solution \hat{z} of (12.21) is then obtained as $\hat{z}_{|S} = x_{|S}$ and $\hat{z}_{|N \setminus S} = 0$, where $S = \text{supp}(\hat{h})$.

This type of constraint can be represented by a group structure with an overall sparsity constraint k , where for each node in the tree we define a group consisting of that node and all its ancestors. When a group is selected, we require that all its elements are selected as well. Problem (12.22) can then be cast as a special case of the Weighted Maximum Coverage problem (12.6). Fortunately, this particular group structure leads to tractable solutions. Indeed, (12.22) can be solved exactly via a dynamic program that runs in polynomial time [5, 24]. For d -regular trees, that is trees for which each node has d children, the algorithm in [5] has complexity $\mathcal{O}(nkd)$.

12.5.2 Convex approaches

The hierarchical structure can also be enforced by convex penalties, based on groups of variables. Given a tree structure T , define groups consisting of a node and all its descendants and let \mathfrak{G}_T represent the set of all these groups. Based on this construction, the hierarchical group lasso penalty [69, 71, 133] imitates the hierarchical sparse model and is defined as follows:

$$\Omega(x)_{\text{HGL}} = \sum_{G \in \mathfrak{G}_T} w_G \|x|_G\|_p, \quad (12.23)$$

where $p \geq 1$, w_G are positive weights and $x|_G$ is the restriction of x to the elements contained in G . Since the nodes lower down in the tree appear in more groups than their ancestors, they will contribute more to $\Omega(x)_{\text{HGL}}$ and therefore will be more encouraged to be zero. The proximity operator of Ω_{HGL} can be computed exactly for $p = 2$ and $p = \infty$ via an active set algorithm [69].

Other convex penalties have been recently proposed in order to favor hierarchical structures, while also allowing for a certain degree of flexibility in deviations from the discrete model. One approach considers groups consisting of all parent–child pairs and uses the latent group lasso penalty (see Section 12.3.2) in order to obtain solutions whose support is the union of few such pairs [103], see Figure 12.9 (left).

An interesting extension is given by the *family* model [13, 132], where the groups consist of a node and all its children, see Figure 12.9 (right). Again the latent groups lasso penalty is used. This model is better suited for wavelet decomposition of images because it better reflects the fact that a large coefficient value implies large coefficients values for all its children at a finer scale.

For both these cases, one can use the duplication strategy to transform the overlapping proximity problem into a block one, which can be efficiently solved in closed-form, see [64] for more details.

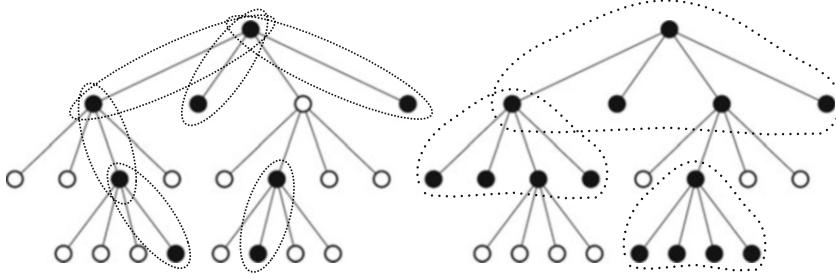


Fig. 12.9 Examples of parent–child and family models. Active groups are indicated by dotted ellipses. The support (black nodes) is given by the union of the active groups. (Left) Parent–child model. (Right) Family model.

12.6 Submodular models

Most of the structures described so far are naturally combinatorial, but the models presented in the previous sections are incapable of capturing more complex structures. A more general approach is to use a set function R to quantify how much a given support set deviates from the desired structure. For instance, we can describe all previous structures by using an indicator function that assigns an infinite value to sets that do not belong to the chosen structure \mathcal{M}_k . Next, we provide a definition of such set functions and the sets that they describe.

Definition 6. Given a set function $R : 2^N \rightarrow \mathbb{R}$, $R \not\equiv 0$, we can define a model \mathcal{R}_τ consisting of all sets S for which $R(S) \leq \tau$:

$$\mathcal{R}_\tau := \{S \mid R(S) \leq \tau\}. \quad (12.24)$$

Unfortunately, computing the projection of a given signal onto this set is generally feasible for any set function, due to its intractable combinatorial nature.

Therefore, it is necessary to restrict our attention to set functions with specific properties that lead to tractable problems. It turns out that in many applications (see, e.g., [2, 3, 6, 25, 56, 72, 89, 109]) the combinatorial penalties have a convenient “diminishing returns” property that we can exploit. These are the so-called *submodular* set functions.

Definition 7. A set function $R : 2^N \rightarrow \mathbb{R}$ is submodular iff $R(S \cup \{v\}) - R(S) \geq R(U \cup \{v\}) - R(U)$, $\forall S \subseteq U \subseteq N$, and $v \in N \setminus U$.

Given a combinatorial penalty that encodes the desired structure, we want to solve an inverse problem of the form

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to } \text{supp}(x) \in \mathcal{R}_\tau. \quad (12.25)$$

To simplify things, one can relax the hard constraint in (12.25) by using the combinatorial penalty as a regularizer:

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda \cdot R(\text{supp}(x)), \quad (12.26)$$

for some regularization parameter $\lambda > 0$.

Unfortunately, the composite minimization of a continuous and discrete function in (12.26) leads to an NP-hard problem, even for submodular set functions. In fact, even in the simple sparsity case in CS setting, where $f(x) := \frac{1}{2} \|y - Ax\|_2^2$, and $R(S) = |S|$, the problem (12.26) is NP-hard, in general.

When submodular functions are paired with continuous functions like in (12.26), the minimization becomes very difficult. However when considered alone, submodular functions can be efficiently minimized, see Section 12.7.1.2. Submodular function minimization (SFM) constitutes a key component in both the convex and discrete approaches to solving (12.26).

12.6.1 The discrete model

In the discrete setting, (12.26) is preserved in an attempt to faithfully encode the discrete model at hand. While such criterion seems cumbersome to solve, it has favorable properties that lead to polynomial-time solvability, irrespective of its combinatorial nature: there are efficient combinatorial algorithms that provide practical solutions to overcome this bottleneck and are guaranteed to converge in polynomial time [39]; see Section 12.7.1.

Within this context, our intention here is to present fundamental algorithmic steps that reveal the underlying structure of (12.26). Here, we assume that f has an L -Lipschitz continuous gradient and, thus, it admits the following upper bound:

$$f(x) \leq f(x^i) + \langle \nabla f(x^i), x - x^i \rangle + \frac{L}{2} \|x - x^i\|_2^2 := Q(x, x^i) \quad \forall x, x^i \in \text{dom}(f).$$

Then, we perform the following iterative Majorization–Minimization (MM) scheme:

$$x^{i+1} \in \operatorname{argmin}_{x \in \mathbb{R}^n} \{Q(x, x^i) + \lambda R(\text{supp}(x))\} = \operatorname{argmin}_{S \subseteq N} \left\{ \min_{x: \text{supp}(x)=S} Q(x, x^i) + \lambda R(S) \right\}. \quad (12.27)$$

If we focus on the inner minimization above and for a given support S , one can compute the optimal minimizer \hat{x}^i as $\hat{x}_{S^c}^i = 0$ and $\hat{x}_S^i = x_S^i - (1/L)\nabla f(x^i)_S$. By substituting \hat{x}_S^i into the upper bound $Q(\cdot, \cdot)$, we obtain a modular majorizer M^i :

$$f(x) + \lambda R(\text{supp}(x)) \leq C - \frac{L}{2} \sum_{j \in \text{supp}(x)} \left(x_j^i - \frac{1}{L} \nabla f(x^i)_j \right)^2 + \lambda R(\text{supp}(x)) \\ := M^i(\text{supp}(x)) + \lambda R(\text{supp}(x)),$$

where C is a constant. Finally, the update (12.27) becomes

$$x^{i+1} = \hat{x}_S^i \quad \text{subject to} \quad S \in \operatorname{argmin}_{S \subseteq N} M^i(S) + \lambda R(S). \quad (12.28)$$

Since the majorizer $M^i + \lambda R$ is submodular, and thus can be done efficiently. The minimization in (12.28) is a *submodular function minimization* problem,

Proposition 2 ([39]). *At each iteration, the new estimate x^{i+1} produced by the Majorization-Minimization algorithm satisfies*

$$f(x^{i+1}) + \lambda R(\text{supp}(x^{i+1})) \leq f(x^i) + \lambda R(\text{supp}(x^i)),$$

which implies convergence in the objective function value.

Note that this scheme is a type of proximal gradient method (cf., Section 12.7.2.1), applied to submodular regularizers, where the proximal operation in (12.41) corresponds to the SFM step in (12.28). We refer the reader to Section 12.7.2 for more information.

12.6.2 Convex approaches

The non-convex problem (12.26) can be converted to a closely related convex problem by replacing the discrete regularizer $g(x) = R(\text{supp}(x))$ by its largest convex lower bound. This is also called convex envelope and is given by the biconjugate g^{**} , i.e. by applying the Fenchel–Legendre conjugate twice [18]. We call this approach a convex relaxation of (12.26). However, for some functions g , there is no meaningful convex envelope. For these cases, one can use the convex closure $R^- : [0, 1]^n \rightarrow \mathbb{R}$, of $R(S)$, which is the point-wise largest lower bound of R [37]. Note that both the convex envelope and the convex closure are “tight” relaxations, but one in the continuous domain, the other in the discrete domain.

Both notions of convex relaxation for submodular functions use the Lovász extension (LE), introduced in [80]. We give here only one of the equivalent definitions of Lovász extension (for the other equivalent formulations, see, e.g., [3]):

Definition 8. Given a set function R such that $R(\emptyset) = 0$, we define its Lovász extension as follows: $\forall x \in \mathbb{R}^N$,

$$r(x) = \sum_{k=1}^N x_{j_k} \left(R(\{j_1, \dots, j_k\}) - R(\{j_1, \dots, j_{k-1}\}) \right), \quad (12.29)$$

where the coordinates are ordered in non-increasing order, $x_{j_1} \geq \dots \geq x_{j_N}$, breaking ties arbitrarily.

The Lovász extension is the convex closure of its corresponding submodular function on the unit hypercube, i.e $R^- = r$ [37]. In this sense, this extension already gives a convex relaxation for any submodular function. However, it turns out that using the Lovász extension as a convexification might not always fully capture the structure of the discrete penalty. We elaborate more on this in Section 12.6.3.

For a monotone⁴ submodular function its convex envelope is also obtained through the LE.

Proposition 3 ([2]). *Given a monotone submodular function R , with $R(\emptyset) = 0$, and with Lovász extension r , the convex envelope of $g(x) = R(\text{supp}(x))$ on the unit ℓ_∞ -ball is given by the norm $\Omega_{\text{sub}}(x) := g^{**}(x) = r(|x|)$.*

The proximity operator (12.2) of Ω_{sub} can be efficiently computed. In fact, computing $\text{prox}_\lambda^{\Omega_{\text{sub}}}(x)$ is shown in [2] to be equivalent to minimizing the submodular function $\lambda R(S) - \sum_{i \in S} |x_i|$ using the minimum-norm point algorithm. Other SFM algorithms can also be used, but then it is necessary to solve a sequence of SFMs. Note that simpler subcases of submodular functions may admit more efficient proximity operators than this general one. As a result, the convex relaxation of (12.26), where $R(\text{supp}(x))$ is replaced by $\Omega_{\text{sub}}(x)$, can be efficiently solved by a proximal gradient method (cf., Section 12.7.2.1).

The monotonicity requirement in Proposition (3) is necessary to guarantee the convexity of $r(|x|)$. Unfortunately, all symmetric⁵ submodular functions, among which undirected cut functions (cf., Section 12.6.3), are not monotone. In fact, the convex envelope of $R(\text{supp}(x))$ on the unit ℓ_∞ -ball, for any submodular function R , is given by [2]

$$g^{**}(x) = \min_{\delta \in [0,1]^n, \delta \geq |x|} r(\delta). \quad (12.30)$$

Thus, when R is monotone, the convex envelope is $r(|x|)$ as stated in Proposition 3. When R is symmetric and $R(N) = R(\emptyset) = 0$ (which is assumed without loss of generality, since addition of constants does not affect the regularization), $g^{**}(x) = 0, \forall x \in \mathbb{R}^n$ and the minimum in (12.30) is attained at any constant value vector δ .

Hence, the convexification of symmetric submodular functions consists of the Lovász extension alone. However the LE is a poor convexification that can significantly modify the intended structure. This can already be seen by the fact that the LE is tight only on the unit hypercube, while most penalties $R(\text{supp}(x))$ are not constrained there, and that $R(\text{supp}(x))$ is symmetric around the origin which is not the case for $r(x)$. Some artifacts of using the LE as a convexification for symmetric functions are illustrated in [39].

⁴A monotone function is a function that satisfies: $\forall S \subseteq T \subseteq N, R(S) \leq R(T)$.

⁵A symmetric function is a function that satisfies: $\forall S \subseteq N, R(S) = R(N \setminus S)$.

For some problems, the submodularity requirement can be relaxed. Convex relaxations of general set functions, when paired with the ℓ_p -norm ($p > 1$) as a continuous prior, are studied in [99].

Proposition 4 ([99]). *Define the norm*

$$\Omega_p(x) := \min_{v \in \mathcal{V}} \left\{ \sum_{S \subseteq N} R(S)^{\frac{1}{q}} \|v^S\|_p \mid \sum_{S \subseteq N} v^S = x \right\}, \quad (12.31)$$

where $\frac{1}{p} + \frac{1}{q} = 1$ and $\mathcal{V} = \{v = (v^S)_{S \subseteq N} \in (\mathbb{R}^n)^{2^N} \text{ s.t. } \text{supp}(v^S) \subseteq S\}$. Then $c\Omega_p$, where $c = (q\lambda)^{1/q}(p\mu)^{1/p}$, is the convex positively homogeneous envelope of the function $\lambda R(\text{supp}(x)) + \mu \|x\|_p^p$, and the convex envelope of $cR(\text{supp}(x))^{\frac{1}{q}} \|x\|_p$.

Note that Ω_p is the same norm as (12.8) (up to a constant factor), with weights d_j given by a set function R , such that $d_j = R(G_j)^{1/q}$, and the group structure \mathfrak{G} is the entire power set of N .

The convex relaxation proposed in Proposition 4, when applied to a submodular function, generalizes the relaxation of Proposition 3 to unit balls with respect to the ℓ_p -norm ($p > 1$) and not only to the ℓ_∞ one.

12.6.3 Examples

We now offer some examples of structures that can be described via submodular functions.

Simple sparsity The most ubiquitous structure in practice is simple sparsity, which corresponds to the set function $R(S) = |S|$ that measures the cardinality of S . The convexification obtained by Proposition 3 is the usual ℓ_1 -norm.

Group sparsity Given a group structure \mathfrak{G} , a group sparse model can be enforced by penalizing the number of groups that intersects with the support: $R_{\cap}(S) = \sum_{S \cap G_i \neq \emptyset} d_i$ where $d_i \geq 0$ is the weight associated with the group $G_i \in \mathfrak{G}$. For example, defining the groups to be nodes and all of its descendants on a tree (see Figure 12.10) enforces the support to form a rooted connected tree, which is characteristic of wavelet coefficients (cf., Section 12.5). The way the groups and their weights are defined leads to different sparsity patterns [3].

The convexification of $R_{\cap}(S)$ is the ℓ_1/ℓ_∞ -norm⁶ over groups $\sum_{G_i \in \mathfrak{G}} d_i \|x\|_\infty$ (even for overlapping groups) [2, 133]. Note that for groups that form a partition of N , $R_{\cap}(S)$ is equivalent to the minimum weight set cover, defined as:

$$R_{sc}(S) = \min_{s \in \mathbb{B}^M} \sum_{i=1}^M d_i s_i \quad \text{s.t.} \quad \sum_{i=1}^M s_i \mathbb{1}_{n, G_i} \geq \mathbb{1}_{n, S} \quad (12.32)$$

⁶Actually, it is a norm iff $N = \bigcup_{d_i > 0} G_i$.

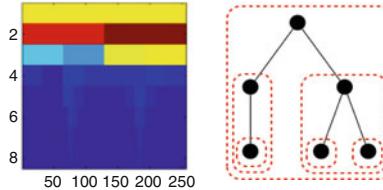


Fig. 12.10 Wavelet coefficients and underlying hierarchical group model.

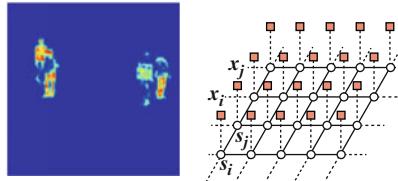


Fig. 12.11 Background subtraction and underlying Ising model.

with the same set of groups \mathfrak{G} and weights d . R_{sc} computes the total weight of the lightest cover for S using the groups in \mathfrak{G} . Note that $R_{sc}(S)$ is not a submodular function,⁷ unless the groups form a partition.

Ising model The Ising model [85] is a model that associates with each coefficient x_i of a signal $x \in \mathbb{R}^n$ a discrete variable $s_i \in \{-1, 1\}$ to represent the state (zero/nonzero) of the coefficient. The Ising penalty enforces clustering over the coefficients, which is a desired structure, for example, in background subtraction in images or videos [25]. It can be naturally encoded on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where the vertices are the coefficients of the signal, i.e. $\mathcal{V} = N$, and the edges connects neighboring coefficients. For example, for images, the coefficients of the signal are the pixels of the image, and edges connect pixels next to each other, forming a lattice (see Figure 12.11). This is called the two-dimensional lattice Ising model. The Ising penalty is then expressed via the following symmetric submodular function:

$$R_{\text{ISING}}(S) = \frac{1}{2} \left(|\mathcal{E}| - \sum_{(i,j) \in \mathcal{E}} s_i s_j \right), \quad (12.33)$$

⁷Consider the following example: Let $N = \{1, 2, 3, 4\}$, and $\mathfrak{G} = \{G_1 = \{1\}, G_2 = \{2, 3\}, G_3 = \{1, 2, 4\}\}$, with weights defined as $d_i = |G_i|$. Then the inequality in Definition 7 is not satisfied for the sets $S = \{1, 2\}$ for which $R_{sc}(S) = 3$, and $U = \{1, 2, 4\}$ for which $R_{sc}(U) = 4$, with the addition of the element $\{3\}$.

where $s \in \mathbb{R}^n$ is an indicator vector for the set S such that $s_i = 1$ if $i \in S$ and $s_i = -1$, otherwise. When $S = \text{supp}(x)$ for a signal $x \in \mathbb{R}^n$, s_i encodes the state of the corresponding coefficient x_i . We can also view the Ising penalty as a cut function: $R_{\text{ISING}}(S)$ counts the number of edges that are cut by the set S . Cut functions with appropriate graphs and weights capture a large subset of submodular functions [66, 72]. Since R_{ISING} is symmetric, its convexification is its Lovász extension, as discussed above, which is shown [39] to be the anisotropic discrete Total Variation semi-norm

$$\|x\|_{\text{TV}} = \sum_{(i,j) \in \mathcal{E}} |x_i - x_j|. \quad (12.34)$$

While the Ising model enforces the clustering of nonzero coefficients, its convexification encodes a different structure by favoring piecewise constant signals. Furthermore, $\|\cdot\|_{\text{TV}}$ penalty cannot be described by any of the structures introduced in the previous section.

Entropy Given n random variables X_1, \dots, X_n , the joint entropy of $X_S = (X_i)_{i \in S}$ defines a submodular function, $R(S) = H(X_S)$. Moreover, the mutual information also defines a symmetric submodular function, $R(S) = I(X_S; X_{N \setminus S})$. Such functions occur, for example, in experimental design [109], in learning Bayesian networks [56], in semi-supervised clustering [89], and in diverse feature selection [33]. Both of these set functions encourage structures that cannot be defined by the models presented in the previous sections.

Supermodular functions Another variant of set function regularizers used in practice, for example in diverse feature selection, are supermodular functions, i.e. negatives of submodular functions. The problem (12.26) with supermodular regularizers can then be cast as a maximization problem with a submodular regularizer. The notion of “approximate” submodularity introduced in [34, 73] is used to reduce the problem to an “approximate” submodular maximization problem, where several efficient algorithms for constrained/unconstrained submodular maximization [21, 22] with provable approximation guarantees can be employed. For more information, we refer the reader to [33].

12.7 Optimization

Apart from the rather mature theory describing the sparse signal approximation problem, the success of sparse optimization techniques lies also in the computational tractability of sparse approximation. In the CS regime, Donoho [36] and Candès et al. [23] utilize simple convex optimization—i.e., linear programming solvers or second-order cone programming methods—and matrix isometry assumptions over sparsely restricted sets, to compute a sparse approximation from a limited set of linear measurements in polynomial time. Alternatively, and in contrast to the

conventional convex relaxation approaches, iterative greedy algorithms [41, 74, 91] rely on the discrete structure of the sparse signal approximation problem to find a solution, and are characterized by identical approximation guarantees as in [23, 36]. Moreover, their overall complexity matches, if not improves, the computational cost of convex approaches.

In the case of structured (and thus more complicated) sparsity models, one needs efficient optimization solutions for structured sparsity problems that scale to high-dimensional settings. From our discussions above, it is apparent that the key actors for this purpose are projection and proximity operations over structured sets, which go beyond simple selection heuristics and towards provable solution quality as well as runtime/space bounds.

Projection operations faithfully follow the underlying combinatorial model, but in most cases, they result in hard-to-solve or even NP-hard optimization problems. Furthermore, model misspecification often results in wildly inaccurate solutions.

Proximity operators of convex sparsity-inducing norms often only partially describe the underlying discrete model and might lead to “rules-of-thumb” in problem solving (e.g., how to set up the regularization parameter). However, such approaches work quite well in practice, and are more robust to deviations from the model, leading to satisfactory solutions.

Here, our intention is to present an overview of the dominant approaches followed in practice. We consider the following three general optimization formulations⁸:

- *Discrete projection formulation:* Given a signal model M_k , let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a closed convex data fidelity/loss function. Here, we focus on the *projected* non-convex minimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad x \in M_k. \quad (12.35)$$

- *Convex proximity formulation:* Given a signal model M_k , let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a closed convex data fidelity/loss function, $g : \mathbb{R}^n \rightarrow \mathbb{R}$ a closed convex regularization term, possibly nonsmooth, that models M_k ⁹ and $\lambda > 0$. In this chapter, we focus on the convex composite minimization problem:

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda \cdot g(x). \quad (12.36)$$

- *Convex structured-norm minimization:* Given a signal model M_k , let $g : \mathbb{R}^n \rightarrow \mathbb{R}$ be a closed convex regularization term, possibly nonsmooth, that models M_k .

⁸We acknowledge that there are other criteria that can be considered in practice; for completeness, in the simple sparsity case, we refer the reader to the ℓ_1 -norm constrained linear regression (a.k.a. Lasso [115])—similarly, there are alternative optimization approaches for the discrete case [127]. However, our intention in this chapter is to use the most prevalent formulations used in practice.

⁹For example, ℓ_1 -norm models well the ℓ_0 -“norm.”

Moreover, let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a closed convex data fidelity/loss function and $\sigma > 0$. We consider the following minimization problem:

$$\min_{x \in \mathbb{R}^n} g(x) \quad \text{subject to} \quad f(x) \leq \sigma. \quad (12.37)$$

In addition, we also briefly mention SFM

$$\min_{S \in 2^N} R(S),$$

where $R : 2^N \rightarrow \mathbb{R}$ is a submodular function; see Definition 7. SFM is at the core of both discrete and convex approaches for solving estimation problems, where the structure is imposed via a submodular set function; see Section 12.6.

Some representative problem instances are given next.

Compressive sensing (CS) In classic CS, we are interested in recovering a high-dimensional signal $x^* \in M_k$ from an underdetermined set of linear observations $y \in \mathbb{R}^m$ ($m < n$), generated via a known $A \in \mathbb{R}^{m \times n}$:

$$y = Ax^* + \varepsilon.$$

Here, $\varepsilon \in \mathbb{R}^m$ denotes an additive noise term with $\|\varepsilon\|_2 \leq \sigma$.

For $f(x) := \frac{1}{2}\|y - Ax\|_2^2$ and using the fact that $x^* \in M_k$, one can formulate the problem at hand as in (12.35); observe that, in the case where $M_k \equiv \Sigma_k$, (12.35) corresponds to the ℓ_0 -“norm” optimization problem [90]. In applications where we have considerably more a priori information on the sparse coefficients, Model-based CS [8] extends CS to leverage this knowledge where $M_k \neq \Sigma_k$; see [4] for efficient implementations using expander matrices as measurement operators. Alternatively, given a function g that encourages solutions to be in M_k , one can solve the CS problem with convex machinery as in (12.36)–(12.37); when $M_k \equiv \Sigma_k$, (12.36)–(12.37) correspond to the Basis Pursuit DeNoising (BPDN) criterion [29].

Sparse graph modeling In Gaussian graphical model selection [32, 58, 59, 76, 121], the sparse-regularized maximum likelihood estimation yields the criterion:

$$\hat{x} = \operatorname{argmin}_{x \in \mathbb{R}^{n^2}} \left\{ -\log \det(\operatorname{mat}(x)) + \operatorname{trace}(\operatorname{mat}(x) \cdot \hat{C}) + \lambda \cdot g(x) \right\}, \quad (12.38)$$

where $x \in \mathbb{R}^{n^2}$ is the vectorized version of the inverse covariance estimate, $\operatorname{mat}(\cdot)$ is the linear matrix operation that converts vectors to matrices and \hat{C} is the sample covariance. Developments in random matrix theory [70] divulge the poor performance of solving (12.38) without regularization; in this case, the solution is usually fully dense and no inference about the dependence graph structure is possible. Choosing a suitable regularizer g [65, 78, 79], that models the connections between the variables, reduces the degrees of freedom, leading to robust and more interpretable results.

Other applications Similar sparsity regularization in discrete or convex settings is found in many diverse disciplines: a non-exhaustive list includes natural language processing [84], multipath radar signals [106], and time series of gene expressions [102].

12.7.1 Greedy and discrete approaches

12.7.1.1 Projected gradient descent and matching pursuit variants

Iterative greedy algorithms maintain the combinatorial nature of (12.35), but instead of tackling it directly, which would be intractable, the algorithms of this class greedily refine a k -model-sparse solution using only “local” information available at the current iteration. Within this context, matching pursuit approaches [83, 122] gradually construct the sparse estimate by greedily choosing the nonzero coefficients that best explain the current residual, e.g. $y - Bx^i$, where x^i is the current iterate. Extensions of such greedy approaches have recently been proposed to accommodate structured models in the selection process [61].

Most of the algorithmic solutions so far concentrate on the non-convex projected gradient descent algorithm, a popular method known for its simplicity and ease of implementation. Per iteration, the total computational complexity is determined by the calculation of the gradient and the projection operation on M_k as in (12.1). Most algorithms in the discrete model can be described or easily deduced by the following simple recursion:

$$x^{i+1} = \mathcal{P}_{M_k} \left(x^i - \frac{\mu}{2} \nabla f(x^i) \right), \quad (12.39)$$

where μ is a step size and $\mathcal{P}_{M_k}(\cdot)$ is the projection onto k -model-sparse signals.

Representative examples for the CS application are hard thresholding methods over simple sparse sets [14, 41, 74, 77, 91]. [8] further extends these ideas to *model-based CS*, where non-overlapping group structures and tree structures are used to perform the projection (12.1). From a theoretical computer science perspective, [4, 62] address the CS problem using *sparse* sensing matrices. Exploiting model-based sparsity in recovery provably reduces the number of measurements m , without sacrificing accuracy. The resulting algorithm reduces the main computational cost of the proposed scheme on the difficulty of projecting onto the model-sparse set, which, as mentioned in Section 12.3, in most relevant cases, can be computed in linear time using dynamic programming, in most relevant cases.¹⁰

¹⁰In the case of CS, an important modification of (12.39) to achieve linear computational time per iteration is the substitution of the gradient with the *median* operator, which is nonlinear and defined component-wise on a vector; for more information, we refer to [41, 47].

12.7.1.2 Submodular function minimization

Submodularity is considered the discrete equivalent of convexity in the sense that it allows efficient minimization. The best combinatorial algorithm for SFM has a proven complexity of $O(n^5E + n^6)$, where E is the function evaluation complexity [101]. For practical purposes however, another algorithm called minimum-norm point algorithm [43] is mostly used. This algorithm has no known worst case complexity but in practice it usually runs in $O(n^2)$. Furthermore, for certain functions [44, 66], submodular minimization becomes equivalent to computing the minimum s–t cut on the corresponding graph $G(\mathcal{V}, \mathcal{E})$, which has time complexity $\tilde{O}(|\mathcal{E}| \min\{|\mathcal{V}|^{2/3}, |\mathcal{E}|^{1/2}\})$ [49], where the notation $\tilde{O}(\cdot)$ ignores log terms.

12.7.2 Convex approaches

12.7.2.1 Proximity methods

Proximity gradient methods are iterative processes that rely on two key structural assumptions: i) f has Lipschitz continuous gradient¹¹ (see Definition 2) and ii) the regularizing term g is endowed with a *tractable* proximity operator.

By the Lipschitz gradient continuity and given a putative solution $x^i \in \text{dom}(f)$, one can locally approximate f around x^i using a quadratic function as

$$f(x) \leq Q(x, x^i) := f(x^i) + \langle \nabla f(x^i), x - x^i \rangle + \frac{L}{2} \|x - x^i\|_2^2, \quad \forall x \in \text{dom}(f).$$

The special structure of this upper bound allows us to consider a majorization–minimization approach: instead of solving (12.36) directly, we solve a sequence of simpler composite quadratic problems:

$$x^{i+1} \in \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \{Q(x, x^i) + g(x)\}. \quad (12.40)$$

In particular, we observe that (12.40) is equivalent to the following iterative *proximity* operation, similar to (12.2):

$$x^{i+1} \in \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| x - \left(x^i - \frac{1}{L} \nabla f(x^i) \right) \right\|_2^2 + \frac{1}{2L} g(x) \right\}. \quad (12.41)$$

Here, the anchor point x in (12.2) is the gradient descent step: $x := x^i - \frac{1}{L} \nabla f(x^i)$.

¹¹In [120], the authors consider a more general class of functions with no *global* Lipschitz constant L over their domain. The description of this material is out of the scope of this chapter and is left to the reader who is interested in deeper convex analysis and optimization.

In the case where $g(x) := \|x\|_1$, the proximity algorithm in (12.41) is known as the Iterative Soft-Thresholding Algorithm (ISTA) [26, 30, 35]. Iterative algorithms can use memory to provide momentum in convergence. Based on Nesterov's optimal gradient methods [10, 95] proves the universality of such acceleration in the composite convex minimization case of (12.36), where $g(x)$ can be any convex norm with tractable proximity operator. However, the resulting optimization criterion in (12.41) is more challenging when g encodes more elaborate sparsity structures.

Within this context, [108, 125] present a new convergence analysis for proximity (accelerated) gradient problems, under the assumption of *inexact proximity evaluations*, and study how these errors propagate into the convergence rate.

An emerging direction for solving composite minimization problems of the form (12.36) is based on the proximity-Newton method [45]. The origins of this method can be traced back to the work of [16, 45], which relies on the concept of *strong regularity* introduced by [107] for generalized equations. In this case, we identify that the basic optimization framework above can be easily adjusted to second-order Newton gradient and quasi-Newton approaches:

$$x^{i+1} \in \operatorname{argmin}_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \|x - (x^i - H_i^{-1} \nabla f(x^i))\|_{H_i}^2 + \frac{1}{2} g(x) \right\}, \quad (12.42)$$

where H_i represents either the actual Hessian of f at x^i (i.e., $\nabla^2 f(x_i)$) or a symmetric positive definite matrix approximating $\nabla^2 f(x^i)$. Given a computationally efficient Newton direction, one can re-use the model-based proximity solutions presented in the previous subsection along with a second order *variable metric* gradient descent scheme, as presented in (12.42) [120].

12.7.2.2 Primal–dual and alternating minimization approaches

Several model-based problems can be solved using the convex structured-norm minimization in (12.37). As a stylized example, consider the noiseless CS problem formulation

$$\min_{x \in \mathbb{R}^n} g(x) \quad \text{subject to} \quad y = Ax, \quad (12.43)$$

where g is a convex structured norm.

Within this context, primal–dual convex optimization methods provide attractive approaches, by exploiting both the primal and the dual formulations at the same time (via Lagrangian dualization). More precisely, primal–dual methods solve the following minimax problem:

$$\min_{x \in \mathbb{R}^n} \max_{q \in \mathbb{R}^m} \{g(x) + \langle Ax - y, q \rangle\}, \quad (12.44)$$

where q is the Lagrange multiplier associated with the equality constraint $y = Ax$. The minimax formulation (12.44) can be considered as a special case of the general formulation in [27, 96].

If the function g is nonsmooth and possesses a closed form proximity operator, then general primal–dual methods such as [27, 54] have been known as powerful tools to tackle problem (12.44). Since g is nonsmooth, one can also use the primal–dual subgradient method [98] as well as the prox-method proposed in [94] for solving (12.43). An alternative approach is to combine primal–dual optimization methods and smoothing techniques, as done in [96, 97, 119], for solving (12.44).

When (12.44) possesses a separable structure $g(x) := \sum_{i=1}^p g_i(x_i)$, distributed approaches can be applied. Recently, the authors in [118, 119] developed a unified primal–dual decomposition framework for solving (12.43), also using ideas from the alternating direction methods of multipliers [19, 30, 123]. Such alternating optimization methods offer a unifying computational perspective on a broad set of large-scale convex estimation problems in machine learning [67] and signal processing [30], including sparse recovery, deconvolution, and de-mixing [50]. They have numerous universality and scalability benefits and, in most cases, they are well suited to distributed convex optimization. In this case, one can borrow ideas and tools from gradient descent and Newton schemes in solving the subproblems of the alternating minimization. While their theoretical convergence guarantees are not optimal, they usually work well in practice.

12.8 Applications

12.8.1 Compressive Imaging

Natural images are usually sparse in wavelet basis. In this experiment, we study the image reconstruction problem from compressive measurements, where structured sparsity ideas are applied in practice.

For this purpose, given a $p \times p$ natural grayscale image $x \in \mathbb{R}^{p^2}$, we use the Discrete Wavelet Transform (DWT) with $\log_2(p)$ levels, based on the Daubechies 4 wavelet, to represent x ; see the Wavelet representation of two images in Figures 12.12 and 12.13. In mathematical terms, the DWT can be described by an operator matrix W , so that x can be represented as $x = Wc$, where $c \in \mathbb{R}^n$, $n := p^2$, are the wavelet coefficients for x . Furthermore, x can be well approximated by using only a limited number of wavelet coefficients $\hat{c} \approx c$ with $\|\hat{c}\|_0 \ll n$.

To exploit this fact in practice, we consider the problem of recovering $x \in \mathbb{R}^n$ from m compressive measurements $y \in \mathbb{R}^m$. The measurements are obtained by applying a sparse matrix $A \in \mathbb{R}^{m \times n}$, where $m = n/8$, to the vectorized image such that

$$y = Ax.$$

Here, A is the adjacency matrix of an expander graph [4] of degree $d = 8$, so that $\|A\|_0 = dn$. Thus, the overall measurement operator on the wavelet coefficients is given by the concatenation of the expander matrix with the DWT: $y = Aw$.

We use the following methods for recovering c from the measurements y :

$$\begin{aligned}
 & \min_{c \in \mathbb{R}^n} \|y - Aw\|_2^2 && \text{(Rooted Connected Tree model (RC))} \\
 & \text{subject to } \text{supp}(c) \in T_k. \\
 & \min_{c \in \mathbb{R}^n} \|c\|_1 && \text{(Basis Pursuit (BP))} \\
 & \text{subject to } y = Aw. \\
 & \min_{c \in \mathbb{R}^n} \|c\|_{\text{HGL}} && \text{(Hierarchical Group Lasso (HGL) pursuit)} \\
 & \text{subject to } y = Aw. \\
 & \min_{c \in \mathbb{R}^n} \|c\|_{\text{PC}} && \text{(Parent-Child Latent Group Lasso (PC) pursuit)} \\
 & \text{subject to } y = Aw. \\
 & \min_{c \in \mathbb{R}^n} \|c\|_{\text{FAM}} && \text{(Family Latent Group Lasso (FAM) pursuit)} \\
 & \text{subject to } y = Aw.
 \end{aligned}$$

The RC model is solved via the improved projected gradient descent given in [74], with the projections computed via the dynamic program proposed in [5]. All of the remaining methods are solved using the primal–dual method described in [117] which relies on the proximity operator of the associated structure-sparsity inducing penalties. For BP the proximity operator is given by the standard soft-thresholding function. For HGL, we use the algorithm and code given by [69]. For the latent group lasso approaches, PC and FAM, we adopt the duplication strategy proposed in [64, 100], for which the proximity operator reduces to the standard block-wise soft-thresholding on the duplicated variables. All algorithms are written in Matlab, except for the HGL proximity operator and the RC projection, which are written in C.

The duplication approach consists in creating a latent vector that contains copies of the original variables. The number of copies is determined by the number of groups that a given variable belongs to. For the parent–child model, each node belongs to the four groups that contain each of its children and the group that contains its father. The root has only three children, corresponding to the roots of the horizontal, vertical, and diagonal wavelet trees. Each leaf belongs only to the group that contains its father. A simple calculation shows that the latent vector for the PC model contains $2(n - 1)$ variables.

For the Family model, instead, each node belongs to only two groups: the group containing its children, and the group containing its siblings and its father. Each

leaf belongs to only the group that contains its siblings and its father. Overall, the number of variables in the latent vector for the Family models is equal to $\frac{5n}{4} - 1$.

The duplication approach does not significantly increase the problem size, while it allows an efficient implementation of the proximity operator. Indeed, given that the proximity operator can be computed in closed form over the duplicated variables, this approach is as fast as the hierarchical group lasso approach, where the proximity operator is computed via C code.

In order to obtain a good performance, both the parent–child and the family model require a proper weighting scheme to penalize groups lower down in the tree, where smaller wavelet coefficients are expected, compared to nodes closer to the root, which normally carry most of the energy of the signals and should be penalized less. We have observed that setting the group weights proportional to L^2 , where L is the level of the node of the group closest to the root, gives good results. In particular, we set the weights equal to L^2 , with 0 being the root level.

12.8.1.1 Results

We perform the compressive imaging experiments on both a 256×256 portrait of a woman and a 2048×2048 mountain landscape. Apart from conversion to grayscale and resizing through the Matlab function `imresize`, no preprocessing is carried out. We run the primal–dual algorithm of [117] up to precision 10^{-5} . We measure the recovery performance in terms of Peak Signal-to-Noise Ratio (PSNR) and relative recovery error in ℓ_2 norm as $\frac{\|\hat{x} - x\|_2}{\|x\|_\infty}$, where \hat{x} is the estimated image and x is the true image.

Figures 12.12 and 12.13 report the recovery results using $m = \frac{n}{8}$, that is, using only 12.5% samples compared to the ambient dimension. The estimated images are on the top two rows, while the third and fourth rows show the estimated wavelet coefficients.

The effect of imposing structured sparsity can be clearly seen for the HGL, PC, and FAM models, where the high values of the coefficients tend to cluster around the root of the wavelet tree (i.e., top-left corner of the image) and their intensity decreases descending the tree. The family model shows the grouping among the siblings, where four leaves are either all zero or all nonzero. For the 256×256 image, despite being coded in C, the discrete model is approximately 160 times slower than the other methods, which are computationally equivalent: e.g., in our tests, the family model took around 60 s, while the RC one required almost 2 h. We therefore did not use the RC model on the larger 2048×2048 mountain image, but we compared also against Total Variation (TV) pursuit, which obtains the best performance on this image.

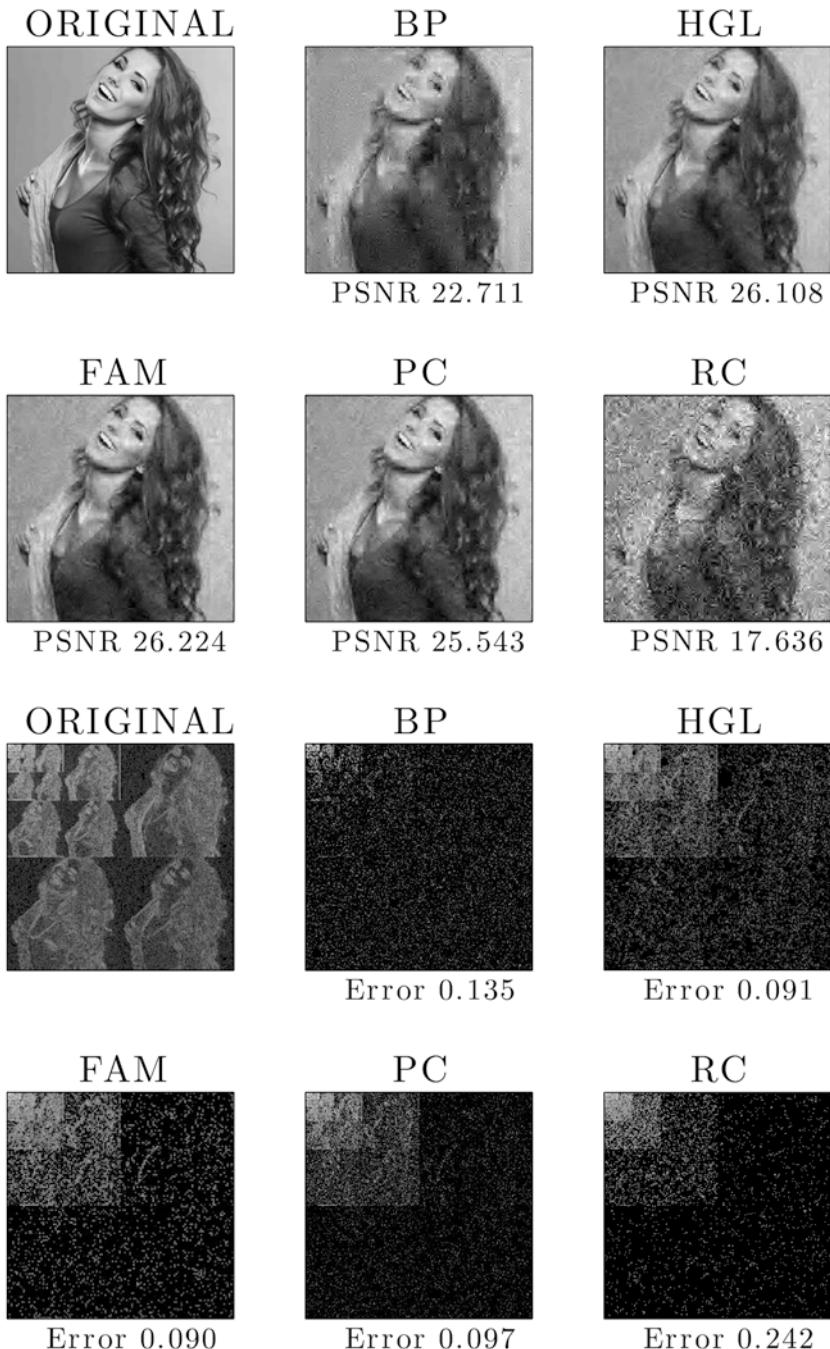


Fig. 12.12 Woman image recovery performance from compressive measurements. Here, $p = 256$. The top two rows show the reconstruction performance in the original domain, along with the PSNR levels achieved. The bottom two rows show the corresponding representations into the wavelet domain.

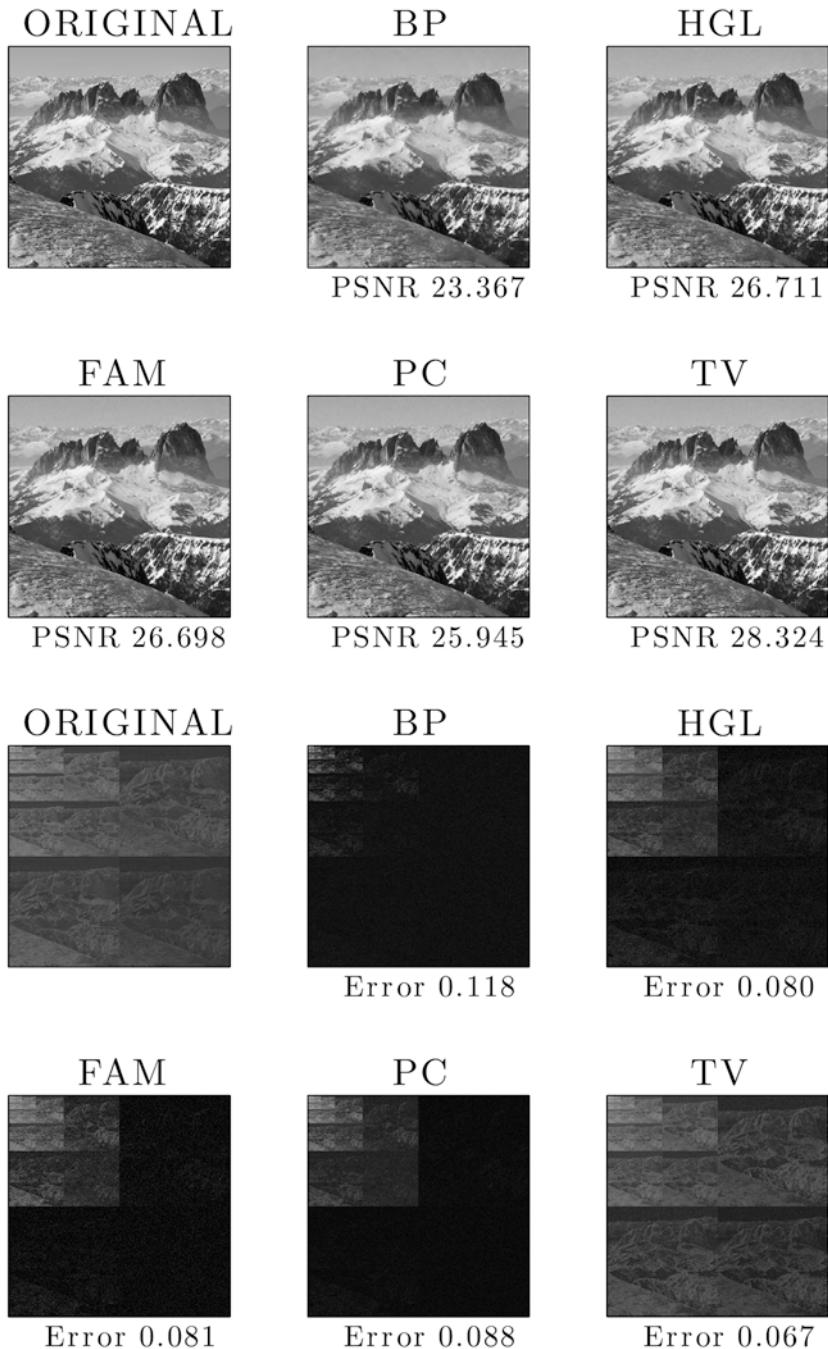


Fig. 12.13 Mountains image recovery performance from compressive measurements. Here, $p = 2048$. The top two rows show the reconstruction performance in the original domain, along with the PSNR levels achieved. The bottom two rows show the corresponding representations into the wavelet domain.

12.8.2 Neuronal spike detection from compressed data

In the experiments that follow, we compare the performance of the following three optimization criteria, assuming the dispersive model D_k .

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad x \in D_k. \quad (\text{Discrete dispersive})$$

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda \cdot \Omega_{\text{exclusive}}(x). \quad (\text{Exclusive norm regularization})$$

$$\min_{x \in \mathbb{R}^n} \Omega_{\text{exclusive}}(x) \quad \text{subject to} \quad f(x) \leq \sigma^2. \quad (\text{Exclusive norm pursuit})$$

Here, we use $p = 2$ to define the exclusive norm, according to (12.20).

Empirical performance on synthetic data Figures 12.14 and 12.15 illustrate the utility of each approach in the compressed sensing setting where $f(x) := \frac{1}{2} \|y - Ax\|_2^2$. That is, we observe $x^* \in \mathbb{R}^n$ through a limited set of linear sketches $y = Ax^* + \varepsilon \in \mathbb{R}^m$ where $A \in \mathbb{R}^{m \times n}$ is a known Gaussian matrix, where each entry is drawn i.i.d. from $\mathcal{N}(0, 1/m)$. Here, we assume $n = 500$ and $m = 70$ for $\|x^*\|_0 = 25$. Without loss of generality, we assume $(x^*)_i \geq 0, \forall i$ and $\Delta^* = 20$.

In the discrete case, we relax the refractory period Δ to model signal structure deviations; here, we assume $\Delta = 15$. The discrete dispersive model [8, 91] clearly outperforms the rest of the approaches under comparison; such behavior is also observed on average over the set of experiments conducted (Figure 12.14). This also

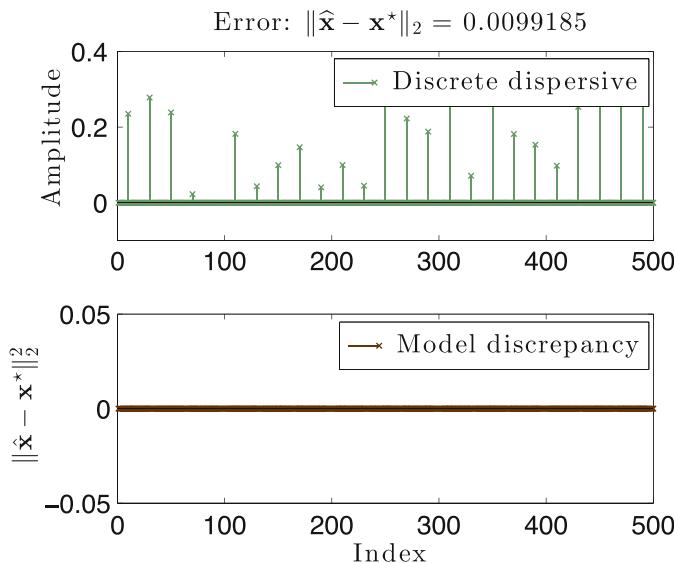


Fig. 12.14 Performance of the discrete dispersive approach for the spike train recovery problem from a limited set of linear measurements.

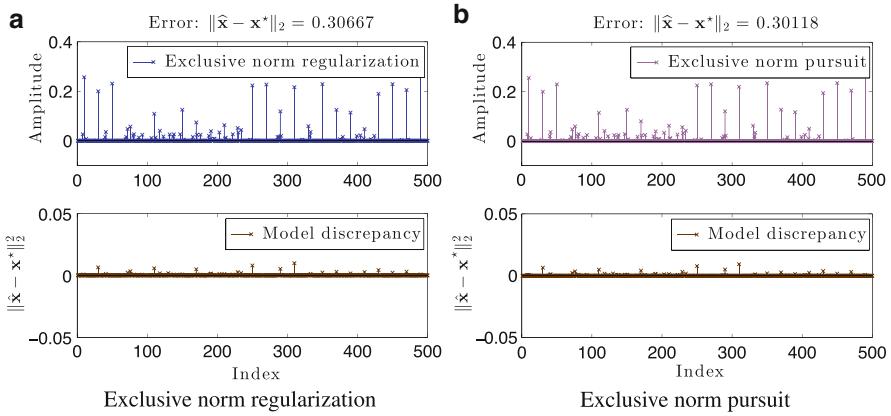


Fig. 12.15 Performance of dispersive convex relaxations for the problem of spike train recovery from a limited set of linear measurements. **(Left panel:)** Exclusive norm regularization approach. **(Right panel:)** Exclusive norm pursuit approach.

implies that the discrete model usually requires fewer measurements for accurate recovery compared to conventional sparse approximation, as long as the underlying signal approximately follows D_k .

On the other hand, due to convex relaxations, convex approaches introduce unnecessary nonzero coefficients that do not comply with the underlying model. However, both approaches show good performance in recovering x^* from limited measurements; see Figure 12.15.

Real neuronal spike data In order to understand the functioning of the human brain, it is necessary to identify and study the behavior of neuronal cell membranes under rapid change in the electric potential. However, to observe such phenomena, electrical activities on neurons need to be recorded using specialized equipment. In this experiment, we perform somatic spike detection of a tufted L5 pyramidal cell responding to in-vivo-like current injected in the apical dendrites and the soma simultaneously (see [63] for the experimental details).

A snapshot of the neuronal spikes waveforms is shown in Figure 12.16. In order to accurately detect the neuronal spikes, a high-frequency sample acquisition equipment is required. Within this context, we apply CS ideas to decrease the number of samples needed to approximately detect the *positions* of the spike train. Let $x^* \in \mathbb{R}^n$ with $n = 832$ represent the signal in Figure 12.16a. Furthermore, let $A \in \mathbb{R}^{m \times n}$ be a Gaussian sensing matrix where each entry is drawn i.i.d. from $\mathcal{N}(0, 1/m)$ and $m = n/4$, i.e., we perform a 75% compression.

We use the proposed schemes to recover the locations of the neuronal spikes under the assumption of the dispersive model with refractory period Δ . Here, Δ is set equal to the *average* period between two consecutive spikes.

Figure 12.16b represents the recovered k -sparse approximation using the discrete dispersive model D_k . Here, k is set to the number of spikes expected to appear for a

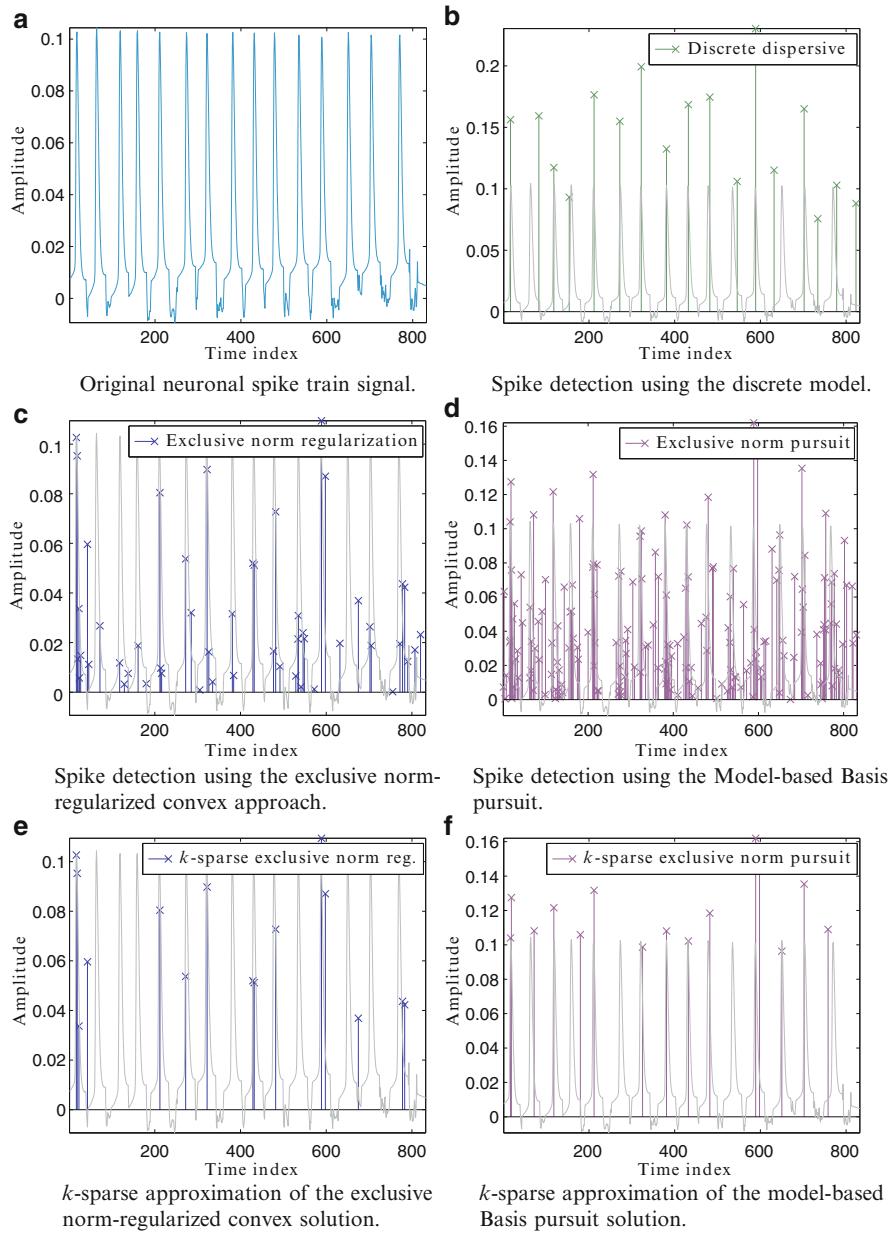


Fig. 12.16 Spike detection in real neuronal data using the dispersive model. Figure 12.16a depicts the original signal x^* . We observe $y = Ax^*$ using only 25% measurements through a linear sketch A. Figures 12.16b-12.16d illustrate the performance of the three approaches under comparison. Figures 12.16e-12.16f show the convex solutions, sparsified to be k -sparse.

given time period—such number can be easily deduced by observing the behavior of a specific neuron type. From Figure 12.16b, we observe that the discrete model approximates the locations of the spikes quite accurately: most of the spike locations are exactly recovered. However, due to the “strictness” of the discrete model, we observe that small deviations from D_k lead to imprecise estimations; e.g., between the 12th and 13th spike of the sequence, a larger (than usual) refractory period is observed that leads to mis-location of the next spike estimation.

Figure 12.16c, d depicts the performance of convex solvers using the exclusive norm as (i) a regularizer and (ii) an objective function. Tweaking the parameter λ in the first case, one can achieve *sparse* solutions that approximate the underlying model (Figure 12.16c). However, one can observe multiple detected spikes with separation less than Δ , violating the assumed model. In the model-based Basis pursuit case, the solver tries to *fit* the solution to the data, which usually leads to less sparse solutions (Figure 12.16d). One can further sparsify the convex solutions to obtain a k -sparse answer as in Figure 12.16e, f; however, in most cases, further processing of the returned signal is required to maintain a D_k -modeled solution. For example, in this case, due to the fact that convex norms force the solution to fully *explain* the observations, the sparsified solution includes more than one spike per true spike location.

12.8.3 Digital Confocal Imaging

Confocal imaging [87, 111] has become one of the best techniques for 3D imaging, due to its ability to reduce the blur caused by out-of-focus scattering by means of focused light and a physical pinhole. The recent work of Goy et al. [51] combines this technique with digital holography, which allows access to the complex field at the recording CCD sensor, and enables a novel and flexible way of eliminating the out-of-focus contribution via a so-called virtual pinhole in the digital domain. Under the assumption that the scattered light is not distorted by the matter in the light cones around the focus, this makes it possible to obtain very sharp images of the sample at all depths.

We are interested in modeling the scattering process so that we can instead leverage the information carried by the out-of-focus scattering for reducing the number of focusing locations that are needed for a faithful imaging of the sample. We create a linear model by adopting the classical first Born approximation for the scattering [17]. Our model takes into account the specific form of the Gaussian beam used to focus the light inside the sample, the numerical aperture of the transmission lens that determines which fraction of the scattered light is captured and the resolution and size of the CCD that records the hologram. For simplicity, we model the scattered field as it would arrive on the CCD and not its interaction with the reference field normally used to obtain the hologram. We use a 3D discrete model for representing the imaged sample, where one 3D pixel, or voxel, contains the average value of the scattering potential in the volume contained therein.

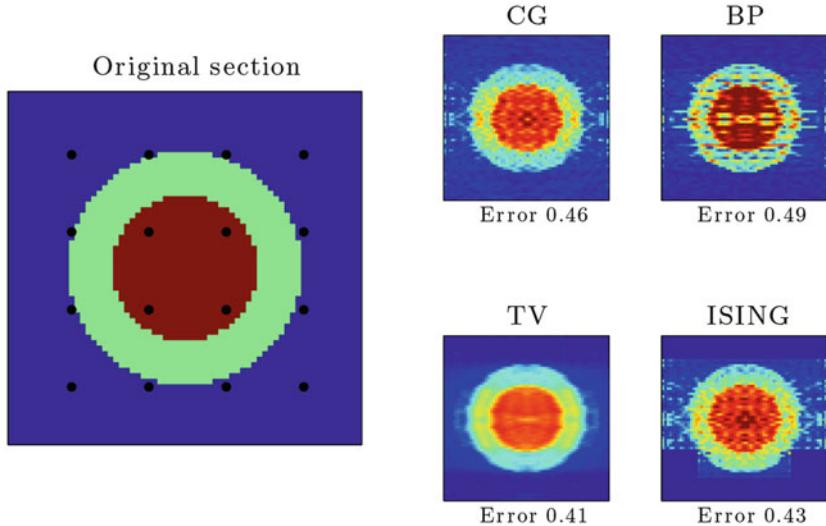


Fig. 12.17 (Left) Central section of the sample to be reconstructed. The black dots represent the 16 focus locations. (Right) Reconstruction results for the four methods.

For a specific focus location i , the measurement model A_i maps a 3D object represented by its scattering potential F to a 2D complex field. The complete measurement model A is then given by the concatenation of A_i for all focus points

$$A(F) = \begin{bmatrix} A_1(F) \\ A_2(F) \\ \vdots \\ A_P(F) \end{bmatrix},$$

where P is the total number of focus points.

For the experiments in this section, we consider a $64 \times 8 \times 64$ volume filled with oil with refractive index 1.52, containing two concentric cylinders with axis aligned with the y axis. Their refractive indices are 1.54 for the outer cylinder and 1.56 for the inner one. A x - z cross section of their scattering potential is shown in Figure 12.17 (Left). Note that F is equal to zero outside the cylinder and has constant value inside each. These properties could be captured via sparsity and clusteredness, for example via ℓ_1 norm or Total Variation minimization or via the Ising discrete model; see Section 12.6. We choose 16 focus locations taken from a uniform grid over the central x - z plane. The focusing lens and the transmission lens have numerical aperture 1 and 0.65, respectively, while the coherent light of the Gaussian beam has wavelength $\lambda = 405 \cdot 10^{-9}$ m. We consider a 91×91 CCD that covers the entire field of view of the transmission lens. We simulate the measurements by applying the measurement model to the synthetic scattering

potential F , that is, $y = A(F)$, where we have $F \in \mathbb{R}^{64 \times 8 \times 64}$ and $y \in \mathbb{C}^{16 \times 91 \times 91}$. Despite the apparent oversampling ratio, the problem is still ill-posed, because each recorded field carries information only about a small neighborhood around each focus point.

We compare four methods for estimating F from the measurements y . Let the data-fit term be the square loss $L(F) = \frac{1}{2} \|y - A(F)\|_2^2$.

$$\min_F L(F) . \quad (\text{Conjugate Gradient (CG)})$$

$$\min_F \|F\|_1 \quad \text{subject to} \quad y = A(F) . \quad (\text{Basis Pursuit (BP)})$$

$$\min_F \|F\|_{TV} \quad \text{subject to} \quad y = A(F) . \quad (\text{Total Variation (TV) pursuit})$$

$$\min_F L(F) + \lambda R_{\text{ISING}}(\text{supp}(F)) + \tau |\text{supp}(F)| , \quad (\text{Ising plus Cardinality (IC)})$$

In the last approach, one has two regularization parameters $\lambda, \tau \geq 0$ to be set.

12.8.3.1 Results

We evaluate the recovery performance of the methods using the relative recovery error $\|\hat{F} - F^*\|_2 / \|F^*\|_2$, where F^* is the synthetic scattering potential used to simulate the output y . We explored several pairs of regularization parameters λ and τ for the discrete IC model, selecting the pairs that gave the best performance. Of course, in practice one cannot resort to this type of parameter exploration and needs to choose them according to knowledge regarding the noise level and the desired amount of clusteredness. Furthermore, this method is very sensitive to changes in the regularization parameters, making it difficult to find the best value for them.

Figure 12.17 (right) reports the results on this experimental setup. Total Variation pursuit achieves the best performance without having to tune any parameter, closely followed by the discrete model. Basis Pursuit enforces too much sparsity, especially in the regions of the sample that lie farthest away from the focus points and hence do not contribute to the output y . It is interesting to note that in this case, enforcing the incorrect structure, such as sparsity by itself, leads to poorer performance than unstructured recovery, i.e., minimizing the square loss via conjugate gradient.

12.9 Conclusions

Recent advances in compressed sensing and machine learning go beyond the simple sparsity models towards *structured* sparsity models. Such models describe the interdependency between the nonzero components of a signal, enabling the interpretability of the results and improving the recovery performance from com-

pressive samples as well as improving noise robustness. In order to better understand the full impact of structured sparsity, this chapter analyzes the discrete models of structured sparsity and their convex relaxations, comparing their relative computational and performance trade-offs in the context of several applications.

While it may be quite natural to describe a sparsity model in terms of non-convex, discrete structures, the ensuing optimization problems can often pose difficult computational challenges. Even in cases when non-convex recovery is efficient, the results typically depend on a very precise specification of the model; see, for instance, the experiments on the neuronal spike detection in Section 12.8.2.

In contrast, although convex relaxations only “favor” the desired structure and obtain approximately interpretable solutions, they typically yield more stable results. Most of the convex problems emerging from structured sparsity models can be efficiently tackled via first-order algorithms that scale gracefully with the problem size. As observed in the compressive imaging experiment, Section 12.8.1, convex methods can be up to two orders of magnitude faster than their discrete counterparts.

Furthermore, discrete models do not usually capture the dependencies between the coefficients’ values. In certain situations, it can be suitable to constrain only the support: for instance, when we desire to cluster the nonzero coefficients in contiguous regions irrespective of their value, the Ising model is appropriate [39]. However, when there is strong prior information about the coefficients’ values, convex methods can perform better. In the compressive imaging results of Section 12.8.1, HGL outperforms the RC model despite both approaches enforce a rooted connected structure, because HGL also favors a decay of the coefficients from root to leaves.

Properly leveraging structural prior information requires a careful, precise, and complete definition of a discrete model or of a convex function that promotes the correct structure. We have in fact observed a peculiar phenomenon, where favoring only part of the signal’s structure turns out to be detrimental. In the digital confocal imaging simulations of Section 12.8.3, even though the signal to be recovered is sparse, Basis Pursuit yields worse results than unstructured recovery. This may be due to the fact that the original signal is characterized by two structures: sparsity and piecewise constancy. Basis Pursuit favors the first, but ignores the second, while TV does the opposite. The numerical results indicate that when the two structures are considered separately, piecewise constancy better describes the signal and yields a superior reconstruction compared to sparsity.

We believe that further research into the interplay between discrete and convex approaches and the related optimization problems promises to reap substantial rewards in terms of, for example, reducing sample complexity for certain classes of signals, and improving the recovery efficiency.

References

- Argyriou, A., Micchelli, C., Pontil, M., Shen, L., Xu, Y.: Efficient first order methods for linear composite regularizers (2000). arXiv preprint arXiv:1104.1436
- Bach, F.: Structured sparsity-inducing norms through submodular functions. In: Proceedings of Neural Information Processing Systems (NIPS) Foundation, pp. 118–126 (2010)
- Bach, F.: Learning with submodular functions: a convex optimization perspective (2011). arXiv preprint arXiv:1111.6453
- Bah, B., Baldassarre, L., Cevher, V.: Model-based sketching and recovery with expanders. In: Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA) (2014)
- Baldassarre, L., Bhan, N., Cevher, V., Kyriolidis, A.: Group-sparse model selection: Hardness and relaxations (2013). arXiv preprint arXiv:1303.3207
- Baraniuk, R.: Optimal tree approximation with wavelets. In: Proceedings of SPIE’s International Symposium on Optical Science, Engineering, and Instrumentation, pp. 196–207. International Society for Optics and Photonics (1999)
- Baraniuk, R., DeVore, R., Kyriazis, G., Yu, X.: Near best tree approximation. *Adv. Comput. Math.* **16**(4), 357–373 (2002)
- Baraniuk, R., Cevher, V., Duarte, M., Hegde, C.: Model-based compressive sensing. *IEEE Trans. Inf. Theory* **56**(4), 1982–2001 (2010)
- Baraniuk, R., Cevher, V., Wakin, M.: Low-dimensional models for dimensionality reduction and signal recovery: a geometric perspective. *Proc. IEEE* **98**(6), 959–971 (2010)
- Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2**(1), 183–202 (2009)
- Bengio, Y.: Learning deep architectures for AI. *Found. Trends Mach. Learn.* **2**(1), 1–127 (2009)
- Bertsekas, D.: Projected Newton methods for optimization problems with simple constraints. *SIAM J. Control Optim.* **20**(2), 221–246 (1982)
- Bhan, N., Baldassarre, L., Cevher, V.: Tractability of interpretability via selection of group-sparse models. In: Proceedings of IEEE International Symposium on Information Theory (ISIT) (2013)
- Blumensath, T., Davies, M.: Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.* **27**(3), 265–274 (2009)
- Blumensath, T., Davies, M.: Sampling theorems for signals from the union of finite-dimensional linear subspaces. *IEEE Trans. Inf. Theory* **55**(4), 1872–1882 (2009)
- Bonnans, J.: Local analysis of Newton-type methods for variational inequalities and nonlinear programming. *Appl. Math. Optim.* **29**, 161–186 (1994)
- Born, M., Wolf, E.: Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light. 7th edn. Cambridge University Press, Cambridge, UK (1999)
- Borwein, J., Lewis, A.: Convex Analysis and Nonlinear Optimization: Theory and Examples. Springer-Verlag, New York, US (2006)
- Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3**(1), 1–122 (2011)
- Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge, UK (2004)
- Buchbinder, N., Feldman, M., Naor, J., Schwartz, R.: A tight linear time $1/2$ -approximation for unconstrained submodular maximization. In: IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS), pp. 649–658 (2012)
- Buchbinder, N., Feldman, M., Naor, J., Schwartz, R.: Submodular maximization with cardinality constraints. In: Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA) (2014)
- Candes, E.: Compressive sampling. In: Proceedings of the International Congress of Mathematicians: Madrid, August 22–30, 2006: Invited Lectures, pp. 1433–1452 (2006)

24. Cartis, C., Thompson, A.: An exact tree projection algorithm for wavelets (2013). arXiv preprint arXiv:1304.4570
25. Cevher, V., Hegde, C., Duarte, M., Baraniuk, R.: Sparse signal recovery using Markov random fields. In: Proceedings of Neural Information Processing Systems (NIPS) Foundation (2009)
26. Chambolle, A., De Vore, R., Lee, N., Lucier, B.: Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Trans. Image Process.* **7**(3), 319–335 (1998)
27. Chambolle, A., Pock, T.: A first-order primal–dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.* **40**(1), 120–145 (2011)
28. Chandrasekaran, V., Recht, B., Parrilo, P., Willsky, A.: The convex geometry of linear inverse problems. *Found. Comput. Math.* **12**, 805–849 (2012)
29. Chen, S., Donoho, D., Saunders, M.: Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20**(1), 33–61 (1998)
30. Combettes, P., Wajs, V.: Signal recovery by proximal forward–backward splitting. *Multiscale Model. Simulat.* **4**(4), 1168–1200 (2005)
31. Crouse, M., Nowak, R., Baraniuk, R.: Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Process.* **46**(4), 886–902 (1998)
32. Dahl, J., Vandenberghe, L., Roychowdhury, V.: Covariance selection for nonchordal graphs via chordal embedding. *Optim. Methods Softw.* **23**(4), 501–520 (2008)
33. Das, A., Dasgupta, A., Kumar, R.: Selecting diverse features via spectral regularization. In: Proceedings of Neural Information Processing Systems (NIPS) Foundation, pp. 1592–1600 (2012)
34. Das, A., Kempe, D.: Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection (2011). arXiv preprint arXiv:1102.3975
35. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pure Appl. Math.* **57**(11), 1413–1457 (2004)
36. Donoho, D.: Compressed sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006)
37. Dugum, S.: Submodular functions: extensions, distributions, and algorithms: a survey (2009). arXiv preprint arXiv:0912.0322
38. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. *Ann. Stat.* **32**(2), 407–499 (2004)
39. El Halabi, M., Baldassarre, L., Cevher, V.: To convexify or not? Regression with clustering penalties on graphs. In: IEEE 5th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pp. 21–24 (2013)
40. Eldar, Y., Mishali, M.: Robust recovery of signals from a structured union of subspaces. *IEEE Trans. Inf. Theory* **55**(11), 5302–5316 (2009)
41. Foucart, S.: Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM J. Numer. Anal.* **49**(6), 2543–2563 (2011)
42. Friedman, J., Hastie, T., Tibshirani, R.: A note on the group lasso and a sparse group lasso (2010). arXiv preprint arXiv:1001.0736
43. Fujishige, S., Isotani, S.: A submodular function minimization algorithm based on the minimum-norm base. *Pac. J. Optim.* **7**(1), 3–17 (2011)
44. Fujishige, S., Patkar, S.: Realization of set functions as cut functions of graphs and hypergraphs. *Discret. Math.* **226**(1), 199–210 (2001)
45. Fukushima, M., Mine, H.: A generalized proximal point algorithm for certain non-convex minimization problems. *Int. J. Syst. Sci.* **12**(8), 989–1000 (1981)
46. Gerstner, W., Kistler, W.: Spiking Neuron Models: Single Neurons, Populations, Plasticity. Cambridge University Press, Cambridge, UK (2002)
47. Gilbert, A., Indyk, P.: Sparse recovery using sparse matrices. *Proc. IEEE* **98**(6), 937–947 (2010)
48. Girosi, F.: An equivalence between sparse approximation and support vector machines. *Neural Comput.* **10**(6), 1455–1480 (1998)
49. Goldberg, A., Rao, S.: Beyond the flow decomposition barrier. *J. ACM* **45**(5), 783–797 (1998)

50. Goldstein, T., Donoghue, B., Setzer, S.: Fast Alternating Direction Optimization Methods. CAM Report, pp. 12–35 (2012)
51. Goy, A., Psaltis, D.: Digital confocal microscope. *Opt. Exp.* **20**(20), 22720 (2012)
52. Gramfort, A., Kowalski, M.: Improving M/EEG source localization with an inter-condition sparse prior. In: Proceedings of IEEE International Symposium on Biomedical Imaging (2009)
53. Guigue, V., Rakotomamonjy, A., Canu, S.: Kernel basis pursuit. In: Machine Learning, pp. 146–157. Springer-Verlag, Berlin, Heidelberg (2005)
54. He, B., Yuan, X.: On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method. *SIAM J. Numer. Anal.* **50**, 700–709 (2012)
55. He, L., Carin, L.: Exploiting structure in wavelet-based Bayesian compressive sensing. *IEEE Trans. Signal Process.* **57**(9), 3488–3497 (2009)
56. Heckerman, D., Geiger, D., Chickering, D.: Learning Bayesian networks: the combination of knowledge and statistical data. *Mach. Learn.* **20**(3), 197–243 (1995)
57. Hegde, C., Duarte, M., Cevher, V.: Compressive sensing recovery of spike trains using a structured sparsity model. In: Signal Processing with Adaptive Sparse Structured Representations (SPARS) (2009)
58. Hsieh, C., Sustik, M., Dhillon, I., Ravikumar, P.: Sparse inverse covariance matrix estimation using quadratic approximation. In: Proceedings of Neural Information Processing Systems (NIPS) Foundation, pp. 2330–2338 (2011)
59. Hsieh, C., Sustik, M., Dhillon, I., Ravikumar, P., Poldrack, R.: BIG & QUIC: Sparse inverse covariance estimation for a million variables. In: Proceedings of Neural Information Processing Systems (NIPS) Foundation, pp. 3165–3173 (2013)
60. Huang, J., Zhang, T.: The benefit of group sparsity. *Ann. Stat.* **38**(4), 1978–2004 (2010)
61. Huang, J., Zhang, T., Metaxas, D.: Learning with structured sparsity. *J. Mach. Learn. Res.* **12**, 3371–3412 (2011)
62. Indyk, P., Razenshteyn, I.: On model-based RIP-1 matrices. In: Automata, Languages, and Programming, pp. 564–575. Springer-Verlag, Berlin, Heidelberg (2013)
63. International Neuroinformatics Coordinating Faculty.: Spike time prediction – challenge C (2009)
64. Jacob, L., Obozinski, G., Vert, J.P.: Group lasso with overlap and graph lasso. In: Proceedings of The 30th International Conference on Machine Learning (ICML) (2009)
65. Jalali, A., Ravikumar, P., Vasuki, V., Sanghavi, S.: On learning discrete graphical models using group-sparse regularization. In: Proceedings of International Conference on Artificial Intelligence and Statistics, pp. 378–387 (2011)
66. Jegelka, S., Lin, H., Bilmes, J.: On fast approximate submodular minimization. In: Proceedings of Neural Information Processing Systems (NIPS) Foundation, pp. 460–468 (2011)
67. Jenatton, R., Audibert, J.-Y., Bach, F.: Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.* **12**, 2777–2824 (2011)
68. Jenatton, R., Gramfort, A., Michel, V., Obozinski, G., Bach, F., Thirion, B.: Multi-scale mining of fMRI data with hierarchical structured sparsity. In: Pattern Recognition in NeuroImaging (PRNI) (2011)
69. Jenatton, R., Mairal, J., Obozinski, G., Bach, F.: Proximal methods for hierarchical sparse coding. *J. Mach. Learn. Res.* **12**, 2297–2334 (2011)
70. Johnstone, I.: On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.* **29**(2), 295–327 (2001)
71. Kim, S., Xing, E.: Tree-guided group lasso for multi-task regression with structured sparsity. In: Proceedings of The 30th International Conference on Machine Learning (ICML), pp. 543–550 (2010)
72. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(2), 147–159 (2004)
73. Krause, A., Cevher, V.: Submodular dictionary selection for sparse representation. In: Proceedings of The 30th International Conference on Machine Learning (ICML), pp. 567–574 (2010)

74. Kyrillidis, A., Cevher, V.: Recipes on hard thresholding methods. In: Proceedings of 4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP) (2011)
75. Kyrillidis, A., Cevher, V.: Combinatorial selection and least absolute shrinkage via the clash algorithm. In: Proceedings of International Symposium on Information Theory Proceedings (ISIT), pp. 2216–2220 (2012)
76. Kyrillidis, A., Cevher, V.: Fast proximal algorithms for self-concordant function minimization with application to sparse graph selection. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6585–6589 (2013)
77. Kyrillidis, A., Puy, G., Cevher, V.: Hard thresholding with norm constraints. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3645–3648 (2012)
78. Lee, J., Hastie, T.: Structure learning of mixed graphical models. In: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, pp. 388–396 (2013)
79. Loh, P., Wainwright, M.: Structure estimation for discrete graphical models: generalized covariance matrices and their inverses. *Ann. Stat.* **41**(6), 3022–3049 (2013)
80. Lovász, L.: Submodular functions and convexity. In: Mathematical Programming The State of the Art, pp. 235–257. Springer-Verlag, Berlin, Heidelberg (1983)
81. Lustig, M., Donoho, D., Pauly, J.: Sparse MRI: the application of compressed sensing for rapid MR imaging. *Magn. Reson. Med.* **58**(6), 1182–1195 (2007)
82. Mallat, S.: A Wavelet Tour of Signal Processing. Academic Press, Burlington, MA, US (1999)
83. Mallat, S., Zhang, Z.: Matching pursuits with time–frequency dictionaries. *IEEE Trans. Signal Process.* **41**(12), 3397–3415 (1993)
84. Martins, A., Smith, N., Aguiar, P., Figueiredo, M.: Structured sparsity in structured prediction. In: proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1500–1511 (2011)
85. McCoy, B., Wu, T.: The Two-Dimensional Ising Model. Harvard University Press, Cambridge, MA, US (1973)
86. Meier, L., Van De Geer, S., Bühlmann, P.: The group lasso for logistic regression. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **70**(1), 53–71 (2008)
87. Minsky, M.: Microscopy Apparatus. US Patent 3,013,467 (1961)
88. Mosci, S., Villa, S., Verri, A., Rosasco, L.: A primal–dual algorithm for group ℓ_1 regularization with overlapping groups. In: Proceedings of Neural Information Processing Systems (NIPS) Foundation (2010)
89. Narasimhan, M., Jojic, N., Bilmes, J.: Q-Clustering. In: Proceedings of Neural Information Processing Systems (NIPS) Foundation (2005)
90. Natarajan, B.: Sparse approximate solutions to linear systems. *SIAM J. Comput.* **24**(2), 227–234 (1995)
91. Needell, D., Tropp, J.: COSAMP: iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.* **26**(3), 301–321 (2009)
92. Nemhauser, G., Wolsey, L.: Integer and Combinatorial Optimization, vol. 18. Wiley, New York (1988)
93. Nemhauser, G., Wolsey, L., Fisher, M.: An analysis of approximations for maximizing submodular set functions. *Math. Program.* **14**(1), 265–294 (1978)
94. Nemirovskii, A.: Proximal-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex–concave saddle point problems. *SIAM J. Optim.* **15**(1), 229–251 (2004)
95. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Sov. Math. Dokl.* **27**, 372–376 (1983)
96. Nesterov, Y.: Excessive gap technique in nonsmooth convex minimization. *SIAM J. Optim.* **16**(1), 235–249 (2005)
97. Nesterov, Y.: Smooth minimization of nonsmooth functions. *Math. Program.* **103**(1), 127–152 (2005)

98. Nesterov, Y.: Primal–dual subgradient methods for convex problems. *Math. Program.* **120**(1, Ser. B), 221–259 (2009)
99. Obozinski, G., Bach, F.: Convex relaxation for combinatorial penalties (2012). arXiv preprint arXiv:1205.1240
100. Obozinski, G., Jacob, L., Vert, J.: Group lasso with overlaps: The latent group lasso approach (2011). arXiv preprint arXiv:1110.0413
101. Orlin, J.: A faster strongly polynomial time algorithm for submodular function minimization. *Math. Program.* **118**(2), 237–251 (2009)
102. Puig, A., Wiesel, A., Zaas, A., Woods, C., Ginsburg, G., Fleury, G., Hero, A.: Order-preserving factor analysis—application to longitudinal gene expression. *IEEE Trans. Signal Process.* **59**, 4447–4458 (2011)
103. Rao, N., Nowak, R., Wright, S., Kingsbury, N.: Convex approaches to model wavelet sparsity patterns. In: Proceedings of 18th IEEE International Conference on Image Processing (ICIP), pp. 1917–1920 (2011)
104. Rao, N., Recht, B., Nowak, R.: Signal recovery in unions of subspaces with applications to compressive imaging (2012). arXiv preprint arXiv:1209.3079
105. Rapaport, F., Barillot, E., Vert, J.: Classification of arrayCGH data using fused SVM. *Bioinformatics* **24**(13), 375–i382 (2008)
106. Rebafka, T., LÃ©vy-Leduc, C., Charbit, M.: OMP-type algorithm with structured sparsity patterns for multipath radar signals (2011). arXiv preprint arXiv:1103.5158
107. Robinson, S.: Strongly regular generalized equations. *Math. Oper. Res.* **5**, 43–62 (1980)
108. Schmidt, M., Roux, N.L., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Proceedings of Neural Information Processing Systems (NIPS) Foundation (2011)
109. Seeger, M.: On the Submodularity of Linear Experimental Design. Technical Report (2009)
110. Shapiro, J.: Embedded image coding using zero trees of wavelet coefficients. *IEEE Trans. Signal Process.* **41**(12), 3445–3462 (1993)
111. Sheppard, C., Shotton, D.: Confocal Laser Scanning Microscopy. BIOS Scientific Publishers, Garland Science, New York, US (1997)
112. Simon, N., Friedman, J., Hastie, T., Tibshirani, R.: A sparse-group lasso. *J. Comput. Graph. Stat.* **22**(2), 231–245 (2013)
113. Stojnic, M., Parvaresh, F., Hassibi, B.: On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Trans. Signal Process.* **57**(8) 3075–3085 (2009)
114. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**(43), 15545–15550 (2005)
115. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **28**(1), 267–288 (1996)
116. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **67**(1), 91–108 (2005)
117. Tran-Dinh, Q., Cevher, V.: An Optimal Primal–Dual Decomposition Framework. Technical Report, LIONS – EPFL (2014)
118. Tran-Dinh, Q., Cevher, V.: A Unified Optimal Primal–Dual Framework for Constrained Convex Minimization. Technical Report, LIONS, pp. 1–32 (2014)
119. Tran-Dinh, Q., Cevher, V.: Constrained convex minimization via model-based excessive gap. In: Proceedings of the Neural Information Processing Systems Foundation Conference (NIPS) (2014)
120. Tran Dinh, Q., Kyriolidis, A., Cevher, V.: Composite self-concordant minimization (2013). arXiv preprint arXiv:1308.2867
121. Tran Dinh, Q., Kyriolidis, A., Cevher, V.: A proximal Newton framework for composite minimization: graph learning without Cholesky decompositions and matrix inversions. In: Proceedings of The 30th International Conference on Machine Learning (ICML), pp. 271–279 (2013)

122. Tropp, J., Gilbert, A.: Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory* **53**(12), 4655–4666 (2007)
123. Tseng, P.: Applications of splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Optim.* **29**, 119–138 (1991)
124. Villa, S., Rosasco, L., Mosci, S., Verri, A.: Proximal methods for the latent group lasso penalty. *Comput. Optim. Appl.* **58**(2), 1–27 (2012)
125. Villa, S., Salzo, S., Baldassarre, L., Verri, A.: Accelerated and inexact forward–backward algorithms. *SIAM J. Optim.* **23**(3), 1607–1633 (2013)
126. Vincent, M., Hansen, N.: Sparse group lasso and high dimensional multinomial classification. *Comput. Stat. Data Anal.* **71**, 771–786 (2014)
127. Wright, S., Nowak, R., Figueiredo, M.: Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.* **57**(7), 2479–2493 (2009)
128. Wright, S., Nocedal, J.: Numerical Optimization. Springer, New York (1999)
129. Yuan, L., Liu, J., Ye, J.: Efficient methods for overlapping group lasso. In: Proceedings of Neural Information Processing Systems (NIPS) Foundation, pp. 352–360 (2011)
130. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **68**(1), 49–67 (2006)
131. Zeng, X., Figueiredo, M.: A novel sparsity and clustering regularization (2013). arXiv preprint arXiv:1310.4945
132. Zhang, Z., Shi, Y., Yin, B.: MR images reconstruction based on TV-group sparse model. In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2013)
133. Zhao, P., Rocha, G., Yu, B.: The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Stat.* **37**(6A), 3468–3497 (2009)
134. Zhou, H., Sehl, M.E., Sinsheimer, J.S., Lange, K.: Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* **26**(19), 2375 (2010)
135. Zhou, Y., Jin, R., Hoi, S.: Exclusive lasso for multi-task feature selection. In: Proceedings of International Conference on Artificial Intelligence and Statistics, pp. 988–995 (2010)

Chapter 13

Explicit Matrices with the Restricted Isometry Property: Breaking the Square-Root Bottleneck

Dustin G. Mixon

Abstract Matrices with the restricted isometry property (RIP) are of particular interest in compressed sensing. To date, the best known RIP matrices are constructed using random processes, while explicit constructions are notorious for performing at the “square-root bottleneck,” i.e., they only accept sparsity levels on the order of the square root of the number of measurements. The only known explicit matrix which surpasses this bottleneck was constructed by Bourgain, Dilworth, Ford, Konyagin, and Kutzarova in Bourgain et al. (Duke Math. J. 159:145–185, 2011). This chapter provides three contributions to advance the groundbreaking work of Bourgain et al.: (i) we develop an intuition for their matrix construction and underlying proof techniques; (ii) we prove a generalized version of their main result; and (iii) we apply this more general result to maximize the extent to which their matrix construction surpasses the square-root bottleneck.

13.1 Introduction

A matrix Φ is said to satisfy the (K, δ) -restricted isometry property (RIP) if

$$(1 - \delta)\|x\|^2 \leq \|\Phi x\|^2 \leq (1 + \delta)\|x\|^2$$

for every K -sparse vector x . RIP matrices are useful when compressively sensing signals which are sparse in some known orthonormal basis. Indeed, if there is a unitary sparsity matrix Ψ such that every signal of interest x has the property that Ψx is K -sparse, then any such x can be stably reconstructed from measurements of the form $y = Ax$ by minimizing $\|\Psi x\|_1$ subject to the measurements, provided $A\Psi^{-1}$ satisfies (K, δ) -RIP with $\delta < 1/3$ [9]. For sensing regimes in which measurements are costly, it is desirable to minimize the number of measurements necessary for signal reconstruction; this corresponds to the number of rows M in the $M \times N$ sensing matrix A . One can apply the theory of Gelfand widths to show that

D.G. Mixon (✉)

Air Force Institute of Technology, Wright-Patterson Air Force Base, OH, USA

e-mail: dustin.mixon@afit.edu

stable reconstruction by L1-minimization requires $K = O(M/\log(N/M))$ [4], and random matrices show that this bound is essentially tight; indeed, $M \times N$ matrices with iid subgaussian entries satisfy $(2K, \delta)$ -RIP with high probability provided $M = \Omega_\delta(K \log(N/K))$ [13].

Unfortunately, random matrices are not always RIP, though the failure rate vanishes asymptotically. In applications, you might wish to verify that your randomly drawn matrix actually satisfies RIP before designing your sensing platform around that matrix, but unfortunately, this is NP-hard in general [2]. As such, one is forced to blindly assume that the randomly drawn matrix is RIP, and admittedly, this is a reasonable assumption considering the failure rate. Still, this is dissatisfying from a theoretical perspective, and it motivates the construction of explicit RIP matrices:

Definition 1. For any $z > 0$, let $\text{ExRIP}[z]$ denote the following statement:

There exists an explicit family of $M \times N$ matrices with arbitrarily large aspect ratio N/M which are (K, δ) -RIP with $K = \Omega(M^{z-\varepsilon})$ for all $\varepsilon > 0$ and $\delta < 1/3$.

Since there exist (non-explicit) matrices satisfying $z = 1$ above, the goal is to prove $\text{ExRIP}[1]$. The most common way to demonstrate that an explicit matrix Φ satisfies RIP is to leverage the pairwise incoherence between the columns of Φ . Indeed, it is straightforward to prove $\text{ExRIP}[1/2]$ by taking Φ to have near-optimally incoherent unit-norm columns and appealing to interpolation of operators or Gershgorin's circle theorem (e.g., see, [1, 11, 12]). The emergence of this “square-root bottleneck” compelled Tao to pose the explicit construction of RIP matrices as an open problem [18]. Since then, only one construction has managed to break the bottleneck: In [8], Bourgain, Dilworth, Ford, Konyagin, and Kutzarova prove $\text{ExRIP}[1/2 + \varepsilon_0]$ for some undisclosed $\varepsilon_0 > 0$. This constant has since been estimated as $\varepsilon_0 \approx 5.5169 \times 10^{-28}$ [15].

Instead of estimating δ in terms of coherence, Bourgain et al. leverage additive combinatorics to construct Φ and to demonstrate certain cancellations in the Gram matrix $\Phi^* \Phi$. Today (three years later), this is the only known explicit construction which breaks the square-root bottleneck, thereby leading to two natural questions:

- What are the proof techniques that Bourgain et al. applied?
- Can we optimize the analysis to increase ε_0 ?

These questions were investigated recently in a series of blog posts [15–17], on which this chapter is based. In the next section, we provide some preliminaries—we first cover the techniques used in [8] to demonstrate RIP, and then we discuss some basic additive combinatorics to motivate the matrix construction. Section 13.3 then describes the construction of Φ , namely a subcollection of the chirps studied in [10], and discusses one method of selecting chirps (i.e., the method of Bourgain et al.). Section 13.4 provides the main result, namely *the BDFKK restricted isometry machine*, which says that a “good” selection of chirps will result in an RIP matrix construction which breaks the square-root bottleneck. This is a generalization of the main result in [8], as it offers more general sufficient conditions for good

chirp selection, but the proof is similar. After generalizing the sufficient conditions, we optimize over these conditions to increase the largest known ε_0 for which ExRIP[1/2 + ε_0] holds:

$$\varepsilon_0 \approx 4.4466 \times 10^{-24}.$$

Of course, any improvement to the chirp selection method will further increase this constant, and hopefully, the BDFKK restricted isometry machine and overall intuition provided in this chapter will foster such progress.¹ Section 13.5 contains the proofs of certain technical lemmas that are used to prove the main result.

13.2 Preliminaries

The goal of this section is to provide some intuition for the main ideas in [8]. We first explain the overall proof technique for demonstrating RIP (this is the vehicle for breaking the square-root bottleneck), and then we introduce some basic ideas from additive combinatorics. Throughout, $\mathbb{Z}/n\mathbb{Z}$ denotes the cyclic group of n elements, S^n denotes the cartesian power of a set S (i.e., the set of n -tuples with entries from S), \mathbb{F}_p denotes the field of p elements (in this context, p will always be prime), and \mathbb{F}_p^* is the multiplicative subgroup of \mathbb{F}_p .

13.2.1 The Big-Picture Techniques

Before explaining how Bourgain et al. broke the square-root bottleneck, let's briefly discuss the more common, coherence-based technique to demonstrate RIP. Let $\Phi_{\mathcal{K}}$ denote the submatrix of Φ whose columns are indexed by $\mathcal{K} \subseteq \{1, \dots, N\}$. Then (K, δ) -RIP equivalently states that, for every \mathcal{K} of size K , the eigenvalues of $\Phi_{\mathcal{K}}^* \Phi_{\mathcal{K}}$ lie in $[1 - \delta, 1 + \delta]$. As such, we can prove that a matrix is RIP by approximating eigenvalues. To this end, if we assume the columns of Φ have unit norm, and if we let μ denote the largest off-diagonal entry of $\Phi^* \Phi$ in absolute value (this is the worst-case coherence of the columns of Φ), then the Gershgorin circle theorem implies that Φ is $(K, (K-1)\mu)$ -RIP. Unfortunately, the coherence can't be too small, due to the Welch bound [20]:

$$\mu \geq \sqrt{\frac{N-M}{M(N-1)}},$$

¹Since writing the original draft of this chapter, K. Ford informed the author that an alternative to the chirp selection method is provided in [7]. We leave the impact on ε_0 for future work.

which is $\Omega(M^{-1/2})$ provided $N \geq cM$ for some $c > 1$. Thus, to get $(K-1)\mu = \delta < 1/2$, we require $K < 1/(2\mu) + 1 = O(M^{1/2})$. This is much smaller than the random RIP constructions which instead take $K = O(M^{1-\varepsilon})$ for all $\varepsilon > 0$, thereby revealing the shortcoming of the Gershgorin technique.

Now let's discuss the alternative techniques that Bourgain et al. use. The main idea is to convert the RIP statement, which concerns all K -sparse vectors simultaneously, into a statement about finitely many vectors:

Definition 2 (flat RIP). We say $\Phi = [\varphi_1 \cdots \varphi_N]$ satisfies (K, θ) -flat RIP if for every disjoint $I, J \subseteq \{1, \dots, N\}$ of size $\leq K$,

$$\left| \left\langle \sum_{i \in I} \varphi_i, \sum_{j \in J} \varphi_j \right\rangle \right| \leq \theta \sqrt{|I||J|}.$$

Lemma 1 (essentially Lemma 3 in [8], cf. Theorem 13 in [3]). *If Φ has (K, θ) -flat RIP and unit-norm columns, then Φ has $(K, 150\theta \log K)$ -RIP.*

Unlike the coherence argument, flat RIP doesn't lead to much loss in K . In particular, [3] shows that random matrices satisfy (K, θ) -flat RIP with $\theta = O(\delta/\log K)$ when $M = \Omega((K/\delta^2)\log^2 K \log N)$. As such, it makes sense that flat RIP would be a vehicle to break the square-root bottleneck. However, in practice, it's difficult to control both the left- and right-hand sides of the flat RIP inequality—it would be much easier if we only had to worry about getting cancellations, and not getting different levels of cancellation for different-sized subsets. This leads to the following:

Definition 3 (weak flat RIP). We say $\Phi = [\varphi_1 \cdots \varphi_N]$ satisfies (K, θ') -weak flat RIP if for every disjoint $I, J \subseteq \{1, \dots, N\}$ of size $\leq K$,

$$\left| \left\langle \sum_{i \in I} \varphi_i, \sum_{j \in J} \varphi_j \right\rangle \right| \leq \theta' K.$$

Lemma 2 (essentially Lemma 1 in [8]). *If Φ has (K, θ') -weak flat RIP and worst-case coherence $\mu \leq 1/K$, then Φ has $(K, \sqrt{\theta'})$ -flat RIP.*

Proof. By the triangle inequality, we have

$$\left| \left\langle \sum_{i \in I} \varphi_i, \sum_{j \in J} \varphi_j \right\rangle \right| \leq \sum_{i \in I} \sum_{j \in J} |\langle \varphi_i, \varphi_j \rangle| \leq |I||J|\mu \leq |I||J|/K.$$

Since Φ also has weak flat RIP, we then have

$$\left| \left\langle \sum_{i \in I} \varphi_i, \sum_{j \in J} \varphi_j \right\rangle \right| \leq \min\{\theta' K, |I||J|/K\} \leq \sqrt{\theta' |I||J|}. \quad \square$$

Unfortunately, this coherence requirement puts K back in the square-root bottleneck, since $\mu \leq 1/K$ is equivalent to $K \leq 1/\mu = O(M^{1/2})$. To rectify this, Bourgain et al. use a trick in which a modest K with tiny δ can be converted to a large K with modest δ :

Lemma 3 (buried in Lemma 3 in [8], cf. Theorem 1 in [14]). *If Φ has (K, δ) -RIP, then Φ has $(sK, 2s\delta)$ -RIP for all $s \geq 1$.*

In [14], this trick is used to get RIP results for larger K when testing RIP for smaller K . For the explicit RIP matrix problem, we are stuck with proving how small δ is when K is on the order of $M^{1/2}$. Note that this trick will inherently exhibit some loss in K . Assuming the best possible scaling for all N, K , and δ is $M = \Theta((K/\delta^2)\log(N/K))$, then if $N = \text{poly}(M)$, you can get $(M^{1/2}, \delta)$ -RIP only if $\delta = \Omega((\log^{1/2} M)/M^{1/4})$. In this best-case scenario, you would want to pick $s = M^{1/4-\varepsilon}$ for some $\varepsilon > 0$ and apply Lemma 3 to get $K = O(M^{3/4-\varepsilon})$. In some sense, this is another manifestation of the square-root bottleneck, but it would still be a huge achievement to saturate this bound.

13.2.2 A Brief Introduction to Additive Combinatorics

In this subsection, we briefly detail some key ideas from additive combinatorics; the reader who is already familiar with the subject may proceed to the next section, whereas the reader who wants to learn more is encouraged to see [19] for a more complete introduction. Given an additive group G and finite sets $A, B \subseteq G$, we can define the sumset

$$A + B := \{a + b : a \in A, b \in B\},$$

the difference set

$$A - B := \{a - b : a \in A, b \in B\},$$

and the additive energy

$$E(A, B) := \#\{(a_1, a_2, b_1, b_2) \in A^2 \times B^2 : a_1 + b_1 = a_2 + b_2\}.$$

These definitions are useful in quantifying the additive structure of a set. In particular, consider the following:

Lemma 4. *A nonempty subset A of some additive group G satisfies the following inequalities:*

- (i) $|A + A| \geq |A|$
- (ii) $|A - A| \geq |A|$
- (iii) $E(A, A) \leq |A|^3$

with equality precisely when A is a translate of some subgroup of G .

Proof. For (i), pick $a \in A$. Then $|A + A| \geq |A + a| = |A|$. Considering

$$A + A = \bigcup_{a \in A} (A + a),$$

we have equality in (i) precisely when $A + A = A + a$ for every $a \in A$. Equivalently, given $a_0 \in A$, then for every $a \in A$, addition by $a - a_0$ permutes the members of $A + a_0$. This is further equivalent to the following: Given $a_0 \in A$, then for every $a \in A$, addition by $a - a_0$ permutes the members of $A - a_0$. It certainly suffices for $H := A - a_0$ to be a group, and it is a simple exercise to verify that this is also necessary.

The proof for (ii) is similar.

For (iii), we note that

$$E(A, A) = \#\{(a, b, c) \in A^3 : a + b - c \in A\} \leq |A|^3,$$

with equality precisely when A has the property that $a + b - c \in A$ for every $a, b, c \in A$. Again, it clearly suffices for $A - a_0$ to be a group, and necessity is a simple exercise. \square

The notion of additive structure is somewhat intuitive. You should think of a translate of a subgroup as having maximal additive structure. When the bounds (i), (ii) and (iii) are close to being achieved by A (e.g., A is an arithmetic progression), you should think of A as having a lot of additive structure. Interestingly, while there are different measures of additive structure (e.g., $|A - A|$ and $E(A, A)$), they often exhibit certain relationships (perhaps not surprisingly). The following is an example of such a relationship which is used throughout the paper by Bourgain et al. [8]:

Lemma 5 (Corollary 1 in [8]). *If $E(A, A) \geq |A|^3/K$, then there exists a set $A' \subseteq A$ such that $|A'| \geq |A|/(20K)$ and $|A' - A'| \leq 10^7 K^9 |A|$.*

In words, a set with a lot of additive energy necessarily has a large subset with a small difference set. This is proved using a version of the Balog–Szemerédi–Gowers lemma [5].

If translates of subgroups have maximal additive structure, then which sets have minimal additive structure? It turns out that random sets tend to (nearly) have this property, and one way to detect low additive structure is Fourier bias:

$$\|A\|_u := \max_{\substack{\theta \in G \\ \theta \neq 0}} |\widehat{1}_A(\theta)|,$$

where the Fourier transform ($\widehat{\cdot}$) used here has a $1/|G|$ factor in front (it is not unitary). For example, if $G = \mathbb{Z}/n\mathbb{Z}$, we take

$$\widehat{f}(\xi) := \frac{1}{|G|} \sum_{x \in G} f(x) e^{-2\pi i x \xi / n}.$$

Interestingly, $\|A\|_u$ captures how far $E(A, A)$ is from its minimal value $|A|^4/|G|$:

Lemma 6. *For any subset A of a finite additive group G , we have*

$$\|A\|_u^4 \leq \frac{1}{|G|^3} \left(E(A, A) - \frac{|A|^4}{|G|} \right) \leq \frac{|A|}{|G|} \|A\|_u^2.$$

In the next section, we will appeal to Lemma 6 to motivate the matrix construction used by Bourgain et al. [8]. We will also use some of the techniques in the proof of Lemma 6 to prove a key lemma (namely, Lemma 7):

Proof (Proof of Lemma 6). We will assume $G = \mathbb{Z}/n\mathbb{Z}$, but the proof generalizes. Denote $e_n(x) := e^{2\pi i x/n}$ and $\lambda_x := \#\{(a, a') \in A^2 : a - a' = x\}$. For the left-hand inequality, we consider

$$\sum_{\theta \in G} |\widehat{1}_A(\theta)|^4 = \sum_{\theta \in G} \left| \frac{1}{|G|} \sum_{a \in A} e_n(-\theta a) \right|^4 = \frac{1}{|G|^4} \sum_{\theta \in G} \left| \sum_{x \in G} \lambda_x e_n(-\theta x) \right|^2,$$

where the last step is by expanding $|w|^2 = w\bar{w}$. Then Parseval's identity simplifies this to $\frac{1}{|G|^3} \|\lambda\|_2^2 = \frac{1}{|G|^3} E(A, A)$. We use this to bound $\|A\|_u^4$:

$$\|A\|_u^4 = \max_{\substack{\theta \in G \\ \theta \neq 0}} |\widehat{1}_A(\theta)|^4 \leq \sum_{\substack{\theta \in G \\ \theta \neq 0}} |\widehat{1}_A(\theta)|^4 = \frac{1}{|G|^3} E(A, A) - \frac{|A|^4}{|G|^4}.$$

For the right-hand inequality, we apply Parseval's identity:

$$E(A, A) = \sum_{x \in G} \lambda_x^2 = \frac{1}{|G|} \sum_{\theta \in G} \left| \sum_{x \in G} \lambda_x e_n(-\theta x) \right|^2 = \frac{|A|^4}{|G|} + \frac{1}{|G|} \sum_{\substack{\theta \in G \\ \theta \neq 0}} \left| \sum_{x \in G} \lambda_x e_n(-\theta x) \right|^2$$

From here, we apply the expansion $|w|^2 = w\bar{w}$

$$\left| \sum_{a \in A} e_n(-\theta a) \right|^2 = \sum_{x \in G} \lambda_x e_n(-\theta x)$$

to continue:

$$\sum_{\substack{\theta \in G \\ \theta \neq 0}} \left| \sum_{x \in G} \lambda_x e_n(-\theta x) \right|^2 = \sum_{\substack{\theta \in G \\ \theta \neq 0}} \left| \sum_{a \in A} e_n(-\theta a) \right|^4 \leq \sum_{\substack{\theta \in G \\ \theta \neq 0}} (|G| \|A\|_u)^2 \left| \sum_{a \in A} e_n(-\theta a) \right|^2.$$

Applying Parseval's identity then gives

$$E(A, A) \leq \frac{|A|^4}{|G|} + (|G| \|A\|_u)^2 \cdot \frac{1}{|G|} \sum_{\theta \in G} \left| \sum_{a \in A} e_n(-\theta a) \right|^2 = \frac{|A|^4}{|G|} + |G|^2 \|A\|_u^2 |A|,$$

which is a rearrangement of the right-hand inequality. \square

13.3 The Matrix Construction

This section combines ideas from the previous section to introduce the matrix construction used by Bourgain et al. [8] to break the square-root bottleneck. The main idea is to construct a Gram matrix $\Phi^* \Phi$ whose entries exhibit cancellations for weak flat RIP (see Definition 3). By Lemma 6, we can control cancellations of complex exponentials

$$\left| \sum_{a \in A} e_n(\theta a) \right| \leq n \|A\|_u, \quad \theta \neq 0$$

in terms of the additive energy of the index set $A \subseteq \mathbb{Z}/n\mathbb{Z}$; recall that $e_n(x) := e^{2\pi i x/n}$. This motivates us to pursue a Gram matrix whose entries are complex exponentials. To this end, consider the following vector:

$$u_{a,b} := \frac{1}{\sqrt{p}} \left(e_p(ax^2 + bx) \right)_{x \in \mathbb{F}_p},$$

where p is prime and \mathbb{F}_p denotes the field of size p . Such vectors are called *chirps*, and they are used in a variety of applications including radar. Here, we are mostly interested in the form of their inner products. If $a_1 = a_2$, then $\langle u_{a_1,b_1}, u_{a_2,b_2} \rangle = \delta_{b_1,b_2}$ by the geometric sum formula. Otherwise, the inner product is more interesting:

$$\langle u_{a_1,b_1}, u_{a_2,b_2} \rangle = \frac{1}{p} \sum_{x \in \mathbb{F}_p} e_p((a_1 - a_2)x^2 + (b_1 - b_2)x).$$

Since $a_1 - a_2 \neq 0$, we can complete the square in the exponent, and changing variables to $y := x + (b_1 - b_2)/(2(a_1 - a_2))$ gives

$$\langle u_{a_1,b_1}, u_{a_2,b_2} \rangle = \frac{1}{p} e_p \left(-\frac{(b_1 - b_2)^2}{4(a_1 - a_2)} \right) \sum_{y \in \mathbb{F}_p} e_p((a_1 - a_2)y^2).$$

Finally, this can be simplified using a quadratic Gauss sum formula:

$$\langle u_{a_1,b_1}, u_{a_2,b_2} \rangle = \frac{\sigma_p}{\sqrt{p}} \left(\frac{a_1 - a_2}{p} \right) e_p \left(-\frac{(b_1 - b_2)^2}{4(a_1 - a_2)} \right),$$

where σ_p is 1 or i (depending on whether p is 1 or 3 mod 4) and $\left(\frac{a_1 - a_2}{p} \right)$ is a Legendre symbol, taking value ± 1 depending on whether $a_1 - a_2$ is a perfect square mod p . Modulo these factors, the above inner product is a complex exponential, and since we want these in our Gram matrix $\Phi^* \Phi$, we will take Φ to have columns of the form $u_{a,b}$ —in fact, the columns will be $\{u_{a,b}\}_{(a,b) \in \mathcal{A} \times \mathcal{B}}$ for some well-designed sets $\mathcal{A}, \mathcal{B} \subseteq \mathbb{F}_p$.

For weak flat RIP, we want to bound the following quantity for every $\Omega_1, \Omega_2 \subseteq \mathcal{A} \times \mathcal{B}$ with $|\Omega_1|, |\Omega_2| \leq \sqrt{p}$:

$$\left| \left\langle \sum_{(a_1, b_1) \in \Omega_1} u_{a_1, b_1}, \sum_{(a_2, b_2) \in \Omega_2} u_{a_2, b_2} \right\rangle \right|.$$

For $i = 1, 2$, define

$$A_i := \{a \in \mathcal{A} : \exists b \in \mathcal{B} \text{ s.t. } (a, b) \in \Omega_i\} \quad \text{and} \quad \Omega_i(a) := \{b \in \mathcal{B} : (a, b) \in \Omega_i\}.$$

These provide an alternate expression for the quantity of interest:

$$\begin{aligned} \left| \sum_{(a_1, b_1) \in \Omega_1} \sum_{(a_2, b_2) \in \Omega_2} \langle u_{a_1, b_1}, u_{a_2, b_2} \rangle \right| &= \left| \sum_{a_1 \in A_1} \sum_{\substack{b_1 \in \Omega_1(a_1) \\ a_2 \in A_2 \\ b_2 \in \Omega_2(a_2)}} \langle u_{a_1, b_1}, u_{a_2, b_2} \rangle \right|, \\ &\leq \sum_{\substack{a_1 \in A_1 \\ a_2 \in A_2}} \left| \sum_{\substack{b_1 \in \Omega_1(a_1) \\ b_2 \in \Omega_2(a_2)}} \langle u_{a_1, b_1}, u_{a_2, b_2} \rangle \right| \\ &= \frac{1}{\sqrt{p}} \sum_{\substack{a_1 \in A_1 \\ a_2 \in A_2}} \left| \sum_{\substack{b_1 \in \Omega_1(a_1) \\ b_2 \in \Omega_2(a_2)}} e_p \left(-\frac{(b_1 - b_2)^2}{4(a_1 - a_2)} \right) \right|. \end{aligned}$$

Pleasingly, it now suffices to bound a sum of complex exponentials, which we feel equipped to do using additive combinatorics. The following lemma does precisely this (it can be viewed as an analog of Lemma 6).

Lemma 7 (Lemma 9 in [8]). *For every $\theta \in \mathbb{F}_p^*$ and $B_1, B_2 \subseteq \mathbb{F}_p$, we have*

$$\left| \sum_{\substack{b_1 \in B_1 \\ b_2 \in B_2}} e_p \left(\theta(b_1 - b_2)^2 \right) \right| \leq |B_1|^{1/2} E(B_1, B_1)^{1/8} |B_2|^{1/2} E(B_2, B_2)^{1/8} p^{1/8}.$$

Proof. First, Cauchy–Schwarz gives

$$\begin{aligned} \left| \sum_{\substack{b_1 \in B_1 \\ b_2 \in B_2}} e_p \left(\theta(b_1 - b_2)^2 \right) \right|^2 &= \left| \sum_{b_1 \in B_1} 1 \cdot \sum_{b_2 \in B_2} e_p \left(\theta(b_1 - b_2)^2 \right) \right|^2 \\ &\leq |B_1| \sum_{b_1 \in B_1} \left| \sum_{b_2 \in B_2} e_p \left(\theta(b_1 - b_2)^2 \right) \right|^2. \end{aligned}$$

Expanding $|w|^2 = w\bar{w}$ and rearranging then gives an alternate expression for the right-hand side:

$$|B_1| \sum_{b_2, b'_2 \in B_2} e_p \left(\theta(b_2^2 - (b'_2)^2) \right) \overline{\sum_{b_1 \in B_1} e_p \left(\theta(2b_1(b_2 - b'_2)) \right)}.$$

Applying Cauchy–Schwarz again, we then have

$$\left| \sum_{\substack{b_1 \in B_1 \\ b_2 \in B_2}} e_p(\theta(b_1 - b_2)^2) \right|^4 \leq |B_1|^2 |B_2|^2 \sum_{b_2, b'_2 \in B_2} \left| \sum_{b_1 \in B_1} e_p(\theta(2b_1(b_2 - b'_2))) \right|^2,$$

and expanding $|w|^2 = w\bar{w}$ this time gives

$$|B_1|^2 |B_2|^2 \sum_{\substack{b_1, b'_1 \in B_1 \\ b_2, b'_2 \in B_2}} e_p(2\theta(b_1 - b'_1)(b_2 - b'_2)).$$

At this point, it is convenient to change variables, namely, $x = b_1 - b'_1$ and $y = b_2 - b'_2$:

$$\left| \sum_{\substack{b_1 \in B_1 \\ b_2 \in B_2}} e_p(\theta(b_1 - b_2)^2) \right|^4 \leq |B_1|^2 |B_2|^2 \sum_{x, y \in \mathbb{F}_p} \lambda_x \mu_y e_p(2\theta xy), \quad (13.1)$$

where $\lambda_x := \#\{(b_1, b'_1) \in B_1^2 : b_1 - b'_1 = x\}$ and similarly for μ_y in terms of B_2 . We now apply Cauchy–Schwarz again to bound the sum in (13.1):

$$\left| \sum_{x \in \mathbb{F}_p} \lambda_x \sum_{y \in \mathbb{F}_p} \mu_y e_p(2\theta xy) \right|^2 \leq \|\lambda\|_2^2 \sum_{x \in \mathbb{F}_p} \left| \sum_{y \in \mathbb{F}_p} \mu_y e_p(2\theta xy) \right|^2,$$

and changing variables $x' := -2\theta x$ (this change is invertible since $\theta \neq 0$), we see that the right-hand side is a sum of squares of the Fourier coefficients of μ . As such, Parseval’s identity gives the following simplification:

$$\left| \sum_{x, y \in \mathbb{F}_p} \lambda_x \mu_y e_p(2\theta xy) \right|^2 \leq p \|\lambda\|_2^2 \|\mu\|_2^2 = p E(B_1, B_1) E(B_2, B_2).$$

Applying this bound to (13.1) gives the result. \square

13.3.1 How to Construct \mathcal{B}

Lemma 7 enables us to prove weak-flat-RIP-type cancellations in cases where $\Omega_1(a_1), \Omega_2(a_2) \subseteq \mathcal{B}$ both lack additive structure. Indeed, the method of [8] is to do precisely this, and the remaining cases (where either $\Omega_1(a_1)$ or $\Omega_2(a_2)$ has more additive structure) will find cancellations by accounting for the dilation weights $1/(a_1 - a_2)$. Overall, we will be very close to proving that Φ is RIP if

most subsets of \mathcal{B} lack additive structure. To this end, Bourgain et al. [8] actually prove something much stronger: They design \mathcal{B} in such a way that all sufficiently large subsets have low additive structure. The following theorem is the first step in the design:

Theorem 1 (Theorem 5 in [8]). *Fix $r, M \in \mathbb{N}$, $M \geq 2$, and define the cube $\mathcal{C} := \{0, \dots, M-1\}^r \subseteq \mathbb{Z}^r$. Let τ denote the solution to the equation*

$$\left(\frac{1}{M}\right)^{2\tau} + \left(\frac{M-1}{M}\right)^\tau = 1.$$

Then for any subsets $A, B \subseteq \mathcal{C}$, we have

$$|A + B| \geq (|A||B|)^\tau.$$

As a consequence of this theorem (taking $A = B$), we have $|A + A| \geq |A|^{2\tau}$ for every $A \subseteq \mathcal{C}$, and since $\tau > 1/2$, this means that large subsets A have $|A + A| \gg |A|$, indicating low additive structure. However, \mathcal{C} is a subset of the group \mathbb{Z}^r , whereas we need to construct a subset \mathcal{B} of \mathbb{F}_p . The trick here is to pick \mathcal{B} so that it inherits the additive structure of \mathcal{C} :

$$\mathcal{B} := \left\{ \sum_{j=1}^r x_j (2M)^{j-1} : x_1, \dots, x_r \in \{0, \dots, M-1\} \right\}. \quad (13.2)$$

Indeed, the $2M$ -ary expansion of $b_1, b_2 \in \mathcal{B}$ reveals the corresponding $c_1, c_2 \in \mathcal{C}$. Also, adding b_1 and b_2 incurs no carries, so the expansion of $b_1 + b_2$ corresponds to $c_1 + c_2$ (even when $c_1 + c_2 \notin \mathcal{C}$). This type of mapping $\mathcal{C} \rightarrow \mathbb{F}_p$ is called a *Freiman isomorphism*, and it's easy to see that Freiman isomorphic sets have the same sized sumsets, difference sets, and additive energy.

We already know that large subsets of \mathcal{C} (and \mathcal{B}) exhibit low additive structure, but the above theorem only gives this in terms of the sumset, whereas Lemma 7 requires low additive structure in terms of additive energy. As such, we will first convert the above theorem into a statement about difference sets, and then apply Lemma 5 to further convert it in terms of additive energy:

Corollary 1 (essentially Corollary 3 in [8]). *Fix r, M and τ according to Theorem 1, take \mathcal{B} as defined in (13.2), and pick s and t such that $(2\tau - 1)s \geq t$. Then every subset $B \subseteq \mathcal{B}$ such that $|B| > p^s$ satisfies $|B - B| > p^t |B|$.*

Proof. First note that $-B$ is a translate of some other set $B' \subseteq \mathcal{B}$. Explicitly, if $b_0 = \sum_{j=1}^r (M-1)(2M)^{j-1}$, then we can take $B' := b_0 - B$. As such, Theorem 1 gives

$$|B - B| = |B + B'| \geq |B|^{2\tau} = |B|^{2\tau-1} |B| > p^{(2\tau-1)s} |B| \geq p^t |B|. \quad \square$$

Corollary 2 (essentially Corollary 4 in [8]). *Fix r , M and τ according to Theorem 1, take \mathcal{B} as defined in (13.2), and pick γ and ℓ such that $(2\tau-1)(\ell-\gamma) \geq 10\gamma$. Then for every $\varepsilon > 0$, there exists P such that for every $p \geq P$, every subset $S \subseteq \mathcal{B}$ with $|S| > p^\ell$ satisfies $E(S, S) < p^{-\gamma+\varepsilon}|S|^3$.*

Proof. Suppose to the contrary that there exists $\varepsilon > 0$ such that there are arbitrarily large p for which there is a subset $S \subseteq \mathcal{B}$ with $|S| > p^\ell$ and $E(S, S) \geq p^{-\gamma+\varepsilon}|S|^3$. Writing $E(S, S) = |S|^3/K$, then $K \leq p^{\gamma-\varepsilon}$. By Lemma 5, there exists $B \subseteq S$ such that, for sufficiently large p ,

$$|B| \geq |S|/(20K) > \frac{1}{20}p^{\ell-\gamma+\varepsilon} > p^{\ell-\gamma},$$

and

$$|B - B| \leq 10^7 K^9 |S| \leq 10^7 K^9 (20K|B|) \leq 10^7 \cdot 20 \cdot p^{10(\gamma-\varepsilon)} |B| < p^{10\gamma} |B|.$$

However, this contradicts the previous corollary with $s = \ell - \gamma$ and $t = 10\gamma$. \square

Notice that we can weaken our requirements on γ and ℓ if we had a version of Lemma 5 with a smaller exponent on K . This exponent comes from a version of the Balog–Szemerédi–Gowers lemma (Lemma 6 in [8]), which follows from the proof of Lemma 2.2 in [5]. (Specifically, take $A = B$, and you need to change $A -_E B$ to $A +_E B$, but this change doesn't affect the proof.) Bourgain et al. indicate that it would be desirable to prove a better version of this lemma, but it is unclear how easy that would be.

13.3.2 How to Construct \mathcal{A}

The previous subsection showed how to construct \mathcal{B} so as to ensure that all sufficiently large subsets have low additive structure. By Lemma 7, this in turn ensures that Φ exhibits weak-flat-RIP-type cancellations for most $\Omega_1(a_1), \Omega_2(a_2) \subseteq \mathcal{B}$. For the remaining cases, Φ must exhibit weak-flat-RIP-type cancellations by somehow leveraging properties of \mathcal{A} .

The next section gives the main result, which requires a subset $\mathcal{A} = \mathcal{A}(p)$ of \mathbb{F}_p for which there exists an even number m as well as an $\alpha > 0$ (both independent of p) such that the following two properties hold:

- (i) $\Omega(p^\alpha) \leq |\mathcal{A}(p)| \leq p^\alpha$.
- (ii) For each $a \in \mathcal{A}$, then $a_1, \dots, a_{2m} \in \mathcal{A} \setminus \{a\}$ satisfies

$$\sum_{j=1}^m \frac{1}{a - a_j} = \sum_{j=m+1}^{2m} \frac{1}{a - a_j} \tag{13.3}$$

only if (a_1, \dots, a_m) and (a_{m+1}, \dots, a_{2m}) are permutations of each other. Here, division (and addition) is taken in the field \mathbb{F}_p .

Unfortunately, these requirements on \mathcal{A} lead to very little intuition compared to our current understanding of \mathcal{B} . Regardless, we will continue by considering how Bourgain et al. constructs \mathcal{A} . The following lemma describes their construction and makes a slight improvement to the value of α chosen in [8]:

Lemma 8. *Take $L := \lfloor p^{1/2m(4m-1)} \rfloor$ and $U := \lfloor L^{4m-1} \rfloor$. Then*

$$\mathcal{A} := \{x^2 + Ux : 1 \leq x \leq L\}$$

satisfies (i) and (ii) above if we take

$$\alpha = \frac{1}{2m(4m-1)}$$

and p is a sufficiently large prime.

The original proof of this result is sketched after the statement of Lemma 2 in [8]. Unfortunately, this proof contains a technical error: The authors conclude that a prime p does not divide some integer $D_1 D_2 V$ since $V \neq 0$, p does not divide D_1 and $|D_2 V| < p$, but the conclusion is invalid since $D_2 V$ is not necessarily an integer. The following alternative proof removes this difficulty:

Proof (Proof of Lemma 8). One may quickly verify (i). For (ii), we first note that multiplication by $\prod_{i=1}^{2m} (a - a_i)$ gives that (13.3) is equivalent to

$$S := \sum_{j=1}^{2m} \lambda_j \prod_{\substack{i=1 \\ i \neq j}}^{2m} (a - a_i) \equiv 0 \pmod{p},$$

where $\lambda_j = 1$ for $j \in \{1, \dots, m\}$ and $\lambda_j = -1$ for $j \in \{m+1, \dots, 2m\}$. Here, we are treating a and the a_i 's as integers in $\{0, \dots, p-1\}$, and so S is an integer. Furthermore,

$$|S| \leq m(2(L^2 + UL))^{2m-1} \leq m(2(p^{1/m(4m-1)} + p^{4m/2m(4m-1)}))^{2m-1} < p$$

for sufficiently large p . As such, we have $S = 0$ (not just $0 \pmod{p}$).

With this, it suffices to show that for any $n \in \{1, \dots, 2m\}$, any distinct $x, x_1, \dots, x_n \in \{1, \dots, L\}$, and any nonzero integers $\lambda_1, \dots, \lambda_n$ such that $|\lambda_1| + \dots + |\lambda_n| \leq 2m$,

$$W = \sum_{j=1}^n \lambda_j \prod_{\substack{i=1 \\ i \neq j}}^n (x - x_i)(x + x_i + U) \tag{13.4}$$

is nonzero (just as before, everything is an integer here). Indeed, taking $a = x^2 + Ux$ and $a_j = x_j^2 + Ux_j$, we see that S has the form of W , and so if W is always nonzero, so must S , which by the above reduction implies (ii).

To prove the claim, note that $x + x_1 + U$ is a factor of the j th term of (13.4) for every $j \geq 2$. As such, W is congruent to the first term modulo $x + x_1 + U$:

$$W \equiv \lambda_1 \prod_{i=2}^n (x - x_i)(x + x_i + U) \pmod{x + x_1 + U}.$$

Each factor of the form $x + x_i + U$ can be further simplified:

$$x + x_i + U = (x_i - x_1) + (x + x_1 + U) \equiv x_i - x_1 \pmod{x + x_1 + U},$$

and so

$$W \equiv W_1 \pmod{x + x_1 + U},$$

where

$$W_1 := \lambda_1 \prod_{i=2}^n (x - x_i)(x_i - x_1).$$

Note that W_1 is nonzero since x and the x_i 's are distinct by assumption. Also,

$$|W_1| \leq 2m(L^2)^{2m-1} \leq U < x + x_1 + U$$

for sufficiently large p . As such, $x + x_1 + U$ does not divide W_1 , and so W is nonzero, as desired. \square

13.4 The Main Result

We are now ready to state the main result of this chapter, which is a generalization of the main result in [8]. Later in this section, we will maximize ε_0 such that this result implies ExRIP[1/2 + ε_0] with the matrix construction from [8].

Theorem 2 (The BDFKK restricted isometry machine). *For every prime p , define subsets $\mathcal{A} = \mathcal{A}(p)$ and $\mathcal{B} = \mathcal{B}(p)$ of \mathbb{F}_p . Suppose there exist constants $m \in 2\mathbb{N}$, $\ell, \gamma > 0$ independent of p such that the following conditions apply:*

(a) *For every sufficiently large p , and for every $a \in \mathcal{A}$ and $a_1, \dots, a_{2m} \in \mathcal{A} \setminus \{a\}$,*

$$\sum_{j=1}^m \frac{1}{a - a_j} = \sum_{j=m+1}^{2m} \frac{1}{a - a_j}$$

only if (a_1, \dots, a_m) and (a_{m+1}, \dots, a_{2m}) are permutations of each other. Here, division (and addition) is taken in the field \mathbb{F}_p .

(b) For every $\varepsilon > 0$, there exists $P = P(\varepsilon)$ such that for every $p \geq P$, every subset $S \subseteq \mathcal{B}(p)$ with $|S| \geq p^\ell$ satisfies $E(S, S) \leq p^{-\gamma+\varepsilon} |S|^3$.

Pick α such that

$$\Omega(p^\alpha) \leq |\mathcal{A}(p)| \leq p^\alpha, \quad |\mathcal{B}(p)| \geq \Omega(p^{1-\alpha+\varepsilon'}) \quad (13.5)$$

for some $\varepsilon' > 0$ and every sufficiently large p . Pick $\varepsilon_1 > 0$ for which there exist $\alpha_1, \alpha_2, \varepsilon, x, y > 0$ such that

$$\varepsilon_1 + \varepsilon < \alpha_1 - \alpha - (4/3)x - \varepsilon, \quad (13.6)$$

$$\ell \leq 1/2 + (4/3)x - \alpha_1 + \varepsilon/2, \quad (13.7)$$

$$\varepsilon_1 + \varepsilon < \gamma/4 - y/4 - \varepsilon, \quad (13.8)$$

$$\alpha_2 \geq 9x + \varepsilon, \quad (13.9)$$

$$c_0y/8 - (\alpha_1/4 + 9\alpha_2/8)/m \leq x/8 - \alpha/4, \quad (13.10)$$

$$\varepsilon_1 + \varepsilon < c_0y/8 - (\alpha_1/4 + 9\alpha_2/8)/m, \quad (13.11)$$

$$my \leq \min\{1/2 - \alpha_1, 1/2 - \alpha_2\}, \quad (13.12)$$

$$3\alpha_2 - 2\alpha_1 \leq (2 - c_0)my. \quad (13.13)$$

Here, $c_0 = 1/10430$ is a constant from Proposition 2 in [6]. Then for sufficiently large p , the $p \times |\mathcal{A}(p)||\mathcal{B}(p)|$ matrix with columns

$$u_{a,b} := \frac{1}{\sqrt{p}} \left(e^{2\pi i (ax^2 + bx)/p} \right)_{x \in \mathbb{F}_p} \quad a \in \mathcal{A}, b \in \mathcal{B}$$

satisfies $(p^{1/2+\varepsilon_1/2-\varepsilon''}, \delta)$ -RIP for any $\varepsilon'' > 0$ and $\delta < \sqrt{2} - 1$, thereby implying ExRIP[1/2 + $\varepsilon_1/2$].

Let's briefly discuss the structure of the proof of this result. As indicated in Section 13.2, the method is to prove flat-RIP-type cancellations, namely that

$$S(A_1, A_2) := \sum_{\substack{a_1 \in A_1 \\ a_2 \in A_2}} \sum_{\substack{b_1 \in \Omega_1(a_1) \\ b_2 \in \Omega_2(a_2)}} \left(\frac{a_1 - a_2}{p} \right) e_p \left(\frac{(b_1 - b_2)^2}{2(a_1 - a_2)} \right) \quad (13.14)$$

has size $\leq p^{1-\varepsilon_1-\varepsilon}$ whenever Ω_1 and Ω_2 are disjoint with size $\leq \sqrt{p}$. (Actually, we get to assume that these subsets and the $\Omega_i(a_i)$'s satisfy certain size constraints since we have an extra $-\varepsilon$ in the power of p ; having this will imply the general case without the ε , as made clear in the proof of Theorem 2.) This bound is proved by considering a few different cases. First, when the $\Omega_i(a_i)$'s are small, (13.14) is small by a triangle inequality. Next, when the $\Omega_i(a_i)$'s are large, then we can apply a triangle inequality over each A_i and appeal to hypothesis (b) in Theorem 2

and Lemma 7. However, this will only give sufficient cancellation when the A_i 's are small. In the remaining case, Bourgain et al. prove sufficient cancellation by invoking Lemma 10 in [8], which concerns the following quantity:

$$T_{a_1}(A_2, B) := \sum_{\substack{b_1 \in B \\ a_2 \in A_2, b_2 \in \Omega_2(a_2)}} \left(\frac{a_1 - a_2}{p} \right) e_p \left(\frac{(b_1 - b_2)^2}{4(a_1 - a_2)} \right). \quad (13.15)$$

Specifically, Lemma 10 in [8] gives that $|T_{a_1}(A_2, B)|$ is small whenever B has sufficient additive structure. In the proof of the main result, they take a maximal subset $B_0 \subseteq \Omega_1(a_1)$ such that $|T_{a_1}(A_2, B_0)|$ is small, and then they use this lemma to show that $\Omega_1(a_1) \setminus B_0$ necessarily has little additive structure. By Lemma 7, this in turn forces $|T_{a_1}(A_2, B_1)|$ to be small, and so $|T_{a_1}(A_2, \Omega_1(a_1))|$ (and furthermore $|S(A_1, A_2)|$) are also small due to a triangle inequality. The reader is encouraged to find more details in the proofs found in Section 13.5.

What follows is a generalized version of the statement of Lemma 10 in [8], which we then use in the hypothesis of a generalized version of Lemma 2 in [8]:

Definition 4. Let $L10 = L10[\alpha_1, \alpha_2, k_0, k_1, k_2, m, y]$ denote the following statement about subsets $\mathcal{A} = \mathcal{A}(p)$ and $\mathcal{B} = \mathcal{B}(p)$ of \mathbb{F}_p for p prime:

For every $\varepsilon > 0$, there exists $P > 0$ such that for every prime $p \geq P$ the following holds:

Take $\Omega_1, \Omega_2 \subseteq \mathcal{A} \times \mathcal{B}$ such that

$$|A_2| \geq p^y, \quad (13.16)$$

and for which there exist powers of two M_1, M_2 such that

$$\frac{M_i}{2} \leq |\Omega_i(a_i)| < M_i \quad (13.17)$$

and

$$|A_i|M_i \leq 2\sqrt{p} \quad (13.18)$$

for $i = 1, 2$ and for every $a_i \in A_i$. Then for every $B \subseteq \mathbb{F}_p$ such that

$$p^{1/2-\alpha_1} \leq |B| \leq p^{1/2} \quad (13.19)$$

and

$$|B - B| \leq p^{\alpha_2}|B|, \quad (13.20)$$

we have that (13.15) satisfies

$$|T_{a_1}(A_2, B)| \leq |B|p^{1/2-\varepsilon_2+\varepsilon} \quad (13.21)$$

with $\varepsilon_2 = k_0y - (k_1\alpha_1 + k_2\alpha_2)/m$ for every $a_1 \in A_1$.

Lemma 9 (generalized version of Lemma 2 in [8]). *Take \mathcal{A} arbitrarily and \mathcal{B} satisfying the hypothesis (b) in Theorem 2, pick α such that $|\mathcal{A}(p)| \leq p^\alpha$ for every sufficiently large p , and pick $\varepsilon, \varepsilon_1, x > 0$ such that L10 holds with (13.6)–(13.9) and*

$$\varepsilon_1 + \varepsilon < \varepsilon_2 \leq x/8 - \alpha/4. \quad (13.22)$$

Then the following holds for every sufficiently large p :

Take $\Omega_1, \Omega_2 \subseteq \mathcal{A} \times \mathcal{B}$ for which there exist powers of two M_1, M_2 such that (13.17) and (13.18) hold for $i = 1, 2$ and every $a_i \in A_i$. Then (13.14) satisfies $|S(A_1, A_2)| \leq p^{1-\varepsilon_1-\varepsilon}$.

The following result gives sufficient conditions for L10, and thus Lemma 9 above:

Lemma 10 (generalized version of Lemma 10 in [8]). *Suppose \mathcal{A} satisfies hypothesis (a) in Theorem 2. Then L10 is true with $k_0 = c_0/8$, $k_1 = 1/4$ and $k_2 = 9/8$ provided (13.12) and (13.13) are satisfied.*

These lemmas are proved in Section 13.5. With these in hand, we are ready to prove the main result:

Proof (Proof of Theorem 2). By Lemma 10, we have that L10 is true with

$$\varepsilon_2 = c_0 y/8 - (\alpha_1/4 + 9\alpha_2/8)/m.$$

As such, (13.10) and (13.11) together imply (13.22), and so the conclusion of Lemma 9 holds. We will use this conclusion to show that the matrix identified in Theorem 2 satisfies $(p^{1/2}, p^{-\varepsilon_1})$ -weak flat RIP. Indeed, this will imply $(p^{1/2}, p^{-\varepsilon_1/2})$ -flat RIP by Lemma 2, $(p^{1/2}, 75p^{-\varepsilon_1/2} \log p)$ -RIP by Lemma 1, and $(p^{1/2+\varepsilon_1/2-\varepsilon''}, 75p^{-\varepsilon''} \log p)$ -RIP for any $\varepsilon'' > 0$ by Lemma 3 (taking $s = p^{\varepsilon_1/2-\varepsilon''}$). Since $75p^{-\varepsilon''} \log p < \sqrt{2} - 1$ for sufficiently large p , this will prove the result.

To demonstrate $(p^{1/2}, p^{-\varepsilon_1})$ -weak flat RIP, pick disjoint $\Omega_1, \Omega_2 \subseteq \mathcal{A} \times \mathcal{B}$ of size $\leq p^{1/2}$. We need to show

$$\left| \left\langle \sum_{(a_1, b_1) \in \Omega_1} u_{a_1, b_1}, \sum_{(a_2, b_2) \in \Omega_2} u_{a_2, b_2} \right\rangle \right| \leq p^{1/2-\varepsilon_1}.$$

Recall that

$$\begin{aligned} & \sum_{(a_1, b_1) \in \Omega_1} \sum_{(a_2, b_2) \in \Omega_2} \langle u_{a_1, b_1}, u_{a_2, b_2} \rangle \\ &= \sum_{\substack{a_1 \in A_1 \\ a_2 \in A_2}} \sum_{\substack{b_1 \in \Omega_1(a_1) \\ b_2 \in \Omega_2(a_2)}} \langle u_{a_1, b_1}, u_{a_2, b_2} \rangle \\ &= \sum_{\substack{a_1 \in A_1 \\ a_2 \in A_2}} \sum_{\substack{b_1 \in \Omega_1(a_1) \\ b_2 \in \Omega_2(a_2)}} \frac{\sigma_p}{\sqrt{p}} \left(\frac{a_1 - a_2}{p} \right) e_p \left(-\frac{(b_1 - b_2)^2}{4(a_1 - a_2)} \right). \end{aligned}$$

As such, we may assume that A_1 and A_2 are disjoint without loss of generality, and it suffices to show that

$$|S(A_1, A_2)| = \left| \sum_{\substack{a_1 \in A_1 \\ a_2 \in A_2}} \sum_{\substack{b_1 \in \Omega_1(a_1) \\ b_2 \in \Omega_2(a_2)}} \left(\frac{a_1 - a_2}{p} \right) e_p \left(-\frac{(b_1 - b_2)^2}{4(a_1 - a_2)} \right) \right| \leq p^{1-\varepsilon_1}.$$

To estimate this sum, it is convenient to partition A_i according to the size of $\Omega_i(a_i)$. For each k , define the set

$$A_i^{(k)} := \{a_i \in A_i : 2^{k-1} \leq |\Omega_i(a_i)| < 2^k\}.$$

Then we have

$$|A_i^{(k)}| 2^{k-1} \leq \sum_{a_i \in A_i^{(k)}} |\Omega_i(a_i)| = |\{(a_i, b_i) \in \Omega_i : a_i \in A_i^{(k)}\}| \leq |\Omega_i| \leq p^{1/2}.$$

As such, taking $M_i = 2^k$ gives that $A_i \leftarrow A_i^{(k)}$ satisfies (13.17) and (13.18), which enables us to apply the conclusion of Lemma 9. Indeed, the triangle inequality and Lemma 9 together give

$$|S(A_1, A_2)| \leq \sum_{k_1=1}^{\lceil \frac{1}{2} \log_2 p \rceil} \sum_{k_2=1}^{\lceil \frac{1}{2} \log_2 p \rceil} |S(A_1^{(k_1)}, A_2^{(k_2)})| \leq p^{1-\varepsilon_1-\varepsilon} \log^2 p \leq p^{1-\varepsilon_1},$$

where the last step takes p sufficiently large, i.e., such that $p^\varepsilon \geq \log^2 p$. \square

To summarize Theorem 2, we may conclude $\text{ExRIP}[1/2 + \varepsilon_1/2]$ if we can find

- (i) $m \in 2\mathbb{N}$ satisfying hypothesis (a),
- (ii) $\ell, \gamma > 0$ satisfying hypothesis (b),
- (iii) α satisfying (13.5), and
- (iv) $\alpha_1, \alpha_2, \varepsilon, x, y > 0$ satisfying (13.6)–(13.13).

Since we want to conclude $\text{ExRIP}[z]$ for the largest possible z , we are inclined to maximize ε_1 subject to (i)–(iv), above. To find m, ℓ, γ, α which satisfy (i)–(iii), we must leverage a particular construction of \mathcal{A} and \mathcal{B} , and so we turn to Lemma 8 and Corollary 2. Indeed, for any given m , Lemma 8 constructs \mathcal{A} satisfying hypothesis (a) such that

$$\alpha = 1/(2m(4m-1)) \tag{13.23}$$

satisfies the first part of (13.5). Next, if we take $\beta := \alpha - \varepsilon'$ and define $r := \lfloor \beta \log_2 p \rfloor$ and $M := 2^{1/\beta-1}$, then (13.2) constructs \mathcal{B} which, by Corollary 2, satisfies hypothesis (b) provided

$$(2\tau - 1)(\ell - \gamma) \geq 10\gamma, \quad (13.24)$$

where τ is the solution to

$$\left(\frac{1}{M}\right)^{2\tau} + \left(\frac{M-1}{M}\right)^\tau = 1.$$

For this construction, $|\mathcal{B}| = M^r \geq \Omega(p^{1-\beta})$, thereby satisfying the second part of (13.5).

It remains to maximize ε_1 for which there exist $m, \varepsilon', \ell, \gamma, \alpha_1, \alpha_2, \varepsilon, x, y$ satisfying (13.6)–(13.13), (13.23), and (13.24). Note that m and ε' determine α and τ , and the remaining constraints (13.6)–(13.13), (13.24) which define the feasible tuples $(\varepsilon_1, \ell, \gamma, \alpha_1, \alpha_2, \varepsilon, x, y)$ are linear inequalities. As such, taking the closure of this feasibility region and running a linear program will produce the supremum of ε_1 subject to the remaining constraints. This supremum increases monotonically as $\varepsilon' \rightarrow 0$, and so we only need to consider the limiting case where $\varepsilon' = 0$. Running the linear program for various values of m reveals what appears to be a largest supremum of $\varepsilon_1 \approx 8.8933 \times 10^{-24}$ at $m = 53,000,000$ (see Fig. 13.1). Dividing by 2 then gives

$$\varepsilon_0 \approx 4.4466 \times 10^{-24}.$$

While this optimization makes a substantial improvement (this is over 8,000 times larger than the original due to Bourgain et al. in [8]), the constant is still tiny!

For this particular construction of \mathcal{A} and \mathcal{B} , the remaining bottlenecks may lie at the very foundations of additive combinatorics. For example, it is currently known that the constant c_0 from Proposition 2 in [6] satisfies $1/10430 \leq c_0 < 1$. If $c_0 = 1/2$ (say), then taking $m = 10,000$ leads to ε_0 being on the order of 10^{-12} . Another source of improvement is Lemma 5 (Corollary 1 in [8]), which is proved using a version of the Balog–Szemerédi–Gowers lemma [5]. Specifically, the power of K in

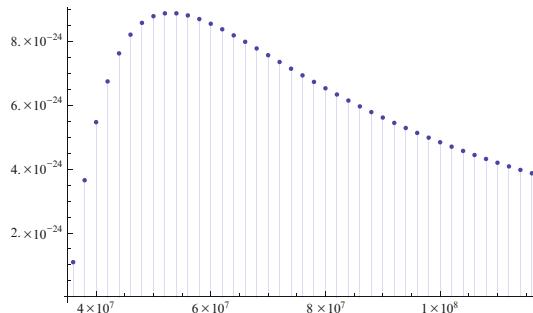


Fig. 13.1 The supremum of ε_1 as a function of m . Taking $\varepsilon' = 0$, we run a linear program to maximize ε_1 subject to the closure of the constraints (13.6)–(13.13), (13.24) for various values of m . A locally maximal supremum of $\varepsilon_1 \approx 8.8933 \times 10^{-24}$ appears around $m = 53,000,000$.

Lemma 5 is precisely the coefficient of x in (13.9), as well as the coefficient of γ (less 1) in the right-hand side of (13.24); as such, decreasing this exponent would in turn enlarge the feasibility region. An alternative construction for \mathcal{A} is proposed in [7], and it would be interesting to optimize this construction as well, though the bottlenecks involving c_0 and the Balog–Szemerédi–Gowers lemma are also present in this alternative.

13.5 Proofs of Technical Lemmas

This section contains the proofs of the technical lemmas (Lemmas 9 and 10) which were used to prove the main result (Theorem 2).

13.5.1 Proof of Lemma 9

First note that $|A_1|M_1 < p^{1/2+(4/3)x+\alpha-\alpha_1+\varepsilon}$ implies that

$$|A_1||\Omega_1(a_1)| < p^{1/2+(4/3)x+\alpha-\alpha_1+\varepsilon}$$

by (13.17), and by (13.18), we also have

$$|A_2||\Omega_2(a_2)| < 2p^{1/2}.$$

As such, the triangle inequality gives that

$$|S(A_1, A_2)| \leq |A_1||A_2||\Omega_1(a_1)||\Omega_2(a_2)| \leq 2p^{1+(4/3)x+\alpha-\alpha_1+\varepsilon} \leq p^{1-\varepsilon_1-\varepsilon},$$

where the last step uses (13.6). Thus, we can assume $|A_1|M_1 \geq p^{1/2+(4/3)x+\alpha-\alpha_1+\varepsilon}$, and so the assumption $|\mathcal{A}| \leq p^\alpha$ gives

$$M_1 \geq \frac{1}{|A_1|}p^{1/2+(4/3)x+\alpha-\alpha_1+\varepsilon} \geq p^{1/2+(4/3)x-\alpha_1+\varepsilon}. \quad (13.25)$$

Applying (13.17) and (13.25) then gives

$$|\Omega_1(a_1)| \geq \frac{M_1}{2} \geq \frac{1}{2}p^{1/2+(4/3)x-\alpha_1+\varepsilon} > p^{1/2+(4/3)x-\alpha_1+\varepsilon/2} \geq p^\ell,$$

where the last step uses (13.7). Note that we can redo all of the preceding analysis by interchanging indices 1 and 2. As such, we also have $|\Omega_2(a_2)| > p^\ell$. (This will enable us to use hypothesis (b) in Theorem 2.) At this point, we bound

$$\left| \sum_{\substack{b_1 \in \Omega_1(a_1) \\ b_2 \in \Omega_2(a_2)}} e_p \left(\frac{(b_1 - b_2)^2}{4(a_1 - a_2)} \right) \right|$$

using Lemma 7:

$$\leq |\Omega_1(a_1)|^{1/2} E(\Omega_1(a_1), \Omega_1(a_1))^{1/8} |\Omega_2(a_2)|^{1/2} E(\Omega_2(a_2), \Omega_2(a_2))^{1/8} p^{1/8}.$$

Next, since $|\Omega_1(a_1)|, |\Omega_2(a_2)| > p^\ell$, hypothesis (b) in Theorem 2 with $\varepsilon \leftarrow 4\varepsilon$ gives

$$\leq |\Omega_1(a_1)|^{7/8} |\Omega_2(a_2)|^{7/8} p^{1/8 - \gamma/4 + \varepsilon}.$$

At this point, the triangle inequality gives

$$\begin{aligned} |S(A_1, A_2)| &\leq \sum_{\substack{a_1 \in A_1 \\ a_2 \in A_2}} \left| \sum_{\substack{b_1 \in \Omega_1(a_1) \\ b_2 \in \Omega_2(a_2)}} e_p \left(\frac{(b_1 - b_2)^2}{4(a_1 - a_2)} \right) \right| \\ &\leq \sum_{\substack{a_1 \in A_1 \\ a_2 \in A_2}} |\Omega_1(a_1)|^{7/8} |\Omega_2(a_2)|^{7/8} p^{1/8 - \gamma/4 + \varepsilon} \end{aligned}$$

which can be further bounded using (13.17) and (13.18):

$$\leq 2^{7/4} |A_1|^{1/8} |A_2|^{1/8} p^{1 - \gamma/4 + \varepsilon}$$

Thus, if $|A_1|, |A_2| < p^y$, then

$$|S(A_1, A_2)| \leq 2^{7/4} p^{1+y/4 - \gamma/4 + \varepsilon} \leq p^{1 - \varepsilon_1 - \varepsilon},$$

where the last step uses (13.8). As such, we may assume that either $|A_1|$ or $|A_2|$ is $\geq p^y$. Without loss of generality, we assume $|A_2| \geq p^y$. (Considering (13.16), this will enable us to use L10.)

At this point, take $B_0 \subseteq \Omega_1(a_1)$ to be a maximal subset satisfying (13.21) for $B \leftarrow B_0$, and denote $B_1 := \Omega_1(a_1) \setminus B_0$. Then the triangle inequality gives

$$|T_{a_1}(A_2, B_1)| \leq \sum_{a_2 \in A_2} \left| \sum_{\substack{b_1 \in B_1 \\ b_2 \in \Omega_2(a_2)}} e_p \left(\frac{(b_1 - b_2)^2}{4(a_1 - a_2)} \right) \right|,$$

and then Lemma 7 gives

$$\leq \sum_{a_2 \in A_2} |B_1|^{1/2} E(B_1, B_1)^{1/8} |\Omega_2(a_2)|^{1/2} E(\Omega_2(a_2), \Omega_2(a_2))^{1/8} p^{1/8}.$$

This can be bounded further by applying $E(\Omega_2(a_2), \Omega_2(a_2)) \leq |\Omega_2(a_2)|^3$, (13.17), (13.18) and the assumption $|\mathcal{A}| \leq p^\alpha$:

$$\leq 2^{7/8} |B_1|^{1/2} E(B_1, B_1)^{1/8} p^{\alpha/8 + 9/16}. \quad (13.26)$$

At this point, we claim that $E(B_1, B_1) \leq p^{-x} M_1^3$. To see this, suppose otherwise. Then $|B_1|^3 \geq E(B_1, B_1) > p^{-x} M_1^3$, implying

$$|B_1| > p^{-x/3} M_1, \quad (13.27)$$

and by (13.17), we also have

$$E(B_1, B_1) > p^{-x} M_1^3 > p^{-x} |\Omega_1(a_1)|^3 \geq p^{-x} |B_1|^3.$$

Thus, Lemma 5 with $K = p^x$ produces a subset $B'_1 \subseteq B_1$ such that

$$|B'_1| \geq \frac{|B_1|}{20p^x} > \frac{M_1}{20p^{(4/3)x}} \geq \frac{1}{20} p^{1/2 - \alpha_1 + \epsilon} \geq p^{1/2 - \alpha_1}$$

where the second and third inequalities follow from (13.27) and (13.25), respectively, and

$$|B'_1 - B'_1| \leq 10^7 p^{9x} |B_1| \leq p^{9x + \epsilon} |B_1| \leq p^{\alpha_2} |B_1|,$$

where the last step follows from (13.9). As such, $|B'_1|$ satisfies (13.19) and (13.20), implying that $B \leftarrow B'_1$ satisfies (13.21) by L10. By the triangle inequality, $B \leftarrow B_0 \cup B'_1$ also satisfies (13.21), contradicting B_0 's maximality.

We conclude that $E(B_1, B_1) \leq p^{-x} M_1^3$, and continuing (13.26) gives

$$|T_{a_1}(A_2, B_1)| \leq 2^{7/8} |B_1|^{1/2} M_1^{3/8} p^{9/16 + \alpha/8 - x/8}.$$

Now we apply (13.21) to $B \leftarrow B_0$ and combine with this to get

$$\begin{aligned} |T_{a_1}(A_2, \Omega_1(a_1))| &\leq |T_{a_1}(A_2, B_0)| + |T_{a_1}(A_2, B_1)| \\ &\leq |B_0| p^{1/2 - \epsilon_1} + 2^{7/8} |B_1|^{1/2} M_1^{3/8} p^{9/16 + \alpha/8 - x/8}. \end{aligned}$$

Applying $|B_0|, |B_1| \leq |\Omega_1(a_1)| \leq M_1$ by (13.17) then gives

$$|T_{a_1}(A_2, \Omega_1(a_1))| \leq M_1 p^{1/2 - \epsilon_1} + 2^{7/8} M_1^{7/8} p^{9/16 + \alpha/8 - x/8}.$$

Now we apply the triangle inequality to get

$$\begin{aligned} |S(A_1, A_2)| &\leq \sum_{a_1 \in A_1} |T_{a_1}(A_2, \Omega_1(a_1))| \\ &\leq |A_1| \left(M_1 p^{1/2 - \epsilon_1} + 2^{7/8} M_1^{7/8} p^{9/16 + \alpha/8 - x/8} \right), \end{aligned}$$

and applying (13.18) and the assumption $|\mathcal{A}| \leq p^\alpha$ then gives

$$\leq 2p^{1-\varepsilon_2} + 2^{7/4}p^{1+\alpha/4-x/8} \leq 2p^{1-\varepsilon_2} + 2^{7/4}p^{1-\varepsilon_2} \leq p^{1-\varepsilon_1-\varepsilon},$$

where the last steps use (13.22). This completes the proof.

13.5.2 Proof of Lemma 10

We start by following the proof of Lemma 10 in [8]. First, Cauchy–Schwarz along with (13.17) and (13.18) gives

$$\begin{aligned} |T_{a_1}(A_2, B)|^2 &= \left| \sum_{\substack{a_2 \in A_2 \\ b_2 \in \Omega_2(a_2)}} \left(\frac{a_1 - a_2}{p} \right) \cdot \sum_{b_1 \in B} e_p \left(\frac{(b_1 - b_2)^2}{4(a_1 - a_2)} \right) \right|^2 \\ &\leq 2\sqrt{p} \sum_{\substack{a_2 \in A_2 \\ b_2 \in \Omega_2(a_2)}} \left| \sum_{b_1 \in B} e_p \left(\frac{(b_1 - b_2)^2}{4(a_1 - a_2)} \right) \right|^2. \end{aligned}$$

Expanding $|w|^2 = w\bar{w}$ and applying the triangle inequality then gives

$$= 2\sqrt{p} \left| \sum_{\substack{a_2 \in A_2 \\ b_2 \in \Omega_2(a_2)}} \sum_{b_1, b \in B} e_p \left(\frac{b_1^2 - b^2}{4(a_1 - a_2)} - \frac{b_2(b_1 - b)}{2(a_1 - a_2)} \right) \right| \leq 2\sqrt{p} \sum_{b_1, b \in B} |F(b, b_1)|,$$

where

$$F(b, b_1) := \sum_{\substack{a_2 \in A_2 \\ b_2 \in \Omega_2(a_2)}} e_p \left(\frac{b_1^2 - b^2}{4(a_1 - a_2)} - \frac{b_2(b_1 - b)}{2(a_1 - a_2)} \right).$$

Next, Hölder's inequality $\|F\|_1 \leq \|F\|_m \|1\|_{1-1/m}$ gives

$$|T_{a_1}(A_2, B)|^2 \leq 2\sqrt{p} |B|^{2-2/m} \left(\sum_{b_1, b \in B} |F(b, b_1)|^m \right)^{1/m}. \quad (13.28)$$

To bound this, we use a change of variables $x := b_1 + b \in B + B$ and $y := b_1 - b \in B - B$ and sum over more terms:

$$\sum_{b_1, b \in B} |F(b, b_1)|^m \leq \sum_{\substack{x \in B + B \\ y \in B - B}} \left| \sum_{\substack{a_2 \in A_2 \\ b_2 \in \Omega_2(a_2)}} e_p \left(\frac{xy}{4(a_1 - a_2)} - \frac{b_2 y}{2(a_1 - a_2)} \right) \right|^m.$$

Expanding $|w|^m = w^{m/2} \bar{w}^{m/2}$ and applying the triangle inequality then gives

$$\begin{aligned}
&= \left| \sum_{\substack{x \in B+B \\ y \in B-B}} \sum_{\substack{a_2^{(i)} \in A_2 \\ b_2^{(i)} \in \Omega_2(a_2^{(i)}) \\ 1 \leq i \leq m}} e_p \left(\sum_{i=1}^{m/2} \left[\frac{xy}{4(a_1 - a_2^{(i)})} - \frac{b_2 y}{2(a_1 - a_2^{(i)})} - \frac{xy}{4(a_1 - a_2^{(i+m/2)})} + \frac{b_2 y}{2(a_1 - a_2^{(i+m/2)})} \right] \right) \right| \\
&\leq \sum_{y \in B-B} \sum_{\substack{a_2^{(i)} \in A_2 \\ b_2^{(i)} \in \Omega_2(a_2^{(i)}) \\ 1 \leq i \leq m}} \left| \sum_{x \in B+B} e_p \left(\frac{xy}{4} \sum_{i=1}^{m/2} \left[\frac{1}{a_1 - a_2^{(i)}} - \frac{1}{a_1 - a_2^{(i+m/2)}} \right] \right) \right|.
\end{aligned}$$

Next, we apply (13.17) to bound the number of m -tuples of $b_2^{(i)}$'s for each m -tuple of $a_2^{(i)}$'s (there are less than M_2^m). Combining this with the bound above, we know there are complex numbers $\varepsilon_{y,\xi}$ of modulus ≤ 1 such that

$$\sum_{b_1, b \in B} |F(b, b_1)|^m \leq M_2^m \sum_{y \in B-B} \sum_{\xi \in \mathbb{F}_p} \lambda(\xi) \varepsilon_{y,\xi} \sum_{x \in B+B} e_p(xy\xi/4), \quad (13.29)$$

where

$$\lambda(\xi) := \left| \left\{ a^{(1)}, \dots, a^{(m)} \in A_2 : \sum_{i=1}^{m/2} \left[\frac{1}{a_1 - a^{(i)}} - \frac{1}{a_1 - a^{(i+m/2)}} \right] = \xi \right\} \right|.$$

To bound the $\xi = 0$ term in (13.29), pick $a^{(1)}, \dots, a^{(m)} \in A_2$ such that

$$\sum_{i=1}^{m/2} \left[\frac{1}{a_1 - a^{(i)}} - \frac{1}{a_1 - a^{(i+m/2)}} \right] = 0. \quad (13.30)$$

Then

$$\sum_{i=1}^{m/2} \frac{1}{a_1 - a^{(i)}} + \frac{m}{2} \cdot \frac{1}{a_1 - a^{(1)}} = \sum_{i=m/2+1}^m \frac{1}{a_1 - a^{(i)}} + \frac{m}{2} \cdot \frac{1}{a_1 - a^{(1)}},$$

and so by hypothesis (a) in Theorem 2, we have that $(a^{(1)}, \dots, a^{(m/2)}, a^{(1)}, \dots, a^{(1)})$ is a permutation of $(a^{(m/2+1)}, \dots, a^{(m)}, a^{(1)}, \dots, a^{(1)})$, which in turn implies that $(a^{(1)}, \dots, a^{(m/2)})$ and $(a^{(m/2+1)}, \dots, a^{(m)})$ are permutations of each other. Thus, all possible solutions to (13.30) are determined by $(a^{(1)}, \dots, a^{(m/2)})$. There are $|A_2|^{m/2}$ choices for this $m/2$ -tuple, and for each choice, there are $(m/2)!$ available permutations for $(a^{(m/2+1)}, \dots, a^{(m)})$. As such,

$$\lambda(0) = (m/2)! |A_2|^{m/2}, \quad (13.31)$$

which we will use later to bound the $\xi = 0$ term. In the meantime, we bound the remainder of (13.29). To this end, it is convenient to define the following functions:

$$\zeta'(z) := \sum_{\substack{y \in B - B \\ \xi \in \mathbb{F}_p^* \\ y\xi = z}} \varepsilon_{y,\xi} \lambda(\xi), \quad \zeta(z) := \sum_{\substack{y \in B - B \\ \xi \in \mathbb{F}_p^* \\ y\xi = z}} \lambda(\xi).$$

Note that $|\zeta'(z)| \leq \zeta(z)$ by the triangle inequality. We use the triangle inequality and Hölder's inequality to bound the $\xi \neq 0$ terms in (13.29):

$$\begin{aligned} & \left| \sum_{y \in B - B} \sum_{\xi \in \mathbb{F}_p^*} \lambda(\xi) \varepsilon_{y,\xi} \sum_{x \in B + B} e_p(xy\xi/4) \right| \\ &= \left| \sum_{\substack{x \in B + B \\ z \in \mathbb{F}_p}} \zeta'(z) e_p(xz/4) \right| \\ &\leq \sum_{x \in \mathbb{F}_p} \left| 1_{B+B}(x) \cdot \sum_{z \in \mathbb{F}_p} \zeta'(z) e_p(xz/4) \right| \\ &\leq |B+B|^{3/4} \left(\sum_{x \in \mathbb{F}_p} \left| \sum_{z \in \mathbb{F}_p} \zeta'(z) e_p(xz/4) \right|^4 \right)^{1/4}. \end{aligned} \quad (13.32)$$

To proceed, note that

$$\begin{aligned} \left(\sum_{z \in \mathbb{F}_p} \zeta'(z) e_p(xz/4) \right)^2 &= \sum_{z, z'' \in \mathbb{F}_p} \zeta'(z) \zeta'(z'') e_p(x(z+z''))/4) \\ &= \sum_{z' \in \mathbb{F}_p} (\zeta' * \zeta')(z') e_p(xz'/4), \end{aligned}$$

where the last step follows from a change of variables $z' = z + z''$. With this and Parseval's identity, we continue (13.32):

$$\begin{aligned} &= |B+B|^{3/4} \left(\sum_{x \in \mathbb{F}_p} \left| \sum_{z' \in \mathbb{F}_p} (\zeta' * \zeta')(z') e_p(xz'/4) \right|^2 \right)^{1/4} \\ &= |B+B|^{3/4} \|\zeta' * \zeta'\|_2^{1/2} p^{1/4} \\ &\leq |B+B|^{3/4} \|\zeta * \zeta\|_2^{1/2} p^{1/4}, \end{aligned} \quad (13.33)$$

where the last step follows from the fact that $|(\zeta' * \zeta')(z)| \leq (\zeta * \zeta)(z)$, which can be verified using the triangle inequality. Since $\zeta(z) = \sum_{\xi \in \mathbb{F}_p^*} 1_{B-B}(z/\xi) \lambda(\xi)$, the triangle inequality gives

$$\begin{aligned}
\|\zeta * \zeta\|_2 &= \left\| \left(\sum_{\xi \in \mathbb{F}_p^*} \lambda(\xi) 1_{\xi(B-B)} \right) * \left(\sum_{\xi' \in \mathbb{F}_p^*} \lambda(\xi') 1_{\xi'(B-B)} \right) \right\|_2 \\
&\leq \sum_{\xi, \xi' \in \mathbb{F}_p^*} \lambda(\xi) \lambda(\xi') \|1_{\xi(B-B)} * 1_{\xi'(B-B)}\|_2 \\
&= \sum_{\xi, \xi' \in \mathbb{F}_p^*} \lambda(\xi) \lambda(\xi') \|1_{B-B} * 1_{(\xi'/\xi)(B-B)}\|_2,
\end{aligned} \tag{13.34}$$

where the last step follows from the (easily derived) fact that $1_{B-B} * 1_{(\xi'/\xi)(B-B)}$ is a dilation of $1_{\xi(B-B)} * 1_{\xi'(B-B)}$.

To bound (13.34), we will appeal to Corollary 2 in [8], which says that for any $A \subseteq \mathbb{F}_p$ and probability measure λ over \mathbb{F}_p ,

$$\sum_{b \in \mathbb{F}_p^*} \lambda(b) \|1_A * 1_{bA}\|_2 \ll (\|\lambda\|_2 + |A|^{-1/2} + |A|^{1/2} p^{-1/2})^{c_0} |A|^{3/2}, \tag{13.35}$$

where \ll is Vinogradov notation; $f \ll g$ means $f = O(g)$. As such, we need to construct a probability measure and understand its 2-norm. To this end, define

$$\lambda_1(\xi) := \frac{\lambda(\xi)}{\|\lambda\|_1} = \frac{\lambda(\xi)}{|A_2|}. \tag{13.36}$$

The sum $\sum_{\xi \in \mathbb{F}_p} \lambda(\xi)^2$ is precisely the number of solutions to

$$\frac{1}{a_1 - a^{(1)}} + \cdots + \frac{1}{a_1 - a^{(m)}} - \frac{1}{a_1 - a^{(m+1)}} - \cdots - \frac{1}{a_1 - a^{(2m)}} = 0,$$

which by hypothesis (a) in Theorem 2, only has trivial solutions. As such, we have

$$\|\lambda\|_2^2 = m! |A_2|^m. \tag{13.37}$$

At this point, define $\lambda'_1(b)$ to be $\lambda_1(\xi'/b)$ whenever $b \neq 0$ and $\lambda_1(0)$ otherwise. Then λ'_1 is a probability measure with the same 2-norm as λ_1 , but it allows us to directly apply (13.35):

$$\begin{aligned}
&\sum_{\xi \in \mathbb{F}_p^*} \lambda_1(\xi) \|1_{B-B} * 1_{(\xi'/\xi)(B-B)}\|_2 \\
&= \sum_{b \in \mathbb{F}_p^*} \lambda'_1(b) \|1_{B-B} * 1_{b(B-B)}\|_2 \\
&\ll (\|\lambda_1\|_2 + |B-B|^{-1/2} + |B-B|^{1/2} p^{-1/2})^{c_0} |B-B|^{3/2}.
\end{aligned} \tag{13.38}$$

At this point, our proof deviates from the proof of Lemma 10 in [8]. By (13.36), (13.37), and (13.16), we have

$$\|\lambda_1\|_2 = |A_2|^{-m} \|\lambda\|_2 \leq \sqrt{m!} |A_2|^{-m/2} \leq \sqrt{m!} p^{-my/2},$$

Next, (13.19) and (13.20) together give

$$|B - B| \geq |B| \geq p^{1/2 - \alpha_1}$$

and

$$|B - B| \leq p^{\alpha_2} |B| \leq p^{1/2 + \alpha_2}.$$

Thus,

$$\begin{aligned} \|\lambda_1\|_2 + |B - B|^{-1/2} + |B - B|^{1/2} p^{-1/2} &\leq \sqrt{m!} p^{-my/2} + p^{\alpha_1/2 - 1/4} + p^{\alpha_2/2 - 1/4} \\ &\leq p^{-my/2 + 4m\epsilon}, \end{aligned} \quad (13.39)$$

where the last step follows from (13.12). So, by (13.34), (13.36), (13.38), and (13.39), we have

$$\begin{aligned} \|\zeta * \zeta\|_2 &\leq |A_2|^{2m} \sum_{\xi' \in \mathbb{F}_p^*} \lambda_1(\xi') \sum_{\xi \in \mathbb{F}_p^*} \lambda_1(\xi) \|1_{B-B} * 1_{(\xi'/\xi)(B-B)}\|_2 \\ &\ll |A_2|^{2m} (\|\lambda_1\|_2 + |B - B|^{-1/2} + |B - B|^{1/2} p^{-1/2})^{c_0} |B - B|^{3/2} \\ &\leq |A_2|^{2m} p^{-(c_0/2)my + 4c_0m\epsilon} |B - B|^{3/2}, \end{aligned} \quad (13.40)$$

and subsequent application of (13.29), (13.31), (13.33), and (13.40) gives

$$\begin{aligned} \sum_{b_1, b \in B} |F(b, b_1)|^m &\leq (\frac{m}{2})! (M_2 |A_2|)^m |A_2|^{-m/2} |B - B| |B + B| \\ &\quad + O(M_2^m |A_2|^m |B - B|^{3/4} |B + B|^{3/4} p^{-(c_0/4)my + 2c_0m\epsilon} p^{1/4}). \end{aligned} \quad (13.41)$$

By Lemma 4 in [8] (which states that $|A + A| \leq |A - A|^2 / |A|$), condition (13.20) implies

$$|B + B| \leq \frac{|B - B|^2}{|B|} \leq p^{2\alpha_2} |B|.$$

We now use this with (13.18), (13.16), and (13.20) to bound (13.41):

$$\ll (\frac{m}{2})! (2\sqrt{p})^m p^{-my/2} p^{3\alpha_2} |B|^2 + (2\sqrt{p})^m p^{(9/4)\alpha_2} |B|^{3/2} p^{-(c_0/4)my + 2c_0m\epsilon} p^{1/4}$$

Next, the left-hand inequality of (13.19) gives that $p^{1/4} \leq |B|^{1/2} p^{\alpha_1/2}$, leading to the following bound:

$$\ll |B|^2 p^{m/2 - my/2 + 3\alpha_2} + |B|^2 p^{m/2 + \alpha_1/2 + (9/4)\alpha_2 - (c_0/4)my + 2c_0m\epsilon}.$$

Overall, we have

$$\sum_{b_1, b \in B} |F(b, b_1)|^m \leq 2^{-m} |B|^2 p^{m/2 + \alpha_1/2 + (9/4)\alpha_2 - (c_0/4)my + 2m\epsilon},$$

since $c_0 < 1$, and $3\alpha_2 - 2\alpha_1 \leq (2 - c_0)my$ (i.e., (13.13)). Thus, (13.28) gives

$$\begin{aligned} |T(A_2, B)|^2 &\leq \sqrt{p} |B|^{2-2/m} (|B|^2 p^{m/2 + \alpha_1/2 + (9/4)\alpha_2 - (c_0/4)my + 2m\epsilon})^{1/m} \\ &= |B|^2 p^{1 - (c_0y/4 - \alpha_1/(2m) - (9\alpha_2)/(4m)) + 2\epsilon}. \end{aligned}$$

Finally, taking square roots produces the result.

Acknowledgements The author thanks the anonymous referees for their helpful suggestions. This work was supported by NSF Grant No. DMS-1321779. The views expressed in this chapter are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

References

1. Applebaum, L., Howard, S.D., Searle, S., Calderbank, R.: Chirp sensing codes: deterministic compressed sensing measurements for fast recovery. *Appl. Comput. Harmon. Anal.* **26**, 283–290 (2009)
2. Bandeira, A.S., Dobriban, E., Mixon, D.G., Sawin, W.F.: Certifying the restricted isometry property is hard. *IEEE Trans. Inf. Theory* **59**, 3448–3450 (2013)
3. Bandeira, A.S., Fickus, M., Mixon, D.G., Wong, P.: The road to deterministic matrices with the restricted isometry property. *J. Fourier Anal. Appl.* **19**, 1123–1149 (2013)
4. Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices. *Constr. Approx.* **28**, 253–263 (2008)
5. Bourgain, J., Garaev, M.Z.: On a variant of sum-product estimates and explicit exponential sum bounds in prime fields. *Math. Proc. Camb. Philos. Soc.* **146**, 1–21 (2009)
6. Bourgain, J., Glibichuk, A.: Exponential sum estimate over subgroup in an arbitrary finite field. http://www.math.ias.edu/files/avi/Bourgain_Glibichuk.pdf (2011)
7. Bourgain, J., Dilworth, S.J., Ford, K., Konyagin, S.V., Kutzarova, D.: Breaking the k^2 barrier for explicit RIP matrices. In: *STOC 2011*, pp. 637–644 (2011)
8. Bourgain, J., Dilworth, S.J., Ford, K., Konyagin, S., Kutzarova, D.: Explicit constructions of RIP matrices and related problems. *Duke Math. J.* **159**, 145–185 (2011)
9. Cai, T.T., Zhang, A.: Sharp RIP bound for sparse signal and low-rank matrix recovery. *Appl. Comput. Harmon. Anal.* **35**, 74–93 (2013)
10. Casazza, P.G., Fickus, M.: Fourier transforms of finite chirps. *EURASIP J. Appl. Signal Process.* **2006**, 7 p (2006)
11. DeVore, R.A.: Deterministic constructions of compressed sensing matrices. *J. Complexity* **23**, 918–925 (2007)
12. Fickus, M., Mixon, D.G., Tremain, J.C.: Steiner equiangular tight frames. *Linear Algebra Appl.* **436**, 1014–1027 (2012)
13. Foucart, S., Rauhut, H.: *A Mathematical Introduction to Compressive Sensing*. Springer, Berlin (2013)
14. Koiran, P., Zouzias, A.: Hidden cliques and the certification of the restricted isometry property. arXiv:1211.0665 (2012)

15. Mixon, D.G.: Deterministic RIP matrices: breaking the square-root bottleneck, short, fat matrices (weblog). <http://www.dustingmixon.wordpress.com/2013/12/02/deterministic-rip-matrices-breaking-the-square-root-bottleneck/> (2013)
16. Mixon, D.G.: Deterministic RIP matrices: breaking the square-root bottleneck, II, short, fat matrices (weblog). <http://www.dustingmixon.wordpress.com/2013/12/11/deterministic-rip-matrices-breaking-the-square-root-bottleneck-ii/> (2013)
17. Mixon, D.G.: Deterministic RIP matrices: breaking the square-root bottleneck, III, short, fat matrices (weblog). <http://www.dustingmixon.wordpress.com/2014/01/14/deterministic-rip-matrices-breaking-the-square-root-bottleneck-iii/> (2013)
18. Tao, T.: Open question: deterministic UUP matrices. What's new (weblog). <http://www.terrytao.wordpress.com/2007/07/02/open-question-deterministic-uup-matrices/> (2007)
19. Tao, T., Vu, V.H.: Additive Combinatorics. Cambridge University Press, Cambridge (2006)
20. Welch, L.R.: Lower bounds on the maximum cross correlation of signals. *IEEE Trans. Inform. Theory* **20**, 397–399 (1974)

Chapter 14

Tensor Completion in Hierarchical Tensor Representations

Holger Rauhut, Reinhold Schneider, and Željka Stojanac

Abstract Compressed sensing extends from the recovery of sparse vectors from undersampled measurements via efficient algorithms to the recovery of matrices of low rank from incomplete information. Here we consider a further extension to the reconstruction of tensors of low multi-linear rank in recently introduced hierarchical tensor formats from a small number of measurements. Hierarchical tensors are a flexible generalization of the well-known Tucker representation, which have the advantage that the number of degrees of freedom of a low rank tensor does not scale exponentially with the order of the tensor. While corresponding tensor decompositions can be computed efficiently via successive applications of (matrix) singular value decompositions, some important properties of the singular value decomposition do not extend from the matrix to the tensor case. This results in major computational and theoretical difficulties in designing and analyzing algorithms for low rank tensor recovery. For instance, a canonical analogue of the tensor nuclear norm is NP-hard to compute in general, which is in stark contrast to the matrix case. In this book chapter we consider versions of iterative hard thresholding schemes adapted to hierarchical tensor formats. A variant builds on methods from Riemannian optimization and uses a retraction mapping from the tangent space of the manifold of low rank tensors back to this manifold. We provide first partial convergence results based on a tensor version of the restricted isometry property (TRIP) of the measurement map. Moreover, an estimate of the number of measurements is provided that ensures the TRIP of a given tensor rank with high probability for Gaussian measurement maps.

H. Rauhut • Ž. Stojanac
RWTH Aachen University, Lehrstuhl C für Mathematik (Analysis),
Templergraben 55, 52062 Aachen, Germany
e-mail: rauhut@mathc.rwth-aachen.de; stojanac@mathc.rwth-aachen.de

R. Schneider (✉)
Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany
e-mail: schneidr@math.tu-berlin.de

14.1 Introduction

As outlined in the introductory chapter of this book, *compressed sensing* allows the recovery of (approximately) sparse vectors from a small number of linear random measurements via efficient algorithms including ℓ_1 -minimization and iterative hard thresholding, see also [15, 19] for introductory material. This theory was later extended to the reconstruction of low rank matrices from random measurements in [10, 11, 24, 52]. An important special case includes the *matrix completion problem*, where one seeks to fill in missing entries of a low rank matrix [9, 11, 24, 53]. Corresponding algorithms include *nuclear norm minimization* [11, 18, 52] and versions of *iterative hard thresholding* [8, 60].

In the present article we pursue a further extension of compressed sensing. We consider the recovery of a low rank tensor from a relatively small number of measurements. In contrast to already existing work in this direction [21, 35], we will understand low rank tensors in the framework of recently introduced *hierarchical tensor formats* [26]. This concept includes the classical *Tucker format* [14, 36, 63] as well as *tensor trains* [46, 47]. These hierarchical tensors can be represented in a data sparse way, i.e., they require only a very low number of data for their representation compared to the dimension of the full tensor space.

Let us recall the setup of low rank matrix recovery first. Given a matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ of rank at most $r \ll \min\{n_1, n_2\}$, the goal of low rank matrix recovery is to reconstruct \mathbf{X} from linear measurements $b_i = 1, \dots, m$, i.e., $\mathbf{b} = \mathcal{A}(\mathbf{X})$, where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}^m$ with $m \ll n_1 n_2$ is a linear sensing operator.

This problem setting can be transferred to the problem to recover higher order tensors $\mathbf{u} \in \mathcal{H}_d := \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d) \mapsto \mathbf{u}(\mu_1, \dots, \mu_d)$ from the linear measurements $\mathbf{b} = \mathcal{A}(\mathbf{u})$, where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d} \rightarrow \mathbb{R}^m$, is a sensing operator $m \ll n_1 n_2 \dots n_d$. Here d denotes the order of the tensor (number of modes of a tensor) and we remark that, for easier readability, we use the notation $\mathbf{u}(\mu_1, \dots, \mu_d)$ referring to the entries of the tensor. In the present article we assume that the tensors to be reconstructed belong to a class of hierarchical tensors of a given low multi-linear rank $\mathbf{r} = (r_j)_{j=1}^p$, where p depends on the specific tensor format [13], see also below. Of particular interest is the special case of *tensor completion*, where the measurement operator samples entries of the tensor, i.e.,

$$(\mathcal{A}\mathbf{u})_i = \mathbf{u}(\boldsymbol{\mu}_i) = \mathbf{u}(\mu_{1,i}, \dots, \mu_{d,i}) = b_i, \quad i = 1, \dots, m,$$

where the $\boldsymbol{\mu}_i \in \Omega$, $|\Omega| = m$, are given (multi-)indices [5, 37, 59]. Tensors, even of high order $d \gg 3$, appear frequently in data and signal analysis. For example, a video signal is a tensor of order $d = 3$. High order tensors of relatively low rank may also arise in vector-tensorization [25, 49] of a low dimensional signal. The present article tries to present a framework for tensor recovery from the perspective of recent developments in tensor product approximation [4, 26, 36], in particular the development of hierarchical tensors [26, 46]. The *canonical format* (CANDECOMP, PARAFAC) representing a tensor of order d as a sum of elementary tensor products, or rank one tensors (see [4, 36])

$$\begin{aligned}
\mathbf{u}(\mu_1, \dots, \mu_d) &= \sum_{k=1}^r (\mathbf{c}_k^1 \otimes \dots \otimes \mathbf{c}_k^d)(\mu_1, \dots, \mu_d) \\
&= \sum_{k=1}^r \mathbf{c}_k^1(\mu_1) \cdots \mathbf{c}_k^d(\mu_d) , \quad \mu_i = 1, \dots, n_i , \quad i = 1, \dots, d,
\end{aligned} \tag{14.1}$$

with $\mathbf{c}_k^i \in \mathbb{R}^{n_i}$, suffers from severe difficulties, unless $d \leq 2$. For example, the tensor rank is not well defined, and the set of tensors of the above form with fixed r is not closed [32] and does not form an algebraic variety. However, we obtain a closed subset, if we impose further conditions. Typical examples for such conditions are, e.g., symmetry [38] or bounds $|\mathbf{c}_k^i| \leq \alpha$ for some fixed α [64].

However, it has been experienced that computations within the *Tucker tensor format* behave relatively robust and stable, whereas the complexity unfortunately still suffers from the curse of dimensionality. A first important observation may be summarized in the fact that the set of Tucker tensors with a Tucker rank at most $\mathbf{r} = (r_1, \dots, r_d)$ forms an algebraic variety, i.e., a set of common zeros of multi-variate polynomials. Recently developed *hierarchical tensors*, introduced by Hackbusch and coworkers (HT tensors) [22, 28] and the group of Tyrtyshnikov (tensor trains, TT) [46, 47], have extended the Tucker format [14, 36, 63] into a multi-level framework, that no longer suffers from high order scaling w.r.t. the order d , as long as the ranks are moderate. For $d = 3$ there is no essential difference, whereas for larger $d \geq 4$ one benefits from the use of the novel formats. This makes the Tucker format [35, 59, 73] and, in particular, its hierarchical generalization [13, 37, 50], the *hierarchical tensor format*, a proper candidate for tensor product approximation in the sense that it serves as an appropriate model class in which we would like to represent or approximate tensors of interest in a data sparse way. Several algorithms developed in compressed sensing and matrix recovery or matrix completion can be easily transferred to this tensor setting (with the exception of nuclear norm minimization, which poses some fundamental difficulties). However, we already note at this point that the analysis of algorithms is much harder for tensors than for matrices as we will see below.

Historically, the hierarchical tensor framework has evolved in the quantum physics community hidden in the renormalization group ideas [71], and became clearly visible in the framework of matrix product and tensor network states [58]. An independent source of these developments can be found in quantum dynamics as the multi-layer multiconfigurational time-dependent Hartree (MCTDH) method [3, 43, 69]. Only after the recent introduction of hierarchical tensor representations in numerics, namely Hierarchical Tucker (HT) [26, 28] and Tensor Trains (TT) [46, 47], its relationship to already existing concepts in quantum physics has been realized [39]. We refer the interested reader to the recent survey articles [23, 27, 29, 39] and the monograph [26]. In the present paper we would like to provide a fairly self-contained introduction, and demonstrate how these concepts can be applied for tensor recovery.

There are several essential difficulties when passing from matrix to tensor recovery. In matrix recovery, the original problem can be reformulated to finding

the solution of the optimization problem

$$\text{minimize } \text{rank}(\mathbf{Z}) \text{ s.t. } \mathcal{A}(\mathbf{Z}) = \mathbf{b}, \mathbf{Z} \in \mathbb{R}^{n_1 \times n_2},$$

i.e., to finding the matrix \mathbf{Z} with the lowest rank consistent with the measurements. While this problem is NP-hard [18], it can be relaxed to the convex optimization problem of constrained nuclear norm minimization

$$\text{minimize } \|\mathbf{Z}\|_* \text{ s.t. } \mathcal{A}(\mathbf{Z}) = \mathbf{b}, \mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}.$$

Here, the nuclear norm is the sum of the singular values $\sigma_j(\mathbf{Z})$ of \mathbf{Z} , i.e., $\|\mathbf{Z}\|_* = \sum_j \sigma_j(\mathbf{Z})$. The minimizer of this problem reconstructs \mathbf{X} exactly under suitable conditions on \mathcal{A} [9, 10, 19, 24, 52].

In the hierarchical tensor setting, we are dealing with a rank tuple $\mathbf{r} = (r_1, \dots, r_p)$, which we would like to minimize simultaneously. However, this is not the only difficulty arising from the nonlinear tensor ansatz. In fact, the tensor nuclear norm is NP-hard to compute [20, 31] and therefore, tensor nuclear norm minimization is computationally prohibitive. Another difficulty arises because in contrast to the matrix case, also the best rank \mathbf{r} -approximation to a given tensor is NP-hard to compute [20, 31].

Our model class of hierarchical tensors of fixed multi-linear rank \mathbf{r} is a smooth embedded manifold, and its closure constitutes an algebraic variety. These are properties on which one can built local optimization methods [1, 44, 57], subsumed under the moniker *Riemannian optimization*. Moreover, for *hierarchical tensor representation* efficient numerical tools for finding at least a quasi-best approximation are available, namely the *higher order singular value decomposition (HOSVD)*, related to the Tucker model [14], or the *hierarchical singular value decomposition (HSVD)*, which is an extension of the HOSVD to hierarchical Tucker models [22, 26, 48, 68]. All these methods proceed via successive computations of the SVD of certain matricizations of the original tensor.

The HSVD (and the HOSVD as a special case) enables us to compute rank \mathbf{r} approximations to a given tensor via truncation of the decomposition. This allows to extend a particular class of greedy type algorithms, namely *iterative hard thresholding algorithms* to the present tensor setting. In a wider sense, this class of techniques includes also related Riemannian manifold techniques [13, 37] and alternating least squares methods [34]. First numerical tests show promising results [13, 37, 50, 51]. For a convergence analysis in the tensor case, and in applications, however, we have to struggle with more and harder difficulties than in the matrix case. The most fundamental of these consists in the fact that truncations of the HSVD only provide quasi-best low rank approximations. Although bounds of the approximation error are known [22], they are not good enough for our purposes, which is the main reason why we are only able to provide partial convergence results in this chapter. Another difficulty with iterative hard thresholding algorithms is that the rank, here a rank tuple $\mathbf{r} = (r_1, \dots, r_p)$, has to be fixed a priori. In practice this rank tuple is not known in advance, and a strategy for specifying appropriate ranks is required. Well-known strategies borrowed from matrix recovery consist in

increasing the rank during the approximation or starting with overestimating the rank and reduce the ranks through the iteration [72]. For our seminal treatment, we simply assume that the multi-linear rank \mathbf{r} is known in advance, i.e., the sought tensor \mathbf{u} is of exact rank \mathbf{r} . Moreover, we assume noiseless measurements $(\mathcal{A}\mathbf{u})_j = b_j \in \mathbb{R}$, $j = 1, \dots, m$. The important issues of adapting ranks and obtaining robustness will be deferred to future research [51].

Our chapter is related to the one on *two algorithms for compressed sensing of sparse tensors* by S. Friedland, Q. Li, D. Schonfeld, and E.E. Bernal. The latter chapter also considers the recovery of mode d -tensors from incomplete information using efficient algorithms. However, in contrast to our chapter, the authors assume usual sparsity of the tensor instead of the tensor being of low rank. The tensor structure is used in order to simplify the measurement process and to speed up the reconstruction rather than to work with the smallest possible number of measurements and to exploit low-rankness.

14.2 Hierarchical Tensors

14.2.1 Tensor product spaces

We start with some preliminaries. In the sequel, we consider only the real field $\mathbb{K} = \mathbb{R}$, but most parts are easy to extend to the complex case as well. We will confine ourselves to finite dimensional linear spaces $V_i = \mathbb{R}^{n_i}$ from which the tensor product space

$$\mathcal{H}_d = \bigotimes_{i=1}^d V_i := \bigotimes_{i=1}^d \mathbb{R}^{n_i},$$

is built [26]. If it is not stated explicitly, the $V_i = \mathbb{R}^{n_i}$ are supplied with the canonical basis $\{\mathbf{e}_1^i, \dots, \mathbf{e}_{n_i}^i\}$ of the vector space \mathbb{R}^{n_i} . Then any $\mathbf{u} \in \mathcal{H}_d$ can be represented as

$$\mathbf{u} = \sum_{\mu_1=1}^{n_1} \dots \sum_{\mu_d=1}^{n_d} \mathbf{u}(\mu_1, \dots, \mu_d) \mathbf{e}_{\mu_1}^1 \otimes \dots \otimes \mathbf{e}_{\mu_d}^d.$$

Using this basis, with a slight abuse of notation, we can identify $\mathbf{u} \in \mathcal{H}_d$ with its representation by a d -variate function, often called hyper matrix,

$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_d) \mapsto \mathbf{u}(\mu_1, \dots, \mu_d) \in \mathbb{R}, \quad \mu_i = 1, \dots, n_i, \quad i = 1, \dots, d,$$

depending on discrete variables, usually called indices $\mu_i = 1, \dots, n_i$, and $\boldsymbol{\mu}$ is called a multi-index. Of course, the actual representation $\mathbf{u}(\dots)$ of $\mathbf{u} \in \mathcal{H}_d$ depends on the chosen bases of V_1, \dots, V_d . With $n = \max\{n_i : i = 1, \dots, d\}$, the number of possibly nonzero entries in the representation of \mathbf{u} is $n_1 \cdots n_d = \mathcal{O}(n^d)$. This is often referred

to as the *curse of dimensions*. We equip the linear space \mathcal{H}_d with the ℓ_2 -norm $\|\mathbf{u}\| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$ and the inner product

$$\langle \mathbf{u}, \mathbf{v} \rangle := \sum_{\mu_1=1}^{n_1} \cdots \sum_{\mu_d=1}^{n_d} \mathbf{u}(\mu_1, \dots, \mu_d) \mathbf{v}(\mu_1, \dots, \mu_d).$$

We distinguish linear operators between vector spaces and their corresponding representation by matrices, which are written by capital bold letters \mathbf{U} . Throughout this chapter, all tensor contractions or various tensor–tensor products are either defined explicitly, by summation over corresponding indices, or by introducing corresponding matricizations of the tensors and performing matrix–matrix products.

14.2.2 Subspace approximation

The essence of the classical Tucker format is that, given a tensor \mathbf{u} and a rank-tuple $\mathbf{r} = (r_j)_{j=1}^d$, one is searching for optimal subspaces $U_i \subset \mathbb{R}^{n_i}$ such that

$$\min \|\mathbf{u} - \mathbf{v}\|, \text{ where } \mathbf{v} \in U_1 \otimes \cdots \otimes U_d,$$

is minimized over U_1, \dots, U_d with $\dim U_i = r_i$. Equivalently, we are looking for corresponding bases $\mathbf{b}_{k_i}^i$ of U_i , which can be written in the form

$$\mathbf{b}_{k_i}^i := \sum_{\mu_i=1}^{n_i} \mathbf{b}^i(\mu_i, k_i) \mathbf{e}_{\mu_i}^i, \quad k_i = 1, \dots, r_i < n_i, \quad (14.2)$$

where $\mathbf{b}^i(k_i, \mu_i) \in \mathbb{R}$, for each coordinate direction $i = 1, \dots, d$. With a slight abuse of notation we often identify the basis vector with its representation

$$\mathbf{b}_{k_i}^i \simeq (\mu_i \mapsto \mathbf{b}^i(\mu_i, k_i)), \quad \mu_i = 1, \dots, n_i, \quad k_i = 1, \dots, r_i,$$

i.e., a discrete function or an n_i -tuple. This concept of subspace approximation can be used either for an approximation \mathbf{u} of a single tensor, as well as for an ensemble of tensors \mathbf{u}_j , $j = 1, \dots, m$, in tensor product spaces. Given the bases $\mathbf{b}_{k_i}^i$, \mathbf{u}_j can be represented by

$$\mathbf{u}_j = \sum_{k_1=1}^{r_1} \cdots \sum_{k_d=1}^{r_d} \mathbf{c}(j, k_1, \dots, k_d) \mathbf{b}_{k_1}^1 \otimes \cdots \otimes \mathbf{b}_{k_d}^d \in \bigotimes_{i=1}^d U_i \subset \mathcal{H}_d = \bigotimes_{i=1}^d \mathbb{R}^{n_i}. \quad (14.3)$$

In case \mathbf{b}^i ’s form orthonormal bases, the core tensor $\mathbf{c} \in \mathbb{R}^m \otimes \bigotimes_{i=1}^d \mathbb{R}^{r_i}$ is given entry-wise by

$$\mathbf{c}(j, k_1, \dots, k_d) = \langle \mathbf{u}_j, \mathbf{b}_{k_1}^1 \otimes \cdots \otimes \mathbf{b}_{k_d}^d \rangle.$$

We call a representation of the form (14.3) with some $\mathbf{b}_{k_i}^i, \mathbf{c}$ a Tucker representation, and the Tucker representations the Tucker format. In this formal parametrization, the upper limit of the sums may be larger than the ranks and $\{\mathbf{b}_{k_i}^i\}_{k_i}$ may not be linearly independent. Noticing that a Tucker representation of a tensor is not uniquely defined, we are interested in some normal form.

Since the core tensor contains $r_1 \cdots r_d \sim r^d$, $r := \max\{r_i : i = 1, \dots, d\}$, possibly nonzero entries, this concept does not prevent the number of free parameters from scaling exponentially with the dimensions $\mathcal{O}(r^d)$. Setting $n := \max\{n_i : i = 1, \dots, d\}$, the overall complexity for storing the required data (including the basis vectors) is bounded by $\mathcal{O}(ndr + r^d)$. Since n_i is replaced by r_i , one obtains a dramatical compression $\frac{r_1}{n_1} \cdots \frac{r_d}{n_d} \sim \left(\frac{r}{n}\right)^d$. Without further sparsity of the core tensors the Tucker format is appropriate for low order tensors $d < 4$.

14.2.3 Hierarchical tensor representation

The *hierarchical Tucker format* (HT) in the form introduced by Hackbusch and Kühn in [28] extends the idea of subspace approximation to a hierarchical or multi-level framework. Let us proceed in a hierarchical way. We first consider $V_1 \otimes V_2 = \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2}$ or preferably the subspaces $U_1 \otimes U_2$ introduced in the previous section. For the approximation of $\mathbf{u} \in \mathcal{H}_d$ we only need a subspace $U_{\{1,2\}} \subset U_1 \otimes U_2$ with dimension $r_{\{1,2\}} < r_1 r_2$. Indeed, $V_{\{1,2\}}$ is defined through a new basis

$$V_{\{1,2\}} = \text{span} \{ \mathbf{b}_{k_{\{1,2\}}}^{\{1,2\}} : k_{\{1,2\}} = 1, \dots, r_{\{1,2\}} \},$$

with basis vectors given by

$$\mathbf{b}_{k_{\{1,2\}}}^{\{1,2\}} = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \mathbf{b}^{\{1,2\}}(k_1, k_2, k_{\{1,2\}}) \mathbf{b}_{k_1}^1 \otimes \mathbf{b}_{k_2}^2, \quad k_{\{1,2\}} = 1, \dots, r_{\{1,2\}}.$$

One may continue in several ways, e.g. by building a subspace $U_{\{1,2,3\}} \subset U_{\{1,2\}} \otimes U_3 \subset U_1 \otimes U_2 \otimes U_3 \subset V_1 \otimes V_2 \otimes V_3$, or $U_{\{1,2,3,4\}} \subset U_{\{1,2\}} \otimes U_{\{3,4\}}$, where $U_{\{3,4\}}$ is defined analogously to $U_{\{1,2\}}$ and so on.

For a systematic treatment, this approach can be cast into the framework of a partition tree, with leaves $\{1\}, \dots, \{d\}$, simply abbreviated here by $1, \dots, d$, and vertices $\alpha \subset D := \{1, \dots, d\}$, corresponding to the partition $\alpha = \alpha_1 \cup \alpha_2$, $\alpha_1 \cap \alpha_2 = \emptyset$. Without loss of generality, we can assume that $i < j$, for all $i \in \alpha_1$, $j \in \alpha_2$. We call α_1, α_2 the *sons* of the *father* α and D is called the *root* of the tree. In the example above we have $\alpha := \{1, 2, 3\} = \alpha_1 \cup \alpha_2 = \{1, 2\} \cup \{3\}$, where $\alpha_1 := \{1, 2\}$ and $\alpha_2 := \{3\}$.

In general, we do not need to restrict the number of sons, and define the *coordination number* by the number of sons +1 (for the father). Restricting to a binary tree so that each node contains two sons for non-leaf nodes (i.e., $\alpha \neq \{i\}$) is often the common choice, which we will also consider here. Let $\alpha_1, \alpha_2 \subset D$ be the two sons of $\alpha \subset D$, then $U_\alpha \subset U_{\alpha_1} \otimes U_{\alpha_2}$ is defined by a basis

$$\mathbf{b}_\ell^\alpha = \sum_{i=1}^{r_{\alpha_1}} \sum_{j=1}^{r_{\alpha_2}} \mathbf{b}^\alpha(i, j, \ell) \mathbf{b}_i^{\alpha_1} \otimes \mathbf{b}_j^{\alpha_2}. \quad (14.4)$$

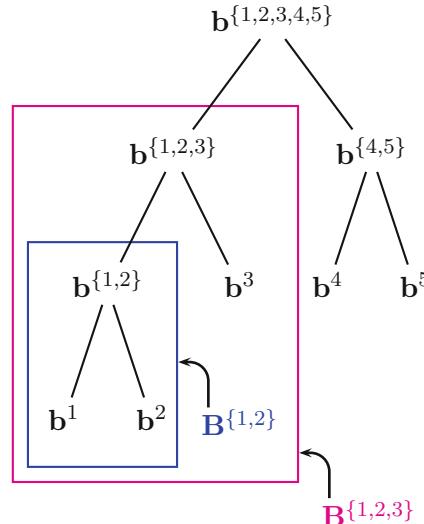
They can also be considered as matrices $(\mu_\alpha, \ell) \mapsto \mathbf{B}^\alpha(\mu_\alpha, \ell) \in \mathbb{R}^{n_\alpha \times r_\alpha}$ with $n_\alpha = \prod_{\ell \in \alpha} n_\ell$, for $\alpha \neq \{i\}$. Without loss of generality, all basis vectors, e.g. $\{\mathbf{b}_\ell^\alpha : \ell = 1, \dots, r_\alpha\}$, can be constructed to be orthonormal, as long as $\alpha \neq D$ is not the root. The tensors $(\ell, i, j) \mapsto \mathbf{b}^\alpha(i, j, \ell)$ will be called *transfer* or *component tensors*. For a leaf $\{i\} \simeq i$, the tensor $(\mu_i, k_i) \mapsto \mathbf{b}^i(\mu_i, k_i)$ in (14.2) denotes a *transfer* or *component tensor*. The component tensor $\mathbf{b}^D = \mathbf{b}^{\{1, \dots, d\}}$ at the root is called the *root tensor*.

Since the matrices \mathbf{B}^α are too large, we avoid computing them. We store only the *transfer* or *component tensors* which, for fixed $\ell = 1, \dots, r_\alpha$, can also be casted into *transfer matrices* $(i, j) \mapsto \mathbf{B}_\alpha(\ell, i, j) \in \mathbb{R}^{r_{\alpha_1} \times r_{\alpha_2}}$.

Proposition 1 ([26]). *A tensor $\mathbf{u} \in \mathcal{H}_d$ is completely parametrized by the transfer tensors \mathbf{b}^α , $\alpha \in \mathbb{T}$, i.e., by a multi-linear function τ*

$$(\mathbf{b}^\alpha)_{\alpha \in \mathbb{T}} \mapsto \mathbf{u} = \tau(\{\mathbf{b}^\alpha : \alpha \in \mathbb{T}\}).$$

Indeed τ is defined by applying (14.4) recursively. Since \mathbf{b}^α depends bi-linearly on \mathbf{b}^{α_1} and \mathbf{b}^{α_2} , the composite function τ is multi-linear in its arguments \mathbf{b}^α .



Hierarchical Tensor representation of an order 5 tensor

Data complexity: Let $n := \max\{n_i : i = 1, \dots, d\}$, $r := \max\{r_\alpha : \alpha \in \mathbb{T}\}$. Then the number of data required for the representation is $\mathcal{O}(n dr + dr^3)$, in particular does not scale exponentially w.r.t. the order d .

14.2.4 Tensor trains and matrix product representation

We now highlight another particular case of hierarchical tensor representations, namely *Tyrtynnikov tensors (TT)* or *tensor trains and matrix product representations* defined by taking $U_{\{1, \dots, p+1\}} \subset U_{\{1, \dots, p\}} \otimes V_{\{p+1\}}$, developed as TT tensors (tensor trains) by [47, 48] and known as matrix product states (MPS) in physics. Therein, we abbreviate $i \simeq \{1, \dots, i\}$ and consider the unbalanced tree $\mathbb{T} = \{\{1\}, \{2\}, \{1, 2\}, \{3\}, \{1, 2, 3\}, \dots, \{d\}, \{1, \dots, d\}\}$ and setting $r_0 = r_d = 1$. The transfer tensor \mathbf{b}^α for a leaf $\alpha \in \{\{2\}, \{3\}, \dots, \{d\}\}$ is usually defined as identity matrix of appropriate size and therefore the tensor $\mathbf{u} \in \mathcal{H}_d$ is completely parametrized by transfer tensors $(\mathbf{b}^\alpha)_{\alpha \in \mathbb{T}'}$, where $\mathbb{T}' = \{1, \dots, d\} = \{\{1\}, \{1, 2\}, \dots, \{1, 2, \dots, d\}\}$. Applying the recursive construction, the tensor \mathbf{u} can be written as

$$\begin{aligned} (\mu_1, \dots, \mu_d) &\mapsto \mathbf{u}(\mu_1, \dots, \mu_d) \\ &= \sum_{k_1=1}^{r_1} \dots \sum_{k_{d-1}=1}^{r_{d-1}} \mathbf{b}^1(\mu_1, k_1) \mathbf{b}^2(k_1, \mu_2, k_2) \dots \mathbf{b}^d(k_{d-1}, \mu_d). \end{aligned} \quad (14.5)$$

Introducing the matrices $\mathbf{B}_i(\mu_i) \in \mathbb{R}^{r_{i-1} \times r_i}$,

$$(\mathbf{B}_i(\mu_i))_{k_{i-1}, k_i} = \mathbf{b}^i(k_{i-1}, \mu_i, k_i), \quad 1 \leq i \leq d,$$

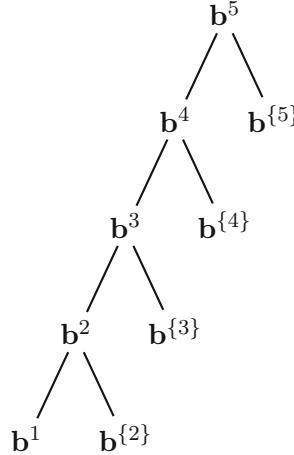
and with the convention $r_0 = r_d = 1$

$$(\mathbf{B}_1(\mu_1))_{k_1}^* = \mathbf{b}^1(\mu_1, k_1) \text{ and } (\mathbf{B}_d(\mu_d))_{k_{d-1}} = \mathbf{b}^d(k_{d-1}, \mu_d),$$

the formula (14.5) can be rewritten entry-wise by matrix–matrix products

$$\mathbf{u}(\mu_1, \dots, \mu_d) = \mathbf{B}_1(\mu_1) \cdots \mathbf{B}_i(\mu_i) \cdots \mathbf{B}_d(\mu_d) = \tau(\mathbf{b}^1, \dots, \mathbf{b}^d). \quad (14.6)$$

This representation is by no means unique. In general, there exist $\mathbf{b}^\alpha \neq \mathbf{c}^\alpha$ such that $\tau(\{\mathbf{b}^\alpha : \alpha \in \mathbb{T}\}) = \tau(\{\mathbf{c}^\alpha : \alpha \in \mathbb{T}\})$.



TT representation of an order 5 tensor with abbreviation $i \simeq \{1, \dots, i\}$

The tree is ordered according to the father–son relation into a hierarchy of levels, where \mathbf{b}^d is the root tensor. Let us observe that we can rearrange the hierarchy in such a way that any node $p = 1, \dots, d$ can form the root of the tree, i.e., \mathbf{b}^p becomes the root tensor. Using only orthogonal basis vectors, which is the preferred choice, this ordering reflects left- and right-hand orthogonalization in matrix product states [33].

A tensor in canonical form

$$\mathbf{u} = \sum_{k=1}^R \mathbf{u}_k^1 \otimes \cdots \otimes \mathbf{u}_k^d$$

can be easily written in the TT form, by setting $r_i = R$, for all $i = 1, \dots, d-1$ and

$$\mathbf{b}^i(k_{i-1}, \mu_i, k_i) = \begin{cases} \mathbf{u}_k^i(\mu_i) & \text{if } k_{i-1} = k_i = k, i = 2, \dots, d-1 \\ 0 & \text{if } k_{i-1} \neq k_i, i = 2, \dots, d-1 \\ \mathbf{u}_k^i(\mu_i) & \text{if } k_i = k, i = 1 \\ \mathbf{u}_k^i(\mu_i) & \text{if } k_{i-1} = k, i = d \end{cases}.$$

Data complexity: Let $n := \max\{n_i : i = 1, \dots, d\}$, $r := \max\{r_j : j = 1, \dots, d-1\}$. Then the number of data required for the presentation is $\mathcal{O}(dnr^2)$. Computing a single entry of a tensor requires the matrix multiplication of d matrices of size at most $r \times r$. This can be performed in $\mathcal{O}(ndr^3)$ operations.

Since the parametrization τ can be written in the simple matrix product form (14.6), we will consider the TT format often as a prototype model, and use it frequently for our explanations. We remark that most properties can easily be

extended to the general hierarchical case with straightforward modifications [26], and we leave those modifications to the interested reader.

14.2.5 Matricization of a tensor and its multi-linear rank

Let \mathbf{u} be a tensor in \mathcal{H}_d . Given a fixed dimension tree \mathbb{T} , for each node $\alpha \in \mathbb{T}$, $\alpha \neq D$, we can build a matrix \mathbf{U}^α from \mathbf{u} by grouping the indices μ_i , with $i \in \alpha$ into a row index I and the remaining indices μ_j with $j \in D \setminus \alpha$ into the column index J of the matrix $\mathbf{U}^\alpha = (\mathbf{U}_{I,J}^\alpha)$. For the root $\alpha = D$ we simply take the vectorized tensor $\mathbf{U}^D \in \mathbb{R}^{n_1 \cdots n_d \times 1}$. Since the rank of this matrix is one, it is often omitted.

For example, in the Tucker case, for $\alpha = \{i\}$ being a leaf, we set $I = \mu_i$ and $J = (\mu_1, \dots, \mu_{i-1}, \mu_{i+1}, \dots, \mu_d)$ providing a matrix

$$\mathbf{U}_{\mu_i;(\mu_1,\dots,\mu_{i-1},\mu_{i+1},\dots,\mu_d)}^\alpha = \mathbf{u}(\mu_1, \dots, \mu_d).$$

Similar, in the TT-format, with the convention that $i \simeq \{1, \dots, i\}$, we obtain matrices $\mathbf{U}^i \in \mathbb{R}^{n_1 \cdots n_i \times n_{i+1} \cdots n_d}$ with entries

$$\mathbf{U}_{(\mu_1,\dots,\mu_i);(\mu_{i+1},\dots,\mu_d)}^i = \mathbf{u}(\mu_1, \dots, \mu_d).$$

Definition 1. Given a dimension tree \mathbb{T} with p nodes, we define the *multi-linear rank* by the p -tuple $\mathbf{r} = (r_\alpha)_{\alpha \in \mathbb{T}}$ with $r_\alpha = \text{rank}(\mathbf{U}^\alpha)$, $\alpha \in \mathbb{T}$. The set of tensors $\mathbf{u} \in \mathcal{H}_d$ of given multi-linear rank \mathbf{r} will be denoted by $\mathcal{M}_\mathbf{r}$. The set of all tensors of rank \mathbf{s} at most \mathbf{r} , i.e., $s_\alpha \leq r_\alpha$ for all $\alpha \in \mathbb{T}$ will be denoted by $\mathcal{M}_{\leq \mathbf{r}}$.

Unlike the matrix case, it is possible that for some tuples \mathbf{r} , $\mathcal{M}_\mathbf{r} = \emptyset$ [12]. However, since our algorithm works on a closed nonempty set $\mathcal{M}_{\leq \mathbf{r}}$, this issue does not concern us.

In contrast to the canonical format (14.1), also known as CANDECOMP/PARAFAC, see [36, 41] and the border rank problem [38], in the present setting the rank is a well-defined quantity. This fact makes the present concept highly attractive for tensor recovery. On the other hand, if a tensor \mathbf{u} is of rank \mathbf{r} , then there exists a component tensor \mathbf{b}^α of the form (14.4) where $\ell = 1, \dots, r_\alpha$.

It is well known that the set of all matrices $\mathbf{A} \in \mathbb{R}^{n \times m}$ of rank at most r is a set of common zeros of multi-variate polynomials, i.e., an algebraic variety. The set $\mathcal{M}_{\leq \mathbf{r}}$ is the set of all tensors $\mathbf{u} \in \mathcal{H}_d$, where the matrices \mathbf{U}^α have a rank at most r_α . Therefore, it is again a set of common zeros of multivariate polynomials.

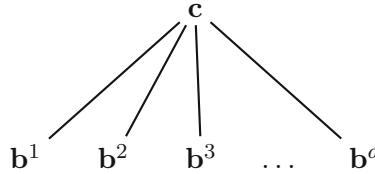
14.2.6 Higher order singular value decomposition

Let us provide more details about the rather classical higher order singular value decomposition. Above we have considered only binary dimension trees \mathbb{T} , but we can extend the considerations also to N -ary trees with $N \geq 3$. The d -ary tree \mathbb{T} (the tree with a root with d sons $i \simeq \{i\}$) induces the so-called Tucker decomposition and the corresponding higher order singular value decomposition (HOSVD). The Tucker decomposition was first introduced by Tucker in 1963 [61] and has been refined later on in many works, see, e.g., [40, 61, 62].

Definition 2 (Tucker decomposition). Given a tensor $\mathbf{u} \in \mathbb{R}^{n_1 \times \dots \times n_d}$, the decomposition

$$\mathbf{u}(\mu_1, \dots, \mu_d) = \sum_{k_1=1}^{r_1} \dots \sum_{k_d=1}^{r_d} \mathbf{c}(k_1, \dots, k_d) \mathbf{b}_{k_1}^1(\mu_1) \dots \mathbf{b}_{k_d}^d(\mu_d),$$

$r_i \leq n_i$, $i = 1, \dots, d$, is called a Tucker decomposition. The tensor $\mathbf{c} \in \mathbb{R}^{r_1 \times \dots \times r_d}$ is called the core tensor and the $\mathbf{b}_{k_i}^i \in \mathbb{R}^{n_i}$, for $i = 1, \dots, d$, form a basis of the subspace $U_i \subset \mathbb{R}^{n_i}$. They can also be considered as transfer or component tensors $(\mu_i, k_i) \mapsto \mathbf{b}^i(\mu_i, k_i) \in \mathbb{R}^{n_i \times r_i}$.



Tucker representation of an order d tensor

Notice that the Tucker decomposition is highly non-unique. For an $i \in \{1, \dots, d\}$ and invertible matrix $\mathbf{Q}_i \in \mathbb{R}^{r_i \times r_i}$, one can define a matrix $\bar{\mathbf{B}}^i = \mathbf{B}^i \mathbf{Q}_i$ and the tensor $\bar{\mathbf{c}}_i$

$$\bar{\mathbf{c}}_i(k_1, \dots, k_d) = \sum_{\bar{k}_i=1}^{r_i} \mathbf{c}_i(k_1, \dots, \bar{k}_i, \dots, k_d) \mathbf{Q}_i^{-1}(\bar{k}_i, k_i)$$

such that the tensor \mathbf{u} can also be written as

$$\mathbf{u}(\mu_1, \dots, \mu_d) = \sum_{k_1=1}^{r_1} \dots \sum_{k_d=1}^{r_d} \bar{\mathbf{c}}_i(k_1, \dots, k_d) \mathbf{b}_{k_1}^1(\mu_1) \dots \bar{\mathbf{b}}_{k_i}^i(\mu_i) \dots \mathbf{b}_{k_d}^d(\mu_d).$$

Similarly to the matrix case and the singular value decomposition, one can impose orthogonality conditions on the matrices \mathbf{B}^i , for all $i = 1, \dots, d$, i.e., we assume that

$\{\mathbf{b}_{k_i}^i : k_i = 1, \dots, r_i\}$ are orthonormal bases. However, in this case one does not obtain a super-diagonal core tensor \mathbf{c} .

Definition 3 (HOSVD decomposition). The HOSVD decomposition of a given tensor $\mathbf{u} \in \mathcal{H}_d$ is a special case of the Tucker decomposition where

- the bases $\{\mathbf{b}_{k_i}^i \in \mathbb{R}^{n_i} : k_i = 1, \dots, r_i\}$ are orthogonal and normalized, for all $i = 1, \dots, d$,
- the tensor $\mathbf{c} \in \mathcal{H}_d$ is all orthogonal, i.e., $\langle \mathbf{c}_{k_i=p}, \mathbf{c}_{k_i=q} \rangle = 0$, for all $i = 1, \dots, d$ and whenever $p \neq q$,
- the subtensors of the core tensor \mathbf{c} are ordered according to their ℓ_2 norm, i.e., $\|\mathbf{c}_{k_i=1}\| \geq \|\mathbf{c}_{k_i=2}\| \geq \dots \geq \|\mathbf{c}_{k_i=n_i}\| \geq 0$.

Here, the subtensor $\mathbf{c}_{k_i=p} \in \mathbb{R}^{n_1 \times \dots \times n_{i-1} \times n_{i+1} \times \dots \times n_d}$ is a tensor of order $d-1$ defined as

$$\mathbf{c}_{k_i=p}(\mu_1, \dots, \mu_{i-1}, \mu_{i+1}, \dots, \mu_d) = \mathbf{c}(\mu_1, \dots, \mu_{i-1}, p, \mu_{i+1}, \dots, \mu_d).$$

The HOSVD can be computed via successive SVDs of appropriate unfoldings or matricizations $\mathbf{U}^{\{i\}} = \mathbf{U}^i$, see, e.g., [36] and below for the more general HSVD. For more information on this decomposition, we refer the interested reader to [14, 26].

14.2.7 Hierarchical singular value decomposition and truncation

The singular value decomposition of the matricization \mathbf{U}^α , $\alpha \in \mathbb{T}$, is factorizing the tensor into two parts. Thereby, we separate the tree into two subtrees. Each part can be treated independently in an analogous way as before by applying a singular value decomposition. This procedure can be continued in a way such that one ends up with an explicit description of the component tensors. There are several sequential orders one can proceed including top-down and bottom-up strategies. We will call every decomposition of the above type a *higher order singular value decomposition (HOSVD)* or in the hierarchical setting a hierarchical singular value decomposition (HSVD). As long as no approximation, i.e., no truncation, has been applied during the corresponding SVDs, one obtains an exact recovery of the original tensor at the end. The situation changes if we apply truncations (via thresholding). Then the result may depend on the way and the order we proceed as well as on the variant of the HSVD.

In order to become more explicit let us demonstrate an HSVD procedure for the model example of a TT-tensor [46], already introduced in [68] for the matrix product representation. Without truncations this algorithm provides an exact reconstruction with a TT representation provided the multi-linear rank $\mathbf{s} = (s_1, \dots, s_{d-1})$ is chosen large enough. In general, the s_i 's can be chosen to be larger than the dimensions n_i . Via inspecting the ranks of the relevant matricizations, the multi-linear rank \mathbf{s} may be determined a priori.

1. Given: $\mathbf{u} \in \mathcal{H}_d$ of multi-linear rank $\mathbf{s} = (s_1, \dots, s_{d-1})$, $s_0 = s_d := 1$.
2. Set $\mathbf{v}^1 = \mathbf{u}$.
3. **For** $i = 1, \dots, d-1$ **do**

- Form matricization $\mathbf{V}^i \in \mathbb{R}^{s_{i-1}n_i \times n_{i+1} \cdots n_d}$ via

$$\mathbf{V}_{(k_{i-1}, \mu_i); (\mu_{i+1}, \dots, \mu_d)}^i = \mathbf{v}^i(k_{i-1}, \mu_i, \mu_{i+1}, \dots, \mu_d).$$

- Compute the SVD of \mathbf{V}^i :

$$\mathbf{V}_{(k_{i-1}, \mu_i); (\mu_{i+1}, \dots, \mu_d)}^i = \sum_{k_i=1}^{s_i} \sigma_{k_i}^i \mathbf{b}^i(k_{i-1}, \mu_i, k_i) \mathbf{d}^{i+1}(k_i, \mu_{i+1}, \dots, \mu_d),$$

where the $\mathbf{b}^i(\cdot, \cdot, k_i)$ and the $\mathbf{d}^{i+1}(k_i, \dots)$ are orthonormal (i.e., the left and right singular vectors) and the $\sigma_{k_i}^i$ are the nonzero singular values of \mathbf{V}^i .

- Set $\mathbf{v}^{i+1}(k_i, \mu_{i+1}, \dots, \mu_d) = \sigma_{k_i}^i \mathbf{d}^{i+1}(k_i, \mu_{i+1}, \dots, \mu_d)$
- 4. Set $\mathbf{b}^d(k_{d-1}, \mu_d) := \mathbf{v}^d(k_{d-1}, \mu_d)$ and $\mathbf{B}_i(\mu_i)_{k_{i-1}, k_i} = \mathbf{b}^i(k_{i-1}, \mu_i, k_i)$ for $i = 1, \dots, d$.
- 5. Decomposition $\mathbf{u}(\boldsymbol{\mu}) = \mathbf{B}_1(\mu_1) \cdots \mathbf{B}_d(\mu_d)$.

Above, the indices k_i run from 1 to n_i and, for notational consistency, $k_0 = k_d = 1$. Let us notice that the present algorithm is not the only way to use multiple singular value decompositions in order to obtain a hierarchical representation of \mathbf{u} for the given tree, here a TT representation. For example, one may start at the right end separating \mathbf{b}^d first and so on. The procedure above provides some *normal form* of the tensor.

Let us now explain hard thresholding on the example of a TT tensor and the HSVD defined above. This procedure remains essentially the same with the only difference that we apply a thresholding to a target rank $\mathbf{r} = (r_i)_{i=1}^{d-1}$ with $r_i \leq s_i$ at each step by setting $\sigma_{k_i}^i = 0$ for all $k_i > r_i$, $i = 1, \dots, d-1$, where the $(\sigma_{k_i}^i)_{k_i}$ is the monotonically decreasing sequence of singular values of \mathbf{V}^i . This leads to an approximate right factor \mathbf{v}_ϵ^i , within a controlled ℓ_2 error $\epsilon_i = \sqrt{\sum_{k_i > r_i} (\sigma_{k_i}^i)^2}$. By the *hard thresholding* HSVD procedure presented above, one obtains a unique approximate tensor

$$\mathbf{u}_\epsilon := \mathbf{H}_\mathbf{r}(\mathbf{u}) \tag{14.7}$$

of multi-linear rank \mathbf{r} within a guaranteed error bound

$$\|\mathbf{u}_\epsilon - \mathbf{u}\| \leq \sum_{i=1}^{d-1} \epsilon_i.$$

In contrast to the matrix case, this approximation \mathbf{u}_ε , however, may not be the best rank \mathbf{r} approximation of \mathbf{u} , which is in fact NP-hard to compute [20, 31]. A more evolved analysis shows the following quasi-optimal error bound.

Theorem 1. *Let $\mathbf{u}_\varepsilon = \mathbf{H}_\mathbf{r}(\mathbf{u})$. Then there exists $C(d) = \mathcal{O}(\sqrt{d})$, such that \mathbf{u}_ε satisfies the quasi-optimal error bound*

$$\inf\{\|\mathbf{u} - \mathbf{v}\| : \mathbf{v} \in \mathcal{M}_{\leq \mathbf{r}}\} \leq \|\mathbf{H}_\mathbf{r}(\mathbf{u})\| \leq C(d) \inf\{\|\mathbf{u} - \mathbf{v}\| : \mathbf{v} \in \mathcal{M}_{\leq \mathbf{r}}\}. \quad (14.8)$$

The constant satisfies $C(d) = \sqrt{d}$ for the Tucker format [22], $C(d) = \sqrt{d-1}$ for the TT format [47] and $C(d) = \sqrt{2d-3}$ for a balanced tree in the HSVD of [22].

The procedure introduced above can be modified to apply for general hierarchical tensor representations. When we consider the HSVD in the sequel, we have in mind that we have fixed our hierarchical SVD method choosing one of the several variants.

14.2.8 Hierarchical tensors as differentiable manifolds

It has been shown that, fixing a tree \mathbb{T} , the set of hierarchical tensors of exactly multi-linear rank \mathbf{r} forms an analytical manifold [17, 33, 44, 65]. We will describe its essential features using the TT format or the matrix product representation. For an invertible $r_1 \times r_1$ matrix \mathbf{G}_1 , it holds

$$\begin{aligned} \mathbf{u}(\boldsymbol{\mu}) &= \mathbf{B}^1(\mu_1)\mathbf{B}^2(\mu_2) \cdots \mathbf{B}^i(\mu_i) \cdots \mathbf{B}^d(\mu_d) \\ &= \mathbf{B}^1(\mu_1)\mathbf{G}_1\mathbf{G}_1^{-1}\mathbf{B}^2(\mu_2) \cdots \mathbf{B}^i(\mu_i) \cdots \mathbf{B}^d(\mu_d) \\ &= \tilde{\mathbf{B}}^1(\mu_1)\tilde{\mathbf{B}}^2(\mu_2) \cdots \mathbf{B}^i(\mu_i) \cdots \mathbf{B}^d(\mu_d), \end{aligned}$$

where $\tilde{\mathbf{B}}^1(\mu_1) = \mathbf{B}^1(\mu_1)\mathbf{G}_1$ and $\tilde{\mathbf{B}}^2(\mu_2) = \mathbf{G}_1^{-1}\mathbf{B}^2(\mu_2)$. This provides two different representations of the same tensor \mathbf{u} . In order to remove the redundancy in the above parametrization of the set $\mathcal{M}_\mathbf{r}$, let us consider the linear space of parameters $(\mathbf{b}^1, \dots, \mathbf{b}^d) \in \mathcal{X} := \times_{i=1}^d X_i$, $X_i := \mathbb{R}^{r_{i-1}r_i n_i}$, or equivalently $\mathcal{U} := (\mathbf{B}^1(\cdot), \dots, \mathbf{B}^d(\cdot))$, together with a Lie group action. For a collection of invertible matrices $\mathcal{G} = (\mathbf{G}_1, \dots, \mathbf{G}_{d-1})$ we define a transitive group action by

$$\mathcal{G} \circ \mathcal{U} := (\mathbf{B}^1\mathbf{G}_1, \mathbf{G}_1^{-1}\mathbf{B}^2\mathbf{G}_2, \dots, \mathbf{G}_{d-1}^{-1}\mathbf{B}^d).$$

One observes that the tensor \mathbf{u} remains unchanged under this transformation of the component tensors. Therefore, we will identify two representations $\mathcal{U}_1 \sim \mathcal{U}_2$, if there exists \mathcal{G} such that $\mathcal{U}_2 = \mathcal{G} \circ \mathcal{U}_1$. It is easy to see that the equivalence classes $[\mathcal{U}] := \{\mathcal{V} : \mathcal{U} \sim \mathcal{V}\}$ define smooth manifolds in \mathcal{X} . We are interested in the quotient manifold \mathcal{X} / \sim , which is isomorphic to $\mathcal{M}_\mathbf{r}$. This construction gives rise

to an embedded analytic manifold [1, 2, 30, 44, 65] where a Riemannian metric is canonically defined.

The *tangent space* $\mathcal{T}_{\mathbf{u}}$ at $\mathbf{u} \in \mathcal{M}_{\mathbf{r}}$ is of importance for calculations. It can be easily determined by means of the product rule as follows. A generic tensor $\delta \mathbf{u} \in \mathcal{T}_{\mathbf{u}}$ of a TT tensor \mathbf{u} is of the form

$$\begin{aligned}\delta \mathbf{u}(\mu_1, \dots, \mu_d) &= \mathbf{t}^1(\mu_1, \dots, \mu_d) + \dots + \mathbf{t}^d(\mu_1, \dots, \mu_d) \\ &= \delta \mathbf{B}^1(\mu_1) \mathbf{B}^2(\mu_2) \cdots \mathbf{B}^d(\mu_d) + \dots \\ &+ \mathbf{B}^1(\mu_1) \cdots \mathbf{B}^{i-1}(\mu_{i-1}) \delta \mathbf{B}^i(\mu_i) \mathbf{B}^{i+1}(\mu_{i+1}) \cdots \mathbf{B}^d(\mu_d) + \dots \\ &+ \mathbf{B}^1(\mu_1) \cdots \mathbf{B}^{d-1}(\mu_{d-1}) \delta \mathbf{B}^d(\mu_d).\end{aligned}$$

This tensor is uniquely determined if we impose *gauging conditions* onto $\delta \mathbf{B}^i$, $i = 1, \dots, d-1$ [33, 65]. There is no gauging condition imposed onto $\delta \mathbf{B}^d$. Typically these conditions are of the form

$$\sum_{k_{i-1}=1}^{r_{i-1}} \sum_{\mu_i=1}^{n_i} \mathbf{b}^i(k_{i-1}, \mu_i, k_i) \delta \mathbf{b}^i(k_{i-1}, \mu_i, k'_i) = 0, \quad \forall k_i, k'_i = 1, \dots, r_i. \quad (14.9)$$

With this condition at hand an orthogonal projection $P_{\mathcal{T}_{\mathbf{u}}}$ onto the tangent space $\mathcal{T}_{\mathbf{u}}$ is well defined and computable in a straightforward way.

The manifold $\mathcal{M}_{\mathbf{r}}$ is open and its closure is $\mathcal{M}_{\leq \mathbf{r}}$, the set of all tensors with ranks at most r_{α} , $\alpha \in \mathbb{T}$,

$$\text{clos}(\mathcal{M}_{\mathbf{r}}) = \mathcal{M}_{\leq \mathbf{r}}.$$

This important result is based on the observation that the matrix rank is an upper semi-continuous function [16]. The singular points are exactly those for which at least one rank $\tilde{r}_{\alpha} < r_{\alpha}$ is not maximal, see, e.g., [57]. We remark that, for the root D of the partition tree \mathbb{T} , there is no gauging condition imposed onto $\delta \mathbf{B}^D$. We highlight the following facts without explicit proofs for hierarchical tensors.

Proposition 2. *Let $\mathbf{u} \in \mathcal{M}_{\mathbf{r}}$. Then*

- (a) *the corresponding gauging conditions (14.9) imply that the tangential vectors \mathbf{t}^i are pairwise orthogonal;*
- (b) *the tensor \mathbf{u} is included in its own tangent space $\mathcal{T}_{\mathbf{u}}$;*
- (c) *the multi-linear rank of a tangent vector is at most $2\mathbf{r}$, i.e., $\delta \mathbf{u} \in \mathcal{M}_{\leq 2\mathbf{r}}$.*

Curvature estimates are given in [44].

14.3 Tensor completion for hierarchical tensors

14.3.1 The low rank tensor recovery problem

We pursue on extending methods for solving optimization problems in the calculus of hierarchical tensors to *low rank tensor recovery* and to *tensor completion* as a special case. The latter builds on ideas from the theory of *compressed sensing* [19], which predicts that sparse vectors can be recovered efficiently from incomplete linear measurements via efficient algorithms. Given a linear measurement mapping $\mathcal{A} : \mathcal{H}_d = \bigotimes_{i=1}^d \mathbb{R}^{n_i} \rightarrow \mathbb{R}^m$ our aim is to recover a tensor $\mathbf{u} \in \mathcal{H}_d$ from $m \ll N := n_1 \cdot n_2 \cdots n_d$ measurements $\mathbf{b} \in \mathbb{R}^m$ given by

$$\mathbf{b} = (b_i)_{i=1}^m = \mathcal{A}\mathbf{u}.$$

Since this problem is underdetermined we additionally assume that \mathbf{u} is of low rank, i.e., given a dimension tree \mathbb{T} and a multi-linear rank \mathbf{r} , we suppose that the tensor is contained in the corresponding tensor manifold, $\mathbf{u} \in \mathcal{M}_{\mathbf{r}}$.

The *tensor completion* problem—generalizing the matrix completion problem [9, 11, 24, 66, 70]—is the special case where the measurement map subsamples entries of the tensor, i.e., $b_i = (\mathcal{A}\mathbf{u})_i = \mathbf{u}(\boldsymbol{\mu}_i) = \mathbf{u}(\mu_{1,i}, \dots, \mu_{d,i})$, $i = 1, \dots, m$, with the multi-indices $\boldsymbol{\mu}_i$ being contained in a suitable index set $\Omega \subset [n_1] \times \cdots \times [n_d]$ of cardinality $m \ll n_1 \cdots n_d$.

We remark that in practice the desired rank \mathbf{r} may not be known in advance and/or the tensor \mathbf{u} is only close to $\mathcal{M}_{\mathbf{r}}$ rather than being exactly contained in $\mathcal{M}_{\mathbf{r}}$. Moreover, the left-hand side \mathbf{b} may not be known exactly because of noise on the measurements. In the present paper we defer from tackling these important stability and robustness issues and focus on the problem in the above form.

The problem of reconstructing $\mathbf{u} \in \mathcal{H}_d$ from $\mathbf{b} = \mathcal{A}\mathbf{u}$ can be reformulated to finding the minimizer of

$$\mathcal{J}(\mathbf{v}) = \frac{1}{2} \|\mathcal{A}\mathbf{v} - \mathbf{b}\|^2 \quad \text{subject to } \mathbf{v} \in \mathcal{M}_{\mathbf{r}}. \quad (14.10)$$

In other words, we are looking for a tensor of multi-linear rank \mathbf{r} , which fits best the given measurements. A minimizer over $\mathcal{M}_{\leq \mathbf{r}}$ always exists, but a solution of the above problem may not exist in general since $\mathcal{M}_{\mathbf{r}}$ is not closed. However, assuming $\mathbf{u} \in \mathcal{M}_{\mathbf{r}}$ and $\mathbf{b} = \mathcal{A}\mathbf{u}$ as above, existence of a minimizer is trivial because setting $\mathbf{v} = \mathbf{u}$ gives $\mathcal{J}(\mathbf{v}) = 0$. We note that finding a minimizer of (14.10) is NP-hard in general [20, 31].

The necessary first order condition for a minimizer of the problem (14.10) can be formulated as follows, see, e.g., [44]. If $\mathbf{u} \in \mathcal{M}_{\mathbf{r}}$ is a solution of $\operatorname{argmin}_{\mathbf{v} \in \mathcal{M}_{\mathbf{r}}} \mathcal{J}(\mathbf{v})$, then

$$\langle \nabla \mathcal{J}(\mathbf{u}), \delta \mathbf{u} \rangle = 0, \quad \text{for all } \delta \mathbf{u} \in \mathcal{T}_{\mathbf{u}},$$

where $\nabla \mathcal{J}$ is the gradient of \mathcal{J} .

14.3.2 Optimization approaches

In analogy to compressed sensing and low rank matrix recovery where convex relaxations (ℓ_1 -minimization and nuclear norm minimization) are very successful, a first idea for a tractable alternative to (14.10) may be to find an analogue of the nuclear norm for the tensor case. A natural approach is to consider the set Q of unit norm rank one tensors in \mathcal{H}_d ,

$$Q = \{\mathbf{u} = \mathbf{b}_1 \otimes \mathbf{b}_2 \cdots \otimes \mathbf{b}_d : \|\mathbf{u}\| = 1\}.$$

Its closed convex hull $B = \overline{\text{conv } Q}$ is taken as the unit ball of the tensor nuclear norm, so that the tensor nuclear norm is the gauge function of B ,

$$\|\mathbf{u}\|_* = \inf\{t : \mathbf{u} \in tB\}.$$

In fact, for the matrix case $d = 2$ we obtain the standard nuclear norm. Unfortunately, for $d \geq 3$, the nuclear tensor norm is NP-hard to compute [20, 31].

The contributions [21, 35, 42] proceed differently by considering the matrix nuclear norm of several unfoldings of the tensor \mathbf{u} . Given a dimension tree $\alpha \in \mathbb{T}$ and corresponding matricizations \mathbf{U}^α of \mathbf{u} , we consider the Schatten norm

$$\|\mathbf{U}^\alpha\|_p := \left(\sum_{k_\alpha} (\sigma_{k_\alpha}^\alpha)^p \right)^{\frac{1}{p}}, \quad \alpha \in \mathbb{T}, \quad 1 \leq p < \infty,$$

where the $\sigma_{k_\alpha}^\alpha$ are the singular values of \mathbf{U}^α , see, e.g., [6, 56]. Furthermore, for $1 \leq q \leq \infty$ and given $a_\alpha > 0$, e.g. $a_\alpha = 1$, a norm on \mathcal{H}_d can be introduced by

$$\|\mathbf{u}\|_{p,q}^q := \sum_{\alpha \in \mathbb{T}} a_\alpha \|\mathbf{U}^\alpha\|_p^q.$$

A prototypical choice of a convex optimization formulation used for tensor recovery consists in finding

$$\operatorname{argmin}\{\mathcal{J}(\mathbf{u}) := \|\mathbf{u}\|_{1,q} : \mathcal{A}\mathbf{u} = \mathbf{b}\}.$$

For $a_\alpha = 1$ and $q = 1$ this functional was suggested in [21, 42] for reconstructing tensors in the Tucker format. Although the numerical results are reasonable, it seems that conceptually this is not the “right” approach and too simple for the present purpose, see corresponding negative results in [45, 55]. For the Tucker case first rigorous results have been shown in [35]. However the approach followed there is based on results from matrix completion and does not use the full potential of tensor decompositions. In fact, their bound on the number of required measurements is $\mathcal{O}(rn^{d-1})$. A more “balanced” version of this approach is considered in [45], where the number of required measurements scales like $\mathcal{O}(r^{d/2}n^{d/2})$, which is better but still far from the expected linear scaling in n , see also Theorem 2 below.

14.3.3 Iterative hard thresholding schemes

Rather than following the convex optimization approach which leads to certain difficulties as outlined above, we consider versions of the iterative hard thresholding algorithm well-known from compressed sensing [7, 19] and low rank matrix recovery [60]. Iterative hard thresholding algorithms fall into the larger class of *projected gradient methods*. Typically one performs a gradient step in the ambient space \mathcal{H}_d , followed by a mapping \mathcal{R} onto the set of low rank tensors \mathcal{M}_r or $\mathcal{M}_{\leq r}$, formally

$$\begin{aligned}\mathbf{y}^{n+1} &:= \mathbf{u}^n - \alpha_n \nabla \mathcal{J}(\mathbf{u}^n) \quad (\text{gradient step}) \\ &= \mathbf{u}^n - \alpha_n (\mathcal{A}^* (\mathcal{A} \mathbf{u}^n - \mathbf{b})) , \\ \mathbf{u}^{n+1} &:= \mathcal{R}(\mathbf{y}^{n+1}) \quad (\text{projection step}).\end{aligned}$$

Apart from specifying the steplength α_n , the above algorithm depends on the choice of the projection operator $\mathcal{R} : \mathcal{H}_d \rightarrow \mathcal{M}_{\leq r}$. An example would be

$$\mathcal{R}(\mathbf{y}^{n+1}) := \operatorname{argmin}\{\|\mathbf{y}^{n+1} - \mathbf{z}\| : \mathbf{z} \in \mathcal{M}_{\leq r}\}.$$

Since this projection is not computable in general [20, 31], we may rather choose the hierarchical singular value (HSVD) thresholding procedure (14.7)

$$\mathbf{u}^{n+1} := \mathcal{R}(\mathbf{y}^{n+1}) = \mathbf{H}_r(\mathbf{y}^{n+1}) \quad (\text{hard thresholding}),$$

which is only quasi-optimal (14.8). We will call this procedure *tensor iterative hard thresholding (TIHT)*, or shortly *iterative hard thresholding (IHT)*.

Another possibility for the projection operator relies on the concept of retraction from differential geometry [1]. A retraction maps $\mathbf{u} + \xi$, where $\mathbf{u} \in \mathcal{M}_r$ and $\xi \in \mathcal{T}_{\mathbf{u}}$, smoothly to the manifold. For $R : (\mathbf{u}, \xi) \rightarrow R(\mathbf{u}, \xi) \in \mathcal{M}_r$ being a retraction it is required that R is twice differentiable and $R(\cdot, \mathbf{0}) = \mathbf{I}$ is the identity. Moreover, a retraction satisfies, for $\|\xi\|$ sufficiently small,

$$\begin{aligned}\|\mathbf{u} + \xi - R(\mathbf{u}, \xi)\| &= \mathcal{O}(\|\xi\|^2), \\ \|\mathbf{u} - R(\mathbf{u}, \xi)\| &= \mathcal{O}(\|\xi\|).\end{aligned}\tag{14.11}$$

Several examples of retractions for hierarchical tensors are known [37, 44], which can be efficiently computed. If a retraction is available, then a nonlinear projection \mathcal{R} can be realized in two steps. First we project (linearly) onto the tangent space $\mathcal{T}_{\mathbf{u}^n}$ at \mathbf{u}^n , and afterwards we apply a *retraction* R . This leads to the so-called *Riemannian gradient iteration method* (RGI) defined formally as

$$\begin{aligned}\mathbf{z}^{n+1} &:= P_{\mathcal{T}_{\mathbf{u}^n}} (\mathbf{u}^n - \alpha_n P_{\mathcal{T}_{\mathbf{u}^n}} (\mathcal{A}^* (\mathcal{A} \mathbf{u}^n - \mathbf{b}))) \quad (\text{projected gradient step}) \\ &= P_{\mathcal{T}_{\mathbf{u}^n}} (\mathbf{u}^n - \alpha_n \mathcal{A}^* (\mathcal{A} \mathbf{u}^n - \mathbf{b})) =: \mathbf{u}^n + \xi^n \\ \mathbf{u}_{n+1} &:= R(\mathbf{u}^n, \mathbf{z}^{n+1} - \mathbf{u}^n) = R(\mathbf{u}^n, \xi^n) \quad (\text{retraction step}).\end{aligned}$$

With a slight abuse of notation we will write

$$\mathcal{R}(\mathbf{y}^{n+1}) = R \circ P_{\mathcal{M}_{\mathbf{u}^n}} \mathbf{y}^{n+1}$$

for the RGI.

It may happen that an iterate \mathbf{u}^n is of lower rank, i.e., $\mathbf{u}^n \in \mathcal{M}_{\mathbf{s}}$ with $s_{\alpha} < r_{\alpha}$ at least for one $\alpha \in \mathbb{T}$. In this case $\mathbf{u}^n \in \mathcal{M}_{\leq \mathbf{r}}$ is a singular point and no longer on our manifold, i.e., $\mathbf{u}^n \notin \mathcal{M}_{\mathbf{r}}$, and our RGI algorithm fails. However, since $\mathcal{M}_{\mathbf{r}}$ is dense in $\mathcal{M}_{\leq \mathbf{r}}$, for arbitrary $\varepsilon > 0$, there exists $\mathbf{u}_{\varepsilon}^n \in \mathcal{M}_{\mathbf{r}}$, with $\|\mathbf{u}^n - \mathbf{u}_{\varepsilon}^n\| < \varepsilon$. Practically such a regularized $\mathbf{u}_{\varepsilon}^n$ is not hard to choose. Alternatively, the algorithm described above may be regularized in a sense that it automatically avoids the situation being trapped in a singular point [37]. Here, we do not go into these technical details.

14.3.4 Restricted isometry property for hierarchical tensors

A crucial sufficient condition for exact recovery in compressed sensing is the *restricted isometry property (RIP)*, see Chapter 1. It has been applied in the analysis of iterative hard thresholding both in the compressed sensing setting [7, 19] and in the low rank matrix recovery setting [60]. The RIP can be easily generalized to the present tensor setting. As common, $\|\cdot\|$ denotes the Euclidean norm below.

Definition 4. Let $\mathcal{A} : \mathcal{H}_d = \bigotimes_{i=1}^d \mathbb{R}^{n_i} \rightarrow \mathbb{R}^m$ be a linear measurement map, \mathbb{T} be a dimension tree and for $\mathbf{r} = (r_{\alpha})_{\alpha \in \mathbb{T}}$, let $\mathcal{M}_{\mathbf{r}}$ be the associated low rank tensor manifold. The tensor restricted isometry constant (TRIC) $\delta_{\mathbf{r}}$ of \mathcal{A} is the smallest number such that

$$(1 - \delta_{\mathbf{r}}) \|\mathbf{u}\|^2 \leq \|\mathcal{A}\mathbf{u}\|^2 \leq (1 + \delta_{\mathbf{r}}) \|\mathbf{u}\|^2, \quad \text{for all } \mathbf{u} \in \mathcal{M}_{\mathbf{r}}. \quad (14.12)$$

Informally, we say that a measurement map \mathcal{A} satisfies the *tensor restricted isometry property (TRIP)* if $\delta_{\mathbf{r}}$ is small (at least $\delta_{\mathbf{r}} < 1$) for some ‘‘reasonably large’’ \mathbf{r} .

Observing that $\mathbf{u} + \mathbf{v} \in \mathcal{M}_{\leq 2\mathbf{r}}$ for two tensors $\mathbf{u}, \mathbf{v} \in \mathcal{M}_{\leq \mathbf{r}}$ the TRIP (14.12) of order $2\mathbf{r}$ implies that \mathcal{J} has a unique minimizer on $\mathcal{M}_{\mathbf{r}}$. Indeed, for two tensors $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{M}_{\mathbf{r}}$ satisfying $\mathcal{A}\mathbf{u}_1 - \mathbf{b} = \mathcal{A}\mathbf{u}_2 - \mathbf{b} = \mathbf{0}$, it follows that $\mathcal{A}(\mathbf{u}_1 - \mathbf{u}_2) = \mathbf{0}$ in contradiction to the TRIP and $\mathbf{u}_1 - \mathbf{u}_2 \in \mathcal{M}_{\leq 2\mathbf{r}}$.

For Gaussian (or more generally subgaussian) measurement maps, the TRIP holds with high probability for both the HOSVD and the TT format under a suitable bound on the number of measurements [50, 51], which basically scales like the number of degrees of freedom of a tensor of multi-linear rank \mathbf{r} (up to a logarithmic factor in d). In order to state these results, let us introduce Gaussian measurement maps. A measurement map $\mathcal{A} : \mathcal{H}_d \rightarrow \mathbb{R}^m$ can be identified with a tensor in $\mathbb{R}^m \otimes \bigotimes_{i=1}^d \mathbb{R}^{n_i}$ via

$$(\mathcal{A}\mathbf{u})_\ell = \sum_{\mu_1=1}^{n_1} \sum_{\mu_2=1}^{n_2} \cdots \sum_{\mu_d=1}^{n_d} \mathbf{a}(\ell, \mu_1, \mu_2, \dots, \mu_d) \mathbf{u}(\mu_1, \dots, \mu_d), \quad \ell = 1, \dots, m.$$

If all entries of \mathcal{A} are independent realizations of normal distributed random variables with mean zero and variance $1/m$, then \mathcal{A} is called a *Gaussian measurement map*.

Theorem 2 ([50, 51]). *For $\delta, \varepsilon \in (0, 1)$, a random draw of a Gaussian measurement map $\mathcal{A} : \mathcal{H}_d = \bigotimes_{i=1}^d \mathbb{R}^{n_i} \rightarrow \mathbb{R}^m$ satisfies $\delta_r \leq \delta$ with probability at least $1 - \varepsilon$ provided*

- *HOSVD format: $m \geq C\delta^{-2} \max \{ (r^d + dnr) \log(d), \log(\varepsilon^{-1}) \}$,*
- *TT format: $m \geq C\delta^{-2} \max \{ (dnr^2) \log(dr), \log(\varepsilon^{-1}) \}$,*

where $n = \max \{n_i : i = 1, \dots, d\}$ and $r = \max \{r_i : i = 1, \dots, d\}$ and $C > 0$ is a universal constant.

The above result extends to subgaussian and in particular to Bernoulli measurement maps, see, e.g., [19, 67] for the definition of subgaussian random variables and matrices. Presently, it is not clear whether the logarithmic factor in d above is necessary or whether it is an artifact of the proof. We conjecture that similar bounds hold also for the general hierarchical tensor format. We note that in practice, the application of Gaussian sensing operators acting on the tensor product space \mathcal{H}_d seems to be computationally too expensive except for relatively small dimensions d , (e.g., $d = 2, 3, 4$), and small n_i . A more realistic measurement map for which TRIP bounds can be shown [51] is the decomposition of random sign flips of the tensor entries, a d -dimensional Fourier transform and random subsampling. All these operations can be performed quickly (exploiting) the FFT. For further details we refer to [50, 51].

We finally remark that the TRIP does not hold in the tensor completion setup because sparse and low rank tensor may belong to the kernel of the measurement map. In this scenario, additional incoherence properties on the tensor to be recovered like in the matrix completion scenario [9, 24, 53] are probably necessary.

14.3.5 Convergence results

Unfortunately, a full convergence (and recovery) analysis of the TIHT and RGI algorithms under the TRIP is not yet available. Nevertheless, we present two partial results. The first concerns the local convergence of the RGI and the second is a convergence analysis of the TIHT under an additional assumption on the iterates.

We assume that $\mathbf{u} \in \mathcal{M}_r$, where the low rank tensor manifold is associated with a fixed hierarchical tensor format. Measurements are given by

$$\mathbf{b} = \mathcal{A}\mathbf{u},$$

where \mathcal{A} is assumed to satisfy the TRIP of order $3\mathbf{r}$ below. Recall that our projected gradient scheme starts with an initial guess \mathbf{u}^0 and forms the iterates

$$\mathbf{y}^{n+1} := \mathbf{u}^n + \mathcal{A}^*(\mathbf{b} - \mathcal{A}\mathbf{u}^n), \quad (14.13)$$

$$\mathbf{u}^{n+1} := \mathcal{R}(\mathbf{y}^{n+1}), \quad (14.14)$$

where either $\mathcal{R}(\mathbf{u}^{n+1}) := \mathbf{H}_\mathbf{r}(\mathbf{y}^{n+1})$ (TIHT) or $\mathcal{R}(\mathbf{u}^{n+1}) := R \circ P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{y}^{n+1}$ (RGI).

We first show local convergence, in the sense that the iterates \mathbf{u}^n converge to the original tensor \mathbf{u} if the initial guess is sufficiently close to the solution $\mathbf{u} \in \mathcal{M}_\mathbf{r}$. Of course, this analysis also applies if one of the later iterates comes close enough to \mathbf{u} .

Theorem 3 (Local convergence). *Let $\mathbf{b} = \mathcal{A}\mathbf{u}$ for $\mathbf{u} \in \mathcal{M}_{\leq \mathbf{r}}$ and let \mathbf{u}^n be the iterates (14.13), (14.14) of the Riemannian gradient iterations, i.e., $\mathcal{R}(\mathbf{u}^{n+1}) := R \circ P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{y}^{n+1}$, where R is a retraction. In addition, let's assume that \mathcal{A} satisfies the TRIP of order $3\mathbf{r}$, i.e., $\delta_{3\mathbf{r}} \leq \delta < 1$. Suppose that*

$$\|\mathbf{u} - \mathbf{u}^0\| \leq \varepsilon$$

is sufficiently small and the distance to the singular points $\varepsilon < \text{dist}(\mathbf{u}^0, \partial \mathcal{M}_\mathbf{r})$ is sufficiently large. Then, there exists $0 < \rho < 1$ (depending on δ and ε) such that the series $\mathbf{u}^n \in \mathcal{M}_{\leq \mathbf{r}}$ converges linearly to $\mathbf{u} \in \mathcal{M}_{\leq \mathbf{r}}$ with rate ρ ,

$$\|\mathbf{u}^{n+1} - \mathbf{u}\| \leq \rho \|\mathbf{u}^n - \mathbf{u}\|.$$

Proof. We consider the orthogonal projection $P_{\mathcal{T}_{\mathbf{u}^n}}$ onto the tangent space $\mathcal{T}_{\mathbf{u}^n}$. There exists $1 < \gamma = \gamma(\varepsilon)$ and $\kappa > 0$ depending on the curvature of $\mathcal{M}_\mathbf{r}$, such that, for all $\|\mathbf{v} - \mathbf{u}^n\| < \varepsilon$, it holds that [44]

$$\gamma^{-1} \|\mathbf{u}^n - \mathbf{v}\| \leq \|P_{\mathcal{T}_{\mathbf{u}^n}}(\mathbf{u}^n - \mathbf{v})\| \leq \gamma \|\mathbf{u}^n - \mathbf{v}\| \quad (14.15)$$

$$\|(I - P_{\mathcal{T}_{\mathbf{u}^n}})(\mathbf{u}^n - \mathbf{v})\| \leq \kappa \|P_{\mathcal{T}_{\mathbf{u}^n}}(\mathbf{u} - \mathbf{v})\|^2. \quad (14.16)$$

Using the triangle inequality we estimate

$$\begin{aligned} \|\mathbf{u}^{n+1} - \mathbf{u}\| &= \|R(\mathbf{u}^n, P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{y}^{n+1} - \mathbf{u}^n) - \mathbf{u}\| \\ &\leq \|R(\mathbf{u}^n, P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{y}^{n+1} - \mathbf{u}^n) - P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{y}^{n+1}\| + \|P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{y}^{n+1} - P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{u}\| + \|P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{u} - \mathbf{u}\|. \end{aligned} \quad (14.17)$$

We will bound each of the three terms in (14.17) separately. We start with the first term, where we exploit the property (14.11) of retractions. Moreover, $W^n := \mathcal{T}_{\mathbf{u}^n} \subset \mathcal{M}_{\leq 2\mathbf{r}}$ by Proposition 2(c) and $P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{y}^{n+1} = P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{u}^n + P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{u}^n \mathcal{A}^* \mathcal{A}(\mathbf{u} - \mathbf{u}^n) = \mathbf{u}^n + P_{\mathcal{T}_{\mathbf{u}^n}} \mathcal{A}^* \mathcal{A}(\mathbf{u} - \mathbf{u}^n)$ because $P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{u}^n = \mathbf{u}^n$ by Proposition 2(a). Since $\mathbf{u} - \mathbf{u}^n \in \mathcal{M}_{\leq 2\mathbf{r}}$ we may apply the TRIP of order $2\mathbf{r} < 3\mathbf{r}$ to obtain

$$\begin{aligned} & \|R(\mathbf{u}^n, P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{y}^{n+1} - \mathbf{u}^n) - P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{y}^{n+1}\| \leq C \|P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{y}^{n+1} - \mathbf{u}^n\|^2 \\ & = C \|P_{\mathcal{T}_{\mathbf{u}^n}} \mathcal{A}^* \mathcal{A}(\mathbf{u} - \mathbf{u}^n)\|^2 \leq C (1 + \delta_{3r})^2 \varepsilon \|\mathbf{u} - \mathbf{u}^n\|, \end{aligned} \quad (14.18)$$

where in the last estimate we also used that $\|\mathbf{u} - \mathbf{u}^n\| \leq \varepsilon$.

For the second term in (14.17) observe that $(I - P_{\mathcal{T}_{\mathbf{u}^n}})(\mathbf{u} - \mathbf{u}^n) = \mathbf{u} - P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{u} \in \mathcal{M}_{\leq 3r}$ by Proposition 2. The TRIP implies therefore that the spectrum of $P_{\mathcal{T}_{\mathbf{u}^n}}(I - \mathcal{A}^* \mathcal{A})|_{W^n}$ is contained in the interval $[-\delta_{3r}, \delta_{3r}]$. With these observations and (14.16) we obtain

$$\begin{aligned} \|P_{\mathcal{T}_{\mathbf{u}^n}}(\mathbf{y}^{n+1} - \mathbf{u})\| &= \|P_{\mathcal{T}_{\mathbf{u}^n}}((\mathbf{u} - \mathbf{u}^n) - \mathcal{A}^*(\mathbf{b} - \mathcal{A}\mathbf{u}^n))\| \\ &\leq \|P_{\mathcal{T}_{\mathbf{u}^n}}((\mathbf{u} - \mathbf{u}^n) - \mathcal{A}^* \mathcal{A}(P_{\mathcal{T}_{\mathbf{u}^n}}(\mathbf{u} - \mathbf{u}^n)))\| \\ &\quad + \|P_{\mathcal{T}_{\mathbf{u}^n}} \mathcal{A}^* \mathcal{A}((I - P_{\mathcal{T}_{\mathbf{u}^n}})(\mathbf{u} - \mathbf{u}^n))\| \\ &\leq \|P_{\mathcal{T}_{\mathbf{u}^n}}((\mathbf{u} - \mathbf{u}^n) - \mathcal{A}^*(\mathcal{A}P_{\mathcal{T}_{\mathbf{u}^n}}(\mathbf{u} - \mathbf{u}^n)))\| \\ &\quad + (1 + \delta_{3r})\|(I - P_{\mathcal{T}_{\mathbf{u}^n}})(\mathbf{u} - \mathbf{u}^n)\| \\ &\leq \|P_{\mathcal{T}_{\mathbf{u}^n}}(I - \mathcal{A}^* \mathcal{A}P_{\mathcal{T}_{\mathbf{u}^n}})(\mathbf{u} - \mathbf{u}^n)\| \\ &\quad + (1 + \delta_{3r})\kappa \|P_{\mathcal{T}_{\mathbf{u}^n}}(\mathbf{u} - \mathbf{u}^n)\|^2 \\ &\leq \delta_{3r} \|P_{\mathcal{T}_{\mathbf{u}^n}}(\mathbf{u} - \mathbf{u}^n)\| + (1 + \delta_{3r})\kappa \|P_{\mathcal{T}_{\mathbf{u}^n}}(\mathbf{u} - \mathbf{u}^n)\|^2. \end{aligned}$$

Hence, for ε sufficiently small, there exists a factor $0 < \tilde{\rho} < 1$ such that

$$\|P_{\mathcal{T}_{\mathbf{u}^n}}(\mathbf{y}^{n+1} - \mathbf{u}^n)\| \leq \tilde{\rho} \|P_{\mathcal{T}_{\mathbf{u}^n}}(\mathbf{u} - \mathbf{u}^n)\|. \quad (14.19)$$

For the third term in (14.17), first notice that by the Pythagorean theorem

$$\|\mathbf{u}^n - \mathbf{u}\|^2 = \|\mathbf{u}^n - P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{u}\|^2 + \|P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{u} - \mathbf{u}\|^2.$$

Since $P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{u}^n = \mathbf{u}^n$, using (14.15) one obtains

$$\|P_{\mathcal{T}_{\mathbf{u}^n}} \mathbf{u} - \mathbf{u}\| \leq \sqrt{1 - \gamma^{-2}} \|\mathbf{u}^n - \mathbf{u}\|. \quad (14.20)$$

Combining the estimates (14.18), (14.19), and (14.20) yields

$$\|\mathbf{u}^{n+1} - \mathbf{u}\| \leq \rho \|\mathbf{u} - \mathbf{u}^n\|,$$

where $\rho = C(1 + \delta_{3r})\varepsilon + \tilde{\rho} + \sqrt{1 - \gamma^{-2}} < 1$ for ε and κ small and γ close enough to 1. Consequently, the sequence \mathbf{u}^n converges linearly to \mathbf{u} .

The weak point of the previous theorem is that this convergence can be guaranteed only in a very narrow neighborhood of the solution. To obtain global convergence, now for TIHT, we ask for an additional assumption on the iterates.

Theorem 4 (Conditionally global convergence). *Let $\mathbf{b} = \mathcal{A}\mathbf{u}$ for $\mathbf{u} \in \mathcal{M}_{\leq r}$ and let \mathbf{u}^n be the iterates (14.13), (14.14) of the tensor iterative hard thresholding algorithm, i.e., $\mathcal{R}(\mathbf{u}^{n+1}) := \mathbf{H}_r(\mathbf{y}^{n+1})$. In addition, let's assume that \mathcal{A} satisfies the TRIP of order $3r$, i.e., $\delta_{3r} \leq 1/2$. We further assume that the iterates satisfy, for all $n \in \mathbb{N}$,*

$$\|\mathbf{u}^{n+1} - \mathbf{y}^{n+1}\| \leq \|\mathbf{u} - \mathbf{y}^{n+1}\|. \quad (14.21)$$

Then the sequence $\mathbf{u}^n \in \mathcal{M}_{\leq r}$ converges linearly to a unique solution $\mathbf{u} \in \mathcal{M}_{\leq r}$ with rate $\rho < 1$, i.e.,

$$\|\mathbf{u}^{n+1} - \mathbf{u}\| \leq \rho \|\mathbf{u}^n - \mathbf{u}\|.$$

For details of the proof, we refer to [51]. We note that the above result can be extended to robustness under noise on the measurements and to tensors being only approximately of low rank, i.e., being close to \mathcal{M}_r but not necessarily on \mathcal{M}_r .

Let us comment on the essential condition (14.21). In the case that \mathcal{R} computes the best rank r approximation then this condition holds since

$$\inf\{\|\mathbf{v} - \mathbf{y}^{n+1}\| : \mathbf{v} \in \mathcal{M}_{\leq r}\} \leq \|\mathbf{u} - \mathbf{y}^{n+1}\|$$

is trivially true. However, the best approximate is not numerically available and the truncated HSVD only ensures the worst case error estimate

$$\|\mathbf{u}^{n+1} - \mathbf{y}^{n+1}\| \leq C(d) \inf\{\|\mathbf{v} - \mathbf{y}^{n+1}\| : \mathbf{v} \in \mathcal{M}_{\leq r}\}.$$

Nevertheless, in practice this bound may be pessimistic for a generic tensor so that (14.21) may be likely to hold. In any case, the above theorem may at least *explain* why we observe recovery by TIHT in practice.

14.3.6 Alternating least squares scheme (ALS)

An efficient and fairly simple method for computing (at least a local) minimizer of $\|\mathbf{u} - \mathbf{v}\|$ subject to $\mathbf{v} \in \mathcal{M}_r$ is based on *alternating least squares* (ALS), which is a variant of block Gauß-Seidel optimization. In contrast to poor convergence experienced with the canonical format (CANDECOMP, PARAFAC) [36], ALS implemented appropriately in the hierarchical formats has been observed to be surprisingly powerful [34]. Furthermore, and quite importantly, it is robust with respect to over-fitting, i.e., allows optimization in the set $\mathcal{M}_{\leq r}$ [34]. As a local optimization scheme, like the Riemannian optimization it converges only to a local minimum at best. This scheme applied to TT tensors is basically a one-site DMRG (density matrix renormalization group) algorithm introduced for quantum spin systems in [58, 71]. The basic idea for computing $\mathbf{u} = \tau(\{\mathbf{b}_\alpha : \alpha \in \mathbb{T}\})$ by the ALS

or Block Gauß–Seidel method is to compute the required components \mathbf{b}_α , one after each other. Fixing the components \mathbf{b}_α , $\alpha \in \mathbb{T}$, $\alpha \neq t$, only the component \mathbf{b}_t is left to be optimized in each iteration step. Before passing to the next iteration step, the new iterate has to be transformed into the normal form \mathbf{b}_t^{n+1} by orthogonalization, e.g. by applying an SVD (without truncation) or simply by QR factorization.

Let us assume that the indices $\alpha \in \mathbb{T}$ are in a linear ordering $<$, which is consistent with the hierarchy. The components given by the present iterate are denoted by \mathbf{b}_α^n , if $\alpha \geq t$, respectively, \mathbf{b}_α^{n+1} for $\alpha < t$. We optimize over \mathbf{b}_t and introduce a corresponding tensor by

$$\mathbf{u}_{-t}^{n+1} := \tau(\{\mathbf{b}_\alpha^{n+1} : \alpha < t\} \cup \{\mathbf{b}_t\} \cup \{\mathbf{b}_\alpha^n : \alpha > t\}) \in \mathcal{H}_d.$$

Since the parametrization $\tau(\{\mathbf{b}_\alpha : \alpha \in \mathbb{T}\}) \in \mathcal{M}_{\leq r}$ is multi-linear in its arguments \mathbf{b}_α , the map τ_t^{n+1} defined by $\mathbf{b}_t \mapsto \tau_t^{n+1}(\mathbf{b}_t) := \mathbf{u}_{-t}^{n+1}$ is linear. The first order optimality condition for the present minimization is

$$0 = \nabla \mathcal{J} \circ \tau_t^n(\mathbf{b}_t) = (\tau_t^n)^* \mathcal{A}^* (\mathcal{A} \tau_t^n(\mathbf{b}_t) - \mathbf{b}),$$

which constitutes a linear equation for the unknown \mathbf{b}_t . It is not hard to show the following facts.

Theorem 5. 1. Suppose that \mathcal{A} satisfies the TRIP of order \mathbf{r} with $TRIC \ \delta_r < 1$ and that $\{\mathbf{b}_\alpha : \alpha \neq t\}$ is orthogonalized as above. Then, since $\|\tau_t^n(\mathbf{b}_t)\| = \|\mathbf{b}_t\|$, the TRIP reads as

$$(1 - \delta_r) \|\mathbf{b}_t\|^2 \leq \|\mathcal{A} \tau_t^n(\mathbf{b}_t)\|^2 \leq (1 + \delta_r) \|\mathbf{b}_t\|^2.$$

In addition, the functional $\mathcal{J} \circ \tau_t^n$ is strictly convex, and $\mathcal{J} \circ \tau_t^n$ possesses a unique minimizer \mathbf{b}_t^n .

2. For $\mathbf{u}_{-t}^n := \tau_t^n(\mathbf{b}_t^n)$, the sequence $\mathbf{J}(\mathbf{u}_{-t}^n)$ is nonincreasing with n , and it is decreasing unless \mathbf{u}_{-t}^n is a stationary point, i.e., $\nabla \mathbf{J}(\mathbf{u}_{-t}^n) \perp \mathcal{T}_{\mathbf{u}_{-t}^n}$: In the latter case the algorithm stagnates.
3. The sequence of iterates \mathbf{u}_{-t}^n is uniformly bounded.

This result implies at least the existence of a convergent subsequence. However, no conclusion can be drawn whether this algorithm recovers the original low rank tensor \mathbf{u} from $\mathbf{b} = \mathcal{A}\mathbf{u}$. For further convergence analysis of ALS, we refer, e.g., to [54].

In [72] convergence of a Block Gauß–Seidel method was shown by means of the Lojasiewicz–Kurdyka inequality. Also nonnegative tensor completion has been discussed there. It is likely that these arguments apply also to the present setting. The ALS is simplified if one rearranges the tree in each micro-iteration step such that one optimizes always the root. This can be easily done for TT tensors with left and right-orthogonalization [33, 34], and can be modified for general hierarchical tensors as well. Often, it is preferable to proceed in an opposite order after the optimization of all components (half-sweep). For the Gauss–Southwell variant, where one optimizes

the component with the largest defect, convergence estimates from gradient based methods can be applied [57]. Although the latter method converges faster, one faces a high computational overhead.

Let us remark that the Block Gauß-Seidel method and ALS strategy can be used in various situations, in particular, as one ingredient in the TIHT and RGI algorithms from the previous section. For instance, ALS can be applied directly after a gradient step defining the operator \mathcal{R} or one can use a simple half-sweep for approximating the gradient correction \mathbf{y}^{n+1} by a rank \mathbf{r} tensor in order to define the nonlinear projection \mathcal{R} .

14.4 Numerical results

For numerical tests, we concentrate on the HOSVD and the tensor iterative hardthresholding (TIHT) algorithm for recovering order $d = 3$ tensors from Gaussian measurement maps $\mathcal{A} : \mathcal{H}_3 = \bigotimes_{i=1}^3 \mathbb{R}^{n_i} \rightarrow \mathbb{R}^m$, i.e., the entries of \mathcal{A} identified with a tensor in $\mathbb{R}^m \otimes \bigotimes_{i=1}^3 \mathbb{R}^{n_i}$ are i.i.d. $\mathcal{N}(0, \frac{1}{m})$ random variables.

For these tests, we generate tensors $\mathbf{u} \in \mathcal{H}_3$ of rank $\mathbf{r} = (r_1, r_2, r_3)$ via its Tucker decomposition. Let us suppose that

$$\mathbf{u}(\mu_1, \mu_2, \mu_3) = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \sum_{k_3=1}^{r_3} \mathbf{c}(k_1, k_2, k_3) \mathbf{b}_{k_1}^1(\mu_1) \mathbf{b}_{k_2}^2(\mu_2) \mathbf{b}_{k_3}^3(\mu_3)$$

is the corresponding Tucker decomposition. Each entry of the core tensor is taken independently from the normal distribution, $\mathcal{N}(0, 1)$, and the component tensors $\mathbf{b}^j \in \mathbb{R}^{n_j \times r_j}$ are the first r_j left singular vectors of a matrix $\mathbf{M}^j \in \mathbb{R}^{n_j \times n_j}$ whose elements are also drawn independently from the normal distribution $\mathcal{N}(0, 1)$.

We then form the measurements $\mathbf{b} = \mathcal{A}\mathbf{u}$ and run the TIHT algorithm with the specified multi-linear rank $\mathbf{r} = (r_1, r_2, r_3)$ on \mathbf{b} . We test whether the algorithm successfully reconstructs the original tensor and say that the algorithm converged if $\|\mathbf{u} - \hat{\mathbf{u}}\| < 10^{-4}$. We stop the algorithm if it did not converge after 5000 iterations.

Figures 14.1–14.3 present the recovery results for low rank tensors of size $10 \times 10 \times 10$ (Figures 14.1 and Figure 14.2) and $6 \times 10 \times 15$ (Figure 14.3). The horizontal axis represents the number of measurements taken with respect to the number of degrees of freedom of an arbitrary tensor of this size. To be more precise, for a tensor of size $n_1 \times n_2 \times n_3$, the number \bar{n} on the horizontal axis represents $m = \lceil n_1 n_2 n_3 \frac{\bar{n}}{100} \rceil$ measurements. The vertical axis represents the percentage of the successful recovery. For fixed tensor dimensions $n_1 \times n_2 \times n_3$, fixed HOSVD-rank $\mathbf{r} = (r_1, r_2, r_3)$ and fixed number of measurements m , we performed 200 simulations.

Table 14.1 complements Figures 14.1–14.3. With $\%_{\max}$ we denote the maximal percentage of measurements for which we did not manage to recover even one tensor out of 200. The minimal percentage of measurements for full recovery is denoted by

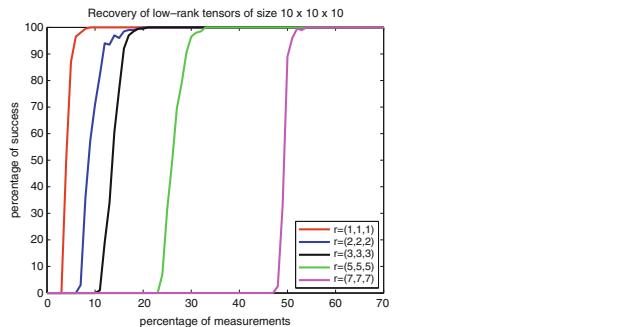


Fig. 14.1 Numerical results for $10 \times 10 \times 10$ tensors with same k -ranks.

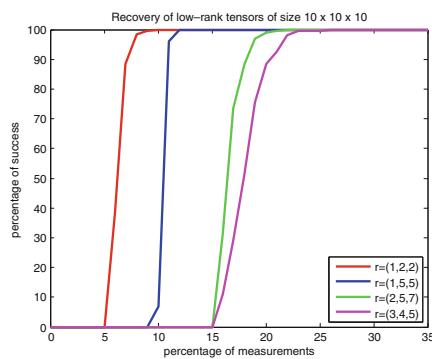


Fig. 14.2 Numerical results for $10 \times 10 \times 10$ tensors with different k -ranks.

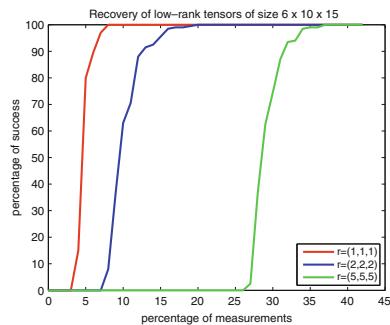


Fig. 14.3 Numerical results for $6 \times 10 \times 15$ tensors with same k -ranks.

Table 14.1 Numerical results for tensor IHT algorithm

$n_1 \times n_2 \times n_3$	rank	$\%_{\max}$	$\%_{\min}$	# of iterations for $\%_{\min}$
$10 \times 10 \times 10$	(1, 1, 1)	3	9	321
$10 \times 10 \times 10$	(2, 2, 2)	6	20	185
$10 \times 10 \times 10$	(3, 3, 3)	10	21	337
$10 \times 10 \times 10$	(5, 5, 5)	23	33	547
$10 \times 10 \times 10$	(7, 7, 7)	47	54	1107
$10 \times 10 \times 10$	(1, 2, 2)	5	10	588
$10 \times 10 \times 10$	(1, 5, 5)	9	12	1912
$10 \times 10 \times 10$	(2, 5, 7)	15	22	696
$10 \times 10 \times 10$	(3, 4, 5)	15	26	384
$6 \times 10 \times 15$	(1, 1, 1)	3	8	511
$6 \times 10 \times 15$	(2, 2, 2)	7	20	214
$6 \times 10 \times 15$	(5, 5, 5)	26	37	501

$\%_{\min}$. The last column represents the number of iterations needed for full recovery with $m = \lceil n_1 n_2 n_3 \frac{\%_{\min}}{100} \rceil$ number of measurements.

14.5 Concluding remarks

In this chapter we considered low rank tensor recovery for hierarchical tensors extending the classical Tucker format to a multi-level framework. For low ranks, this model can break the curse of dimensionality. Its number of degrees of freedom scale like $\mathcal{O}(ndr + dr^3) \ll \mathcal{O}(n^d)$ and $\mathcal{O}(ndr^2)$ for TT tensors instead of $\mathcal{O}(n^d)$. Under the assumption of a tensor restricted isometry property, we have shown local convergence for Riemannian gradient iterations and global convergence under a certain condition on the iterates of the tensor iterative hard thresholding algorithm for hierarchical tensors, including the classical Tucker format as well as tensor trains. For instance for TT tensors, an estimate of the TRIP for Gaussian measurement maps was provided that requires the number of measurements to scale like $m \sim ndr^2 \log(dr)$. However, it is still not clear whether the logarithmic factor is needed.

Let us finally mention some open problems. One important task is to establish global convergence to the original tensor of any of the discussed algorithms, without additional assumptions such as (14.21) on the iterates. In addition, robustness and stability for the Riemannian gradient method are still open. Further, also the TRIP

related to general HT tensors for Gaussian measurement maps is not yet established. Since the TRIP does not hold for the completion problem, it is not clear yet whether a low rank tensor can be recovered from less than $\mathcal{O}(n^{d/2})$ entries.

References

1. Absil, P.-A., Mahony, R.E., Sepulchre, R.: Optimization algorithms on matrix manifolds. *Found. Comput. Math.* **10**, 241–244 (2010)
2. Arnold, A., Jahnke, T.: On the approximation of high-dimensional differential equations in the hierarchical Tucker format. *BIT Numer. Math.* **54**, 305–341 (2014)
3. Beck, M.H., Jäckle, A., Worth, G.A., Meyer, H.-D.: The multi-configuration time-dependent Hartree (MCTDH) method: a highly efficient algorithm for propagating wavepackets. *Phys. Rep.* **324**, 1–105 (2000)
4. Beylkin, G., Mohlenkamp, M.J.: Algorithms for numerical analysis in high dimensions. *SIAM J. Sci. Comput.* **26**, 2133–2159 (2005)
5. Beylkin, G., Garecke, J., Mohlenkamp, M.J.: Multivariate regression and machine learning with sums of separable functions. *SIAM J. Sci. Comput.* **31**, 1840–1857 (2009)
6. Bhatia, R.: *Matrix Analysis*. Graduate Texts in Mathematics, vol. 169. Springer, New York (1997)
7. Blumensath, T., Davies, M.: Iterative thresholding for sparse approximations. *J. Fourier Anal. Appl.* **14**, 629–654 (2008)
8. Blumensath, T., Davies, M.: Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.* **27**, 265–274 (2009)
9. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. *Found. Comput. Math.* **9**, 717–772 (2009)
10. Candès, E.J., Plan, Y.: Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Trans. Inf. Theory* **57**, 2342–2359 (2011)
11. Candès, E.J., Tao, T.: The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inf. Theory* **56**, 2053–2080 (2010)
12. Carlini, E., Kleppe, J.: Ranks derived from multilinear maps. *J. Pure Appl. Algebra* **215**, 1999–2004 (2011)
13. Da Silva, C., Herrmann, F.J.: Hierarchical Tucker tensor optimization - applications to tensor completion. In: *Proceedings of 10th International Conference on Sampling Theory and Applications* (2013)
14. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **21**, 1253–1278 (2000)
15. Eldar, Y.C., Kutyniok, K. (eds.): *Compressed Sensing : Theory and Applications*. Cambridge University Press, Cambridge (2012)
16. Falcó, A., Hackbusch, W.: On minimal subspaces in tensor representations. *Found. Comput. Math.* **12**, 765–803 (2012)
17. Falcó, A., Hackbusch, W., Nouy, A.: Geometric structures in tensor representations. *Technical Reports*, vol. 9. MPI MIS Leipzig (2013)
18. Fazel, M.: Matrix rank minimization with applications. Ph.D. thesis, Stanford University, CA (2002)
19. Foucart, S., Rauhut, H.: *A Mathematical Introduction to Compressive Sensing. Applied and Numerical Harmonic Analysis*. Birkhäuser, New York (2013)
20. Friedland, S., Lim, L.-H.: Computational complexity of tensor nuclear norm, preprint, ArXiv:1410.6072 (2014)
21. Gandy, S., Recht, B., Yamada, I.: Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Probl.* **27**, 025010 (2011)

22. Grasedyck, L.: Hierarchical singular value decomposition of tensors. *SIAM. J. Matrix Anal. Appl.* **31**, 2029–2054 (2010)
23. Grasedyck, L., Kressner, D., Tobler, C.: A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen* **36**, 53–78 (2013)
24. Gross, D.: Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inf. Theory* **57**, 1548–1566 (2011)
25. Hackbusch, W.: Tensorisation of vectors and their efficient convolution. *Numer. Math.* **119**, 465–488 (2011)
26. Hackbusch, W.: *Tensor Spaces and Numerical Tensor Calculus*. Springer Series in Computational Mathematics, vol. 42. Springer, New York (2012)
27. Hackbusch, W.: Numerical tensor calculus. *Acta Numerica* **23**, 651–742 (2014)
28. Hackbusch, W., Kühn, S.: A new scheme for the tensor representation. *J. Fourier Anal. Appl.* **15**, 706–722 (2009)
29. Hackbusch, W., Schneider, R.: Tensor spaces and hierarchical tensor representations, In: Dahlke, S., Dahmen, W., Griebel, M., Hackbusch, W., Ritter, K., Schneider, R., Schwab, C., Yserentant, H. (eds.), *Extraction of quantifiable information from complex systems, Lecture notes in computational science and engineering*, vol. 102, publisher, Springer, New York, pp. 237–361 (2014)
30. Haegeman, J., Osborne, T., Verstraete, F.: Post-matrix product state methods: to tangent space and beyond. *Phys. Rev. B* **88**, 075133 (2013)
31. Hastad, J.: Tensor rank is NP-complete. *J. Algorithms* **11**, 644–654 (1990)
32. Hillar, C.J., Lim, L.-H.: Most tensor problems are NP hard. *J. ACM* **60**, 45:1–45:39 (2013)
33. Holtz, S., Rohwedder, T., Schneider, R.: On manifolds of tensors of fixed TT rank. *Numer. Math.* **120**, 701–731 (2012)
34. Holtz, S., Rohwedder, T., Schneider, R.: The alternating linear scheme for tensor optimisation in the tensor train format. *SIAM J. Sci. Comput.* **34**, A683–A713 (2012)
35. Huang, B., Mu, C., Goldfarb, D., Wright, J.: Provable low-rank tensor recovery. http://www.optimization-online.org/DB_FILE/2014/02/4252.pdf (2014)
36. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. *SIAM Rev.* **51**, 455–500 (2009)
37. Kressner, D., Steinlechner, M., Vandereycken, B.: Low-rank tensor completion by Riemannian optimization. *BIT Numer. Math.* **54**, 447–468 (2014)
38. Landsberg, J.M.: *Tensors: Geometry and Applications*. Graduate Studies in Mathematics, vol. 128. AMS, Providence (2012)
39. Legeza, Ö., Rohwedder, T., Schneider, R., Szalay, S.: Tensor product approximation (DMRG) and coupled cluster method in quantum chemistry. In: Bach, V., Delle Site, L. (eds.) *Many-Electron Approaches in Physics, Chemistry and Mathematics*, pp. 53–76. Springer, Switzerland (2014)
40. Levin, J.: Three-mode factor analysis. Ph.D. thesis, University of Illinois, Urbana (1963)
41. Lim, L.-H., De Silva, V.: Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM J. Matrix Anal. Appl.* **30**, 1084–1127 (2008)
42. Liu, J., Musalski, P., Wonka, P., Ye, J.: Tensor completion for estimating missing values in visual data. *Trans. Pattern Anal. Mach. Intell. (PAMI)* **35**, 208–220 (2012)
43. Lubich, C.: *From Quantum to Classical Molecular Dynamics: Reduced Methods and Numerical Analysis*. Zürich Lectures in Advanced Mathematics, vol. 12. EMS, Zürich (2008)
44. Lubich, C., Rohwedder, T., Schneider, R., Vandereycken, B.: Dynamical approximation by hierarchical Tucker and tensor-train tensors. *SIAM J. Matrix Anal. Appl.* **34**, 470–494 (2013)
45. Mu, C., Huang, B., Wright, J., Goldfarb, D.: Square deal: lower bounds and improved relaxations for tensor recovery. arxiv.org/abs/1307.5870v2 (2013)
46. Oseledets, I.V.: A new tensor decomposition. *Dokl. Math.* **80**, 495–496 (2009)

47. Oseledets, I.V.: Tensor-train decomposition. *SIAM J. Sci. Comput.* **33**, 2295–2317 (2011)
48. Oseledets, I.V., Tyrtyshnikov, E.E.: Breaking the curse of dimensionality, or how to use SVD in many dimensions. *SIAM J. Sci. Comput.* **31**, 3744–3759 (2009)
49. Oseledets, I.V., Tyrtyshnikov, E.E.: Algebraic wavelet transform via quantics tensor train decomposition. *SIAM J. Sci. Comput.* **33**, 1315–1328 (2011)
50. Rauhut, H., Schneider, R., Stojanac, Ž.: Tensor recovery via iterative hard thresholding. In: Proceedings of 10th International Conference of Sampling Theory and Applications (2013)
51. Rauhut, H., Schneider, R., Stojanac, Ž.: Low rank tensor recovery via iterative hard thresholding (in preparation)
52. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solution of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52**, 471–501 (2010)
53. Recht, B.: A simpler approach to matrix completion. *J. Mach. Learn. Res.* **12**, 3413–3430 (2011)
54. Rohwedder, T., Uschmajew, A.: On local convergence of alternating schemes for optimization of convex problems in the tensor train format. *SIAM J. Numer. Anal.* **51**, 1134–1162 (2013)
55. Romera-Paredes, B., Pontil, M.: A new convex relaxation for tensor completion. *NIPS* **26**, 2967–2975 (2013)
56. Schneider, R., Uschmajew, A.: Approximation rates for the hierarchical tensor format in periodic Sobolev spaces. *J. Complexity* **30**, 56–71 (2014)
57. Schneider, R., Uschmajew, A.: Convergence results for projected line-search methods on varieties of low-rank matrices via Lojasiewicz inequality. *SIAM J. Optim.*, **25**(1), 622–646 (2015)
58. Schollwöck, U.: The density-matrix renormalization group in the age of matrix product states. *Ann. Phys. (NY)* **326**, 96–192 (2011)
59. Signoretto, M., De Lathauwer, L., Suykens, J.A.K.: Nuclear norms for tensors and their use for convex multilinear estimation. International Report 10–186, ESAT-SISTA, K. U. Leuven (2010)
60. Tanner, J., Wei, K.: Normalized iterative hard thresholding for matrix completion. *SIAM J. Sci. Comput.* **35**, S104–S125 (2013)
61. Tucker, L.R.: Implications of factor analysis of three-way matrices for measurement of change. In: Harris, C.W. (ed.) *Problems in Measuring Change*, pp. 122–137. University of Wisconsin Press, Madison (1963)
62. Tucker, L.R.: The extension of factor analysis to three-dimensional matrices. In: Gulliksen, H., Frederiksen, N. (eds.) *Contributions to Mathematical Psychology*, pp. 110–127. Holt, Rinehart & Winston, New York (1964)
63. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. *Psychometrika* **31**, 279–311 (1966)
64. Uschmajew, A.: Well-posedness of convex maximization problems on Stiefel manifolds and orthogonal tensor product approximations. *Numer. Math.* **115**, 309–331 (2010)
65. Uschmajew, A., Vandeheycken, B.: The geometry of algorithms using hierarchical tensors. *Linear Algebra Appl.* **439**, 133–166 (2013)
66. Vandeheycken, B.: Low-rank matrix completion by Riemannian optimization. *SIAM J. Optim.* **23**, 1214–1236 (2013)
67. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. In: Eldar, C.Y., Kutyniok, G. (eds.) *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, Cambridge (2012)
68. Vidal, G.: Efficient classical simulation of slightly entangled quantum computations. *Phys. Rev. Lett.* **91**, 147902 (2003)
69. Wang, H., Thoss, M.: Multilayer formulation of the multi-configuration time-dependent Hartree theory. *J. Chem. Phys.* **119**, 1289–1299 (2003)

70. Wen, Z., Yin, W., Zhang, Y.: Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Math. Program. Comput.* **4**, 333–361 (2012)
71. White, S.: Density matrix formulation for quantum renormalization groups. *Phys. Rev. Lett.* **69**, 2863–2866 (1992)
72. Xu, Y., Yin, W.: A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM J. Imaging Sci.* **6**, 1758–1789 (2013)
73. Xu, Y., Hao, R., Yin, W., Su, Z.: Parallel matrix factorisation for low-rank tensor completion. *UCLA CAM*, 13–77 (2013)

Chapter 15

Compressive Classification: Where Wireless Communications Meets Machine Learning

Miguel Rodrigues, Matthew Nokleby, Francesco Renna,
and Robert Calderbank

Abstract This chapter introduces Shannon-inspired performance limits associated with the classification of low-dimensional subspaces embedded in a high-dimensional ambient space from compressive and noisy measurements. In particular, it introduces the *diversity-discrimination tradeoff* that describes the interplay between the number of classes that can be separated by a compressive classifier—measured via the *discrimination gain*—and the performance of such a classifier—measured via the *diversity gain*—and the relation of such an interplay to the underlying problem geometry, including the ambient space dimension, the subspaces dimension, and the number of compressive measurements. Such a fundamental limit on performance is derived from a syntactic equivalence between the compressive classification problem and certain wireless communications problems. This equivalence provides an opportunity to cross-pollinate ideas between the wireless information theory domain and the compressive classification domain. This chapter also demonstrates how theory aligns with practice in a concrete application: face recognition from a set of noisy compressive measurements.

M. Rodrigues (✉)

Department of Electronic and Electrical Engineering, University College London, London, UK
e-mail: m.rodrigues@ucl.ac.uk

M. Nokleby • R. Calderbank

Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA
e-mail: matthew.nokleby@duke.edu; robert.calderbank@duke.edu

F. Renna

Department of Computer Science, Instituto de Telecomunicações,
University of Porto, Porto, Portugal
e-mail: frarennna@dcc.fc.up.pt

15.1 Introduction

The reliable classification of high-dimensional signals from low-dimensional features is an increasingly crucial task in the age of the data deluge. Such compressive classification problems are relevant to a myriad of applications that involve massive datasets, ranging from face and hand-written digit recognition [1, 20, 24, 26, 27, 55] to tumor classification [3, 40].

Compressive classification appears in the machine learning literature as feature extraction or dimension reduction. For example, linear dimension reduction methods based on geometrical characterizations of the source have been developed, with linear discriminant analysis (LDA) and principal component analysis (PCA) just depending on second-order statistics [17]. Linear dimension reduction methods based on higher-order statistics of the data have also been developed [12, 14, 21, 25, 29, 31, 44–46, 52]. In particular, mutual information is used to extract features from data in [12, 14, 52] and approximations to mutual information based on the Rényi entropy are used in [21, 45, 46]. Nonlinear (supervised) dimension reduction methods have also become popular recently [47, 53].

Compressive classification also appears in the signal and image processing literature in view of recent advances in compressive sensing [9–11, 16] (see also chapter 1). Compressive information processing, recently proposed by Davenport et al. [15], advocates the resolution of various information processing tasks such as detection, classification, or pattern recognition directly in the compressive measurement domain rather than the original high-dimensional signal domain, using conventional random compressive measurement strategies that are agnostic to the exact form of the original signal - so applicable to a large class of signals. Other instances of compressive classification appear under the guise of sparse support recovery in compressive sensing [37, 38, 48–51].

With the sensing, analysis and processing of high-dimensional data now at the heart of multiple fields, it is becoming clear that there is the need for rigorous mathematical frameworks that shed light on fundamental performance limits associated with the classification of high-dimensional data from low-dimensional data features. This chapter provides insight about the performance of the classification of signals based on a set of linear compressive and noisy measurements, with a focus on phenomena that can be modelled by a union of low-dimensional subspaces embedded in a high-dimensional ambient space.

This chapter argues that, in view of a syntactic equivalence between the compressive classification problem and certain wireless communications problems, it is possible to translate ideas from the wireless information theory domain to the machine learning domain in order to construct Shannon-inspired performance characterizations of the compressive classification problem. The chapter then introduces the diversity-discrimination tradeoff (DDT)—a construction reminiscent of the diversity-multiplexing tradeoff (DMT) in wireless communications—that captures the interplay between the number of classes that can be separated by a compressive classifier and the performance of such a classifier in terms of the

underlying problem geometry. It is shown that the DDT unveils not only ultimate performance limits in compressive classification but also optimal signal geometries for classification, a fact that has implications on how an “engineer” may design features from data. It is also shown that theory aligns with practice in real-world problems, namely, in face recognition applications.

Finally, we note that, beyond the DDT, our framework can also be used to construct other wireless information theory and wireless communications theory inspired characterizations, as described in [32, 36].

15.2 The Compressive Classification Problem

We consider the problem of classification from compressive and noisy measurements. In particular, we use the standard measurement model given by:

$$y = Ax + z, \quad (15.1)$$

where $y \in R^m$ represents the measurement vector, $x \in R^n$ represents the random signal vector, $A \in R^{m \times n}$ represents the random measurement matrix or kernel,¹ and $z \sim \mathcal{N}(0, \sigma^2 \cdot I) \in R^m$ represents standard white Gaussian noise. We also consider the (energy) constraints given by:

$$E[\|A\|_2^2] \leq 1 \quad (15.2)$$

$$E[\|x\|_2^2] \leq m \quad (15.3)$$

where $\|\cdot\|_2$ is the Euclidean norm of a vector or the spectral norm of a matrix, so that the signal-to-noise ratio (SNR) becomes $\text{SNR} = 1/\sigma^2$. We also suppose, in keeping with our interest in the compressive regime, that $n \geq m$ throughout.

We assume that the signal x follows a certain distribution q , i.e. $x \sim q$, where the distribution q is drawn equiprobably from a class alphabet or distribution alphabet \mathcal{P} with cardinality L , which is known to the classifier. The classification problem is to produce an estimate $\hat{q} \in \mathcal{P}$ of the true underlying distribution $q \in \mathcal{P}$, based on the signal measurements, so that the average misclassification probability is given by:

$$P_e = \frac{1}{L} \sum_{q \in \mathcal{P}} \Pr(\hat{q} \neq q | x \sim q). \quad (15.4)$$

¹We constrain the random measurement kernel to be full row rank. We also constrain the distribution of the random measurement kernel to be invariant to rotations. These constraints are obeyed by the standard Gaussian i.i.d. random kernels in compressive sensing.

In the sequel, and motivated by applications of compressive classification of subspaces in modern signal processing, image processing, and machine learning problems [7, 18, 20, 28, 42, 43, 54], we will exclusively concentrate on the classification of Gaussian distributions that are supported on a low-dimensional linear subspace (of dimension k) embedded in the higher-dimensional ambient space (of dimension n). In particular, we assume that the alphabet of distributions \mathcal{P} is contained in a distribution constraint set \mathcal{Q} given by:²

$$\mathcal{Q} = \{q : q = \mathcal{N}(0, \Psi \Psi^T), \Psi \in \mathbb{R}^{n \times k}, \alpha \leq \text{eig}(\Psi^T \Psi) \leq m/k\}, \quad (15.5)$$

for some $0 < \alpha \leq m/k$, where the lower bound on the eigenvalues ensures that the subspace associated with a distribution $q \in \mathcal{Q}$ has k non-trivial dimensions whereas the upper bound on the eigenvalues enforces the energy constraint. We also concentrate on classification problems where $m > k$.

15.3 Dualities: Compressive Classification and Multiple-Antenna Wireless Communications

We derive insight into classification through a duality between the classification of low-rank Gaussian distributions from compressive and noisy linear measurements and the communication of codewords over Rayleigh fading multiple-antenna non-coherent channels (a communications model studied in detail in [30, 57]). This channel model, which consists of a transmitter having k antennas, a receiver having l antennas, and a channel matrix unknown to the transmitter and receiver that persists for m symbol times, can be modelled as follows:³

$$Y = HX + Z, \quad (15.6)$$

where $Y \in \mathbb{R}^{l \times m}$ represents the received codeword, $X \in \mathbb{R}^{k \times m}$ represents the transmitted codeword, $H \in \mathbb{R}^{l \times k}$ represents an i.i.d. zero-mean unit-variance Gaussian matrix that expresses the gains between the different transmit and receive antennas, and $Z \in \mathbb{R}^{l \times m}$ represents i.i.d. Gaussian noise.

We can unveil the duality by re-writing the measurement model (15.1) as follows:

$$y = A\Psi c + z \quad (15.7)$$

²The final inequalities hold element-wise.

³We assume that the channel model is real rather than complex as in previous treatments in the wireless communications literature (e.g., [30, 57]). This distinction is unimportant, as it is straightforward to adapt arguments over complex-valued channels to real-valued ones and vice-versa.

or

$$y^T = c^T \Psi^T A^T + z^T. \quad (15.8)$$

where the vector $c \sim \mathcal{N}(0, I) \in R^k$ “drives” the matrix $\Psi \in R^{n \times k}$ to produce the signal vector $x \sim \mathcal{N}(0, \Psi \Psi^T) \in R^n$. We can thus conclude that the classification of a k -dimensional Gaussian distribution embedded in an n -dimensional space based on m noisy measurements is equivalent to the identification of the space-time codeword $\Psi^T A^T \in R^{k \times m}$ communicated over the non-coherent channel with k transmit antennas, a single receive antenna, and a coherence time of m (in multiples of symbol duration).

This syntactic equivalence provides an opportunity to translate methods used to characterize performance of multiple-antenna wireless communication to the compressive classification domain. In particular, we seek fundamental limits in compressive classification that capture the tradeoff between the rate of increase in the number of classes and the rate of decay in the probability of misclassification that—akin to the DMT in wireless communications [57, 58]—we will dub as the DDT.⁴

However, and beyond the syntactic equivalence disclosed in (15.7) and (15.8), it is also relevant to briefly reflect on the operational implications of the new characterization. The DMT establishes that there are no transmission schemes that are able to achieve a certain rate of growth of codewords and rate of decay of the error probability outside the region bounded by the DMT curve. It also establishes the geometry of DMT-achieving transmission schemes. This provides concrete optimal design guidelines for a communications system designer.

The DDT—as the DMT—also offers both a fundamental characterization of the tradeoff between the rate of growth of classes and the rate of decay of the misclassification probability in a classification problem together with optimal geometries for classification problems. The DDT can then be seen as a fundamental framework that provides an “engineer” with the insight to optimally design low-dimensional features from high-dimensional data—an aspect that has been at the forefront of recent contributions [33–35].

⁴The DMT was introduced in the context of wireless communications to characterize the high-SNR performance of fading coherent MIMO channels [57, 58]. It shows that the spatial flexibility provided by multiple antennas can simultaneously increase the achievable rate and decrease the probability of error in a wireless communications channel, but only according to a tradeoff that is tightly characterized at high SNR.

15.4 Wireless Communications Inspired Performance Characterizations

We now use the duality to derive the DDT expressing the tradeoff between the number of classes in the compressive classification problem (measured via the discrimination gain) and the performance of the compressive classification problem (measured via the diversity gain), in the low-noise regime.

15.4.1 The Diversity-Discrimination Tradeoff (DDT)

The characterization of such a tradeoff—as in wireless communications— involves constructing a sequence of class or distribution alphabets indexed by the SNR, in order to examine the associated discrimination and diversity gains. In particular, let us fix the distribution constraint set \mathcal{Q} . Let $\{\mathcal{P}^{(\text{SNR})}\}$ be a sequence of distribution alphabets indexed by the SNR, where each alphabet $\mathcal{P}^{(\text{SNR})} \subset \mathcal{Q}$, and let $\{P_e^{(\text{SNR})}\}$ be the associated sequence of misclassification probabilities indexed by the SNR. The number of distributions $L^{(\text{SNR})}$ associated with each alphabet can vary with the SNR but the dimensions n , m , and k associated with each alphabet are fixed. We can now define the discrimination gain, the diversity gain and thereby the DDT of a classification problem.

Definition 1. Fix the distribution constraint set \mathcal{Q} . A sequence of class alphabets $\{\mathcal{P}^{(\text{SNR})}\}$, where $\mathcal{P}^{(\text{SNR})} \subset \mathcal{Q}$, is said to have discrimination gain r and diversity gain d if

$$\lim_{\text{SNR} \rightarrow \infty} \frac{\log L^{(\text{SNR})}}{\frac{1}{2} \log(\text{SNR})} = r \quad (15.9)$$

and

$$\lim_{\text{SNR} \rightarrow \infty} -\frac{\log(P_e^{(\text{SNR})})}{\frac{1}{2} \log(\text{SNR})} = d. \quad (15.10)$$

The sequence of class alphabets $\{\mathcal{P}^{(\text{SNR})}\}$ is also said to have diversity-discrimination function $d(r)$.

In simple terms, the sequence of class alphabets $\{\mathcal{P}^{(\text{SNR})}\}$ has approximately $\text{SNR}^{r/2}$ classes and affords a misclassification probability of approximately $\text{SNR}^{-d/2}$ at high SNR.

Definition 2. Fix the distribution constraint set \mathcal{Q} . The diversity-discrimination tradeoff is defined as the supremum over the diversity-discrimination functions of all permissible sequences of class alphabets $\{\mathcal{P}^{(\text{SNR})}\}$, where $\mathcal{P}^{(\text{SNR})} \subset \mathcal{Q}$, or

$$d^*(r) = \sup_{\{\mathcal{P}^{(\text{SNR})}\}, \mathcal{P}^{(\text{SNR})} \in \mathcal{Q}} d(r). \quad (15.11)$$

This definition implies that, at high SNR, a sequence of class alphabets $\{\mathcal{P}^{(\text{SNR})}\}$ with SNR^r classes cannot have a misclassification probability decaying faster than $\text{SNR}^{-d^*(r)}$. The practical significance of these definitions is associated with the fact that there exist sequences of class alphabets—with a precisely specified geometry—that satisfy such scaling laws.

The following Theorem now characterizes the DDT associated with the classification of low-rank Gaussian distributions from compressive and noisy measurements.

Theorem 1. *Given the measurement model in (15.1) subject to the constraints in (15.2) and (15.3) and given the distribution constraint set in (15.5), the DDT is bounded above by*

$$d(r) \leq k \left[1 - \frac{r}{m-k} \right]^+ \quad (15.12)$$

and is bounded below by

$$d(r) \geq [\min\{m-k, k\} - r]^+. \quad (15.13)$$

where $[.]^+ = \max(0, \cdot)$.

Proof. We only outline a proof of the Theorem. To prove the outer bound, we adapt the outage-based argument of [56]. When the norm of the “driving” process $\|c\|_2^2$ is small, the mutual information between the signal classes and the signal measurements is small, so Fano’s inequality guarantees that the probability of misclassification is bounded away from zero. It is straightforward to show that the event that $\|c\|_2^2$ is too small corresponds to the DDT outer bound. To prove the inner bound, we construct an ensemble of classes drawn uniformly from the appropriate Grassmann manifold. Using the Bhattacharyya bound on probability of misclassifying Gaussian classes, we bound the probability of confusing any two subspaces. Applying the union bound to this pairwise error probability, we obtain the result. The full proof appears in [32]. \square

Fig. 15.1 depicts the upper and lower bounds to the DDT embodied in Theorem 1. Note that Theorem 1 only provides a partial characterization of the DDT, since the upper and lower bounds do not match in general. Nevertheless, the lower bound achieves full diversity gain when $r = 0$ in the regime $m \geq 2k$ and full discrimination gain when $d = 0$ in the regime $m \leq 2k$. Note also that Theorem 1 reveals the relationship between some of the problem dimensions and high-SNR performance. The maximum possible discrimination gain is $r = m - k$, which corresponds to $d = 0$, or very slowly decaying error probability. By contrast, the maximum possible diversity gain is $d = k$, corresponding to $r = 0$, or a constant

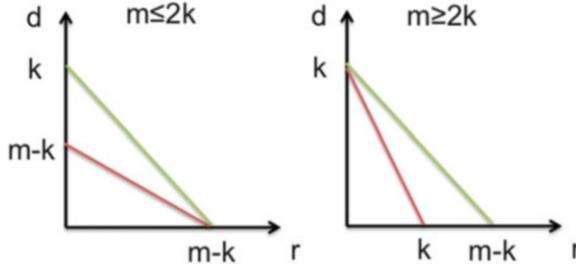


Fig. 15.1 Upper bound (green) and lower bound (red) to the diversity-discrimination tradeoff for $m \leq 2k$ (left) and $m \geq 2k$ (right).

(or very slowly increasing) number of classes. For fixed subspace dimension $k < m$, an increase in the number of measurements permits either an increase in the number of discernible classes or a decrease in the misclassification probability. Similarly, for a fixed number of measurements $m > k$, it is easier to discriminate between lower-dimensional subspaces than those of higher dimension. Note finally that the DDT does not depend on the ambient space dimension n ; only the relationship between the number of measurements and the subspace dimension governs the high-SNR performance.

The proof of Theorem 1 also reveals the lower bound is achieved by an ensemble of subspaces uniformly distributed over the Grassmann manifold. We conjecture that this latter ensemble is DDT-optimal, but our analysis cannot establish this claim because the Chernoff bound used to estimate the misclassification probability is not always tight.

15.4.2 The Diversity Gain at Zero Discrimination Gain

It has been shown that certain class geometries are able to achieve the fundamental limit portrayed by the DDT. However, it is also of interest to augment the analysis in order to understand how other class geometries may affect the performance of the compressive classification problem.

We now consider a class or distribution alphabet with fixed cardinality that does not scale with the SNR, so that the discrimination gain is zero, given by:

$$\mathcal{P} = \{q_1 = \mathcal{N}(0, \Psi_1 \Psi_1^T), \dots, q_L = \mathcal{N}(0, \Psi_L \Psi_L^T)\} \quad (15.14)$$

where $\Psi_i \in \mathbb{R}^{n \times k}, i = 1, \dots, L$ and $\alpha \leq \text{eig}(\Psi_i^T \Psi_i) \leq m/k, i = 1, \dots, L$ for some $0 < \alpha \leq m/k$. We also consider a high-SNR asymptotic characterization of the probability of misclassification associated with the optimal maximum *a posteriori* (MAP) classifier via the diversity gain. However, we use an upper bound to the

misclassification probability P_e^{UB} in lieu of the true misclassification probability P_e , which results from using the union bound in conjunction with the Bhattacharyya bound [17], given by:⁵

$$P_e(\text{SNR}) \leq P_e^{UB}(\text{SNR}) = \sum_{i \neq j} \frac{1}{L} \cdot e^{-K_{ij}(\text{SNR})}, \quad (15.15)$$

where

$$K_{ij}(\text{SNR}) = \frac{1}{2} \log \frac{\det \left(A \left(\frac{\Psi_i \Psi_i^T + \Psi_j \Psi_j^T}{2} \right) A^T + \frac{1}{\text{SNR}} \cdot I \right)}{\sqrt{\det(A \Psi_i \Psi_i^T A^T + \frac{1}{\text{SNR}} \cdot I) \det(A \Psi_j \Psi_j^T A^T + \frac{1}{\text{SNR}} \cdot I)}}. \quad (15.16)$$

The following Theorem offers a high-SNR asymptotic characterization of the behavior of the (upper bound to) the misclassification probability. In particular, we express such an asymptotic behavior as a function of certain geometrical quantities associated with the class alphabet, including:

$$r_i = \text{rank}(\Psi_i \Psi_i^T) \quad (15.17)$$

which relates to the dimension of the subspace spanned by the signal in class i ,

$$r_{ij} = \text{rank}(\Psi_i \Psi_i^T + \Psi_j \Psi_j^T) \quad (15.18)$$

which relates to the dimension of the direct sum of subspaces spanned by the signals in classes i or j , and

$$no_{ij} = r_{ij} - [(r_i + r_j) - r_{ij}] = 2 \cdot (r_{ij} - k) \quad (15.19)$$

that relates to the difference between the dimension of the subspaces spanned by signals in classes i or j and the dimension of the intersection of such subspaces. This can also be interpreted as the number of non-overlapping dimensions between the subspaces spanned by the eigenvectors corresponding to the non-zero eigenvalues of the covariance matrices pertaining to the two classes [36]. Note that subspaces i and j completely overlap if and only if $r_{ij} = \frac{r_i + r_j}{2} = k \Leftrightarrow no_{ij} = 0$.

Theorem 2. *Consider the measurement model in (15.1) subject to the constraints in (15.2) and (15.3). Consider also the class or distribution alphabet in (15.14). Then, the upper bound to the probability of misclassification behaves as:*

$$P_e(\text{SNR}) \leq P_e^{UB}(\text{SNR}) = c \cdot \frac{1}{\text{SNR}^{d/2}} + o\left(\frac{1}{\text{SNR}^{d/2}}\right), \quad \text{SNR} \rightarrow \infty \quad (15.20)$$

⁵By working with the upper bound to the misclassification probability rather than the true one, we obtain a lower bound to the diversity gain rather than the exact one.

where c is a constant that does not depend on the SNR, $d = \min_{i \neq j} d_{ij}$, and

- If $\frac{r_i+r_j}{2} = r_{ij}$, then $d_{ij} = 0$
- If $\frac{r_i+r_j}{2} < r_{ij}$, then

$$d_{ij} = \left(m - \frac{r_i + r_j}{2} \right) = (m - k) \quad (15.21)$$

for $r_{ij} > m > r_i = r_j = k$ and

$$d_{ij} = \left(r_{ij} - \frac{r_i + r_j}{2} \right) = (r_{ij} - k) \quad (15.22)$$

for $m \geq r_{ij} > r_i = r_j = k$.

Proof. We also only outline the proof of the Theorem. Consider the eigenvalue decompositions of the positive semidefinite matrices:

$$A\Psi_i\Psi_i^T A^T = V_i \text{diag}(\lambda_{i_1}, \dots, \lambda_{i_{s_i}}, 0, \dots, 0) V_i^T \quad (15.23)$$

$$A\Psi_j\Psi_j^T A^T = V_j \text{diag}(\lambda_{j_1}, \dots, \lambda_{j_{s_j}}, 0, \dots, 0) V_j^T \quad (15.24)$$

$$A(\Psi_i\Psi_i^T + \Psi_j\Psi_j^T)A^T = V_{ij} \text{diag}(\lambda_{ij_1}, \dots, \lambda_{ij_{s_{ij}}}, 0, \dots, 0) V_{ij}^T, \quad (15.25)$$

where $s_i = \text{rank}(A\Psi_i\Psi_i^T A^T)$, $s_j = \text{rank}(A\Psi_j\Psi_j^T A^T)$ and $s_{ij} = \text{rank}(A(\Psi_i\Psi_i^T + \Psi_j\Psi_j^T)A^T)$. It follows immediately that the upper bound to the misclassification probability in (15.15) and (15.16) can be rewritten as follows:

$$P_e^{UB}(\text{SNR}) = \sum_{i \neq j} \frac{1}{L} \left(\frac{1}{\text{SNR}} \right)^{\frac{1}{2} \left(s_{ij} - \frac{s_i + s_j}{2} \right)} 2^{s_{ij}/2} \sqrt{\frac{\sqrt{\prod_{k=1}^{s_i} (\lambda_{i_k} + 1/\text{SNR}) \prod_{k=1}^{s_j} (\lambda_{j_k} + 1/\text{SNR})}}{\prod_{k=1}^{s_{ij}} (\lambda_{ij_k} + 2/\text{SNR})}}. \quad (15.26)$$

Finally, we can conclude the proof by noting that, with probability 1, $s_i = \min\{m, r_i\}$, $s_j = \min\{m, r_j\}$ and $s_{ij} = \min\{m, r_{ij}\}$. The full proof appears in [36]. \square

The expansions to the (upper bound to) the probability of misclassification embodied in Theorem 2 encapsulate intuitive operational aspects associated with the compressive classification problem:

- When there is at least a pair of subspaces that overlaps, i.e. there exists a pair of subspaces such that $\frac{r_i+r_j}{2} = r_{ij}$, then the upper bound to the probability of misclassification exhibits an error floor;
- When there is no pair of subspaces that overlaps, i.e. for all pairs of subspaces $\frac{r_i+r_j}{2} < r_{ij}$, then the upper bound to the probability of misclassification (and the probability of misclassification) does not exhibit an error floor. In addition, in view of the fact that

$$d = \min_{i \neq j} d_{ij} = \min_{i \neq j} (\min (m - k; r_{ij} - k)) = \min \left(m - k; \min_{i \neq j} r_{ij} - k \right) \quad (15.27)$$

we can also conclude that the overall diversity gain associated with the (upper bound to) the probability of misclassification is a function of the interplay between the measurements and the geometry of the worst (closest) pair of subspaces, i.e. the pair of subspaces i^* and j^* such that $\{i^*, j^*\} = \operatorname{argmin}_{i \neq j} r_{ij}$ or equivalently $\{i^*, j^*\} = \operatorname{argmin}_{i \neq j} no_{ij}$. Concretely,

- by gradually increasing the number of measurements m from $k + 1$ to $\min_{i \neq j} r_{ij}$ it is possible to increase the diversity gain from 1 to $\frac{1}{2} \cdot \min_{i \neq j} no_{ij}$; in fact, as m ranges from $k + 1$ to $\min_{i \neq j} r_{ij}$ we can gradually unveil the degree of non-overlap between the original subspaces associated with the worst pair of classes because the number of non-overlapping dimensions in the projected domain gradually approaches the number of non-overlapping dimensions in the original domain, achieving it when $m = \min_{i \neq j} r_{ij}$;
- however, by increasing the number of measurements m beyond $\min_{i \neq j} r_{ij}$ it is not possible to increase the diversity gain any further; here, for $m > \min_{i \neq j} r_{ij}$ the projection of the subspaces from the original space to the measurement space does not result in an increase in the number of original non-overlapping dimensions;
- the degree of overlap between the closest pair of subspaces then defines a threshold on the number of measurements for the performance of classification—measured via the diversity gain—in the measurement domain to equal that in the data domain. One then understands the role of measurement as a way to probe the differences between the classes.

Notice also that in circumstances where the subspaces are uniformly distributed over the Grassmann manifold then $r_i = k, \forall i$, $r_{ij} = 2k, \forall i \neq j$, and $no_{ij} = 2k, \forall i \neq j$ with probability one, so that one can achieve k diversity gain with exactly $2k$ measurements. This corresponds to the diversity gain associated with zero discrimination gain offered by the DDT, which can be proven to be achievable by distributing the subspaces uniformly over the Grassmann manifold (with $m \geq 2k$).

Overall, the results suggest that the optimal geometry is such that the low-dimensional features constructed from the high-dimensional data ought to be associated with uniformly distributed subspaces on the Grassmann manifold. In practice, the role of the engineer is to learn such features from data that strike a balance between two objectives: i) nearly optimal discrimination and diversity geometry, consistent with the DDT; and ii) data representation fidelity. Constructions of nearly optimal Grassmannian packings for multiple-antenna non-coherent wireless communication problems may thus also inform how to learn nearly optimal features from data for compressive classification problems [2, 4–6, 8, 22, 23].

15.5 Results

We now explore how our theory aligns with practice in a concrete real-world application. In particular, we consider a face recognition problem where the orientation of the face relative to the camera remains fixed, but the illumination of the face varies. By supposing the faces themselves to be approximately convex and to reflect light according to Lambert's law, it is shown via spherical harmonics that the set of images of an individual face is well approximated by a nine-dimensional subspace [7]. Therefore, the face recognition task reduces to a (9-dimensional) subspace classification task.

It is only natural to ask whether our analysis can inform system design for such a real-world problem. Our classification experiments are based on 38 256-dimensional cropped faces from the Extended Yale Face Database B. This database, which is described in [19, 27], contains a few dozen greyscale photographs under a variety of illumination conditions for each face (see Fig. 15.2). We conduct the experiments by using two different data models, an approximately low-rank model (associated with the original dataset) and an exactly low-rank model (associated with a manipulation of the original dataset), that are learnt from a training set as follows:

1. *Approximately low-rank model*—The approximately low-rank model is obtained directly from the training set by learning the covariances associated with each face based on the set of face realizations under the different illumination conditions, using standard maximum likelihood (ML) estimators;
2. *Exactly low-rank model*—The exactly low-rank model is obtained from the approximately low-rank model as follows: *i*) we obtain new covariance matrices from the original ones by retaining only the nine largest eigenvalues; *ii*) we then project the faces onto the nine-dimensional subspace spanned by the eigenvectors associated with the nine largest eigenvalues of the corresponding class conditioned covariance.



Fig. 15.2 Two sample images from the Extended Yale Face Database B. These images are associated with the same face under different illumination conditions.

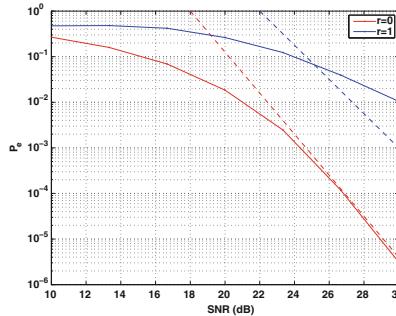


Fig. 15.3 Probability of misclassification versus SNR for a discrimination gain of $r = 0$ and a discrimination gain of $r = 1$ and for $m = 15$ measurements (solid lines: simulation; dashed lines: theory's slope).

We also conduct the experiments by proceeding as follows: for either data model, we randomly select a face from a testing set, vectorize it, linearly project it using a realization of an isotropically distributed measurement matrix, add i.i.d. Gaussian noise with the appropriate signal-to-noise ratio, and classify it using a MAP classifier that assumes that the class conditioned distributions are Gaussian with covariances corresponding to the learnt ones.⁶

Fig. 15.3 shows the probability of misclassification versus the signal-to-noise ratio for $r = 0$ and $r = 1$ discrimination gains and for $m = 15$ measurements, for the exactly low-rank model. Note that the number of classes (faces) in the classification problem scales with the signal-to-noise ratio as per the discrimination gain. Both theoretical and experimental results show that, as expected, the diversity gain is a function of the discrimination gain. In particular, theory predictions are aligned with the practical results: theory predicts a maximum diversity gain equal to $d = 9$ and $d = 9 \cdot (1 - 1/(m - k)) = 15/2$ for $r = 0$ and $r = 1$, respectively. For $r = 0$, experimental performance, as seen by the slope of the error curve, indeed exhibits the expected diversity order. For $r = 1$, experimental performance is consistent with theory in that the slope of the error curve is shallower than for $r = 0$; however, for the signal-to-noise ratios considered in our experiments, the error performance does not decay as fast as predicted. Because the number of classes increases so sharply in the SNR, and because there are only 38 faces in the database to classify, it is difficult to test high-SNR performance for nonzero discrimination gain.

Fig. 15.4 shows the probability of misclassification versus the signal-to-noise ratio for a fixed number of 38 classes (null discrimination gain) for $5 \leq m \leq 20$ measurements, both for the exactly low-rank model and the approximately low-rank

⁶Note that this classifier is mismatched in view of the fact that the class conditioned distributions are not necessarily Gaussian. In fact, it is immediate to demonstrate that face samples within each class do not pass the Royston's multivariate normality test [41], as they return p-values below 10^{-3} for all classes.

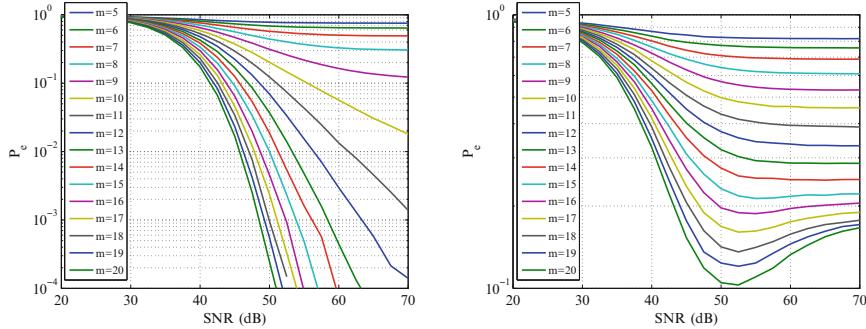


Fig. 15.4 Probability of misclassification versus SNR for a fixed number of 38 classes (null discrimination gain) for $5 \leq m \leq 20$ measurements, for exactly low-rank model (left) and approximately low-rank model (right).

model. Note now that the number of classes (faces) in the classification problem does not scale with the signal-to-noise ratio. Consider first the performance associated with the exactly low-rank model. It is possible to show—via direct examination of the data—that the dimension of the direct sum of pairs of nine-dimensional subspaces is equal to 18, i.e. the number of non-overlapping dimensions between pairs of nine-dimensional subspaces is equal to 18.⁷ It turns out that the experimental results are therefore consistent with theoretical ones. By inspecting the figure one concludes that: *i*) the maximum diversity gain is approximately below 9 always, as suggested by theory; *ii*) increasing the number of measurements from $m = 9$ to $m = 18$ results approximately in increases in the diversity gain from $d = 0$ to $d = 9$, in steps equal to 1 per additional measurement; *iii*) however, increasing the number of measurements beyond $m = 18$ does not result in additional marked increases in the diversity gain. Theory can also show—although not done here—that it suffices to take $m > k = 9$ measurements to eliminate the probability of misclassification floor. This is also confirmed by Fig. 15.4.

Consider now the performance associated with the approximately low-rank model. It can also be shown that, although approximately low-rank, the dimension of the individual subspaces and the dimension of the direct sum of pairs of subspaces now tends to be equal to the dimension of the ambient space, so that the misclassification probability also tends to exhibit an error floor.

However, we argue that the theory can still inform practice. In view of the fact that it is typical to describe the covariance matrices associated with an approximately low-rank model Σ_i in terms of the covariance matrices associated with the exactly low-rank model $\bar{\Sigma}_i$ as $\Sigma_i = \bar{\Sigma}_i + \varepsilon I$, where the matrix εI accounts for deviations between the approximately low-rank and the exactly low-rank model [13, 39],

⁷Note that this suggests that nature may tend to approximately distribute the subspaces uniformly on the Grassmann manifold.

then the measurement model in (15.1) with noise given by $z \sim \mathcal{N}(0, \sigma^2 I)$ and an approximately low-rank model with covariances $\Sigma_i, i = 1, \dots, L$ is mathematically equivalent to the measurement model in (15.1) with noise given by $z \sim \mathcal{N}(0, \sigma^2 I + \varepsilon \Phi \Phi^T)$ and an exactly low-rank model with covariances $\bar{\Sigma}_i, i = 1, \dots, L$. The misclassification performance associated with the approximately low-rank model with noise level $1/\sigma^2$ then corresponds to the misclassification performance associated with the exactly low-rank model with noise level $1/(\sigma^2 + \varepsilon)$ since $\Phi \Phi^T = I$ for a randomly generated Haar measurement matrix. By letting $\sigma^2 \rightarrow 0$ then the diversity gain informs how low the misclassification probability floor is and how fast the misclassification probability associated with the approximately low-rank model tends to the misclassification probability associated with the exactly low-rank model at $\text{SNR} = 1/\varepsilon$. It is also clear that the better the exactly low-rank model fits the approximately low-rank one (i.e., the lower the ε), the better our theory aligns with practice.

15.6 Conclusions

The focus has been on the classification of low-dimensional subspaces embedded in a high-dimensional ambient space from compressive and noisy measurements. By unveiling dualities between the compressive classification of low-rank Gaussian distributions and the communication of codewords over multiple-antenna fading non-coherent channels, we have argued that it is possible to import methods from the wireless information theory and the wireless communications theory domains to the compressive classification domain in order to unveil fundamental performance characterizations and geometries for classification problems [32, 36]. The DDT—a construction reminiscent of the DMT in wireless communications—is one such characterization that articulates about the tradeoff between the number of classes that can be separated by a compressive classifier and the performance of such a classifier: its relevance derives from the fact that it informs an “engineer” about ultimate performance limits and about how to learn features from data with nearly optimal discrimination and diversity geometries.

The connections and dualities between problems in the wireless communications domain and problems in the sensing and machine learning domains remain largely unexplored. We believe that Shannon theory, in the same way that has played an important role in unveiling fundamental limits and optimal designs for wireless communications systems in the past few decades, will also likely cast further insight in the limits of sensing, analysis, and processing of high-dimensional data in coming years.

Acknowledgements This work was supported by the Royal Society International Exchanges Scheme IE120996. The work of Robert Calderbank and Matthew Nokleby is also supported in part by the Air Force Office of Scientific Research under the Complex Networks Program.

References

1. Adini, Y., Moses, Y., Ullman, S.: Face recognition: the problem of compensating for changes in illumination direction. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(7), 721–732 (1997)
2. Aggarwal, V., Ashikhmin, A., Calderbank, R.: A Grassmannian packing based on the Nordstrom-Robinson code. In: *IEEE Information Theory Workshop*, Chengdu, China, pp. 1–5 (2006)
3. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.* **96**(12), 6745–6750 (1999)
4. Ashikhmin, A., Calderbank, R.: Space-time Reed-Muller codes for noncoherent MIMO transmission. In: *IEEE International Symposium on Information Theory*, Adelaide, Australia, 1952–1956 (2005)
5. Ashikhmin, A., Calderbank, R., Kewlin, W.: Multidimensional second order Reed-Muller codes as Grassmannian packings. In: *IEEE International Symposium on Information Theory*, Seattle, WA, USA, pp. 1001–1005 (2006)
6. Ashikhmin, A., Calderbank, R.: Grassmannian packings from operator Reed-Muller codes. *IEEE Trans. Inf. Theory* **56**(10), 5689–5714 (2010)
7. Basri, R., Jacobs, D.W.: Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(2), 218–233 (2003)
8. Calderbank, R., Hardin, R.H., Rains, E.M., Shor, P.W., Sloane, N.J.A.: A group-theoretic framework for the construction of packings in Grassmannian spaces. *J. Algebraic Comb.* **9**, 129–140 (1999)
9. Candès, E., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**(2), 489–509 (2006)
10. Candès, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59**(8), 1207–1223, 2006.
11. Candès, E., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inf. Theory* **52**(12), 5406–5425 (2006)
12. Carson, W.R., Chen, M., Rodrigues, M.R.D., Calderbank, R., Carin, L.: Communications-inspired projection design with application to compressive sensing. *SIAM J. Imaging Sci.* **5**(4), 1185–1212 (2012)
13. Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D., Carin, L.: Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: algorithm and performance bounds. *IEEE Trans. Signal Process.* **58**(12), 6140–6155 (2010)
14. Chen, M., Carson, W.R., Rodrigues, M.R.D., Calderbank, R., Carin, L.: Communication-inspired linear discriminant analysis. In: *International Conference on Machine Learning*, Edinburgh, UK, pp. 919–926 (2012)
15. Davenport, M., Boufounos, P., Wakin, M., Baraniuk, R.: Signal processing with compressive measurements. *IEEE J. Sel. Top. Sign. Process.* **4**(2), 445–460 (2010)
16. Donoho, D.: Compressed sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006)
17. Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification*, 2nd edn. Wiley-Interscience, New York, NY (2000)
18. Elhamifar, Vidal, R.: Sparse subspace clustering: algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(11), 2765–2781 (2013)
19. Georgiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(6), 643–660 (2001)
20. Hastie, T., Simard, P.Y.: Metrics and models for handwritten character recognition. *Stat. Sci.* **13**(1), 54–65 (1998)
21. Hild, K., Erdogmus, D., Torkkola, K., Principe, J.: Feature extraction using information-theoretic learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(9), 1385–1392 (2006)

22. Hochwald, B.M., Marzetta, T.L.: Unitary space-time modulation for multiple-antenna communications in Rayleigh flat fading. *IEEE Trans. Inf. Theory* **46**(2), 543–564 (2000)
23. Hochwald, B.M., Marzetta, T.L., Richardson, T.J., Sweldens, W., Urbanke, R.: Systematic design of unitary space-time constellations. *IEEE Trans. Inf. Theory* **46**(6), 1962–1973 (2000)
24. Hull, J.J.: A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.* **16**(5), 550–554 (1994)
25. Kaski, S., Peltonen, J.: Informative discriminant analysis. In: International Conference on Machine Learning, Washington, DC, USA, pp. 329–336 (2003)
26. LeCun, Y., Jackel, L., Bottou, L., Brunot, A., Cortes, C., Denker, J., Drucker, H., Guyon, I., Muller, U., Sackinger, E., Simard, P., Vapnik, V.: Comparison of learning algorithms for handwritten digit recognition. In: International Conference on Artificial Neural Networks, Warsaw, Poland, pp. 53–60 (1995)
27. Lee, K.-C., Ho, J., Kriegman, D.J.: Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(5), 684–698 (2005)
28. Liu, G., Lin, Z., Yu, Y.: Robust subspace segmentation by low-rank representation. In: International Conference on Machine Learning, Haifa, Israel, pp. 663–670 (2010)
29. Liu, L., Fieguth, P.: Texture classification from random features. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(3), 574–586 (2012)
30. Marzetta, T.L., Hochwald, B.M.: Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading. *IEEE Trans. Inf. Theory* **45**(1), 139–157 (1999)
31. Nenadic, Z.: Information discriminant analysis: feature extraction with an information-theoretic objective. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(8), 1394–1407 (2007)
32. Nokleby, M., Rodrigues, M.R.D., Calderbank, R.: Discrimination on the Grassmann manifold: fundamental limits of subspace classifiers. Available at <http://arxiv.org/abs/1404.5187>
33. Qiu, Q., Sapiro, G.: Learning robust subspace clustering. Available at <http://arxiv.org/abs/1308.0273>
34. Qiu, Q., Sapiro, G.: Learning transformations for clustering and classification. Available at <http://arxiv.org/abs/1309.2074>
35. Qiu, Q., Sapiro, G.: Learning transformations for classification forests. Available at <http://arxiv.org/abs/1312.5604>
36. Reboredo, H., Renna, F., Calderbank, R., Rodrigues, M.R.D.: Compressive classification of a mixture of Gaussians: analysis, designs and geometrical interpretation. Available at <http://arxiv.org/abs/1401.6962>
37. Reeves, G., Gastpar, M.: The sampling rate distortion tradeoff for sparsity pattern recovery in compressed sensing. *IEEE Trans. Inf. Theory* **58**(5), 3065–3092 (2012)
38. Reeves, G., Gastpar, M.: Approximate sparsity pattern recovery: information-theoretic lower bounds. *IEEE Trans. Inf. Theory* **59**(6), 3451–3465 (2013)
39. Renna, F., Calderbank, R., Carin, L., Rodrigues, M.R.D.: Reconstruction of signals drawn from a Gaussian mixture from noisy compressive measurements. *IEEE Trans. Signal Process.* **62**(9), 2265–2277 (2014)
40. Ross, D.T., et al.: Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* **24**(3), 227–235 (2000)
41. Royston, J.P.: Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *Appl. Stat.* **32**(2), 121–133 (1983)
42. Soltanolkotabi, M., Candès, E.: A geometric analysis of subspace clustering with outliers. *Ann. Stat.* **40**(4), 2195–2238 (2012)
43. Soltanolkotabi, M., Elhamifar, E., Candès, E.: Robust subspace clustering. Available at arxiv.org/abs/1301.2603
44. Tao, D., Li, X., Wu, X., Maybank, S.: Geometric mean for subspace selection. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(2), 260–274 (2009)
45. Torkkola, K.: Learning discriminative feature transforms to low dimensions in low dimensions. In: Advances in Neural Information Processing Systems, Vancouver, Canada, pp. 969–976 (2001)

46. Torkkola, K.: Feature extraction by non-parametric mutual information maximization. *J. Mach. Learn. Res.* **3**, 1415–1438 (2003)
47. Tenenbaum, J., de Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. *Science* **290**(5500), 2319–2323 (2000)
48. Tulino, A., Caire, G., Verdú, S., Shamai, S.: Support recovery with sparsely sampled free random matrices. *IEEE Trans. Inf. Theory* **59**(7), 4243–4271 (2013)
49. Wainwright, M.: Sharp thresholds for high-dimensional and noisy sparsity recovery using 11-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theory* **55**(5), 2183–2202 (2009)
50. Wainwright, M.: Information-theoretic limits on sparsity recovery in the high dimensional and noisy setting. *IEEE Trans. Inf. Theory* **55**(12), 5728–5741 (2009)
51. Wang, W., Wainwright, M., Ramchandran, K.: Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Trans. Inf. Theory* **56**(6), 2967–2979 (2010)
52. Wang, L., Carlson, D., Rodrigues, M.R.D., Wilcox, D., Calderbank, R., Carin, L.: Designed measurements for vector count data. In: *Advances in Neural Information Processing Systems*, Lake Tahoe, NV, USA, pp. 1142–1150 (2013)
53. Wang, L., Razi, A., Rodrigues, M.R.D., Calderbank, R., Carin, L.: Nonlinear information-theoretic compressive measurement design. In: *International Conference on Machine Learning*, Beijing, China, pp. 1161–1169 (2014)
54. Wang, Y., Xu, H.: Noisy sparse subspace clustering. In: *International Conference on Machine Learning*, Atlanta, GA, USA, pp. 89–97 (2013)
55. Wright, J., Yang, M., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(2), 210–227 (2009)
56. Zheng, L., Tse, D.: The diversity-multiplexing tradeoff for non-coherent multiple antenna channels. In: *The Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, USA, pp. 1011–1020 (2002)
57. Zheng, L., Tse, D.: Communication on the Grassmann manifold: a geometric approach to the noncoherent multiple-antenna channel. *IEEE Trans. Inf. Theory* **48**(2), 359–383 (2002)
58. Zheng, L., Tse, D.: Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels. *IEEE Trans. Inf. Theory* **49**(5), 1073–1096 (2003)

Applied and Numerical Harmonic Analysis (69 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)
- O. Bratteli, P. Jorgensen, and B. Treadway: *Wavelets Through a Looking Glass* (ISBN 978-0-8176-4280-80)
- H.G. Feichtinger and T. Strohmer: *Advances in Gabor Analysis* (ISBN 978-0-8176-4239-6)
- O. Christensen: *An Introduction to Frames and Riesz Bases* (ISBN 978-0-8176-4295-2)
- L. Debnath: *Wavelets and Signal Processing* (ISBN 978-0-8176-4235-8)
- G. Bi and Y. Zeng: *Transforms and Fast Algorithms for Signal Analysis and Representations* (ISBN 978-0-8176-4279-2)
- J.H. Davis: *Methods of Applied Mathematics with a MATLAB Overview* (ISBN 978-0-8176-4331-7)
- J.J. Benedetto and A.I. Zayed: *Modern Sampling Theory* (ISBN 978-0-8176-4023-1)
- E. Prestini: *The Evolution of Applied Harmonic Analysis* (ISBN 978-0-8176-4125-2)
- L. Brandolini, L. Colzani, A. Iosevich, and G. Travaglini: *Fourier Analysis and Convexity* (ISBN 978-0-8176-3263-2)
- W. Freeden and V. Michel: *Multiscale Potential Theory* (ISBN 978-0-8176-4105-4)
- O. Christensen and K.L. Christensen: *Approximation Theory* (ISBN 978-0-8176-3600-5)
- O. Calin and D.-C. Chang: *Geometric Mechanics on Riemannian Manifolds* (ISBN 978-0-8176-4354-6)
- J.A. Hogan and J.D. Lakey: *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 G. 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)
- R. Balan, M. Begue, J. Benedetto, W. Czaja, and K.A. Okoudjou: *Excursions in Harmonic Analysis, Volume 3* (ISBN 978-3-319-13229-7)

- 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* (ISBN 978-3-319-08800-6)
- R. Balan, M. Begué, J.J. Benedetto, W. Czaja, and K.A. Okoudjou: *Excursions in Harmonic Analysis, Volume 3: The February Fourier Talks at the Norbert Wiener Center* (ISBN 978-3-319-13229-7)
- S. Dahlke, F. De Mari, P. Grohs, and D. Labate: *Harmonic and Applied Analysis: From Groups to Signals* (ISBN 978-3-319-18862-1)
- H. Boche, R. Calderbank, G. Kutyniok, and J. Vybíral: *Compressed Sensing and its Applications: MATHEON Workshop 2013* (ISBN 978-3-319-16041-2)

For an up-to-date list of ANHA titles, please visit <http://www.springer.com/series/4968>