Quantized Compressive Sampling for Structured Signal Estimation

Von der Fakultät für Elektrotechnik und Informationstechnik
der Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

genehmigte Dissertation

vorgelegt von

Niklas Koep, M.Sc.

aus Birkesdorf

Berichter: Univ.-Prof. Dr. rer. nat. Rudolf Mathar
Univ.-Prof. Dr. rer. nat. Holger Rauhut

Tag der mündlichen Prüfung: 05. Juni 2019

Niklas Koep
Quantized Compressive Sampling for Structured Signal Estimation

ISBN: 978-3-95886-291-3
1. Auflage 2019

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über www.dnb.ddb.de abrufbar.

Vertrieb:

© Verlagshaus Mainz GmbH Aachen
Süsterfeldstr. 83, 52072 Aachen
Tel. 0241 / 87 34 34 00
www.Verlag-Mainz.de

Herstellung:

Druckerei Mainz GmbH Aachen
Süsterfeldstraße 83
52072 Aachen
www.DruckereiMainz.de

Satz: nach Druckvorlage des Autors
Umschlaggestaltung: Jennifer Flohr

printed in Germany

D82 (Diss. RWTH Aachen University, 2019)

Abstract

This thesis investigates different approaches to enable the use of *compressed sensing* (CS)-based acquisition devices in resource-constrained environments relying on cheap, energy-efficient sensors. We consider the acquisition of structured *low-complexity* signals from excessively quantized 1-bit observations, as well as *partial* compressive measurements collected by one or multiple sensors. In both scenarios, the central goal is to alleviate the complexity of sensing devices in order to enable signal acquisition by simple, inexpensive sensors.

In the first part of the thesis, we address the reconstruction of signals with a sparse Fourier transform from 1-bit time domain measurements. We propose a modification of the *binary iterative hard thresholding* algorithm, which accounts for the conjugate symmetric structure of the underlying signal space. In this context, a modification of the *hard thresholding operator* is developed, whose use extends to various other (quantized) CS recovery algorithms. In addition to undersampled measurements, we also consider oversampled signal representations, in which case the measurement operator is deterministic rather than constructed randomly. Numerical experiments verify the correct behavior of the proposed methods.

The remainder of the thesis focuses on the reconstruction of *group-sparse signals*, a signal class in which nonzero components are assumed to appear in nonoverlapping coefficient groups. We first focus on 1-bit quantized Gaussian observations and derive theoretical guarantees for several reconstruction schemes to recover target vectors with a desired level of accuracy. We also address recovery based on dithered quantized observations to resolve the scale ambiguity inherent in the 1-bit CS model to allow for the recovery of both direction and magnitude of group-sparse vectors.

In the last part, the acquisition of group-sparse vectors by a collection of independent sensors, which each observe a different portion of a target vector, is considered. Generalizing earlier results for the canonical sparsity model, a bound on the number of measurements required to allow for stable and robust signal recovery is established. The proof relies on a powerful concentration bound on the suprema of chaos processes. In order to establish our main result, we develop an extension of *Maurey's empirical method* to bound the covering number of sets which can be represented as convex combinations of elements in compact convex sets.

Preface

This thesis was written during my time as a research assistant at the Institute for Theoretical Information Technology of RWTH Aachen University.

I would like to sincerely thank my supervisor Prof. Dr. Rudolf Mathar for giving me the opportunity to pursue a doctoral degree at his institute and for allowing me to always follow my own research interests. I also want to express my deepest gratitude to Prof. Dr. Holger Rauhut for agreeing to act as second reviewer of my thesis. Moreover, I am grateful to Prof. Dr. Peter Vary and Prof. Dr. Dorit Merhof for serving as members of my defense committee.

A very special thank you also goes to my former and present colleagues of TI, whose company I cherished both academically and socially in my five years at the institute. My roommate Arya Bangun, who I shared an office with for almost four years, deserves special mention for always maintaining a cheerful attitude even at times when mine was anything but. I am deeply indebted to my friend and colleague Arash Behboodi, who always had an open door for me and without whom this thesis might never have been written. I will forever be grateful for his unwavering support, optimism and seemingly endless patience.

Mein größter Dank gilt schließlich meiner Familie, insbesondere meinen Eltern Richard und Maria Koep, sowie meiner Schwester Jessica. Nur durch eure uneingeschränkte Unterstützung und Ermutigung während meines Studiums und meiner Promotion habe ich es bis zu diesem Punkt geschafft. Ich verdanke euch alles.

#happytoyou

Aachen, July 2019 Niklas Koep

Contents

List of Figures

1

Introduction

A fundamental challenge in contemporary data and signal processing applications is the efficient acquisition of exceedingly high-dimensional signals carrying only a limited amount of information. This information is usually captured by some low-complexity structure inherent to a particular signal class. The most common low-complexity structure by far manifests itself in the form of *sparsity* of finite-dimensional vectors in a suitable basis or more generally an overcomplete dictionary [**RSV08**] or frame [**CK12**]. The field of *compressed sensing* (CS) emerged from the very idea that the number of samples required to acquire and reconstruct such signals should be on the order of the information-theoretic rather than the linear-algebraic dimension of the ambient signal space. This was the result of a series of landmark papers due to Candès, Tao, Romberg [**CT06a**; **CT05**; **CRT06b**; **CRT06a**; **CT06b**] and Donoho [**Don06c**]. Their seminal works showed that every vector $\mathbf{x} \in \mathbb{C}^d$ containing at most s nonzero coefficients can be perfectly reconstructed from $m = \Omega(s \log(d/s))$ linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x}$, provided that the measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times d}$ satisfies certain structural conditions. Unfortunately, the deterministic construction of such matrices with provably optimal scaling in terms of the information dimension s remains a yet unsolved problem. The situation changes drastically, however, if one turns to random measurement ensembles. In this case, it can be shown that certain random matrices capture just enough information about sparse signals to allow for them

to be reconstructed by tractable and efficient recovery schemes with overwhelmingly high probability. In the years since its inception, a vibrant research activity has developed CS into an elegant and mature theory at the intersection of applied mathematics and engineering. Today, concepts from CS already have far-reaching consequences in fields like diagnostic imaging, where compressed sensing is now actively used to provide significant scan time reductions in *magnetic resonance imaging* (MRI) systems [**Sie16**; **Gen17**; **Phi18**].

A common thread throughout the canonical theory of compressed sensing is the assumption that measurements are essentially available at infinite precision during reconstruction. Given the prevalence of digital signal processing, however, practical CS-based systems also have to take into consideration the effects of *analog-to-digital converters* (ADCs) used to map continuous measurements to the digital domain by means of *quantization*. While the canonical compressed sensing theory accounts for both additive noise and model uncertainties of the signal class, such robustness properties are generally not sufficient to appropriately model signal-dependent quantization noise introduced by ADCs operating at lower bit depths. To reduce the influence of quantization noise, classical quantization theory therefore employs the idea of *oversampled representations*, where signals are sampled at super-Nyquist rates. Since this increases the signal bandwidth while keeping the quantization noise fixed, the approach can be used to hide quantization noise in less relevant frequency bands by means of *noise shaping*.

The notion of oversampling in compressive sampling systems is slightly at odds with the key idea of reducing the number of samples by exploiting signal sparsity for reconstruction from undersampled representations without information loss. Recent years have therefore seen an increase in research efforts to establish both specially tailored quantization-aware compressive sampling schemes, as well as efficient reconstruction algorithms to close the gap between theoretical and practical benefits of CS-based acquisition systems. One particular line of research, which has attracted particular interest due to its favorable implications in terms of hardware complexity and energy efficiency, is the so-called 1-*bit compressed sensing* measurement model. In this sampling scheme, the only information retained about the linear measurements $\mathbf{Ax} = (\langle \mathbf{a}_i, \mathbf{x} \rangle)_{i=1}^m$ is the sign of each component $\langle \mathbf{a}_i, \mathbf{x} \rangle$. Despite this excessive quantization strategy, it has been demonstrated that faithful signal estimation is still possible within limits when target signals are sparse. The appeal of this approach is threefold. First, the simplicity of the 1-bit quantization operation allows for the utilization of cheap and energy-efficient sampling devices based on comparators operating at fixed voltage levels. Secondly, the energy-efficient nature of the quantizer enables ADCs to possibly operate at super-Nyquist rates while still keeping the total bit budget below comparable sampling devices with higher bit depths. Thirdly, the highly nonlinear nature of the sign function renders 1-bit quantizers impervious to certain monotone nonlinearities of the acquisition system. Additionally, pre-quantization noise is only registered by a memoryless 1-bit sampling device if additive perturbations of the linear measurements before quantization are significant enough to result in an effective bit flip in the quantized measurements.

To address the growing demand for efficient and low-cost sensing devices, which are of central importance in technological advances such as the *internet of things* (IoT) and *Industry 4.0*, we consider various approaches of structured signal recovery with a focus on energy efficiency. To that end, we investigate different compressive acquisition paradigms exploiting prior knowledge about certain structural properties of the class of target signals

or the underlying measurement model. In particular, we consider the recovery of signals with a sparse Fourier transform from binary measurements, as well as the reconstruction of group-sparse signals from both quantized and unquantized partial observations. The Fourier transform represents a ubiquitous sparsity basis in various domains such as wireless communication, radar localization, medical imaging and speech recognition. Similarly, group-sparse signal structures are frequently employed to model clustered sparsity phenomena, which typically arise in certain areas of wireless communication, medical and natural imaging, genetics and facial recognition.

In the following section, we give an overview of the work presented in this thesis. We then highlight the main contributions of each chapter in Section 1.2.[1]

1.1 Thesis Outline

We begin by reviewing some of the central results in the theory of compressed sensing in Chapter 2. We also briefly discuss the comparatively young area of quantized compressed sensing and conclude the background chapter with an overview of some key results in the 1-bit compressed sensing literature.

We then consider the problem of estimating signals with a frequency-sparse representation from 1-bit quantized time domain measurements in Chapter 3. Rather than relying on purely random measurement ensembles drawn, *e.g.*, from the Gaussian distribution, we instead consider structured measurement matrices based on randomly subsampled discrete Fourier systems. We review the so-called *binary iterative hard thresholding* (BIHT) algorithm and discuss the necessary modifications to extend the algorithm to the setting of conjugate symmetric sparse signal recovery. We also consider the reconstruction from oversampled binary measurements, which requires a specialized construction of the measurement matrix that does not rely on random subsampling. The chapter concludes with a numerical study benchmarking the performance of the proposed algorithms against convex programming techniques for 1-bit signal recovery, which we adopt for our purposes.

In Chapter 4, we turn our attention to a different low-complexity signal class characterized by signal coefficients which appear in nonoverlapping groups, giving rise to the so-called *group-sparsity* model. This signal class forms the basis for the remainder of the thesis. After introducing the particulars of the signal model, we analyze three recovery procedures to estimate the direction of group-sparse signals from 1-bit observations of Gaussian projections. We then compare their empirical performance and benchmark each method in various scenarios against regular 1-bit recovery schemes, which ignore the underlying group structure of the signal class.

While the quantization scheme considered in Chapter 3 and the first half of Chapter 4 is scale-invariant, allowing only for the estimation of signals up to a global scale factor, we also consider the problem of recovery from *dithered* observations. In this model, one intentionally introduces a known noise-like offset to the linear measurements prior to quantization. Surprisingly, this common extension to the 1-bit quantization model provides sufficient additional information about the measurement process to allow for both direction and norm estimation subject to an a priori norm constraint on the signal class. In this context, we present and analyze six different recovery schemes, which relate the problem of

[1] Parts of this thesis and related works have previously appeared in [**KM17**; **KM18**] and [**KBM19a**], while [**KBM19b**] has been accepted for publication prior to the preparation of this dissertation.

group-sparse vector recovery to the task of direction recovery in a lifted signal space. We conclude the chapter with an empirical study of the numerical behavior of the considered recovery schemes.

Dispensing with quantized observations in Chapter 5, we investigate a measurement model for the compressive acquisition of high-dimensional group-sparse vectors intended to reduce the energy consumption of sensors in a distributed setting. More precisely, we consider a measurement model in which one or more sensors observe distinct portions of a group-sparse target vector. Such acquisition systems can be conveniently modeled by block diagonal measurement matrices where the blocks are either identical or independent copies of a prototype random matrix. Following a common narrative in compressed sensing, we establish recovery guarantees for the respective acquisition models by appealing to a group-sparse variant of the *restricted isometry property* (RIP). This is then used to establish stable and robust recovery results for group-sparse vector recovery. We first consider block diagonal measurement matrices with independent subgaussian blocks and reformulate the condition that such matrices satisfy the group-sparse restricted isometry property in terms of the supremum of a particular chaos process. We then discuss how to bound certain geometric objects associated with said chaos process in order to derive a condition on the number of measurements for the group-sparse RIP to hold with high probability. The analysis is then repeated with a few minor modifications for measurement matrices with identical copies of a single subgaussian random matrix. After relating our obtained bounds to results in the literature, we close out the chapter by numerically verifying the predicted recovery behavior with a series of phase transition diagrams.

In Chapter 6, we finally conclude the thesis with a summary of the main results presented in this work, including a discussion of open problems and suggested future research directions.

1.2 Contributions

In this section, we briefly summarize the main contributions of each chapter.

Chapter 3: Estimation of Frequency-Sparse Signals from Binary Measurements

While a considerable amount of research exists which deals with the recovery of sparse vectors from 1-bit measurements of Gaussian projections, the body of work on more structured measurement matrices is severely limited. To address this issue, the main goal of Chapter 3 is to establish through a series of numerical experiments that recovery of sparse signals from structured observations is in fact possible. Due to the central importance of the Fourier basis in engineering domains, we focus our attention on measurement matrices consisting of randomly subsampled discrete Fourier transform matrices. More precisely, we consider the recovery of conjugate symmetric sparse vectors from 1-bit measurements of time domain signals.

While various methods for the recovery of real-valued sparse signals from compressive 1-bit measurements have been proposed in the literature, the BIHT algorithm remains one of the most accurate ones to date. In order to make sense of the single-bit quantization operation during reconstruction, we modify the BIHT algorithm to account for the particular structure of the underlying conjugate symmetric signal space. As a necessary

byproduct, we propose a variation of the so-called *hard thresholding operator* to project vectors on the set of conjugate symmetric sparse vectors. In addition to the BIHT algorithm, this also enables other well-known algorithms proposed for linear compressed sensing such as the *iterative hard thresholding* (IHT) [**BD09**], *hard thresholding pursuit* (HTP) [**Fou11**] and *compressive sampling matching pursuit* (CoSaMP) [**NT09**] algorithms to be used for the recovery of frequency-sparse signals from real-valued compressive time domain measurements.

While we first consider measurements based on random subsampling in discrete Fourier systems, we also extend the acquisition model to oversampled time domain representations. Assuming a fixed signal bandwidth, the idea is to remove those frequency coefficients from the vector of *discrete Fourier transform* (DFT) coefficients which correspond to frequencies we know to be absent in the target signal. To that end, we propose to model the associated measurement matrix based on the idea of *exact interpolation*, a frequency domain zero-padding scheme for interpolation in the time domain. Considering that the resulting measurement operator is purely deterministic, the proposed sampling scheme is particularly hardware-friendly as it does not rely on any form of randomness in the acquisition process. While no theoretical results concerning the reconstruction fidelity are established, we present a series of numerical experiments which validate the correct behavior of the proposed methods empirically.

Chapter 4: Single-Bit Group-Sparse Signal Recovery

In Chapter 4, we address the problem of group-sparse signal estimation from 1-bit quantized Gaussian projections. This particular variation of the canonical sparsity model has been extensively studied in the context of linear compressed sensing, as well as sparse model selection and regression in the statistics literature. However, the model has not yet been thoroughly studied within the framework of 1-bit compressed sensing.

In the first part of Chapter 3, we aim to fill this gap in the literature by establishing recovery guarantees for three different reconstruction schemes modeled after existing methods for sparse recovery from binary measurements. While the general analysis strategy for these approaches carries over from the respective theory developed for the canonical sparsity model, we also point out some new connections. For instance, we establish a nonuniform recovery guarantee for a simple group hard thresholding algorithm with robustness to both additive pre-quantization and adversarial post-quantization noise. We also put a particular emphasis on numerical experiments to confirm some of the theoretical claims such as robustness to measurement noise, and to assess how close predicted error decay rates are to those observed empirically. Even in the canonical sparsity setting, such numerical investigations seem to be entirely absent from the literature.

In the second half of Chapter 4, we consider six different dithering-based reconstruction schemes, which we relate to direction recovery results established in the first half of the chapter. Again, we present novel noise robustness results for two group hard thresholding algorithms, as well as two convex approaches which maximize the correlation between quantized and unquantized observations. As in the first half of the chapter, we conduct a thorough numerical study of the proposed recovery schemes to confirm our theoretical results.

Chapter 5: Recovery of Group-Sparse Vectors with Block Diagonal Measurement Operators

In Chapter 5, we establish a particular variant of the restricted isometry property for block diagonal subgaussian random matrices. This is then used to provide bounds on the number of measurements required to guarantee robust and stable recovery of group-sparse vectors. As outlined in Section 1.1, establishing a group-sparse version of the RIP is achieved by relating the associated group-RIP constant to a particular chaos process. The general proof strategy follows the example of a closely related result for the canonical sparsity model [**Eft+15**] by appealing to a powerful tail bound for suprema of chaos processes. The main difficulty in this context is bounding Talagrand's γ_2-*functional* by means of a metric entropy integral, which in turn requires estimating the covering number of our signal set w. r. t. an induced norm depending on the acquisition model. At small scales, we bound the covering number by means of a standard volumetric estimate. At higher scales, however, this bound is not effective enough. Instead, one may appeal to alternative tools such as *Sudakov minoration* or *Maurey's empirical method*. As we will demonstrate in Chapter 5, the former technique provides the correct scaling of the number of measurements in terms of the group-sparsity level to establish our recovery result. The resulting bound, however, is independent of the underlying sparsity basis. Due to the block diagonal nature of the acquisition model, this is clearly suboptimal. In order to resolve this issue, we develop an extension of *Maurey's empirical method*. While Maurey's method only applies to convex polytopes, our extension can be used to bound the covering number of sets whose elements can be represented as convex combinations of compact subsets. As in the canonical sparsity setting, our resulting bound on the number of measurements depends on a coherence-like parameter of the sparsity basis. In the most favorable scenario of sparsity in the discrete Fourier transform basis, we establish almost optimal scaling in the system parameters up to logarithmic factors. In this case, our result shows that the number of measurements per sensor can be reduced without affecting the reconstruction fidelity if more sensors are added to the system so that the total number of measurements remains fixed.

2

Background

In this chapter, we discuss some fundamental concepts in the theory of compressed sensing. We mainly limit ourselves to topics which bear direct relevance to the subsequent chapters of the thesis and refer interested readers to more general treatments of the subject in the literature such as [**Dav+12**; **Kut13**], as well as the extended monograph [**FR13**]. In the interest of keeping this introduction short, we delegate a discussion about general mathematical preliminaries to Appendix A. There we also collect a few common definitions and well-known results in probability theory, as well as convex and geometric functional analysis. We begin by fixing notation for the rest of the thesis.

2.1 Notation

Throughout this work, we denote matrices by uppercase boldface letters, vectors by lowercase boldface letters and scalars by regular type symbols. For an integer $d \in \mathbb{N}$, we use the common shorthand notation $[d] := \{1, \dots, d\} = [1, d] \cap \mathbb{N}$ and write $|U|$ for the cardinality of a set U. We use $\mathbb{1}_E$ to denote the binary indicator function of an event E with $\mathbb{1}_E = 1$ if E occurs and 0 otherwise. The support of a vector $\mathbf{x} \in \mathbb{C}^d$ is defined as $\operatorname{supp}(\mathbf{x}) := \{i \in [d] : x_i \neq 0\}$. Given a subset $S \subset U$, we denote the complement of S

w. r. t. U by $\overline{S} := U \setminus S$. Given a norm $\|\cdot\|_\theta$ on $\mathbb{C}^d$ depending on some abstract parameter θ, we write $\mathbb{B}_\theta^d$ for the norm ball associated with $\|\cdot\|_\theta$, *i.e.*, $\mathbb{B}_\theta^d := \{\mathbf{x} \in \mathbb{C}^d : \|\mathbf{x}\|_\theta \leq 1\}$. For $\theta = p$, we reserve the notation $\mathbb{B}_p^d$ to denote the closed unit balls of the family of ℓ_p-norms on $\mathbb{C}^d$ defined as

$$\|\mathbf{x}\|_p := \begin{cases} \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}, & 1 \leq p < \infty, \\ \max_{i \in [d]} |x_i|, & p = \infty. \end{cases}$$

Even though we mainly work in $\mathbb{C}^d$, we denote by $\langle \cdot, \cdot \rangle \colon \mathbb{C}^d \to \mathbb{C}^d$ the bilinear form defined as $\langle \mathbf{a}, \mathbf{b} \rangle := \sum_{i=1}^d a_i b_i$ for $\mathbf{a}, \mathbf{b} \in \mathbb{C}^d$. With this, the i-th entry of the matrix-vector product $\mathbf{G}\mathbf{a}$ for $\mathbf{G} \in \mathbb{C}^{m \times d}$ and $\mathbf{a} \in \mathbb{C}^d$ is $\langle \mathbf{g}_i, \mathbf{a} \rangle$ where $\mathbf{g}_i \in \mathbb{C}^d$ denotes the i-th row of $\mathbf{G}$. The canonical sesquilinear inner product on $\mathbb{C}^d$ is instead denoted by $\langle \cdot, \cdot \rangle_{\mathbb{C}}$, *i.e.*, $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbb{C}} := \sum_{i=1}^d a_i \overline{b_i}$. As such, the canonical ℓ_2-norm on $\mathbb{C}^d$ is induced by $\|\mathbf{a}\|_2^2 = \langle \mathbf{a}, \overline{\mathbf{a}} \rangle = \langle \overline{\mathbf{a}}, \mathbf{a} \rangle = \langle \mathbf{a}, \mathbf{a} \rangle_{\mathbb{C}}$ where $\overline{\mathbf{a}}$ denotes the complex conjugate of $\mathbf{a} \in \mathbb{C}^d$. We denote the (complex) unit Euclidean sphere in $\mathbb{C}^d$ by $\mathbb{S}^{d-1}$. The identity matrix on $\mathbb{C}^d$ is generally denoted by Id_d, but we sometimes simply write Id for notational brevity. The all ones and zero vector is denoted by $\mathbf{1}$ and $\mathbf{0}$, respectively. Given a matrix $\mathbf{G} \in \mathbb{C}^{m \times d}$ and an index set $U \subset [d]$, we denote by $\mathbf{G}_U$ the matrix of size $m \times |U|$ consisting of the columns of $\mathbf{G}$ indexed by U. This notation also extends to vectors where—depending on context—we sometimes abuse notation and write $\mathbf{x}_U$ for the vector of length $|U|$ corresponding to the restriction of $\mathbf{x}$ to U, or the d-vector $\mathbf{x}_U \in \mathbb{C}^d$ agreeing with $\mathbf{x}$ on U and vanishing identically on $\overline{U}$. For two vectors $\mathbf{x}, \mathbf{z} \in \mathbb{C}^d$, we denote by $\mathbf{x} \circ \mathbf{z}$ the Hadamard product with $(\mathbf{x} \circ \mathbf{z})_i = x_i z_i$. Moreover, we assume that matrix-vector multiplication binds before $\circ$ and write $\mathbf{y} \circ \mathbf{A}\mathbf{x}$ to mean $\mathbf{y} \circ (\mathbf{A}\mathbf{x})$ for $\mathbf{A} \in \mathbb{C}^{m \times d}$ and $\mathbf{y} \in \mathbb{C}^m$. Finally, to ease notation, we will make frequent use of the following asymptotic notation: given two scalars $a, b \in \mathbb{R}$, we write $a \lesssim b$ if there exists an absolute constant $C > 0$ such that $a \leq Cb$ holds. Similarly, we write $a \gtrsim b$ to mean $a \geq Cb$. If the implicit constant depends on some parameter τ, we will sometimes also write $a \lesssim_\tau b$ and $a \gtrsim_\tau b$ to indicate such a dependence.

2.2 Compressed Sensing

At its core, *compressed sensing* (CS) is concerned with the question under which conditions a vector $\mathbf{x} \in \mathbb{C}^d$ can be uniquely determined from linear measurements of the form $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$, where the matrix $\mathbf{A} \in \mathbb{C}^{m \times d}$ is assumed to have full rank. For arbitrary vectors $\mathbf{x}$, linear algebra dictates that this is only possible if $\mathrm{rank}(\mathbf{A}) = d$, implying most crucially that $m \geq d$. In the parlance of sampling theory, we say that one has to acquire more measurements than the dimension of the ambient signal space to undo the measurement procedure modeled by the *measurement* or *sensing matrix* $\mathbf{A}$. Unfortunately, in many practical applications the signal space containing $\mathbf{x}$ might be of very high dimension. It is therefore highly desirable to establish conditions on the signal space which would allow for vectors to be recovered from $m < d$ measurements.

The key insight of compressed sensing is that the rank requirement stated above turns out to be overly pessimistic if the vector one tries to recover from knowledge of the measurements $\mathbf{y}$ and the matrix $\mathbf{A}$ exhibits a sparse low-complexity structure. In fact, if $\mathbf{x}$ has at most s nonzero components, then it can be shown that there always exists a

matrix $\mathbf{A} \in \mathbb{C}^{(s+1)\times d}$ which uniquely determines $\mathbf{x}$ from its measurements $\mathbf{y} = \mathbf{A}\mathbf{x}$ [FR13, Theorem 2.16]. Such a result is usually referred to as *nonuniform* since it does not imply that the same matrix $\mathbf{A}$ also uniquely determines any other vector $\mathbf{x}' \neq \mathbf{x}$ with at most s nonzero entries. In contrast, a *uniform* result is one which holds for the entire class of s-sparse vectors, which we generally denote by

$$\Sigma_s := \left\{\mathbf{x} \in \mathbb{C}^d : \|\mathbf{x}\|_0 \leq s\right\}$$

with $\|\mathbf{x}\|_0 := |\mathrm{supp}(\mathbf{x})| = |\{i \in [d] : x_i \neq 0\}|$ denoting the so-called ℓ_0-pseudonorm[1]. We will sometimes use the shorthand notation $\Sigma_s(\mathcal{V}) := \Sigma_s \cap \mathcal{V}$ for some linear subspace $\mathcal{V} \subset \mathbb{C}^d$ to emphasize the base space of sparse vectors.

Motivated by the above observations, it is natural to formulate the so-called ℓ_0-minimization problem to recover a vector $\mathring{\mathbf{x}} \in \Sigma_s$ from its linear measurements $\mathbf{A}\mathring{\mathbf{x}}$:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_0 \\ \text{s.t.} \quad & \mathbf{A}\mathring{\mathbf{x}} = \mathbf{A}\mathbf{x}. \end{aligned} \tag{P_0}$$

From this formulation, one immediately concludes that Problem (P_0) admits a unique solution if and only if $\ker(\mathbf{A}) \cap \Sigma_{2s} = \{\mathbf{0}\}$, meaning that the null space of $\mathbf{A}$ must not contain any other $2s$-sparse vectors beside the zero vector. This yields the fundamental condition $m \geq 2s$, which must be satisfied by any matrix $\mathbf{A} \in \mathbb{C}^{m\times d}$ to reconstruct every s-sparse vector (see, *e.g.*, [Don06a, Lemma 2.1] or [FR13, Section 2.2]). Unfortunately, the combinatorial nature of Problem (P_0) renders the recovery strategy computationally intractable since it requires solving $\sum_{i=0}^{d} \binom{d}{i} \geq \sum_{i=0}^{d} (d/i)^i$ linear systems. Luckily, the story does not end here.

Decades before the advent of compressed sensing, researchers in statistics and seismology had already observed that sparse solutions to linear systems could be obtained by solving the following equality-constrained ℓ_1-minimization problem:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \mathbf{A}\mathring{\mathbf{x}} = \mathbf{A}\mathbf{x}. \end{aligned} \tag{P_1}$$

This tractable program, which can be solved via linear programming in the real setting and via second-order cone programming in the complex case, corresponds to the closest convex relaxation of Problem (P_0). It is most commonly referred to as the *basis pursuit* (BP) problem, a term originally coined in [CD94]. While the concept of ℓ_1-regularization found widespread adoption in a variety of domains such as model selection in statistics and image denoising, the exact mathematical connections between Problem (P_0) and Problem (P_1) did not become clear until the seminal works of Candès, Romberg, Tao and Donoho. In the modern theory of compressed sensing, the equivalence between the two programs can be conveniently established via the so-called *null space property* introduced in [CDD09].

Definition 2.1 (Null space property)**.** *A matrix $\mathbf{A} \in \mathbb{C}^{m\times d}$ is said to satisfy the* null space property *(NSP) of order s if for any $S \subset [d]$ of size $|S| \leq s$, one has*

$$\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\overline{S}}\|_1 \quad \forall \mathbf{v} \in \ker(\mathbf{A}) \setminus \{\mathbf{0}\}.$$

[1] While $\|\cdot\|_0$ can be interpreted as the limit of $\|\cdot\|_q^q$ for $q \to 0$, it is neither a semi- nor a quasinorm since it is clearly not homogeneous.

While this property was not originally used by Candès *et al.*, it can be shown to be both necessary and sufficient to establish the equivalence between Problem (P_0) and Problem (P_1). Let us first point out that the null space property implies the previously stated condition $\ker(\mathbf{A}) \cap \Sigma_{2s} = \{\mathbf{0}\}$, establishing that every s-sparse vector corresponds to the unique solution of Problem (P_0). To that end, let $\mathbf{v} \in \ker(\mathbf{A}) \cap \Sigma_{2s}$, and consider the decomposition $\mathbf{v} = \mathbf{v}_{S_1} + \mathbf{v}_{S_2}$ where $S_1, S_2 \subset S := \operatorname{supp}(\mathbf{v})$ are arbitrary index sets with $|S_1| = |S_2| = s$. Now assume that $\mathbf{v} \neq \mathbf{0}$. Then by the NSP we have $\|\mathbf{v}_{S_1}\|_1 < \|\mathbf{v}_{\overline{S_1}}\|_1 = \|\mathbf{v}_{S \setminus S_1}\|_1 = \|\mathbf{v}_{S_2}\|_1$ and similarly $\|\mathbf{v}_{S_2}\|_1 < \|\mathbf{v}_{S_1}\|_1$, which is a contradiction. The NSP of order s therefore guarantees that the only $2s$-sparse vector in the null space of $\mathbf{A}$ is the zero vector.

The following result establishes the equivalence between Problem (P_0) and (P_1) conditioned on the null space property.

Theorem 2.2. *Let $\mathbf{A} \in \mathbb{C}^{m \times d}$ and fix $s \in [d]$. Then every s-sparse vector $\mathring{\mathbf{x}} \in \mathbb{C}^d$ is the unique minimizer of Problem (P_1) if and only if $\mathbf{A}$ satisfies the NSP of order s.*

Proof. We first show that the NSP implies uniqueness. To that end, let $\mathbf{x} \neq \mathring{\mathbf{x}}$ be feasible for Problem (P_1). Set $S = \operatorname{supp}(\mathring{\mathbf{x}})$, and define the vector $\mathbf{v} := \mathring{\mathbf{x}} - \mathbf{x} \in \ker(\mathbf{A}) \setminus \{\mathbf{0}\}$, which we decompose as $\mathbf{v}_S = \mathring{\mathbf{x}}_S - \mathbf{x}_S = \mathring{\mathbf{x}} - \mathbf{x}_S$ and $\mathbf{v}_{\overline{S}} = \mathbf{x}_{\overline{S}}$. By the null space property, this implies

$$\|\mathring{\mathbf{x}}\|_1 = \|\mathring{\mathbf{x}} - \mathbf{x}_S + \mathbf{x}_S\|_1 \leq \|\mathbf{v}_S\|_1 + \|\mathbf{x}_S\|_1 < \|\mathbf{v}_{\overline{S}}\|_1 + \|\mathbf{x}_S\|_1 = \|\mathbf{x}_{\overline{S}}\|_1 + \|\mathbf{x}_S\|_1 = \|\mathbf{x}\|_1.$$

This establishes $\mathring{\mathbf{x}}$ as the unique minimizer of Problem (P_1).

For the opposite direction, let $\mathbf{v} \in \ker(\mathbf{A}) \setminus \{\mathbf{0}\}$, and denote by $S \subset [d]$ an arbitrary index set with $|S| \leq s$. Since $\mathbf{v}_S$ is s-sparse, it is also the unique optimal solution of Problem (P_1). Moreover, since $\mathbf{A}\mathbf{v}_S = \mathbf{A}(-\mathbf{v}_{\overline{S}})$, the vector $-\mathbf{v}_{\overline{S}}$ is feasible for Problem (P_1). This implies

$$\|\mathbf{v}_S\|_1 < \|-\mathbf{v}_{\overline{S}}\|_1 = \|\mathbf{v}_{\overline{S}}\|_1,$$

which is the null space property since the choice of S was arbitrary. □

While the NSP provides a necessary and sufficient condition to establish perfect recovery of s-sparse vectors from compressive measurements, it is not general enough to establish robustness to additive noise in the observations and sparsity defect. For this, one may either appeal to a generalized version of the NSP (see, *e.g.*, [**FR13**, Definition 4.21]) or appeal to the infamous *restricted isometry property* due to Candès and Tao [**CT05**].

Definition 2.3 (Restricted isometry property)**.** *A matrix $\mathbf{A} \in \mathbb{C}^{m \times d}$ is said to satisfy the* restricted isometry property *(RIP) of order s with constant $0 < \delta_s \leq \delta < 1$ if*

$$(1 - \delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta)\|\mathbf{x}\|_2^2 \quad \forall \mathbf{x} \in \Sigma_s.$$

To see that this property implies that RIP matrices act injectively on Σ_s, consider a matrix $\mathbf{A} \in \mathbb{C}^{m \times d}$ satisfying the RIP of order $2s$ with constant $\delta_s < 1$. Now consider two vectors $\mathbf{x}, \mathbf{z} \in \Sigma_s$ with $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{z}$ such that $\mathbf{v} := \mathbf{x} - \mathbf{z} \in \ker(\mathbf{A})$. Then the RIP implies that

$$(1 - \delta_s)\|\mathbf{v}\|_2^2 \leq \|\mathbf{A}\mathbf{v}\|_2^2 = 0.$$

This only holds for $\mathbf{v} = \mathbf{0}$ and therefore $\mathbf{x} = \mathbf{z}$, meaning that no two s-sparse vectors are mapped to the same element of $\mathbf{A}\Sigma_s$ if the matrix satisfies the RIP of order $2s$.

Before stating a stable[2] and robust recovery guarantee based on the RIP, we require the so-called *best s-term approximation error* defined as

$$\sigma_s(\mathbf{x})_p = \inf_{\mathbf{u}\in\Sigma_s} \|\mathbf{x} - \mathbf{u}\|_p.$$

The error $\sigma_s(\mathbf{x})_p$ quantifies the model mismatch or sparsity defect of a vector $\mathbf{x}$ relative to Σ_s. Equipped with this definition, one may establish the following recovery result due to Candès.

Theorem 2.4 ([Can08, Theorem 1.3]). *Assume the matrix $\mathbf{A} \in \mathbb{C}^{m\times d}$ satisfies the restricted isometry property of order $2s$ with $\delta_{2s} < \sqrt{2}-1$. Then under the noisy measurement model $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} + \mathbf{e}$ with $\mathring{\mathbf{x}} \in \mathbb{C}^d$ and $\mathbf{e} \in \mathbb{C}^m$ with $\|\mathbf{e}\|_2 \leq \nu$, every minimizer $\hat{\mathbf{x}}$ of*

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} && \|\mathbf{x}\|_1 \\ &\text{s.t.} && \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \nu \end{aligned}$$

satisfies

$$\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \leq C_0 \frac{\sigma_s(\mathring{\mathbf{x}})_1}{\sqrt{s}} + C_1\nu$$

where $C_0, C_1 > 0$ are universal constants which only depend on δ_{2s}.

This general result implies perfect recovery in the noiseless setting with $\nu = 0$ if $\mathring{\mathbf{x}}$ is exactly s-sparse such that $\sigma_s(\mathring{\mathbf{x}})_1 = 0$. As pointed out before, a similar result can be established by a generalized version of the null space property (cf. [FR13, Theorem 4.22]). Moreover, we emphasize that the RIP is only sufficient to guarantee stable and robust recovery. This means that the RIP implies the NSP while the other direction is not generally true. Finally, certifying whether a matrix $\mathbf{A}$ satisfies the NSP or RIP turns out to belong to the complexity class NP-hard [TP14]. This unfortunate circumstance might seem like a considerable roadblock regarding the practical relevance of compressed sensing. Luckily, however, researchers realized early on that RIP matrices abound if one turns to random designs. This includes both random matrices whose rows are isotropic subgaussian random vectors (see Definition A.3), as well as matrices constructed from randomly subsampled basis functions of *bounded orthonormal systems* (BOSs). This includes trigonometric polynomials or discrete orthobases such as the *discrete Fourier transform* (DFT), *discrete cosine transform* (DCT) or Haar basis, as well as wavelet systems. In general, it suffices to show that a random matrix $\mathbf{A}$ satisfies the concentration inequality

$$\mathbb{P}\Big(\big|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \geq t\|\mathbf{x}\|_2^2\Big) \leq 2\exp(-ct^2m) \quad \forall \mathbf{x} \in \mathbb{C}^d \tag{2.1}$$

for $t \in (0,1)$ and $c > 0$ a constant to establish the RIP. This holds, for instance, for subgaussian random matrices whose rows $\mathbf{Y} \in \mathbb{R}^d$ are isotropic subgaussian random vectors (see, *e.g.*, [FR13, Lemma 9.8]). Most crucially, the isotropy assumption does not require the entries of $\mathbf{Y}$ to be independent. With this, one establishes the following result.

[2] Informally speaking, a recovery procedure is called *stable* if it does not go completely astray if the target vector $\mathbf{x} \in \mathbb{C}^d$ is not exactly sparse. In particular, one requires that a stable reconstruction map recovers $\mathbf{x}$ exactly if $\|\mathbf{x}\|_0 \leq s$.

Theorem 2.5 ([FR13, Theorem 9.11]). *Let $\mathbf{A} \in \mathbb{C}^{m\times d}$ be a random matrix satisfying (2.1). Then with probability at least $1-\eta$, the matrix $\mathbf{A}$ satisfies the RIP of order s with constant $\delta_s \leq \delta$, provided that*

$$m \geq C\delta^{-2}\left[s\log(d/s) + \log(\eta^{-1})\right]$$

where the constant $C > 0$ only depends on c in (2.1).

This result in combination with Theorem 2.4 implies that any vector $\mathring{\mathbf{x}} \in \mathbb{C}^d$ can be stably and robustly recovered from $m = \Omega(s\log(d/s))$ measurements with overwhelmingly high probability. As hinted at in the introduction of this section, the number of measurements depends almost linearly on s, the information dimension of the target signal. Moreover, the dependence of m on the parameters d and s is known to be optimal, meaning most crucially that the logarithmic factor cannot be removed. In a sense, this logarithmic dependence on d and s is the price one has to pay for the fact that the support of the signal is unknown. This surprising fact was already established in one of Donoho's earliest papers on compressed sensing [Don06c] by appealing to the theory of so-called *Gelfand numbers*. As recently argued by Foucart in [Fou16, Section 6.3], however, the theory of compressed sensing is nowadays largely self-contained in the sense that the optimality of $\Omega(s\log(d/s))$ for stable recovery can be established without the concept of Gelfand numbers. By *stable* we mean that for a pair $(\mathbf{A}, \Delta)$ with $\mathbf{A} \in \mathbb{C}^{m\times d}$ a measurement matrix and $\Delta\colon \mathbb{C}^m \to \mathbb{C}^d$ a (nonlinear) recovery map, one has

$$\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_2 \leq C\frac{\sigma_s(\mathbf{x})_p}{\sqrt{s}} \quad \forall \mathbf{x} \in \mathbb{C}^d$$

for some absolute constant $C > 0$. In particular, if $(\mathbf{A}, \Delta)$ is a stable pair, then $m \geq cs\log(d/s)$ where $c > 0$ only depends on C above [FR13, Proposition 10.7].

In closing, we point out that it is also possible to directly establish the null space property or its stable and robust variant directly for random ensembles rather than appealing to the restricted isometry property first. This is sometimes desirable (and necessary) since for certain ensembles, there exist examples (see, *e.g.*, [Ada+11]) which provably lead to suboptimal RIP-based lower bounds on the required number of measurements for stable recovery. A probabilistic proof of the stable NSP was first provided for Gaussian random matrices in [FR13, Section 4.2] by estimating the probability that the least singular value restricted to a particular cone is bounded away from zero via *Gordon's escape through a mesh* theorem [Gor88]. In order to overcome unfavorable scaling in the number of measurements for certain random matrices, this result was later extended in [DLR18] to more heavy-tailed ensembles via Mendelson's *small ball method* [Men14].

2.3 Quantized Compressed Sensing

We now turn to the topic of *quantized compressed sensing* (QCS), an area of research which has gained a lot of traction in recent years as it sits at the intersection of theory and practice of compressed sensing. Since its inception, the field of compressed sensing has developed into an elegant and mature theory for efficient signal acquisition and reconstruction. However, most theoretical results reported in the literature initially focused exclusively on

measurements belonging to a continuum $\mathbb{K}^m$ with $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. While convenient from a theoretical perspective, this ignores the fact that the contemporary world of signal processing is distinctly digital. This means that linear projections $\mathbf{Ax} \in \mathbb{K}^m$ considered by compressive acquisition systems have to be subsequently digitized for further processing, transmission and storage. To that end, one considers nonlinear quantization maps $Q\colon \mathbb{K}^m \to \mathcal{A}^m$ which map the individual components of real- or complex-valued vectors onto a finite set of quantization points $\mathcal{A} \subset \mathbb{K}$, the so-called *quantization alphabet*. Naively, the resulting quantization noise $Q(\mathbf{Ax}) - \mathbf{Ax}$ introduced by this lossy (and fundamentally irreversible) mapping can be modeled as an additive perturbation, which can be readily accounted for within the existing theory. Unfortunately, this approach is only justified under the so-called *high-resolution assumption* (HRA), which roughly states that the largest cell width $\theta > 0$ of a scalar quantizer is small compared to its dynamic range. For simplicity, we mainly focus on the case of uniform quantization in this section and refer the interested reader to the detailed survey [Bou+15], which also features an in-depth discussion of more advanced quantization schemes. This includes nonuniform quantization with companding like A- and μ-law encoding, as well as a frame-theoretic treatment of recursive $\Sigma\Delta$ quantization schemes in compressed sensing.

Before going any further, we first introduce an important concept in quantized compressed sensing known as *quantization consistency* and present some fundamental error bounds which hold for any reconstruction procedure based on quantized compressive measurements.

Definition 2.6 (Quantization consistency)**.** *Given a measurement matrix $\mathbf{A} \in \mathbb{K}^{m\times d}$ and a quantization operator $Q\colon \mathbb{K}^m \to \mathcal{A}^m$, a reconstruction map $\Delta\colon \mathcal{A}^m \to \mathbb{K}^d$ is said to be* quantization-consistent *if*

$$Q(\mathbf{A}\Delta(Q(\mathbf{Ax}))) = Q(\mathbf{Ax}) \quad \forall \mathbf{x} \in \mathbb{K}^d.$$

In words, a recovery map Δ is called quantization-consistent if any estimate $\hat{\mathbf{x}} = \Delta(Q(\mathbf{Ax}))$ yields the same quantized measurements as the vector $\mathbf{x}$ one aims to recover. It turns out that consistent reconstruction is a key enabler for accurate signal recovery from quantized observations, at least in the noiseless setting. Such a constraint is natural as it exploits all available information about the measurement system. Moreover, for multi-bit quantization schemes with $B \geq 2$, quantization consistency is easily guaranteed by imposing a simple convex constraint in the context of convex programming.

For illustrative purposes, consider a B-bit quantizer Q and measurements of the form $\mathbf{y} = Q(\mathbf{Ax})$ with $\mathbf{A} \in \mathbb{K}^{m\times d}$ and $\mathbf{x}$ belonging to an s-dimensional subspace $\mathcal{V}$ of $\mathbb{K}^d$. Since the measurement space $\mathbf{A}\mathcal{V}$ forms an s-dimensional linear subspace of $\mathbb{K}^m$, only a fraction of all possible $(2^B)^m = 2^{mB}$ quantization cells are actually needed to represent the set $Q(\mathbf{A}\mathcal{V})$. The same holds true if the subspace $\mathcal{V}$ is replaced with a low-complexity signal set $\mathcal{K}$ such as the set of sparse vectors $\Sigma_s = \Sigma_s(\mathbb{K}^d)$. This observation allows for deriving a lower bound on the worst-case reconstruction error of $\mathbf{x}$ from its quantized measurements $\mathbf{y}$ among all recovery maps $\Delta\colon \mathcal{A}^m \to \mathbb{K}^d$. The key idea is to select a finite subset $\mathcal{Q} \subset \mathcal{K}$ which minimizes the error

$$\epsilon_{\text{opt}} := \sup_{\mathbf{x}\in\mathcal{K}} \inf_{\mathbf{q}\in\mathcal{Q}} \|\mathbf{x} - \mathbf{q}\|_2.$$

For $\mathcal{K} = \Sigma_s$, it can be shown that this yields [Bou+15]

$$\epsilon_{\text{opt}} \gtrsim \frac{2^{-B}s}{m}, \tag{2.2}$$

meaning that the reconstruction error $\|\mathbf{x} - \Delta(Q(\mathbf{Ax}))\|_2$ for any $\mathbf{x} \in \Sigma_s$ decays at most linearly in the number of measurements. Most crucially, quantization-consistent reconstruction schemes have previously been shown to achieve the lower bound (2.2) within logarithmic factors in the Gaussian setting [Jac16].

In the following, we consider a scalar quantizer $Q\colon \mathbb{K} \to \mathcal{A}$ which acts element-wise on the linear measurements $\mathbf{Ax}$. For simplicity, we assume that the dynamic range $\mathcal{R} \subset \mathbb{R}$ of the quantizer Q is large enough so that the input $\mathbf{Ax}$ does not saturate, *i.e.*, $[\min_{i\in[m]} \langle \mathbf{a}_i, \mathbf{x}\rangle, \max_{i\in[m]} \langle \mathbf{a}_i, \mathbf{x}\rangle] \subseteq \mathcal{R}$. Consider, for instance, a uniform quantizer Q with dynamic range $\mathcal{R}$ and quantization alphabet $\mathcal{A} = \theta(\mathbb{Z} + 1/2) \cap \mathcal{R}$ for some fixed cell width $\theta > 0$. The mapping Q takes its input y to the staircase according to the quantization rule $Q(y) = \theta(\lfloor y/\theta \rfloor + 1/2)$.[3] Assuming a symmetric dynamic range $\mathcal{R} = [-y_{\max}, y_{\max}]$ with $0 < y_{\max} < \infty$, a B-bit uniform quantizer partitions the interval $\mathcal{R}$ into 2^B quantization cells of equal width $\theta = 2y_{\max}/2^B = y_{\max}2^{-B+1}$. Since this model constrains the quantization error to the interval $[-\theta/2, \theta/2]$, the error term $\mathbf{q} := Q(\mathbf{Ax}) - \mathbf{Ax} \in \mathbb{R}^m$ of a uniform scalar quantizer Q acting individually on each coordinate of the linear measurement vector $\mathbf{Ax}$ belongs to the scaled unit ball $\theta/2\mathbb{B}_\infty^m \subset \sqrt{m}\theta/2\mathbb{B}_2^m$. However, this bound actually turns out to be slightly too pessimistic if B is large enough for Q to satisfy the HRA. In this case, it is natural to consider a probabilistic model of the quantizer noise in terms of independent uniformly distributed random variables $q_i \sim_{\text{i.i.d.}} \mathsf{U}([-\theta/2, \theta/2])$. It then follows that $\mathbb{E}\|\mathbf{q}\|_2^2 \leq (m\theta^2/2 + \zeta\sqrt{m}\theta^2/\sqrt{5})/6 =: \xi^2$ where $\zeta > 0$ is a small universal constant [Bou+15]. Both observations suggest to solve the following *quadratically-constrained basis pursuit* (QCBP) problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \|\mathbf{Ax} - \mathbf{y}\|_2 \leq \varepsilon_2 \end{aligned} \tag{$\mathrm{P}_{2.1}$}$$

for $\varepsilon_2 = \xi$ or $\varepsilon_2 = \sqrt{m}\theta/2$ to recover $\mathring{\mathbf{x}} \in \mathbb{K}^m$ from its quantized measurements $\mathbf{y} = Q(\mathbf{A}\mathring{\mathbf{x}})$. The recovery quality of a minimizer $\mathbf{x}^\star$ of Problem ($\mathrm{P}_{2.1}$) then follows by Theorem 2.4. Unfortunately, since minimizers $\mathbf{x}^\star$ of Problem ($\mathrm{P}_{2.1}$) are not necessarily feasible for the problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \|\mathbf{Ax} - \mathbf{y}\|_\infty \leq \theta/2, \end{aligned} \tag{$\mathrm{P}_{2.2}$}$$

this approach is suboptimal considering the aforementioned importance of quantization consistency. These observations led to the advent of the so-called *basis pursuit dequantizer* framework [JHF11], which considers the family of problems

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \|\mathbf{Ax} - \mathbf{y}\|_p \leq \varepsilon_p \end{aligned} \tag{$\mathrm{P}_{2.3}$}$$

[3] Such a quantization rule is generally referred to as a *mid-rise quantizer* due to its behavior around $y = 0$ where a mid-rise quantizer exhibits a rising edge.

for an appropriate choice of $p \geq 1$ and ε_p. In particular, ε_p is chosen based on a high-probability bound on $\|\mathbf{q}\|_p$ under the HRA with $q_i \sim_{\text{i.i.d.}} \mathsf{U}([-\theta/2, \theta/2])$. The analysis of this problem hinges on a generalized variant of the restricted isometry property suited to yield error guarantees of the form $\|\mathring{\mathbf{x}} - \mathbf{x}^\star\|_2 \lesssim \sigma_s(\mathring{\mathbf{x}})_1/\sqrt{s} + \varepsilon_p$. Under these conditions, it can be shown that the recovery error decays as $\mathcal{O}(\theta/\sqrt{p+1})$ [JHF11]. Note that this result only holds for finite p, meaning that consistent reconstruction is never strictly enforced by Problem ($\mathrm{P}_{2.3}$). Moreover, this error decay behavior comes at a price. While Gaussian random matrices can be shown to satisfy the generalized RIP condition with high probability, the required number of measurements scales exponentially in p. Most importantly, this means that one sacrifices the linear dependence of m on the sparsity level s. Expressed in terms of m, one finds that the error decays as $\mathcal{O}(\theta/\sqrt{\log(m)})$, which is far from the optimal linear decay rate implied by (2.2).

Sacrificing a recovery error which decays with m, it was shown in [DLR18] that consistent reconstruction allows for θ-accurate signal recovery in the sparse case with the same number of measurements as required for exact recovery. This result relies on a generalized null space property for Gaussian random matrices rather than the RIP. It is shown that for $\mathring{\mathbf{x}} \in \Sigma_s$ and $\mathbf{x}^\star$ denoting a minimizer of Problem ($\mathrm{P}_{2.2}$), one has $\|\mathring{\mathbf{x}} - \mathbf{x}^\star\|_2 = \mathcal{O}(\theta)$ with high probability, provided that $m = \Omega(s \log(d/s))$.

Appealing to random pre-quantization dithering of the form $Q(\mathbf{A}\mathbf{x} + \boldsymbol{\tau})$, it was later shown in [Mos+16] that consistent reconstruction based on subgaussian measurement matrices is possible with an error decay of $\mathcal{O}(m^{-1/4} + \kappa)$ where $\kappa > 0$ denotes a constant which vanishes if the measurement matrix is Gaussian. While still far from optimal, this improves upon [DLR18] in the Gaussian setting.

One of the strongest recovery results in QCS to date was first presented in [XJ18]. The work leverages the classical restricted isometry property and combines it with the so-called *limited projection distortion* property, establishing that a simple noniterative recovery procedure achieves an error decay of $\mathcal{O}((1+\theta)m^{-1/2})$. One key observation in the result is that choosing the dithering vector $\boldsymbol{\tau} \in \mathbb{R}^m$ in the dithered observation model $\mathbf{y} = Q(\mathbf{A}\mathbf{x} + \boldsymbol{\tau})$ as $\tau_i \sim_{\text{i.i.d.}} \mathsf{U}([0, \theta])$, one has that dithering removes the effect of quantization in expectation, *i.e.*, $\mathbb{E}_{\boldsymbol{\tau}} Q(\mathbf{w} + \boldsymbol{\tau}) = \mathbf{w}$.

In the next section, we turn to an extreme version of QCS, which will feature heavily in the remainder of this thesis: the 1-bit compressed sensing model.

2.3.1 One-Bit Compressed Sensing

As hinted at in the introduction, the 1-bit compressed sensing model has a number of convenient benefits over higher-order quantization schemes such as memoryless multi-bit quantization or more complicated designs like $\Sigma\Delta$ quantizers. While 1-bit CS can be regarded as a special case of multi-bit quantization with $B = 1$ as discussed above, the model has its own idiosyncrasies, which require special attention.

The 1-bit CS model for $\mathbf{x} \in \mathbb{R}^d$ considers measurements of the form[4]

$$\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathbf{x}) \in \{\pm 1\}^m \tag{2.3}$$

where $\mathbf{A} \in \mathbb{R}^{m \times d}$ models the linear part of the acquisition system as before, and the quantization function is chosen as $Q = \operatorname{sgn}$. One immediate consequence of this extreme

[4] By convention, we define $\operatorname{sgn}(0) = 1$.

quantization model is that there is no hope of recovering anything more than the direction of $\mathbf{x}$ from knowledge of the binary measurements $\mathbf{y}$ and the measurement matrix $\mathbf{A}$. This follows directly from the scale invariance of the sgn-operator since $\operatorname{sgn}(\mathbf{Ax}) = \operatorname{sgn}(\alpha\mathbf{Ax})$ for any $\alpha > 0$. We immediately point out that this limitation can be lifted by considering dithered observations $\mathbf{y} = \operatorname{sgn}(\mathbf{Ax}+\boldsymbol{\tau})$ as briefly mentioned above. For simplicity, however, we limit our current discussion to the case $\boldsymbol{\tau} = \mathbf{0}$ and refer to Section 4.4 for more details. To remove the scale ambiguity from the problem, we concern ourselves with estimating vectors $\mathbf{x}$ which belong to some structured signal set $\mathcal{K} \subset \mathbb{S}^{d-1}$.

Geometrically, the measurement procedure amounts to a tessellation of the unit sphere. More precisely, every row $\mathbf{a}_i$ of $\mathbf{A}$ defines a hyperplane with normal $\mathbf{a}_i$ which partitions the sphere $\mathbb{S}^{d-1}$ into two distinct spherical caps. A measurement of the form $y_i = \operatorname{sgn}(\langle \mathbf{a}_i, \mathbf{x}\rangle)$ therefore determines which side of the hyperplane $\mathbf{x}$ lies on. The collection of all sign measurements $(\operatorname{sgn}(\langle \mathbf{a}_i, \mathbf{x}\rangle))_{i=1}^m$ consequently yields an encoding of which sphere patch $\mathbf{x}$ belongs to. Intuitively, a new measurement vector $\mathbf{a}_j$ only provides any new information if the hyperplane $E_j := \{\mathbf{z} \in \mathbb{R}^d : \langle \mathbf{a}_j, \mathbf{z}\rangle = 0\}$ intersects the set $C_{\mathbf{x}} := \{\mathbf{z} : \operatorname{sgn}(\mathbf{Ax}) = \operatorname{sgn}(\mathbf{Az})\}$ of quantization-consistent vectors for a fixed vector $\mathbf{x} \in \mathbb{R}^d$. If the vectors $\mathbf{a}_i$ are drawn from the Haar measure on $\mathbb{S}^{d-1}$, then the probability that a new measurement adds new information decreases every time a hyperplane shrinks the quantization region. This already hints at the fact that the reconstruction error cannot decay faster than polynomially in m unless one considers adaptive quantization schemes [Bar+17b].

Before addressing the issue of signal recovery, we first remark on a convenient property of model (2.3). Apart from its beneficial implications w. r. t. hardware complexity and energy consumption, the 1-bit CS model has the added advantage of robustness against gross nonlinearities and saturation effects of the measurement system, as well as additive perturbations of the measurements [BB08; Bou10]. To demonstrate the error resilience of the model, consider noisy measurements of the form $\tilde{\mathbf{y}} = \operatorname{sgn}(\mathbf{Ax} + \mathbf{e})$ with $\mathbf{e} \in \mathbb{R}^m$ a noise vector consisting of i.i.d. $\mathsf{N}(0, \sigma^2)$ random variables. If we take $\mathbf{A}$ to be a standard Gaussian random matrix and acquire measurements of a fixed signal $\mathbf{x} \in \mathbb{R}^d$, we have $y_i = \langle \mathbf{a}_i, \mathbf{x}\rangle + e_i \sim \mathsf{N}(0, \|\mathbf{x}\|_2^2 + \sigma^2)$. It was shown in [Jac+13, Lemma 4] that under these assumptions, the probability $\mathbb{P}(\langle \mathbf{a}_i, \mathbf{x}\rangle y_i < 0) =: \tilde{p}$ of a sign flip is bounded by $\tilde{p} \leq \frac{1}{2}\sigma(\|\mathbf{x}\|_2^2 + \sigma^2)^{-1/2} = \frac{1}{2}(1 + d\rho)^{-1/2}$ with $\rho := \|\mathbf{x}\|_2^2/(d\sigma^2)$ denoting the *signal-to-noise ratio* (SNR). At a signal space dimension of $d = 1000$ and an SNR of $-10\,\mathrm{dB}$, this implies a sign flip probability of less than 0.05. Intuitively, this error robustness is due to the fact that the noise level has to eclipse the unquantized measurements (in addition to having opposite sign) to result in an effective bit flip.

Similar to the situation in multi-bit QCS, it is possible to derive a lower bound on the attainable reconstruction error among all recovery maps by choosing an appropriate subset $\mathcal{Q}$ of $\widetilde{\Sigma}_s := \Sigma_s(\mathbb{R}^d) \cap \mathbb{S}^{d-1}$ which minimizes the worst-case error

$$\epsilon_{\mathrm{opt}} = \sup_{\mathbf{x} \in \widetilde{\Sigma}_s} \inf_{\mathbf{q} \in \mathcal{Q}} \|\mathbf{x} - \mathbf{q}\|_2.$$

In particular, it was shown in [Jac+13] that for any measurement matrix $\mathbf{A} \in \mathbb{R}^{m \times d}$, the union of s-dimensional subspaces $\mathbf{A}\widetilde{\Sigma}_s$ of $\mathbb{R}^m$ intersects at most $|\mathcal{Q}| = 2^s \binom{d}{s} \binom{m}{s}$ orthants

identified by $\operatorname{sgn}(\mathbf{A}\widetilde{\Sigma}_s)$. Using this, they derive the lower bound

$$\epsilon_{\text{opt}} \geq \frac{s}{2em + \sqrt{2}s^{3/2}}, \tag{2.4}$$

which shows that the reconstruction error of any decoder Δ again decays at most linearly in m, *i.e.*, $\|\mathbf{x} - \Delta(\operatorname{sgn}(\mathbf{A}\mathbf{x}))\|_2 = \mathcal{O}(s/m)$ for $\mathbf{x} \in \widetilde{\Sigma}_s$ and any reconstruction map $\Delta\colon \{\pm 1\}^m \to \mathbb{R}^d$. Moreover, they establish ([Jac+13, Theorem 2]) that any quantization-consistent reconstruction map Δ according to Definition 2.6 satisfies $\|\mathbf{x} - \Delta(Q(\mathbf{A}\mathbf{x}))\|_2 \leq \varepsilon$ with probability at least $1 - \eta$ if the entries of $\mathbf{A} \in \mathbb{R}^{m\times d}$ are drawn independently from $\mathsf{N}(0, 1)$ and

$$m \geq \frac{2}{\varepsilon}\Big[2s\log(d) + 4s\log(17/\varepsilon) + \log(\eta^{-1})\Big]. \tag{2.5}$$

For a fixed failure probability η, this corresponds to an error decay of $\|\mathbf{x} - \Delta(Q(\mathbf{A}\mathbf{x}))\|_2 = \mathcal{O}(s/m\log(dm/s))$, which is optimal up to the logarithmic factor. Additionally, [JDV13] establishes that with almost the same choice of m and $\mathbf{A}$ as before, two s-sparse vectors $\mathbf{x}$ and $\mathbf{z}$ whose quantized measurements differ in at most t positions are at most $\|\mathbf{x} - \mathbf{z}\|_2 \leq \varepsilon(1 + t/s)$ apart if $m \gtrsim \varepsilon^{-1}s\log(md)$. These results emphasize the importance of quantization consistency in the context of 1-bit compressed sensing.

While [Jac+13] also introduces an efficient recovery procedure—the *binary iterative hard thresholding* (BIHT) algorithm—fitting into the above category, a theoretical analysis of the proposed scheme remains an open problem. The same holds true for a series of other iterative reconstruction algorithms, which have been proposed over the years, such as the *renormalized fixed-point iteration* (RFPI) [BB08], the *matched sign pursuit* (MSP) [Bou09], the *restricted-step shrinkage* (RSS) [Las+11] and the *sign-truncated matching pursuit* (STrMP) algorithm [LGX16]. The strongest theoretical guarantees to date are those established for convex recovery schemes, which we will discuss next.

Given the central importance of the basis pursuit problem in CS to seek sparse solutions of underdetermined linear systems, a natural recovery approach to reconstruct a sparse vector $\mathring{\mathbf{x}} \in \widetilde{\Sigma}_s$ from its binary measurements is given by the problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}) = \operatorname{sgn}(\mathbf{A}\mathbf{x}) \\ & \|\mathbf{x}\|_2 = 1. \end{aligned} \tag{$\mathrm{P}_{2.4}$}$$

While minimizers of this program are (by construction) quantization-consistent, the problem is clearly nonconvex due to the sgn-operator in the first constraint and the unit-norm constraint. Baraniuk and Boufounos therefore suggest to relax the constraint $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathbf{x})$ to $\mathbf{y} \circ \mathbf{A}\mathbf{x} \geq \mathbf{0}$ [BB08]. While any feasible vector $\mathbf{x} \in C_{\mathring{\mathbf{x}}} = \{\mathbf{x} : \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}) = \operatorname{sgn}(\mathbf{A}\mathbf{x})\}$ clearly satisfies $\mathbf{y} \circ \mathbf{A}\mathbf{x} \geq \mathbf{0}$, so does the zero vector (and any vector in the null space of $\mathbf{A}$ for that matter). To remove this drawback, one imposes the norm constraint $\|\mathbf{x}\|_2 = 1$, leading to the program

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \mathbf{y} \circ \mathbf{A}\mathbf{x} \geq \mathbf{0} \\ & \|\mathbf{x}\|_2 = 1. \end{aligned} \tag{$\mathrm{P}_{2.5}$}$$

While also nonconvex, this program turns out to be more amenable to theoretical analysis. The formulation, which also forms the starting point for the derivation of the RSS algorithm [**BB08**], was first analyzed theoretically by Plan and Vershynin in [**PV13a**]. In particular, they showed that for $\mathbf{A} \in \mathbb{R}^{m \times d}$ a standard Gaussian matrix, any minimizer $\mathbf{x}^\star$ of Problem ($\mathrm{P}_{2.5}$) satisfies $\|\mathring{\mathbf{x}} - \mathbf{x}^\star\|_2 \leq \varepsilon$ with probability at least $1 - C\exp(-c\varepsilon m)$, provided that $m \geq C\varepsilon^{-5} s \log(2d/s)$. Moreover, they show that the guarantee extends to the convex program

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \mathbf{y} \circ \mathbf{A}\mathbf{x} \geq \mathbf{0} \\ & \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle = c_0 \end{aligned} \tag{$\mathrm{P}_{2.6}$}$$

for an arbitrary positive constant c_0 at the expense of m now depending on a polylogarithmic factor, requiring $m = \Omega(\varepsilon^{-5} s \log(d/s)^2)$ for ε-accurate recovery. Both results also apply to the more general signal model $\widetilde{\mathcal{E}}_s := \{\mathbf{x} \in \mathbb{S}^{d-1} : \|\mathbf{x}\|_1 \leq \sqrt{s}\}$ of *effectively s-sparse vectors* on the unit sphere. To keep this introduction brief, we delay a more detailed discussion of Problem ($\mathrm{P}_{2.6}$) and several related recovery schemes to Chapter 4.

We conclude this section by remarking that the analysis in [**PV13a**] implies an error decay of $\mathcal{O}(m^{-1/5})$ if d and s are fixed. This seems to contradict the (almost) linear decay rate implied by (2.5) since minimizers of Problem ($\mathrm{P}_{2.6}$) are in fact quantization-consistent. The difference is rooted in the fact that the bound (2.5) only holds under the assumption that target vectors are genuinely s-sparse rather than effectively s-sparse. Since the set $\widetilde{\mathcal{E}}_s$ is significantly larger than $\widetilde{\Sigma}_s$, this discrepancy in the predicted decay rate is to be expected.

3

Frequency-Sparse Signal Recovery from Binary Measurements

In a variety of domains such as medical imaging and wireless communication, it is natural to assume that certain signals of interest admit a sparse representation in the frequency domain. In fact, even some of the earliest works in compressed sensing by Candès, Romberg and Tao were originally motivated by the observation that diagnostic measurements acquired by magnetic resonance imaging systems exhibit sparse representations in the 2- or 3-dimensional Fourier domain, also referred to as k-space [**CRT06b**; **LDP07**; **Lus+08**; **GBK08**]. In wireless communication, sparse structures arise from different aspects of the underlying communication channel [**TH08**; **Ber+09**; **Ber+10**; **Baj+10**]. A domain of particular interest in this context is the field of *spectrum sensing* [**YA09**]. Despite the fact that most of the usable spectrum has been licensed off to dedicated license holders, resulting in the so-called *spectrum scarcity problem*, large portions of the available spectrum remain effectively un- or underused depending on geographical location, carrier frequency or time of day. This led to the emergence of the concept of the so-called *cognitive radio*

Parts of this chapter have been published in [**KM17**].

and in particular the related notion of *opportunistic spectrum access*. These methodologies aim to enable unlicensed users to monitor a wireless channel for communication by a primary license holder to facilitate unlicensed communication in vacant frequency bands. Due to exceedingly high bandwidths commonly encountered in wireless communication, compressed sensing and spectrum sensing are a natural fit to reduce the number of samples required to assess whether a communication channel is occupied or not [**TG07**; **Pol+09**; **Wan+09**; **ZLT11**; **AL12**]. Finally, sparsity-aware signal processing finds widespread adoption in radar imaging and localization [**HS09**; **End10**; **FSY10**] where sparsity typically arises in the time-frequency plane of the *short-time Fourier transform* (STFT).

While sparsity in the Fourier domain represents a ubiquitous assumption at the heart of numerous applications, most sensing models operate under the assumption of infinite precision measurements. As discussed in Section 2.3, this assumption is problematic as practical sampling devices must subsequently quantize compressive samples for processing, transmission or storage. This bears the potential to cause significant artifacts during reconstruction if the high-resolution assumption is violated. On the other hand, rising demands in energy-efficient sampling devices generally limit the use of high-resolution *analog-to-digital converters* (ADCs) as quantizers often constitute the main source of power draw in analog-to-digital converters [**Wal99**]. Moreover, the resolution of a quantizer represents a major limiting factor in its attainable sampling rate, which decays exponentially in the number of bits per measurement [**Le+05**]. For these reasons, it is desirable to combine compressive measurement models which exploit sparsity in the frequency domain with coarse quantization schemes such as recursive $\Sigma\Delta$ quantization schemes or the extreme 1-bit quantization paradigm discussed in Section 2.3. While the class of $\Sigma\Delta$ quantizers generally enjoys more favorable error decay behavior [**HS18**], this improvement comes at the cost of increased hardware complexity due to the need to track state variables during quantization. In this chapter, we therefore consider the reconstruction of frequency-sparse signals from memoryless 1-bit observations.

While there exists a substantial body of research addressing the 1-bit compressed sensing acquisition model with Gaussian observations, the amount of work dealing with structured ensembles is significantly more limited. Moreover, existing work on structured ensembles such as [**DJR17**; **DM18a**; **DM18b**] focuses almost exclusively on subsampled random convolutions based on partial circulant matrices generated by Gaussian or subgaussian random vectors. One notable exception is the recent work by Maly and Palzer [**MP19**] who analyze the *distributed compressed sensing* (DCS) model with binary observations by establishing a generalized restricted isometry property for block diagonal Gaussian random matrices. This is in stark contrast to linear compressed sensing where general recovery results for a large class of structured random matrices based on bounded orthonormal system have been established (see, *e.g.*, [**Rau10**]). In a sense, the situation is akin to the prevalent gap between sparse recovery results for binary observations from Gaussian projections and more general subgaussian designs. This gap is fundamentally rooted in certain unresolvable measurement ambiguities caused by subgaussian observations, which require appealing to pre-quantization *dithering* techniques to allow for more sophisticated analyses as presented in [**DM18a**; **DM18b**].

While no theoretical results exist to date which establish guarantees for the recovery of frequency-sparse signals from memoryless binary measurements, the 1-bit acquisition model has recently found application in the area of sparse *direction of arrival* (DOA) estimation [**Stö+15**; **Yu+16**; **LV17**; **Gao+17**; **HXL18**; **CGH18**]. In DOA estimation,

it is often assumed that a limited number of narrow-band signals impinges on a microphone or antenna array in the far-field of the respective excitation sources. Under this assumption, the relative delay of arrival of the superposition of signals on an individual sensor is fully determined by the incident directions of the individual wavefronts, as well as the relative position of sensors in the array. The phase shift caused by the delayed arrival of the signal superposition at each sensor can consequently be modeled as a multiplication by a complex exponential depending on the so-called *spatial frequency*, which determines the *angle of arrival* (AOA) of each signal. Discretizing the valid range of spatial frequencies then results in a system of equations where the associated measurement matrix—the so-called *steering* or *array manifold matrix*—is of Fourier type. In this context, [**Stö+15**] proposes a complex variant of the binary iterative hard thresholding algorithm [**Jac+13**], which quantizes real and imaginary parts separately and reconstructs the target signal under a joint-sparsity assumption on the real and imaginary parts of the signal.

Chapter Outline

The chapter is structured as follows. In Section 3.1, we introduce the frequency-sparse signal and acquisition model considered throughout, as well as its specialization to conjugate symmetric signals with a real-valued inverse Fourier transform. We then review the binary iterative hard thresholding algorithm in Section 3.2 and present our modification for conjugate symmetric signal recovery in Section 3.3. We also discuss an extension of the measurement model to oversampled time domain representations. Before concluding the chapter in Section 3.5, we present several numerical experiments in Section 3.4 to confirm the correct behavior of the proposed recovery methods. We also empirically investigate the impact of inaccurate prior information about the sparsity level of target vectors. Finally, we consider the influence of adversarial post-quantization noise on the recovery performance and demonstrate how to harden the proposed algorithms to render them noise-resilient, provided one has access to an estimate on the number of erroneous measurements.

3.1 Compressive Sampling of Frequency-Sparse Signals

The starting assumption in compressed sensing and its extensions is that a vector $\mathbf{z} \in \mathbb{C}^d$ exhibits some type of low-complexity structure, which one aims to exploit in order to reconstruct $\mathbf{z}$ from $m < d$ measurements. Oftentimes, it is further assumed that this low-complexity structure only reveals itself after expressing $\mathbf{z}$ in a suitable basis. Some of the most prominent examples of low-complexity bases (or more generally frames or overcomplete dictionaries) are the DFT, DCT or wavelet bases like the Haar basis, as well as the extended family of $*$-let transforms such as curvelets [**CD99**], noiselets [**CGM01**] or shearlets [**KL12**]. In this chapter, we assume that $\mathbf{z}$ represents the discretized version of a time domain signal after analog-to-digital conversion. Moreover, we assume that signals of interest have a sparse representation in the (continuous) frequency domain. It will therefore prove useful for our purposes to first consider the continuous representation of $\mathbf{z}$. Denote by $\mathcal{F}\colon \mathcal{T} \to \mathcal{T}$ the Fourier transform operator on the space of tempered distributions[1]

[1] A tempered distribution is a complex-valued continuous linear functional on the space of Schwartz functions $\mathcal{S}(\mathbb{R}) := \{\varphi\colon \mathbb{R} \to \mathbb{C} \mid \forall m, n \in \mathbb{N} : x^m \partial_x^n \varphi(x) \to 0 \text{ as } x \to \pm\infty\}$.

$\mathcal{T} = \{T\colon \mathcal{S}(\mathbb{R}) \to \mathbb{C}\}$, and consider the space of complex-valued band-limited functions

$$B_{\mathcal{F}}([-f_{\mathrm{b}}, f_{\mathrm{b}}]) := \{u\colon \mathbb{R} \to \mathbb{C} \mid \mathcal{F}u(f) = 0 \; \forall |f| > f_{\mathrm{b}}\}.$$

According to the Nyquist-Shannon sampling theorem, every function $u \in B_{\mathcal{F}}([-f_{\mathrm{b}}, f_{\mathrm{b}}])$ can be reconstructed from a discrete-time representation of the signal if it is sampled at a rate $f_{\mathrm{r}} \geq 2f_{\mathrm{b}}$. To sample functions in $B_{\mathcal{F}}([-f_{\mathrm{b}}, f_{\mathrm{b}}])$, we now define the sampling operator $\mathcal{A}_{d,f_{\mathrm{r}}}\colon B_{\mathcal{F}}([-f_{\mathrm{b}}, f_{\mathrm{b}}]) \to \mathbb{C}^d$ with sampling rate $f_{\mathrm{r}} = 1/T_{\mathrm{r}}$ to produce vectors of the form $\mathbf{z}_t = \mathcal{A}_{d,f_{\mathrm{r}}}u(t) = (u(t + nT_{\mathrm{r}}))_{n=0}^{d-1} \in \mathbb{C}^d$. In the following, we ignore the time instant t at which sampling begins and simply write $\mathbf{z}_t = \mathbf{z}$. Moreover, we assume that we acquire Nyquist-rate samples of elements in $B_{\mathcal{F}}([-f_{\mathrm{b}}, f_{\mathrm{b}}])$, meaning that we choose the sampling rate $f_{\mathrm{r}} = 2f_{\mathrm{b}}$ so that $B_{\mathcal{F}}([-f_{\mathrm{b}}, f_{\mathrm{b}}]) = B_{\mathcal{F}}([-f_{\mathrm{r}}/2, f_{\mathrm{r}}/2])$. We consider signals $u(t)$ which are formed as weighted superpositions of s complex exponentials according to

$$u(t) = \sum_{\nu=1}^{s} a_\nu e^{i2\pi f_\nu t} e^{i\varphi_\nu} \quad \text{with} \quad a_\nu \in \mathbb{R}, f_\nu \in \left[-\frac{f_{\mathrm{r}}}{2}, \frac{f_{\mathrm{r}}}{2}\right], \varphi_\nu \in [0, 2\pi)$$

and acquire d samples of $u(t)$ via $\mathcal{A}_{d,f_{\mathrm{r}}}$. If the frequencies $\{f_\nu\}_{\nu=1}^{s} \subset [-f_{\mathrm{r}}/2, f_{\mathrm{r}}/2]$ are integer multiples of the frequency resolution f_{r}/d, then the discrete Fourier transform of $\mathbf{z}$ is s-sparse, *i.e.*, $\mathbf{x} = \mathbf{F}_d^* \mathbf{z} = \mathbf{F}_d^* \mathcal{A}_{d,f_{\mathrm{r}}} u(t) \in \Sigma_s(\mathbb{C}^d)$ with

$$\mathbf{F}_d = \frac{1}{\sqrt{d}} (\exp(i2\pi\mu\nu/d))_{0 \leq \mu,\nu \leq d-1} \tag{3.1}$$

denoting the orthogonal DFT matrix.

In order to model the action of compressively measuring such vectors $\mathbf{z}$, the bulk of the literature on compressed sensing considers multiplication of $\mathbf{z}$ with a random matrix following a suitable distribution such as the Gaussian distribution or more generally subgaussian ensembles. As discussed in Section 2.2, such measurement operators are generally supported by strong theoretical foundations, establishing high-probability bounds for stable and robust recovery of sparse vectors. Unfortunately, measurement operators based on such unstructured random ensembles are generally hard to realize in physical systems as they significantly complicate the necessary hardware circuitry. However, randomness in CS acquisition systems has over the years proven an invaluable ingredient for constructing effective measurement operators with favorable empirical and theoretical reconstruction performance. If $u(t)$ is known to be band-limited, a natural sensing model more geared towards practical hardware implementations may assume that a sensing device randomly selects elements from a sequence of Nyquist-rate samples to produce a compressively sampled representation of $u(t)$. This is modeled by a random subsampling matrix $\mathbf{R}_\Omega := \mathrm{Id}_\Omega^\top \in \mathbb{R}^{|\Omega| \times d}$, where the index set $\Omega \subset [d]$ with $|\Omega| = m$ is chosen uniformly at random from $[d]$ without replacement. This results in the measurements

$$\mathbf{y} = \mathbf{R}_\Omega \mathbf{z} = \mathbf{R}_\Omega \mathbf{F}_d \mathbf{x} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m,$$

where the resulting measurement matrix $\mathbf{A} := \mathbf{R}_\Omega \mathbf{F}_d$ consists of m randomly selected rows of the DFT matrix $\mathbf{F}_d$. This is an example of a measurement matrix generated via random subsampling in a bounded orthonormal system (see [**FR13**, Chapter 12]).

3.1.1 Conjugate Symmetric Frequency-Sparse Signals

In the general setup described above, the band-limited signals u were assumed to be complex-valued. If we restrict attention to real-valued band-limited functions instead, then there is an additional structure that one may subsequently exploit in order to reduce the search space during the recovery of $\mathbf{x}$. This structure arises due to the fact that the Fourier transform of a real-valued signal exhibits a conjugate symmetric spectrum. This fact naturally extends to the sampled frequency domain representation $\mathbf{x}$ of $u(t)$, leading to the following definition.

Definition 3.1. *An even length vector* $\mathbf{x} = (x_1, \dots, x_d)^\top \in \mathbb{C}^d$ *is called* conjugate symmetric *if* $x_1 \in \mathbb{R}$ *and* $x_{i+1} = \overline{x_{d-i+1}}$ $\forall i = 1, \dots, d-1$.

Since d is assumed to be even, the coefficient $x_{d/2}$, also known as the *Nyquist coefficient*, is always real. Moreover, the set of all conjugate symmetric vectors, which we denote by $\mathbb{X}_d$, forms a linear subspace of $\mathbb{C}^d$. Most importantly, this means that linear combinations of elements in $\mathbb{X}_d$ remain in the space. This convenient fact will allow us to define an iterative recovery procedure which ensures that we never leave the search space $\mathbb{X}_d$. Note that Definition 3.1 implies a particular ordering of the elements of $\mathbf{x} \in \mathbb{X}_d$, where the first coefficient corresponds to the signal's DC component, followed by the positive frequency coefficients, followed in turn by the negative frequency coefficients in reverse order. This is in line with the definition of the DFT matrix defined in (3.1). Due to the fact that $\mathbb{X}_d$ forms a linear subspace of $\mathbb{C}^d$, which we may isomorphically identify with a corresponding subspace of $\mathbb{R}^{2d}$, enforcing a membership constraint of $\mathbb{X}_d$ via convex programming amounts to a simple linear constraint. This means that any recovery procedure proposed in the 1-bit compressed sensing literature based on convex programming is easily extended to the problem of conjugate symmetric vector recovery. This class of reconstruction schemes will form the baseline during our numerical experiments in Section 3.4.

3.2 Binary Iterative Hard Thresholding

As outlined in Section 2.3, the concept of quantization consistency is of key importance in the theory of quantized compressed sensing as it represents a sufficient condition for sparse vectors with the same quantized compressive measurements to be close to each other in the Euclidean sense. It is therefore desirable to find a way to express data fidelity in a way that can be promoted in a tractable manner during recovery. The convex programming approach due to Plan and Vershynin described in Section 2.3.1 solves this issue by appropriately relaxing the nonconvex constraint $\operatorname{sgn}(\mathbf{Ax}) = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}})$ into a convex constraint that is easily enforcible in a convex program. A similar line of research was presented in [**PV16**] for the generalized linear model $\mathbf{y} = f(\mathbf{Ax})$ with $f\colon \mathbb{R}^m \to \mathbb{R}^m$ denoting a (possibly random) nonlinearity. It is shown that if $\mathring{\mathbf{x}}$ belongs to some structure-promoting set $\mathcal{K} \subset \mathbb{R}^d$, then the vector $\mu\mathring{\mathbf{x}}$ minimizes the expected error $\mathbb{E}\|\mathbf{Ax} - f(\mathbf{A}\mathring{\mathbf{x}})\|_2^2$, provided that $\mathbf{A} \in \mathbb{R}^{m\times d}$ is a standard Gaussian matrix. The scaling parameter μ only depends on the nonlinearity f. This observation suggests solving the so-called *generalized LASSO*

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{Ax} - \mathbf{y}\|_2 \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{K} \end{aligned} \tag{$\mathrm{P}_{3.1}$}$$

as analyzed in [PV16]. For $f = \operatorname{sgn}$ acting element-wise on vectors, the corresponding analysis implies that minimizers of the above program are quantization-consistent in expectation by the scale invariance of the sgn-operator if $\mathcal{K}$ is a linear cone. Unfortunately, the interpretation of $\mu\mathring{\mathbf{x}}$ as a minimizer of Problem ($\mathrm{P}_{3.1}$) breaks down if $\mathbf{A}$ is not a Gaussian random matrix. We will therefore turn to an alternative way to promote quantization consistency for our purposes, which was initially proposed in [BB08] and later adopted in [Jac+13].

While convex programs such as Problem ($\mathrm{P}_{2.6}$) and Problem ($\mathrm{P}_{3.1}$) can be solved in polynomial time, the respective solvers employed for the task are usually still not efficient enough to enable time-critical reconstruction unless they are specifically designed for a particular problem instance. In the broader compressed sensing literature, this fueled research into more specialized iterative reconstruction schemes based, *e.g.*, on projected gradient methods like the well-known *iterative hard thresholding* (IHT) algorithm [BD09]. Given an estimate of the sparsity level s and linear measurements $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} \in \mathbb{C}^m$ of an s-sparse vector $\mathring{\mathbf{x}} \in \mathbb{C}^d$, the IHT algorithm aims to solve the nonconvex program

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \\ \text{s.t.} \quad & \mathbf{x} \in \Sigma_s(\mathbb{C}^d). \end{aligned}$$

To that end, IHT repeats the iteration rule

$$\mathbf{x}^{(n+1)} = \underset{\mathbf{u} \in \Sigma_s(\mathbb{C}^d)}{\operatorname{argmin}} \left\| \mathbf{x}^{(n)} - \lambda_n \mathbf{A}^*(\mathbf{A}\mathbf{x}^{(n)} - \mathbf{y}) - \mathbf{u} \right\|_2, \tag{3.2}$$

which corresponds to the orthogonal projection of the gradient descent update $\mathbf{x}^{(n)} - \lambda_n \nabla_{\mathbf{x}^{(n)}}(\frac{1}{2}\|\mathbf{A}\mathbf{x}^{(n)} - \mathbf{y}\|_2^2) = \mathbf{x}^{(n)} - \lambda_n \mathbf{A}^*(\mathbf{A}\mathbf{x}^{(n)} - \mathbf{y})$ with step-size λ_n on the set of s-sparse vectors. Despite the nonconvexity of the set $\Sigma_s(\mathbb{C}^d)$ as a union of s-dimensional coordinate subspaces, the projection admits a closed-form solution via the hard thresholding operator $\mathcal{H}_s \colon \mathbb{C}^d \to \Sigma_s(\mathbb{C}^d)$. One way to formalize the construction of $\mathcal{H}_s(\mathbf{x})$, which will be useful later on, is by considering the nonincreasing rearrangement $\check{\mathbf{x}}$ of $\mathbf{x}$ characterized by $\check{x}_1 \geq \check{x}_2 \geq \ldots \geq \check{x}_d$ with $\check{x}_i := |x_{\pi(i)}|$ and $\pi \colon [d] \to [d]$ a permutation. The vector $\mathcal{H}_s(\mathbf{x})$—the best s-sparse approximation of $\mathbf{x}$—then corresponds to the vector which agrees with $\mathbf{x}$ on the index set $S = \{\pi(1), \ldots, \pi(s)\}$ and vanishes identically on $\overline{S} = [d] \setminus S$. This construction also holds if the ℓ_2-norm in (3.2) is replaced by an arbitrary ℓ_p-norm with $p \geq 1$. This leads to the IHT update rule

$$\mathbf{x}^{(n+1)} = \mathcal{H}_s\Big(\mathbf{x}^{(n)} - \lambda_n \mathbf{A}^*(\mathbf{A}\mathbf{x}^{(n)} - \mathbf{y})\Big). \tag{3.3}$$

We emphasize that general convergence results of the projected gradient (and subgradient) method rely on the convexity assumption of both the objective function and the feasible set.[2] This means that convergence of methods which project on Σ_s is not generally guaranteed and—if possible at all—has to be established by additional assumptions on the objective function. For the IHT algorithm, this is possible under the assumption that

[2] Convergence of the projected gradient method can be established via convergence of the *proximal gradient method* [PB14] for functionals of the form $f(\mathbf{x}) + g(\mathbf{x})$ where $f, g \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ are closed proper convex functions, and f is differentiable. The proximal gradient method then repeats the update rule $\mathbf{x}^{(n+1)} = \operatorname{prox}_{\lambda_n g}(\mathbf{x}^{(n)} - \lambda_n \nabla f(\mathbf{x}^{(n)}))$, where the proximal operator is defined as $\operatorname{prox}_{\lambda g}(\mathbf{x}) = \operatorname{argmin}_{\mathbf{u}} \{g(\mathbf{u}) + 1/(2\lambda)\|\mathbf{x} - \mathbf{u}\|_2^2\}$. For $g = \iota_C$ with ι_C denoting the indicator function of a convex set $C \subset \mathbb{R}^d$, this algorithm reduces to the projected gradient method.

$\mathbf{A}$ satisfies the restricted isometry property with small enough RIP constant δ. In this case one can show that (3.3) converges to $\mathring{\mathbf{x}}$ as $n \to \infty$ for various different choices of the step-size parameter λ_n controlling the convergence rate of the algorithm [Blu12].

Inspired by this reconstruction algorithm, Jacques *et al.* set out to define an appropriate data fidelity measure in the 1-bit CS setting which admits a similar iterative algorithm. To that end, they propose to minimize the so-called *one-sided ℓ_1-norm* of the vector $\mathbf{y} \circ \mathbf{A}\mathbf{x} = \operatorname{diag}\{\mathbf{y}\}\mathbf{A}\mathbf{x}$ serving as a quantization consistency indicator and define the objective function $J_1(\mathbf{x}) := \|[\mathbf{y} \circ \mathbf{A}\mathbf{x}]_-\|_1$ where $[\cdot]_- = \min\{0, \cdot\}$ denotes the negative part of a real number.[3] Given a vector $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}) \in \{\pm 1\}^m$ of binary quantized measurements, the function J_1 accumulates all coordinates of the linear measurements $\mathbf{A}\mathbf{x}$ whose signs disagree with the corresponding entries in $\mathbf{y}$. Clearly, the composition of the nondifferentiable function $[\cdot]_-$ and the ℓ_1-norm results in a nondifferentiable function. Since the function J_1 is convex in $\mathbf{x}$, however, Jacques *et al.* propose a *projected subgradient method* to recover $\mathring{\mathbf{x}}$ from its binary observations.

The idea of subgradients and the associated concept of *subdifferentials* generalize the notion of gradients to nondifferentiable functions. The definition is motivated by the first-order convexity condition of smooth convex functions $f\colon D \subseteq \mathbb{R}^d \to \mathbb{R}$, which states that for a fixed vector $\mathbf{x} \in D$, the linear function $f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle$ is a global underestimator of f, *i.e.*,

$$f(\mathbf{z}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle \quad \forall \mathbf{z} \in D.$$

Dropping the smoothness requirement on f, this leads to the following definition [Roc15].

Definition 3.2 (Subgradient and subdifferential)**.** *A vector $\mathbf{g} \in \mathbb{R}^d$ is called a* subgradient *of $f\colon D \subseteq \mathbb{R}^d \to \mathbb{R}$ at $\mathbf{x} \in D$ if*

$$f(\mathbf{z}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{z} - \mathbf{x} \rangle \quad \forall \mathbf{z} \in D. \tag{3.4}$$

The collection of all vectors $\mathbf{g}$ satisfying condition (3.4), denoted $\partial f(\mathbf{x})$, is called the subdifferential *of f at $\mathbf{x} \in D$:*

$$\partial f(\mathbf{x}) := \Big\{ \mathbf{g} \in \mathbb{R}^d : f(\mathbf{z}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{z} - \mathbf{x} \rangle \ \forall \mathbf{z} \in D \Big\}.$$

Remark 3.3. *If f is convex and differentiable at a point $\mathbf{x}$, the gradient $\nabla f(\mathbf{x})$ is the unique subgradient satisfying condition (3.4), i.e., the subdifferential of f at $\mathbf{x}$ is the singleton set $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$. Moreover, while every convex function admits a nonempty subdifferential set, the same does not hold for arbitrary functions. This is in stark contrast to smooth nonconvex functions, which always admit a gradient by definition.*

Equipped with this concept, Jacques *et al.* establish the following result in [Jac+13] for which we provide a simple constructive proof in the interest of self-containedness.

Lemma 3.4. *The vector $p(\mathbf{x}) = \frac{1}{2}\mathbf{A}^\top(\operatorname{sgn}(\mathbf{A}\mathbf{x}) - \mathbf{y})$ is a subgradient of the functional $J_1(\mathbf{x}) = \|[\mathbf{y} \circ \mathbf{A}\mathbf{x}]_-\|_1$.*

[3] When applied to vectors, we use the convention that $[\cdot]_-$ acts element-wise.

Proof. We start by rewriting the functional J_1 at a point $\mathbf{x} \in \mathbb{R}^d$ as

$$J_1(\mathbf{x}) = \left\| [\mathbf{y} \circ \mathbf{A}\mathbf{x}]_- \right\|_1 = \sum_{i=1}^m \left| [y_i \langle \mathbf{a}_i, \mathbf{x} \rangle]_- \right| = -\sum_{i=1}^m [y_i \langle \mathbf{a}_i, \mathbf{x} \rangle]_- = \sum_{i=1}^m [-y_i \langle \mathbf{a}_i, \mathbf{x} \rangle]_+$$

where $[\cdot]_+ = \max\{0, \cdot\}$ denotes the positive part. Since one generally has for convex functions $f_1, \ldots, f_m \colon \mathbb{R}^d \to \mathbb{R}$ that $\partial(f_1 + \ldots + f_m)(\mathbf{x}) = \partial f_1(\mathbf{x}) + \ldots + \partial f_m(\mathbf{x})$[4] (see for instance [Roc15, Chapter 23]), this yields

$$\partial J_1(\mathbf{x}) = \sum_{i=1}^m \partial [-y_i \langle \mathbf{a}_i, \mathbf{x} \rangle]_+ = -\sum_{i=1}^m y_i \mathbf{a}_i \, \partial [z]_+ \Big|_{z = -y_i \langle \mathbf{a}_i, \mathbf{x} \rangle}$$

where the last step follows by the chain rule of subdifferential calculus for affine transformations [Roc15]. Clearly, the subdifferential of the (convex) function $[\cdot]_+ = \max\{0, \cdot\}$ is given by

$$\partial [z]_+ = \begin{cases} \{0\}, & z < 0, \\ [0, 1], & z = 0, \\ \{1\}, & z > 0. \end{cases}$$

It therefore follows that the vector $p(\mathbf{x}) := -\sum_{i=1}^m y_i \mathbf{a}_i \mathbb{1}_{\{-y_i \langle \mathbf{a}_i, \mathbf{x} \rangle > 0\}}$ is an element of the subdifferential $\partial J_1(\mathbf{x})$ where

$$\begin{aligned} p(\mathbf{x}) &= -\sum_{i=1}^m y_i \mathbf{a}_i \mathbb{1}_{\{-y_i \langle \mathbf{a}_i, \mathbf{x} \rangle > 0\}} \\ &= -\sum_{i=1}^m y_i \mathbf{a}_i \mathbb{1}_{\{y_i \langle \mathbf{a}_i, \mathbf{x} \rangle < 0\}} \\ &= -\mathbf{A}^\top \operatorname{diag}\{y_i\}_{i=1}^m \left(\mathbb{1}_{\{y_i \langle \mathbf{a}_i, \mathbf{x} \rangle < 0\}} \right)_{i=1}^m \\ &= -\frac{1}{2} \mathbf{A}^\top \operatorname{diag}\{\mathbf{y}\} (\mathbf{1} - \operatorname{sgn}(\mathbf{y} \circ \mathbf{A}\mathbf{x})) \\ &= \frac{1}{2} \mathbf{A}^\top (\mathbf{y} \circ \operatorname{sgn}(\mathbf{y} \circ \mathbf{A}\mathbf{x}) - \mathbf{y}) \\ &= \frac{1}{2} \mathbf{A}^\top (\operatorname{sgn}(\mathbf{A}\mathbf{x}) - \mathbf{y}) \end{aligned}$$

as announced. □

The *binary iterative hard thresholding* (BIHT) algorithm now proceeds by alternating between a subgradient step of the objective function J_1 and a projection step on the set of s-sparse vectors $\Sigma_s(\mathbb{R}^d)$ by means of the hard thresholding operator $\mathcal{H}_s$. The algorithm automatically terminates once a quantization-consistent vector $\mathbf{x}^{(n+1)}$ has been found since one has $p(\mathbf{x}^{(n)}) = \mathbf{0}$ in that case, which means BIHT repeats the iteration $\mathbf{x}^{(n+1)} = \mathcal{H}_s(\mathbf{x}^{(n)})$. Considering that any s-sparse vector is a fixed-point of $\mathcal{H}_s$, this implies that the algorithm automatically stalls if the normalized *Hamming distance* $\Delta_{\mathrm{H}}(\mathbf{y}, \operatorname{sgn}(\mathbf{A}\mathbf{x}^{(n)}))$ with

$$\Delta_{\mathrm{H}} \colon \{\pm 1\}^m \times \{\pm 1\}^m \to [0, 1] \colon (\mathbf{u}, \mathbf{v}) \mapsto \Delta_{\mathrm{H}}(\mathbf{u}, \mathbf{v}) := \frac{1}{m} \sum_{i=1}^m \mathbb{1}_{\{u_i \neq v_i\}} = \frac{1}{2m} \|\mathbf{u} - \mathbf{v}\|_1$$

[4] The first sum is to be understood in a pointwise fashion, while the second one corresponds to the Minkowski sum $A + B := \{a + b : a \in A, b \in B\}$ of sets.

vanishes, yielding a natural stopping criterion for the BIHT algorithm. The full algorithm listing is given in Algorithm 1. Note that the step-size $\lambda_n > 0$ of the subgradient update $\mathbf{u}^{(n+1)} = \mathbf{x}^{(n)} - \lambda_n p(\mathbf{x}^{(n)})$ in Algorithm 1 is assumed to be fixed as $\lambda_n = \lambda = 2$ since the algorithm appears to be independent of the choice of step-size according to numerical experiments conducted in [Jac+13] and [JDV13]. As pointed out before, the idea of using the sign-violation vector $\mathbf{y} \circ \mathbf{A}\mathbf{x}$ in the objective function goes back to the original work [BB08] due to Boufounos and Baraniuk. They propose to minimize the smooth convex objective function $J_2(\mathbf{x}) := \frac{1}{2}\|[\mathbf{y} \circ \mathbf{A}\mathbf{x}]_-\|_2^2$ with gradient $\nabla J_2(\mathbf{x}) = \mathbf{A}^\top \operatorname{diag}\{\mathbf{y}\}[\mathbf{y} \circ \mathbf{A}\mathbf{x}]_-$ in the context of their renormalized fixed-point iteration algorithm. This objective function has also been considered by Jacques *et al.* in the formulation of what is commonly referred to as BIHT-ℓ_2 [Jac+13]. Unfortunately, theoretical performance analyses or even convergence results of either algorithm have so far eluded researchers. That being said, overwhelming numerical evidence indicates that the BIHT algorithm generally terminates after less than 100 iterations in the noiseless setting. The same does not necessarily hold when one considers either additive pre-quantization noise or adversarial post-quantization bit flips. However, in this situation one may turn to the so-called *adaptive outlier pursuit* (AOP) algorithm [YYO12] to adaptively identify and subsequently correct possibly erroneous bit positions to regain some of the performance lost due to noise.

Algorithm 1 Binary Iterative Hard Thresholding (BIHT)

Input: $\mathbf{A} \in \mathbb{R}^{m \times d}$, $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}) \in \{-1, 1\}^m$, $s \in [d]$
Initialize: $\mathbf{x}^{(0)} \leftarrow \mathbf{0}$, $n \leftarrow 0$
  do
    $\mathbf{u}^{(n+1)} \leftarrow \mathbf{x}^{(n)} - \mathbf{A}^\top(\operatorname{sgn}(\mathbf{A}\mathbf{x}^{(n)}) - \mathbf{y})$ ▷ Subgradient step
    $\mathbf{x}^{(n+1)} \leftarrow \mathcal{H}_s(\mathbf{u}^{(n+1)})$ ▷ Projection on $\Sigma_s(\mathbb{R}^d)$
    $n \leftarrow n + 1$
  while $\Delta_{\mathrm{H}}(\mathbf{y}, \operatorname{sgn}(\mathbf{A}\mathbf{x}^{(n)})) > 0$ **and** $n < n_{\max}$
Output: $\mathbf{x}^{(n)}/\|\mathbf{x}^{(n)}\|_2$ ▷ Projection onto the unit sphere

3.3 Conjugate Symmetric Binary Iterative Hard Thresholding

We now turn to the task of developing an extension of the BIHT algorithm for the recovery of sparse conjugate symmetric vectors. Again, we assume that we acquire Nyquist-rate samples of elements in $B_{\mathcal{F}}([-f_{\mathrm{b}}, f_{\mathrm{b}}])$, *i.e.*, we fix the sampling rate to be $f_{\mathrm{r}} = 2f_{\mathrm{b}}$ so that $B_{\mathcal{F}}([-f_{\mathrm{b}}, f_{\mathrm{b}}]) = B_{\mathcal{F}}([-f_{\mathrm{r}}/2, f_{\mathrm{r}}/2])$. The signal class of sparse conjugate symmetric vectors originates from sampling band-limited functions $u \in B_{\mathcal{F}}([-f_{\mathrm{r}}/2, f_{\mathrm{r}}/2])$ of the form

$$u(t) = \sum_{\nu=1}^{l} a_\nu \cos(2\pi f_\nu t + \phi_\nu),\ a_\nu \in \mathbb{R},\ f_\nu \in \frac{f_{\mathrm{r}}}{d}\left\{-\frac{d}{2}, \ldots, \frac{d}{2}\right\},\ \varphi_\nu \in [0, 2\pi).$$

Sampling such signals via the operator $\mathcal{A}_{d,f_{\mathrm{r}}}$ results in time domain vectors $\mathbf{z}$ with an s-sparse representation $\mathbf{x} = \mathbf{F}_d^* \mathbf{z} \in \Sigma_s(\mathbb{X}_d)$ with $s = 2l$ due to the symmetry structure on $\mathbb{X}_d$. As remarked before, it is easy to verify that $\mathbb{X}_d$ forms a linear subspace of $\mathbb{C}^d$. In that sense, any iterative algorithm of the form $\mathbf{x}^+ = \mathbf{x} + \lambda\mathbf{p}$, where $\mathbf{x} \in \mathbb{X}_d$ is the current iterate,

$\mathbf{p} \in \mathbb{X}_d$ a search direction and $\lambda > 0$ a step-size, will yield a valid conjugate symmetric vector $\mathbf{x}^+$. This will be a crucial property in our conjugate symmetric modification of the BIHT algorithm. We adopt the idea of subsampled bounded orthonormal systems to model the acquisition in the time domain as outlined in Section 3.1. This means that we take measurements of the form $\mathbf{R}_\Omega \mathbf{F}_d \mathbf{x} = \mathrm{Id}_\Omega^\top \mathbf{F}_d \mathbf{x} = \mathbf{A}\mathbf{x}$ with $\Omega \subset [d]$ denoting a random index set of size $|\Omega| = m$ on which we conceptually subsample the Nyquist-rate sequence $(z_k)_{k=1}^d$ represented by the vector $\mathbf{z} \in \mathbb{R}^d$ of time domain samples. In order to ease energy demands of the ADCs employed in the sampling system, we adopt the 1-bit CS paradigm by only retaining the sign information about the subsampled waveform by acquiring measurements

$$\mathbf{y} = \operatorname{sgn}(\mathbf{R}_\Omega \mathbf{F}_d \mathbf{x}) = \operatorname{sgn}(\mathbf{A}\mathbf{x}) \in \{\pm 1\}^m.$$

We now turn to the modification of the BIHT algorithm for the recovery of such vectors $\mathbf{x} \in \mathbb{X}_d$ from their corresponding 1-bit time domain measurements. There are two issues to overcome in this regard. First, the concept of subdifferentials does not directly translate to real-valued functionals $f\colon \mathbb{C}^d \to \mathbb{R}$ defined on $\mathbb{C}^d$ since one needs to make sense of statements of the form $f(\mathbf{x}) \geq f(\mathbf{z}) + \langle \mathbf{g}, \mathbf{x} - \mathbf{z} \rangle$ to define subgradients. This is not well-defined for vectors $\mathbf{x}, \mathbf{z}, \mathbf{g} \in \mathbb{C}^d$ since $\mathbb{C}$ is not totally ordered. Secondly, the iterative nature of BIHT necessitates that the hard thresholding operator $\mathcal{H}_s$ is appropriately modified so as to produce s-sparse vectors belonging to the subspace $\mathbb{X}_d$ rather than $\mathbb{C}^d$. Without such a modification, $\mathcal{H}_s$ may produce vectors $\mathbf{x}'$ outside of $\mathbb{X}_d$ such that the operation $\operatorname{sgn}(\mathbf{A}\mathbf{x}')$ may not be well-defined. This is rooted in the fact that $\mathbf{A} = \mathbf{R}_\Omega \mathbf{F}_d$ acts as a subsampled inverse DFT, which might result in $\mathbf{A}\mathbf{x}'$ being complex. These issues are addressed in the following two sections.

3.3.1 Reformulation of the Subgradient Iteration

In order to adapt the BIHT algorithm for the recovery of frequency-sparse signals, we use the natural vector space identification of $\mathbb{C}^d$ with $\mathbb{R}^{2d}$. The BIHT algorithm aims to minimize the functional

$$J_1(\mathbf{x}) = \left\| [\mathbf{y} \circ \mathbf{A}\mathbf{x}]_- \right\|_1 = -\sum_{k=1}^m [y_k \langle \mathbf{a}_k, \mathbf{x} \rangle]_-$$

by means of the projected subgradient method where the operator $[\cdot]_- = \min\{0, \cdot\}$ is only defined for real arguments. However, since the measurement matrix considered here is of the form $\mathbf{A} = \mathrm{Id}_S^\top \mathbf{F}_d$, we have for $\mathbf{x} \in \mathbb{X}_d$ and with $\langle \cdot, \cdot \rangle$ denoting the standard inner product on $\mathbb{R}^d$ extended to $\mathbb{C}^d$ that

$$\begin{aligned}
\langle \mathbf{a}_k, \mathbf{x} \rangle &= \langle \mathfrak{Re}(\mathbf{a}_k), \mathfrak{Re}(\mathbf{x}) \rangle - \langle \mathfrak{Im}(\mathbf{a}_k), \mathfrak{Im}(\mathbf{x}) \rangle + i \underbrace{\left(\langle \mathfrak{Im}(\mathbf{a}_k), \mathfrak{Re}(\mathbf{x}) \rangle + \langle \mathfrak{Re}(\mathbf{a}_k), \mathfrak{Im}(\mathbf{x}) \rangle \right)}_{=0 \text{ since } \mathbf{x} \in \mathbb{X}_d} \\
&= \langle \mathfrak{Re}(\mathbf{a}_k), \mathfrak{Re}(\mathbf{x}) \rangle - \langle \mathfrak{Im}(\mathbf{a}_k), \mathfrak{Im}(\mathbf{x}) \rangle \\
&= \left\langle \begin{pmatrix} \mathfrak{Re}(\mathbf{a}_k) \\ -\mathfrak{Im}(\mathbf{a}_k) \end{pmatrix}, \begin{pmatrix} \mathfrak{Re}(\mathbf{x}) \\ \mathfrak{Im}(\mathbf{x}) \end{pmatrix} \right\rangle \\
&=: \langle \hat{\mathbf{a}}_k, \hat{\mathbf{x}} \rangle.
\end{aligned}$$

With the lifted objective function

$$\hat{J}_1 \colon \mathbb{R}^{2d} \to \mathbb{R} \colon \hat{\mathbf{x}} \mapsto -\sum_{k=1}^{m} [y_k \langle \hat{\mathbf{a}}_k, \hat{\mathbf{x}} \rangle]_-,$$

it follows from Lemma 3.4 for $\hat{\mathbf{A}} \in \mathbb{R}^{m \times 2d}$ with rows $\{\hat{\mathbf{a}}_k\}_{k=1}^m$ that

$$\hat{p}(\hat{\mathbf{x}}) = \frac{1}{2} \hat{\mathbf{A}}^\top \big(\operatorname{sgn}(\hat{\mathbf{A}} \hat{\mathbf{x}}) - \mathbf{y} \big)$$

is a subgradient of $\hat{J}_1$, which translates to

$$\begin{aligned} p_{\mathbb{X}}(\mathbf{x}) &:= \frac{1}{2} \mathbf{A}^* (\operatorname{sgn}(\mathfrak{Re}(\mathbf{A}) \mathfrak{Re}(\mathbf{x}) - \mathfrak{Im}(\mathbf{A}) \mathfrak{Im}(\mathbf{x})) - \mathbf{y}) \\ &= \frac{1}{2} \mathbf{A}^* (\operatorname{sgn}(\mathbf{A}\mathbf{x}) - \mathbf{y}) \end{aligned}$$

in the complex domain since $\mathbf{x} \in \mathbb{X}_d$ and therefore $\mathbf{A}\mathbf{x} \in \mathbb{R}^m$. Unsurprisingly, this is in direct accordance to the real-valued case discussed in Section 3.2 where the adjoint operator $\mathbf{A}^*$ is replaced with the transpose $\mathbf{A}^\top$. In the same fashion, the conjugate symmetric gradient ∇J_2 of the smooth BIHT-ℓ_2 objective function extended to the domain $\mathbb{X}_d$ can be derived as $\nabla J_2(\mathbf{x}) = \mathbf{A}^* \operatorname{diag}\{\mathbf{y}\}[\mathbf{y} \circ \mathbf{A}\mathbf{x}]_-$. Also note that, given a measurement matrix of the form $\mathbf{A} = \mathbf{R}_\Omega \mathbf{F}_d = \mathrm{Id}_\Omega^\top \mathbf{F}_d$, the subgradient $p_{\mathbb{X}}(\mathbf{x})$ can be interpreted as the DFT of the real-valued signal $\frac{1}{2}\mathrm{Id}_\Omega(\operatorname{sgn}(\mathbf{A}\mathbf{x}) - \mathbf{y})$, which implies $p_{\mathbb{X}}(\mathbf{x}) \in \mathbb{X}_d$. Again, a similar argument holds for $\nabla J_2(\mathbf{x})$.

3.3.2 The Hard Thresholding Operator for Conjugate Symmetric Vectors

If the signal of interest $\mathbf{x}$ belongs to the set $\Sigma_s(\mathbb{X}_d)$, the support of its real and imaginary parts are identical[5], *i.e.*, we are looking for $2s$-sparse vectors when minimizing the objective function $\hat{J}_1$ over $\hat{\mathbf{x}} \in \mathbb{R}^{2d}$. In other words, we might naively choose the next iterate $\hat{\mathbf{x}}^+$ of the modified BIHT algorithm as the best $2s$-sparse approximation of the subgradient update $\hat{\mathbf{x}} - \hat{p}(\hat{\mathbf{x}})$, namely $\hat{\mathbf{x}}^+ = \mathcal{H}_{2s}(\hat{\mathbf{x}} - \hat{p}(\hat{\mathbf{x}}))$. Unfortunately, this strategy does not respect the conjugate symmetric structure in the solution such that the support of the first and last d components of $\hat{\mathbf{x}}^+$ (the real and imaginary parts of $\mathbf{x}^+$) obtained in this way will rarely agree. Instead, we must find the best s-sparse approximation in the space $\mathbb{X}_d$ explicitly. In this case, we need to modify the thresholding strategy so as not to destroy the conjugate symmetry of the input vector. Formally, we need to evaluate the operator $\mathcal{H}_{s,p}^{\mathbb{X}} \colon \mathbb{X}_d \to \Sigma_s(\mathbb{X}_d)$ with

$$\mathcal{H}_{s,p}^{\mathbb{X}}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{u} \in \Sigma_s(\mathbb{X}_d)} \|\mathbf{x} - \mathbf{u}\|_p.$$

Unlike for $\mathcal{H}_s$, however, the optimum will not be attained at the same $\mathbf{x} \in \Sigma_s(\mathbb{X}_d)$ for every $p \geq 1$, which we indicate by the subscript p in the notation $\mathcal{H}_{s,p}^{\mathbb{X}}$. In the following, we refer to the element z_1 and $z_{d/2+1}$ of $\mathbf{x} \in \mathbb{X}_d$ as *DC* and *Nyquist* coefficient, respectively. Given the definition of $\mathbb{X}_d$, both coefficients are always real-valued and therefore do not appear

[5] This follows immediately from the definition of the space $\mathbb{X}_d$.

in symmetric pairs. To construct $\mathcal{H}_{s,p}^{\mathbb{X}}(\mathbf{x})$, we turn to the nonincreasing rearrangement $\check{\mathbf{x}}$ of $\mathbf{x}$ with $\check{x}_1 \geq \check{x}_2 \geq \ldots \geq \check{x}_d$ and $\check{x}_i = |x_{\pi(i)}|$ for $\pi\colon [d] \to [d]$.

For even s, it is necessary to guarantee that the permutation π of the nonincreasing rearrangement always maps the DC and Nyquist coefficients to consecutive indices. Since all other entries appear in conjugate symmetric pairs of equal modulus, selecting the first s components then guarantees that we either end up with $s/2$ or $(s-2)/2 = s/2-1$ conjugate symmetric pairs in $\mathcal{H}_{s,p}^{\mathbb{X}}(\mathbf{x})$. To that end, we start with the nonincreasing rearrangement[6] $\check{\mathbf{x}}$ and change the permutation π to π' such that the pair $(|x_1|, |x_{d/2+1}|)$ appears before the smallest index $i \in [d] \setminus \{\pi(1), \pi(d/2+1)\}$ which satisfies $|x_1|^p + |x_{d/2+1}|^p \geq 2\check{x}_i^p$. This condition is motivated by the fact that the contribution of the pair $(|x_1|, |x_{d/2+1}|)$ to the best conjugate symmetric s-term approximation error $\sigma_s^{\mathbb{X}}(\mathbf{x})_p := \inf_{\mathbf{u}\in\Sigma_s(\mathbb{X}_d)} \|\mathbf{x}-\mathbf{u}\|_p$ is $|x_1|^p + |x_{d/2+1}|^p$ if it is not included in the approximation. If no index i exists for which the above condition holds, the pair of DC and Nyquist coefficients is located at the end of the rearrangement $\check{\mathbf{x}}$ such that the permutation π does not need to be modified. This also holds if DC and Nyquist coefficients already happen to be positioned at consecutive coordinates in $\check{\mathbf{x}}$.

For odd s, it is necessary that either the DC or Nyquist coefficient is included in the approximation but never both. This can be seen as follows. Assume that the indices $i_{\mathrm{DC}} := \pi(1)$ and $i_{\mathrm{Nyq}} := \pi(d/2+1)$ of the DC and Nyquist coefficient in the nonincreasing rearrangement satisfy $i_{\mathrm{DC}}, i_{\mathrm{Nyq}} \leq s$ or $i_{\mathrm{DC}}, i_{\mathrm{Nyq}} > s$. Then either both the DC and Nyquist coefficient will be included in the approximation or neither of them. Since s is odd this means that we would invariably split up one conjugate symmetric pair by only retaining the first s coefficients of $\check{\mathbf{x}}$, thereby destroying the required conjugate symmetry of the approximation. It is therefore necessary to modify the permutation π as follows. If we have $\min\{|x_1|, |x_{d/2+1}|\} = \check{x}_i$ with $i \leq s$, then the permutation π is changed so that the element $\check{x}_i$ is mapped to some index $i' > s$ that is not included in the approximation, say $i' = d$. On the other hand, if we have $\max\{|x_1|, |x_{d/2+1}|\} = \check{x}_j$ with $j > s$, then π is modified such that the element $\check{x}_j$ is located before the smallest index $j' \in [d]$ for which $\check{x}_j \geq 2\check{x}_{j'}$ holds. This strategy guarantees that we always retain the coefficient in the approximation that reduces the best s-sparse approximation error the most. Note that the construction for odd sparsity levels s is independent of the ℓ_p-norm used in the definition of $\mathcal{H}_{s,p}^{\mathbb{X}}$ so that we have $\mathcal{H}_{s,p}^{\mathbb{X}} = \mathcal{H}_s^{\mathbb{X}}\ \forall p \geq 1$.

While the operator $\mathcal{H}_{s,p}^{\mathbb{X}}$ was introduced here for the use in the BIHT algorithm, it is also immediately applicable to most iterative CS recovery algorithms that rely on hard thresholding to enforce a certain sparsity level of solutions. While conjugate symmetry is easily enforced via convex programming (cf. Section 3.4), the same does not necessarily hold for greedy or iterative algorithms. However, it is easy to see that swapping out the operator $\mathcal{H}_s$ for $\mathcal{H}_{s,p}^{\mathbb{X}}$ in the *compressive sampling matching pursuit* (CoSaMP) [**NT09**], *quantized iterative hard thresholding* (QIHT) [**JDV13**], IHT [**BD09**] and *hard thresholding pursuit* (HTP) [**Fou11**] algorithms allows them to be used for the recovery of frequency-sparse signals from compressive time domain measurements without further modifications.

We summarize the resulting algorithm, which we term *conjugate symmetric binary iterative hard thresholding* (CS-BIHT), in Algorithm 2. Moreover, we refer to the ℓ_2-

[6] When implementing the operator $\mathcal{H}_{s,p}^{\mathbb{X}}$, it is vital to only consider the first $d/2+1$ coefficients of $\mathbf{x}$ in the nonincreasing rearrangement. This is necessary to avoid that different conjugate pairs with the same modulus get mixed up by the sorting algorithm used to construct $\check{\mathbf{x}}$.

smoothed variant of CS-BIHT as CS-BIHT-ℓ_2. This algorithm replaces the subgradient update $\mathbf{u}^+ = \mathbf{x} - \lambda\mathbf{A}^*(\operatorname{sgn}(\mathbf{Ax}) - \mathbf{y})$ in Algorithm 2 with the gradient update $\mathbf{u}^+ = \mathbf{x} - \lambda\mathbf{A}^*\operatorname{diag}\{\mathbf{y}\}[\mathbf{y} \circ \mathbf{Ax}]_-$ corresponding to the smooth one-sided objective function $J_2(\mathbf{x}) = \frac{1}{2}\|[\mathbf{y} \circ \mathbf{Ax}]_-\|_2^2$.

Algorithm 2 Conjugate Symmetric Binary Iterative Hard Thresholding (CS-BIHT)

Input: $\mathbf{A} = \mathbf{R}_\Omega\mathbf{F}_d \in \mathbb{C}^{m\times d}$, $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}) \in \{\pm 1\}^m$, $s \in [d]$
Initialize: $\mathbf{x}^{(0)} \leftarrow \mathbf{0}$, $n \leftarrow 0$
do
$\mathbf{u}^{(n+1)} \leftarrow \mathbf{x}^{(n)} - \mathbf{A}^*(\operatorname{sgn}(\mathbf{Ax}^{(n)}) - \mathbf{y})$ ▷ Subgradient step
$\mathbf{x}^{(n+1)} \leftarrow \mathcal{H}^{\mathbb{X}}_{s,p}(\mathbf{u}^{(n+1)})$ ▷ Projection onto $\Sigma_s(\mathbb{X}_d)$
$n \leftarrow n + 1$
while $\Delta_{\mathrm{H}}(\mathbf{y}, \operatorname{sgn}(\mathbf{Ax}^{(n)})) > 0$ **and** $n < n_{\max}$
Output: $\mathbf{x}^{(n)}/\|\mathbf{x}^{(n)}\|_2$ ▷ Projection onto the unit sphere

3.3.3 Extension to Oversampled Time Domain Measurements

While acquiring more measurements than the dimension of the signal space is of no interest in classical (linear) compressed sensing, the same does not apply to the 1-bit or more generally the quantized compressed sensing observation model. This is due to the fact that any nonlinear quantization function $Q\colon \mathbb{R}^m \to \mathbb{R}^m$ acting on the linear measurements $\mathbf{Ax} \in \mathbb{R}^m$ prevents inversion of the equation system $\mathbf{y} = Q(\mathbf{Ax})$ to solve for $\mathbf{x} \in \mathbb{R}^d$ even if $m > d$. This is precisely the situation we find ourselves in if Q corresponds to the binary sgn function. Despite the fact that acquiring more than d samples does not imply that the inverse problem can be solved exactly, adding more sign measurements is still expected to improve estimation accuracy. This is especially desirable in case of the 1-bit acquisition model where sampling devices are assumed to be highly energy-efficient such that acquiring more measurements does not cause an excessive increase in power consumption. For these reasons, we now consider the reconstruction of frequency-sparse band-limited signals from super-Nyquist real-valued time domain measurements. In case that measurement matrices are based on purely random, unstructured ensembles such as Gaussian random matrices, the idea of oversampling is as simple as drawing independent random vectors from the respective distribution and appending them as additional rows to an existing measurement matrix. Unfortunately, the situation is less straightforward if the measurement matrix $\mathbf{A} \in \mathbb{C}^{m\times d}$ for $m < d$ is based on a unitary matrix like the DFT matrix. This was the case in the previous section where the measurement matrix was formed by randomly selecting m rows from the orthogonal DFT matrix $\mathbf{F}_d$. When m exceeds d, however, it becomes necessary to reinterpret the measurement process since we cannot simply add more rows to the DFT matrix $\mathbf{F}_d$ to create an appropriate oversampling operator $\mathbf{A}$.

As before, we assume that the signal class $\{u\colon \mathbb{R} \to \mathbb{R} \mid u \in B_{\mathcal{F}}([-f_{\mathrm{b}}, f_{\mathrm{b}}])\}$ consists of superpositions of band-limited sinusoids with $f_{\mathrm{b}} = f_{\mathrm{r}}/2$, where f_{b} corresponds to the Nyquist frequency. In contrast to the previous setting, we now aim to reconstruct the spectrum of such signals from oversampled representations. To that end, we consider $m > d$ time domain measurements of $u(t)$. Conceptually, this means that we sample each signal $u(t)$ via the sampling operator $\mathcal{A}_{m,f_{\mathrm{r}}'}$ with sampling rate $f_{\mathrm{r}}' := f_{\mathrm{r}}m/d$. Since $u(t)$

is band-limited to $|f| \leq f_r/2$, this implies that the discrete Fourier transform $\hat{\mathbf{x}} \in \mathbb{X}_m$ of $\mathbf{z} = \mathcal{A}_{m,f_r'} u(t) \in \mathbb{R}^m$ is sparse if every frequency in $u(t)$ is an integer multiple of f_r/d and contains no nonzero coefficients in the index set[7]

$$\begin{aligned} U &= \left\{ \frac{d}{2}+2, \frac{d}{2}+3, \ldots, \frac{m}{2}+1, \frac{m}{2}+2, \ldots, \frac{m}{2}+1+\frac{m-d}{2}-1 \right\} \\ &= \left\{ \frac{d}{2}+2, \frac{d}{2}+3, \ldots, \frac{m}{2}+1, \frac{m}{2}+2, \ldots, m-\frac{d}{2} \right\} \end{aligned} \tag{3.5}$$

corresponding to frequencies above the Nyquist frequency of $u(t)$. Since the CS-BIHT algorithm alternates between estimating the spectral representation of $u(t)$ and constructing an approximation of the corresponding sampled time domain vector, the measurement operator $\mathbf{A} \in \mathbb{C}^{m \times d}$ needs to act as an inverse DFT operator, which produces a time-interpolated discrete signal.

To construct $\mathbf{A}$, we adopt the notion of *exact interpolation* (see, *e.g.*, [Lyo04, Section 13.28.1] or [Fra89]). The idea of exact interpolation is to zero-pad a vector $\mathbf{x} \in \mathbb{X}_d$ such that the resulting vector $\hat{\mathbf{x}} \in \mathbb{X}_m$ remains conjugate symmetric and passes through every sampling point of $\mathbf{F}_d\mathbf{x}$ in the time domain. This includes splitting the Nyquist coefficient of $\mathbf{x}$ by re-weighting and assigning it to both the positive and negative frequency spectrum of the interpolated vector $\hat{\mathbf{x}} \in \mathbb{X}_m$. The construction is detailed in Algorithm 3. Note that unlike the procedure proposed in [Lyo04], we re-weight the Nyquist coefficient by a factor of $1/\sqrt{2}$ rather than $1/2$ so that $\|\hat{\mathbf{x}}\|_2 = \|\mathbf{x}\|_2$. Since the exact interpolation procedure induces a linear map on $\mathbb{X}_d$, there exists a linear operator $\mathbf{P}_{\text{int}} \in \mathbb{R}^{m \times d}$ which acts on any vector $\mathbf{x} \in \mathbb{X}_d$ as described in Algorithm 3. Equipped with this operator, we now define the matrix $\tilde{\mathbf{A}} = \mathbf{F}_m \mathbf{P}_{\text{int}}$ and finally the measurement matrix $\mathbf{A}$ by normalizing the columns of $\tilde{\mathbf{A}}$ to unit norm. We emphasize that the explicit construction of $\mathbf{A}$ is not necessary for the CS-BIHT algorithm and its variations since one may instead capitalize on highly optimized implementations of the *fast Fourier transform* (FFT) algorithm to implement the action of $\mathbf{A}$. This also applies to modern solvers for convex programs such as interior-point methods [Van12]. However, highly-tuned implementations of such algorithms require substantial work and experience while an efficient implementation of CS-BIHT involving $\mathbf{F}_m$ and $\mathbf{P}_{\text{int}}$ is fairly straightforward.

Algorithm 3 Exact Time Domain Interpolation

Input: $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{X}_d$, $m \in \mathbb{N}$ with $m > d$

Initialize: $\hat{\mathbf{x}} \leftarrow \mathbf{0} \in \mathbb{C}^m$

$\hat{x}_i \leftarrow x_i$ for $i \in \{1, \ldots, d/2\}$ ▷ Copy DC and positive frequency coefficients

$\hat{x}_{m-d/2+i} \leftarrow \overline{x_i} = x_{d/2+i}$ for $i \in \{2, \ldots, d/2\}$ ▷ Copy negative frequency coefficients

$\hat{x}_{d/2+1} \leftarrow 2^{-1/2} x_{d/2+1}$ ▷ Duplicate Nyquist coefficient

$\hat{x}_{m-d/2+1} \leftarrow 2^{-1/2} x_{d/2+1}$

Output: $\hat{\mathbf{x}} \in \mathbb{X}_m$

[7] This follows from the arrangement of DFT coefficients in $\hat{\mathbf{x}}$ implied by Definition 3.1.

3.4 Numerical Evaluation

We now turn to investigating the empirical recovery performance of the conjugate symmetric binary iterative hard thresholding algorithm and its variants.

3.4.1 Simulation Setup

In all our simulations, we consider a similar setup to the original work [Jac+13] which first introduced the BIHT algorithm. That is, we consider the recovery of conjugate symmetric s-sparse vectors in $\mathbb{C}^d$ with $s = 20$ and $d = 1000$. In order to generate the target signals, we choose an index set S' with $|S'| = s/2 = 10$ elements from the set $\{1, \dots, d/2\}$ at random. If S' includes the DC coefficient, the entry $d/2+1$ is added to S' as well to make sure that the constructed vector is exactly s-sparse. The nonzero elements supported on $S' \setminus \{1, d/2+1\}$ are then drawn independently from the circularly symmetric complex standard Gaussian distribution[8], while DC and Nyquist coefficients are drawn from the real standard Gaussian distribution. The positive frequency coefficients are then mirrored such that the resulting vector $\tilde{\mathbf{x}}$ is conjugate symmetric. We then set $\mathring{\mathbf{x}} = \tilde{\mathbf{x}}/\|\tilde{\mathbf{x}}\|_2$.

For the moment, we limit our attention to the regime $m \leq d$. In classical compressed sensing with linear observations, this is the only regime of interest as acquiring more measurements than the ambient dimension of the space yields overdetermined systems such that recovery reduces to a simple least-squares problem. Note, however, that depending on the application at hand, the oversampled regime $m > d$ still might be of interest in 1-bit CS, given the assumption of cheap low-complexity 1-bit samplers, which may well be able to operate at super-Nyquist rates due to the reduced demands on hardware complexity and data rates. Due to the differences involved in constructing the measurement operator $\mathbf{A}$ for $m > d$, we defer the numerical analysis of recovery from oversampled measurements to a separate discussion in Section 3.4.4. Throughout all our experiments, we choose $p = 1$ for the operator $\mathcal{H}^{\mathbb{X}}_{s,p}$ as initial experiments indicate that the choice is inconsequential to the overall performance of the CS-BIHT algorithm. For each parameter combination, we consider $n_{\mathrm{MC}} = 1000$ *Monte Carlo* (MC) instances for which we independently redraw any random quantities. As discussed in Section 3.2, the CS-BIHT algorithm is terminated once a quantization-consistent solution is found or the iteration count exceeds $n_{\max} = 3000$.

In addition to the regular CS-BIHT algorithm, we also investigate the performance of the CS-BIHT-ℓ_2 algorithm. Note that for this ℓ_2-variant of CS-BIHT, it is necessary to adjust the step-size λ_n according to the gradient $\nabla J_2(\mathbf{x}) = \mathbf{A}^* \mathrm{diag}\{\mathbf{y}\}[\mathbf{y} \circ \mathbf{A}\mathbf{x}]_-$. In particular, we choose $\lambda_n = \|\nabla J_2(\mathbf{x}^{(n)})\|_2^{-1}$, which seems to perform well in practice. Moreover, while we initialize CS-BIHT with $\mathbf{x}^{(0)} = \mathbf{0}$ (cf. Algorithm 2), this initial value causes CS-BIHT-ℓ_2 to stall in the first iteration since $\nabla J_2(\mathbf{0}) = \mathbf{0}$. We therefore initialize the smoothed version of the algorithm with a randomly drawn (and densely populated) conjugate symmetric vector. More precisely, the DC an Nyquist coefficients are drawn independently from $\mathsf{N}(0,1)$, while the positive frequency coefficients $x_2^{(0)}, \dots, x_{d/2}^{(0)}$ are drawn as independent circularly symmetric complex Gaussian random variables with unit variance. The conjugate complements of these coefficients are then mirrored to create a conjugate symmetric vector, followed by normalizing the vector to unit ℓ_2-norm.

[8] To construct a circularly symmetric standard Gaussian random variable u one simply sets $u = 2^{-1/2}(g + ih)$ with $g, h \sim_{\text{i.i.d.}} \mathsf{N}(0,1)$.

We also benchmark the performance of the two CS-BIHT algorithms against their respective *support-oracle* variants which always retain the correct coefficients of the (sub)gradient update according to the ground truth support set $S \subset [d]$ of the vector $\mathring{\mathbf{x}} \in \Sigma_s(\mathbb{X}_d)$ we aim to recover. In other words, the hard thresholding operator $\mathcal{H}^{\mathbb{X}}_{s,p}$ in both algorithms is replaced with the restriction operator $\mathbf{R}_S$ which we interpret here with slight abuse of notation as the identity matrix whose columns indexed by $\overline{S}$ are replaced by the zero vector, *i.e.*, $\mathbf{R}_S$ acts on a vector $\mathbf{x}$ such that $(\mathbf{R}_S\mathbf{x})_i = x_i \mathbb{1}_{\{i \in S\}}$ for $i \in [d]$. By removing the issue of support identification, the performance of these support-oracle-assisted versions gives some insight into how well the individual cost functions manage to differentiate between more dominating and less pronounced coefficients in the target vectors.

For comparison, we also consider two convex programs proposed by Plan and Vershynin which we modify for the recovery of conjugate symmetric signals. The first reconstruction scheme is based on the convex program proposed in [**PV13a**] for which uniform recovery of effectively s-sparse vectors was shown to hold with high probability on the draw of a standard Gaussian random matrix $\mathbf{A}$. We extend this program to the situation of conjugate symmetric signal recovery with $\mathbf{A} = \mathbf{R}_\Omega \mathbf{F}_d$ by solving the program[9]

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \mathbf{y} \circ \mathbf{A}\mathbf{x} \geq 0 \\ & \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle = 1 \\ & \mathbf{x} \in \mathbb{X}_d. \end{aligned} \tag{PV$_1$}$$

As mentioned before, the second constraint in the formulation removes the null space of $\mathbf{A}$ from the feasible set so that most importantly Problem (PV$_1$) does not admit a trivial solution at $\mathbf{x}^\star = \mathbf{0}$. The last constraint enforces the conjugate symmetry of solutions. Due to the subspace nature of $\mathbb{X}_d$, the constraint is linear and hence easily enforced (cf. Definition 3.1). Secondly, we consider the maximization problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle \\ \text{s.t.} \quad & \|\mathbf{x}\|_2 \leq 1 \\ & \|\mathbf{x}\|_1 \leq \sqrt{s} \\ & \mathbf{x} \in \mathbb{X}_d \end{aligned} \tag{PV$_2$}$$

as originally proposed in [**PV13b**] for the recovery of effectively sparse vectors in $\mathbb{R}^d$ from Gaussian observations. This time the objective function aims to maximize the quantization consistency of solutions $\mathbf{x}^\star$, while the feasible set encodes a membership constraint for the set of *effectively s-sparse vectors*[10] inside the unit ball, which is motivated by the relation $\|\mathbf{x}\|_1 \leq \sqrt{\|\mathbf{x}\|_0}\|\mathbf{x}\|_2$ due to the Cauchy-Schwarz inequality. We will delay a more detailed discussion of the program until in Section 4.3.2 where we consider a natural

[9] The first constraint of Problem (PV$_1$) is technically not well-defined for complex $\mathbf{A}$ and $\mathbf{x}$ since $\mathbb{C}$ is not totally ordered. In practice, we therefore explicitly enforce $\mathbf{y} \circ \mathfrak{Re}(\mathbf{A}\mathbf{x}) \geq \mathbf{0}$ and $\mathbf{y} \circ \mathfrak{Im}(\mathbf{A}\mathbf{x}) = \mathbf{0}$ instead. The latter constraint is also implicit in the constraint $\langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle = 1$.

[10] Technically, the set of effectively s-sparse vectors is given by the set $\mathcal{E}_s := \{\mathbf{x} \in \mathbb{C}^d : \|\mathbf{x}\|_1 \leq \sqrt{s}\|\mathbf{x}\|_2\}$ such that we actually need to constrain the search space of Problem (PV$_2$) to the set $\mathcal{E}_s \cap \mathbb{B}_2^d = \{\mathbf{x} : \|\mathbf{x}\|_1 \leq \sqrt{s}\|\mathbf{x}\|_2, \|\mathbf{x}\|_2 \leq 1\}$. However, as we will discuss in Section 4.3.2, the set $\{\mathbf{x} : \|\mathbf{x}\|_1 \leq \sqrt{s}, \|\mathbf{x}\|_2 \leq 1\}$ in Problem (PV$_2$) corresponds to the convex hull of the set $\mathcal{E}_s \cap \mathbb{B}_2^d$. Since we are maximizing a linear function over a compact set, the optimal values of the respective programs therefore coincide (cf. Proposition A.11).

extension of Problem (PV_2) in the context of *group-sparse* recovery. Instead, we merely point out that the program can be shown to exhibit a remarkable error resilience when the entries of the measurement matrix $\mathbf{A}$ are again drawn independently from the standard Gaussian distribution. As in the case of the CS-BIHT variants, we also consider support-oracle-assisted variants of both programs where we add the linear constraint $\mathbf{x}_{\overline{S}} = \mathbf{0}$ to Problem (PV_1) and (PV_2).

The first criterion we employ to evaluate the performance of the various recovery strategies is the average reconstruction SNR in dB denoted by $\rho\,[\mathrm{dB}] := 20\log_{10}(\overline{\rho})$ with

$$\overline{\rho} := \frac{1}{n_{\mathrm{MC}}}\sum_{i=1}^{n_{\mathrm{MC}}}\frac{\|\mathring{\mathbf{x}}_i\|_2}{\|\mathring{\mathbf{x}}_i - \mathbf{x}_i^\star\|_2} = \frac{1}{n_{\mathrm{MC}}}\sum_{i=1}^{n_{\mathrm{MC}}}\|\mathring{\mathbf{x}}_i - \mathbf{x}_i^\star\|_2^{-1}$$

where $\mathring{\mathbf{x}}_i \in \mathbb{X}_d \cap \mathbb{S}^{d-1}$ and $\mathbf{x}_i^\star$ denote the i-th signal (out of n_{MC} draws) and its reconstructed version by a particular recovery method, respectively. As emphasized earlier, recovery of sparse vectors from their 1-bit observations is only possible up to a positive scale factor due to the scale invariance of the sgn-operator. We therefore limit our attention to the recovery of signals with unit ℓ_2-norm. As a result, the normalized geodesic distance $\Delta_{\mathbb{S}}\colon \mathbb{S}^{d-1}\times\mathbb{S}^{d-1}\to[0,1]$ on the complex unit sphere $\mathbb{S}^{d-1} = \{\mathbf{x}\in\mathbb{C}^d : \|\mathbf{x}\|_2 = 1\}$ with

$$\Delta_{\mathbb{S}}(\mathbf{x},\mathbf{z}) = \frac{1}{\pi}\arccos\left(\mathfrak{Re}\langle\mathbf{x},\mathbf{z}\rangle_{\mathbb{C}}\right)$$

could serve as a natural measure of similarity between two vectors $\mathbf{x},\mathbf{z}\in\mathbb{S}^{d-1}$. However, since

$$\|\mathbf{x}-\mathbf{z}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{z}\|_2^2 - 2\mathfrak{Re}\langle\mathbf{x},\mathbf{z}\rangle_{\mathbb{C}} = 2(1-\mathfrak{Re}\langle\mathbf{x},\mathbf{z}\rangle_{\mathbb{C}}),$$

we have

$$\|\mathbf{x}-\mathbf{z}\|_2^2 = 2(1-\cos(\pi\Delta_{\mathbb{S}}(\mathbf{x},\mathbf{z}))) = 4\sin\left(\frac{\pi}{2}\Delta_{\mathbb{S}}(\mathbf{x},\mathbf{z})\right)^2$$

and therefore

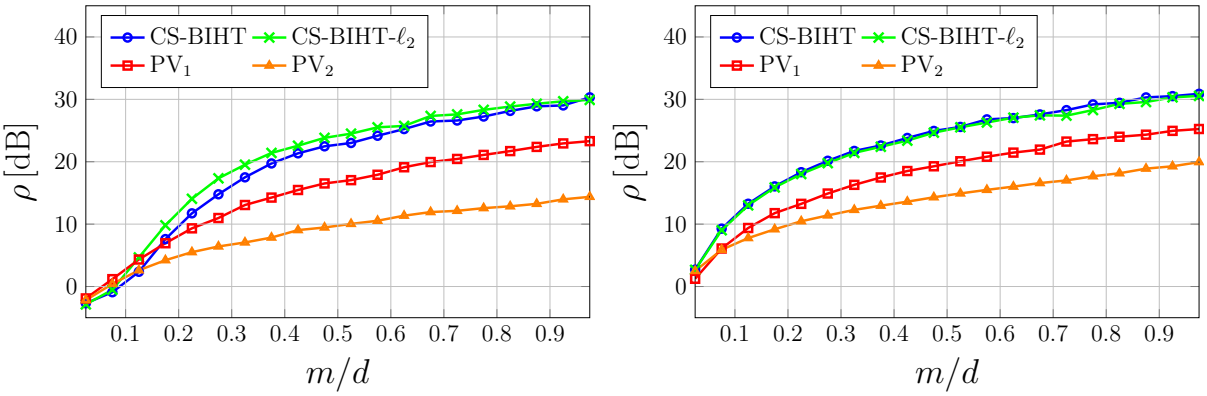
$$\Delta_{\mathbb{S}}(\mathbf{x},\mathbf{z}) = \frac{2}{\pi}\arcsin\left(\frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_2\right)$$

given that $\sin(x)$ is injective (and nonnegative) on $[\pi/2,\pi]$. Geodesic distance and SNR as defined above are therefore directly related by a monotonically increasing function such that the metric $\Delta_{\mathbb{S}}$ does not reveal any new information about the reconstruction quality. Instead of the geodesic distance, we will evaluate the support recovery performance by considering the cardinality of the symmetric set difference between the true support S of the ground truth signal $\mathring{\mathbf{x}}$ and the support $S^\star$ of its estimate $\mathbf{x}^\star$:

$$\Delta_{\mathrm{supp}}(S,S^\star) := |S\Delta S^\star| = |(S\setminus S^\star)\cup(S^\star\setminus S)|.$$

Since we assume perfect knowledge of the sparsity level s, we have $\Delta_{\mathrm{supp}}(S,S^\star)\le 2s$ in the case of CS-BIHT. Unfortunately, Problem (PV_1) and (PV_2) generally do not produce genuinely sparse vectors such that we have to employ a thresholding scheme to estimate the support set $S^\star$ which contains the majority of a signal's energy. Given a vector $\mathbf{x}^\star$ produced by some recovery algorithm, we consider the sequence $(\mathcal{H}^{\mathbb{X}}_{t,p}(\mathbf{x}^\star))_{t=1}^d$ of progressively more accurate conjugate symmetric t-sparse approximations of $\mathbf{x}^\star$ and set $S^\star = \mathrm{supp}(\mathcal{H}^{\mathbb{X}}_{t^\star,p}(\mathbf{x}^\star))$ where $t^\star$ denotes the smallest sparsity level t such that

$$\frac{\left\|\mathbf{x}^\star - \mathcal{H}^{\mathbb{X}}_{t,p}(\mathbf{x}^\star)\right\|_2}{\|\mathbf{x}^\star\|_2} \le 10^{-3}.$$



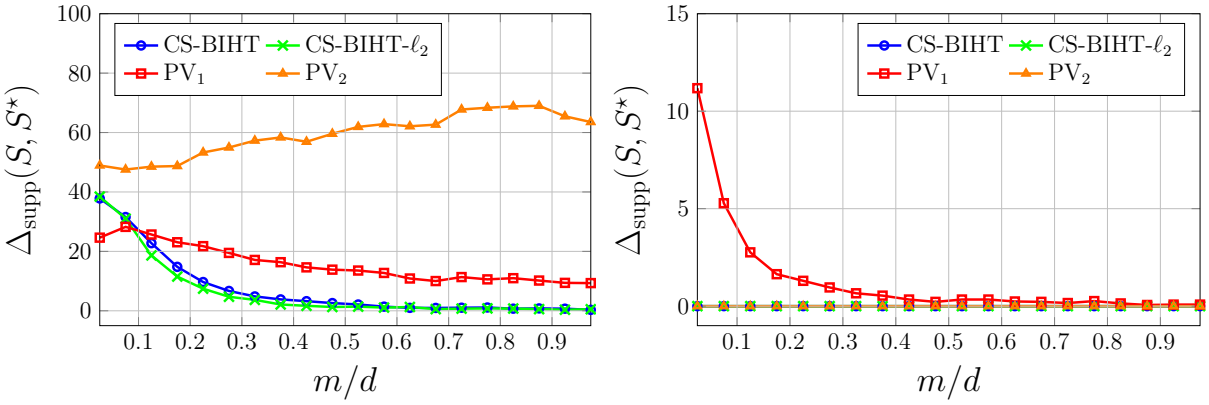
(a) Support-unaware variants **(b)** Support-oracle variants

Figure 3.1: Average reconstruction SNR vs. number of measurements for $d = 1000$ and $s = 20$

3.4.2 Noiseless Recovery

In the first experiment, we investigate the recovery of conjugate symmetric sparse vectors from noiseless 1-bit observations. The average SNR over the number of measurements normalized to the signal dimension is shown Figure 3.1a. As in the real-valued case, which we do not present here, the conjugate symmetric version of the BIHT algorithm generally outperforms Problem (PV$_1$) by 5–7 dB on average at even moderate numbers of measurements. Additionally, the smooth variant CS-BIHT-ℓ_2 slightly outperforms CS-BIHT at lower values of m with the gap closing more and more as m increases. This is in stark contrast to what was previously reported in [**Jac+13**] in the Gaussian setting where the ℓ_2-smoothed variant generally falls significantly behind the nonsmooth BIHT algorithm. The performance of Problem (PV$_2$) is generally another 5 to 8 dB lower than that of Problem (PV$_1$), resulting in a gap of as much as 15 dB compared to CS-BIHT-ℓ_2 when the number of measurements approaches d. For the support-oracle variants of CS-BIHT and CS-BIHT-ℓ_2 (Figure 3.1b), the performance gap almost vanishes entirely, which suggests that CS-BIHT-ℓ_2 is slightly more effective at identifying the support of the target vector. With m approaching d, however, the performance of the regular and support-oracle versions are almost identical, emphasizing the effectiveness of both CS-BIHT algorithms in general. While the gap between the individual methods slightly closes when the support is known a prior, the relative relation of the recovery performance between CS-BIHT(-ℓ_2), Problem (PV$_1$) and (PV$_2$) remains the same. The biggest jump in performance is observed for Problem (PV$_2$) whose average SNR improves by around 6 dB for $m = d$.

Next, we turn to the support identification problem, the results of which are depicted in Figure 3.2. Once again, CS-BIHT and CS-BIHT-ℓ_2 outperform both convex programs and manage to accurately identify the correct support at modest numbers of measurements with the misidentification rate dropping to 0 beyond $m \geq 800$ measurements. Even though the constraint $\mathbf{y} \circ \mathbf{Ax} \geq 0$ in Problem (PV$_1$) promotes strictly quantization-consistent solutions, the recovery scheme's overall ability to identify the true signal support is rather poor and only improves marginally as the number of measurements increases. Most importantly, in contrast to both CS-BIHT algorithms, the misidentification rate does not drop to 0 in the regime considered for m. Unsurprisingly, Problem (PV$_2$) falls even further behind the three other methods. Even more striking is the fact that the number of



(a) Support-unaware variants

(b) Support-oracle variants

Figure 3.2: Support recovery error vs. number of measurements for $d = 1000$ and $s = 20$

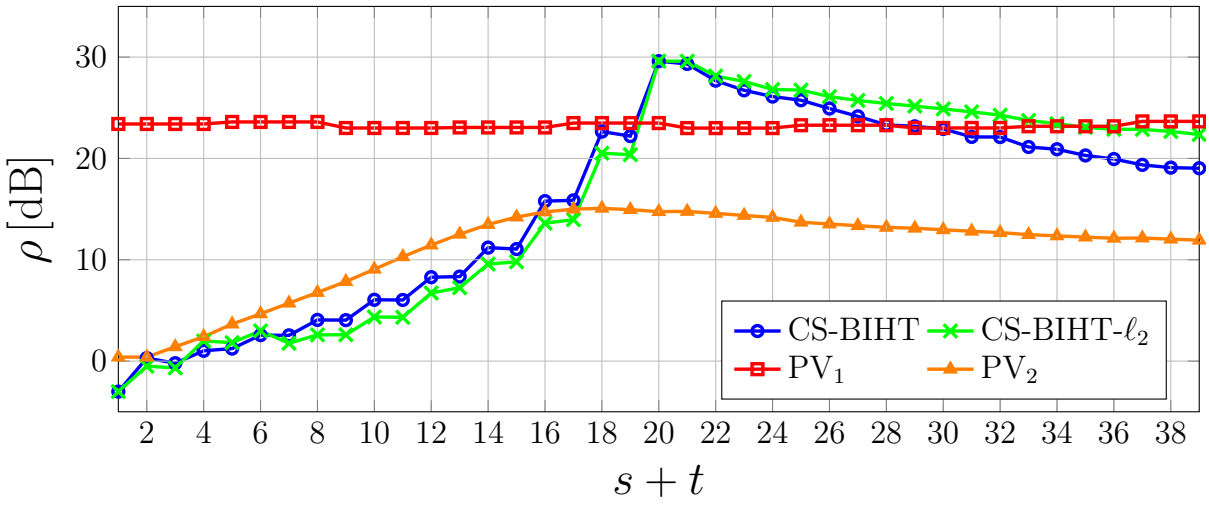


Figure 3.3: Average reconstruction SNR vs. sparsity offset when recovery methods are provided with incorrect prior information about the sparsity level of target vectors ($d = 1000, s = 20$)

erroneously selected support indices seems to increase as more measurements are acquired. Finally, while one would expect the support error to be fixed at 0 for the support-oracle-assisted versions of the considered recovery procedures, this is surprisingly not the case for Problem (PV$_1$) as shown in Figure 3.2b. Since solutions of Problem (PV$_1$) are forced to be identically zero on $\overline{S}$ for $S = \mathrm{supp}(\mathring{\mathbf{x}})$, this implies that the program actually concentrates the signal energy on fewer than the remaining $s = 20$ possible nonzero coefficients.

With the exception of the ℓ_1-minimization approach (PV$_1$), the considered recovery schemes require prior information about the sparsity level of the target vector. In order to examine how well these methods cope with inaccurate sparsity information, we now provide each method with a modified sparsity level $s + t$ with $s = 20$ as before and t ranging from -19 to 19. The results of this experiment are shown in Figure 3.3. Given the independence of Problem (PV$_1$) from $s + t$, the recovery performance remains fixed for all values of t as pointed out above. On the other hand, the reconstruction fidelity of both versions of the CS-BIHT algorithm deteriorates substantially when no accurate information about the sparsity level is available. In this case, the algorithms are even

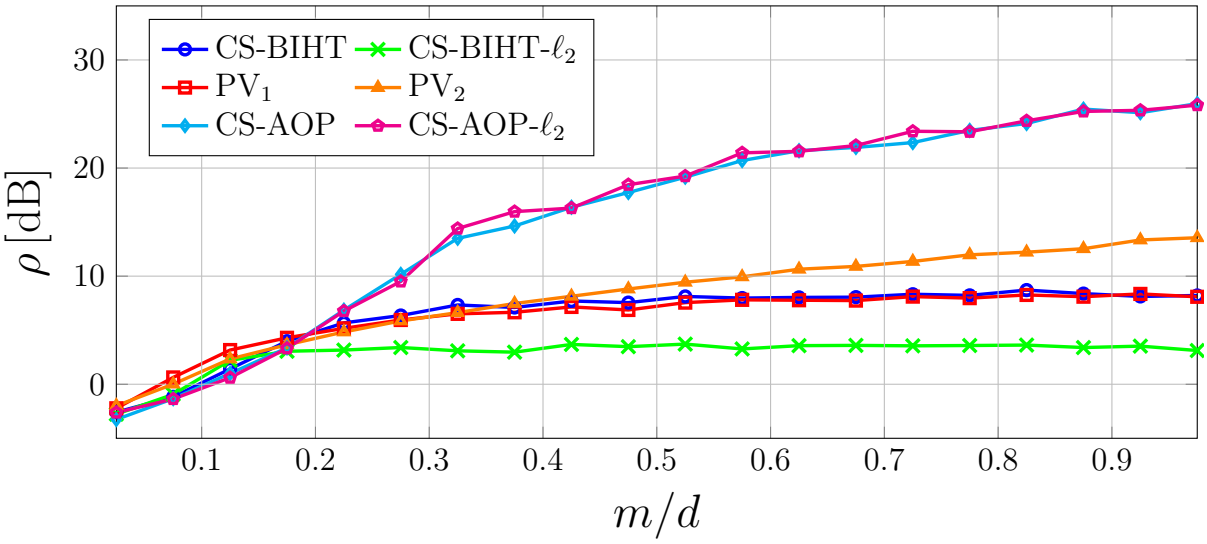


Figure 3.4: Average reconstruction SNR vs. number of measurements for $d = 1000$ and $s = 20$ when 3 % of all sign measurements are flipped

outperformed by Problem (PV_2) despite the fact that the convex maximization method relies on the same inaccurate sparsity information. However, since the search space $\{\mathbf{x} \in \mathbb{X}_d : \|\mathbf{x}\|_1 \leq \sqrt{s+t}, \|\mathbf{x}\|_2 \leq 1\}$ of Problem (PV_2) is significantly larger than that of CS-BIHT and CS-BIHT-ℓ_2, namely $\Sigma_{s+t}(\mathbb{X}_d)$, this is to be expected. Moreover, while the loss in reconstruction SNR for CS-BIHT and CS-BIHT-ℓ_2 is significant when the sparsity level is underestimated, the effect is less detrimental once t moves past the point where the sparsity level is overestimated. Since the two algorithms outperform Problem (PV_1) even when the sparsity level is overestimated by as much as 50 %, this gives the hard thresholding approaches some leeway in case a lower estimate of the sparsity level is available.

3.4.3 Noisy Recovery

Next, we investigate how the different recovery strategies fare with noisy observations. The assumption of noiseless sampling considered in the previous section represents a highly idealized scenario. While the previous experiments provide some indication of what kind of reconstruction fidelity to expect, it is also necessary to investigate how well the respective recovery schemes are able to cope with noisy observations. Unfortunately, the BIHT algorithm turns out to be highly sensitive to noise in that only a few bit flips can already lead to significantly slower (empirical) convergence behavior and reconstruction quality. To combat such effects, Yan, Yang and Osher proposed the so-called *adaptive outlier pursuit* (AOP) algorithm—a variant of BIHT which adaptively tries to identify and subsequently correct potentially erroneous bit positions in a noisy measurement vector $\tilde{\mathbf{y}} \in \{\pm 1\}^m$ [**YYO12**]. Given an upper estimate $\beta \in [m]$ on the number of corrupted entries in the vector $\tilde{\mathbf{y}}$, the AOP algorithm identifies possible bit error positions by solving a combinatorial optimization problem.

To make the method precise, consider the noisy 1-bit measurement model $\tilde{\mathbf{y}} = \mathbf{f} \circ \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}})$ for some $\mathring{\mathbf{x}} \in \mathbb{R}^d$. The binary vector $\mathbf{f} \in \{\pm 1\}^m$ with $m\Delta_{\mathrm{H}}(\mathbf{1}, \mathbf{f}) = \beta \in [m]$ is used to model adversarial post-quantization bit flips in the quantization device. The idea

of the AOP method is to estimate and update a vector $\mathbf{b} \in \{\pm 1\}^m$ to iteratively correct possible bit errors during reconstruction. This is done in an alternating fashion which alternates between moving closer to a quantization-consistent solution via a regular BIHT iteration and—fixing the current estimate $\mathbf{x}^{(n)}$—determining a corrected measurement vector $\tilde{\mathbf{y}} \circ \mathbf{b}$ which better matches the measurements $\operatorname{sgn}(\mathbf{A}\mathbf{x}^{(n)})$ of the iterate $\mathbf{x}^{(n)}$. To that end, one considers for the current estimate $\mathbf{x}^{(n)}$ in iteration n of the BIHT algorithm the combinatorial optimization problem

$$\begin{aligned} \underset{\mathbf{b}}{\text{minimize}} \quad & \mathcal{L}^{(n)}(\tilde{\mathbf{y}} \circ \mathbf{b}) \\ \text{s.t.} \quad & \|[\mathbf{b}]_-\|_1 \leq \beta \\ & \mathbf{b} \in \{\pm 1\}^m \end{aligned} \tag{P$_{3.2}$}$$

where $\mathcal{L}^{(n)}$ denotes either the nonsmooth loss function $\mathcal{L}_1^{(n)}(\mathbf{y}) = \|[\mathbf{y} \circ \mathbf{A}\mathbf{x}^{(n)}]_-\|_1$ or the smooth loss function $\mathcal{L}_2^{(n)}(\mathbf{y}) = \|[\mathbf{y} \circ \mathbf{A}\mathbf{x}^{(n)}]_-\|_2^2$ of the BIHT algorithm. This combinatorial optimization problem admits a closed-form solution $\mathbf{b}^\star$ established in [**YYO12**] as

$$b_i^\star = \begin{cases} -1, & \tilde{y}\left\langle \mathbf{a}_i, \mathbf{x}^{(n)} \right\rangle \geq \tau, \\ 1, & \text{otherwise} \end{cases} \tag{3.6}$$

with τ denoting the β-th largest entry of $|[\tilde{\mathbf{y}} \circ \mathbf{A}\mathbf{x}^{(n)}]_-|$ where $|\cdot|$ is applied element-wise. The vector $\mathbf{b}^\star$ is then used in the next iteration of the BIHT algorithm to flip the (hopefully correctly) identified bit error positions in $\tilde{\mathbf{y}}$.[11] In other words, the update $\mathbf{x}^{(n+1)}$ is constructed as $\mathbf{x}^{(n+1)} = \mathcal{H}_s(\mathbf{x}^{(n)} - \lambda_n p_{\mathbf{b}^\star}(\mathbf{x}^{(n)}))$ with $p_{\mathbf{b}^\star}(\mathbf{x}) = \mathbf{A}^\top(\operatorname{sgn}(\mathbf{A}\mathbf{x}) - \tilde{\mathbf{y}} \circ \mathbf{b}^\star)$ or $p_{\mathbf{b}^\star}(\mathbf{x}) = \mathbf{A}^\top \operatorname{diag}\{\tilde{\mathbf{y}} \circ \mathbf{b}^\star\}[\tilde{\mathbf{y}} \circ \mathbf{b}^\star \circ \mathbf{A}\mathbf{x}]_-$ in the nonsmooth and smooth case, respectively. The bit error vector $\mathbf{b}$ is updated throughout the algorithm if the normalized Hamming distance Δ_{H} between the noisy observations $\tilde{\mathbf{y}}$ and the quantized measurements $\operatorname{sgn}(\mathbf{A}\mathbf{x}^{(n)})$ of the current estimate $\mathbf{x}^{(n)}$ reduces. Since the algorithm assumes that β bits are flipped, the procedure is terminated once said Hamming distance is less β.

We adopt this method for the recovery of sparse conjugate symmetric vectors and repeat the previous experiment for the AOP versions of CS-BIHT and CS-BIHT-ℓ_2 when 3 % of all measurements are flipped. The full algorithm, which we dub *conjugate symmetric adaptive outlier pursuit* (CS-AOP), is detailed in Algorithm 4. For simplicity, we provide both the CS-AOP algorithm and its smooth variant CS-AOP-ℓ_2 with the exact number of bit flips. Note that while we had to normalize the gradient update of CS-BIHT-ℓ_2 by choosing the step-size as $\lambda_n = \|\nabla J_2(\mathbf{x}^{(n)})\|_2^{-1}$, this choice leads to highly suboptimal performance in case of the CS-AOP-ℓ_2 algorithm. Instead, we simply choose $\lambda_n = 1$ as for CS-BIHT and CS-AOP. Again, we benchmark the performance against the convex programs (PV$_1$) and (PV$_2$).

The results of this experiment are shown in Figure 3.4. As indicated above, the performance of both CS-BIHT variants without bit error corrections deteriorates significantly with CS-BIHT-ℓ_2 merely achieving a reconstruction SNR of 3 dB when acquiring measurements on the order of the ambient dimension. Despite the fact that the smooth and nonsmooth version of the CS-BIHT algorithm performed equally well at moderately

[11] In the original formulation of the AOP algorithm, the vector $\mathbf{b}^\star$ was used to estimate the uncorrupted bit positions and ignore the likely corrupted ones when constructing the next iterate $\mathbf{x}^{(n+1)}$. This reduces the effective number of measurements used during reconstruction. The variant we discuss here in which potentially erroneous bits are flipped was instead termed *AOP with flips* in [**YYO12**].

Algorithm 4 Conjugate Symmetric Adaptive Outlier Pursuit (CS-AOP)

Input: $\mathbf{A} = \mathbf{R}_\Omega \mathbf{F}_d \in \mathbb{C}^{m\times d}$, $\tilde{\mathbf{y}} = \mathbf{f} \circ \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}) \in \{\pm 1\}^m$, $s \in [d]$, $\beta \in [m]$
Initialize: $\mathbf{x}^{(0)} \leftarrow \mathbf{0}$, $\mathbf{b} \leftarrow \mathbf{1}$, $\gamma_{\text{tol}} \leftarrow 1$, $n \leftarrow 0$
do
  $\mathbf{u}^{(n+1)} \leftarrow \mathbf{x}^{(n)} - \mathbf{A}^*(\operatorname{sgn}(\mathbf{A}\mathbf{x}^{(n)}) - \tilde{\mathbf{y}} \circ \mathbf{b})$ ▷ Subgradient step
  $\mathbf{x}^{(n+1)} \leftarrow \mathcal{H}^{\mathbb{X}}_{s,p}(\mathbf{u}^{(n+1)})$ ▷ Projection on $\Sigma_s(\mathbb{X}_d)$
  if $\Delta_{\text{H}}(\tilde{\mathbf{y}}, \operatorname{sgn}(\mathbf{A}\mathbf{x}^{(n+1)})) \leq \gamma_{\text{tol}}$ **then** ▷ Update $\mathbf{b}$ if Δ_{H} improved
    $\gamma_{\text{tol}} \leftarrow \Delta_{\text{H}}(\tilde{\mathbf{y}}, \operatorname{sgn}(\mathbf{A}\mathbf{x}^{(n+1)}))$
    $\mathbf{b} \leftarrow \mathbf{b}^\star$ according to (3.6)
  end if
  $n \leftarrow n + 1$
while $\Delta_{\text{H}}(\tilde{\mathbf{y}} \circ \mathbf{b}, \operatorname{sgn}(\mathbf{A}\mathbf{x}^{(n)})) > 0$ **and** $\gamma_{\text{tol}} \geq \beta/m$ **and** $n < n_{\max}$
Output: $\mathbf{x}^{(n)}/\|\mathbf{x}^{(n)}\|_2$ ▷ Projection on the unit sphere

large numbers of measurements, there now exists a performance gap indicating a slightly higher noise robustness for the CS-BIHT algorithm compared to its smooth counterpart. Given the fact that both algorithms do not account for potential bit errors and blindly pursue solutions which are quantization-consistent with the noisy observations $\tilde{\mathbf{y}}$, such a significant drop in performance is to be expected. The same argument applies to the first convex programming approach, which results in the performance of CS-BIHT and Problem (PV_1) to be on par in the noisy regime. On the other hand, the noise robustness of Problem (PV_2), which was theoretically established in the Gaussian setting in [**PV13b**], seems to extend to measurement matrices constructed from subsampled Fourier systems as considered in this chapter. This is due to the fact that while the objective function of Problem (PV_2) promotes solutions whose linear measurements are correlated with the (noisy) observations $\tilde{\mathbf{y}}$, it does not strictly enforce quantization consistency. Finally, the conjugate symmetric versions of the AOP algorithm consistently manage to outperform any competing method for $m \geq 200$. While they do not quite catch up to the performance of CS-BIHT and CS-BIHT-ℓ_2 in the noiseless setting, they still attain around 26 dB reconstruction SNR for $m = d$.

The effectiveness of the conjugate symmetric outlier pursuit algorithms is also reflected in their empirical support recovery performance depicted in Figure 3.5 whose support detection error tends to 0 for $m \to d$. The next best methods for noisy support identification turn out to be the nonsmooth and smooth CS-BIHT algorithms whose performance is now on par with that of Problem (PV_2) in the noiseless case. Note, however, that with an average of 8 and 16 support elements misidentified for $m = d$, respectively, this constitutes a significant error at a sparsity level of $s = 20$, which renders the two CS-BIHT algorithms unsuited for support detection at nonnegligible noise levels. The situation is even worse for Problem (PV_1) and (PV_2) whose support error increases with m with the amount of misidentified support indices increasing at an even higher rate for Problem (PV_1) than for Problem (PV_2). We emphasize though that the behavior of Problem (PV_2) is in line with its support recovery performance in the noiseless setting. A natural explanation for this phenomenon is the fact that the convex maximization problem does not enforce strictly sparse but merely effectively sparse solutions, leading to a significant amount of signal energy being spread across $[d]$. Since every sparse vector is also effectively sparse by the

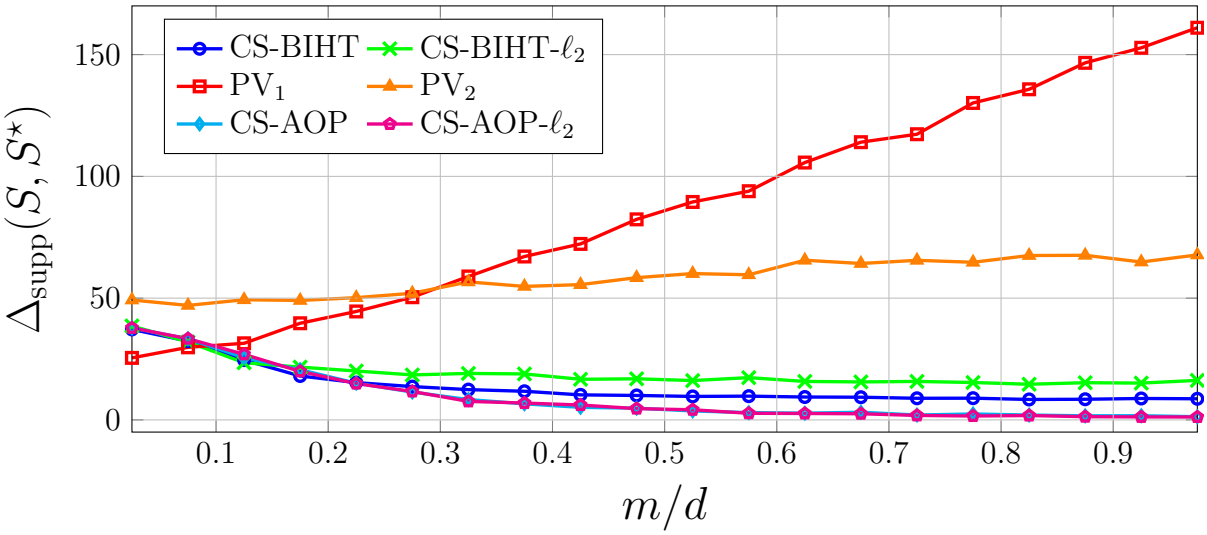


Figure 3.5: Average support identification error vs. number of measurements for $d = 1000$ and $s = 20$ when 3 % of all sign measurements are flipped

Cauchy-Schwarz inequality, the set of effectively sparse vectors is significantly larger than the set of genuinely sparse vectors such that Problem (PV$_2$) seems to almost always select an element outside of $\Sigma_s(\mathbb{X}_d)$.

3.4.4 Recovery from Oversampled Measurements

Before concluding this chapter, we consider the reconstruction of sparse conjugate symmetric vectors from oversampled measurements. Based on the signal model described in Section 3.3, we generate conjugate symmetric vectors $\mathring{\mathbf{x}} \in \mathbb{X}_d$ with $\|\mathring{\mathbf{x}}\|_0 \leq s$ as in the previous experiments. Instead of randomly selecting m rows from the orthogonal DFT matrix, however, we now construct the measurement matrix $\mathbf{A}$ according to the exact interpolation procedure outlined in Section 3.3.3. To implement the two CS-BIHT variants, we use a fast implementation based on the FFT algorithm, while for the convex approaches we provide $\mathbf{A}$ in full, which considerably increases computational complexity, especially at higher oversampling rates.

The results of this experiment are shown in Figure 3.6 and Figure 3.7 where we compare the average reconstruction SNR and support identification error as a function of m, respectively. The experiments confirm the claim that acquiring samples beyond the Nyquist rate generally improves reconstruction fidelity for each reconstruction scheme. While qualitatively, these improvements are universal, the quantitative improvement for CS-BIHT is considerably higher than for competing methods. This is especially surprising considering the fact that CS-BIHT and CS-BIHT-ℓ_2 appears to converge to the same reconstruction behavior as m approaches d in the undersampled regime. Instead, at an oversampling factor of $m/d = 4$, the CS-BIHT algorithm now outperforms its ℓ_2-variant by around 7 dB, the same as Problem (PV$_1$), which in turn maintains a consistent performance gap of around 7 dB to the CS-BIHT algorithm across the entire range $m/d \in [1, 4]$. Trailing furthest behind remains Problem (PV$_2$) whose reconstruction fidelity improves by a mere 2 dB from $m = d$ to $m = 4d$.

The support recovery performance of the individual recovery schemes emphasizes these

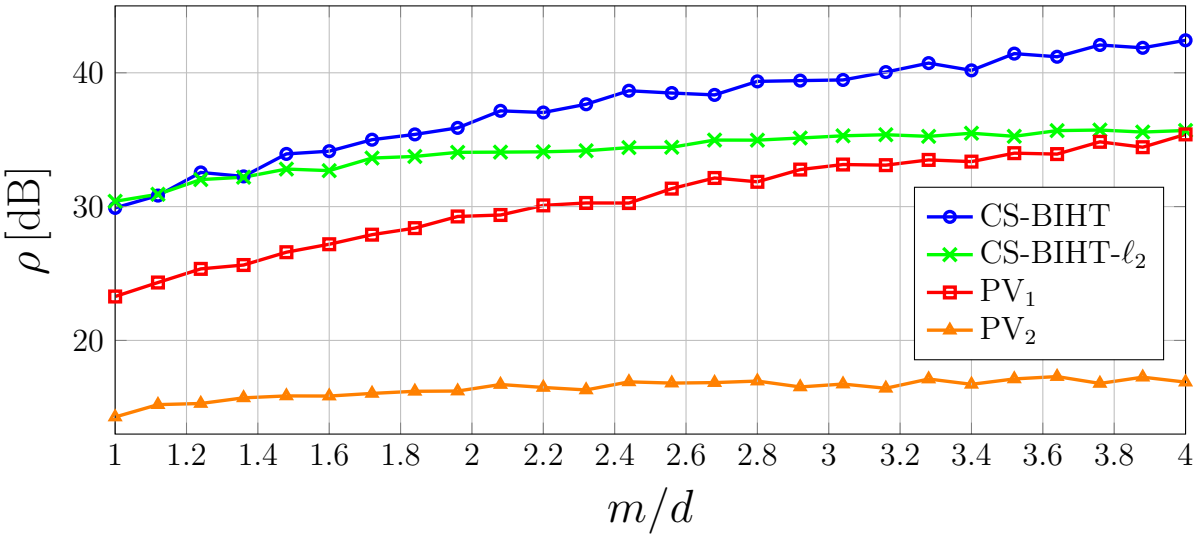


Figure 3.6: Average reconstruction SNR vs. number of measurements for $d = 1000$ and $s = 20$

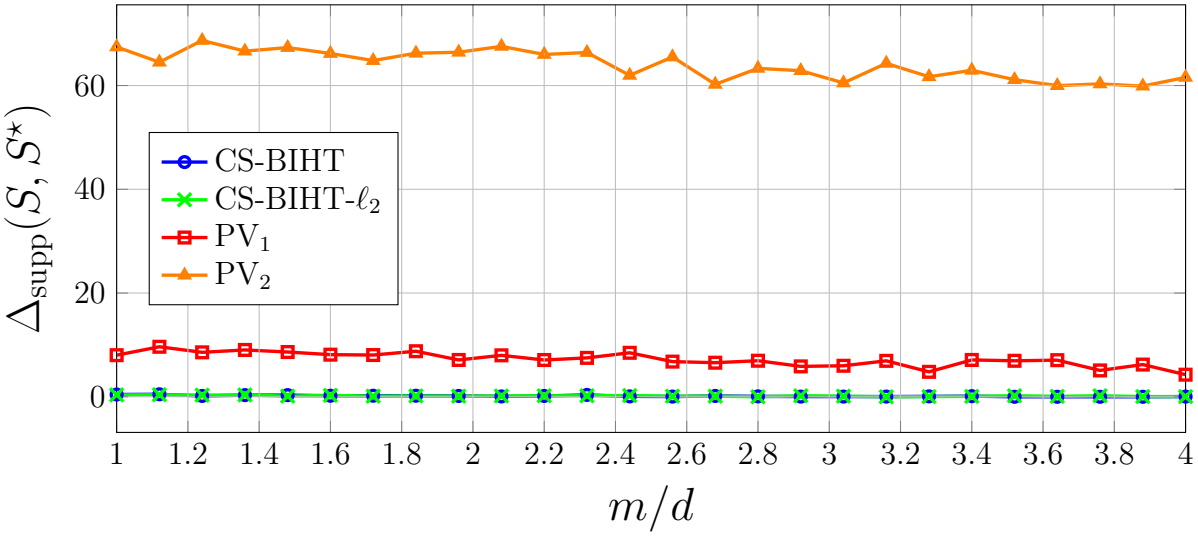


Figure 3.7: Average support identification error vs. number of measurements for $d = 1000$ and $s = 20$

observations with Problem (PV_2) erroneously identifying more than 60 support indices across the considered range of measurements. Despite their considerable performance gap in terms of reconstruction SNR, the support recovery behavior of CS-BIHT and CS-BIHT-ℓ_2 are virtually identical, almost perfectly identifying all $s = 20$ active support elements. Trailing slightly behind with $\Delta_{\mathrm{supp}}(S, S^\star)$ averaging around 4 is Problem (PV_1). This reinforces the notion that CS-BIHT—if provided with the correct sparsity level s—is both highly effective at identifying and estimating (up to global scaling) the individual elements of sparse conjugate symmetric vectors. Given the fact that the oversampled acquisition model does not involve any random sampling operations, the proposed approach is particularly hardware-friendly and admits fast and efficient reconstruction algorithms which capitalize on fast realizations of the measurement operator $\mathbf{A}$ in terms of the FFT algorithm.

3.5 Conclusion

In this chapter, we proposed a modification of the *binary iterative hard thresholding* (BIHT) algorithm for the recovery of conjugate symmetric frequency-sparse vectors from 1-bit quantized time domain measurements. Such vectors arise from sampling superpositions of sinusoidal signals, whose frequencies correspond to integer multiples of the frequency resolution implied by the sampling rate. The underlying acquisition model comes with three significant advantages over purely randomized designs. First, while linear mixing of analog signals with random sequences based on, *e.g.*, Gaussian ensembles is difficult to realize in hardware, the measurement model considered here is based on sub- or oversampled representations of the analog time domain signals. This renders the acquisition model particularly hardware-friendly. Secondly, each individual measurement is represented by a single information bit corresponding to the sign of the corresponding linear time domain sample. This allows for the analog-to-digital converters of the acquisition system to be implemented in the form of energy-efficient comparators w. r. t. a fixed voltage threshold. This in turn enables sensing devices to oversample signals without significantly increasing the complexity and associated cost of the necessary hardware circuitry. Thirdly, since the measurement matrix $\mathbf{A}$ modeling the acquisition system is based on the DFT matrix, one may exploit highly efficient implementations of the Cooley-Tukey FFT algorithm and its extensions during reconstruction.

Our proposed extension of the BIHT algorithm to the conjugate symmetric setting involves two parts. First, the subgradient update of the objective function, which penalizes inconsistent sign measurements, was reformulated to account for the complex nature of the underlying signal space. This subgradient step is intended to move the current iterate into a direction with improved quantization consistency. Secondly, we proposed a necessary modification of the so-called hard thresholding operator to project the subgradient update on the set $\Sigma_s(\mathbb{X}_d)$ of s-sparse conjugate symmetric vectors to exploit the additional structure of target vectors. This modification also guarantees that the subgradient update in the next iteration is well-defined, ensuring that any iterate produced by the algorithm has a real-valued inverse Fourier transform. In fact, the conjugate symmetric hard thresholding operator presented in this chapter enables any iterative algorithm that enforces sparsity by means of hard thresholding to be used during recovery of signals with a sparse conjugate symmetric discrete Fourier transform. This includes algorithms such as the QIHT algorithm

(a generalization of BIHT to multi-bit quantization schemes), as well as the IHT and HTP algorithms.

In order to extend the acquisition model to oversampled measurements, we adopted the notion of *exact interpolation*. This represents a simple zero-padding construction, which can be used to generate a sampling matrix $\mathbf{A} \in \mathbb{C}^{m \times d}$ with $m > d$ modeling the effects of oversampling in the time domain. The resulting matrix acts on conjugate symmetric vectors $\mathbf{x} \in \mathbb{X}_d$ to yield an interpolated representation $\mathbf{A}\mathbf{x} \in \mathbb{R}^m$ of the associated Nyquist-rate sampled time domain signal $\mathbf{z} = \mathbf{F}_d\mathbf{x} \in \mathbb{R}^d$. The implied sampling procedure does not rely on any form of randomness, as samples are assumed to be recorded in regular intervals, which makes the model particularly favorable for hardware implementations.

To gauge the effectiveness of the proposed *conjugate symmetric binary iterative hard thresholding* (CS-BIHT) algorithm, as well as a closely related smooth variant termed CS-BIHT-ℓ_2, we investigated their empirical recovery performance on synthetic data. In this context, we benchmarked both algorithms against existing convex programming approaches, which we extended to our setting. The experiments confirm the correct behavior of the CS-BIHT and CS-BIHT-ℓ_2 algorithms and demonstrate their superior performance over convex approaches. This mirrors the behavior of BIHT in the Gaussian setting. At moderately large numbers of measurements, both algorithms also turn out to be highly effective at identifying active support indices of target signals. We also compared the performance of each reconstruction method with their corresponding oracle-assisted versions, which provide each scheme with the true support of the target signal. These experiments show that the CS-BIHT algorithms are significantly more effective in estimating the individual nonzero entries of sparse conjugate symmetric vectors (up to global scaling) than alternative approaches. While both algorithms rely on an estimate of the number of nonzero coefficients, they were demonstrated to still outperform competing convex programming approaches if the sparsity level is not overestimated too much. Finally, since the BIHT algorithm is highly susceptible to adversarial post-quantization bit flips, which also extends to our conjugate symmetric formulation, we combined the proposed CS-BIHT algorithm with the so-called *adaptive outlier pursuit* (AOP) algorithm. This method tries to adaptively identify and subsequently correct possibly erroneous bit positions during reconstruction to reduce the reconstruction task to a noiseless recovery problem. While the performance does not quite catch up to the noiseless setting, the availability of prior information about the expected number of erroneous measurements was shown to be highly valuable in practice to partially combat the performance loss due to measurement noise.

4

One-Bit Compressed Sensing of Group-Sparse Signals

In Chapter 3, we addressed the issue of estimating vectors with a sparse discrete Fourier transform from sub- or oversampled binary observations. Such sparsity models frequently appear in various domains of engineering such as wireless communication, audio and image signal processing and speech detection. While the numerical experiments of the previous chapter provide strong evidence that similar guarantees should hold for both 1-bit Gaussian and (random) Fourier measurements (at least on an average-case basis), theoretical results supporting this claim remain elusive.

In this chapter, we turn our attention to another common low-complexity signal model with widespread application in a variety of different domains. In particular, we will assume throughout the rest of this thesis that signals of interest exhibit sparsity not in terms of individual coefficients but w. r. t. nonoverlapping coefficient groups. Naturally, the elements of such signal ensembles are generally referred to as *group-sparse* signals. This sparsity model arises in various domains of science and engineering such as facial recognition

Parts of this chapter have been published in [**KBM19a**].

[JCM12], magnetic resonance imaging [CH12], subspace clustering [EV09], measurement of gene expression levels [Sim+13], as well as wireless communication [Qin+18]. To make the notion of group-sparsity precise, we consider vectors $\mathbf{x} \in \mathbb{R}^d$ for which the potential support set $[d]$ is decomposed into G nonoverlapping coefficient groups according to the following natural definition.

Definition 4.1 (Group partition)**.** *A collection* $\mathcal{I} = \{\mathcal{I}_1, \dots, \mathcal{I}_G\}$ *of subsets* $\mathcal{I}_i \subseteq [d] = \{1, \dots, d\}$ *is called a* group partition *of* $[d]$ *if* $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset \ \forall i \neq j$ *and* $\bigcup_{i=1}^G \mathcal{I}_i = [d]$*. We call a group partition* $\mathcal{I} = \{\mathcal{I}_1, \dots, \mathcal{I}_G\}$ ascending *if, for* $\mathcal{I}_i = \{j_{i,1}, \dots, j_{i,|\mathcal{I}_i|}\}$*, it holds that* $j_{i,k+1} = j_{i,k} + 1$ *for all* $k \in [|\mathcal{I}_i| - 1]$ *and* $\min \mathcal{I}_{i+1} = \max \mathcal{I}_i + 1$ *for all* $i \in [G-1]$*.*

Given a group partition $\mathcal{I}$, we denote as usual by $\mathbf{x}_{\mathcal{I}_i} \in \mathbb{R}^d$ the restriction of $\mathbf{x}$ to the indices in $\mathcal{I}_i$, *i.e.*, $(\mathbf{x}_{\mathcal{I}_i})_j = x_j \cdot \mathbb{1}_{\{j \in \mathcal{I}_i\}}$ for $j \in [d]$. A signal $\mathbf{x}$ is called *s-group-sparse* (w. r. t. the group partition $\mathcal{I}$) if it is supported on at most s groups. Note that Definition 4.1 does not necessarily assume that the elements in $\mathcal{I}_i$ are consecutive indices, nor that the cardinality of each individual subset $\mathcal{I}_i$ is identical. Imposing both these restrictions gives rise to the closely related notion of *block-sparsity*, where the support set $[d]$ is usually assumed to be decomposed into equisized groups of the form

$$\mathcal{I} = \{\{1, \dots, d/G\}, \{d/G + 1, \dots, 2d/G\}, \dots, \{(G-1)d/G + 1, \dots, d\}\} \tag{4.1}$$

with G assumed to divide d without remainder. The notion of group-sparsity therefore generalizes both sparse- and block-sparse signal models.

Lastly, we want to point out a third model which is itself closely related to the block-sparsity model. Assuming that a vector $\mathbf{x} \in \mathbb{R}^d$ is partitioned into G nonoverlapping blocks of size g according to the partition in Equation (4.1), the so-called *fusion-frame sparsity* model further assumes that each individual subvector $\mathbf{x}_{\mathcal{I}_i} \in \mathbb{R}^g$ belongs to some k_i-dimensional subspace $\mathcal{V}_i \subseteq \mathbb{R}^g$. Fusion-frame sparsity receives its name from a close connection to fusion frames, which constitute a generalization of the classical frame concept. In this context, a *fusion frame* is a collection $(\mathcal{V}_i, w_i)_{i=1}^G$ of subspaces $\mathcal{V}_i \subseteq \mathbb{R}^g$ and scalar weights w_i which satisfy the fusion frame condition [CK12, Chapter 13]

$$A\|\mathbf{z}\|_2^2 \leq \sum_{i=1}^G w_i^2 \|\Pi_{\mathcal{V}_i} \mathbf{z}\|_2^2 \leq B\|\mathbf{z}\|_2^2 \quad \forall \mathbf{z} \in \mathbb{R}^g$$

for some constants $0 < A \leq B < \infty$ with $\Pi_{\mathcal{V}_i}$ denoting the orthogonal projector on $\mathcal{V}_i$. The classical frame definition can be recovered from this generalization for $k_i = 1 \ \forall i$ in which case

$$\|\Pi_{\mathcal{V}_i} \mathbf{z}\|_2^2 = \left\| \frac{\boldsymbol{\varphi}_i \langle \boldsymbol{\varphi}_i, \mathbf{z} \rangle}{\|\boldsymbol{\varphi}_i\|_2^2} \right\|_2^2 = \frac{|\langle \boldsymbol{\varphi}_i, \mathbf{z} \rangle|^2}{\|\boldsymbol{\varphi}_i\|_2^2}$$

and $w_i = \|\boldsymbol{\varphi}_i\|_2$ with $\boldsymbol{\varphi}_i \in \mathbb{R}^g$ denoting the i-th frame vector. Given a fusion frame, one may now define the Hilbert space

$$\mathcal{H} := \left\{ (\mathbf{z}_i)_{i=1}^G : \mathbf{z}_i \in \mathcal{V}_i \right\} \subseteq \mathbb{R}^{gG} = \mathbb{R}^d$$

with a vector $\mathbf{z} \in \mathcal{H}$ being referred to as *s-fusion-frame-sparse* if at most s coefficient vectors $\mathbf{z}_i$ are different from zero. Note that some works addressing fusion-frame sparsity

like [ADR16; Aya18] actually do not require $(\mathcal{V}_i, w_i)_{i=1}^G$ to define a proper fusion frame. Instead, they merely ask that each set $\mathcal{V}_i$ forms a subspace of dimension $k_i = k$ and impose an additional incoherence assumption between individual subspaces.

The group-sparsity model described above is oftentimes referred to as the *nonoverlapping group-sparsity model* to contrast it with models which allow for coefficient groups to share common indices. Such models are often natural in the context of group model selection [Bal+16]. A classical example application is in genetics, in particular gene expression analysis, where certain genes might be present in multiple biological pathways and might therefore contribute to different (not necessarily mutually exclusive) genetic traits [HBM12]. Other examples include applications in cognitive neuroscience where *functional magnetic resonance imaging* (fMRI) is used to identify active regions (represented by voxels) in the brain associated with different cognitive states due to excitation caused by external stimuli [Rao+13]. Due to anatomical differences in the subjects' brains, activation regions will generally differ yet share common traits, which can be used to identify relevant voxel neighborhoods in subjects. In contrast, a typical example of nonoverlapping group-sparsity in wireless communication is in cognitive radio and in particular in dynamic spectrum sensing. In this context, the spectrum assigned to a single user in a *frequency-division multiple access* (FDMA) system might be composed of multiple nonconsecutive frequency subcarriers, giving rise to a nonoverlapping group sparsity structure of the spectrum. Since subbands can only be assigned to one individual user at a time in a particular time slot, the grouping structure is inherently nonoverlapping to avoid signal interference, which would be caused by multiple users communicating on shared subcarriers [Qin+18].

Chapter Outline

The remainder of this chapter is structured as follows. In Section 4.1, we introduce some basic notation for the nonoverlapping group-sparsity model considered in the rest of this thesis. We also present the acquisition model used in the first half of this chapter. In Section 4.2, we discuss the surprisingly limited body of pre-existing literature on block- and group-sparse signal recovery in the context of 1-bit compressed sensing. Similar to the acquisition model considered in the previous chapter, we will address the problem of recovering group-sparse signals on the sphere in Section 4.3. In particular, we consider three different recovery strategies modeled after existing schemes in the literature and analyze their theoretical performance. The section concludes with a numerical study to investigate potential performance gaps between the theoretically predicted and empirically observed behavior of the individual recovery schemes. By appealing to a well-known dithering strategy, we will lift the restriction of recovering unit-norm group-sparse vectors in Section 4.4 and address the issue of recovering both norm and direction of vectors inside scaled unit balls. In particular, we consider six recovery strategies, whose analyses depend on the guarantees established in Section 4.3. After a numerical study, we finally conclude the chapter in Section 4.5.

4.1 Signal and Acquisition Model

In the remainder of this thesis, we consider vectors with a group-sparse or *approximately* group-sparse representation w. r. t. an arbitrary orthonormal basis. In this chapter in

particular, we further restrict ourselves to the real setting. To put the notion of group-sparsity into well-defined terms, we first define the following family of mixed norms w. r. t. a given group partition $\mathcal{I}$.

Definition 4.2 (Group ℓ_p-norms)**.** *Let $\mathbf{x} \in \mathbb{R}^d$. Then the group ℓ_p-norm on $\mathbb{R}^d$ with $p \geq 1$ is defined as*

$$\|\mathbf{x}\|_{\mathcal{I},p} := \left(\sum_{i=1}^{G} \|\mathbf{x}_{\mathcal{I}_i}\|_2^p\right)^{1/p}.$$

Given the prevalence of the ℓ_1-norm in the canonical theory of compressed sensing, it comes as no surprise that the group ℓ_1-norm will take the place of the ℓ_1-norm in the group-sparse setting. Naturally, we also extend the notation $\|\cdot\|_{\mathcal{I},p}$ to $p = 0$ in which case $\|\cdot\|_{\mathcal{I},0}$ corresponds to the group ℓ_0-pseudonorm which merely measures the number of groups a vector is supported on. In this context, we will sometimes say that a group $\mathcal{I}_i \in \mathcal{I}$ is *active* if a vector $\mathbf{x} \in \mathbb{R}^d$ has at least one nonzero entry inside group $\mathcal{I}_i$. This leads to the following definition of the group ℓ_0-pseudonorm:

$$\|\mathbf{x}\|_{\mathcal{I},0} := |\{i \in [G] : \mathbf{x}_{\mathcal{I}_i} \neq \mathbf{0}\}|.$$

Due to the central importance of this signal ensemble in the remainder of this work, we introduce the shorthand notation

$$\Sigma_{\mathcal{I},s} := \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_{\mathcal{I},0} \leq s\right\}$$

to denote the set of s-group-sparse vectors w. r. t. the group partition $\mathcal{I}$. From a practical perspective, it is oftentimes more realistic, however, to model signals of interest as *approximately* group-sparse or *group-compressible*. An informal definition of this notion simply requires a signal $\mathbf{x} \in \mathbb{R}^d$ to be well-approximated by elements in $\Sigma_{\mathcal{I},s}$. With the definition of the so-called *best s-term group approximation error*

$$\sigma_s(\mathbf{x})_{\mathcal{I},p} := \inf_{\mathbf{z} \in \Sigma_{\mathcal{I},s}} \|\mathbf{x} - \mathbf{z}\|_{\mathcal{I},p},$$

a vector $\mathbf{x}$ is called *group-compressible* if $\sigma_s(\mathbf{x})_{\mathcal{I},p}$ rapidly decays as s increases. A notion closely related to group-compressibility is that of *effective group-sparsity*. In particular, we will heavily rely on the signal ensemble

$$\mathcal{E}_{\mathcal{I},s} := \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_{\mathcal{I},1} \leq \sqrt{s}\|\mathbf{x}\|_2\right\}.$$

Note that by the Cauchy-Schwarz inequality, every s-group-sparse vector is naturally effectively s-group-sparse and hence $\Sigma_{\mathcal{I},s} \subset \mathcal{E}_{\mathcal{I},s}$ (see also Lemma 4.9). Moreover, note that neither $\Sigma_{\mathcal{I},s}$ nor $\mathcal{E}_{\mathcal{I},s}$ are convex sets, which bears the potential to cause computational issues when trying to enforce the respective signal structure during recovery. However, due to the nature of the recovery algorithms considered in this chapter, enforcing such membership constraints turns out to be a relatively straightforward task.

In general, we consider measurements of the form

$$\mathbf{y} = Q(\mathbf{A}\mathbf{x}) = Q(\mathbf{\Phi}\mathbf{\Psi}\mathbf{x}) \in \{\pm 1\}^m$$

where $\mathbf{\Phi} \in \mathbb{R}^{m\times d}$ denotes the linear measurement operator modeling the physical acquisition system (before quantization), $\mathbf{\Psi} \in \mathrm{O}(d) := \{\mathbf{Q} \in \mathbb{R}^{d\times d} : \mathbf{Q}^\top\mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathrm{Id}_d\}$ denotes the sparsity basis[1], and $Q\colon \mathbb{R}^m \to \{\pm 1\}^m$ is a 1-bit quantization map which models both the quantization process as well as possible measurement noise. Note that in a real system, one would only have access to the (generally non-group-sparse) expansion $\mathbf{z} = \mathbf{\Psi}\mathbf{x}$ through the measurement operator $\mathbf{\Phi}$. This was the case in the previous chapter where $\mathbf{\Psi} = \mathbf{F}_d$ was assumed to be an orthogonal DFT matrix, and $\mathbf{\Phi}$ was chosen as a random subsampler. Sparsity in the frequency domain was then exploited by acquiring 1-bit time domain samples below (or above the Nyquist rate) since direct access to the frequency domain representation is not available due to the acoustic or electromagnetic nature of signals of interest in many engineering applications. For $\mathbf{\Psi} \neq \mathrm{Id}_d$, reconstruction of $\mathbf{z}$ from its compressive measurements $\mathbf{y} \in \{\pm 1\}^m$ therefore proceeds in a two-step process where one first estimates the group-sparse expansion $\hat{\mathbf{x}}$ before synthesizing the estimate $\hat{\mathbf{z}} = \mathbf{\Psi}\hat{\mathbf{x}}$. However, since we are mainly interested in error estimates of the form $\|\mathring{\mathbf{z}} - \hat{\mathbf{z}}\|_2 \leq \varepsilon$ for some prescribed accuracy ε, it suffices to consider $\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2$ due to orthogonality of $\mathbf{\Psi}$. For simplicity, we therefore refer to the composite matrix $\mathbf{A} = \mathbf{\Phi}\mathbf{\Psi}$ in the sequel of this chapter as the measurement matrix of the system rather than making a distinction between $\mathbf{\Phi}$ and $\mathbf{A}$.

4.2 Prior Work

As alluded to in the introduction of this chapter, the number of works in the literature which address the issue of block- or group-sparse signal reconstruction from 1-bit or higher-order quantized measurements is surprisingly limited. A notable exception is the work by Zeng and Figueiredo [**ZF14b**] who consider the recovery of block-sparse signals by combining the binary iterative hard thresholding algorithm discussed in the previous chapter with an intermediate projection on a scaled *total variation* (TV) (semi)norm ball. Rather than explicitly exploiting the group structure in the signal support, their method, termed *binary fused compressive sensing* (BFCS), instead assumes that only the total sparsity level is known a priori. The underlying group structure assumed in the target signal is then enforced by "fusing" neighboring coefficients together by a projection on a TV-ball $\nu\mathbb{B}^d_{\mathrm{TV}}$ of radius ν for some appropriately chosen value $\nu > 0$. In particular, they consider the update equation

$$\mathbf{x}^+ = (\Pi_{\nu\mathbb{B}^d_{\mathrm{TV}}} \circ \mathcal{H}_s)(\mathbf{x} - \mu J(\mathbf{x})) \tag{4.2}$$

with $J(\mathbf{x})$ denoting either a subgradient of the functional $\|[\mathbf{y} \circ \mathbf{A}\mathbf{x}]_-\|_1$ or the gradient of $\frac{1}{2}\|[\mathbf{y} \circ \mathbf{A}\mathbf{x}]_-\|_2^2$ (cf. Section 3.2). Note that in their formulation, s denotes the total number of nonzero coefficients in a vector rather than the number of active groups. As pointed out in [**ZF14a**], the operators $\mathcal{H}_s$ and $\Pi_{\nu\mathbb{B}^d_{\mathrm{TV}}}$ do not commute. However, the projection on a scaled TV-ball preserves the sparsity structure of a vector such that any new iterate $\mathbf{x}^+$ is guaranteed to belong to the set of target signals $\nu\mathbb{B}^d_{\mathrm{TV}} \cap \Sigma_s$. We point out that projecting on a TV-seminorm ball inherently assumes that signal coefficients exhibit a certain regularity in the sense that they have the same sign pattern and are close

[1] We abuse terminology and usually refer to an orthogonal matrix $\mathbf{Q} \in \mathrm{O}(d)$ as a *basis* for $\mathbb{R}^d$ rather than to the collection $\{\mathbf{q}_i\}_{i=1}^d$ of its columns $\mathbf{q}_i$ which span $\mathbb{R}^d$.

to each other in magnitude. Without this structure in the original signal, the methods proposed in [ZF14b; ZF14a] fail to accurately recover general group-sparse vectors.

To exploit prior knowledge on the coefficient groups, the authors extended their method in [ZF14a] by replacing the projection on $\nu\mathbb{B}_{\mathrm{TV}}^d$ with a projection on a set whose subvectors belong to scaled TV-balls. In particular, given an ascending group partition $\mathcal{I} = \{\mathcal{I}_1, \dots, \mathcal{I}_G\}$, they define the set

$$\mathcal{T}_\lambda := \left\{ \mathbf{x} = (\mathbf{x}_{\mathcal{I}_i})_{i=1}^G \in \mathbb{R}^d : \mathbf{x}_{\mathcal{I}_i} \in \frac{\lambda}{|\mathcal{I}_i| - 1} \mathbb{B}_{\mathrm{TV}}^{|\mathcal{I}_i|} \right\}$$

of vectors whose subvectors indexed by the partitions $\mathcal{I}_i$ belong to scaled TV-balls. Note that with the choice of the singleton partition $\mathcal{I} = \{[d]\}$, projection on $\mathcal{T}_\lambda$ coincides with a projection on $\nu\mathbb{B}_{\mathrm{TV}}^d$ with $\nu = \lambda/(d-1)$ as used in the update rule (4.2). In addition to explicitly exploiting prior knowledge of possible group structures, the authors also include a mechanism to correct possible adversarial bit flips of the binary measurements during signal reconstruction. Their approach combines the modified BFCS algorithm with the AOP method due to Yan, Yang and Osher [YYO12] previously discussed in Section 3.4.3. Due to the added noise-robustness, the resulting method is termed *robust binary fused compressive sensing* (RoBFCS).

Since both the BFCS and RoBFCS algorithm utilize at their core the BIHT algorithm, no theoretical recovery guarantee nor convergence results are available for the presented algorithms. Moreover, the projection on the sets $\nu\mathbb{B}_{\mathrm{TV}}^d$ and $\mathcal{T}_\lambda$ are nontrivial and do not admit simple closed-form solutions. As a result, one needs to solve a convex program at each iteration of the respective algorithm, which significantly impacts performance. While the authors of [ZF14b; ZF14a] utilize an iterative projection scheme to ease computational burdens, informal experiments conducted by the present author revealed that the method is not competitive even if an efficient convex solver such as MOSEK [Mos10] is employed to solve the projection subproblem. The added time complexity therefore renders the resulting algorithms intractable for moderately-sized problem instances.

On the theoretical side, results are mainly limited to the work by Rao *et al.* who consider classification of group-sparse vectors with overlapping groups from binary measurements in [Rao+13; Rao+14]. They appeal to the general framework proposed by Plan and Vershynin in [PV13b] to deal with a large class of measurement nonlinearities and derive bounds on the mean width of a specifically designed constraint set which models group-sparse signals with overlapping groups and within-group sparsity. A similar result will also be presented below in Section 4.3.2 in the case of nonoverlapping group partitions.

4.3 Direction Recovery of Group-Sparse Signals

In this section, we first consider the recovery of (effectively) group-sparse vectors from quantized measurements of the form

$$\mathbf{x} \mapsto \operatorname{sgn}(\mathbf{A}\mathbf{x}) = \mathbf{y}.$$

We emphasize again that due to the scale invariance of the sgn-operator, there is no hope to recover anything more than the direction of a vector. For notational convenience, we

will write

$$\widetilde{\Sigma}_{\mathcal{I},s} := \Sigma_{\mathcal{I},s} \cap \mathbb{S}^{d-1} = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_{\mathcal{I},0} \leq s, \|\mathbf{x}\|_2 = 1\right\}$$

and

$$\widetilde{\mathcal{E}}_{\mathcal{I},s} := \mathcal{E}_{\mathcal{I},s} \cap \mathbb{S}^{d-1} = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_{\mathcal{I},1} \leq \sqrt{s}, \|\mathbf{x}\|_2 = 1\right\}$$

throughout this chapter to denote the sets of genuinely and effectively group-sparse vectors with unit Euclidean norm, respectively.

Before considering concrete recovery schemes for the sensing model described above, it is natural to first revisit the question of the worst-case reconstruction error for arbitrary (possibly intractable) recovery maps as discussed in Section 2.3.1. In the sparse case, the proof of (2.4) provided in [Jac+13] is based on an intricate covering argument of the $\binom{d}{s}$ s-dimensional unit spheres contained in $\widetilde{\Sigma}_s$. For the group-sparse signal model with a group partition $\mathcal{I}$ with $|\mathcal{I}| = G$, one instead considers $\binom{G}{s}$ unit spheres in coordinate subspaces of dimension at most sg. In other words, the number of groups G replaces the ambient dimension d in the canonical sparsity case. While this changes the maximal number of quantization cells occupied by the set $\operatorname{sgn}(\mathbf{A}\widetilde{\Sigma}_{\mathcal{I},s})$ for an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{m\times d}$ from $2^s\binom{m}{s}\binom{d}{s}$ to $2^{sg}\binom{m}{sg}\binom{G}{s}$ (cf. [Jac+13, Lemma 1]), the worst-case reconstruction error remains independent of G. In fact, reviewing the arguments presented in [Jac+13, Appendix B], it immediately follows that the worst-case reconstruction error

$$\epsilon_{\mathrm{opt}} = \sup_{\mathbf{x}\in\widetilde{\Sigma}_{\mathcal{I},s}} \inf_{\mathbf{q}\in\mathcal{Q}} \|\mathbf{x} - \mathbf{q}\|_2.$$

for an optimal subset $\mathcal{Q} \subset \widetilde{\Sigma}_{\mathcal{I},s}$ only depends on the dimension of the individual coordinate subspaces and is therefore bounded below by

$$\epsilon_{\mathrm{opt}} \gtrsim \frac{sg}{(sg)^{3/2} + m}.$$

As before, this implies that for a fixed group-sparsity level, the reconstruction error of any nonadaptive method decays at most linearly in the number of measurements.

4.3.1 Quantization-Consistent Reconstruction

The first recovery approach we consider in this section is inspired by the convex program

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \mathbf{y} \circ \mathbf{A}\mathbf{x} \geq \mathbf{0} \\ & \langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle = m\sqrt{2/\pi}, \end{aligned} \tag{P$_{4.1}$}$$

which was already briefly discussed in Section 2.3.1. It was originally proposed by Plan and Vershynin in [PV13a] to estimate effectively s-sparse vectors $\mathring{\mathbf{x}} \in \mathbb{S}^{d-1}$ from single-bit quantized measurements $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}})$. Moreover, it was the first reconstruction scheme with provably accurate reconstruction performance with an almost optimal dependence on d and s. In particular, the authors showed via results on random hyperplane tessellations that faithful signal recovery of unit-normalized effectively sparse vectors is possible from

$\Omega(s\log(d/s)^2)$ measurements. The dependence on the factor $\log(d/s)^2$ was later improved to $\log(d/s)$ by Foucart, which is known to be optimal [**Fou16**]. In his work, Foucart also significantly reduces the proof complexity of the result by appealing to a variation of the classical restricted isometry property which had previously been employed in [**JHF11**] in the context of signal recovery from uniformly quantized multi-bit measurements.

Recall from Section 2.3.1 that the first constraint of Problem ($\mathrm{P}_{4.1}$) is motivated by quantization consistency (cf. Definition 2.6) and is essentially equivalent to the nonconvex constraint $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathbf{x})$ as we have

$$\big\{\mathbf{x} \in \mathbb{R}^d : \mathbf{y} \circ \mathbf{A}\mathbf{x} \geq \mathbf{0}\big\} = \big\{\mathbf{x} \in \mathbb{R}^d : \operatorname{sgn}(\mathbf{A}\mathbf{x}) = \mathbf{y}\big\} \cup \ker(\mathbf{A}).$$

Due to the convention that $\operatorname{sgn}(x) = -1$ if $x < 0$ and $\operatorname{sgn}(x) = 1$ if $x \geq 0$, any vector in the null space of $\mathbf{A}$ is mapped to the all-ones vector $\mathbf{1}_m$ under $\operatorname{sgn}(\mathbf{A}\cdot)$. As a consequence, vectors in the null space of $\mathbf{A}$ can never be distinguished from one another. At first sight, this suggests that Problem ($\mathrm{P}_{4.1}$) may actually be too optimistic to yield meaningful estimates since its search space includes the entire null space of $\mathbf{A}$ despite the fact that such vectors are only quantization-consistent with the measurement vector $\mathbf{y} = \mathbf{1}_m$. Since Problem ($\mathrm{P}_{4.1}$) seeks to minimize a norm (and hence a positive definite function), however, the only immediately problematic vector in the feasible set is the zero vector. While enforcing $\mathbf{x}$ to be different from zero is difficult in practice, we may instead impose a constraint that requires $\mathbf{x}$ to miss the null space of $\mathbf{A}$ altogether if we also impose that $\mathbf{x}$ must *not* be orthogonal to $\mathbf{A}^\top\mathbf{y}$. This is the purpose of the second constraint of Problem ($\mathrm{P}_{4.1}$), which results in the feasible set

$$\begin{aligned}\big\{\mathbf{x} \in \mathbb{R}^d : \mathbf{y} \circ \mathbf{A}\mathbf{x} \geq \mathbf{0}, \langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle = c_0\big\} &= \big\{\mathbf{x} \in \mathbb{R}^d : \mathbf{y} \circ \mathbf{A}\mathbf{x} > \mathbf{0}, \langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle = c_0\big\} \\ &= \big\{\mathbf{x} \in \mathbb{R}^d : \mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathbf{x}), \langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle = c_0\big\}\end{aligned}$$

for some arbitrary constant $c_0 > 0$. Since elements in the null space of $\mathbf{A}$ cannot be distinguished based on their binary observations in the first place, the first restriction that $\mathbf{x} \notin \ker(\mathbf{A})$ is clearly justified. On the other hand, the additional condition that $\mathbf{x} \notin (\mathbf{A}^\top\mathbf{y})^\perp$ is a natural requirement since $\langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle = \langle \mathbf{A}^\top\mathbf{y}, \mathbf{x}\rangle$ measures the correlation between the quantized and unquantized measurements, which naturally should be different from zero for any minimizer $\mathbf{x}^\star$ (cf. Section 4.3.2). The constant $c_0 = m\sqrt{2/\pi}$ on the right-hand side of the second constraint of Problem ($\mathrm{P}_{4.1}$) is ultimately arbitrary due to the scale invariance of the problem as it only affects $\mathbf{x}^\star$ but not $\mathbf{x}^\star/\|\mathbf{x}^\star\|_2$. The particular choice above can be motivated by the observation that for any $\mathbf{x}$ satisfying the condition $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathbf{x})$, the second constraint is simply $\langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle = \|\mathbf{A}\mathbf{x}\|_1$. If $\mathbf{A}$ consists of i.i.d. standard Gaussian entries, it follows that $\mathbb{E}\|\mathbf{A}\mathbf{x}\|_1 = m\sqrt{2/\pi}\|\mathbf{x}\|_2$, *i.e.*, in expectation, any solution of Problem ($\mathrm{P}_{4.1}$) lies on a scaled ℓ_2-ball.

Motivated by this program, we formulate the following natural recovery procedure

$$\begin{aligned}\underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_{\mathcal{I},1} \\ \text{s.t.} \quad & \mathbf{y} \circ \mathbf{A}\mathbf{x} \geq \mathbf{0} \\ & \langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle = 1\end{aligned} \tag{$\mathrm{P}_{4.2}$}$$

to estimate group-sparse vectors from their quantized projections with which we associate the recovery map $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}} \colon \{\pm 1\}^m \to \mathbb{R}^d$ defined as

$$\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}(\mathbf{y}) = \operatorname*{arginf}_{\mathbf{x}} \big\{\|\mathbf{x}\|_{\mathcal{I},1} : \mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathbf{x}), \|\mathbf{A}\mathbf{x}\|_1 = 1\big\}.$$

In keeping with our notation for (effectively) group-sparse vectors on the sphere, we will generally denote recovery maps for the sets $\widetilde{\Sigma}_{\mathcal{I},s}$ and $\widetilde{\mathcal{E}}_{\mathcal{I},s}$ by $\widetilde{\Delta}$. Given a measurement matrix $\mathbf{A}$ and quantized measurements $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}) \in \{\pm 1\}^m$, the operator $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ can be realized in the same fashion as Problem ($\mathrm{P}_{4.1}$) after rewriting the constraints into convex form.

The original analysis of Problem ($\mathrm{P}_{4.1}$) by Plan and Vershynin was based on a rather complicated hyperplane tessellation argument, in addition to a counting argument of the number of vertices of the feasible polytope. In this work, we will follow the simplified proof strategy suggested by Foucart in [**Fou16**]. Due to the close connection between Problem ($\mathrm{P}_{4.2}$) and Problem ($\mathrm{P}_{4.1}$), the performance analysis of the recovery map $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ will therefore heavily rely on the following variation of the restricted isometry property for genuine or effectively group-sparse vectors. The version we state here also allows for group-sparsity bases other than the canonical one.

Definition 4.3 ((ℓ_2, ℓ_1)-group restricted isometry property)**.** *A matrix $\mathbf{A} = \mathbf{\Phi}\mathbf{\Psi} \in \mathbb{R}^{m \times d}$ with $\mathbf{\Phi} \in \mathbb{R}^{m \times d}$ and $\mathbf{\Psi} \in \mathrm{O}(d) = \{\mathbf{Q} \in \mathbb{R}^{d \times d} : \mathbf{Q}^\top \mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathrm{Id}_d\}$ is said to satisfy the* (ℓ_2, ℓ_1) group restricted isometry property *(group-RIP) of order s if*

$$(1 - \delta)\|\mathbf{x}\|_2 \leq \|\mathbf{\Phi}\mathbf{\Psi}\mathbf{x}\|_1 \leq (1 + \delta)\|\mathbf{x}\|_2 \ \forall \mathbf{x} \in \Sigma_{\mathcal{I},s} \tag{4.3}$$

for some $\delta \in (0, 1)$. The smallest constant $\delta_s \leq \delta$ for which (4.3) holds is called the group restricted isometry constant *(group-RIC) of $\mathbf{A}$.*

By replacing the set $\Sigma_{\mathcal{I},s}$ with $\mathcal{E}_{\mathcal{I},s}$ in (4.3), we obtain the definition of the group-RIP matrices for the set of effectively s-group-sparse vectors. In this case, we also say that $\mathbf{A}$ satisfies the *effective group-RIP*.

The performance analysis of the recovery strategy will proceed along the lines of the analysis in the case of sparse vectors as demonstrated by Foucart in [**Fou16**]. First, we show in Lemma 4.4 that—given a measurement matrix $\mathbf{A} \in \mathbb{R}^{m \times d}$ satisfying the group-RIP of order t with constant δ_t—every convex combination $(1 - \lambda)\mathring{\mathbf{x}} + \lambda\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}(\operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}))$ with $\lambda \in [0, 1]$ is effectively t-group-sparse for an appropriate choice of $t > s$. This will immediately imply an upper bound on the recovery error $\|\mathring{\mathbf{x}} - \widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}(\operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}))\|_2$ in terms of the isometry constant δ_t (cf. Lemma 4.6). We then show in Lemma 4.11 that a scaled version of the matrix $\mathbf{A} = \mathbf{\Phi}\mathbf{\Psi}$ with $\mathbf{\Psi} \in \mathrm{O}(d)$ and $\mathbf{\Phi}$ consisting of independent standard Gaussian entries satisfies the (ℓ_2, ℓ_1) group restricted isometry property with high probability on the draw of $\mathbf{\Phi}$. The final recovery guarantee will then be a simple consequence of Lemma 4.6, Lemma 4.11 and Remark 4.5(iii) below.

We want to point out that the above strategy ultimately hinges on the ability to show that random measurement matrices satisfy the (ℓ_2, ℓ_1)-group-RIP with high probability. In light of the vast number of results which establish the classical restricted isometry property for many types of random measurement ensembles in the canonical theory, including subgaussian designs and measurement matrices constructed from subsampled bounded orthonormal systems, one might wonder why such results are not readily available in the 1-bit CS setting. In short, the existence of subgaussian, say, (ℓ_2, ℓ_1)-RIP matrices would contradict the observation that one can easily construct certain distinct sparse vectors with the same sign pattern when observed by the map $\mathbf{u} \mapsto \operatorname{sgn}(\mathbf{A}\mathbf{u})$. The same argument applies to group-sparse vectors which is why we limit our attention to Gaussian ensembles in this chapter.

We begin by first showing the following intermediate result, which we state here in a slightly more general form than its counterpart for sparse or effectively sparse vectors first presented in [Fou16].

Lemma 4.4. *Let $\mathring{\mathbf{x}} \in \widetilde{\Sigma}_{\mathcal{I},s}$, and assume $\mathbf{A} \in \mathbb{R}^{m\times d}$ satisfies the group-RIP of order $t > s$ with constant $\delta_t < (\sqrt{t}-\sqrt{s})/(\sqrt{t}+\sqrt{s})$. Let further $\hat{\mathbf{x}} = \widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}(\operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}))$. Then it holds for $\bar{\mathbf{x}} = (1-\lambda)\mathring{\mathbf{x}} + \lambda\hat{\mathbf{x}}$ with $\lambda \in [0,1]$ that*

$$\frac{\|\bar{\mathbf{x}}\|_{\mathcal{I},1}}{\|\bar{\mathbf{x}}\|_2} \leq \frac{(1+\delta_t)\sqrt{st}}{(1-\delta_t)\sqrt{t}-(1+\delta_t)\sqrt{s}} = \frac{\sqrt{st}}{\frac{1-\delta_t}{1+\delta_t}\sqrt{t}-\sqrt{s}}, \tag{4.4}$$

i.e., $\bar{\mathbf{x}}$ is effectively group-sparse.

Remark 4.5. *(i) Note that (4.4) in Lemma 4.4 shows that the smaller the group-RIP constant δ_t of $\mathbf{A}$, the "more" effectively group-sparse any convex combination $\bar{\mathbf{x}}$ will become. As usual in compressed sensing, this (desirable) behavior comes at a price as the number of measurements required for $\mathbf{A}$ to satisfy the group-RIP with high probability scales with δ_t^{-2} (cf. Lemma 4.11 below).*

(ii) The statement of Lemma 4.4 still holds if $\mathring{\mathbf{x}}$ is effectively s-group-sparse once $\mathbf{A}$ satisfies the effective group-RIP of order t with constant δ_t. In fact, the recovery guarantee in Lemma 4.6 below depends on this exact variation of Lemma 4.4.

(iii) In Foucart's original proof, the parameters t and δ_t were chosen such that any convex combination $\bar{\mathbf{x}}$ is effectively t-sparse. Under this requirement and in light of (4.4), we can express t as a function of s and δ_t, which yields the condition

$$t = 4s\left(\frac{1+\delta_t}{1-\delta_t}\right)^2 \tag{4.5}$$

*for convex combinations $\bar{\mathbf{x}}$ to be effectively t-group-sparse. For instance, the choice $\delta_t = 1/5$ as in [**Fou16**] yields $t = 9s$.*

Proof of Lemma 4.4. The proof of the result follows the example of Foucart's original proof of Lemma 4 in [Fou16]. In order to bound the effective sparsity of a vector $\bar{\mathbf{x}} = (1-\lambda)\mathring{\mathbf{x}} + \lambda\hat{\mathbf{x}}$, we aim to establish an upper bound on the group ℓ_1-norm of $\bar{\mathbf{x}}$, as well as a lower bound on its ℓ_2-norm (which coincides with the group ℓ_2-norm). From the triangle inequality, we immediately have

$$\|\bar{\mathbf{x}}\|_{\mathcal{I},1} \leq (1-\lambda)\|\mathring{\mathbf{x}}\|_{\mathcal{I},1} + \lambda\|\hat{\mathbf{x}}\|_{\mathcal{I},1} \leq (1-\lambda)\sqrt{s} + \lambda\|\hat{\mathbf{x}}\|_{\mathcal{I},1}$$

where the last step follows from the Cauchy-Schwarz inequality $\|\cdot\|_{\mathcal{I},1} \leq \|\cdot\|_{\mathcal{I},0}^{1/2}\|\cdot\|_{\mathcal{I},2}$ and the fact that $\mathring{\mathbf{x}} \in \widetilde{\Sigma}_{\mathcal{I},s} \subset \mathbb{S}^{d-1}$. Next, since $\hat{\mathbf{x}} = \widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}(\operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}))$ and the vector $\mathring{\mathbf{x}}/\|\mathbf{A}\mathring{\mathbf{x}}\|_1$ is clearly feasible for Problem ($\mathrm{P}_{4.2}$), we have

$$\|\hat{\mathbf{x}}\|_{\mathcal{I},1} \leq \left\|\frac{\mathring{\mathbf{x}}}{\|\mathbf{A}\mathring{\mathbf{x}}\|_1}\right\|_{\mathcal{I},1} \leq \frac{\sqrt{s}\|\mathring{\mathbf{x}}\|_{\mathcal{I},2}}{(1-\delta_t)\|\mathring{\mathbf{x}}\|_2} = \frac{\sqrt{s}}{1-\delta_t}$$

where in the second step we used the fact that $\mathbf{A}$ is an (ℓ_2, ℓ_1) group restricted isometry matrix with constant δ_t. Overall, this yields

$$\|\bar{\mathbf{x}}\|_{\mathcal{I},1} \leq \sqrt{s}\left(1 + \frac{\delta_t \lambda}{1 - \delta_t}\right). \tag{4.6}$$

To construct a suitable lower bound on $\|\bar{\mathbf{x}}\|_2$, first note that for $a, b \in \mathbb{R} \setminus \{0\}$, we have $|a + b| = |a| + |b|$ if $\operatorname{sgn}(a) = \operatorname{sgn}(b)$. By feasibility of $\hat{\mathbf{x}}$, we have $\operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}) = \operatorname{sgn}(\mathbf{A}\hat{\mathbf{x}})$. With the previous observation, this yields

$$\|\mathbf{A}\bar{\mathbf{x}}\|_1 = (1 - \lambda)\|\mathbf{A}\mathring{\mathbf{x}}\|_1 + \lambda\|\mathbf{A}\hat{\mathbf{x}}\|_1 \geq (1 - \lambda)(1 - \delta_t) + \lambda = 1 + \delta_t(\lambda - 1) \tag{4.7}$$

where again we used that $\mathbf{A}$ satisfies the group-RIP with constant δ_t in combination with the fact that as a minimizer of Problem (P$_{4.2}$), $\hat{\mathbf{x}}$ satisfies $\|\mathbf{A}\hat{\mathbf{x}}\|_1 = 1$. It remains to obtain an upper bound on $\|\mathbf{A}\bar{\mathbf{x}}\|_1$ in terms of $\|\bar{\mathbf{x}}\|_2$. To that end, one follows a common methodology in compressed sensing.

Given a vector $\mathbf{x} \in \mathbb{R}^d$, denote by $\mathcal{T}_1 \subset \mathcal{I}$ the t groups with largest ℓ_2-norms, by $\mathcal{T}_2$ the t groups with next largest ℓ_2-norms and so on. With slight abuse of notation we then write $\mathbf{x}_{\mathcal{T}_i}$ to mean the vector that agrees with $\mathbf{x}$ on the index set $\bigcup_{S \in \mathcal{T}_i} S$ and vanishes identically otherwise. Then we have by the triangle inequality and an application of the group-RIP condition that

$$\|\mathbf{A}\bar{\mathbf{x}}\|_1 \leq \sum_{i \geq 1} \|\mathbf{A}\bar{\mathbf{x}}_{\mathcal{T}_i}\|_1 \leq (1 + \delta_t)\left(\|\bar{\mathbf{x}}_{\mathcal{T}_1}\|_2 + \sum_{i \geq 2} \|\bar{\mathbf{x}}_{\mathcal{T}_i}\|_2\right).$$

Next, note that we have with $\|\bar{\mathbf{x}}_S\|_2 \leq \|\bar{\mathbf{x}}_R\|_2$ for all $S \in \mathcal{T}_i$ and $R \in \mathcal{T}_{i-1}$ that

$$\|\bar{\mathbf{x}}_S\|_2 \leq \frac{1}{t} \sum_{R \in \mathcal{T}_{i-1}} \|\bar{\mathbf{x}}_R\|_2$$

and therefore

$$\begin{aligned} \|\bar{\mathbf{x}}_{\mathcal{T}_i}\|_2 &= \left(\sum_{S \in \mathcal{T}_i} \|\bar{\mathbf{x}}_S\|_2^2\right)^{1/2} \leq \frac{1}{t} \sum_{R \in \mathcal{T}_{i-1}} \|\bar{\mathbf{x}}_R\|_2 \left(\sum_{S \in \mathcal{T}_i} 1\right)^{1/2} \\ &= \frac{1}{\sqrt{t}} \sum_{R \in \mathcal{T}_{i-1}} \|\bar{\mathbf{x}}_R\|_2 = \frac{1}{\sqrt{t}} \left\|\mathbf{x}_{\mathcal{T}_{i-1}}\right\|_{\mathcal{I},1}. \end{aligned} \tag{4.8}$$

Overall, this yields

$$\|\mathbf{A}\bar{\mathbf{x}}\|_1 \leq (1 + \delta_t)\left(\|\bar{\mathbf{x}}_{\mathcal{T}_1}\|_2 + \frac{1}{\sqrt{t}} \sum_{i \geq 2} \left\|\bar{\mathbf{x}}_{\mathcal{T}_{i-1}}\right\|_{\mathcal{I},1}\right) \leq (1 + \delta_t)\left(\|\bar{\mathbf{x}}\|_2 + \frac{1}{\sqrt{t}} \|\bar{\mathbf{x}}\|_{\mathcal{I},1}\right).$$

Combining this estimate with (4.6) and (4.7), we find

$$1 + \delta_t(\lambda - 1) \leq (1 + \delta_t)\left(\|\bar{\mathbf{x}}\|_2 + \sqrt{\frac{s}{t}}\left(1 + \frac{\delta_t \lambda}{1 - \delta_t}\right)\right).$$

Solving for $\|\bar{\mathbf{x}}\|_2$, this yields with (4.6) that

$$\frac{\|\bar{\mathbf{x}}\|_{\mathcal{I},1}}{\|\bar{\mathbf{x}}\|_2} \leq \frac{\sqrt{s}\left(1 + \frac{\delta_t \lambda}{1 - \delta_t}\right)}{\frac{1 + \delta_t(\lambda - 1)}{1 + \delta_t} - \sqrt{\frac{s}{t}}\left(1 + \frac{\delta_t \lambda}{1 - \delta_t}\right)} = \frac{\sqrt{st}}{\frac{1 - \delta_t}{1 + \delta_t}\sqrt{t} - \sqrt{s}},$$

which completes the proof. □

With effective group-sparsity of convex combinations established, we now turn to characterizing the recovery performance of the map $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$.

Lemma 4.6. *Let $\mathring{\mathbf{x}} \in \mathbb{S}^{d-1}$ be effectively s-group-sparse, i.e., $\|\mathring{\mathbf{x}}\|_{\mathcal{I},1} \leq \sqrt{s}$. Fix a value $\delta_t \in (0, 1/5]$, and assume that $\mathbf{A} \in \mathbb{R}^{m\times d}$ satisfies the effective group-RIP of order t with constant δ_t for t chosen according to (4.5). Then*

$$\left\|\mathring{\mathbf{x}} - \widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}(\operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}))\right\|_2 \leq 4\sqrt{\delta_t}.$$

Proof. Denote as before by $\hat{\mathbf{x}} = \widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}(\operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}))$. From the parallelogram identity, we immediately have that

$$\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2^2 + \|\mathring{\mathbf{x}} + \hat{\mathbf{x}}\|_2^2 = 2\Big(1 + \|\hat{\mathbf{x}}\|_2^2\Big)$$

and thus

$$\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2^2 = 2\Big(1 + \|\hat{\mathbf{x}}\|_2^2\Big) - 4\left\|\frac{\mathring{\mathbf{x}} + \hat{\mathbf{x}}}{2}\right\|_2^2. \tag{4.9}$$

Moreover, since $\mathbf{A}$ satisfies the effective group-RIP of order t, the average $(\mathring{\mathbf{x}} + \hat{\mathbf{x}})/2$ is effectively t-group-sparse by Lemma 4.4. We therefore have

$$\left\|\frac{\mathring{\mathbf{x}} + \hat{\mathbf{x}}}{2}\right\|_2 \geq \frac{\|\mathbf{A}(\mathring{\mathbf{x}} + \hat{\mathbf{x}})\|_1}{2(1+\delta_t)} = \frac{\|\mathbf{A}\mathring{\mathbf{x}}\|_1 + \|\mathbf{A}\hat{\mathbf{x}}\|_1}{2(1+\delta_t)} \geq \frac{(1-\delta_t)\|\mathring{\mathbf{x}}\|_2 + 1}{2(1+\delta_t)} = \frac{1-\delta_t/2}{1+\delta_t}.$$

Next, observe that by Lemma 4.4, the vector $\hat{\mathbf{x}}$ is effectively t-group-sparse and therefore by the definition of the group-RIP condition for effectively group-sparse vectors we find

$$\|\hat{\mathbf{x}}\|_2 \leq \frac{\|\mathbf{A}\hat{\mathbf{x}}\|_1}{1-\delta_t} = \frac{1}{1-\delta_t}$$

where the last step follows from the fact that $\hat{\mathbf{x}}$ is feasible for Problem ($\mathrm{P}_{4.2}$). Combining the previous two estimates with (4.9) therefore yields

$$\begin{aligned}\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2^2 &\leq 2\left(1 + \frac{1}{(1-\delta_t)^2}\right) - 4\left(\frac{1-\delta_t/2}{1+\delta_t}\right)^2 \\ &= \delta_t \frac{\delta_t^3 + 6\delta_t^2 - 15\delta_t + 16}{(1-\delta_t^2)^2}.\end{aligned}$$

One easily verifies that the fractional polynomial is monotonically decreasing in δ_t on $(0, 1/5]$ so that overall we find $\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \leq 4\sqrt{\delta_t}$ as claimed. □

Remark 4.7. *Note that in general the estimator $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ does not produce unit-normalized vectors. Since undithered sign-measurements do not carry any information about the signal energy, we are therefore mainly interested in quantifying the reconstruction quality based solely on the estimated direction of the signal. To that end, note that by the triangle inequality, it follows immediately from the conclusion of Lemma 4.6 that*

$$\left\|\mathring{\mathbf{x}} - \frac{\hat{\mathbf{x}}}{\|\hat{\mathbf{x}}\|_2}\right\|_2 \leq \|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 + \left\|\hat{\mathbf{x}} - \frac{\hat{\mathbf{x}}}{\|\hat{\mathbf{x}}\|_2}\right\|_2 \leq 2\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \leq 8\sqrt{\delta_t},$$

which holds because $\hat{\mathbf{x}}/\|\hat{\mathbf{x}}\|_2$ is a better ℓ_2-normalized approximation of $\hat{\mathbf{x}}$ than $\mathring{\mathbf{x}}$ since $\hat{\mathbf{x}}/\|\hat{\mathbf{x}}\|_2 = \operatorname{argmin}_{\mathbf{x}\in\mathbb{S}^{d-1}} \|\mathbf{x} - \hat{\mathbf{x}}\|_2 = \Pi_{\mathbb{S}^{d-1}}(\hat{\mathbf{x}})$.

The results so far have only been concerned with the deterministic group restricted isometry property of the measurement matrix $\mathbf{A}$. As is by now a well-established fact, verifying whether a matrix satisfies the classical RIP and by extension the group-RIP is known to be an NP-hard problem [**TP14**]. This unfortunate fact quickly led researchers to consider certain classes of random measurement ensembles for which the restricted isometry property can be shown to hold with overwhelmingly high probability. Of particular importance in this context are measurement matrices whose rows are drawn independently from isotropic subgaussian distributions, which include any discrete or bounded distribution such as the Bernoulli or uniform distribution, as well as Gaussian and Steinhaus[2] distributions. Random matrices $\mathbf{A} \in \mathbb{R}^{m\times d}$ formed from such random ensembles require $m = \Omega(s\log(d/s))$ measurements to satisfy the classical RIP (cf. Definition 2.3) characterized by

$$(1-\delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta)\|\mathbf{x}\|_2^2 \quad \forall \mathbf{x} \in \Sigma_s \tag{4.10}$$

for $\delta \in (0,1)$ with high probability. As alluded to before, however, the same conclusion does not hold if we ask instead for restricted isometric embeddings of $\ell_1^d := (\mathbb{R}^d, \|\cdot\|_1)$ into the space ℓ_2^m via the (ℓ_2,ℓ_1)-RIP defined by

$$(1-\delta)\|\mathbf{x}\|_1 \leq \|\mathbf{A}\mathbf{x}\|_2 \leq (1+\delta)\|\mathbf{x}\|_1 \quad \forall \mathbf{x} \in \Sigma_s. \tag{4.11}$$

This is rooted in the fact that irregardless of $\mathbf{A}$ satisfying the (ℓ_2,ℓ_1)-RIP (or more generally the (ℓ_2,ℓ_1)-group-RIP), one can easily construct distinct sparse or group-sparse vectors that are mapped to the same quantization cell. A classical adversarial example in this context are the 2-sparse vectors $\mathbf{x} = (1 \quad 1/2 \quad 0 \quad \dots \quad 0)^\top$ and $\mathbf{x}' = (1 \quad -1/2 \quad 0 \quad \dots \quad 0)^\top$. If the rows of $\mathbf{A}$ are Bernoulli random vectors $\mathbf{a}_i$, then both vectors produce the measurements $y_i = \langle \mathbf{a}_i, \mathbf{x}\rangle = \operatorname{sgn}(a_{i1} + a_{i2}/2) = \operatorname{sgn}(a_{i1}) = \operatorname{sgn}(a_{i1} - a_{i2}/2) = \langle \mathbf{a}_i, \mathbf{x}'\rangle$, independent of the realization of $\mathbf{A}$. This simple counterexample demonstrates why probabilistic results of the form (4.11) are unattainable for general subgaussian ensembles. Similar counterexamples can be constructed for measurement matrices constructed from subsampled basis functions of bounded orthonormal systems. We want to point out, however, that under additional (local) requirements on vectors $\mathring{\mathbf{x}} \in \Sigma_s \cap \mathbb{S}^{d-1}$ which prevent $\mathring{\mathbf{x}}$ from being too sparse, recovery of $\mathring{\mathbf{x}}$ from subgaussian measurements can still be achieved by an alternative convex programming approach we will discuss in Section 4.3.2. In particular, it was shown in [**AFN12**] that in this case the reconstruction error $\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2^2$ is additively bounded by $\|\mathring{\mathbf{x}}\|_\infty^{1/2}$ such that accurate recovery of sparse vectors from 1-bit observations via subgaussian measurement matrices is still possible if $\|\mathring{\mathbf{x}}\|_\infty \ll 1$.

For the reasons outlined above, we limit our discussion to Gaussian measurement matrices in this chapter and show that measurement matrices of the form $\mathbf{A} = \boldsymbol{\Phi}\boldsymbol{\Psi}$ with $\varphi_{ij} \sim_{\text{i.i.d.}} \mathsf{N}(0,1)$ and $\boldsymbol{\Psi} \in \mathrm{O}(d)$ satisfy the (ℓ_2,ℓ_1) group restricted isometry property both w. r. t. group-sparse and effectively group-sparse vectors. We will use the following result due to Plan and Vershynin, which is a simplified version of Lemma 2.1 in [**PV14**], to assert the (effective) group-RIP of $\mathbf{A}$. Equipped with this result, establishing the group-RIP of a Gaussian random matrix amounts to estimating the mean width w (Definition A.20) of the respective low-complexity signal set.

[2] A Steinhaus random variable is a complex random variable which is uniformly distributed on the complex unit circle.

Lemma 4.8. *Let $\mathbf{A} \in \mathbb{R}^{m\times d}$ be a standard Gaussian random matrix, and let $\mathcal{K} \subset \mathbb{S}^{d-1}$. Fix $\delta \in (0,1)$. Then it holds with probability at least $1-\eta$ that*

$$\sup_{\mathbf{x}\in\mathcal{K}} \left| \frac{1}{m}\sqrt{\frac{\pi}{2}}\|\mathbf{A}\mathbf{x}\|_1 - 1 \right| \leq \delta,$$

provided that

$$m \gtrsim \delta^{-2}\Big[w(\mathcal{K})^2 + \log(\eta^{-1})\Big].$$

To see how this result implies the (effective) (ℓ_2,ℓ_1)-group-RIP, consider an arbitrary set $\mathcal{U} \subset \mathbb{R}^d$. Then the condition

$$(1-\delta)\|\mathbf{x}\|_2 \leq \left\| \frac{1}{m}\sqrt{\frac{\pi}{2}}\mathbf{A}\mathbf{x} \right\|_1 \leq (1+\delta)\|\mathbf{x}\|_2 \quad \forall \mathbf{x} \in \mathcal{U}$$

is equivalent to

$$(1-\delta) \leq \frac{1}{m}\sqrt{\frac{\pi}{2}}\|\mathbf{A}\tilde{\mathbf{x}}\|_1 \leq (1+\delta) \quad \forall \tilde{\mathbf{x}} \in \mathcal{U} \cap \mathbb{S}^{d-1},$$

which in turn can be written as

$$\left| \frac{1}{m}\sqrt{\frac{\pi}{2}}\|\mathbf{A}\tilde{\mathbf{x}}\|_1 - 1 \right| \leq \delta \quad \forall \tilde{\mathbf{x}} \in \mathcal{U} \cap \mathbb{S}^{d-1}.$$

Recall that the mean width of a set $\mathcal{U} \subset \mathbb{R}^d$ is defined as

$$w(\mathcal{U}) = \mathbb{E}_{\mathbf{g}} \sup_{\mathbf{x}\in\mathcal{U}} \langle \mathbf{g}, \mathbf{x} \rangle \tag{4.12}$$

where $\mathbf{g} \in \mathbb{R}^d$ denotes a standard Gaussian random vector. Note that by Proposition A.11, this definition immediately implies that w is invariant under application of the convex hull such that

$$w(\mathcal{U}) = w(\operatorname{conv}(\mathcal{U})). \tag{4.13}$$

In order to estimate the mean width of the sets $\widetilde{\Sigma}_{\mathcal{I},s}$ and $\widetilde{\mathcal{E}}_{\mathcal{I},s}$, respectively, we will make use of the following connection between the two sets, which extends an earlier result due to Plan and Vershynin to the group-sparse setting.

Lemma 4.9. *It holds that*

$$\operatorname{conv}(\widetilde{\Sigma}_{\mathcal{I},s}) \subset \widetilde{\mathcal{E}}_{\mathcal{I},s} \subset 2\operatorname{conv}(\widetilde{\Sigma}_{\mathcal{I},s}).$$

Proof. The proof follows the example of the proof of Lemma 2.1 in [PV13a]. For the first inclusion, let $\mathbf{x} \in \operatorname{conv}(\widetilde{\Sigma}_{\mathcal{I},s})$, and write

$$\mathbf{x} = \sum_{i=1}^k \lambda_i \mathbf{x}_i \quad \text{with} \quad k \in \mathbb{N}, \lambda_i \geq 0, \sum_{i=1}^k \lambda_i = 1, \mathbf{x}_i \in \widetilde{\Sigma}_{\mathcal{I},s}.$$

Then

$$\|\mathbf{x}\|_{\mathcal{I},1} = \left\| \sum_{i=1}^k \lambda_i \mathbf{x}_i \right\|_{\mathcal{I},1} \leq \sum_{i=1}^k \lambda_i \|\mathbf{x}_i\|_{\mathcal{I},1} \leq \sqrt{s} \sum_{i=1}^k \lambda_i = \sqrt{s}$$

by Cauchy-Schwarz.

To establish the second inclusion, one reuses the technique already employed in the proof of Lemma 4.4. Let $\mathbf{x} \in \widetilde{\mathcal{E}}_{\mathcal{I},s}$, and denote again by $\mathcal{T}_1 \subset \mathcal{I}$ the set of s groups of largest ℓ_2-norm of $\mathbf{x}$, $\mathcal{T}_2$ the next s largest ℓ_2-norm groups and so on. Moreover, we abuse notation and write $\mathbf{x}_{\mathcal{T}_i} := \mathbf{x}_{\cup_{S\in\mathcal{T}_i} S} \in \mathbb{R}^d$. With these definitions in place, we decompose $\mathbf{x} \in \widetilde{\mathcal{E}}_{\mathcal{I},s}$ as

$$\mathbf{x} = \sum_{i\geq 1} \mathbf{x}_{\mathcal{T}_i} = \sum_{i\geq 1} \|\mathbf{x}_{\mathcal{T}_i}\|_2 \underbrace{\frac{\mathbf{x}_{\mathcal{T}_i}}{\|\mathbf{x}_{\mathcal{T}_i}\|_2}}_{\in \widetilde{\Sigma}_{\mathcal{I},s}}$$

where by convention we terminate the summation at the first index $i \geq i'$ with $\mathbf{x}_{\mathcal{T}_{i'}} = \mathbf{0}$. To complete the proof, it now remains to show that

$$\sum_{i\geq 1} \|\mathbf{x}_{\mathcal{T}_i}\|_2 \leq 2.$$

To that end, first note that

$$\sum_{i\geq 1} \|\mathbf{x}_{\mathcal{T}_i}\|_2 = \|\mathbf{x}_{\mathcal{T}_1}\|_2 + \sum_{i\geq 2} \|\mathbf{x}_{\mathcal{T}_i}\|_2 \leq 1 + \sum_{i\geq 2} \|\mathbf{x}_{\mathcal{T}_i}\|_2,$$

which immediately follows from $\|\mathbf{x}_{\mathcal{T}_1}\|_2 \leq \|\mathbf{x}\|_2 = 1$ since $\mathbf{x} \in \widetilde{\mathcal{E}}_{\mathcal{I},s} \subset \mathbb{S}^{d-1}$. Next, recall from (4.8) that

$$\|\mathbf{x}_{\mathcal{T}_i}\|_2 \leq \frac{1}{\sqrt{s}} \left\|\mathbf{x}_{\mathcal{T}_{i-1}}\right\|_{\mathcal{I},1}$$

and consequently

$$\sum_{i\geq 1} \|\mathbf{x}_{\mathcal{T}_i}\|_2 \leq 1 + \frac{1}{\sqrt{s}} \sum_{i\geq 2} \left\|\mathbf{x}_{\mathcal{T}_{i-1}}\right\|_{\mathcal{I},1} \leq 1 + \frac{1}{\sqrt{s}} \|\mathbf{x}\|_{\mathcal{I},1} \leq 2$$

since $\mathbf{x} \in \widetilde{\mathcal{E}}_{\mathcal{I},s}$ and therefore $\|\mathbf{x}\|_{\mathcal{I},1} \leq \sqrt{s}$. The claim follows. □

By homogeneity of the mean width in combination with (4.13), this result immediately implies that

$$w(\widetilde{\mathcal{E}}_{\mathcal{I},s}) \leq w(2\operatorname{conv}(\widetilde{\Sigma}_{\mathcal{I},s})) = 2w(\widetilde{\Sigma}_{\mathcal{I},s}). \tag{4.14}$$

The scaling requirements on m for a standard Gaussian random matrix $\mathbf{A} \in \mathbb{R}^{m\times d}$ to satisfy, with high probability, either the genuine or effective group-RIP will consequently both be determined by the mean width of $\widetilde{\Sigma}_{\mathcal{I},s}$. We establish a bound on this crucial quantity in the following result.

Lemma 4.10 (Mean width of group-sparse vectors)**.** *It holds that*

$$w(\widetilde{\Sigma}_{\mathcal{I},s}) \leq \sqrt{2s\log(2eG/s)} + \sqrt{sg}.$$

Proof. First, note that

$$w(\widetilde{\Sigma}_{\mathcal{I},s}) = \mathbb{E}_{\mathbf{g}} \sup_{\mathbf{x}\in\widetilde{\Sigma}_{\mathcal{I},s}} \langle \mathbf{x}, \mathbf{g}\rangle = \mathbb{E} \max_{\mathcal{T}} \sup_{\mathbf{x}\in\mathbb{S}^{d-1}_{\mathcal{T}}} \langle \mathbf{x}, \mathbf{g}\rangle = \mathbb{E} \max_{\mathcal{T}} \|\mathbf{g}_{\mathcal{T}}\|_2$$
$$\leq \max_{\mathcal{T}} \mathbb{E}\|\mathbf{g}_{\mathcal{T}}\|_2 + \mathbb{E} \max_{\mathcal{T}} \big|\|\mathbf{g}_{\mathcal{T}}\|_2 - \mathbb{E}\|\mathbf{g}_{\mathcal{T}}\|_2\big|$$

where the subsequent maxima are taken over all possible group index sets $\mathcal{T} \subset \mathcal{I}$ with $|\mathcal{T}| = s$. For the first term, we have by Jensen's inequality that $\max_{\mathcal{T}} \mathbb{E}\|\mathbf{g}_{\mathcal{T}}\|_2 \leq \sqrt{sg}$. For the second term note that $\|\cdot\|_2$ is 1-Lipschitz continuous by definition. The Gaussian concentration inequality therefore shows that the centered random variable

$$X_{\mathcal{T}} := \|\mathbf{g}_{\mathcal{T}}\|_2 - \mathbb{E}\|\mathbf{g}_{\mathcal{T}}\|_2$$

is subgaussian since by Theorem A.5, we have for $\lambda \in \mathbb{R}$ that

$$\mathbb{E}\exp(\lambda X_{\mathcal{T}}) \leq \exp\left(\frac{1}{2}\lambda^2\right).$$

By a common bound on the expected maximum of a sequence of independent subgaussian random variables (Proposition A.7), this implies

$$\mathbb{E}\max_{\mathcal{T}} |X_{\mathcal{T}}| = \mathbb{E}\max_{\mathcal{T}} \big|\|\mathbf{g}_{\mathcal{T}}\| - \mathbb{E}\|\mathbf{g}_{\mathcal{T}}\|_2\big| \leq \sqrt{2\log\left(2\binom{G}{s}\right)} \leq \sqrt{2s\log\left(\frac{2eG}{s}\right)},$$

which yields the announced result. □

The following result which establishes the group-RIP for standard Gaussian random matrices now follows from Lemma 4.8 and Lemma 4.10, as well as rotation invariance of the Gaussian distribution.

Lemma 4.11. *Let $\mathbf{A} = \mathbf{\Phi}\mathbf{\Psi} \in \mathbb{R}^{m\times d}$ with $\mathbf{\Phi} \in \mathbb{R}^{m\times d}$ denoting a standard Gaussian random matrix and $\mathbf{\Psi} \in \mathrm{O}(d)$. Then with probability at least $1-\eta$, the scaled matrix $m^{-1}\sqrt{\pi/2}\mathbf{A}$ satisfies the (ℓ_2,ℓ_1)-group-RIP with constant $\delta_s \leq \delta$, provided that*

$$m \gtrsim \delta^{-2}\big[s\log(G/s) + sg + \log(\eta^{-1})\big].$$

Remark 4.12. *The dependence of m on the parameters s, g and G matches recent results in the theory of Gelfand numbers, which are commonly used in the theory of compressed sensing to establish lower bounds on the number of measurements required among arbitrary encoder-decoder pairs to guarantee stable recovery of sparse vectors. More concretely, Dirksen and Ullrich establish in [**DU18**] that $m = \Omega(s\log(G/s) + sg)$ measurements are necessary to guarantee stable recovery of block-sparse vectors for any linear measurement and (generally nonlinear) recovery map. Since the (ℓ_2,ℓ_1)-RIP (which is a special case of the (ℓ_2,ℓ_1)-group-RIP) can be used to establish such stable recovery guarantees in the canonical CS theory (see for instance [**JHF11**, Theorem 2]), the bound in Lemma 4.11 is optimal in terms of s, g and G.*

We will also frequently require the (ℓ_2,ℓ_1) group restricted isometry property for matrices acting on effectively group-sparse vectors. As pointed out above, barring a transformation of the implicit constant hidden in the notation, the following result establishes the group-RIP for $\mathcal{E}_{\mathcal{I},s}$ with the same scaling requirement on m as in Lemma 4.11.

Lemma 4.13. *Let* $\mathbf{A} = \mathbf{\Phi}\mathbf{\Psi} \in \mathbb{R}^{m\times d}$ *with* $\mathbf{\Phi}$ *and* $\mathbf{\Psi}$ *as in Lemma 4.11. Then with probability at least* $1-\eta$*, the scaled matrix* $m^{-1}\sqrt{\pi/2}\mathbf{A}$ *satisfies the effective* (ℓ_2,ℓ_1)*-group-RIP with constant* $\delta_s \leq \delta$*, provided that*

$$m \gtrsim \delta^{-2}\Big[s\log(G/s) + sg + \log(\eta^{-1})\Big].$$

At this point, we are prepared to establish the following probabilistic recovery guarantee characterizing the performance of Problem ($\mathrm{P}_{4.2}$) in the Gaussian setting.

Theorem 4.14. *Let* $\mathbf{A} \in \mathbb{R}^{m\times d}$ *be a standard Gaussian random matrix. Fix a value* $\varepsilon \leq 8/\sqrt{5}$*, and assume*

$$m \gtrsim \varepsilon^{-4}\Big[s\log(G/s) + sg + \log(\eta^{-1})\Big].$$

Then with probability at least $1-\eta$*, the following event occurs: every vector* $\mathring{\mathbf{x}} \in \widetilde{\mathcal{E}}_{\mathcal{I},s}$ *can be approximated from its measurements* $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}})$ *by the normalized minimizer* $\hat{\mathbf{z}} = \hat{\mathbf{x}}/\|\hat{\mathbf{x}}\|_2$ *of*

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \|\mathbf{x}\|_{\mathcal{I},1} \\ \text{s.t.} & \mathbf{y}\circ\mathbf{A}\mathbf{x} \geq \mathbf{0} \\ & \langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle = 1 \end{array} \tag{$\mathrm{P}_{4.3}$}$$

with

$$\|\mathring{\mathbf{x}} - \hat{\mathbf{z}}\|_2 \leq \varepsilon.$$

Proof. We invoke Lemma 4.6 with $\delta_t = \varepsilon^2/64$, which implies for $\delta_t \leq 1/5$ (and thus $\varepsilon \leq 8/\sqrt{5}$) that every minimizer $\hat{\mathbf{x}}$ of Problem ($\mathrm{P}_{4.3}$) satisfies

$$\left\|\mathring{\mathbf{x}} - \frac{\hat{\mathbf{x}}}{\|\hat{\mathbf{x}}\|_2}\right\|_2 \leq 8\sqrt{\delta_t} = \varepsilon.$$

For the matrix $\mathbf{A}$ to satisfy the group-RIP on $\widetilde{\mathcal{E}}_{\mathcal{I},t}$ as required by Lemma 4.6, it suffices to choose

$$m \gtrsim \varepsilon^{-4}\Big[t\log(G/t) + tg + \log(\eta^{-1})\Big]$$

with $t = 4s((1+\delta_t)/(1-\delta_t))^2 = 4s((64+\varepsilon^2)/(64-\varepsilon^2))^2 \leq 9s$ according to Theorem 4.14. Absorbing the constants in the notation therefore yields the claim. □

The dependence of m on δ_t^{-2} to satisfy the group-RIP implies that the optimal error decay rate of $\varepsilon = \mathcal{O}(m^{-1})$ established in [**Jac+13**] can only be obtained from Lemma 4.6 if the error could be improved from $\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \lesssim \sqrt{\delta_t}$ to $\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \lesssim \delta_t^2$. Whether this is possible remains an open problem. However, the dependence of m on ε in Theorem 4.14 can be improved by appealing to a particular result on random hyperplane tessellations from [**BL15**]. To make this precise, we introduce the following property of a measurement matrix acting on a subset of the unit Euclidean sphere.

Definition 4.15 (Hyperplane tessellation)**.** *A matrix* $\mathbf{A} \in \mathbb{R}^{m\times d}$ *is said to induce an* ε*-tessellation on a set* $\mathcal{K} \subset \mathbb{S}^{d-1}$ *if*

$$\forall \mathbf{x},\mathbf{z} \in \mathcal{K} \text{ with } \operatorname{sgn}(\mathbf{A}\mathbf{x}) = \operatorname{sgn}(\mathbf{A}\mathbf{z}) : \|\mathbf{x}-\mathbf{z}\|_2 \leq \varepsilon.$$

This property guarantees that $\mathbf{A}$ uniformly tessellates subsets of the unit Euclidean sphere in the sense that any pair of vectors $\mathbf{x}, \mathbf{z} \in \mathcal{K}$ with identical measurements under the map $\mathbf{u} \mapsto \operatorname{sgn}(\mathbf{A}\mathbf{u})$ are at most ε apart from one another. Considering that measurement consistency is enforced by the first constraint of Problem ($\mathrm{P}_{4.3}$), the definition of ε-tessellations immediately yields a recovery guarantee for Problem ($\mathrm{P}_{4.3}$). More precisely, if $\mathbf{A}$ induces an ε-tessellation on $\widetilde{\mathcal{E}}_{\mathcal{I},s}$, then we have that $\mathring{\mathbf{x}} \in \widetilde{\mathcal{E}}_{\mathcal{I},s}$ and ℓ_2-normalized minimizers $\hat{\mathbf{z}}$ of Problem ($\mathrm{P}_{4.3}$) are at most ε apart (in the Euclidean sense), provided that $\hat{\mathbf{z}}$ is effectively group-sparse. Since Lemma 4.4 establishes the effective group-sparsity of minimizers of Problem ($\mathrm{P}_{4.3}$) under the group-RIP, one immediately obtains a uniform recovery guarantee of the sets $\widetilde{\Sigma}_{\mathcal{I},s}$ and $\widetilde{\mathcal{E}}_{\mathcal{I},s}$, respectively, under the assumption that $\mathbf{A}$ additionally induces an ε-tessellation on $\widetilde{\Sigma}_{\mathcal{I},s}$ or $\widetilde{\mathcal{E}}_{\mathcal{I},s}$. We record the result in the following lemma.

Lemma 4.16. *Let $\mathring{\mathbf{x}} \in \widetilde{\mathcal{E}}_{\mathcal{I},s}$. Fix $\delta_t \in (0,1)$, and assume $\mathbf{A} \in \mathbb{R}^{m\times d}$ satisfies the (ℓ_2,ℓ_1)-group-RIP on $\widetilde{\mathcal{E}}_{\mathcal{I},t}$ with $t = 4s(1+\delta_t)^2/(1-\delta_t)^2$. Moreover, assume that $\mathbf{A}$ induces an ε-tessellation on $\widetilde{\mathcal{E}}_{\mathcal{I},t}$. Then every minimizer $\hat{\mathbf{x}}$ of Problem ($\mathrm{P}_{4.3}$) satisfies*

$$\left\| \mathring{\mathbf{x}} - \frac{\hat{\mathbf{x}}}{\|\hat{\mathbf{x}}\|_2} \right\|_2 \leq \varepsilon.$$

While testing whether a matrix induces an ε-tessellation on a particular set remains an open problem, tessellation matrices exist in abundance if we pass to random measurement ensembles. These results are fully determined by geometric summary parameters of a signal set $\mathcal{K}$ such as the Gaussian mean width $w(\mathcal{K})$ or the covering number $\mathfrak{N}(\mathcal{K}, \|\cdot\|, \varepsilon)$ (Definition A.15). The first instance of such a probabilistic guarantee goes back to the work of Plan and Vershynin in [**PV14**]. In particular, they demonstrate for $\mathcal{K} \subset \mathbb{S}^{d-1}$ that $m \gtrsim \varepsilon^{-6} w(\mathcal{K})^2$ measurements suffice for a standard Gaussian random matrix $\mathbf{A} \in \mathbb{R}^{m\times d}$ to induce an ε-tessellation on $\mathcal{K}$ with probability at least $1 - 2\exp(-c\varepsilon^2 m)$. In this work, we appeal to the following tightened result due to Bilyk and Lacey, which improves the dependence of m on ε and instead requires an estimate of the covering number of $\mathcal{K}$ rather than its mean width.

Theorem 4.17 ([**BL15**, Theorem 1.5])**.** *Let $\mathbf{A} \in \mathbb{R}^{m\times d}$ be a standard Gaussian random matrix. Then there exist constants $0 < c < 1 < C$ such that with probability at least $1 - (2\mathfrak{N}(\mathcal{K}, \|\cdot\|_2, c\varepsilon))^{-2}$, the matrix $\mathbf{A}$ induces an ε-tessellation on the set $\mathcal{K} \subset \mathbb{S}^{d-1}$, provided that $m \geq C\varepsilon^{-1} \log \mathfrak{N}(\mathcal{K}, \|\cdot\|_2, c\varepsilon)$.*

Estimating the covering number $\mathfrak{N}(\mathcal{K}, u) = \mathfrak{N}(\mathcal{K}, \|\cdot\|_2, u)$ of a set $\mathcal{K}$ usually proceeds in one of three ways or a combination thereof: a volume comparison argument, Maurey's empirical method or (dual) Sudakov minoration (cf. Lemma A.21). Since we are concerned with estimating $\mathfrak{N}(\widetilde{\mathcal{E}}_{\mathcal{I},s}, u)$ in order to establish a tessellation on the set of effectively group-sparse vectors, the first two techniques cannot be applied to our scenario. This is due to the fact that $\widetilde{\mathcal{E}}_{\mathcal{I},s}$ can neither be written as a union of spheres restricted to lower-dimensional coordinate subspaces as required for the volume comparison argument, nor represented as the convex hull of a finite set to apply Maurey's empirical method (see also the discussion in Section 5.3.3). This leaves us with Sudakov's inequality, which relates the covering number of a set to its mean width.

Lemma 4.18. *The covering number of $\widetilde{\mathcal{E}}_{\mathcal{I},s}$ w. r. t. the metric induced by the ℓ_2-norm is bounded by*

$$\begin{aligned}\mathfrak{N}(\widetilde{\mathcal{E}}_{\mathcal{I},s}, \|\cdot\|_2, u) &\lesssim \exp\left(4u^{-2}\left(\sqrt{2s\log(2eG/s)} + \sqrt{sg}\right)^2\right)\\ &\lesssim \exp\left(cu^{-2}(2s\log(2eG/s) + sg)\right).\end{aligned}$$

Proof. Denote by $\mathcal{N} \subset \widetilde{\mathcal{E}}_{\mathcal{I},s}$ the smallest u-net of $\widetilde{\mathcal{E}}_{\mathcal{I},s}$ w. r. t. the Euclidean metric such that $|\mathcal{N}| = \mathfrak{N}(\widetilde{\mathcal{E}}_{\mathcal{I},s}, \|\cdot\|_2, u)$. By Sudakov minoration and (4.14), it then immediately follows that

$$|\mathcal{N}| \lesssim \exp\left[\left(\frac{w(\widetilde{\mathcal{E}}_{\mathcal{I},s})}{u}\right)^2\right] \leq \exp\left[4\left(\frac{\sqrt{2s\log(2eG/s)} + \sqrt{sg}}{u}\right)^2\right].$$

□

We are now ready to state the improved recovery guarantee for Problem ($\mathrm{P}_{4.3}$).

Theorem 4.19. *Fix $\varepsilon \leq 1/3$, and let $\mathbf{A} \in \mathbb{R}^{m\times d}$ be a standard Gaussian random matrix with*

$$m \gtrsim \varepsilon^{-3}(s\log(G/s) + sg).$$

Then with probability at least

$$1 - c\exp\left(-\tilde{c}\left(s\log\left(\frac{G}{16s}\right) + sg\right)\right),$$

the following holds: for all $\mathring{\mathbf{x}} \in \widetilde{\mathcal{E}}_{\mathcal{I},s}$, every normalized minimizer $\hat{\mathbf{z}} = \hat{\mathbf{x}}/\|\hat{\mathbf{x}}\|_2$ of ($\mathrm{P}_{4.3}$) with $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}})$ satisfies $\|\mathring{\mathbf{x}} - \hat{\mathbf{z}}\|_2 \leq \varepsilon$.

Proof. Set $\gamma_\delta := ((1+\delta)/(1-\delta))^2$. According to Lemma 4.16, it suffices to show that for

$$t = 4s\left(\frac{1+\delta}{1-\delta}\right)^2 = 4s\gamma_\delta,$$

the matrix $\mathbf{A}$ both acts as a group-RIP on the set $\widetilde{\mathcal{E}}_{\mathcal{I},t}$ and induces an ε-tessellation on $\widetilde{\mathcal{E}}_{\mathcal{I},t}$ with the announced probability. For the former, we have by Lemma 4.13 that

$$\begin{aligned}m &\gtrsim \delta^{-2}\left(t\log\left(\frac{G}{t}\right) + tg\right)\\ &= \delta^{-2}4\gamma_\delta\left(s\log\left(\frac{G}{4s\gamma_\delta}\right) + sg\right)\end{aligned}$$

measurements suffice for $\mathbf{A}$ to satisfy the group-RIP on $\widetilde{\mathcal{E}}_{\mathcal{I},t}$ with failure probability at most

$$\exp\left(-4\gamma_\delta\left(s\log\left(\frac{G}{4s\gamma_\delta}\right) + sg\right)\right).$$

On the other hand, Theorem 4.17 establishes that $m \gtrsim \varepsilon^{-3} w(\widetilde{\mathcal{E}}_{\mathcal{I},t})^2$ measurements suffice for $\mathbf{A}$ to induce an ε-tessellation on $\widetilde{\mathcal{E}}_{\mathcal{I},t}$ with failure probability at most

$$c_0 \exp(-c_1 \varepsilon^{-2} w(\widetilde{\mathcal{E}}_{\mathcal{I},t})^2).$$

By a union bound over both events, this implies that for

$$\begin{aligned} m &\gtrsim \max\left\{\delta^{-2}\gamma_\delta, \varepsilon^{-3}\gamma_\varepsilon\right\}\left(s\log\left(\frac{G}{s}\right) + sg\right) \\ &\gtrsim \max\left\{\delta^{-2}\gamma_\delta\left(s\log\left(\frac{G}{4s\gamma_\delta}\right) + sg\right), \varepsilon^{-3}\gamma_\varepsilon\left(s\log\left(\frac{G}{4s\gamma_\varepsilon}\right) + sg\right)\right\}, \end{aligned}$$

the matrix $\mathbf{A}$ satisfies the group-RIP and ε-tessellation property with failure probability at most

$$\begin{aligned} &\exp\left(-4\gamma_\delta\left(s\log\left(\frac{G}{4s\gamma_\delta}\right) + sg\right)\right) + c_0 \exp\left(-c_1\varepsilon^{-2}\gamma_\varepsilon\left(s\log\left(\frac{G}{4s\gamma_\varepsilon}\right) + sg\right)\right) \\ &\lesssim c_0' \exp\left(-c_1'\gamma_\delta\left(s\log\left(\frac{G}{4s\gamma_\delta}\right) + sg\right)\right). \end{aligned}$$

With the choice $\delta = \varepsilon$ and the required restriction that $\varepsilon \leq 1/3$, we have that $1 \leq \gamma_\varepsilon \leq 4$ due to monotonicity of γ_ε as a function of ε. It therefore follows that the conclusion of Theorem 4.19 holds with failure probability at most

$$c_0' \exp\left(-c_1'\gamma_\varepsilon\left(s\log\left(\frac{G}{4s\gamma_\varepsilon}\right) + sg\right)\right) \leq c_0' \exp\left(-c_1'\left(s\log\left(\frac{G}{16s}\right) + sg\right)\right),$$

provided that

$$\begin{aligned} m &\gtrsim 4\varepsilon^{-3}\left(s\log\left(\frac{G}{s}\right) + sg\right) \\ &\gtrsim \varepsilon^{-3}\gamma_\varepsilon\left(s\log\left(\frac{G}{s}\right) + sg\right) \\ &\gtrsim \gamma_\varepsilon \max\left\{\varepsilon^{-2}, \varepsilon^{-3}\right\}\left(s\log\left(\frac{G}{s}\right) + sg\right). \end{aligned}$$

This concludes the proof. □

Remark 4.20. *(i) In order for the failure probability in the proof of Theorem 4.19 to yield useful values, it is necessary for $g - \log(4s\gamma_\varepsilon/G)$ to remain positive. This implies the very mild condition*

$$\varepsilon \leq 1 - 2\left(\exp\left(\frac{1}{2}\left[\log\left(\frac{G}{4s}\right) + g\right]\right) + 1\right)^{-1}.$$

(ii) The failure probability in Theorem 4.19 does not *directly depend on the recovery quality ε. This slightly counterintuitive behavior is rooted in the result by Bilyk and Lacey from [**BL15**], which suffers from the same drawback. Note, however, that this is not too different from the previous group-RIP result established in Theorem 4.14, which, for the choice $\eta = \exp(-[s\log(G/s) + sg])$, implies that $\Omega(\varepsilon^{-4}[s\log(G/s) + sg])$ measurements suffice for ε-accurate recovery of effectively group-sparse vectors with probability at least $1 - \exp(-[s\log(G/s) + sg])$. In that sense, Theorem 4.19 therefore improves upon the previous group-RIP-based result by improving the dependence of m on ε.*

The results discussed up until this point have one significant drawback in that they do not establish the behavior of $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ in the presence of noise in the observations. In fact, it remains an open problem how to modify the recovery technique defined by Problem ($\mathrm{P}_{4.3}$) in order to harden it against additive pre-quantization and adversarial post-quantization noise. As we will discuss in the next section, the situation can be remedied by reversing, as it were, the roles of the objective function and the constraint set of Problem ($\mathrm{P}_{4.3}$). More precisely, the recovery strategy discussed in the sequel will enforce the desired group-sparse signal structure via an appropriate choice of the feasible set, while promoting quantization consistency in the objective function.

4.3.2 Correlation Maximization

As pointed out in Section 2.3.1, the concept of quantization or measurement consistency is of key importance in 1-bit compressed sensing as it guarantees that target vectors and their estimates which fall into the same quantization cell are also close in the Euclidean sense. However, recovery schemes which enforce quantization-consistent solutions come with a fundamental drawback. If instead of the measurements $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}) \in \{\pm 1\}^m$ one observes the corrupted vector $\tilde{\mathbf{y}}$ with $\Delta_{\mathrm{H}}(\mathbf{y}, \tilde{\mathbf{y}}) > 0$, the effects of bit errors on the reconstruction quality when using $\tilde{\mathbf{y}}$ in place of $\mathbf{y}$ can be detrimental. Recall from the definition of ε-tessellations, that a matrix $\mathbf{A}$ inducing an ε-tessellation on a set $\mathcal{K} \subset \mathbb{S}^{d-1}$ ensures that no quantization cell has Euclidean diameter exceeding ε. This in turn implies that in the worst case even a single erroneous bit in the measurements $\tilde{\mathbf{y}}$ could result in a recovery error exceeding ε. As a natural remedy to circumvent issues concerning noise robustness, a few recovery strategies have been proposed over the years which aim to minimize an appropriate data fidelity measure, subject to set membership constraints which enforce the assumed signal structure (see, *e.g.*, [**PV13b**; **PV16**; **PVY17**]). One approach of particular importance proposed in this context aims at minimizing the Hamming distance $\Delta_{\mathrm{H}}(\tilde{\mathbf{y}}, \operatorname{sgn}(\mathbf{A}\mathbf{x}))$ between the (possibly noise corrupted) measurements $\tilde{\mathbf{y}}$ and $\operatorname{sgn}(\mathbf{A}\mathbf{x})$ for some estimate $\mathbf{x}$ of $\mathring{\mathbf{x}}$, subject to $\mathbf{x} \in \mathcal{K} \subset \mathbb{S}^{d-1}$. Unfortunately, due to the nonlinear and nonconvex nature of the Hamming metric Δ_{H}, this naive recovery approach is intractable. To relax the problem into a tractable form, note that the normalized Hamming distance may also be expressed as

$$\Delta_{\mathrm{H}}(\tilde{\mathbf{y}}, \operatorname{sgn}(\mathbf{A}\mathbf{x})) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{\tilde{y}_i \neq \operatorname{sgn}(\langle \mathbf{a}_i, \mathbf{x}\rangle)\}} = \frac{1}{2m}\sum_{i=1}^{m}(1 - \tilde{y}_i \operatorname{sgn}(\langle \mathbf{a}_i, \mathbf{x}\rangle)).$$

Given the last identity, a natural way to convexify this function is by replacing the sgn-operator by the identity map since sign violations will make $-\tilde{y}_i\langle \mathbf{a}_i, \mathbf{x}\rangle$ large and therefore penalize solutions which disagree with the sign measurements $\tilde{\mathbf{y}}$ too much. Dropping any remaining constant terms from the penalty function obtained by this convexification and imposing that $\mathbf{x}$ belong to some structure-promoting set $\mathcal{K} \subset \mathbb{S}^{d-1}$, one arrives at the problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & \langle \tilde{\mathbf{y}}, \mathbf{A}\mathbf{x}\rangle \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{K} \end{aligned} \tag{$\mathrm{P}_{4.4}$}$$

where we also expressed the problem in a more natural formulation as a maximization problem. In a sense, this particular objective function measures the correlation between

the quantized measurements $\tilde{\mathbf{y}}$ and the linear observations $\mathbf{Ax}$. Since the formulation does not enforce strict adherence to the sign information postulated by $\tilde{\mathbf{y}}$, the program instead seeks for a vector in $\mathcal{K}$ which explains the data with as few sign violations as possible. Rather than explicitly incorporating a mechanism to correct possible bit errors as for instance pursued in the AOP algorithm [**YYO12**], Problem ($\mathrm{P}_{4.4}$) instead employs a type of inherent majority voting strategy in which less importance is placed on sign observations which go against the majority of other consistent measurements.

Problem ($\mathrm{P}_{4.4}$) was first proposed by Plan and Vershynin in [**PV13b**] and at the time was the first recovery scheme with provable robustness to both pre-quantization and adversarial post-quantization noise in the acquisition system. In particular, the authors consider arbitrary nonlinear observations of the form

$$y_i = Q(\langle \mathbf{a}_i, \mathbf{x} \rangle)$$

for independent standard Gaussian random vectors $\mathbf{a}_i$ and require that, conditioned on $\mathbf{a}_i$, the map Q satisfies the condition

$$\mathbb{E}_Q y_i = \Theta(\langle \mathbf{a}_i, \mathbf{x} \rangle)$$

for some function $\Theta \colon \mathbb{R} \to [-1, 1]$ where $\mathbb{E}_Q$ denotes expectation w. r. t. any random elements of Q independent from the ensemble $\{\mathbf{a}_i\}_{i=1}^m$. This general formulation includes a variety of interesting special cases, including most prominently the 1-bit observation model, as well as pre- and post-quantization noise, and the logistic regression model where Q has the form

$$Q(t) = \frac{1}{1 + e^{-t}}.$$

Given a standard Gaussian random variable g, the recovery performance of each of these models depends on the constant

$$\lambda := \mathbb{E}[g\Theta(g)],$$

which, roughly speaking, measures the average correlation between the linear and nonlinear observations. For instance, if Q is a function taking a constant value in $[-1, 1]$, then $Q(\langle \mathbf{a}_i, \mathbf{x} \rangle)$ carries no information about the linear observation $\langle \mathbf{a}_i, \mathbf{x} \rangle$, which is reflected in the fact that $\mathbb{E}[gQ(g)] = 0$.

Adopted to the setting of group-sparse recovery, we suggest to solve the program

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & \langle \mathbf{y}, \mathbf{Ax} \rangle \\ \text{s.t.} \quad & \|\mathbf{x}\|_{\mathcal{I},1} \leq \sqrt{s} \\ & \|\mathbf{x}\|_2 \leq 1, \end{aligned} \tag{$\mathrm{P}_{4.5}$}$$

which now maximizes the correlation between quantized and unquantized observations over the set of effectively s-group-sparse vectors. For simplicity of notation, we will sometimes express solutions of Problem ($\mathrm{P}_{4.5}$) using the recovery map $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}} \colon \{\pm 1\}^m \to \mathbb{R}^d$ which maps quantized vectors $\mathbf{y}$ to maximizers $\hat{\mathbf{x}}$ of Problem ($\mathrm{P}_{4.5}$). Note that the program above actually considers the bigger set $\sqrt{s}\mathbb{B}_{\mathcal{I},1}^d \cap \mathbb{B}_2^d$ rather than the set of effectively s-group-sparse vectors inside the unit ℓ_2-ball. However, since Problem ($\mathrm{P}_{4.5}$) maximizes a

linear function on a compact set, we may replace the feasible set by its convex hull without changing the solution. In light of the following result which establishes that the feasible set of Problem ($P_{4.5}$) corresponds to the convex hull of $\widetilde{\mathcal{E}}_{\mathcal{I},s}$, the recovery guarantee we will establish below will therefore also hold true for the bigger set $\sqrt{s}\mathbb{B}_{\mathcal{I},1}^d \cap \mathbb{B}_2^d$.

Proposition 4.21. *The set $\sqrt{s}\mathbb{B}_{\mathcal{I},1}^d \cap \mathbb{B}_2^d$ is the convex hull of the set of effectively group-sparse vectors with unit Euclidean norm.*

Proof. We need to establish that

$$\begin{aligned}\operatorname{conv}(\widetilde{\mathcal{E}}_{\mathcal{I},s}) &= \operatorname{conv}\left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_{\mathcal{I},1} \leq \sqrt{s}\|\mathbf{x}\|_2, \|\mathbf{x}\|_2 \leq 1\right\} \\ &= \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_{\mathcal{I},1} \leq \sqrt{s}, \|\mathbf{x}\|_2 \leq 1\right\} \\ &= \sqrt{s}\mathbb{B}_{\mathcal{I},1}^d \cap \mathbb{B}_2^d \\ &=: U.\end{aligned}$$

Clearly, every vector $\mathbf{x} \in \operatorname{conv}(\widetilde{\mathcal{E}}_{\mathcal{I},s})$ is contained in U by the Cauchy-Schwarz inequality, hence $\operatorname{conv}(\widetilde{\mathcal{E}}_{\mathcal{I},s}) \subseteq U$. It therefore remains to show the reverse inclusion $U \subseteq \operatorname{conv}(\widetilde{\mathcal{E}}_{\mathcal{I},s})$ to establish equality of both sets. To that end, it is enough to consider the extreme points of $\operatorname{conv}(\widetilde{\mathcal{E}}_{\mathcal{I},s})$ and U. This follows because by [**Roc15**, Corollary 18.5.1], every compact convex set $C \subseteq \mathbb{R}^d$ can be expressed as the convex hull of its extreme points. An extreme point of C is a vector $\mathbf{x} \in C$ which is not contained in an open line segment in C. We therefore need to show that the extreme points of U are included in $\operatorname{conv}(\widetilde{\mathcal{E}}_{\mathcal{I},s})$. From the definition of extreme points, it clearly follows that no extreme point of U can belong to the interior of U. This means that we only need to consider vectors $\mathbf{x} \in U$ satisfying either $\|\mathbf{x}\|_{\mathcal{I},1} = \sqrt{s}$ or $\|\mathbf{x}\|_2 = 1$. However, any vector $\mathbf{x} \in U$ with $\|\mathbf{x}\|_2 = 1$ is certainly contained in $\operatorname{conv}(\widetilde{\mathcal{E}}_{\mathcal{I},s})$. To complete the proof, we now claim that no vector $\mathbf{x}$ with $\|\mathbf{x}\|_{\mathcal{I},1} = \sqrt{s}$ and $\|\mathbf{x}\|_2 < 1$ can be an extreme point of U. To see this, note that every vector $\mathbf{x}$ with $\|\mathbf{x}\|_{\mathcal{I},1} = \sqrt{s}$ can be written as a convex combination of G vectors in $\sqrt{s}\mathbb{B}_{\mathcal{I},1}^d$ since

$$\mathbf{x} = \sum_{i=1}^{G} \frac{\|\mathbf{x}_{\mathcal{I}_i}\|_2}{\sqrt{s}} \frac{\sqrt{s}\mathbf{x}_{\mathcal{I}_i}}{\|\mathbf{x}_{\mathcal{I}_i}\|_2}$$

with

$$\sum_{i=1}^{G} \frac{\|\mathbf{x}_{\mathcal{I}_i}\|_2}{\sqrt{s}} = \frac{\|\mathbf{x}\|_{\mathcal{I},1}}{\sqrt{s}} = 1 \quad \text{and} \quad \left\|\frac{\sqrt{s}\mathbf{x}_{\mathcal{I}_i}}{\|\mathbf{x}_{\mathcal{I}_i}\|_2}\right\|_{\mathcal{I},1} = \sqrt{s}.$$

Since extreme points are points which cannot be written as convex combination of two other points of a set, we have that every extreme point of $\sqrt{s}\mathbb{B}_{\mathcal{I},1}^d$ is 1-group-sparse w. r. t. $\mathcal{I}$. In other words, we have $\|\mathbf{x}\|_{\mathcal{I},1} = \|\mathbf{x}\|_2 = \sqrt{s}$. Since this violates our assumption that $\|\mathbf{x}\|_2 < 1$, no point $\mathbf{x} \in U$ with $\|\mathbf{x}\|_{\mathcal{I},1} = \sqrt{s}$ and $\|\mathbf{x}\|_2 < 1$ can be an extreme point. The claim follows. □

The performance of $\widetilde{\Delta}_{\mathcal{I}}^{\text{corr}}$ is again determined by the mean width of the signal set $\sqrt{s}\mathbb{B}_{\mathcal{I},1} \cap \mathbb{B}_2^d$ (which is identical to the mean width of $\widetilde{\mathcal{E}}_{\mathcal{I},s}$ due to invariance of the mean width w. r. t. the convex hull). We will make repeated use of the following result due to Plan and Vershynin, which we state here in a form with simplified failure probability.

Theorem 4.22 ([PV13b, Theorem 1.1]). *Let $\mathring{\mathbf{x}} \in \mathcal{K} \subset \mathbb{S}^{d-1}$. Denote by $\mathbf{A} \in \mathbb{R}^{m\times d}$ a standard Gaussian random matrix. Then with probability at least $1-\eta$, any solution $\hat{\mathbf{x}}$ of Problem ($\mathrm{P}_{4.5}$) with $\mathbf{y} = Q(\mathbf{A}\mathring{\mathbf{x}})$ satisfies $\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \leq \varepsilon$, provided that*

$$m \gtrsim \varepsilon^{-4}\lambda^{-2}\Big[w(\mathcal{K})^2 + \log(\eta^{-1})\Big].$$

We will apply Theorem 4.22 to analyze the performance of Problem ($\mathrm{P}_{4.5}$) subject to the noisy measurement model

$$\mathbf{y} = Q(\mathbf{A}\mathring{\mathbf{x}}) = \mathbf{f} \circ \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\nu}) \tag{4.15}$$

where $\mathbf{f} = (f_i)_{i=1}^m \sim \mathsf{B}_m(p)$ denotes an i.i.d. random Bernoulli vector with $\mathbb{P}(f_i = 1) = p > 1/2$, and $\boldsymbol{\nu} \sim \mathsf{N}(\mathbf{0}, \sigma^2\mathrm{Id}_m)$ denotes an additive pre-quantization noise vector independent from both $\mathbf{f}$ and $\mathbf{A}$. To that end, one first needs to determine the function Θ and the associated correlation parameter λ, which was already done by Plan and Vershynin in [PV13b, Section 3.1] (see also Section 4.4.2). In particular, one easily verifies that

$$\begin{aligned}\mathbb{E}_Q y_i &= \mathbb{E}_{f_i} f_i \mathbb{E}_{\nu_i} \operatorname{sgn}(\langle \mathbf{a}_i, \mathring{\mathbf{x}}\rangle + \nu_i)\\ &= (2p-1)(1 - \mathbb{P}(\nu_i \leq -\langle \mathbf{a}_i, \mathring{\mathbf{x}}\rangle))\\ &=: \Theta(\langle \mathbf{a}_i, \mathring{\mathbf{x}}\rangle),\end{aligned}$$

which therefore yields

$$\lambda = \mathbb{E}[g\Theta(g)] = (2p-1)\sqrt{\frac{2}{\pi(\sigma^2+1)}}.$$

The following result is now a simple consequence of Theorem 4.22, as well as Lemma 4.9 in combination with Lemma 4.10.

Theorem 4.23. *Let $\mathring{\mathbf{x}} \in \widetilde{\mathcal{E}}_{\mathcal{I},s}$, and denote by $\mathbf{A} \in \mathbb{R}^{m\times d}$ a standard Gaussian random matrix. Set $\mathbf{y} = Q(\mathbf{A}\mathring{\mathbf{x}})$ with Q corresponding to the noisy measurement operator (4.15). Then with probability at least $1-\eta$, every normalized minimizer $\hat{\mathbf{z}} = \hat{\mathbf{x}}/\|\hat{\mathbf{x}}\|_2$ of Problem ($\mathrm{P}_{4.5}$) with $\hat{\mathbf{x}} = \widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}(\mathbf{y})$ satisfies $\|\mathring{\mathbf{x}} - \hat{\mathbf{z}}\|_2 \leq \varepsilon$, provided that*

$$m \gtrsim \varepsilon^{-4}(\sigma^2+1)(2p-1)^{-2}\Big(s\log(G/s) + sg + \log(\eta^{-1})\Big).$$

4.3.3 Group Hard Thresholding

In this section, we briefly discuss a simple recovery procedure adopted from [Fou16] for group-sparse signal reconstruction, which does not rely on convex programming. In particular, Foucart establishes that

$$\hat{\mathbf{x}} = \mathcal{H}_s(\mathbf{A}^\top \mathbf{y})$$

is an accurate estimator for genuinely sparse vectors $\mathring{\mathbf{x}} \in \widetilde{\Sigma}_s = \Sigma_s \cap \mathbb{S}^{d-1}$ from their binary measurements $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}})$ conditioned on the mixed (ℓ_2, ℓ_1) restricted isometry property of $\mathbf{A}$ (cf. [Fou16, Theorem 8]). Recall from Section 3.2 that the binary iterative hard thresholding algorithm repeats the update equation

$$\mathbf{x}^{(n+1)} = \mathcal{H}_s(\mathbf{x}^{(n)} - \mathbf{A}^\top(\operatorname{sgn}(\mathbf{A}\mathbf{x}^{(n)}) - \mathbf{y}))$$

until either the Hamming distance between $\mathbf{y}$ and $\operatorname{sgn}(\mathbf{A}\mathbf{x}^{(n+1)})$ vanishes or some predefined iteration count is reached. Applying this procedure to the shifted measurement vector $\bar{\mathbf{y}} := \mathbf{y} + \mathbf{1}$ with starting point $\mathbf{x}^{(0)} = \mathbf{0}$, the first iteration of the algorithm therefore yields

$$\mathbf{x}^+ = \mathcal{H}_s(-\mathbf{A}^\top(\operatorname{sgn}(\mathbf{0}) - \bar{\mathbf{y}})) = \mathcal{H}_s(\mathbf{A}^\top(\bar{\mathbf{y}} - \mathbf{1})) = \mathcal{H}_s(\mathbf{A}^\top \mathbf{y}) \tag{4.16}$$

where we used the fact that $\operatorname{sgn}(\mathbf{0}) = \mathbf{1}$ by convention.

Another interpretation of the recovery procedure can be given from the convex programming formulation considered in the previous section. Assume for the moment that we aim to solve the problem

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{maximize}} & \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle \\ \text{s.t.} & \mathbf{x} \in \mathcal{K} \end{array} \tag{$\mathrm{P}_{4.6}$}$$

for a structure-promoting set $\mathcal{K} \subset \mathbb{R}^d$ which corresponds to the unit ball of an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$. Then the optimal value of Problem ($\mathrm{P}_{4.6}$) coincides with the dual norm $\|\cdot\|_*$ of $\|\cdot\|$ (cf. Definition A.13), *i.e.*, given a maximizer $\hat{\mathbf{x}}$ of Problem ($\mathrm{P}_{4.6}$), we have

$$\langle \mathbf{y}, \mathbf{A}\hat{\mathbf{x}} \rangle = \left\langle \mathbf{A}^\top \mathbf{y}, \hat{\mathbf{x}} \right\rangle = \sup_{\|\mathbf{z}\| \leq 1} \left\langle \mathbf{A}^\top \mathbf{y}, \mathbf{z} \right\rangle = \left\| \mathbf{A}^\top \mathbf{y} \right\|_*.$$

If the target vector $\mathring{\mathbf{x}} \in \mathbb{S}^{d-1}$ we aim to recover is known to be s-sparse, one would ideally choose $\mathcal{K} = \widetilde{\Sigma}_s$ to enforce exact sparsity of solutions of Problem ($\mathrm{P}_{4.6}$). The nonconvexity of $\mathcal{K}$, however, renders the resulting optimization problem intractable. On the other hand, due to the linearity of the objective function and compactness of $\mathcal{K}$, the optimal value of the problem remains unchanged if one replaces the constraint $\mathbf{x} \in \widetilde{\Sigma}_s$ with $\mathbf{x} \in \operatorname{conv}(\widetilde{\Sigma}_s)$. Considering that $\operatorname{conv}(\widetilde{\Sigma}_s)$ is a symmetric convex body[3], a classical result in convex analysis now states that the associated gauge function of $\operatorname{conv}(\widetilde{\Sigma}_s)$ defines a norm on $\mathbb{R}^d$ [**Ver12**, Proposition 2.1]. To determine which norm exactly the Minkowski gauge corresponds to, we first define the set $\mathcal{U}_s := \{I \subset [d] : |I| \leq s\}$ of all subsets of $[d]$ of size at most s. This allows us to rewrite the gauge function as

$$\begin{aligned}
\gamma_{\operatorname{conv}(\widetilde{\Sigma}_s)}(\mathbf{x}) &= \inf\left\{t > 0 : \mathbf{x} \in t \operatorname{conv}(\widetilde{\Sigma}_s)\right\} \\
&= \inf\left\{t > 0 : \mathbf{x} \in t \operatorname{conv}\left(\bigcup_{I \in \mathcal{U}_s} \mathbb{S}_I^{d-1}\right)\right\} \\
&= \inf\left\{t > 0 : \mathbf{x} = t \sum_{I \in \mathcal{U}_s} \lambda_I \mathbf{v}_I, \mathbf{v}_I \in \mathbb{S}_I^{d-1}, \lambda_I \geq 0, \sum_{I \in \mathcal{U}_s} \lambda_I = 1\right\} \\
&= \inf\left\{\sum_{I \in \mathcal{U}_s} \|\mathbf{u}_I\|_2 : \mathbf{x} = \sum_{I \in \mathcal{U}_s} \mathbf{u}_I, \mathbf{u}_I \in \mathbb{R}_I^d\right\} \\
&=: \|\mathbf{x}\|_{(s)}
\end{aligned}$$

with $\mathbb{R}_I^d$ denoting the coordinate subspace of $\mathbb{R}^d$ supported on the index set $I \subset [d]$. The norm $\|\cdot\|_{(s)}$ is also known as the *s-support-norm* [**AFN12**]. Since every convex body is

[3] A *convex body* is a compact convex set $C \subset \mathbb{R}^d$ with non-empty relative interior. If the set C is also origin-symmetric, *i.e.*, $\mathbf{x} \in C$ if and only if $-\mathbf{x} \in C$, then C is called a *symmetric convex body*.

a star domain[4], the gauge function allows one to write the set $\operatorname{conv}(\widetilde{\Sigma}_s)$ in terms of its 1-sublevel set [Roc15]:

$$\begin{aligned}\operatorname{conv}(\widetilde{\Sigma}_s) &= \left\{\mathbf{x} \in \mathbb{R}^d : \gamma_{\operatorname{conv}(\widetilde{\Sigma}_s)}(\mathbf{x}) \leq 1\right\} \\ &= \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_{(s)} \leq 1\right\} \\ &=: \mathbb{B}^d_{(s)}.\end{aligned}$$

This in turn implies that we may replace the set membership constraint $\mathbf{x} \in \operatorname{conv}(\widetilde{\Sigma}_s)$ in Problem ($\mathrm{P}_{4.6}$) by the norm constraint $\|\mathbf{x}\|_{(s)} \leq 1$. Note, however, that an implementation of Problem ($\mathrm{P}_{4.6}$) is still intractable in terms of $\|\cdot\|_{(s)}$ since it requires an enumeration of all $\sum_{i=0}^{s} \binom{d}{i}$ subsets contained in $\mathcal{U}_s$. Fortunately, despite the complicated and intractable formulation of the s-support-norm, its dual norm $\|\cdot\|^*_{(s)}$ admits a simple expression in terms of the best s-term approximation of a vector. Denote by $\check{\mathbf{x}}$ the nonincreasing rearrangement of a vector $\mathbf{x}$ characterized by $\check{x}_1 \geq \check{x}_2 \geq \ldots \geq \check{x}_d$ with $\check{x}_i := |x_{\pi(i)}|$ and $\pi\colon [d] \to [d]$ a permutation. Then the norm $\|\cdot\|^*_{(s)}$ is given by

$$\|\mathbf{x}\|^*_{(s)} = \left(\sum_{i=1}^{s} \check{x}_i^2\right)^{1/2} = \|\mathcal{H}_s(\mathbf{x})\|_2$$

(see [AFN12, Section 2.1]). Equipped with this identity, it follows that Problem ($\mathrm{P}_{4.6}$) with $\mathcal{K} = \widetilde{\Sigma}_s$ admits a closed-form solution as

$$\hat{\mathbf{x}} = \frac{\mathcal{H}_s(\mathbf{A}^\top \mathbf{y})}{\left\|\mathcal{H}_s(\mathbf{A}^\top \mathbf{y})\right\|_2}. \tag{4.17}$$

Up to normalization, this corresponds precisely to the hard thresholding scheme considered in [Fou16]. Normalizing reconstructed vectors to unit norm, however, is natural in this context due to the fact that any norm information is irrevocably lost in the acquisition process. This connection between the hard thresholding approach and Problem ($\mathrm{P}_{4.6}$) was first made in [CB15]. In particular, it suggests that the recovery performance of the estimator (4.17) and solutions to Problem ($\mathrm{P}_{4.6}$) when benchmarked on vectors $\mathring{\mathbf{x}} \in \widetilde{\Sigma}_s$ should be roughly equivalent.

To see why (4.17) holds and in order to extend the idea to the group-sparse setting, we choose $\mathcal{K} = \widetilde{\Sigma}_{\mathcal{I},s}$ in Problem ($\mathrm{P}_{4.6}$). Denoting by

$$\mathcal{G}_s := \{\mathcal{T} \subset \mathcal{I} : |\mathcal{T}| \leq s\}$$

the collection of subpartitions of $\mathcal{I}$ of size at most s, we then find for the dual gauge

[4] A set $S \subset \mathbb{R}^d$ is called a *star domain* (or *star-shaped* or *star-convex*) if there exists a point $\hat{\mathbf{x}} \in S$ such that for every $\mathbf{x} \in S$, the closed line segment $\lambda\hat{\mathbf{x}} + (1-\lambda)\mathbf{x}$ with $\lambda \in [0,1]$ lies in S.

function of $\operatorname{conv}(\widetilde{\Sigma}_{\mathcal{I},s})$ that

$$\begin{aligned}\gamma^*_{\operatorname{conv}(\widetilde{\Sigma}_{\mathcal{I},s})}(\mathbf{u}) &= \sup\left\{\langle \mathbf{u}, \mathbf{x}\rangle : \gamma_{\operatorname{conv}(\widetilde{\Sigma}_{\mathcal{I},s})}(\mathbf{x}) \leq 1\right\} \\ &= \sup\left\{\langle \mathbf{u}, \mathbf{x}\rangle : \mathbf{x} \in \bigcup_{\mathcal{T}\in\mathcal{G}_s} \mathbb{S}^{d-1}_{\mathcal{T}}\right\} \\ &= \sup\left\{\left\langle \mathbf{u}, \frac{\mathbf{u}_{\mathcal{T}}}{\|\mathbf{u}_{\mathcal{T}}\|_2}\right\rangle : \mathcal{T} \in \mathcal{G}_s\right\} \\ &= \sup\left\{\|\mathbf{u}_{\mathcal{T}}\|_2 : \mathcal{T} \in \mathcal{G}_s\right\} \\ &= \|\mathcal{H}_{\mathcal{I},s}(\mathbf{u})\|_2.\end{aligned}$$

From the derivation, it immediately follows that the vector

$$\hat{\mathbf{x}} = \frac{\mathcal{H}_{\mathcal{I},s}(\mathbf{u})}{\|\mathcal{H}_{\mathcal{I},s}(\mathbf{u})\|_2}$$

attains the supremum in the definition of the dual gauge function and is therefore a maximizer of Problem ($\mathrm{P}_{4.6}$) for $\mathcal{K} = \widetilde{\Sigma}_{\mathcal{I},s}$. In other words, to solve Problem ($\mathrm{P}_{4.6}$) with $\mathcal{K} = \widetilde{\Sigma}_{\mathcal{I},s}$, we replace the regular hard thresholding operator $\mathcal{H}_s$ by its group-sparse counterpart $\mathcal{H}_{\mathcal{I},s} \colon \mathbb{R}^d \to \Sigma_{\mathcal{I},s}$, namely the so-called *group hard thresholding operator* which only retains the entries belonging to the s groups in $\mathcal{I}$ with largest ℓ_2-norm. Given a group-sparse vector $\mathring{\mathbf{x}} \in \Sigma_{\mathcal{I},s}$ and its quantized projections $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}})$, we therefore define the recovery map

$$\widetilde{\Delta}^{\mathrm{ht}}_{\mathcal{I}}(\mathbf{y}) = \mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top \mathbf{y}) = \operatorname*{argmin}_{\mathbf{x}\in\Sigma_{\mathcal{I},s}} \left\|\mathbf{x} - \mathbf{A}^\top \mathbf{y}\right\|_{\mathcal{I},1}. \tag{4.18}$$

The performance of this procedure is summarized in the following result, which was first shown by Foucart for the case of sparse vectors [**Fou16**]. The proof of the result is identical to the proof of Theorem 8 in [**Fou16**] as soon as one replaces the sets S and T by the respective group support sets $\mathcal{S} := \{I \in \mathcal{I} : \mathring{\mathbf{x}}_I \neq \mathbf{0}\}$ and $\mathcal{T} := \{I \in \mathcal{I} : \hat{\mathbf{x}}_I \neq \mathbf{0}\}$.

Lemma 4.24. *Let $\mathring{\mathbf{x}} \in \widetilde{\Sigma}_{\mathcal{I},s}$, and assume the matrix $\mathbf{A} \in \mathbb{R}^{m\times d}$ satisfies the group-RIP of order $2s$ with constant δ_{2s}. Then it holds for $\hat{\mathbf{x}} = \widetilde{\Delta}^{\mathrm{ht}}_{\mathcal{I}}(\operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}))$ that*

$$\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \leq 2\sqrt{5}\sqrt{\delta_{2s}} \quad \textit{and} \quad \left\|\mathring{\mathbf{x}} - \frac{\hat{\mathbf{x}}}{\|\hat{\mathbf{x}}\|_2}\right\|_2 \leq 4\sqrt{5}\sqrt{\delta_{2s}}.$$

As in the case of Lemma 4.6 and Lemma 4.16, the analysis of Lemma 4.24 assumes that one has access to the noise-free measurements $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}})$. Similar to the performance analysis of Problem ($\mathrm{P}_{4.3}$), it remains an open problem how to modify the proof of Lemma 4.24 to harden the result against pre- and/or post-quantization noise. In light of the previous discussion which connects $\widetilde{\Delta}^{\mathrm{ht}}_{\mathcal{I}}$ with solutions of Problem ($\mathrm{P}_{4.6}$), however, we may simply reuse the analysis of Theorem 4.23 to establish a noise-robust recovery guarantee of the hard thresholding procedure $\widetilde{\Delta}^{\mathrm{ht}}_{\mathcal{I}}$ at the expense of losing uniformity of the result over the set $\operatorname{conv}(\widetilde{\mathcal{E}}_{\mathcal{I},s})$. To that end, we consider again the noisy observation model

$$Q(\mathbf{A}\mathring{\mathbf{x}}) = \mathbf{f} \circ \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\nu}) \tag{4.19}$$

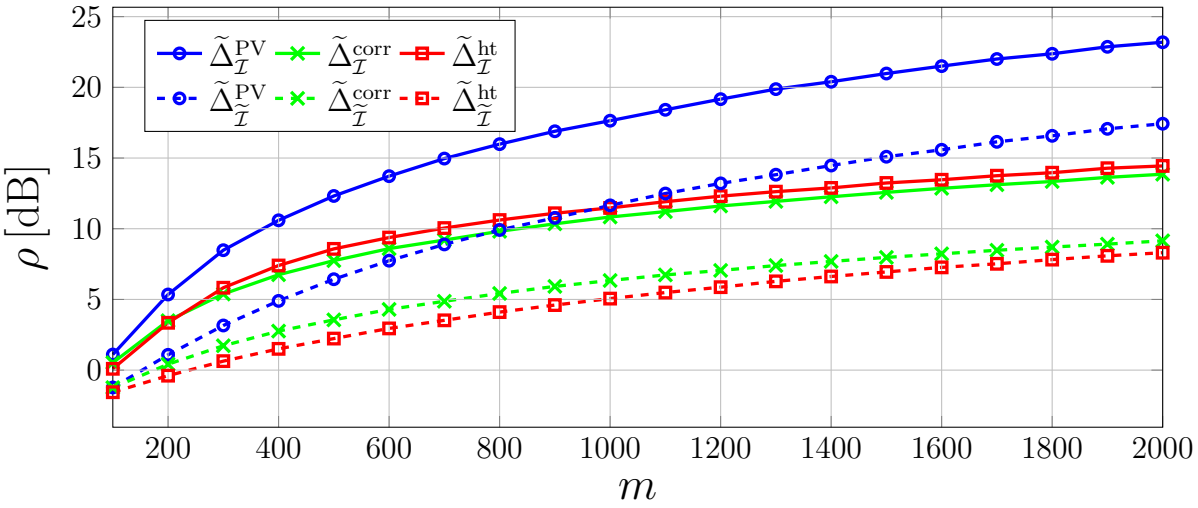


Figure 4.1: SNR vs. number of measurements. The dashed lines represent the performance when the group-sparsity structure is ignored. In this case, each algorithm assumes the trivial group partition $\widetilde{\mathcal{I}} = \{\{1\}, \dots, \{d\}\}$.

with $\mathbf{f} \sim \mathsf{B}_m(p)$ and $\boldsymbol{\nu} \sim \mathsf{N}(\mathbf{0}, \sigma^2 \mathrm{Id}_m)$ as before. Due to invariance of the mean width w. r. t. the convex hull and the fact that

$$w(\widetilde{\Sigma}_{\mathcal{I},s}) \leq \sqrt{2s \log(2eG/s)} + \sqrt{sg}$$

according to Lemma 4.10, the scaling requirements on m to guarantee ε-accurate reconstruction quality of the recovery map $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$ are identical to the requirements of Theorem 4.23 up to a multiplicative constant. For completeness, we repeat the statement of the result for the recovery map $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$ below.

Theorem 4.25. *Let $\mathring{\mathbf{x}} \in \widetilde{\Sigma}_{\mathcal{I},s}$, and denote by $\mathbf{A} \in \mathbb{R}^{m \times d}$ a standard Gaussian random matrix. Then with probability at least $1 - \eta$, every normalized vector $\hat{\mathbf{z}} = \hat{\mathbf{x}}/\|\hat{\mathbf{x}}\|_2$ with $\hat{\mathbf{x}} = \widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}(Q(\mathbf{A}\mathring{\mathbf{x}}))$ and Q denoting the noisy measurement operator defined by (4.19) satisfies $\|\mathring{\mathbf{x}} - \hat{\mathbf{z}}\|_2 \leq \varepsilon$, provided that*

$$m \gtrsim \varepsilon^{-4}(\sigma^2 + 1)(2p - 1)^{-2}\Big(s \log(G/s) + sg + \log(\eta^{-1})\Big).$$

4.3.4 Numerical Evaluation

In this section, we conduct a numerical study to compare the recovery performance of the discussed methods in terms of their estimation accuracy of group-sparse vectors on the unit sphere. In particular, we aim to gain an insight as to how the dependence of the number of measurements on ε compares to the theoretical behavior. Throughout, we consider a signal dimension of $d = 1000$ and split the support set $[d]$ into $G = 100$ nonoverlapping groups. Moreover, we consider $s = 5$ active groups chosen uniformly at random such that each realization contains $g \cdot s = 10 \cdot 5 = 50$ nonzero coefficients with each individual nonzero entry drawn i.i.d. from the standard Gaussian distribution. Finally, we project each vector on $\mathbb{S}^{d-1}$ by normalizing it to unit norm. For each parameter combination, we run 1000 Monte Carlo trials. We compare the three presented group-sparse recovery

methods with their sparse counterparts which we obtain by replacing $\mathcal{I}$ with the trivial partition $\widetilde{\mathcal{I}} := \{\{1\}, \dots, \{d\}\}$. Note that since $\widetilde{\Delta}^{\mathrm{corr}}_{\widetilde{\mathcal{I}}}$ and $\widetilde{\Delta}^{\mathrm{ht}}_{\widetilde{\mathcal{I}}}$ require an estimate of the sparsity level, we provide each recovery method with the total sparsity level $s \cdot g$.

Noiseless Reconstruction Performance

In the first experiment, we consider the recovery performance in terms of the average SNR according to $\rho\,[\mathrm{dB}] = -20 \log_{10} \|\mathring{\mathbf{x}} - \hat{\mathbf{x}}/\|\hat{\mathbf{x}}\|_2\|_2$ where $\hat{\mathbf{x}}$ denotes the output of some recovery method. The average of the Monte Carlo trials is evaluated in the linear domain. The results are shown in Figure 4.1. Unsurprisingly, each method's performance improves with increasing m with $\widetilde{\Delta}^{\mathrm{PV}}_{\mathcal{I}}$ clearly outperforming its competitors. More surprisingly, however, is the fact $\widetilde{\Delta}^{\mathrm{PV}}_{\widetilde{\mathcal{I}}}$ starts to outperform both the hard thresholding approach and $\widetilde{\Delta}^{\mathrm{corr}}_{\mathcal{I}}$ for $m \geq 1000$. In general, the experiments confirm the relation between the different scaling behaviors required for each method as established in Theorem 4.19, 4.23 and 4.25. More precisely, the dependence on ε is most favorable for $\widetilde{\Delta}^{\mathrm{PV}}_{\mathcal{I}}$, while $\widetilde{\Delta}^{\mathrm{corr}}_{\mathcal{I}}$ and $\widetilde{\Delta}^{\mathrm{ht}}_{\mathcal{I}}$ fall slightly behind due to their dependence on ε^{-4} rather than ε^{-3}.

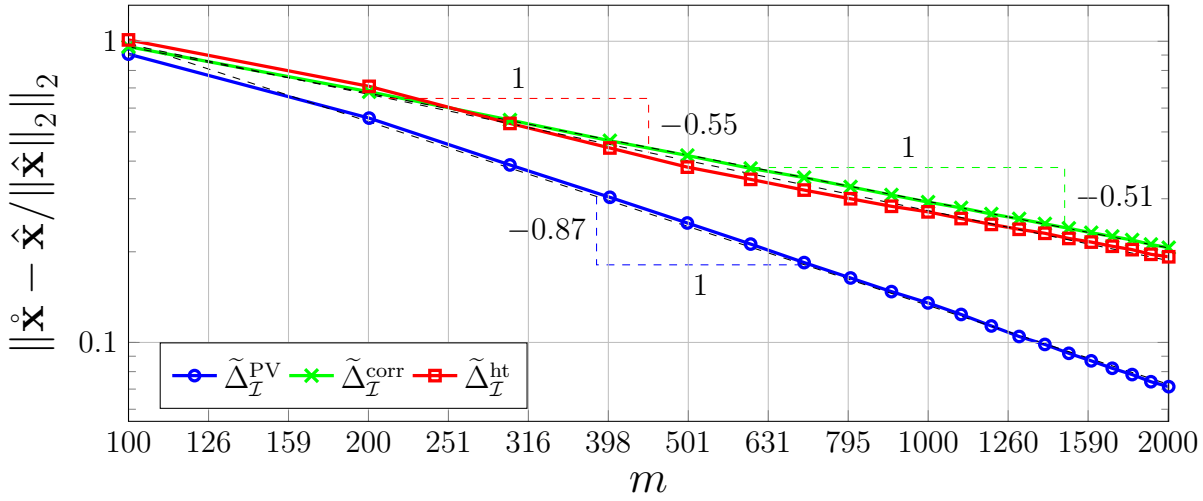
While the experiment already confirms that $\widetilde{\Delta}^{\mathrm{PV}}_{\mathcal{I}}$ outperforms $\widetilde{\Delta}^{\mathrm{corr}}_{\mathcal{I}}$ and $\widetilde{\Delta}^{\mathrm{ht}}_{\mathcal{I}}$, we are not able to infer the exact decay rate of each method from Figure 4.2. To that end, we display the normalized ℓ_2-error as a function of m on a doubly-logarithmic scale in Figure 4.2. In this representation, the polynomial decay rate of the normalized reconstruction error for $\widetilde{\Delta}^{\mathrm{PV}}_{\mathcal{I}}, \widetilde{\Delta}^{\mathrm{corr}}_{\mathcal{I}}$ and $\widetilde{\Delta}^{\mathrm{ht}}_{\mathcal{I}}$ is clearly visible. Moreover, as predicted by Theorem 4.23 and Theorem 4.25, the performance of $\widetilde{\Delta}^{\mathrm{corr}}_{\mathcal{I}}$ and $\widetilde{\Delta}^{\mathrm{ht}}_{\mathcal{I}}$ is almost identical. This is explained by the fact that both algorithms are connected to each other through the problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{K} \end{aligned} \tag{$\mathrm{P}_{4.7}$}$$

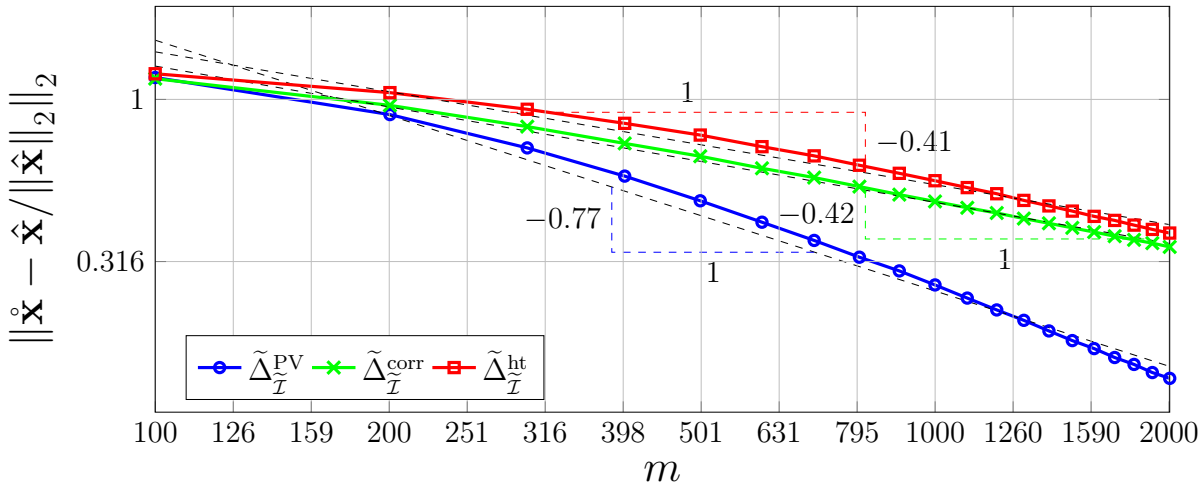
with $\widetilde{\Delta}^{\mathrm{corr}}_{\mathcal{I}}$ corresponding to the choice $\mathcal{K} = \widetilde{\mathcal{E}}_{\mathcal{I},s}$ and $\mathcal{K} = \widetilde{\Sigma}_{\mathcal{I},s}$ for $\widetilde{\Delta}^{\mathrm{ht}}_{\mathcal{I}}$ (cf. Section 4.3.3). As indicated by the annotations in Figure 4.2a, the empirical decay rate of $\widetilde{\Delta}^{\mathrm{corr}}_{\mathcal{I}}$ and $\widetilde{\Delta}^{\mathrm{ht}}_{\mathcal{I}}$ is $\mathcal{O}(m^{-1/2})$ rather than the slower rate of $\mathcal{O}(m^{-1/4})$ predicted by Proposition 4.44. While the predicted decay rate of $\mathcal{O}(m^{-1/3})$ is also slightly off for $\widetilde{\Delta}^{\mathrm{PV}}_{\mathcal{I}}$, its empirical rate of $\mathcal{O}(m^{-4/5})$ is closest to the provably optimal rate of $\mathcal{O}(m^{-1})$ for nonadaptive measurements among all possible reconstruction maps as established in [**Jac+13**]. The situation changes, however, when trying to estimate the empirical error rate of the recovery maps $\widetilde{\Delta}^{\mathrm{PV}}_{\widetilde{\mathcal{I}}}, \widetilde{\Delta}^{\mathrm{corr}}_{\widetilde{\mathcal{I}}}$ and $\widetilde{\Delta}^{\mathrm{ht}}_{\widetilde{\mathcal{I}}}$, which all operate under a model mismatch. As depicted in Figure 4.2b, the error no longer decays linearly in m (in the logarithmic representation). This is due to the fact that no method exploits the true underlying signal structure and is therefore governed by a different error decay than predicted by Proposition 4.44.

Group Support Identification

In many practical applications, rather than aiming at recovering the individual components of a vector, one might instead only be interested in identifying the active groups of a signal. Historically, this problem has only received limited attention in the compressed sensing literature due to the fact that in the absence of quantization, support recovery and vector recovery are equally hard. More precisely, once the support of a signal is known, vector recovery reduces to a simple least-squares projection. In the presence of



(a) Performance of group-sparse direction recovery methods and their associated empirical error decay rates



(b) Performance of sparse direction recovery methods applied to group-sparse signal estimation when the group-sparsity structure is treated as regular signal sparsity

Figure 4.2: Normalized ℓ_2-error vs. number of measurements on a doubly-logarithmic scale. The dashed lines correspond to the linear regression line of each respective curve in combination with their slopes indicating the decay rate of the reconstruction error. The fact that each graph in Figure 4.2a is close to its regression line suggests that the dependence of each method's performance on the system parameters s, g and G is accurately captured by the theory. This is in contrast to Figure 4.2b where the error does not decay log-linearly since there is a mismatch between the recovery schemes and the underlying signal set. As predicted by Theorem 4.23 and Theorem 4.25, the performance of $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$ and $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$ are virtually identical. However, both methods exhibit a faster error decay rate than their predicted rate of $1/4$. Similarly, the real decay rate of $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ is closer to the provably optimal rate of 1 among all nonadaptive recovery schemes than to its predicted rate of $1/3$ according to Theorem 4.19.

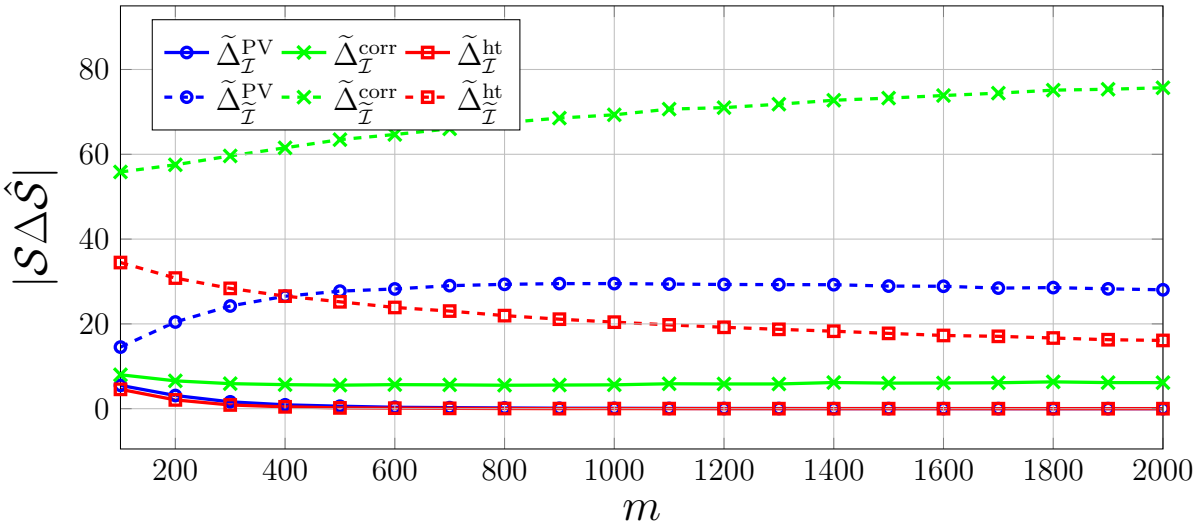


Figure 4.3: Support estimation error. Dashed lines correspond to the performance w. r. t. $\widetilde{\mathcal{I}} = \{\{1\}, \dots, \{d\}\}$.

quantization or any other type of nonlinearity in the acquisition system, the situation changes since knowing the support still does not enable us to perfectly recover sparse or group-sparse vectors. We therefore conduct a simple experiment to gauge how well the individual recovery schemes fare in the context of support identification from highly nonlinear observations of low-complexity signals. For simplicity, we restrict our attention to genuinely group-sparse vectors again.

To investigate how well the individual reconstruction methods manage to identify the group support set $\mathcal{S} = \operatorname{supp}_{\mathcal{I}}(\mathring{\mathbf{x}}) := \{i \in [G] : \mathring{\mathbf{x}}_{\mathcal{I}_i} \neq \mathbf{0}\}$, we consider the symmetric set difference $\mathcal{S} \Delta \hat{\mathcal{S}} := (\mathcal{S} \setminus \hat{\mathcal{S}}) \cup (\hat{\mathcal{S}} \setminus \mathcal{S})$ between the ground truth $\mathcal{S}$ and the estimated group index set $\hat{\mathcal{S}} \subset [G]$, respectively. This metric captures both the *false alarm* and *missed detection rate* by quantifying how many groups were erroneously selected by a recovery method and how many nonzero groups in the ground truth signal $\mathring{\mathbf{x}} \in \widetilde{\Sigma}_{\mathcal{I},s}$ were missed. Despite the fact that Problem (P$_{4.3}$) employs the group ℓ_1-norm to promote group-sparse quantization-consistent solutions, empirical experiments show that minimizers are mostly not exactly group-sparse but rather compressible in the sense that the best s-term group approximation error $\sigma_{\mathcal{I},1}(\hat{\mathbf{x}})_s$ decays quickly with s. This is also mirrored in Lemma 4.4, which predicts that the estimator $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ produces merely effectively rather than genuinely group-sparse vectors. The same holds true for estimates produced by $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$, which explicitly allows for effectively group-sparse solutions. In order to properly compare the group support identification performance, we therefore need a way to estimate the dominant groups of a vector. To that end, we employ the following thresholding strategy. Given an estimator $\hat{\mathbf{x}}$, we denote by $\check{\mathcal{I}} = \{\check{\mathcal{I}}_1, \dots, \check{\mathcal{I}}_G\}$ the nonincreasing group rearrangement of $\hat{\mathbf{x}}$ such that $\|\hat{\mathbf{x}}_{\check{\mathcal{I}}_1}\|_2 \geq \dots \geq \|\hat{\mathbf{x}}_{\check{\mathcal{I}}_G}\|_2$ with $\check{\mathcal{I}}_i = \mathcal{I}_{\pi(i)}$ and $\pi \colon [G] \to [G]$ a permutation. Next, define the group index set $\mathcal{J}^{(n)} := \{\check{\mathcal{I}}_1, \dots, \check{\mathcal{I}}_n\}$. With slight abuse of notation, we write $\hat{\mathbf{x}}_{\mathcal{J}^{(n)}}$ for the restriction of $\hat{\mathbf{x}}$ to the index set $\bigcup_{J \in \mathcal{J}^{(n)}} J$. We then iterate over $n = 1, \dots, G-1$ until the stopping criterion

$$\frac{\|\hat{\mathbf{x}}_{\mathcal{J}^{(n+1)}} - \hat{\mathbf{x}}_{\mathcal{J}^{(n)}}\|_2}{\|\hat{\mathbf{x}}_{\mathcal{J}^{(n+1)}}\|_2} \leq \mu$$

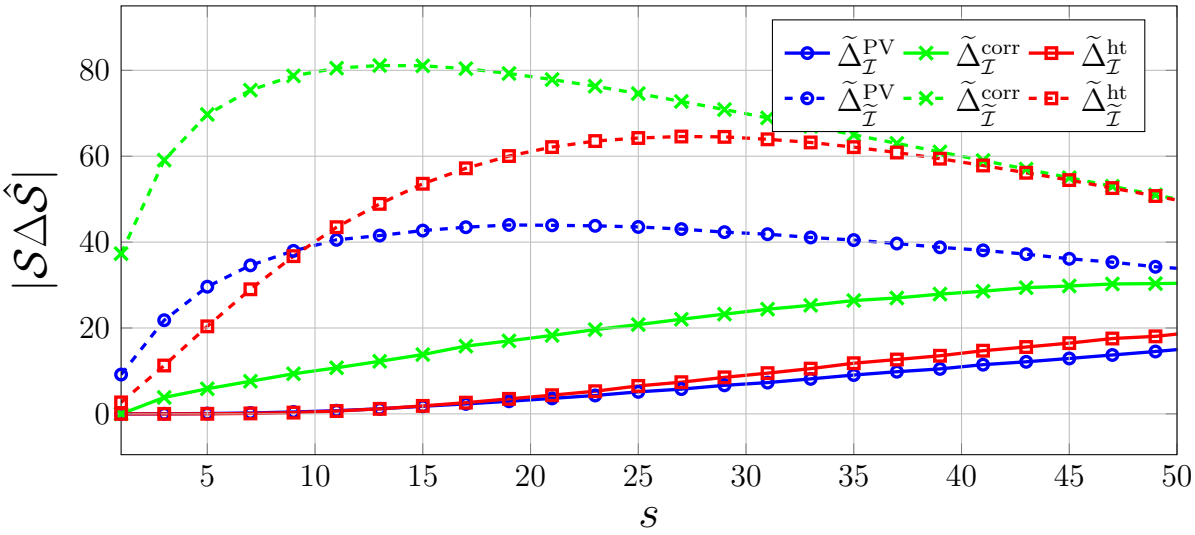


Figure 4.4: Group support error vs. group-sparsity level for $m = 1000$. The dashed lines represent the performance when the group-sparsity structure is ignored. In this case, each algorithm assumes the trivial group partition $\widetilde{\mathcal{I}} = \{\{1\}, \ldots, \{d\}\}$.

is satisfied at some iteration $n^\star$ for the prescribed tolerance $\mu = 10^{-3}$. We then set $\hat{\mathcal{S}} = \{\pi(1), \ldots, \pi(n^\star)\}$.

The results of this experiment are shown in Figure 4.3. While both $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ and $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$ manage to almost perfectly recover the exact group support for $m \geq 400$, $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$ consistently misidentifies around 6 groups even as the number of measurements increases. On the other hand, despite outperforming both $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$ and $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$ in the previous experiment, $\widetilde{\Delta}_{\widetilde{\mathcal{I}}}^{\mathrm{PV}}$ fails to properly identify the group support if the inherent group structure is not explicitly exploited. This can be explained as follows. Despite being oblivious to the underlying group structure, the estimator $\widetilde{\Delta}_{\widetilde{\mathcal{I}}}^{\mathrm{PV}}$ manages to identify the most important coefficients inside each active group. However, on average the program also selects several coefficients which do not belong to one of the active groups, which in turn leads to a large group misidentification rate.

We also consider the behavior when the number of measurements is fixed at $m = 1000$ and we in turn vary the group-sparsity level s. The results depicted in Figure 4.4 reveal that both $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ and $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$ are very competitive and accurate up to $s = 15$ active groups corresponding to $s \cdot g = 150$ nonzero coefficients. This is a substantial increase over standard methods which treat the group-sparsity structure as canonical sparsity and are only accurate at very low group-sparsity levels. Surprisingly, the support recovery performance of $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$ deteriorates much faster than expected considering its similar performance to $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$ in terms of the attained SNR. However, the finding explains the constant bias of $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$ in estimating the support for a fixed group-sparsity level as considered in Figure 4.3. Since $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$ optimizes over the set $\widetilde{\mathcal{E}}_{\mathcal{I},s}$, a bigger set than $\widetilde{\Sigma}_{\mathcal{I},s}$ as considered by $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$, it suggests that despite the fact that $\widetilde{\Sigma}_{\mathcal{I},s} \subset \widetilde{\mathcal{E}}_{\mathcal{I},s}$, $\widetilde{\mathcal{E}}_{\mathcal{I},s}$ contains vectors whose linear measurements are more strongly correlated with the quantized measurements than elements in $\widetilde{\Sigma}_{\mathcal{I},s}$ exclusively.

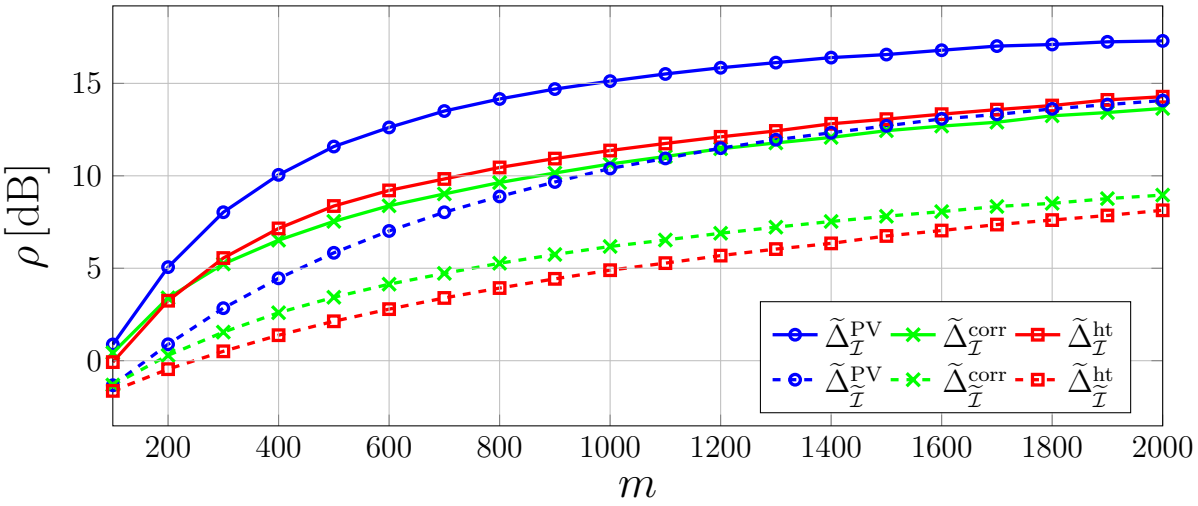


Figure 4.5: SNR vs. number of measurements when measurements are disturbed by additive Gaussian pre-quantization noise with standard deviation $\sigma = 0.2$, corresponding to around 10 % of all sign measurements being flipped. The dashed lines represent the performance when the group-sparsity structure is ignored, *i.e.*, each algorithm assumes the trivial group partition $\widetilde{\mathcal{I}} = \{\{1\}, \dots, \{d\}\}$ while being provided with the total sparsity level (if required).

Direction Recovery and Support Identification from Noisy Observations

We now repeat the previous two experiments in the presence of additive pre-quantization noise. In particular, we consider measurements of the form $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\nu})$ with $\boldsymbol{\nu} \sim \mathsf{N}(\mathbf{0}, \sigma^2 \mathrm{Id}_m)$ and $\sigma = 0.2$. Note that according to [Jac+13, Lemma 4], a noise standard deviation of $\sigma = 0.2$ implies that roughly 10 % of all sign measurements are flipped, which bears the potential to substantially degrade the performance of each recovery method. In particular, Jacques *et al.* show that the expected Hamming distance between noiseless and noisy quantized measurements is bounded by

$$\mathbb{E}\Delta_{\mathrm{H}}(\operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}), \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\nu})) \leq \frac{1}{2} \frac{\sigma}{\sqrt{\sigma^2 + 1}}$$

for $\mathring{\mathbf{x}} \in \mathbb{S}^{d-1}$ and $\mathbf{A}$ a standard Gaussian random matrix. By the Chernoff bound, the binomial random variable $m\Delta_{\mathrm{H}}(\operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}}), \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\nu}))$ then concentrates sharply around its mean.

The results of the experiment are shown in Figure 4.5. While the relative performance of each method w. r. t. other methods remains the same with $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ outperforming both $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$ and $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$, the performance of $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ drops off by around 5 dB compared to the noiseless case. On the other hand, the SNR of both $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$ and $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$ in the noisy setting remains almost unchanged compared to the noiseless setting, emphasizing the error resilience of both methods whose performance is once again on par due to their close connection through Problem ($\mathrm{P}_{4.7}$) and their corresponding performance analyses.

We also consider the support identification accuracy in the noisy setting whose results are depicted in Figure 4.6. As one would expect, the performance of both $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$ and $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$ once again remain unchanged. This means that even in highly noisy scenarios the group hard thresholding approach yields surprisingly accurate support detection performance considering the simplicity of its implementation. We also once again observe the constant

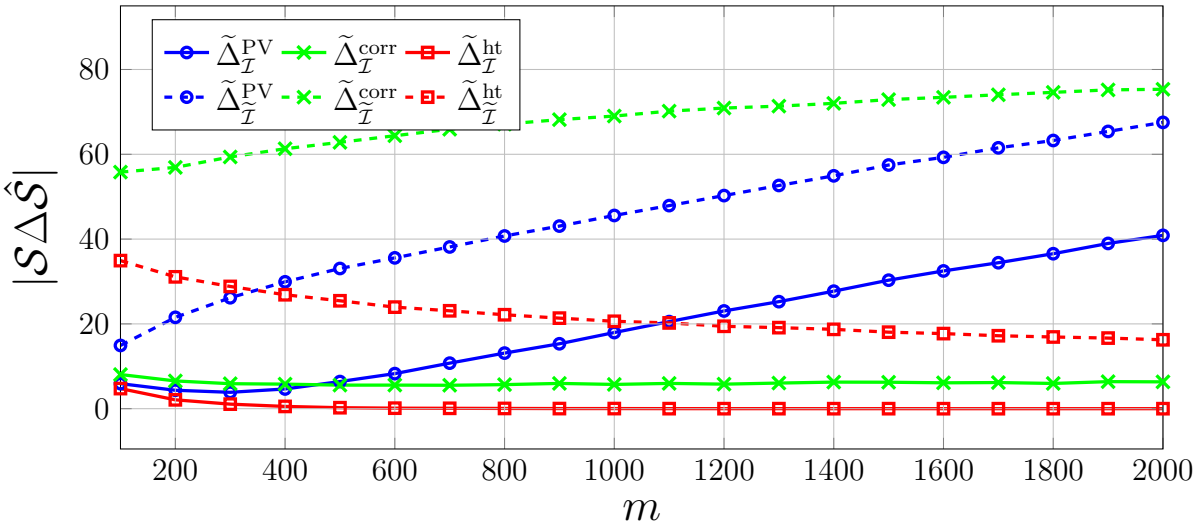


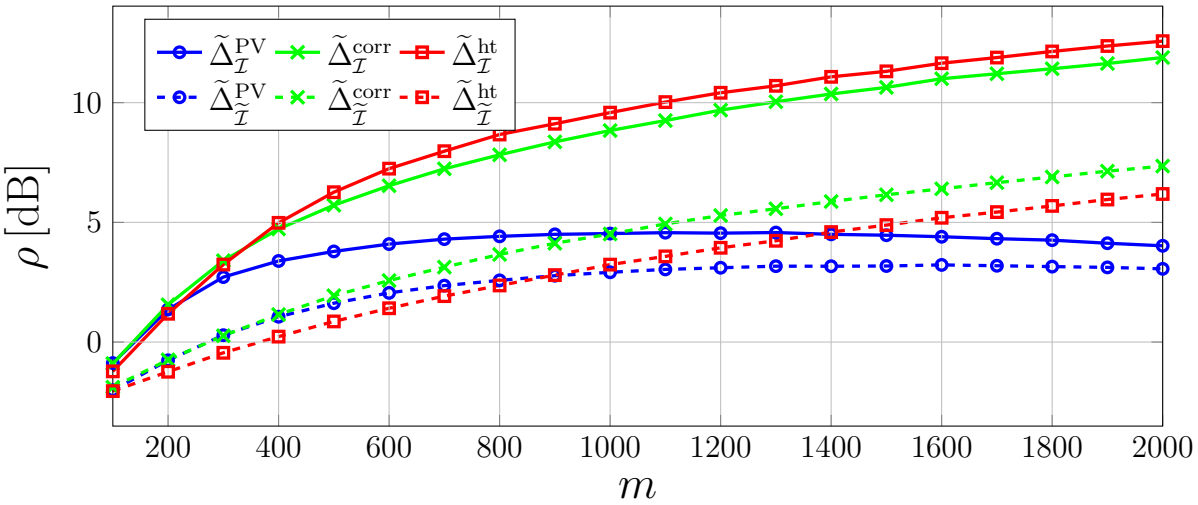
Figure 4.6: Support estimation error in the noisy regime. Dashed lines correspond to the performance w. r. t. $\widetilde{\mathcal{I}} = \{\{1\}, \ldots, \{d\}\}$.

bias of $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$, which consistently misidentifies around 5 to 6 groups according to our thresholding rule to determine active groups in a reconstructed vector. On the other hand, while surprisingly accurate at smaller m, the performance of $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ quickly deteriorates as m increases to the point where as many as 40 groups are erroneously selected. Again, one likely explanation for this phenomenon is rooted in the fact that while $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ correctly identifies the active groups, resulting in a competitive reconstruction performance as shown in Figure 4.5, the scheme also selects a substantial number of other groups with considerably lower but nonnegligible energy.

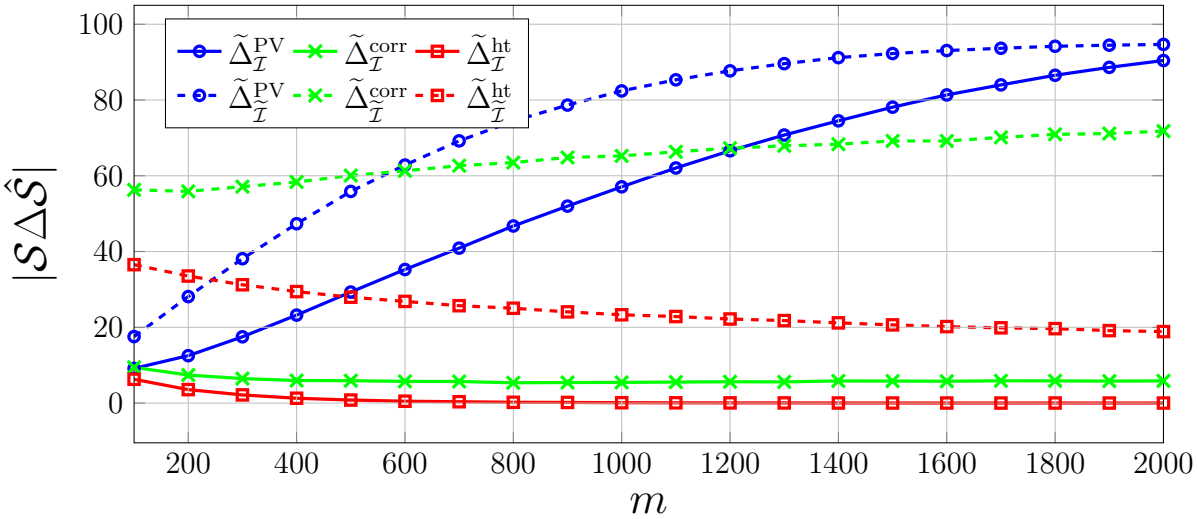
In the last experiment, we repeat the above investigation in the case of adversarial post-quantization noise. This means that we now consider measurements of the form $\mathbf{y} = \mathbf{f} \circ \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}})$ where $\mathbf{f} \sim \mathsf{B}_m(p)$ with $p = 0.9$, meaning that again on average 10 % of all measurements are flipped. Both the performance in terms of average SNR and support identification are presented in Figure 4.7. The graphs tell a familiar story with one crucial exception: the performance of $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ in the face of adversarial post-quantization noise drops considerably compared to its behavior in the presence of pre-quantization noise. While both $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$ and $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$ also exhibit slightly worse performance compared to the previous setting, they generally perform on par with $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$ again slightly edging out $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$ both in terms of average reconstruction error and support detection.

4.4 Recovery from Dithered Observations

In this section, we lift the restriction that signals of interest belong to the unit Euclidean sphere. This change alone has no effect on the recovery results presented in the previous section as the sgn-operator is invariant under positive scaling of its argument. However, as is a well-established fact in the literature of 1-bit compressed sensing at this point, we can circumvent this shortcoming in the measurement procedure by shifting the linear projections by a so-called dithering vector $\boldsymbol{\tau} \in \mathbb{R}^m$ prior to quantization. In particular, we

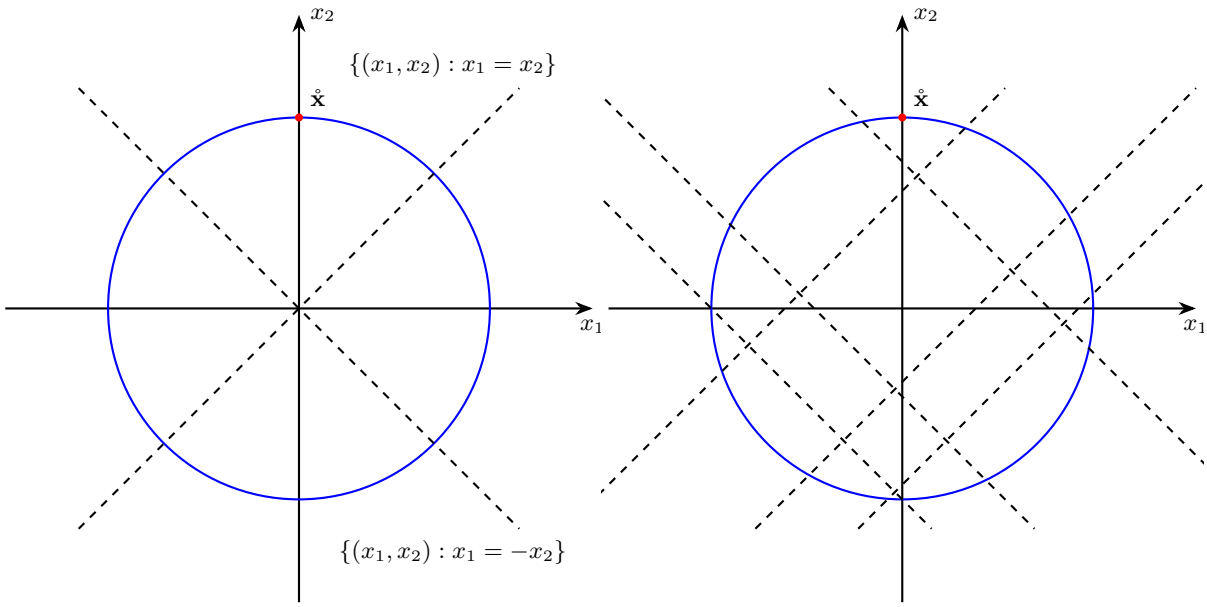


(a) Average SNR vs. number of measurements



(b) Average support identification error vs. number of measurements

Figure 4.7: SNR and support detection error performance in the presence of adversarial post-quantization noise with a sign-flip probability of $1 - p = 0.1$.



(a) Recovery of a 1-sparse vector without dithering **(b)** Recovery of a 1-sparse vector with dithering

Figure 4.8: Recovery of vectors in $\mathbb{R}^2$ from 1-bit observations when the measurement matrix $\mathbf{A}$ is populated with i.i.d. Bernoulli random variables.

now consider for $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{A} \in \mathbb{R}^{m \times d}$ measurements of the form

$$\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathbf{x} + \boldsymbol{\tau}). \tag{4.20}$$

Clearly, the only way we can hope to sensibly estimate the norm of $\mathbf{x}$ in addition to its direction is by choosing the (nonadaptive) threshold vector $\boldsymbol{\tau}$ as a function of the norm of $\mathbf{x}$. As such, we assume that vectors of interest are contained in a scaled unit Euclidean ball of radius r, *i.e.*, we now consider recovery over the sets $\Sigma_{\mathcal{I},s} \cap r\mathbb{B}_2^d$ and $\mathcal{E}_{\mathcal{I},s} \cap r\mathbb{B}_2^d$, respectively.

The idea of dithering was also recently used to overcome the problems outlined in Section 4.3.1 when the underlying measurement matrix is drawn from a subgaussian distribution, which could potentially lead to distinct sparse vectors that are mapped to the same bit string (see, *e.g.*, [**DM18a**]). Conceptually, it is easy to see why dithering can be used to overcome the problem of too little variation in the choice of hyperplanes used to partition a signal set into quantization cells. Consider for instance the recovery of sparse vectors in $\mathbb{R}^2$ from binary observations of Bernoulli projections as depicted in Figure 4.8. In two dimensions, there are only four possible hyperplanes which tessellate the unit sphere into four[5] disjoint quantization cells (Figure 4.8a). As a result, the size of the quantization cell containing $\mathring{\mathbf{x}}$ in Figure 4.8a cannot be further reduced by taking more measurements.[6] The situation changes drastically, however, if one incorporates dithering

[5]Note that one hyperplane alone always defines two quantization cells (*i.e.*, half-spaces) so the two hyperplanes depicted in Figure 4.8a define all possible quantization cells since the other two hyperplanes are collinear to the ones depicted.

[6]Note that in two dimensions this is not an issue since there are only four distinct 1-sparse vectors on the unit sphere $\mathbb{S}^1$. In higher dimensions, however, it causes issues when sparse vectors are not separated by at least one hyperplane. In the Gaussian setting, this happens with high probability on the draw of the vectors $\mathbf{a}_i$ (cf. [**PV13a**, Theorem 4.2]) but not in the Bernoulli setting.

of the form $y_i = \operatorname{sgn}(\langle \mathbf{a}_i, \mathring{\mathbf{x}} \rangle + \tau_i)$. By adding a (known) offset τ_i prior to quantization and thus shifting the quantization point, it becomes possible to tessellate the signal set in a more meaningful way when the choice of hyperplanes is limited. As demonstrated in Figure 4.8b, shifting the two hyperplanes from Figure 4.8a by varying offsets from the origin not only allows to reduce the size of the quantization region containing $\mathring{\mathbf{x}}$ (or any other 1-sparse vector on the boundary of $\mathbb{S}^1$) but also opens up the possibility to approximate the length of a vector.

4.4.1 Reconstruction via Quantization Consistency

In this section, we consider the recovery of group-sparse vectors by means of various convex programming techniques. Most of the results to follow build on guarantees previously established for undithered observations. The first theoretical result of this type which established an error bound under the measurement model (4.20) goes back to the work of Knudson, Saab and Ward [**KSW16**] who suggest to solve an augmented convex program to recover sparse vectors from dithered 1-bit observations for $\mathbf{A} \in \mathbb{R}^{m \times d}$ with standard normal rows $\mathbf{a}_i \sim_{\text{i.i.d.}} \mathsf{N}(\mathbf{0}, \mathrm{Id}_d)$. To gain an additional degree of freedom when estimating vectors outside of the unit sphere, they introduce an auxiliary variable by replacing the linear measurements $\mathbf{A}\mathbf{x} + \boldsymbol{\tau}$ with $\mathbf{A}\mathbf{x} + w\boldsymbol{\tau}/\theta$ during recovery where $\boldsymbol{\tau} \sim \mathsf{N}(\mathbf{0}, \theta^2 \mathrm{Id}_m)$ is assumed to be a Gaussian dithering vector.[7] The motivation for this step is obvious: if $\mathbf{A}$ consists of independent copies of a standard Gaussian random variable, then by introducing the variable w, the dithered measurements $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathbf{x} + w\boldsymbol{\tau}/\theta)$ can be expressed as the undithered measurements of an augmented low-complexity vector with an additional nonzero entry. In other words, one relates the problem of estimating a vector $\mathring{\mathbf{x}} \in r\mathbb{B}_2^d$ to the problem of estimating the extended vector $(\mathring{\mathbf{x}}^\top \quad \theta)^\top$ with the same measurements $\mathbf{y}$ by *lifting* the problem into a slightly higher-dimensional space. This transformation will be a common theme throughout the remainder of this chapter. Before turning our attention to the analysis of a dithered variant of Problem ($\mathrm{P}_{4.3}$) using the technique outlined above, however, we first address recovery of group-sparse vectors inside scaled Euclidean balls by a more straightforward approach. This particular formulation of the problem was also used in [**Bar+17a**] to prove a similar result in the case where signals are analysis-cosparse w. r. t. a tight frame.

Norm-Constrained Cone Programming

Recall from the discussion at the beginning of Section 4.3.1 that our main motivation in the formulation of Problem ($\mathrm{P}_{4.3}$) for the constraint $\langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle = 1$ was to remove the null space of $\mathbf{A}$ from the feasible set. In fact, the constraint not just removes $\ker(\mathbf{A})$ but also the subspace $(\mathbf{A}^\top \mathbf{y})^\perp$ from the search space. This step was necessary as the relaxation of the set of quantization-consistent vectors $\{\mathbf{x} : \mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathbf{x})\}$ to $\{\mathbf{x} : \mathbf{y} \circ \mathbf{A}\mathbf{x} \geq \mathbf{0}\}$ would otherwise render the resulting optimization problem trivial given the optimal solution of $\mathbf{x}^\star = \mathbf{0}$. Under the dithered observation model, however, the zero vector is not trivially feasible. Moreover, a constraint of the form $\langle \mathbf{y}, \mathbf{A}\mathbf{x} + \boldsymbol{\tau} \rangle = c_0$ for some constant $c_0 > 0$

[7] Technically, as we will see below, it suffices for the matrix $(\mathbf{A} \quad \boldsymbol{\tau}/\theta)$ to satisfy certain deterministic properties such as the group-RIP or the ε-tessellation property. However, by requiring $\boldsymbol{\tau}$ to be a Gaussian random vector we may immediately appeal to the probabilistic results from Section 4.3 to characterize the probability for these events to hold.

already imposes an energy constraint on $\mathbf{x}$ by forcing it to belong to the affine hyperplane $\{\mathbf{x} : \langle \mathbf{A}^\top \mathbf{y}, \mathbf{x}\rangle = c_0 - \langle \mathbf{y}, \boldsymbol{\tau}\rangle\}$. In other words, the constant c_0 would have to be tuned according to the unknown norm of target vector $\mathring{\mathbf{x}}$. Inspired by previous results in this direction [**Bar+17a**; **Bar+17b**; **DJR17**], we therefore consider the problem

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \|\mathbf{x}\|_{\mathcal{I},1} \\ \text{s.t.} & \mathbf{y} \circ (\mathbf{A}\mathbf{x} + \boldsymbol{\tau}) \geq \mathbf{0} \\ & \|\mathbf{x}\|_2 \leq r, \end{array} \tag{P$_{4.8}$}$$

which corresponds to Problem (P$_{4.3}$) with the constraint $\langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle = 1$ replaced by $\|\mathbf{x}\|_2 \leq r$. Again, we associate with Problem (P$_{4.8}$) a recovery map $\Delta_{\mathcal{I}}^{\mathrm{nc}} \colon \{\pm 1\}^m \to \mathbb{R}^d$ of the form

$$\Delta_{\mathcal{I}}^{\mathrm{nc}}(\mathbf{y}) = \operatorname*{arginf}_{\mathbf{x}} \left\{ \|\mathbf{x}\|_{\mathcal{I},1} : \mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathbf{x} + \boldsymbol{\tau}), \|\mathbf{x}\|_2 \leq r \right\}$$

where the notation $\Delta_{\mathcal{I}}^{\mathrm{nc}}$ is used as a hint that the operator is based on the norm-constrained *second-order cone program* (SOCP) (P$_{4.8}$). We point out that while Problem (P$_{4.8}$) does not in general have a trivial solution at $\mathbf{x}^\star = \mathbf{0}$, the program still outputs the zero vector with nontrivial probability. To see this, one needs to estimate the probability that $\mathbf{0}$ is feasible for Problem (P$_{4.8}$), which happens to be the case if $\mathbf{y} \circ \boldsymbol{\tau} \leq \mathbf{0}$. By independence of the entries of $\boldsymbol{\tau}$, this means

$$\mathbb{P}(\boldsymbol{\tau} \circ \mathbf{y} \leq \mathbf{0}) = \prod_{i=1}^{m} \mathbb{P}(\tau_i y_i \leq 0) \underset{(m\to\infty)}{\longrightarrow} 0.$$

In other words, while nontrivial, the probability that Problem (P$_{4.8}$) has a trivial solution is exponentially small, which in turn justifies not including a constraint of the form $\langle \mathbf{y}, \mathbf{A}\mathbf{x} + \boldsymbol{\tau}\rangle = c_0$ in the problem.

Note that since $\boldsymbol{\tau}$ has independent Gaussian entries with variance θ^2, the first constraint of the above problem is equivalent to

$$\operatorname{sgn}\left(\begin{pmatrix}\mathbf{A} & \mathbf{g}\end{pmatrix}\begin{pmatrix}\mathring{\mathbf{x}} \\ \theta\end{pmatrix}\right) = \operatorname{sgn}\left(\begin{pmatrix}\mathbf{A} & \mathbf{g}\end{pmatrix}\begin{pmatrix}\mathbf{x} \\ \theta\end{pmatrix}\right)$$

where $\mathbf{g} \in \mathbb{R}^m$ is a standard Gaussian random vector independent of $\mathbf{A}$. As alluded to before, the performance analysis of Problem (P$_{4.8}$) will consequently rely on the (ℓ_2, ℓ_1) group restricted isometry property of matrices of size $m \times (d+1)$. In general, the analysis proceeds along the lines of the proof of Lemma 4.4. Given an (effectively) s-group-sparse vector $\mathring{\mathbf{x}} \in \mathbb{R}^d$ with $\|\mathring{\mathbf{x}}\|_2 \leq r$ and a minimizer $\hat{\mathbf{x}}$ of Problem (P$_{4.8}$), we start by establishing that certain convex combinations of the augmented vectors

$$\mathring{\mathbf{u}} = \begin{pmatrix}\mathring{\mathbf{x}} \\ \theta\end{pmatrix} \quad \text{and} \quad \hat{\mathbf{u}} = \begin{pmatrix}\hat{\mathbf{x}} \\ \theta\end{pmatrix} \tag{4.21}$$

are effectively group-sparse.

Lemma 4.26. *Let $\mathring{\mathbf{x}} \in \mathcal{K} \cap r\mathbb{B}_2^d$ with $\mathcal{K} = \Sigma_{\mathcal{I},s}$ or $\mathcal{K} = \mathcal{E}_{\mathcal{I},s}$, and assume $\tilde{\mathbf{A}} = (\mathbf{A} \quad \mathbf{g}) \in \mathbb{R}^{m\times(d+1)}$ satisfies the (ℓ_2, ℓ_1) group restricted isometry property of order*

$$t = 4(s+1)\left(1 + \frac{r^2}{\theta^2}\right)\left(\frac{1+\delta_t}{1-\delta_t}\right)^2 \tag{4.22}$$

with constant δ_t. Denote by $\hat{\mathbf{x}}$ a minimizer of Problem ($\mathrm{P}_{4.8}$) given the measurements $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$ with $\boldsymbol{\tau} = \theta\mathbf{g}$. Then with $\mathring{\mathbf{u}}$ and $\hat{\mathbf{u}}$ as defined in (4.21) and $\lambda \in [0,1]$, the convex combination

$$\bar{\mathbf{u}} = (1-\lambda)\frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} + \lambda\frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2} \in \mathbb{R}^{d+1}$$

is effectively t-group-sparse w. r. t. the extended group partition $\widetilde{\mathcal{I}} := \mathcal{I} \cup \{\{d+1\}\}$.

Proof. We begin the proof by establishing that $\mathring{\mathbf{u}}$ and $\hat{\mathbf{u}}$ are effectively group-sparse. To that end, note that by definition of $\mathring{\mathbf{u}}$, we have

$$\begin{aligned}
\|\mathring{\mathbf{u}}\|_{\widetilde{\mathcal{I}},1} &= \|\mathring{\mathbf{x}}\|_{\mathcal{I},1} + \theta \\
&\leq \sqrt{s}\|\mathring{\mathbf{x}}\|_2 + \theta \\
&\leq r\sqrt{s} + \theta \\
&\leq \left\langle \begin{pmatrix} \sqrt{s} \\ 1 \end{pmatrix}, \begin{pmatrix} r \\ \theta \end{pmatrix} \right\rangle \\
&\leq \sqrt{s+1}\sqrt{r^2+\theta^2}
\end{aligned}$$

where we invoked the Cauchy-Schwarz inequality in the last step. Since we trivially have $\|\mathring{\mathbf{u}}\|_2, \|\hat{\mathbf{u}}\|_2 \geq \theta$, it follows that

$$\frac{\|\mathring{\mathbf{u}}\|_{\widetilde{\mathcal{I}},1}}{\|\mathring{\mathbf{u}}\|_2} \leq \sqrt{s+1}\sqrt{1+\frac{r^2}{\theta^2}}.$$

Hence, $\mathring{\mathbf{u}}$ is effectively $(s+1)(1+r^2/\theta^2)$-group-sparse w. r. t. $\widetilde{\mathcal{I}}$. Moreover, optimality of $\hat{\mathbf{x}}$ for Problem ($\mathrm{P}_{4.8}$) implies that

$$\|\hat{\mathbf{u}}\|_{\widetilde{\mathcal{I}},1} = \|\hat{\mathbf{x}}\|_{\mathcal{I},1} + r \leq \|\mathring{\mathbf{x}}\|_{\mathcal{I},1} + r = \|\mathring{\mathbf{u}}\|_{\widetilde{\mathcal{I}},1}$$

so that $\hat{\mathbf{u}}$ is also effectively $(s+1)(1+r^2/\theta^2)$-group-sparse. From the triangle inequality, it therefore follows for $\bar{\mathbf{u}}$ that

$$\|\bar{\mathbf{u}}\|_{\widetilde{\mathcal{I}},1} \leq (1-\lambda)\frac{\|\mathring{\mathbf{u}}\|_{\widetilde{\mathcal{I}},1}}{\|\mathring{\mathbf{u}}\|_2} + \lambda\frac{\|\hat{\mathbf{u}}\|_{\widetilde{\mathcal{I}},1}}{\|\hat{\mathbf{u}}\|_2} \leq \sqrt{s+1}\sqrt{1+\frac{r^2}{\theta^2}}. \tag{4.23}$$

Next we bound $\|\bar{\mathbf{u}}\|_2$ from below by the same technique as employed in the proof of Lemma 4.4. Since $\tilde{\mathbf{A}}$ was assumed to satisfy the (ℓ_2, ℓ_1)-group-RIP of order

$$t = 4(s+1)\left(1+\frac{r^2}{\theta^2}\right)\left(\frac{1+\delta_t}{1-\delta_t}\right)^2,$$

we have by feasibility of $\mathring{\mathbf{x}}$ and $\hat{\mathbf{x}}$ for Problem ($\mathrm{P}_{4.8}$), *i.e.*, $\operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau}) = \operatorname{sgn}(\mathbf{A}\hat{\mathbf{x}} + \boldsymbol{\tau})$ and hence $\operatorname{sgn}(\tilde{\mathbf{A}}\mathring{\mathbf{u}}) = \operatorname{sgn}(\tilde{\mathbf{A}}\hat{\mathbf{u}})$, that

$$\left\|\tilde{\mathbf{A}}\bar{\mathbf{u}}\right\|_1 = (1-\lambda)\left\|\frac{\tilde{\mathbf{A}}\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2}\right\|_1 + \lambda\left\|\frac{\tilde{\mathbf{A}}\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2}\right\|_1 \geq (1-\lambda)(1-\delta_t) + \lambda(1-\delta_t) = 1-\delta_t.$$

Next, we bound $\|\tilde{\mathbf{A}}\bar{\mathbf{u}}\|_1$ from above in terms of the ℓ_2-norm of $\bar{\mathbf{u}}$. With the usual notation where $\mathcal{T}_1 \subset \widetilde{\mathcal{I}}$ denotes the t groups with largest ℓ_2-norm of $\bar{\mathbf{u}}$, $\mathcal{T}_2$ the t next largest groups, and so on, we find with the triangle inequality and the group-RIP condition of $\tilde{\mathbf{A}}$ that

$$\begin{aligned}\left\|\tilde{\mathbf{A}}\bar{\mathbf{u}}\right\|_1 &\leq \sum_{i\geq 1}\left\|\tilde{\mathbf{A}}\bar{\mathbf{u}}_{\mathcal{T}_i}\right\|_1 \leq (1+\delta_t)\left(\|\bar{\mathbf{u}}_{\mathcal{T}_1}\|_2 + \sum_{i\geq 2}\|\bar{\mathbf{u}}_{\mathcal{T}_i}\|_2\right)\\ &\leq (1+\delta_t)\left(\|\bar{\mathbf{u}}\|_2 + \frac{\|\bar{\mathbf{u}}\|_{\widetilde{\mathcal{I}},1}}{\sqrt{t}}\right)\\ &\leq (1+\delta_t)\|\bar{\mathbf{u}}\|_2 + (1+\delta_t)\frac{\sqrt{s+1}}{\sqrt{t}}\sqrt{1+\frac{r^2}{\theta^2}}\end{aligned}$$

where again we invoked (4.8), followed by (4.23) in the last step. In combination with the lower bound on $\|\tilde{\mathbf{A}}\bar{\mathbf{u}}\|_1$ and solving for $\|\bar{\mathbf{u}}\|_2$, this yields

$$\|\bar{\mathbf{u}}\|_2 \geq \frac{1-\delta_t}{1+\delta_t} - \frac{\sqrt{s+1}}{\sqrt{t}}\sqrt{1+\frac{r^2}{\theta^2}}.$$

With (4.23), we arrive at

$$\frac{\|\bar{\mathbf{u}}\|_{\widetilde{\mathcal{I}},1}}{\|\bar{\mathbf{u}}\|_2} \leq \frac{\sqrt{s+1}}{\frac{1-\delta_t}{1+\delta_t} - \frac{\sqrt{s+1}}{\sqrt{t}}\sqrt{1+\frac{r^2}{\theta^2}}}\sqrt{1+\frac{r^2}{\theta^2}}.$$

Substituting in our choice for t, it then follows that the vector $\bar{\mathbf{u}}$ is effectively t-group-sparse, which concludes the proof. □

Remark 4.27. *Note that the previous result only holds for $\mathring{\mathbf{x}}$ belonging to $\mathcal{E}_{\mathcal{I},s} \cap r\mathbb{B}_2^d$ (or $\Sigma_{\mathcal{I},s} \cap r\mathbb{B}_2^d$) rather than the more general signal set $r\sqrt{s}\mathbb{B}_{\mathcal{I},1}^d \cap r\mathbb{B}_2^d$. This means that $\mathring{\mathbf{u}}$ is not just effectively κ-group-sparse with $\kappa := (s+1)(1+r^2/\theta^2)$ but actually $(s+1)$-group-sparse since*

$$\frac{\|\mathring{\mathbf{u}}\|_{\widetilde{\mathcal{I}},1}}{\|\mathring{\mathbf{u}}\|_2} = \frac{\|\mathring{\mathbf{x}}\|_{\mathcal{I},1} + \theta}{\sqrt{\|\mathring{\mathbf{x}}\|_2^2 + \theta^2}} \leq \frac{\sqrt{s}\|\mathring{\mathbf{x}}\|_2 + \theta^2}{\sqrt{\|\mathring{\mathbf{x}}\|_2^2 + \theta^2}} \leq \sqrt{s+1}$$

by Cauchy-Schwarz. However, since the same does not hold for $\hat{\mathbf{u}}$, the proof of Lemma 4.26 simplifies if one instead uses that both $\mathring{\mathbf{u}}$ and $\hat{\mathbf{u}}$ are effectively κ-group-sparse.

To establish the recovery guarantee of Problem ($\mathrm{P}_{4.8}$), we will also need the following technical lemma from [Bar+17a], which will also be crucial to establish several other recovery guarantees in this chapter.

Lemma 4.28 ([Bar+17a, Lemma 8])**.** *Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$ with $\alpha, \beta \neq 0$. Denote by $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$ the extended versions of $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^{n+1}$ with α and β as their last coordinates, respectively. Then*

$$\left\|\frac{\mathbf{a}}{\alpha} - \frac{\mathbf{b}}{\beta}\right\|_2 \leq \frac{\|\tilde{\mathbf{a}}\|_2 \cdot \|\tilde{\mathbf{b}}\|_2}{|\alpha| \cdot |\beta|}\left\|\frac{\tilde{\mathbf{a}}}{\|\tilde{\mathbf{a}}\|_2} - \frac{\tilde{\mathbf{b}}}{\|\tilde{\mathbf{b}}\|_2}\right\|_2.$$

Lemma 4.29. *Let $\mathring{\mathbf{x}} \in \mathcal{E}_{\mathcal{I},s}$ with $\|\mathring{\mathbf{x}}\|_2 \leq r$, and assume $\tilde{\mathbf{A}} = (\mathbf{A} \quad \mathbf{g}) \in \mathbb{R}^{m\times(d+1)}$ satisfies the effective group-RIP of order t with constant t for $t > s$ chosen according to (4.22). Given measurements of the form $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$ with $\boldsymbol{\tau} = \theta\mathbf{g}$, every minimizer $\hat{\mathbf{x}}$ of Problem ($\mathrm{P}_{4.8}$) satisfies*

$$\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \leq 4\sqrt{\delta_t}\frac{r^2+\theta^2}{\theta}.$$

Proof. We immediately invoke Lemma 4.28 for the vectors $\mathring{\mathbf{x}}$ and $\hat{\mathbf{x}}$ with $\alpha = \beta = \theta$. With the naming convention established in (4.21), this yields

$$\begin{aligned}\left\|\frac{\mathring{\mathbf{x}}}{\theta} - \frac{\hat{\mathbf{x}}}{\theta}\right\|_2 &\leq \frac{\|\mathring{\mathbf{u}}\|_2 \cdot \|\hat{\mathbf{u}}\|_2}{\theta^2}\left\|\frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2}\right\|_2 \\ &= \frac{\sqrt{\|\mathring{\mathbf{x}}\|_2^2+\theta^2}\cdot\sqrt{\|\hat{\mathbf{x}}\|_2^2+\theta^2}}{\theta^2}\left\|\frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2}\right\|_2 \\ &\leq \frac{r^2+\theta^2}{\theta^2}\left\|\frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2}\right\|_2. \end{aligned} \tag{4.24}$$

For the right-hand side, we find by the parallelogram identity that

$$\left\|\frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2}\right\|_2^2 = 4\left(1 - \left\|\frac{1}{2}\left(\frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} + \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2}\right)\right\|_2^2\right).$$

Since the argument of the ℓ_2-norm on the right-hand side is a convex combination with $\lambda = 1/2$ as required by Lemma 4.26, the (ℓ_2,ℓ_1)-group-RIP condition of $\tilde{\mathbf{A}}$ yields

$$\left\|\frac{1}{2}\left(\frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} + \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2}\right)\right\|_2 \geq \frac{\left\|\tilde{\mathbf{A}}\left(\frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} + \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2}\right)\right\|_1}{2(1+\delta_t)} \geq \frac{\left\|\tilde{\mathbf{A}}\frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2}\right\|_1 + \left\|\tilde{\mathbf{A}}\frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2}\right\|_1}{2(1+\delta_t)} \geq \frac{1-\delta_t}{1+\delta_t}.$$

Rearranging (4.24) for $\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2$ and combining the previous estimates, we finally arrive at

$$\begin{aligned}\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 &\leq 2\frac{r^2+\theta^2}{\theta}\sqrt{1 - \left(\frac{1-\delta_t}{1+\delta_t}\right)^2} \\ &= 2\frac{r^2+\theta^2}{\theta}\sqrt{\frac{(1+\delta_t)^2 - (1-\delta_t)^2}{(1+\delta_t)^2}} \\ &= 4\frac{r^2+\theta^2}{\theta}\frac{\sqrt{\delta_t}}{1+\delta_t} \\ &\leq 4\frac{r^2+\theta^2}{\theta}\sqrt{\delta_t}. \end{aligned}$$

The claim follows. □

With Lemma 4.29 established, we are now prepared to state a corresponding probabilistic recovery result for Problem ($\mathrm{P}_{4.8}$).

Theorem 4.30. *Let* $(\mathbf{A} \quad \mathbf{g}) \in \mathbb{R}^{m\times(d+1)}$ *be a standard Gaussian random matrix. Then the following holds with probability at least* $1-\eta$*: given a vector* $\mathring{\mathbf{x}} \in \mathcal{E}_{\mathcal{I},s} \cap r\mathbb{B}_2^d$ *and its measurements* $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \theta\mathbf{g})$*, every minimizer* $\hat{\mathbf{x}}$ *of Problem* (P$_{4.8}$) *satisfies*

$$\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \leq \varepsilon,$$

provided that

$$m \gtrsim \varepsilon^{-4}\left(\frac{r^2+\theta^2}{\theta}\right)^4\left(1+\frac{r^2}{\theta^2}\right)\left[s\log(G/s) + sg + \log(\eta^{-1})\right]$$

with

$$\varepsilon \leq 2\sqrt{2}\frac{r^2+\theta^2}{\theta}.$$

Proof. According to Lemma 4.29, the recovery error of Problem (P$_{4.8}$) is bounded by

$$\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \leq 4\sqrt{\delta_t}\frac{r^2+\theta^2}{\theta} =: \varepsilon, \tag{4.25}$$

provided that $\tilde{\mathbf{A}}$ satisfies the effective group-RIP on $\mathcal{E}_{\mathcal{I},t}$ with constant δ_t and

$$t = 4(s+1)\left(1+\frac{r^2}{\theta^2}\right)\left(\frac{1+\delta_t}{1-\delta_t}\right)^2.$$

With Lemma 4.13, this event occurs with probability at least $1-\eta$ if

$$m \gtrsim \delta_t^{-2}\left[t\log(G/t) + tg + \log(\eta^{-1})\right].$$

Solving for δ_t in (4.25) and substituting it into the expression for t means that t is upper bounded by

$$t = 36(s+1)\left(1+\frac{r^2}{\theta^2}\right)$$

if $\varepsilon \leq 2\sqrt{2}(r^2+\theta^2)/\theta$ since in that case

$$\frac{1+\delta_t}{1-\delta_t} = \frac{16(r^2+\theta^2)^2 + \theta^2\varepsilon^2}{16(r^2+\theta^2)^2 - \theta^2\varepsilon^2} \leq \frac{24(r^2+\theta^2)^2}{8(r^2+\theta^2)^2} = 3.$$

This implies that

$$m \gtrsim \varepsilon^{-4}\left(\frac{r^2+\theta^2}{\theta}\right)^4\left(1+\frac{r^2}{\theta^2}\right)\left[s\log(G/s) + sg + \log(\eta^{-1})\right]$$

measurements suffice for the conclusion of Theorem 4.30 to hold with probability at least $1-\eta$. □

Remark 4.31. *The optimal choice* $\theta = r$ *implies that* m *scales with* r^4*. This dependence on* r *if the standard deviation* θ *of the quantization thresholds is chosen on the order of* r *is common to most recovery guarantees we will discuss in the remainder of this chapter.*

Quantization-Consistent Recovery via Lifting

As previously announced, we now turn to the analysis of the following recovery procedure based on a lifting reformulation of Problem ($P_{4.3}$) for dithered measurements. In particular, we consider the program

$$\begin{aligned} \underset{\mathbf{x},w}{\text{minimize}} \quad & \|\mathbf{x}\|_{\mathcal{I},1} + |w| \\ \text{s.t.} \quad & \mathbf{y} = \operatorname{sgn}\left(\mathbf{A}\mathbf{x} + \tfrac{w}{\theta}\boldsymbol{\tau}\right) \\ & \left\|\mathbf{A}\mathbf{x} + \tfrac{w}{\theta}\boldsymbol{\tau}\right\|_1 = 1. \end{aligned} \tag{$P_{4.9}$}$$

Note that in the undithered setting, the constant on the right-hand side of the last constraint was ultimately arbitrary and only chosen as 1 for convenience of analysis. However, without the additional variable w, the above problem could potentially become infeasible if the constraint were replaced by $\|\mathbf{A}\mathbf{x} + \boldsymbol{\tau}\|_1 = 1$ as it would indirectly impose a norm on minimizers $\hat{\mathbf{x}}$. The variable w is therefore used to remedy this restriction such that no explicit constraint involving an estimate of the radius r of the ℓ_2-ball which contains $\mathring{\mathbf{x}}$ is necessary. Given a minimizer $(\hat{\mathbf{x}}, \hat{w})$ of Problem ($P_{4.9}$), we associate with it for convenience the recovery map implicitly defined as

$$\Delta_{\mathcal{I}}^{\mathrm{PV}}(\mathbf{y}) := \frac{\theta}{\hat{w}}\hat{\mathbf{x}}. \tag{4.26}$$

The following result establishes a reconstruction quality of $\Delta_{\mathcal{I}}^{\mathrm{PV}}$ conditioned on the effective group-RIP.

Lemma 4.32. *Let $\mathring{\mathbf{x}} \in \mathcal{E}_{\mathcal{I},s}$ with $\|\mathring{\mathbf{x}}\|_2 \leq r$, and assume that $\tilde{\mathbf{A}} = (\mathbf{A} \quad \mathbf{g}) \in \mathbb{R}^{m\times(d+1)}$ satisfies the effective group-RIP on $\mathcal{E}_{\mathcal{I},t}$ with constant $\delta/1024$ for some $\delta \in (0,1)$ and*

$$t = 4(s+1)\left(\frac{1+\delta}{1-\delta}\right)^2.$$

Given measurements of the form $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$ with $\boldsymbol{\tau} = \theta\mathbf{g}$, every minimizer $(\hat{\mathbf{x}}, \hat{w})$ of Problem ($P_{4.9}$) satisfies

$$\left\|\mathring{\mathbf{x}} - \frac{\theta}{\hat{w}}\hat{\mathbf{x}}\right\|_2 \leq \theta\sqrt{\delta},$$

provided that

$$\delta \leq \frac{(3\theta^2 - r^2)^2}{\theta^2(r^2+\theta^2)}$$

with $\theta > r/\sqrt{3}$.

Proof. The proof proceeds along the lines of the proof of Theorem 9 in [Bar+17a]. Given an optimal solution $(\hat{\mathbf{x}}, \hat{w})$, we begin by defining the vectors

$$\hat{\mathbf{u}} = \begin{pmatrix} \hat{\mathbf{x}} \\ \hat{w} \end{pmatrix} \quad \text{and} \quad \mathring{\mathbf{u}} = \begin{pmatrix} \mathring{\mathbf{x}} \\ \theta \end{pmatrix}.$$

We then have by feasibility of $(\hat{\mathbf{x}}, \hat{w})$ for Problem ($\mathrm{P}_{4.9}$) that

$$\mathbf{y} = \operatorname{sgn}\left(\mathbf{A}\hat{\mathbf{x}} + \frac{\hat{w}}{\theta}\boldsymbol{\tau}\right) = \operatorname{sgn}(\mathbf{A}\hat{\mathbf{x}} + \hat{w}\mathbf{g}) = \operatorname{sgn}(\tilde{\mathbf{A}}\hat{\mathbf{u}})$$

and

$$\left\|\mathbf{A}\hat{\mathbf{x}} + \frac{\hat{w}}{\theta}\boldsymbol{\tau}\right\|_1 = \left\|\tilde{\mathbf{A}}\hat{\mathbf{u}}\right\|_1 = 1.$$

This means that the vector $\hat{\mathbf{u}}$ is feasible for Problem ($\mathrm{P}_{4.3}$). Next, note that the normalized augmented vector $\mathring{\mathbf{u}}/\|\mathring{\mathbf{u}}\|_2$ is also feasible for Problem ($\mathrm{P}_{4.3}$). Since by Remark 4.27, the vector $\mathring{\mathbf{u}}$ is effectively $(s+1)$-group-sparse w. r. t. the augmented group partition $\widetilde{\mathcal{I}} = \mathcal{I} \cup \{\{d+1\}\}$ and the matrix $\tilde{\mathbf{A}}$ satisfies the effective group-RIP of order t with constant $\delta/1024$, Lemma 4.6 and Remark 4.7 imply that

$$\left\|\frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2}\right\|_2 \leq 8\sqrt{\frac{\delta}{1024}} = \frac{\sqrt{\delta}}{4},$$

which in turn yields

$$\left|\frac{\theta}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{w}}{\|\hat{\mathbf{u}}\|_2}\right| \leq \frac{\sqrt{\delta}}{4}.$$

An application of the reverse triangle inequality therefore implies

$$\frac{|\hat{w}|}{\|\hat{\mathbf{u}}\|_2} \geq \frac{\theta}{\|\mathring{\mathbf{u}}\|_2} - \frac{\sqrt{\delta}}{4} \geq \frac{\theta}{\sqrt{r^2+\theta^2}} - \frac{\sqrt{\delta}}{4} = \frac{4\theta - \sqrt{\delta}\sqrt{r^2+\theta^2}}{4\sqrt{r^2+\theta^2}}.$$

With Lemma 4.28, it now follows that

$$\begin{aligned}\left\|\frac{\mathring{\mathbf{x}}}{\theta} - \frac{\frac{\theta}{\hat{w}}\hat{\mathbf{x}}}{\theta}\right\|_2 = \left\|\frac{\mathring{\mathbf{x}}}{\theta} - \frac{\hat{\mathbf{x}}}{\hat{w}}\right\|_2 &\leq \frac{\|\mathring{\mathbf{u}}\|_2 \cdot \|\hat{\mathbf{u}}\|_2}{\theta \cdot |\hat{w}|}\left\|\frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2}\right\|_2 \\ &\leq \frac{\sqrt{r^2+\theta^2}}{\theta}\frac{4\sqrt{r^2+\theta^2}}{4\theta - \sqrt{\delta}\sqrt{r^2+\theta^2}}\frac{\sqrt{\delta}}{4} \\ &= \sqrt{\delta}\frac{r^2+\theta^2}{\theta(4\theta - \sqrt{\delta}\sqrt{r^2+\theta^2})}.\end{aligned}$$

With our required assumption on δ, the fractional term is now bounded from above by 1. The recovery error of the estimator $\theta\hat{\mathbf{x}}/\hat{w}$ is consequently bounded by

$$\left\|\mathring{\mathbf{x}} - \frac{\theta}{\hat{w}}\hat{\mathbf{x}}\right\|_2 \leq \theta\sqrt{\delta},$$

which concludes the proof. □

Remark 4.33. *The condition on δ in Lemma 4.32 seems rather unwieldy. However, the condition turns out to be very mild. For instance, the optimal choice $\theta = r$ yields the trivial condition $\delta \leq 2$.*

Based on Lemma 4.32, we may now derive the following probabilistic recovery guarantee for minimizers of Problem ($\mathrm{P}_{4.9}$).

Theorem 4.34. *Let* $\tilde{\mathbf{A}} = (\mathbf{A} \quad \mathbf{g}) \in \mathbb{R}^{m\times(d+1)}$ *be a standard Gaussian random matrix, and fix*

$$\varepsilon \leq \min\left\{16\sqrt{2}\theta, \frac{3\theta^2 - r^2}{\sqrt{r^2+\theta^2}}\right\}.$$

Then with probability at least $1-\eta$*, every vector* $\mathring{\mathbf{x}} \in \mathcal{E}_{\mathcal{I},s} \cap r\mathbb{B}_2^d$ *can approximated from its measurements* $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \theta\mathbf{g})$ *by a minimizer* $(\hat{\mathbf{x}}, \hat{w})$ *of Problem* ($\mathrm{P}_{4.9}$) *such that*

$$\left\|\mathring{\mathbf{x}} - \frac{\theta}{\hat{w}}\hat{\mathbf{x}}\right\|_2 \leq \varepsilon,$$

provided that

$$m \gtrsim \varepsilon^{-4}\theta^4\left[s\log(G/s) + sg + \log(\eta^{-1})\right].$$

Proof. Given that the vector $\theta\hat{\mathbf{x}}/\hat{w}$ approximates $\mathring{\mathbf{x}}$ with accuracy $\theta\sqrt{\delta}$ (in terms of the Euclidean distance) if $\tilde{\mathbf{A}}$ satisfies the effective group-RIP of order

$$t = 4(s+1)\left(\frac{1+\delta}{1-\delta}\right)^2,$$

it suffices to choose m large enough for this event to occur with the desired probability. To that end, we set $\delta = (\varepsilon/\theta)^2$ in Lemma 4.32. By Lemma 4.13, we then have with probability $1-\eta$ that the matrix $m^{-1}\sqrt{\pi/2}\tilde{\mathbf{A}}$ satisfies the effective group-RIP of order t with constant $\delta/\gamma = \varepsilon^2/(\gamma\theta^2)$ for $\gamma = 1024$, provided that

$$\begin{aligned} m &\gtrsim \left(\frac{\delta}{\gamma}\right)^{-2}\left(\frac{1+\delta/\gamma}{1-\delta/\gamma}\right)^2\left[(s+1)\log\left(\frac{G}{s+1}\right) + sg + \log(\eta^{-1})\right] \\ &= \gamma^2\varepsilon^{-4}\theta^4\left(\frac{\gamma\theta^2+\varepsilon^2}{\gamma\theta^2-\varepsilon^2}\right)^2\left[(s+1)\log\left(\frac{G}{s+1}\right) + (s+1)g + \log(\eta^{-1})\right]. \end{aligned}$$

With the condition $\varepsilon \leq \theta\sqrt{\gamma/2} = 16\sqrt{2}\theta$, the fractional term is bounded by 9 so that the conclusion of Theorem 4.34 holds with probability at least $1-\eta$ if

$$\begin{aligned} m &\gtrsim \varepsilon^{-4}\theta^4\left[s\log\left(\frac{G}{s}\right) + sg + \log(\eta^{-1})\right] \\ &\gtrsim \varepsilon^{-4}\theta^4\left[(s+1)\log\left(\frac{G}{s+1}\right) + (s+1)g + \log(\eta^{-1})\right] \end{aligned}$$

where we absorbed the constant γ^2 in the notation. □

The implicit constant γ^2 in the number of measurements is certainly not optimal and may be substantially improved by other techniques. It is ultimately rooted in the proof technique employed for Lemma 4.6, as well as Remark 4.7. However, the asymptotic behavior of m in terms of $\varepsilon, \theta, r, s, g$ and G is as expected by other techniques. One way to circumvent this issue is by following the same strategy as in the proof of Lemma 4.16 and require $\tilde{\mathbf{A}}$ to simultaneously satisfy the group-RIP and ε-tessellation property. The same strategy was also employed in [**Bar+17a**] to establish the recovery of dictionary-sparse vectors from random dithered measurements.

Lemma 4.35. *Let $\mathring{\mathbf{x}} \in \mathcal{E}_{\mathcal{I},s} \cap r\mathbb{B}_2^d$, and assume $\tilde{\mathbf{A}} \in \mathbb{R}^{m\times(d+1)}$ satisfies the group-RIP on $\mathcal{E}_{\mathcal{I},t}$ with constant δ and*

$$t = 4(s+1)\left(\frac{1+\delta}{1-\delta}\right)^2.$$

Assume further that $\tilde{\mathbf{A}}$ induces an $(\varepsilon/4)$-tessellation on $\widetilde{\mathcal{E}}_{\mathcal{I},t}$. Then with

$$\varepsilon \leq \frac{3\theta^2 - r^2}{\theta\sqrt{r^2+\theta^2}},$$

every minimizer $(\hat{\mathbf{x}}, \hat{w})$ of Problem (P4.9) satisfies

$$\left\| \mathring{\mathbf{x}} - \frac{\theta}{\hat{w}}\hat{\mathbf{x}} \right\|_2 \leq \theta\varepsilon.$$

Proof. We begin as in the previous proof and note that the normalized augmented vector $\mathring{\mathbf{u}}/\|\mathring{\mathbf{u}}\|_2$ is effectively $(s+1)$-group-sparse w. r. t. $\widetilde{\mathcal{I}} = \mathcal{I} \cup \{\{d+1\}\}$. Since $(\hat{\mathbf{x}}, \hat{w})$ is a minimizer of Problem (P4.9), it is also a minimizer of Problem (P4.3). Moreover, $\tilde{\mathbf{A}}$ simultaneously satisfies the group-RIP and the ε-tessellation property on $\mathcal{E}_{\mathcal{I},t}$, which implies with Lemma 4.16 that

$$\left\| \frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2} \right\|_2 \leq \frac{\varepsilon}{4}.$$

From here on, the proof is identical to the proof of Lemma 4.32 with $\sqrt{\delta}$ replaced by ε. □

The following result now establishes that every vector $\mathring{\mathbf{x}} \in \mathcal{E}_{\mathcal{I},s} \cap r\mathbb{B}_2^d$ can be ε-estimated by minimizers of Problem (P4.5) with high probability on the draw of the measurement matrix $\mathbf{A}$ and the dithering vector $\boldsymbol{\tau}$. The result follows by combining Lemma 4.35 and Theorem 4.17, using the same arguments as in the proof of Theorem 4.19. As in the last application of the tessellation property, the dependence of m improves from ε^{-4} to ε^{-3}. Additionally, the technique also allows us to reduce the scaling w. r. t. r from r^4 as before to r^3.

Theorem 4.36. *Let $\mathbf{A} \in \mathbb{R}^{m\times d}$ be a standard Gaussian random matrix and denote by $\boldsymbol{\tau} \sim \mathsf{N}(\mathbf{0}, \theta^2\mathrm{Id}_m)$ a random dithering vector. Fix a value $\varepsilon \leq (3\theta^2 - r^2)/(\theta\sqrt{\theta^2+r^2})$, and assume that*

$$m \gtrsim \varepsilon^{-3}\theta^3(s\log(G/s) + sg).$$

Then with probability at least

$$1 - c\exp\left(-c\left[s\log\left(\frac{G}{16s}\right) + sg\right]\right),$$

the following holds: every vector $\mathring{\mathbf{x}} \in \mathcal{E}_{\mathcal{I},s} \cap r\mathbb{B}_2^d$ can be estimated from its measurements $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$ by minimizers $\hat{\mathbf{x}}$ of Problem (P4.3) with

$$\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \leq \varepsilon.$$

4.4.2 Correlation Maximization

Before moving on to the group hard thresholding approach, we first extend the correlation maximization strategy outlined in Section 4.3.2 to dithered measurements. In this context, we will discuss two different approaches. The first one is based on the familiar lifting technique which embeds the dithered measurement model into a $(d+1)$-dimensional undithered setting. The second approach instead uses a regularization technique to deal with issues arising in the proof of an appropriate variant of Theorem 4.23 when the target vector $\mathring{\mathbf{x}}$ no longer belongs to the Euclidean unit sphere. We begin with the lifting approach.

Recovery via Lifting

Motivated by the approach taken in Problem ($\mathrm{P}_{4.9}$), we consider the problem

$$\begin{aligned} \underset{\mathbf{x},w}{\text{maximize}} \quad & \left\langle \mathbf{y}, \mathbf{A}\mathbf{x} + w\frac{\boldsymbol{\tau}}{\theta} \right\rangle \\ \text{s.t.} \quad & \left\| \begin{pmatrix} \mathbf{x} \\ w \end{pmatrix} \right\|_{\widetilde{\mathcal{I}},1} \leq \sqrt{s+1}\sqrt{r^2+\theta^2} \\ & \left\| \begin{pmatrix} \mathbf{x} \\ w \end{pmatrix} \right\|_2 \leq \sqrt{r^2+\theta^2}. \end{aligned} \tag{$\mathrm{P}_{4.10}$}$$

Similar to the recovery map $\Delta_{\mathcal{I}}^{\mathrm{PV}}$ discussed in the previous section, we associate with maximizers $(\hat{\mathbf{x}}, \hat{w})$ of Problem ($\mathrm{P}_{4.10}$) the recovery map

$$\Delta_{\mathcal{I}}^{\mathrm{corr}}(\mathbf{y}) := \frac{\theta}{\hat{w}}\hat{\mathbf{x}}, \tag{4.27}$$

where the notation $\Delta_{\mathcal{I}}^{\mathrm{corr}}$ indicates that optimal solutions of Problem ($\mathrm{P}_{4.10}$) aim to maximize the correlation between quantized and unquantized observations. The motivation for the constraints is obvious from the previous discussion of the dithered measurement model: given a dithering vector $\boldsymbol{\tau} = \theta\mathbf{g}$, the extended vector

$$\mathring{\mathbf{u}} = \begin{pmatrix} \mathring{\mathbf{x}} \\ \theta \end{pmatrix}$$

has the same sign pattern under the map $\mathbf{u} \mapsto \operatorname{sgn}((\mathbf{A} \quad \mathbf{g})\mathbf{u})$ as the original measurements $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$ since

$$\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau}) = \operatorname{sgn}\left(\begin{pmatrix} \mathbf{A} & \frac{\boldsymbol{\tau}}{\theta} \end{pmatrix} \begin{pmatrix} \mathring{\mathbf{x}} \\ \theta \end{pmatrix} \right) = \operatorname{sgn}(\tilde{\mathbf{A}}\mathring{\mathbf{u}})$$

where we defined $\tilde{\mathbf{A}} := (\mathbf{A} \quad \mathbf{g})$. Moreover, by Remark 4.27, the vector $\mathring{\mathbf{u}}$ is effectively $(s+1)$-group-sparse w. r. t. the group partition $\widetilde{\mathcal{I}} = \mathcal{I} \cup \{\{d+1\}\}$ and $\|\mathring{\mathbf{u}}\|_2 = (\|\mathring{\mathbf{x}}\|_2^2 + \theta^2)^{1/2} \leq (r^2+\theta^2)^{1/2}$.

As before, the main idea in the analysis is to relate the recovery performance for the vector $\mathring{\mathbf{u}}$ to the performance of Problem ($\mathrm{P}_{4.5}$) in $d+1$ dimensions. The following recovery guarantee for Problem ($\mathrm{P}_{4.10}$) then follows from Theorem 4.23. We emphasize again that the considered signal set is actually the bigger set $\sqrt{s}r\mathbb{B}_{\mathcal{I},1}^d \cap r\mathbb{B}_2^d$ rather than $\mathcal{E}_{\mathcal{I},s} \cap r\mathbb{B}_2^d$. This is due to the fact that $\mathbf{x} \in \mathcal{E}_{\mathcal{I},s} \cap r\mathbb{B}_2^d$ is not a convex constraint

and can therefore not be directly imposed in a convex program. However, by the same arguments as in Section 4.3.3, the solution of Problem ($\mathrm{P}_{4.10}$) does not change if we replace the constraint by $(\mathbf{x}, w) \in \mathcal{E}_{\widetilde{\mathcal{I}},s+1} \cap \sqrt{r^2+\theta^2}\mathbb{B}_2^{d+1}$ since $\operatorname{conv}(\mathcal{E}_{\widetilde{\mathcal{I}},s+1} \cap \sqrt{r^2+\theta^2}\mathbb{B}_2^{d+1}) = \sqrt{s+1}\sqrt{r^2+\theta^2}\mathbb{B}_{\widetilde{\mathcal{I}},1}^{d+1} \cap \sqrt{r^2+\theta^2}\mathbb{B}_2^{d+1}$.

Theorem 4.37. *Let $\tilde{\mathbf{A}} = (\mathbf{A} \quad \mathbf{g}) \in \mathbb{R}^{m\times(d+1)}$ be a standard Gaussian random matrix, and denote by $\boldsymbol{\tau} = \theta\mathbf{g}$ a dithering vector. Given a vector $\mathring{\mathbf{x}} \in \sqrt{s}r\mathbb{B}_{\mathcal{I},1}^d \cap r\mathbb{B}_2^d$ and the quantized measurements $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$, any solution $(\hat{\mathbf{x}}, \hat{w})$ of Problem ($\mathrm{P}_{4.10}$) satisfies*

$$\left\| \mathring{\mathbf{x}} - \frac{\theta\hat{\mathbf{x}}}{\hat{w}} \right\|_2 \leq \varepsilon$$

with probability at least $1-\eta$, provided that

$$m \gtrsim \varepsilon^{-4}\left(\frac{r^2+\theta^2}{\theta}\right)^4 \Big[s\log(G/s) + sg + \log(\eta^{-1}) \Big]$$

and $\varepsilon \leq \sqrt{r^2+\theta^2}$.

Proof. First note that solving Problem ($\mathrm{P}_{4.10}$) is—up to scaling of the optimal point—equivalent to solving the problem

$$\begin{aligned} \underset{\mathbf{x},w}{\text{maximize}} \quad & \left\langle \mathbf{y}, \tilde{\mathbf{A}}\begin{pmatrix}\mathbf{x}\\ w\end{pmatrix} \right\rangle \\ \text{s.t.} \quad & \left\| \begin{pmatrix}\mathbf{x}\\ w\end{pmatrix} \right\|_{\widetilde{\mathcal{I}},1} \leq \sqrt{s+1} \\ & \left\| \begin{pmatrix}\mathbf{x}\\ w\end{pmatrix} \right\|_2 \leq 1. \end{aligned} \tag{$\mathrm{P}_{4.11}$}$$

In particular, any solution $(\hat{\mathbf{x}}', \hat{w}')$ of Problem ($\mathrm{P}_{4.11}$) implies that the vector

$$\hat{\mathbf{u}} := \begin{pmatrix}\hat{\mathbf{x}}\\ \hat{w}\end{pmatrix} = \sqrt{r^2+\theta^2}\begin{pmatrix}\hat{\mathbf{x}}'\\ \hat{w}'\end{pmatrix} =: \sqrt{r^2+\theta^2}\hat{\mathbf{u}}'$$

is a solution of Problem ($\mathrm{P}_{4.10}$). Next, recall that the measurement map $\mathbf{u} \mapsto \operatorname{sgn}(\tilde{\mathbf{A}}\mathbf{u})$ is invariant under positive scaling of its argument. We may therefore invoke Theorem 4.22 for the vectors $\mathring{\mathbf{u}}/\|\mathring{\mathbf{u}}\|_2$ and $\hat{\mathbf{u}}'/\|\hat{\mathbf{u}}'\|_2$ in combination with (4.14), which immediately implies that with probability at least $1-\eta$, we have

$$\left\| \frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}'}{\|\hat{\mathbf{u}}'\|_2} \right\|_2 = \left\| \frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2} \right\|_2 \leq \tilde{\varepsilon}, \tag{4.28}$$

provided that

$$m \gtrsim \tilde{\varepsilon}^{-4}\Big[(s+1)\log(G/(s+1)) + (s+1)g + \log(\eta^{-1})\Big].$$

In particular, if the event in (4.28) occurs, the bound implies for the last coordinate that

$$\left| \frac{\theta}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{w}}{\|\hat{\mathbf{u}}\|_2} \right| \leq \tilde{\varepsilon}$$

and therefore by the reverse triangle inequality that

$$\frac{|\hat{w}|}{\|\hat{\mathbf{u}}\|_2} \geq \frac{\theta}{\|\mathring{\mathbf{u}}\|_2} - \tilde{\varepsilon} = \frac{\theta - \tilde{\varepsilon}\|\mathring{\mathbf{u}}\|_2}{\|\mathring{\mathbf{u}}\|_2} \geq \frac{\theta - \tilde{\varepsilon}\sqrt{r^2+\theta^2}}{\sqrt{r^2+\theta^2}}.$$

An application of Lemma 4.28 thus yields

$$\begin{aligned}\left\|\mathring{\mathbf{x}} - \frac{\theta\hat{\mathbf{x}}}{\hat{w}}\right\|_2 &= \theta\left\|\frac{\mathring{\mathbf{x}}}{\theta} - \frac{\hat{\mathbf{x}}}{\hat{w}}\right\|_2 \leq \theta\frac{\|\mathring{\mathbf{u}}\|_2}{\theta}\frac{\|\hat{\mathbf{u}}\|_2}{|\hat{w}|}\left\|\frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2}\right\|_2 \\ &\leq \sqrt{r^2+\theta^2}\frac{\tilde{\varepsilon}\sqrt{r^2+\theta^2}}{\theta - \tilde{\varepsilon}\sqrt{r^2+\theta^2}}.\end{aligned}$$

Choosing $\tilde{\varepsilon} = \varepsilon'/\sqrt{r^2+\theta^2}$ and requiring that $\varepsilon' \leq \theta/2$, we arrive at

$$\left\|\mathring{\mathbf{x}} - \frac{\theta\hat{\mathbf{x}}}{\hat{w}}\right\|_2 \leq \frac{\varepsilon'\sqrt{r^2+\theta^2}}{\theta - \varepsilon'} \leq \frac{2\varepsilon'\sqrt{r^2+\theta^2}}{\theta} \overset{!}{=} \varepsilon.$$

Eliminating $\tilde{\varepsilon}$ and ε', this means that for $\varepsilon \leq \sqrt{r^2+\theta^2}$, any maximizer $(\hat{\mathbf{x}}, \hat{w})$ of ($\mathrm{P}_{4.10}$) satisfies

$$\left\|\mathring{\mathbf{x}} - \frac{\theta\hat{\mathbf{x}}}{\hat{w}}\right\|_2 \leq \varepsilon$$

with probability at least $1-\eta$, provided that

$$m \gtrsim \varepsilon^{-4}\left(\frac{r^2+\theta^2}{\theta}\right)^4\left[s\log(G/s) + sg + \log(\eta^{-1})\right].$$

□

Remark 4.38. *(i) It is easy to check that the choice $\theta = r$ is optimal in terms of the number of measurements. This particular choice for the standard deviation of the quantization thresholds therefore implies that m scales with r^4, which is slightly worse than what was established in Theorem 4.36.*

(ii) While the formulation of Problem ($\mathrm{P}_{4.10}$) appears to require exact knowledge of the radius r of the ℓ_2-ball which contains $\mathring{\mathbf{x}}$ to estimate $\mathring{\mathbf{x}}$, this is in fact not the case. As pointed out in the proof of Theorem 4.37, solving Problem ($\mathrm{P}_{4.10}$) is equivalent to solving Problem ($\mathrm{P}_{4.11}$) up to appropriate rescaling of the optimal point. However, since we use $\theta\hat{\mathbf{x}}/\hat{w}$ as estimator for $\mathring{\mathbf{x}}$, the scaling constant is ultimately irrelevant as it gets canceled out in the division by $\hat{w}$. In other words, it suffices to solve Problem ($\mathrm{P}_{4.11}$) for which no explicit estimate of r is required. Nevertheless, as mentioned in Remark 4.38(i), we would ideally choose $\theta = r$ to obtain optimal error decay in terms of m. The original constraint would therefore still be justified under the assumption that an estimate of r is available in any real-world application.

As before, we may extend the previous result to the noisy observation model

$$\mathbf{y} = Q(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau}) = \mathbf{f} \circ \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau} + \boldsymbol{\nu}) \tag{4.29}$$

where as usual $\boldsymbol{\nu} \sim \mathsf{N}(\mathbf{0}, \sigma^2 \mathrm{Id}_m)$ and $\mathbf{f} \sim \mathsf{B}_m(p)$ are independent pre- and post-quantization noise vectors, respectively. However, in the analysis of the associated recovery program we now have to consider the scaled noise vector $\boldsymbol{\nu}/\|\mathring{\mathbf{u}}\|_2$ to mimic measurement of the normalized vector $\mathring{\mathbf{u}}/\|\mathring{\mathbf{u}}\|_2$. More precisely, if $\hat{\mathbf{u}}$ denotes an optimal solution of Problem ($\mathrm{P}_{4.10}$) under the noisy observation model, then we have as before that the normalized vector $\mathring{\mathbf{u}}/\|\mathring{\mathbf{u}}\|_2$ is feasible and $\hat{\mathbf{u}}' := \hat{\mathbf{u}}/\sqrt{r^2+\theta^2}$ is optimal for the program

$$\begin{array}{ll} \underset{\mathbf{u}}{\text{maximize}} & \left\langle \mathbf{f} \circ \operatorname{sgn}(\tilde{\mathbf{A}}\mathring{\mathbf{u}} + \boldsymbol{\nu}), \tilde{\mathbf{A}}\mathbf{u} \right\rangle \\ \text{s.t.} & \|\mathbf{u}\|_{\widetilde{\mathcal{I}},1} \le \sqrt{s+1} \\ & \|\mathbf{u}\|_2 \le 1. \end{array} \tag{$\mathrm{P}_{4.12}$}$$

In order to invoke Theorem 4.23 for the vectors $\mathring{\mathbf{u}}/\|\mathring{\mathbf{u}}\|_2$ and $\hat{\mathbf{u}}'/\|\hat{\mathbf{u}}'\|_2$, however, we have to match the vector which generated the measurements $\mathbf{y}$ to $\mathring{\mathbf{u}}/\|\mathring{\mathbf{u}}\|$. In other words, when invoking Theorem 4.23, we have to interpret the objective function as

$$\left\langle \mathbf{f} \circ \operatorname{sgn}(\tilde{\mathbf{A}}\mathring{\mathbf{u}} + \boldsymbol{\nu}), \tilde{\mathbf{A}}\mathbf{u} \right\rangle = \left\langle \mathbf{f} \circ \operatorname{sgn}\left(\tilde{\mathbf{A}} \frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} + \frac{\boldsymbol{\nu}}{\|\mathring{\mathbf{u}}\|_2} \right), \tilde{\mathbf{A}}\mathbf{u} \right\rangle$$

and consider the scaled noise vector $\nu/\|\mathring{\mathbf{u}}\|_2 \sim \mathsf{N}(\mathbf{0}, \sigma^2/\|\mathring{\mathbf{u}}\|_2^2 \mathrm{Id}_m)$ in place of the random vector $\boldsymbol{\nu}$. With this change, Theorem 4.23 then implies that

$$\left\| \frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}'}{\|\hat{\mathbf{u}}'\|_2} \right\|_2 = \left\| \frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2} \right\|_2 \le \tilde{\varepsilon}$$

with probability at least $1-\eta$ if

$$\begin{aligned} m &\gtrsim \tilde{\varepsilon}^{-4}(2p-1)^{-2}\left(\frac{\sigma^2}{\theta^2}+1\right)\left[s\log(G/s) + sg + \log(\eta^{-1})\right] \\ &\ge \tilde{\varepsilon}^{-4}(2p-1)^{-2}\left(\frac{\sigma^2}{\|\mathring{\mathbf{u}}\|_2^2}+1\right)\left[s\log(G/s) + sg + \log(\eta^{-1})\right]. \end{aligned}$$

From here on the argument follows the proof of Theorem 4.37. We therefore arrive at the following noise-robust version of the lifted recovery program.

Theorem 4.39. *Let $\mathbf{A}$ and $\boldsymbol{\tau}$ be as defined in Theorem 4.37. Given a vector $\mathring{\mathbf{x}} \in \sqrt{s}r\mathbb{B}_{\mathcal{I},1}^d \cap r\mathbb{B}_2^d$ and noisy quantized measurements of the form $\mathbf{y} = Q(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$ according to (4.29), any solution $(\hat{\mathbf{x}}, \hat{w})$ of Problem ($\mathrm{P}_{4.10}$) satisfies*

$$\left\| \mathring{\mathbf{x}} - \frac{\theta\hat{\mathbf{x}}}{\hat{w}} \right\|_2 \le \varepsilon$$

with probability at least $1-\eta$, provided that

$$m \gtrsim \varepsilon^{-4}(2p-1)^{-2}\left(\frac{r^2+\theta^2}{\theta}\right)^4\left(\frac{\sigma^2}{\theta^2}+1\right)\left[s\log(G/s) + sg + \log(\eta^{-1})\right]$$

and $\varepsilon \le \sqrt{r^2+\theta^2}$.

Remark 4.40. *Unlike in the noiseless case, the optimal choice for θ is not as straightforward as before. However, it is easy to verify that the bound on m depends on θ in a convex fashion and that it attains its optimum at*

$$\begin{aligned}\theta &= \frac{1}{2}\sqrt{2r^2 - \sigma^2 + \sqrt{\sigma^4 + 4r^4 + 20\sigma^2 r^2}} \\ &\geq \frac{1}{2}\sqrt{2r^2 - \sigma^2 + \sqrt{\sigma^4 + 4r^4 + 4\sigma^2 r^2}} \\ &= r.\end{aligned}$$

Due to convexity of the map

$$\theta \mapsto \left(\frac{r^2 + \theta^2}{\theta}\right)^4 \left(\frac{\sigma^2}{\theta^2} + 1\right),$$

this means that

$$m = \Omega\left(\varepsilon^{-4}(2p-1)^{-2} r^4 \left(\frac{\sigma^2}{r^2} + 1\right)\left[s\log(G/s) + sg + \log(\eta^{-1})\right]\right)$$

measurements suffice to accurately recover effectively s-group-sparse vectors contained in $r\mathbb{B}_2^d$ via Problem ($\mathrm{P}_{4.10}$) from noisy observations. In that sense, the term σ^2/r^2 in the choice for m roughly acts as a reciprocal signal-to-noise ratio *(SNR), i.e., for a fixed r and accuracy ε, the number of measurements required to obtain an ε-accurate reconstruction increases as r^2/σ^2 decreases.*

Regularized Recovery via Projection

Given the quantized measurements $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$, we now aim to maximize the correlation $\langle \mathbf{y}, \mathbf{A}\mathbf{x} + \boldsymbol{\tau}\rangle = \langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle + \langle \mathbf{y}, \boldsymbol{\tau}\rangle$ between the quantized and unquantized observations. Since $\langle \mathbf{y}, \boldsymbol{\tau}\rangle$ is constant w. r. t. $\mathbf{x}$, one option would be to solve the problem

$$\begin{aligned}\underset{\mathbf{x}}{\text{maximize}} \quad & \langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle \\ \text{s.t.} \quad & \|\mathbf{x}\|_{\mathcal{I},1} \leq \sqrt{s} r \\ & \|\mathbf{x}\|_2 \leq r.\end{aligned}$$

However, the proof technique employed in [PV13b] breaks down when trying to adopt it to the dithered observation setting. We will therefore consider the following ℓ_2-regularized problem instead, which was recently proposed by Dirksen and Mendelson in [DM18a] to make the problem amenable to the proof technique of [PV13b]. In particular, we solve the problem

$$\begin{aligned}\underset{\mathbf{x}}{\text{maximize}} \quad & \langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle - \mu\|\mathbf{x}\|_2^2 \\ \text{s.t.} \quad & \|\mathbf{x}\|_{\mathcal{I},1} \leq \sqrt{s} r \\ & \|\mathbf{x}\|_2 \leq r\end{aligned} \tag{$\mathrm{P}_{4.13}$}$$

where $\mu > 0$ is a regularization parameter. Due to the fact that the objective function contains both a quadratic and linear term in $\mathbf{x}$, the problem above can be rewritten in

terms of the orthogonal projector $\Pi_{\mathcal{K}}$ on the set $\mathcal{K} = \sqrt{s}r\mathbb{B}_{\mathcal{I},1} \cap r\mathbb{B}_2^d$ (see also Section 4.4.3). For this reason, we also denote maximizers of Problem ($\mathrm{P}_{4.13}$) by

$$\Delta_{\mathcal{I}}^{\Pi} := \operatorname*{argsup}_{\mathbf{x}} \Big\{ \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle - \mu \|\mathbf{x}\|_2^2 : \mathbf{x} \in \sqrt{s}r\mathbb{B}_{\mathcal{I},1}^d \cap r\mathbb{B}_2^d \Big\}.$$

The following result establishes a recovery guarantee for maximizers of Problem ($\mathrm{P}_{4.13}$) from dithered observations.

Theorem 4.41. *Let $\mathring{\mathbf{x}} \in \mathbb{R}^d$ with $\|\mathring{\mathbf{x}}\|_{\mathcal{I},1} \leq \sqrt{s}r$ and $\|\mathring{\mathbf{x}}\|_2 \leq r$. Let further $\mathbf{A} \in \mathbb{R}^{m \times d}$ be a standard Gaussian random matrix, and denote by $\boldsymbol{\tau} \sim \mathsf{N}(\mathbf{0}, \theta^2 \mathrm{Id}_m)$ a random vector independent from $\mathbf{A}$. Then with probability at least $1-\eta$, every maximizer $\hat{\mathbf{x}}$ of Problem ($\mathrm{P}_{4.13}$) with $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$ and $\mu \geq m/\sqrt{2\pi\theta^2}$ satisfies*

$$\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \leq \varepsilon,$$

provided that

$$m \gtrsim \varepsilon^{-4}(\theta^2 + r^2)\Big[r^2(s\log(G/s) + sg) + \log(\eta^{-1})\Big].$$

Remark 4.42. *Choosing the threshold standard deviation θ on the order of r as usual, Theorem 4.41 establishes that effectively s-group-sparse vectors can be estimated up to a fidelity of ε from $m = \Omega(\varepsilon^{-4}r^4(s\log(G/s) + sg))$ measurements with probability at least $1 - \exp(-r^2(s\log(G/s) + sg))$. Note that this result exhibits worse scaling dependence of m on ε and r than Theorem 4.36. However, unlike in Theorem 4.36, the failure probability additionally decays exponentially in r^2.*

We point out that Problem ($\mathrm{P}_{4.13}$) does *not* require explicit knowledge of the quantization thresholds $\boldsymbol{\tau}$. According to Theorem 4.41, it suffices to have access to the threshold variance θ^2, which is needed to appropriately choose the regularization parameter μ. This is in stark contrast to the lifted reconstruction scheme ($\mathrm{P}_{4.10}$), where the exact threshold vector $\boldsymbol{\tau}$ has to be known at the decoder.

Note that the results in [**DM18a**] could also be used to derive a similar result to Theorem 4.41 in a more general setting where the rows of $\mathbf{A}$ are allowed to be isotropic subgaussian random vectors, and one can account for both pre- and post-quantization noise. However, the quantization thresholds in Theorem 1.7 in [**DM18a**] are uniformly distributed in the interval $[-1/(2\mu), 1/(2\mu)]$ rather than drawn from the Gaussian distribution where the regularization parameter μ depends on unspecified constants, which are hard to calculate explicitly. As a consequence, employing their recovery technique requires experimentation to appropriately tune the parameter μ, while in our setting the choice for μ is simple.

In general, the proof of Theorem 4.41 proceeds along the lines of [**PV13b**, Theorem 1.1]. We begin by defining the scaled loss function

$$\begin{aligned} L_{\mathring{\mathbf{x}}}(\mathbf{x}) &:= \frac{1}{m}\Big(\langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle - \mu\|\mathbf{x}\|_2^2\Big) \\ &= \frac{1}{m}\Big(\langle \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau}), \mathbf{A}\mathbf{x} \rangle - \mu\|\mathbf{x}\|_2^2\Big) \\ &= \frac{1}{m}\sum_{i=1}^{m}\Big(\operatorname{sgn}(\langle \mathbf{a}_i, \mathring{\mathbf{x}} \rangle + \tau_i)\langle \mathbf{a}_i, \mathbf{x} \rangle - \mu x_i^2\Big), \end{aligned}$$

where the index $\mathring{\mathbf{x}}$ indicates the dependence of $L_{\mathring{\mathbf{x}}}$ on $\mathring{\mathbf{x}}$ through $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$. We will then use a result which establishes that the random variable $L_{\mathring{\mathbf{x}}}(\mathbf{x})$ concentrates sharply around its mean as demonstrated in [PV13b]. To that end, we first calculate $L_{\mathring{\mathbf{x}}}(\mathbf{x})$ in expectation.

Lemma 4.43. *Let $\mathring{\mathbf{x}}, \mathbf{x} \in \mathbb{R}^d$. Then*

$$\mathbb{E} L_{\mathring{\mathbf{x}}}(\mathbf{x}) = \lambda_{\mathring{\mathbf{x}}} \langle \mathbf{x}, \mathring{\mathbf{x}} \rangle - \frac{\mu}{m} \|\mathbf{x}\|_2^2$$

with

$$\lambda_{\mathring{\mathbf{x}}} := \sqrt{\frac{2}{\pi(\theta^2 + \|\mathring{\mathbf{x}}\|_2^2)}}.$$

Proof. Since the rows of $\mathbf{A}$ are identically distributed and independent from $\boldsymbol{\tau}$, we immediately have

$$\mathbb{E} L_{\mathring{\mathbf{x}}}(\mathbf{x}) = \mathbb{E}[\operatorname{sgn}(\langle \mathbf{a}_i, \mathring{\mathbf{x}} \rangle + \tau_i) \langle \mathbf{a}_i, \mathbf{x} \rangle] - \frac{\mu}{m} \|\mathbf{x}\|_2^2$$

where $i \in [m]$ is an arbitrary index. Next, we decompose $\mathbf{x}$ into orthogonal components, *i.e.*, we write $\mathbf{x} = \langle \mathring{\mathbf{x}}, \mathbf{x} \rangle \mathring{\mathbf{x}} / \|\mathring{\mathbf{x}}\|_2^2 + \mathbf{x}^\perp$ with $\langle \mathbf{x}^\perp, \mathring{\mathbf{x}} \rangle = 0$. Substituting this decomposition into the previous expression for $\mathbb{E} L_{\mathring{\mathbf{x}}}(\mathbf{x})$ therefore yields

$$\begin{aligned}
\mathbb{E} L_{\mathring{\mathbf{x}}}(\mathbf{x}) &= \mathbb{E}\left[\operatorname{sgn}(\langle \mathbf{a}_i, \mathring{\mathbf{x}} \rangle + \tau_i)\left(\langle \mathbf{a}_i, \mathring{\mathbf{x}} \rangle \frac{\langle \mathring{\mathbf{x}}, \mathbf{x} \rangle}{\|\mathring{\mathbf{x}}\|_2^2} + \left\langle \mathbf{a}_i, \mathbf{x}^\perp \right\rangle\right)\right] - \frac{\mu}{m} \|\mathbf{x}\|_2^2 \\
&= \frac{\langle \mathring{\mathbf{x}}, \mathbf{x} \rangle}{\|\mathring{\mathbf{x}}\|_2^2} \mathbb{E}[g \operatorname{sgn}(g + \tau)] - \frac{\mu}{m} \|\mathbf{x}\|_2^2
\end{aligned}$$

where $g \sim \mathsf{N}(0, \|\mathring{\mathbf{x}}\|_2^2)$ is independent from $\tau \sim \mathsf{N}(0, \theta^2)$, and the last term in the parentheses disappears due to independence of $\langle \mathbf{a}_i, \mathring{\mathbf{x}} \rangle$ and $\langle \mathbf{a}_i, \mathbf{x}^\perp \rangle$ by orthogonality of $\mathring{\mathbf{x}}$ and $\mathbf{x}^\perp$. Conditioning on g, we now have

$$\begin{aligned}
\mathbb{E}_g[g \mathbb{E}_\tau \operatorname{sgn}(g + \tau)] &= \mathbb{E}_g[g(\mathbb{P}(g + \tau \geq 0) - \mathbb{P}(g + \tau < 0))] \\
&= \mathbb{E}_g[g \mathbb{P}(g + \tau \geq 0) - g(1 - \mathbb{P}(g + \tau \geq 0))] \\
&= 2 \mathbb{E}_g[g \mathbb{P}(\tau \geq -g)] \\
&= 2 \int_{-\infty}^{\infty} g \phi_g(g) \mathbb{P}(\tau \geq -g) \mathrm{d}g
\end{aligned}$$

where ϕ_g denotes the density function of g. Denote by ϕ_τ the corresponding density function of τ, *i.e.*, $\mathbb{P}(\tau \geq -g) = \int_{-g}^{\infty} \phi_\tau(u) \mathrm{d}u$. Since $\phi_g'(g) = -g \phi_g(g) / \|\mathring{\mathbf{x}}\|_2^2$, integration by parts in combination with Leibniz's rule yields

$$\begin{aligned}
\mathbb{E}[g \operatorname{sgn}(g + \tau)] &= -2\|\mathring{\mathbf{x}}\|_2^2 [\phi_g(g) \mathbb{P}(\tau \geq -g)]_{g=-\infty}^{\infty} + 2\|\mathring{\mathbf{x}}\|_2^2 \int_{-\infty}^{\infty} \phi_g(g) \phi_\tau(-g) \mathrm{d}g \\
&= \sqrt{\frac{2}{\pi}} \|\mathring{\mathbf{x}}\|_2^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\theta \|\mathring{\mathbf{x}}\|_2} \exp\left(-\frac{g^2}{2}\left[\frac{1}{\theta^2} + \frac{1}{\|\mathring{\mathbf{x}}\|_2^2}\right]\right) \mathrm{d}g \\
&= \sqrt{\frac{2}{\pi}} \|\mathring{\mathbf{x}}\|_2^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\theta \|\mathring{\mathbf{x}}\|_2} \exp\left(-\frac{g^2}{2\frac{\theta^2 \|\mathring{\mathbf{x}}\|_2^2}{\theta^2 + \|\mathring{\mathbf{x}}\|_2^2}}\right) \mathrm{d}g \\
&= \sqrt{\frac{2}{\pi(\theta^2 + \|\mathring{\mathbf{x}}\|_2^2)}} \|\mathring{\mathbf{x}}\|_2^2.
\end{aligned}$$

This concludes the proof. □

The following result due to Plan and Vershynin now establishes the desired concentration behavior of $L_{\mathring{\mathbf{x}}}(\mathbf{x})$. Note that the result in [PV13b] is technically concerned with the concentration behavior of the unregularized loss function $\mathbf{x} \mapsto \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle / m$ rather than $L_{\mathring{\mathbf{x}}}(\mathbf{x})$. However, since the regularizer $\mu \|\mathbf{x}\|_2^2$ is deterministic, linearity of expectation and $L_{\mathbf{x}}$ implies that the result also holds for $L_{\mathring{\mathbf{x}}}(\mathbf{x})$, which is the version we state below.

Proposition 4.44 ([PV13b, Proposition 4.2])**.** *Let $\mathcal{K} \subset \mathbb{R}^d$, and fix $\mathbf{x} \in \mathbb{R}^d$. Then for each $t \geq 0$, it holds that*

$$\mathbb{P}\left(\sup_{\mathbf{z} \in \mathcal{K}} |L_{\mathbf{x}}(\mathbf{z}) - \mathbb{E} L_{\mathbf{x}}(\mathbf{z})| \leq 4 \frac{w(\mathcal{K})}{\sqrt{m}} + t \right) \geq 1 - 4 \exp\left(-\frac{m t^2}{8} \right).$$

Proof of Theorem 4.41. We apply Proposition 4.44 for the vector $\mathring{\mathbf{x}} \in \mathcal{K} := \operatorname{conv}(\mathcal{E}_{\mathcal{I},s} \cap r\mathbb{B}_2^d) = \sqrt{s} r \mathbb{B}_{\mathcal{I},1}^d \cap r \mathbb{B}_2^d$ and a maximizer $\hat{\mathbf{x}} \in \mathcal{K}$ of Problem (P$_{4.13}$). This means that with probability at least $1 - 4\exp(-mt^2/8)$, we have from optimality of $\hat{\mathbf{x}}$ for Problem (P$_{4.13}$) that

$$\begin{aligned} 0 &\leq L_{\mathring{\mathbf{x}}}(\hat{\mathbf{x}}) - L_{\mathring{\mathbf{x}}}(\mathring{\mathbf{x}}) \\ &\leq L_{\mathring{\mathbf{x}}}(\hat{\mathbf{x}} - \mathring{\mathbf{x}}) \\ &\leq \mathbb{E} L_{\mathring{\mathbf{x}}}(\hat{\mathbf{x}} - \mathring{\mathbf{x}}) + 4 \frac{w(\mathcal{K} - \mathcal{K})}{\sqrt{m}} + t \\ &\leq \lambda_{\mathring{\mathbf{x}}} \Big(\langle \mathring{\mathbf{x}}, \hat{\mathbf{x}} \rangle - \|\mathring{\mathbf{x}}\|_2^2 \Big) - \frac{\mu}{m} \Big(\|\hat{\mathbf{x}}\|_2^2 - \|\mathring{\mathbf{x}}\|_2^2 \Big) + 8 \frac{w(\mathcal{K})}{\sqrt{m}} + t \end{aligned} \tag{4.30}$$

where in the last step we invoked Lemma 4.43. Next, note that we have

$$\begin{aligned} \frac{\lambda_{\mathring{\mathbf{x}}}}{2} \|\hat{\mathbf{x}} - \mathring{\mathbf{x}}\|_2^2 &= \frac{\lambda_{\mathring{\mathbf{x}}}}{2} \|\hat{\mathbf{x}}\|_2^2 + \frac{\lambda_{\mathring{\mathbf{x}}}}{2} \|\mathring{\mathbf{x}}\|_2^2 - \lambda_{\mathring{\mathbf{x}}} \langle \hat{\mathbf{x}}, \mathring{\mathbf{x}} \rangle \\ &= -\lambda_{\mathring{\mathbf{x}}} \Big(\langle \hat{\mathbf{x}}, \mathring{\mathbf{x}} \rangle - \|\mathring{\mathbf{x}}\|_2^2 \Big) + \frac{\lambda_{\mathring{\mathbf{x}}}}{2} \|\hat{\mathbf{x}}\|_2^2 - \frac{\lambda_{\mathring{\mathbf{x}}}}{2} \|\mathring{\mathbf{x}}\|_2^2. \end{aligned}$$

In light of (4.30), this implies

$$\begin{aligned} 0 &\leq \frac{\lambda_{\mathring{\mathbf{x}}}}{2} \Big(\|\hat{\mathbf{x}}\|_2^2 - \|\mathring{\mathbf{x}}\|_2^2 - \|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2^2 \Big) - \frac{\mu}{m} \Big(\|\hat{\mathbf{x}}\|_2^2 - \|\mathring{\mathbf{x}}\|_2^2 \Big) + 8 \frac{w(\mathcal{K})}{\sqrt{m}} + t \\ &\leq -\frac{\lambda_{\mathring{\mathbf{x}}}}{2} \|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2^2 + 8 \frac{w(\mathcal{K})}{\sqrt{m}} + t \end{aligned}$$

since by our assumption on μ we have

$$\frac{\lambda_{\mathring{\mathbf{x}}}}{2} = \frac{1}{\sqrt{2\pi(\theta^2 + \|\mathring{\mathbf{x}}\|_2^2)}} \leq \frac{1}{\sqrt{2\pi\theta^2}} \leq \frac{\mu}{m}.$$

Rearranging for $\|\hat{\mathbf{x}} - \mathring{\mathbf{x}}\|_2^2$ therefore yields

$$\|\hat{\mathbf{x}} - \mathring{\mathbf{x}}\|_2^2 \leq \frac{16}{\lambda_{\mathring{\mathbf{x}}}} \frac{w(\mathcal{K})}{\sqrt{m}} + \frac{2t}{\lambda_{\mathring{\mathbf{x}}}} \overset{!}{=} \varepsilon^2,$$

which, after solving for t, gives

$$t = \frac{\lambda_{\mathring{\mathbf{x}}}}{2}\varepsilon^2 - 8\frac{w(\mathcal{K})}{\sqrt{m}}.$$

Substituting this expression into the failure probability of Proposition 4.44 and requiring this probability to be bounded by η, we find by solving for m that

$$\begin{aligned} m &\geq C\lambda_{\mathring{\mathbf{x}}}^{-2}\varepsilon^{-4}\Big[w(\mathcal{K})^2 + \log(\eta^{-1})\Big] \\ &\geq 4\lambda_{\mathring{\mathbf{x}}}^{-2}\varepsilon^{-4}\left[8w(\mathcal{K}) + \sqrt{8\log\left[\left(\frac{\eta}{4}\right)^{-1}\right]}\right]^2 \end{aligned}$$

measurements suffice for the conclusion of Theorem 4.41 to hold with probability at least $1-\eta$ for an appropriately chosen absolute constant $C > 0$.

It remains to bound the mean width of the set $\mathcal{K} = \operatorname{conv}(\mathcal{E}_{\mathcal{I},s} \cap r\mathbb{B}_2^d)$. Since $\mathcal{K}$ is compact, the supremum $\sup_{\mathbf{x}\in\mathcal{K}} \langle \mathbf{g}, \mathbf{x}\rangle$ for a fixed vector $\mathbf{g}$ is attained on the boundary of $\mathcal{K}$. We therefore have

$$\begin{aligned} w(\mathcal{K}) &= w(\operatorname{conv}(\mathcal{E}_{\mathcal{I},s} \cap r\mathbb{B}_2^d)) \\ &= \mathbb{E} \sup_{\mathbf{x}\in\mathcal{E}_{\mathcal{I},s}\cap r\mathbb{B}_2^d} \langle \mathbf{g}, \mathbf{x}\rangle \\ &= r\mathbb{E} \sup_{\mathbf{x}\in\mathcal{E}_{\mathcal{I},s}\cap \mathbb{S}^{d-1}} \langle \mathbf{g}, \mathbf{x}\rangle \\ &= r w(\widetilde{\mathcal{E}}_{\mathcal{I},s}) \\ &\lesssim r\left(\sqrt{s\log(2eG/s)} + \sqrt{sg}\right) \end{aligned}$$

where the last estimate follows from Lemma 4.10. Noting that for $\mathring{\mathbf{x}} \in \mathcal{K} \subset r\mathbb{B}_2^d$, we have $\lambda_{\mathring{\mathbf{x}}}^{-2} \leq \frac{\pi}{2}(\theta^2 + r^2)$, choosing m according to

$$m \gtrsim \varepsilon^{-4}(\theta^2 + r^2)\Big[r^2(s\log(G/s) + sg) + \log(\eta^{-1})\Big]$$

thus ensures the conclusion of Theorem 4.41 holds with the announced success probability. □

From the proof above, it is immediately obvious that if we consider the noisy measurement model $\mathbf{y} = \mathbf{f} \circ \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau} + \boldsymbol{\nu})$ with $\boldsymbol{\nu} \sim \mathsf{N}(\mathbf{0}, \sigma^2\mathrm{Id}_m)$ and $\mathbf{f} \sim \mathsf{B}_m(p)$ denoting again a pair of independent pre- and post-quantization noise vectors, respectively, we obtain the following straightforward extension of Theorem 4.41. The influence of the post-quantization bit flip probability $1-p$ on the required number of measurements follows from the discussion in Section 4.3.2.

Theorem 4.45. *Let $\mathring{\mathbf{x}} \in \sqrt{s}r\mathbb{B}_{\mathcal{I},1}^d \cap r\mathbb{B}_2^d$. Let further $\mathbf{A} \in \mathbb{R}^{m\times d}$ be a standard Gaussian random matrix, $\mathbf{f} \in \{\pm 1\}^m$ a Bernoulli random vector with $f_i \sim_{\text{i.i.d.}} \mathsf{B}(p)$, and $\boldsymbol{\tau} \sim \mathsf{N}(\mathbf{0}, \theta^2\mathrm{Id}_m)$ and $\boldsymbol{\nu} \sim \mathsf{N}(\mathbf{0}, \sigma^2\mathrm{Id}_m)$ two Gaussian random vectors with $\mathbf{A}, \boldsymbol{\tau}, \mathbf{f}$ and $\boldsymbol{\nu}$ pairwise independent from each other. Then with probability at least $1-\eta$, every maximizer $\hat{\mathbf{x}}$ of Problem ($\mathrm{P}_{4.13}$) for $\mathbf{y} = \mathbf{f} \circ \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau} + \boldsymbol{\nu})$ and $\mu \geq m(2p-1)/\sqrt{2\pi(\theta^2+\sigma^2)}$ satisfies*

$$\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \leq \varepsilon,$$

provided that

$$m \gtrsim \varepsilon^{-4}(2p-1)^{-2}(\theta^2+\sigma^2+r^2)\Big[r^2(s\log(G/s)+sg)+\log(\eta^{-1})\Big].$$

Remark 4.46. *(i) In addition to an estimate of the signal energy r, one now also requires information about the noise level σ and bit flip probability $1-p$ in order to choose the regularization parameter μ when solving Problem (P4.13). This stands in contrasts to Problem (P4.10) where we only required an estimate of r and knowledge of the variance of the quantization thresholds. While this seems like a significant drawback of Problem (P4.13), it is a common assumption that prior knowledge on the noise energy is available. Consider for instance the QCBP problem*

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \|\mathbf{x}\|_1 \\ \text{s.t.} & \|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2 \leq \vartheta \end{array}$$

for $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\nu}$ from classical compressed sensing, which also requires prior information about the likely noise energy in the form of an upper bound $\|\boldsymbol{\nu}\|_2 \leq \vartheta$. Depending on the application, estimates on the noise level are oftentimes fairly easy to obtain in a reliable fashion. For instance, in wireless communication, estimates of the background noise energy can be obtained by sounding a communication channel during transmission pauses at regular intervals to update the current estimate of the channel state. Similarly, since the bit flip probability of a 1-bit quantizer is usually independent of the input and mostly determined by intrinsic hardware characteristics, the bit flip probability may be estimated empirically by feeding a known input to the device and observing how many times the output of the quantizer disagrees with the ground truth input.

(ii) Eliminating the failure probability η from the lower bound on m by choosing $\eta = \exp(-r^2(s\log(G/s)+sg))$, we find that

$$m = \Omega\left(\varepsilon^{-4}\left(1+\left(\frac{r}{\sigma}\right)^{-2}+\left(\frac{r}{\theta}\right)^{-2}\right)r^4(s\log(G/s)+sg)\right)$$

measurements suffice for ε-accurate reconstruction of effectively s-group-sparse vectors from noisy 1-bit observations. The term $(r/\sigma)^2$ corresponds to the SNR between the (unquantized) linear measurements $\mathbf{A}\mathring{\mathbf{x}}$ and the noise vector $\boldsymbol{\nu}$. A lower SNR therefore implies that an acquisition system has to take more measurements to obtain a constant reconstruction fidelity. Eliminating ε, we furthermore see that the reconstruction error is bounded by

$$\|\mathring{\mathbf{x}}-\hat{\mathbf{x}}\|_2 \lesssim r\left(\frac{\left(1+\left(\frac{r}{\sigma}\right)^{-2}+\left(\frac{r}{\theta}\right)^{-2}\right)(s\log(G/s)+sg)}{m}\right)^{1/4}.$$

This shows that regardless of the noise level or the mismatch between the upper bound r on the signal energy and the standard deviation θ of the dithering vector, one can always compensate for inaccurate guesses of r and σ by increasing the number of measurements.

The nonuniform nature of Theorem 4.41 is rooted in the fact the supremum in the concentration bound in Proposition 4.44 is only taken over $\mathbf{z}$ but not over $\mathbf{x}$. In order to extend the result to hold uniformly over the entire signal ensemble, we could appeal to another more general result from [**PV13b**]. Note, however, that this result significantly worsens the scaling dependency on the reconstruction fidelity ε, which is why we skip the result here.

4.4.3 Group Hard Thresholding

In this section, we derive a recovery guarantee for a group hard thresholding method from dithered observations inspired once again by [**Bar+17a**]. The method naturally extends the ideas from Section 4.3.3 to analyze the hard thresholding scheme in the presence of pre- and post-quantization noise at the expense of losing the uniform nature of the recovery guarantee over the entire signal ensemble. In particular, we consider the recovery map

$$\Delta_{\mathcal{I}}^{\mathrm{ht}}(\mathbf{y}) := \frac{\theta^2}{\langle \mathbf{y}, \boldsymbol{\tau} \rangle} \mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top \mathbf{y}). \tag{4.31}$$

Too see why this formulation is natural in the current context, we consider similar to Section 4.3.2 and Section 4.4.2 the (nonconvex) program

$$\begin{aligned} \underset{\mathbf{x},w}{\text{maximize}} \quad & \left\langle \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau}), \mathbf{A}\mathbf{x} + w\frac{\boldsymbol{\tau}}{\theta} \right\rangle \\ \text{s.t.} \quad & \begin{pmatrix} \mathbf{x} \\ w \end{pmatrix} \in \Sigma_{\widetilde{\mathcal{I}},s+1} \cap \sqrt{r^2+\theta^2}\mathbb{B}_2^{d+1} \end{aligned} \tag{$\mathrm{P}_{4.14}$}$$

with $\boldsymbol{\tau} = \theta\mathbf{g}$ and $\mathbf{g} \in \mathbb{R}^m$. The motivation for the constraint follows immediately from the fact that the extended vector

$$\mathring{\mathbf{u}} := \begin{pmatrix} \mathring{\mathbf{x}} \\ \theta \end{pmatrix} \tag{4.32}$$

is $(s+1)$-group-sparse w. r. t. the augmented group partition $\widetilde{\mathcal{I}} = \mathcal{I} \cup \{\{d+1\}\}$ if $\mathring{\mathbf{x}} \in \Sigma_{\mathcal{I},s}$, in addition to the assumption that $\|\mathring{\mathbf{x}}\|_2 \leq r$. As before, the above program is not tractable in its current form. However, due to linearity of the cost function, we may replace the feasible set by its convex hull (cf. Proposition A.11), turning the constraint set into a symmetric convex body whose Minkowski functional induces a norm on $\mathbb{R}^{d+1}$ in the form of the $(s+1)$-group-support-norm. By the same arguments as in Section 4.3.3, Problem ($\mathrm{P}_{4.14}$) therefore admits a closed-form solution as

$$\hat{\mathbf{u}} := \begin{pmatrix} \hat{\mathbf{x}} \\ \hat{w} \end{pmatrix} = \sqrt{r^2+\theta^2} \frac{\mathcal{H}_{\widetilde{\mathcal{I}},s+1}(\tilde{\mathbf{A}}^\top \mathbf{y})}{\left\| \mathcal{H}_{\widetilde{\mathcal{I}},s+1}(\tilde{\mathbf{A}}^\top \mathbf{y}) \right\|_2} \tag{4.33}$$

where again we set $\tilde{\mathbf{A}} = (\mathbf{A} \quad \mathbf{g})$. Assuming for the moment that $\hat{w} \neq 0$, we claim that the vector

$$\mathbf{x}^\star := \frac{\theta}{\hat{w}}\hat{\mathbf{x}} = \theta \frac{\sqrt{r^2+\theta^2} \frac{\mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top \mathbf{y})}{\|\mathcal{H}_{\widetilde{\mathcal{I}},s+1}(\tilde{\mathbf{A}}^\top \mathbf{y})\|_2}}{\sqrt{r^2+\theta^2} \frac{\langle \mathbf{y}, \boldsymbol{\tau}/\theta \rangle}{\|\mathcal{H}_{\widetilde{\mathcal{I}},s+1}(\tilde{\mathbf{A}}^\top \mathbf{y})\|_2}} = \theta^2 \frac{\mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top \mathbf{y})}{\langle \mathbf{y}, \boldsymbol{\tau} \rangle}$$

is a good approximation for vectors in $\Sigma_{\mathcal{I},s} \cap r\mathbb{B}_2^d$.

Lemma 4.47. *Let $\mathring{\mathbf{x}} \in \Sigma_{\mathcal{I},s}$ with $\|\mathring{\mathbf{x}}\|_2 \leq r$, and assume that $\tilde{\mathbf{A}} = (\mathbf{A} \quad \mathbf{g}) \in \mathbb{R}^{m\times(d+1)}$ satisfies for $t = 2(s+1)$ the group-RIP on $\Sigma_{\mathcal{I},t}$ with constant $\delta_t = \varepsilon^2/1280$ for $\varepsilon \leq 4\theta/\sqrt{\theta^2+r^2}$. Given measurements of the form $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$ with $\boldsymbol{\tau} = \theta\mathbf{g}$, the vector*

$$\hat{\mathbf{x}} = \frac{\theta^2}{\langle \boldsymbol{\tau}, \mathbf{y}\rangle} \mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top \mathbf{y})$$

satisfies

$$\|\mathring{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \leq \theta\varepsilon.$$

Proof. The proof once again follows ideas from [Bar+17a]. In particular, we appeal to the representation of the recovery problem in terms of the augmented vector

$$\mathring{\mathbf{u}} = \begin{pmatrix} \mathring{\mathbf{x}} \\ \theta \end{pmatrix}$$

in which case we have $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau}) = \operatorname{sgn}(\tilde{\mathbf{A}}\mathring{\mathbf{u}})$. Since $\mathring{\mathbf{u}} \in \mathbb{R}^{d+1}$ is $(s+1)$-sparse and $\tilde{\mathbf{A}}$ satisfies the (ℓ_2, ℓ_1)-group-RIP of order $t = 2(s+1)$, we have by Lemma 4.24 with the augmented group partition $\widetilde{\mathcal{I}} = \mathcal{I} \cup \{\{d+1\}\}$ and

$$\hat{\mathbf{u}} = \mathcal{H}_{\widetilde{\mathcal{I}},s+1}(\tilde{\mathbf{A}}^\top \mathbf{y})$$

that

$$\left\| \frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2} \right\|_2 \leq 4\sqrt{5}\sqrt{\delta_t} = 4\sqrt{5}\sqrt{\frac{\varepsilon^2}{1280}} = \frac{\varepsilon}{4}.$$

As in the proof of Lemma 4.32, this yields for the last coordinate that

$$\frac{|\hat{u}_{d+1}|}{\|\hat{\mathbf{u}}\|_2} \geq \frac{4\theta - \varepsilon\sqrt{r^2+\theta^2}}{\sqrt{r^2+\theta^2}}.$$

For

$$\varepsilon \leq \frac{4\theta}{\sqrt{r^2+\theta^2}},$$

this implies that the last entry of $\tilde{\mathbf{A}}^\top \mathbf{y}$ survives the hard thresholding step such that

$$\hat{\mathbf{u}} = \mathcal{H}_{\widetilde{\mathcal{I}},s+1}(\tilde{\mathbf{A}}^\top \mathbf{y}) = \begin{pmatrix} \mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top \mathbf{y}) \\ \left\langle \frac{\boldsymbol{\tau}}{\theta}, \mathbf{y} \right\rangle \end{pmatrix}.$$

With this, Lemma 4.28 yields

$$\begin{aligned} \left\| \frac{\mathring{\mathbf{x}}}{\theta} - \frac{\hat{\mathbf{x}}}{\theta} \right\|_2 = \left\| \frac{\mathring{\mathbf{x}}}{\theta} - \frac{\mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top \mathbf{y})}{\left\langle \frac{\boldsymbol{\tau}}{\theta}, \mathbf{y} \right\rangle} \right\|_2 &\leq \frac{\|\mathring{\mathbf{u}}\|_2 \cdot \|\hat{\mathbf{u}}\|_2}{\theta \cdot |\hat{u}_{d+1}|} \left\| \frac{\mathring{\mathbf{u}}}{\|\mathring{\mathbf{u}}\|_2} - \frac{\hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_2} \right\|_2 \\ &\leq \frac{\sqrt{r^2+\theta^2}}{\theta} \frac{\sqrt{r^2+\theta^2}}{4\theta - \varepsilon\sqrt{r^2+\theta^2}} \frac{\varepsilon}{4} \leq \varepsilon\theta, \end{aligned}$$

provided that $\varepsilon \leq (3\theta^2 - r^2)/(\theta\sqrt{\theta^2+r^2})$, which is guaranteed by our previous assumption that $\varepsilon \leq 4\theta/\sqrt{\theta^2+r^2}$. Multiplying both sides of the inequality by θ completes the proof. □

Appealing to the group-RIP, the following result, which we state without proof, is immediate with Lemma 4.47 and Lemma 4.11.

Theorem 4.48. *Let $\mathbf{A} \in \mathbb{R}^{m\times d}$ be a standard Gaussian matrix, and denote by $\boldsymbol{\tau} = \theta\mathbf{g}$ a dithering vector with $\mathbf{g} \sim \mathsf{N}(\mathbf{0}, \theta^2\mathrm{Id}_m)$. Fix a value $\varepsilon \leq 4\theta/\sqrt{\theta^2 + r^2}$, and assume*

$$m \gtrsim \varepsilon^{-4}\theta^4\Big[s\log(G/s) + sg + \log(\eta^{-1})\Big].$$

Then it holds with probability at least $1-\eta$ that every vector $\mathring{\mathbf{x}} \in \Sigma_{\mathcal{I},s}$ with $\|\mathring{\mathbf{x}}\|_2 \leq r$ can be approximated from its quantized measurements $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$ by $\Delta_{\mathcal{I}}^{\mathrm{ht}}$ such that

$$\left\|\mathring{\mathbf{x}} - \Delta_{\mathcal{I}}^{\mathrm{ht}}(\mathbf{y})\right\|_2 \leq \varepsilon.$$

To close out this section, we now turn to establishing a noise-robust recovery guarantee for the group hard thresholding algorithm, reusing ideas from Section 4.4.2. To that end, assume for $\mathring{\mathbf{x}} \in \mathbb{R}^d$ that we aim to solve the problem

$$\begin{aligned} \underset{\mathbf{x},w}{\text{maximize}} \quad & \left\langle \mathbf{f} \circ \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau} + \boldsymbol{\nu}), \mathbf{A}\mathbf{x} + w\frac{\boldsymbol{\tau}}{\theta} \right\rangle \\ \text{s.t.} \quad & \begin{pmatrix}\mathbf{x}\\ w\end{pmatrix} \in \Sigma_{\widetilde{\mathcal{I}},s+1} \cap \sqrt{r^2+\theta^2}\,\mathbb{B}_2^{d+1} \end{aligned} \tag{$\mathrm{P}_{4.15}$}$$

corresponding to a noisy variant of Problem ($\mathrm{P}_{4.14}$) under the noisy measurement model

$$\mathbf{y} = Q(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau}) = \mathbf{f} \circ \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau} + \boldsymbol{\nu}) \tag{4.34}$$

with $\mathbf{f} \in \mathsf{B}_m(p)$ and $\boldsymbol{\nu} \sim \mathsf{N}(\mathbf{0}, \sigma^2\mathrm{Id}_m)$ as usual. By the discussion at the beginning of this section, Problem ($\mathrm{P}_{4.15}$) admits a closed-form solution given by (4.33). Once again, before invoking Theorem 4.22 with

$$\lambda = (2p-1)\sqrt{\frac{2}{\pi(\sigma^2+1)}}$$

(see Section 4.3.2), we have to frame the objective function in the correct context. In particular, we need to treat the problem as if we were to estimate a unit-normalized vector from its measurements

$$\mathbf{y} = \mathbf{f} \circ \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau} + \boldsymbol{\nu}) = \mathbf{f} \circ \operatorname{sgn}\left(\begin{pmatrix}\mathbf{A} & \frac{\boldsymbol{\tau}}{\theta}\end{pmatrix}\frac{\begin{pmatrix}\mathring{\mathbf{x}}\\ \theta\end{pmatrix}}{\sqrt{\|\mathring{\mathbf{x}}\|_2^2+\theta^2}} + \frac{\boldsymbol{\nu}}{\sqrt{\|\mathring{\mathbf{x}}\|_2^2+\theta^2}}\right),$$

i.e., we scale down the noise variance according to the true norm of the augmented vector $\mathring{\mathbf{u}}$ given in (4.32). The rest of the proof then follows the example of Theorem 4.39 and Lemma 4.47. The only difference is that instead of an estimate of the mean width of $\widetilde{\mathcal{E}}_{\mathcal{I},s}$ we instead require an estimate of $w(\widetilde{\Sigma}_{\mathcal{I},s})$. However, due to the fact that $\widetilde{\Sigma}_{\mathcal{I},s} \subset \widetilde{\mathcal{E}}_{\mathcal{I},s}$ and consequently $w(\widetilde{\Sigma}_{\mathcal{I},s}) \leq w(\widetilde{\mathcal{E}}_{\mathcal{I},s})$, the scaling requirements for m in terms of the system parameters are identical to Theorem 4.39. As a consequence, we obtain the following result, which is now a mere corollary of Theorem 4.39. We emphasize again the nonuniform nature of the result in contrast to Theorem 4.48, which holds uniformly over the entire signal ensemble $\Sigma_{\mathcal{I},s} \cap r\mathbb{B}_2^d$.

Theorem 4.49. *Let* $\mathbf{A}$ *and* $\boldsymbol{\tau}$ *be as defined in Theorem 4.48. Given a vector* $\mathring{\mathbf{x}} \in \Sigma_{\mathcal{I},s}$ *with* $\|\mathring{\mathbf{x}}\|_2 \leq r$ *and its noisy measurements* $\mathbf{y} = Q(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$ *according to (4.34), it holds with probability at least* $1 - \eta$ *that*

$$\left\| \mathring{\mathbf{x}} - \frac{\theta^2 \mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top \mathbf{y})}{\langle \boldsymbol{\tau}, \mathbf{y} \rangle} \right\|_2 \leq \varepsilon,$$

provided that

$$m \gtrsim \varepsilon^{-4}(2p-1)^{-2}\left(\frac{r^2+\theta^2}{\theta}\right)^4 \left(\frac{\sigma^2}{\theta^2}+1\right)\left[s\log(G/s) + sg + \log(\eta^{-1})\right]$$

for $\varepsilon \leq \sqrt{r^2+\theta^2}$.

Regularized Group Hard Thresholding

As alluded to in Section 4.4.2, it was recently pointed out in [DM18b] that the regularized correlation maximization problem discussed in Section 4.4.2 can be reformulated as a simple projection problem. To that end, consider the program

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle - \mu \|\mathbf{x}\|_2^2 \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{K} \end{aligned} \tag{$\mathrm{P}_{4.16}$}$$

for some structure-promoting signal set $\mathcal{K} \subset \mathbb{R}^d$. Factoring out μ, we may rewrite the objective function as

$$\begin{aligned} \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle - \mu \|\mathbf{x}\|_2^2 &= \mu \left(\left\langle \frac{1}{\mu}\mathbf{A}^\top \mathbf{y}, \mathbf{x} \right\rangle - \|\mathbf{x}\|_2^2 \right) \\ &= \mu \left(\left\langle \frac{1}{\mu}\mathbf{A}^\top \mathbf{y}, \mathbf{x} \right\rangle - \|\mathbf{x}\|_2^2 - \left\| \frac{1}{2\mu}\mathbf{A}^\top \mathbf{y} \right\|_2^2 + \left\| \frac{1}{2\mu}\mathbf{A}^\top \mathbf{y} \right\|_2^2 \right) \\ &= \mu \left(\left\| \frac{1}{2\mu}\mathbf{A}^\top \mathbf{y} \right\|_2^2 - \left\| \mathbf{x} - \frac{1}{2\mu}\mathbf{A}^\top \mathbf{y} \right\|_2^2 \right). \end{aligned}$$

Since the first term in the parentheses is constant, this means that Problem ($\mathrm{P}_{4.16}$) can be rewritten without changing the optimal point as

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \left\| \mathbf{x} - \frac{1}{2\mu}\mathbf{A}^\top \mathbf{y} \right\|_2 \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{K}. \end{aligned}$$

This corresponds to the orthogonal projection of $\mathbf{A}^\top \mathbf{y}/(2\mu)$ on the set $\mathcal{K}$. If we now have $\mathcal{K} = \Sigma_{\mathcal{I},s} \cap r\mathbb{B}_2^d$, the optimal point $\hat{\mathbf{x}}$ admits a closed-form solution given by the group hard thresholding operator. Without the restriction to the ball $r\mathbb{B}_2^d$, the optimal point is given by the vector attaining the best s-group approximation error. With the additional requirement that $\|\mathbf{x}\|_2 \leq r$, we subsequently need to project $\mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top \mathbf{y}/(2\mu))$ on $r\mathbb{B}_2^d$. This

allows us to define the recovery map

$$\begin{aligned}\Delta_{\mathcal{I}}^{\Pi\text{-ht}}(\mathbf{y}) &:= \Pi_{r\mathbb{B}_2^d}\left(\mathcal{H}_{\mathcal{I},s}\left(\frac{1}{2\mu}\mathbf{A}^\top\mathbf{y}\right)\right)\\ &= \begin{cases}\mathcal{H}_{\mathcal{I},s}\left(\frac{1}{2\mu}\mathbf{A}^\top\mathbf{y}\right), & \left\|\mathcal{H}_{\mathcal{I},s}\left(\frac{1}{2\mu}\mathbf{A}^\top\mathbf{y}\right)\right\|_2 \le r,\\ r\dfrac{\mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top\mathbf{y})}{\left\|\mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top\mathbf{y})\right\|_2}, & \text{otherwise}\end{cases}\\ &= \min\left\{\frac{1}{2\mu}, \frac{r}{\left\|\mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top\mathbf{y})\right\|_2}\right\}\mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top\mathbf{y}) \qquad (4.35)\end{aligned}$$

where the notation $\Delta_{\mathcal{I}}^{\Pi\text{-ht}}$ emphasizes the fact that the map constitutes a projected hard thresholding scheme. We point out that the regularization term $\mu\|\mathbf{x}\|_2^2$ in Problem ($\mathrm{P}_{4.16}$) is only useful in case $\mathcal{K}$ is not just the set of group-sparse vectors on the boundary of a scaled ℓ_2-ball as the term would otherwise be constant. For this reason, the formulation above does not extend to the undithered setting considered in Section 4.3 as we limit our attention to the signal sets $\widetilde{\Sigma}_{\mathcal{I},s}$ and $\widetilde{\mathcal{E}}_{\mathcal{I},s}$ due to the scale invariance of the problem. The performance analysis of $\Delta_{\mathcal{I}}^{\Pi\text{-ht}}$ proceeds—up to estimating the mean width of the assumed signal set—in the same way as the analysis of Problem ($\mathrm{P}_{4.13}$) in Section 4.4.2. Since the mean width of $\sqrt{s}r\mathbb{B}_{\mathcal{I},1}\cap r\mathbb{B}_2^d$ and $\Sigma_{\mathcal{I},s}\cap r\mathbb{B}_2^d$ differs only by a constant, the following result is therefore a direct corollary of Theorem 4.45.

Theorem 4.50. *Let $\mathring{\mathbf{x}} \in \Sigma_{\mathcal{I},s}\cap r\mathbb{B}_2^d$. Assume that $\mathbf{A}$ and $\boldsymbol{\tau}$ are as defined in Theorem 4.48, and assume we acquire measurements $\mathbf{y} = Q(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$ according to the noisy measurement model (4.34). Then with the same assumption on m and μ as in Theorem 4.45, it holds with probability at least $1-\eta$ that*

$$\left\|\mathring{\mathbf{x}} - \Delta_{\mathcal{I}}^{\Pi\text{-ht}}(\mathbf{y})\right\|_2 \le \varepsilon.$$

4.4.4 Numerical Evaluation

In this section, we conduct a few empirical experiments for dithered recovery similar to the ones carried out earlier in the context of direction recovery of group-sparse vectors. We begin by assessing the error decay rates of the five recovery schemes discussed for this purpose. Due to the fact that all recovery guarantees share a polynomial dependence of the number of measurements on r, we already expect performance to fall behind substantially compared to the direction recovery problem. Inspired by the numerical evaluation in [**KSW16**], we therefore reduce the ambient dimension to $d = 300$ to allow for oversampling by a factor of 16 without increasing the computational load too much. Moreover, we now consider $G = 60$ groups of size $g = 5$. We still consider $s = 5$ active groups such that in total every vector $\mathring{\mathbf{x}}$ has 25 nonzero entries drawn from a standard Gaussian distribution. We then rescale $\mathring{\mathbf{x}}$ such that its norm is distributed uniformly over the interval $[r_0, r] = [10, 20]$. Each recovery strategy is provided with the upper bound $r = 20$ if necessary. Moreover, we choose $\theta = r$ as this choice is optimal for most methods considered in our experiments.

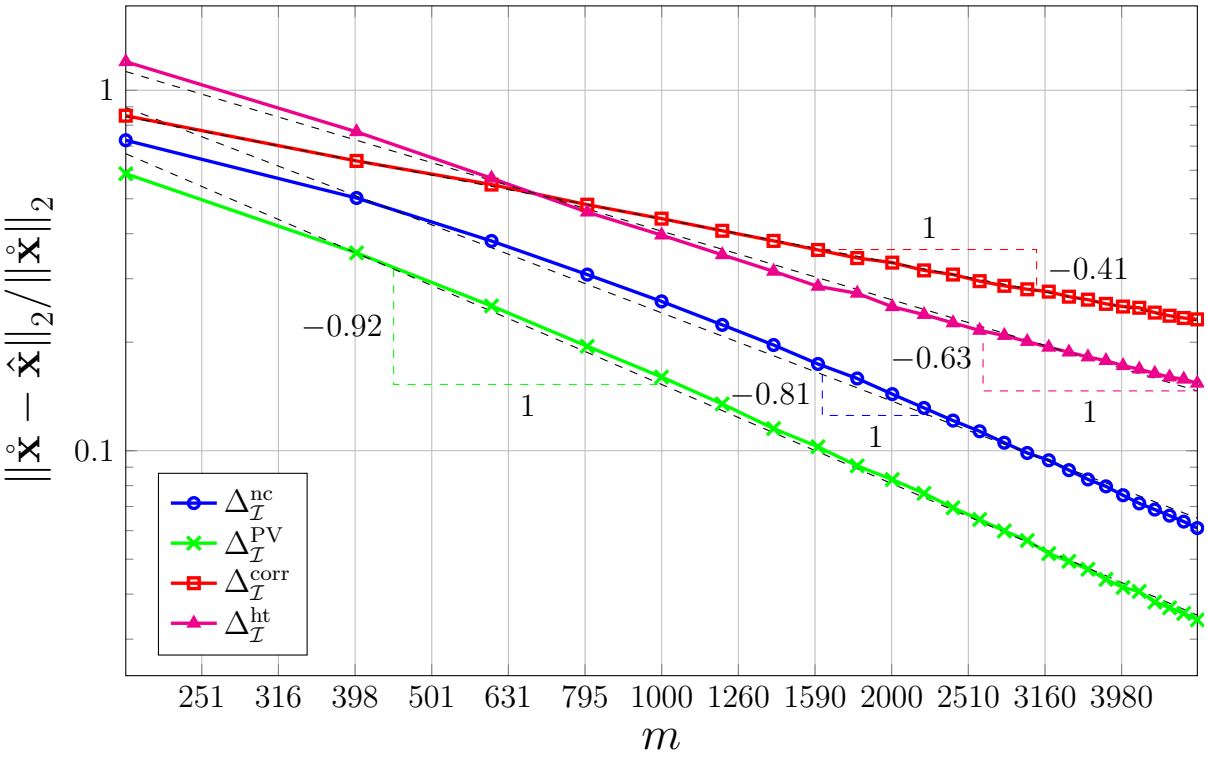


Figure 4.9: Performance of dithered group-sparse recovery methods and their associated empirical error decay rates

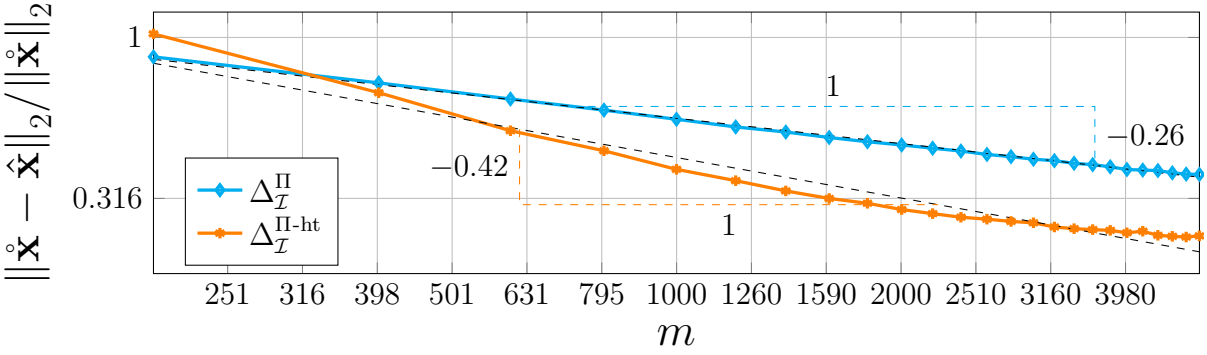
Noiseless Signal Estimation

We first consider the reconstruction of group-sparse vectors from noiseless dithered observations. The results of the first experiment, the normalized ℓ_2-error as a function of m, are shown in Figure 4.9. Considering that both $\Delta_{\mathcal{I}}^{\mathrm{nc}}$ and $\Delta_{\mathcal{I}}^{\mathrm{PV}}$ are based on an augmented version of Problem ($\mathrm{P}_{4.3}$), they inherit the decay behavior of $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$. In particular, they exhibit empirical error decay rates of approximately $\mathcal{O}(m^{-4/5})$ and $\mathcal{O}(m^{-9/10})$, respectively, which are again close to the optimal behavior $\mathcal{O}(m^{-1})$. Surprisingly, there now exists a gap between the performance of $\Delta_{\mathcal{I}}^{\mathrm{corr}}$ and $\Delta_{\mathcal{I}}^{\mathrm{ht}}$ despite their connection pointed out in the discussion in Section 4.4.3 and the fact that both methods are based on the same analysis strategy. More precisely, the lifted group hard thresholding method exhibits a slightly better error decay in the dithered setting, while the decay rate of $\Delta_{\mathcal{I}}^{\mathrm{corr}}$ reduces to $\mathcal{O}(m^{-2/5})$.

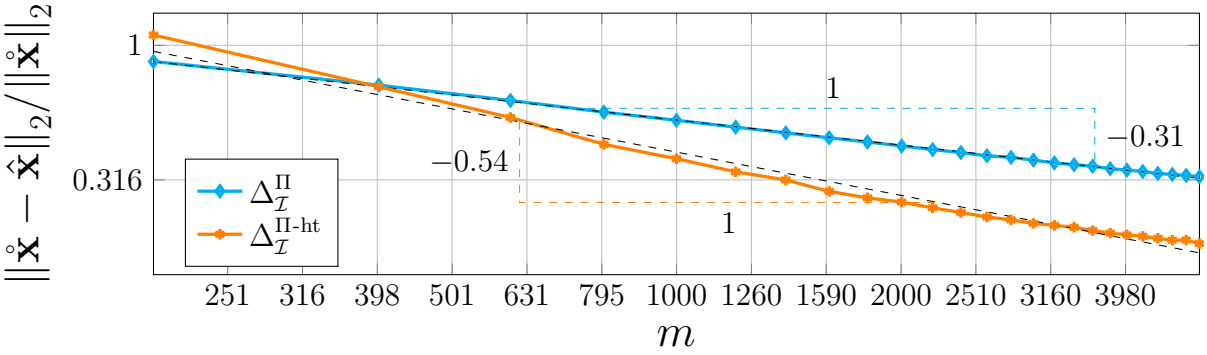
Note that Figure 4.9 does not include the performance profiles of $\Delta_{\mathcal{I}}^{\Pi}$ or $\Delta_{\mathcal{I}}^{\Pi\text{-ht}}$ introduced at the end of Section 4.4.2 and Section 4.4.3, respectively. This is due to the fact that both recovery maps turn out to be highly sensitive to the choice of the regularization parameter μ. According to the proof of Theorem 4.41, it suffices to choose

$$\mu \geq \frac{m}{\sqrt{2\pi(\theta^2 + \|\mathring{\mathbf{x}}\|_2^2)}}$$

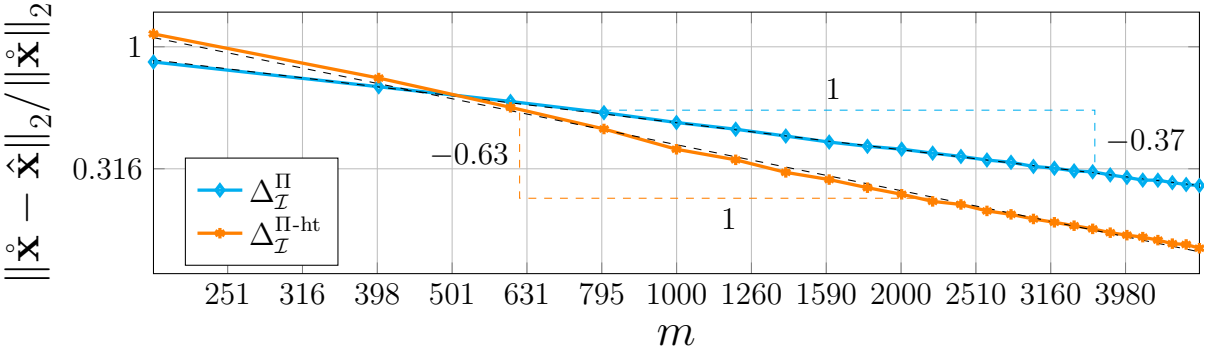
for the conclusion of Theorem 4.41 and in turn the conclusion of Theorem 4.50 to hold. Since $\|\mathring{\mathbf{x}}\|_2$ is unknown, we first choose $\mu = m/\sqrt{2\pi\theta^2}$ in our simulations. This choice is clearly satisfactory for $\Delta_{\mathcal{I}}^{\Pi}$ as shown by the empirical error decay show in Figure 4.10a.



(a) Performance with suboptimal hyperparameter choice $\mu = m/\sqrt{2\pi r^2}$



(b) Performance with $\mu = m/\sqrt{2\pi(r^2 + r_0^2)}$ for $\|\mathring{\mathbf{x}}\|_2 \geq r_0$



(c) Performance with optimal parameter choice $\mu = m/\sqrt{2\pi(r^2 + \|\mathring{\mathbf{x}}\|_2^2)}$

Figure 4.10: Error decay of the projection and projection-based group hard thresholding strategy for different choices of the hyperparameter μ

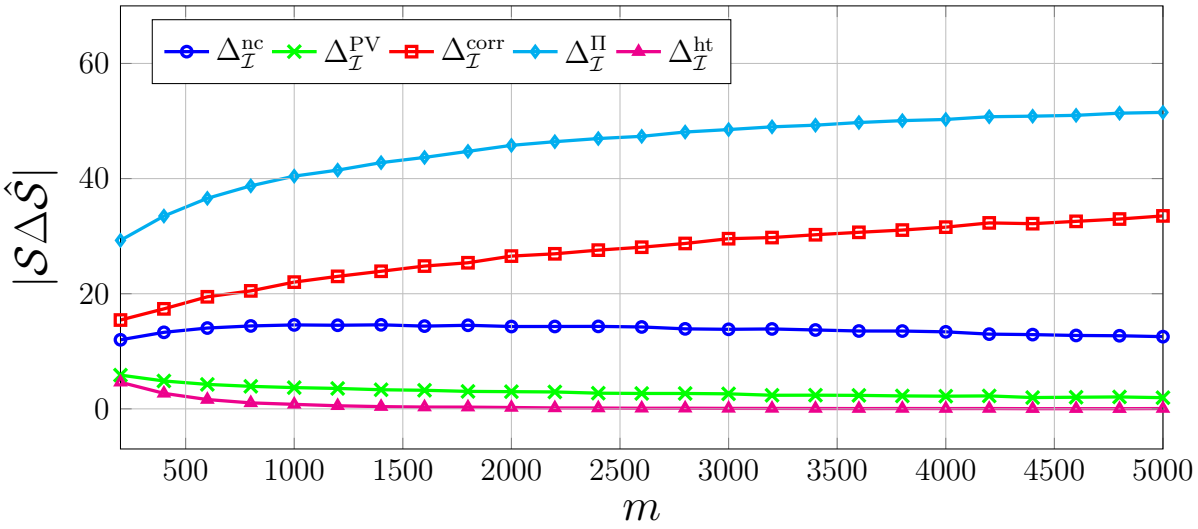


Figure 4.11: Group support detection vs. number of measurements

However, for $\Delta_{\mathcal{I}}^{\Pi\text{-ht}}$, this particular choice of μ leads to a suboptimal error behavior which does not show the polynomial error decay. Since we choose the norm of $\mathring{\mathbf{x}}$ uniformly from the interval $[r_0, r] = [10, 20]$ for each draw of $\mathring{\mathbf{x}}$, a better choice for μ would be

$$\mu = \frac{m}{\sqrt{2\pi(r^2 + r_0^2)}}.$$

As depicted in Figure 4.10b, this choice slightly improves the decay behavior of both methods with the graph corresponding to $\Delta_{\mathcal{I}}^{\Pi\text{-ht}}$ flattening out as desired. In other words, if a sensible lower bound on the expected signal energy is available in practice, the performance of $\Delta_{\mathcal{I}}^{\Pi}$ and $\Delta_{\mathcal{I}}^{\Pi\text{-ht}}$ can be improved by a tighter choice of the regularization parameter μ. On the other hand, if we assume for the moment that $\|\mathring{\mathbf{x}}\|_2$ were known beforehand and choose instead

$$\mu = \frac{m}{\sqrt{2\pi(\theta^2 + \|\mathring{\mathbf{x}}\|_2^2)}}, \tag{4.36}$$

the decay behavior of both recovery schemes improves as depicted in Figure 4.10c. In particular, the decay rate of the regularized group hard thresholding algorithm $\Delta_{\mathcal{I}}^{\Pi\text{-ht}}$ now coincides with the empirical error decay of $\mathcal{O}(m^{-0.63})$ of the lifted group hard thresholding scheme $\Delta_{\mathcal{I}}^{\mathrm{ht}}$ as shown in Figure 4.9. The empirical behavior of $\Delta_{\mathcal{I}}^{\Pi}$ on the other hand moves closer to the behavior observed by $\Delta_{\mathcal{I}}^{\mathrm{corr}}$, exhibiting an error decay of $\mathcal{O}(m^{-0.37})$. This dependence on the hyperparameter μ constitutes a significant drawback of both recovery schemes. Considering that the lifted hard thresholding algorithm does not require any additional parameter tuning, the method always exhibits the same error decay. Moreover, since both methods merely rescale the group hard thresholding estimate $\mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top \mathbf{y})$ their performance in terms of group support recovery is identical. Overall, this means that the regularized group hard thresholding scheme is generally not competitive and should therefore *not* be chosen over $\Delta_{\mathcal{I}}^{\mathrm{ht}}$ in practice.

Next, we turn our attention to the support recovery problem whose results are presented in Figure 4.11. Since $\Delta_{\mathcal{I}}^{\mathrm{ht}}$ and $\Delta_{\mathcal{I}}^{\Pi\text{-ht}}$ only differ in their scaling of $\mathcal{H}_{\mathcal{I},s}(\mathbf{A}^\top \mathbf{y})$, they yield

the same performance for the task at hand, which is why we only include results for $\Delta_{\mathcal{I}}^{\mathrm{ht}}$. Once again, we note the significant performance gap that now exists between $\Delta_{\mathcal{I}}^{\mathrm{corr}}$ and $\Delta_{\mathcal{I}}^{\mathrm{ht}}$ with the latter one achieving almost perfect group support recovery at even moderate numbers of measurements. Despite its almost optimal error decay rate, $\Delta_{\mathcal{I}}^{\mathrm{nc}}$ falls behind $\Delta_{\mathcal{I}}^{\mathrm{PV}}$, which is once again in line with Lemma 4.26, establishing that estimates produced by $\Delta_{\mathcal{I}}^{\mathrm{nc}}$ are merely effectively rather than genuinely group-sparse. This goes along with the difference in their empirical error decay rates as estimated in the previous experiment. Surprisingly, $\Delta_{\mathcal{I}}^{\mathrm{PV}}$ now takes the role of $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$ in the undithered setting by closely following behind $\Delta_{\mathcal{I}}^{\mathrm{ht}}$ with $\Delta_{\mathcal{I}}^{\mathrm{PV}}$ consistently misidentifying around 2 to 3 groups on average but generally performing well. Trailing behind the most by far is the regularized correlation maximization strategy $\Delta_{\mathcal{I}}^{\Pi}$. This is also consistent with the previous findings which show that $\Delta_{\mathcal{I}}^{\Pi}$ suffers from a slower empirical error decay of around $\mathcal{O}(m^{-2/5})$ if the regularization parameter μ is not properly adjusted to the unknown norm of $\mathring{\mathbf{x}}$.

Recovery from Noisy Observations

In the last experiment, we investigate the noise resilience of the proposed recovery schemes. To that end, recall from [Jac+13] that in expectation,

$$\beta := \frac{1}{2}\frac{\sigma}{\sqrt{\|\mathring{\mathbf{x}}\|_2^2 + \theta^2 + \sigma^2}}$$

measurements are flipped due to the influence of pre-quantization noise if we consider observations of the form $\mathbf{y} = \operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau} + \boldsymbol{\nu})$ with $\boldsymbol{\nu} \sim \mathsf{N}(\mathbf{0}, \sigma^2 \mathrm{Id}_m)$ as usual. Solving for σ, we therefore choose the noise variance according to

$$\sigma^2 = \frac{4\beta^2\left(\|\mathring{\mathbf{x}}\|_2^2 + \theta^2\right)}{1 - 4\beta^2},$$

where $\beta \in [0, 1]$ now corresponds to the expected normalized Hamming distance between $\operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau})$ and $\operatorname{sgn}(\mathbf{A}\mathring{\mathbf{x}} + \boldsymbol{\tau} + \boldsymbol{\nu})$. As in Section 4.3.4, we choose $\beta = 0.1$ so that on average 10 % of all sign measurements are flipped. Unfortunately, the convex programs Problem ($\mathrm{P}_{4.8}$) and Problem ($\mathrm{P}_{4.9}$) associated with $\Delta_{\mathcal{I}}^{\mathrm{nc}}$ and $\Delta_{\mathcal{I}}^{\mathrm{PV}}$, respectively, turned out to be highly sensitive to noise to the point where both programs became unstable, consistently resulting in infeasible problem instances. For this reason, we were unable to complete a proper benchmark of $\Delta_{\mathcal{I}}^{\mathrm{nc}}$ and $\Delta_{\mathcal{I}}^{\mathrm{PV}}$ in the noisy regime. This unfortunate circumstance severely limits the usefulness of both recovery methods in any practical context. We point out that such stability issues were not encountered in Section 4.3.4 when testing the methods' counterpart Problem ($\mathrm{P}_{4.3}$) in the context of group-sparse recovery on the sphere from undithered observations. Based on the previous discussion, we also consider oracle variants of $\Delta_{\mathcal{I}}^{\Pi}$ and $\Delta_{\mathcal{I}}^{\Pi\text{-ht}}$ by choosing the regularization parameter μ according to (4.36). We emphasize again that this choice of μ is impossible in practice since the problem of estimating the norm of $\mathring{\mathbf{x}}$ lies at the heart of the dithered observation model. We merely assume knowledge of $\|\mathring{\mathbf{x}}\|_2$ here to obtain a best-case performance profile of recovery maps based on the regularized recovery program ($\mathrm{P}_{4.16}$).

The simulations, whose results are shown in Figure 4.12, confirm again that the error decay of $\Delta_{\mathcal{I}}^{\mathrm{corr}}$ does not change in the noisy setting as predicted by Theorem 4.41. Similarly, choosing the regularization parameter μ in an oracle-optimal way according to (4.36) by

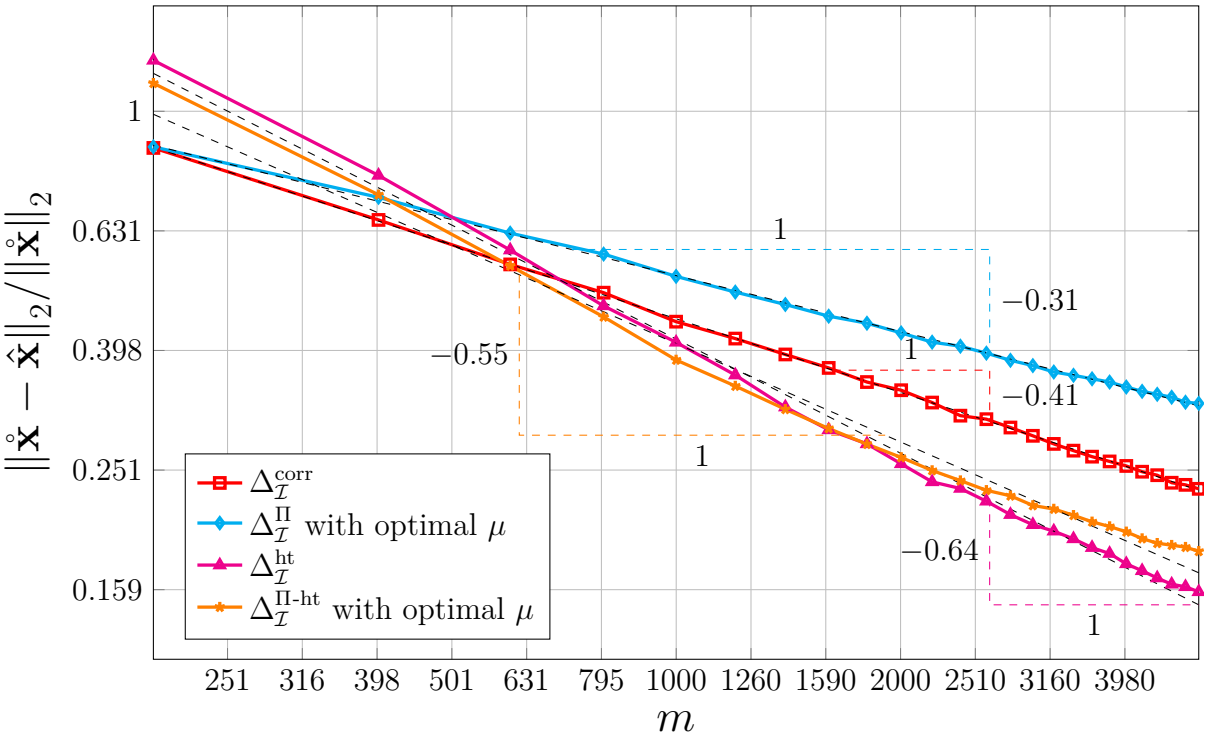


Figure 4.12: Normalized recovery error vs. number of measurements for noisy observations with an average of 10 % of all bits flipped

assuming exact knowledge of $\|\mathring{\mathbf{x}}\|_2$, the error decay of $\Delta_{\mathcal{I}}^{\Pi}$ matches the previous behavior in the noiseless setting. Unlike $\Delta_{\mathcal{I}}^{\text{corr}}$ and $\Delta_{\mathcal{I}}^{\Pi}$, however, the recovery error of the hard thresholding algorithms $\Delta_{\mathcal{I}}^{\text{ht}}$ and $\Delta_{\mathcal{I}}^{\Pi\text{-ht}}$ now visibly deviates from their expected log-linear behavior. For $\Delta_{\mathcal{I}}^{\text{ht}}$, this might be rooted in the fact that according to Theorem 4.49 and Remark 4.40, the optimal choice of θ depends not only on r but also on the noise variance σ^2. While this observation technically also applies to $\Delta_{\mathcal{I}}^{\text{corr}}$ according to Theorem 4.39, the situation is comparable to the previous discussion about the discrepancy between $\Delta_{\mathcal{I}}^{\Pi}$ and $\Delta_{\mathcal{I}}^{\Pi\text{-ht}}$. Overall, these observations suggest that hard thresholding schemes are generally more sensitive to slightly suboptimal parameter configurations. We point out, however, that with more than $m = 1500$ measurements, the lifted hard thresholding algorithm $\Delta_{\mathcal{I}}^{\text{ht}}$ outperforms any competing algorithms with its empirical error decay matching its rate in the noiseless setting.

To close out this chapter, we benchmark the group support detection performance in the noisy regime. Once again, the results shown in Figure 4.13 confirm the noise resilience of the group hard thresholding scheme $\Delta_{\mathcal{I}}^{\text{ht}}$. As in the undithered case, both correlation maximization strategies (excluding $\Delta_{\mathcal{I}}^{\text{ht}}$) are not competitive even at considerably higher values of m. In fact, the number of misidentified groups increases the more measurements we acquire, which is in stark contrast to the hard thresholding approach whose detection error drops to zero beyond $m = 1500$ measurements. Overall, its simple and numerically efficient implementation, noise robustness and support detection accuracy render $\Delta_{\mathcal{I}}^{\text{ht}}$ the most promising recovery scheme.

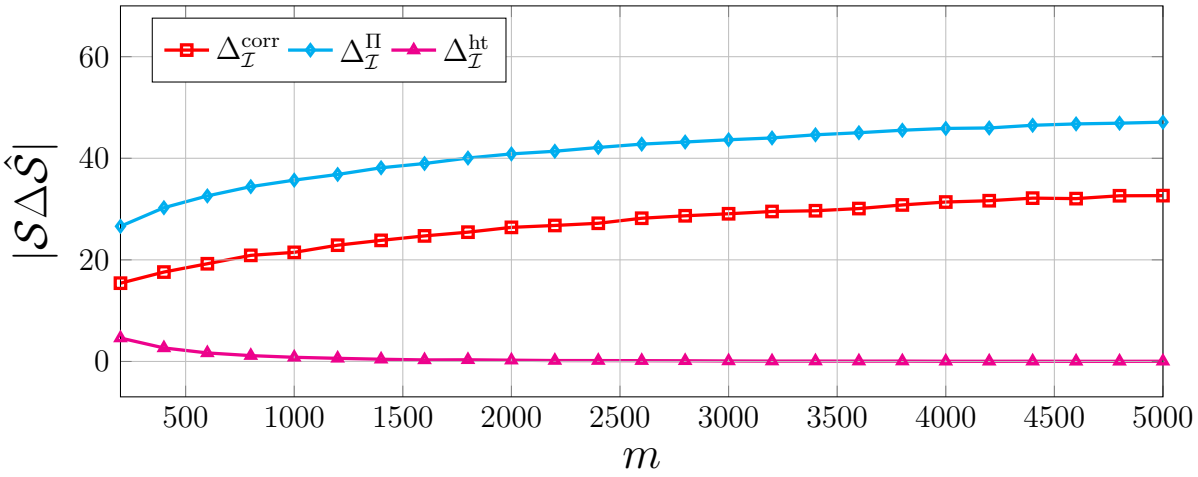


Figure 4.13: Group support detection rate vs. number of measurements when roughly 10 % of all sign measurements are wrong

4.5 Conclusion

In this chapter, we considered the reconstruction of group-sparse vectors from 1-bit observations of the form $\operatorname{sgn}(\mathbf{A}\mathbf{x}) \in \{\pm 1\}^m$. Since any information about the norm of $\mathbf{x}$ is lost in the acquisition process due to the scale invariance of the sgn-operator, we first limited ourselves to the direction recovery problem on the Euclidean unit sphere. We established theoretical reconstruction guarantees for three recovery strategies modeled after existing schemes in the 1-bit compressed sensing literature. In particular, we established that $\Omega(\varepsilon^{-\alpha}(s \log(G/s) + sg))$ measurements suffice to estimate group-sparse signals up to ε-fidelity, where the integer power $\alpha \in \{3, 4\}$ depends on the choice of the recovery procedure.

We complemented our theoretical findings with a series of numerical experiments to gauge how close the predicted error decay rates are to their empirical rates. Overall, we found that every method exhibits slightly faster empirical error decay than predicted by the accompanying theory with the group-sparse variant $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ of a recovery strategy due to Plan and Vershynin coming closest to the provably optimal rate of $\mathcal{O}(m^{-1})$ among all possible decoding maps.

We also considered the performance in the noisy regime and found that two of the considered methods—a correlation maximization scheme $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$ and an efficient group hard thresholding algorithm $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$—exhibit remarkable robustness to both additive pre-quantization noise, as well as adversarial post-quantization bit flips. On the other hand, the performance of $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$, while still competitive, was shown to deteriorate substantially in the presence of noise. Lastly, we considered the problem of identifying the active support of group-sparse vectors. In this context, it was observed that $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$ was only competitive at moderate group-sparsity levels despite its generally accurate signal estimation performance.

Overall, the numerical results put the inexpensive group hard thresholding strategy $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$ above its competitors, which both rely on more complicated convex programming techniques and are therefore not suited for use in time-critical real-time applications. With as many as 10 % of all measured bits flipped, the group hard thresholding algorithm still managed to almost always perfectly identify the active groups of group-sparse vectors,

even in highly undersampled regimes. The main drawback of the algorithm is its seemingly slow error decay rate of $\mathcal{O}(m^{-1/2})$, as well as the fact that $\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$ requires prior information on the group-sparsity level of the vector one aims to recover.

Next we considered the problem of estimating both the direction of a signal as well as its ℓ_2-norm under the assumption that signals of interest are contained in a ball of radius r. To that end, we considered a well-known dithering strategy from quantization theory, which had previously been adopted in the quantized compressed sensing literature to overcome the problem of limited variation in random measurement ensembles. We presented six different reconstruction schemes and analyzed their theoretical performance by appealing to results established in the undithered setting. As in the direction recovery problem, we complemented the theoretical results with a numerical study to test how close the performance of each recovery scheme comes to its predicted accuracy. Once again we found the performance of the two strategies based on the idea of enforcing quantization consistency to outperform its competitors in the noiseless regime, exhibiting almost optimal error decay. We point out, however, that due to the predicted polynomial dependence on the radius of the signal ensemble, one generally has to acquire significantly more measurements than in the previous setting to appropriately tessellate the signal set into small enough quantization cells.

In the noisy regime, we once again confirmed the remarkable error resilience of four of the six considered reconstruction approaches. Unfortunately, the two approaches based on enforcing strict quantization consistency turned out to be highly sensitive to even moderate levels of noise, resulting in consistently infeasible program instances, which could therefore not be included in our experiments. This unfortunate circumstance, which severely limits the usefulness of the respective recovery schemes in practical applications, had previously not been reported in the literature. Overall, the lifted group hard thresholding strategy modeled after its undithered counterpart outperformed any other method considered in our experiments. Due to the fact that the algorithm is highly efficient in terms of its numerical complexity, in addition to its accurate group support identification performance in the presence of both pre- and post-quantization noise, the group hard thresholding strategy remains one of the most promising reconstruction methods to date.

Open Problems

In closing, we would like to point out some open problems and potential future research directions. As established in our numerical experiments, there remains a distinct performance gap between the predicted and empirically observed error decay of most methods discussed in this chapter. While the theoretical decay behavior of most methods is predicted to be $\mathcal{O}(m^{-1/\alpha})$ for $\alpha \in \{3, 4\}$, the numerical experiments indicate that at least some methods are close to the optimal rate of $\mathcal{O}(m^{-1})$. Up until this point, however, it is not clear how to close this gap even for well-studied ensembles like Gaussian distributions.

As demonstrated by the numerical experiments carried out in this work, accurate signal recovery does not necessarily go hand in hand with accurate support detection. In fact, arguably the most accurate method based on enforcing strict quantization consistency exhibits rather unsatisfactory support identification performance. Despite the fact that many of the reconstruction methods considered in this chapter provide accurate group support recovery, there are currently no theoretical results corroborating this behavior. As pointed out before, the problem of identifying the active support and reconstructing sparse

or group-sparse vectors are equally hard in linear compressed sensing. This is not true, however, when one considers nonlinear observations such as single-bit measurements. While the pertinent literature includes a handful of results which address the support recovery problem in the context of 1-bit CS (see, *e.g.*, [**GNR10**; **HB11**; **Gop+13**; **ABK17**]), all of these works require specialized (random) constructions of measurement matrices to allow for theoretical analyses to be carried out. It is therefore highly desirable to close the current gap in the literature by establishing support recovery guarantees for more classical measurement ensembles such as Gaussian or subgaussian distributions.[8]

Finally, the group-sparsity structure considered in this chapter was limited to nonoverlapping group partitions. Conceptually, this restriction is justified in various applications while in others, the need to allow for overlapping groups is of central importance to enable more realistic signal modeling. Unfortunately, in these situations, even selecting an appropriate objective function to promote group-sparsity is a nontrivial task since the decomposition of a vector into subvectors supported on individual groups is no longer unique. For instance, simply minimizing $\|\cdot\|_{\mathcal{I},1}$ w. r. t. an overlapping group partition $\mathcal{I}$ may lead to degenerate minimizers, which are supported on the complement of a union of groups [**JOV09**; **OJV11**]. In other words, minimizing $\|\cdot\|_{\mathcal{I},1}$ with overlapping groups may lead to sparse rather than group-sparse solutions. In order to circumvent such issues, one typical approach is to appeal to more advanced group penalty functions such as the *graph* least-absolute shrinkage selection operator *(LASSO)* [**JOV09**], the *group LASSO with overlap* or the *latent group LASSO* [**OJV11**]. It would therefore be of significant interest to extended the results presented in this chapter to the recovery of group-sparse signals with overlapping groups based on these alternative group penalty functions.

[8] In the latter case, such results would naturally require an additional dithering step as pointed out several times throughout this chapter.

5

Group-Sparse Signal Recovery with Block Diagonal Matrices

While unstructured random matrices as considered in the previous chapter are highly desirable from a theoretical perspective, system designers are usually not free to choose measurement operators at a whim. Instead, in most engineering applications, the fundamental structural properties of a measurement system are generally predetermined by the particular problem domain. This was the original motivation for the sensing model considered in Chapter 3, where it was assumed that elements of the signal class exhibit a sparse representation in the frequency domain. A natural way to compressively sample such signals is by acquiring randomly selected samples of their time domain expansion, followed by exploiting sparsity in the DFT basis during reconstruction. The limiting factor in scenarios like this is often the energy consumption of high-resolution sensing devices. Assuming that the signal bandwidth is so high as to render even sub-Nyquist sampling impractical, this led us to consider coarsely quantized randomly subsampled time domain measurements to keep energy consumption at bay. A classic application

Parts of this chapter have been accepted for publication and will appear in [**KBM19b**].

area where energy efficiency is of key importance is in *wireless sensor networks* (WSNs) [Aky+02]. Such networks commonly rely on low-power (and often more importantly low-cost) sensing devices in order to reliably observe certain environmental phenomena at different geographical locations. Typical applications of WSNs are in healthcare monitoring, wildfire and earthquake detection, flood early warning systems, smart grids, as well as quality control in manufacturing plants. In all these applications, the sensors deployed in a WSN face the challenge of continuously acquiring potentially high-rate data streams, which subsequently need to be analyzed or forwarded to a dedicated fusion center. Moreover, in many applications, sensing nodes of WSNs are commonly subjected to extreme environmental conditions. In such situations, it is desirable to deploy sensors in a redundant fashion to reduce the need for system maintenance, while ensuring reliable system operability. For these reasons, cheap and energy-efficient sensing devices are of central importance for the successful utilization of wireless sensor networks.

In this chapter, we consider an alternative strategy to reduce energy demands of sensors without resorting to extreme quantization paradigms such as 1-bit compressed sensing as considered in the first two chapters of this thesis. To that end, we consider a particular class of structured random matrices at the intersection of purely random ensembles and highly structured matrices such as those generated by subsampled basis functions of bounded orthonormal systems. More precisely, we consider *block diagonal* measurement matrices whose blocks are either independent or identical copies of a dense subgaussian random matrix. These measurement ensembles play a central role in applications where global data aggregation may be prohibitive due to the underlying data rates required to reliably reconstruct a signal. Such rate regimes are typically encountered in streaming applications, where it might be necessary to operate on lower-dimensional data chunks rather than the entire data stream at once [Asi+10; BA10a; BA10b; WC17]. For instance, in reconstruction of video sequences, it is natural to operate either on individual or a limited number of consecutive frames rather than treating the entire video as one high-dimensional vector [PW09].

Block diagonal measurement operators also appear naturally in various acquisition models like *distributed compressed sensing* (DCS) [Sar+05] and the so-called *multiple measurement vector* (MMV) model. In the latter case, one obtains multiple independent snapshots of a signal, whose low-complexity structure is assumed to be stationary in time [REC04; CH06]. Consider for instance an s-sparse ground truth target signal $\mathring{\mathbf{x}} \in \mathbb{C}^d$ observed by L spatially distributed sensors. Due to environmental effects or propagation characteristics of the communication channel between the location $\mathring{\mathbf{x}}$ originates from and the individual sensor nodes, it is natural to assume that each of the L sensors observes a potentially different vector $\mathbf{x}_l$, which all share the inherent low-complexity structure of $\mathring{\mathbf{x}}$. Assuming that each sensor implements the same measurement procedure modeled by the linear operator $\mathbf{\Phi} \in \mathbb{C}^{m\times d}$, one obtains the collection $\{\mathbf{y}_l\}_{l=1}^L = \{\mathbf{\Phi}\mathbf{x}_l\}_{l=1}^L \subset \mathbb{C}^m$ of measurement vectors. Under the assumption that $\mathring{\mathbf{x}} \in \mathbb{C}^d$ is s-sparse and $\operatorname{supp}(\mathbf{x}_l) = \operatorname{supp}(\mathring{\mathbf{x}}) = S \subset [d]$, one may compactly express this measurement system as

$$\mathbf{Y} := \begin{pmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_L \end{pmatrix} = \mathbf{\Phi}\begin{pmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_L \end{pmatrix} =: \mathbf{\Phi}\mathbf{X},$$

where $\mathbf{X} \in \mathbb{C}^{d\times L}$ is a *row-sparse* matrix since at most s rows of $\mathbf{X}$ feature nonzero entries.[1]

[1] The concept of row-sparsity naturally extends to *row-compressibility*, which assumes that the energy contained in any rows indexed by $[d] \setminus S$ is negligible.

This is known as the multiple measurement vector model. Alternatively, one may vectorize the above equation by stacking the columns of $\mathbf{Y}$ and $\mathbf{X}$ into vectors of size mL and dL, respectively, leading to

$$\begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_L \end{pmatrix} = \begin{pmatrix} \mathbf{\Phi} & & \\ & \ddots & \\ & & \mathbf{\Phi} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_L \end{pmatrix}$$

where the vector on the right-hand side now exhibits a periodic sparsity pattern. Such periodicities in the nonzero support indices (or equivalently the row-sparsity structure) can then be exploited via mixed-norm minimization to reconstruct the ensemble $\{\mathbf{x}_l\}_l \subset \mathbb{C}^d$ [**BF09**].

In the image processing literature, the block diagonal measurement model has garnered a lot of attention under the name *block-based compressed sensing* (BCS) [**MF09**; **Luo+09**; **Yan+09**; **MF11**; **MF12**; **FMT12**; **Adl+16**; **Cui+18**], originally introduced in [**Gan07**]. The central idea in BCS is to consider one measurement operator $\mathbf{\Phi}_B \in \mathbb{R}^{m_B \times B^2}$ for image patches of size $B \times B$ and acquire measurements of the form $\mathbf{y}_i = \mathbf{\Phi}_B \mathbf{x}_i \in \mathbb{R}^{m_B}$, where $\mathbf{x}_i \in \mathbb{R}^{B^2}$ denotes the vectorized version of the i-th image block (according to a particular scanning pattern), and m_B denotes the number of compressive measurements of each image patch. Due to the prevalence of most modern image and video codecs operating on rectangular image patches of possibly varying sizes [**Sul+12**], block-based acquisition models are a natural fit. Moreover, in so-called *hybrid video codecs*, which combine spatial and motion-compensated temporal prediction methods to reduce redundancy in weighted differences between individual frames, residual blocks are inherently sparse or compressible.

Chapter Outline

The theoretical performance of the block diagonal acquisition model for sparse recovery was previously addressed in [**Yap+11**; **Eft+15**]. These works establish lower bounds on the number of measurements for subgaussian block diagonal matrices to satisfy the classical restricted isometry property, implying stability and robustness guarantees for recovery of sparse vectors. In this chapter, we extend the results of [**Eft+15**] to more structured signal sets, namely those whose nonzero coefficients appear in groups as considered in the previous chapter. To establish recovery guarantees in this setting, we appeal to the so-called *group restricted isometry property* (group-RIP), a generalization of the restricted isometry property for matrices acting on group-sparse vectors. Unlike the isometry property considered in the previous chapter, the group-RIP considered here constitutes an embedding between two complex Euclidean spaces rather than one from Euclidean space into ℓ_1^m.

We consider two distinct variations of block diagonal measurement matrices. First, we assume that each block of a measurement matrix is an independent copy of a subgaussian random matrix. In the second scenario, only a single block is drawn randomly from a subgaussian distribution. This block is then copied to each block entry, resulting in a block diagonal random matrix with constant block diagonal. Appealing to the group restricted isometry property, it is shown that group-sparse vectors can be stably and robustly reconstructed from partial observations obtained via block diagonal measurement operators. Like the results in [**Eft+15**], the obtained bounds depend on parameters

which control the coherence between the underlying sparsity basis and the canonical basis of the ambient space. If the sparsity basis is highly incoherent, we establish that the scaling behavior required for subgaussian block diagonal matrices to satisfy the group-RIP almost matches, up to logarithmic factors, fundamental lower bounds on the number of measurements required to establish instance-optimal stability results for block-sparse signal recovery. Furthermore, we show that our bounds reduce to the results reported in [**Eft+15**] when interpreting genuinely sparse as group-sparse vectors w. r. t. a trivial group partition. We relate the problem of establishing the group-RIP to estimating certain geometric quantities associated with the suprema of chaos processes involving Talagrand's γ_2-functional. Since the methods employed in [**Eft+15**] do not directly apply to the group-sparse setting, we propose an alternative technique to estimate the covering number of a specific matrix set at higher scales. In particular, we extend Maurey's empirical method to sets which do not admit a polytope representation. As a side effect of our bound on the γ_2-functional, we provide a generalization of Maurey's lemma to provide new bounds on the covering number of sets that consist of finite convex combinations of compact sets.

The rest of the chapter is organized as follows. In Section 5.1, we detail the signal and acquisition model considered in this chapter, while also fixing notation for the remainder of the thesis. Additionally, we introduce the version of the group restricted isometry property used to establish stable and robust recovery via general group-RIP matrices. In Section 5.2, we summarize a few relevant results from the pertinent literature. Focusing on general subgaussian block diagonal matrices first, we derive a lower bound on the number of measurements for measurement matrices to satisfy the group-RIP with high probability in Section 5.3. The case of block diagonal matrices with constant block diagonal is treated in Section 5.4. In Section 5.5, we discuss our obtained bounds and put them in context of earlier results presented in the literature. Before concluding the chapter in Section 5.7, we present a series of numerical experiments in Section 5.6 to empirically investigate the effect of the number of sensors on the reconstruction quality in an average case analysis.

5.1 Signal Recovery with Block Diagonal Group-RIP Matrices

As in the previous chapter, we consider the problem of recovering signals with a low-complexity structure in the form of group-sparsity w. r. t. a nonoverlapping group partition $\mathcal{I}$. However, we now consider linear rather than excessively quantized observations collected by different sensors which each observe a different portion of the signal. These partial observations are modeled by means of block diagonal measurement matrices. In particular, we assume a vector $\mathbf{x} \in \mathbb{C}^D$ which we decompose into G nonoverlapping groups is observed by L sensors. For simplicity, we require D to be an integer multiple of L such that $D = dL$ with $d \in \mathbb{N}$, giving rise to the decomposition of $\mathbf{x}$ into L signal blocks:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_L \end{pmatrix} \in \mathbb{C}^D, \quad \mathbf{x}_l \in \mathbb{C}^d \; \forall l \in [L].$$

Moreover, we redefine the set of group-sparse vectors w. r. t. a group partition $\mathcal{I}$ as

$$\Sigma_{\mathcal{I},s} = \left\{\mathbf{x} \in \mathbb{C}^D : \|\mathbf{x}\|_{\mathcal{I},0} \leq s\right\}$$

where we adjusted the dimension and now consider elements over the complex base field. We consider a measurement system in which we observe an s-group-sparse or compressible signal $\mathbf{x}$ by means of a block diagonal matrix $\mathbf{A}$ consisting of L blocks, namely

$$\mathbf{A} = \begin{pmatrix} \boldsymbol{\Phi}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Phi}_L \end{pmatrix}.$$

However, we assume that we only have access to the signal $\mathbf{x} \in \Sigma_{\mathcal{I},s}$ in terms of its basis expansion $\mathbf{z}$ in a unitary basis $\boldsymbol{\Psi} \in \mathrm{U}(D)$ where

$$\mathrm{U}(D) := \left\{\mathbf{Q} \in \mathbb{C}^{D\times D} : \mathbf{Q}^*\mathbf{Q} = \mathbf{Q}\mathbf{Q}^* = \mathrm{Id}_D\right\}$$

denotes the unitary group. The measurement model therefore reads

$$\mathbf{y} = \operatorname{diag}\left\{\boldsymbol{\Phi}_l\right\}_{l=1}^{L}\mathbf{z} = \operatorname{diag}\left\{\boldsymbol{\Phi}_l\right\}_{l=1}^{L}\boldsymbol{\Psi}\mathbf{x} = \mathbf{A}\boldsymbol{\Psi}\mathbf{x}. \tag{5.1}$$

We will also consider an alternative measurement model in which each sensor is equipped with a copy of the same matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m\times d}$, *i.e.*, $\boldsymbol{\Phi}_l = \boldsymbol{\Phi}$ $\forall l \in [L]$. Ultimately, our goal in this chapter is to provide a sufficient condition for stable and robust recovery of group-sparse signals by establishing a suitable RIP property of block diagonal matrices acting on group-sparse vectors.

The analysis of both sensing models relies on the so-called group restricted isometry property—a generalization of the well-known restricted isometry property modeled on the block-sparse RIP first introduced in [EM09].

Definition 5.1 (Group restricted isometry property)**.** *A matrix $\mathbf{A}\boldsymbol{\Psi} \in \mathbb{C}^{M\times D}$ with $\mathbf{A} \in \mathbb{R}^{M\times D}$ and $\boldsymbol{\Psi} \in \mathrm{U}(D)$ is said to satisfy the* group restricted isometry property *(group-RIP) of order s if, for $\delta \in (0,1)$,*

$$(1-\delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\boldsymbol{\Psi}\mathbf{x}\|_2^2 \leq (1+\delta)\|\mathbf{x}\|_2^2 \quad \forall \mathbf{x} \in \Sigma_{\mathcal{I},s}. \tag{5.2}$$

The smallest constant $\delta_s \leq \delta$ for which (5.2) holds is called the group restricted isometry constant *(group-RIC) of $\mathbf{A}\boldsymbol{\Psi}$.*

In combination with the above definition, a result due to Gao and Ma established in [GM17], which we will introduce next, then implies stable and robust recovery of group-sparse signals. While the signal model employed in [GM17] assumes that $\mathcal{I}$ is an ascending group partition with equisized groups (see Definition 4.1), meaning that the signals are assumed to be block- rather than group-sparse, the proof of Theorem 1 in [GM17] does not explicitly rely on this structure. Furthermore, the result was originally proven in the real setting, but the proof is easily extended to the complex case. These results therefore also extend to more general group partitions as defined in Definition 4.1. Note that such a stability and robustness result in the block-sparse case had previously been established in the seminal work of Eldar and Mishali [EM09], albeit with the necessary condition $\delta_{2s} < \sqrt{2} - 1$ on the block-RIP constant. The precise statement of our generalization is stated in the following result. For the sake of self-containedness, we provide a proof in Appendix B.

Theorem 5.2. *Let $\tilde{\mathbf{A}} \in \mathbb{C}^{M\times D}$ be a matrix satisfying the group restricted isometry property of order $2s$ with constant $\delta_{2s} < 4/\sqrt{41}$. Then for any $\mathring{\mathbf{x}} \in \mathbb{C}^D$ and $\mathbf{y} = \tilde{\mathbf{A}}\mathring{\mathbf{x}} + \mathbf{e}$ with $\|\mathbf{e}\|_2 \leq \nu$, any solution $\mathbf{x}^\star$ of the program*

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \|\mathbf{x}\|_{\mathcal{I},1} \\ \text{s.t.} & \left\|\tilde{\mathbf{A}}\mathbf{x} - \mathbf{y}\right\|_2 \leq \nu \end{array} \tag{$\mathrm{P}_{\mathcal{I},1}$}$$

satisfies

$$\|\mathring{\mathbf{x}} - \mathbf{x}^\star\|_2 \leq C_0 \frac{\sigma_s(\mathring{\mathbf{x}})_{\mathcal{I},1}}{\sqrt{s}} + C_1\nu$$

where the constants $C_0, C_1 > 0$ only depend on δ_{2s}.

Remark 5.3. *(i) In the noiseless setting with $\nu = 0$, the above result immediately implies perfect recovery of all group-sparse signals as the s-term approximation error $\sigma_s(\mathring{\mathbf{x}})_{\mathcal{I},1}$ vanishes as soon as $\mathring{\mathbf{x}} \in \Sigma_{\mathcal{I},s}$.*

(ii) If desired, it is also possible to characterize the recovery quality in terms of the group ℓ_1-norm in which case one obtains

$$\|\mathring{\mathbf{x}} - \mathbf{x}^\star\|_{\mathcal{I},1} \leq C_0'\sigma_s(\mathring{\mathbf{x}})_{\mathcal{I},1} + C_1'\sqrt{s}\nu$$

*for $C_0', C_1' > 0$ which still only depend on δ_{2s} [**GM17**].*

5.2 Prior Work

A common technique to show that the classical restricted isometry property holds with high probability for dense measurement matrices populated by independent copies of a subgaussian random variable is by establishing concentration results of the form

$$\mathbb{P}\Big(\big|\|\tilde{\mathbf{A}}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \geq t\|\mathbf{x}\|_2^2\Big) \leq C\exp(-ct^2m) \quad \forall \mathbf{x} \in \mathbb{C}^d \tag{5.3}$$

with $\tilde{\mathbf{A}} \in \mathbb{C}^{m\times d}$ and universal constants $C, c > 0$ (see Theorem 2.5). Such concentration results can in turn be established by appealing to Bernstein's inequality for subexponential random variables (see, *e.g.*, [**FR13**, Chapter 9]).

The earliest efforts to establish concentration results for block diagonal random matrices go back to the work in [**Wak+10**] and [**Roz+10**]. In particular, Wakin *et al.* [**Wak+10**] consider block diagonal matrices

$$\mathbf{A} = \begin{pmatrix} \boldsymbol{\Phi}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Phi}_L \end{pmatrix} \in \mathbb{R}^{M\times D}$$

with independent copies $\boldsymbol{\Phi}_l$ of a dense subgaussian random matrix with $\mathbb{E}(\boldsymbol{\Phi}_l)_{ij}^2 = 1$ and show that for $\mathbf{x} \in \mathbb{R}^D$,

$$\begin{aligned} &\mathbb{P}\Big(\big|M^{-1}\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \geq t\|\mathbf{x}\|_2^2\Big) \\ &\leq 2\begin{cases} \exp\left(-ct^2M\dfrac{\|\lambda(\mathbf{x})\|_1^2}{\|\lambda(\mathbf{x})\|_2^2}\right), & 0 \leq t \leq \dfrac{\tilde{c}\|\lambda(\mathbf{x})\|_2^2}{\|\lambda(\mathbf{x})\|_1\|\lambda(\mathbf{x})\|_\infty} \\ \exp\left(-c'tM\dfrac{\|\lambda(\mathbf{x})\|_1}{\|\lambda(\mathbf{x})\|_\infty}\right), & t \geq \dfrac{\tilde{c}\|\lambda(\mathbf{x})\|_2^2}{\|\lambda(\mathbf{x})\|_1\|\lambda(\mathbf{x})\|_\infty} \end{cases} \end{aligned}$$

with $\lambda(\mathbf{x})$ denoting the vector

$$\lambda(\mathbf{x}) := \begin{pmatrix} \|\mathbf{x}_1\|_2^2 \\ \vdots \\ \|\mathbf{x}_L\|_2^2 \end{pmatrix} \in \mathbb{R}^L.$$

Unlike the concentration inequality (5.3) for dense operators, the concentration result above depends on the local properties of the individual component signals through the vector $\lambda(\mathbf{x})$. More precisely, when t is small as assumed in most applications, the speed at which the tail probability decays is controlled by the ratio $\Lambda(\mathbf{x}) := \|\lambda(\mathbf{x})\|_1^2 / \|\lambda(\mathbf{x})\|_2^2$ which measures how much the signal energy concentrates on individual signal blocks $\mathbf{x}_l \in \mathbb{R}^d$. The behavior is most favorable in case each component signal $\mathbf{x}_l$ contains the same energy $\|\mathbf{x}_l\|_2 = \|\mathbf{x}_k\|_2 \; \forall k, l \in [L]$ in which case $\Lambda(\mathbf{x}) = L$. This behavior is intuitively expected since the signal energy is uniformly spread across the entire signal $\mathbf{x}$ such that each sensor always captures a certain portion of the signal energy, implying that each measurement vector $\mathbf{y}_l = \mathbf{\Phi}_l \mathbf{x}_l$ contributes information about the compound signal $\mathbf{x}$. At the other extreme where $\mathbf{x}_l = \mathbf{0}$ for all but one index $l \in [L]$ and thus $\Lambda(\mathbf{x}) = 1$, the decay behavior is least favorable since only a single vector $\mathbf{y}_l$ carries information about $\mathbf{x}$ which has to be compensated for by acquiring more measurements. This behavior is confirmed in [Wak+10] via numerical experiments.

Assuming that the elements $\mathbf{\Phi}_l$ of the block diagonal matrix $\mathbf{A}$ are replaced by a copy of the same subgaussian random matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times d}$ drawn once, Rozell *et al.* establish in [Roz+10] the almost identical concentration bound

$$\begin{aligned} &\mathbb{P}\Big(\big|M^{-1}\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \geq t\|\mathbf{x}\|_2^2\Big) \\ &\leq 2\begin{cases} \exp\left(-ct^2 M \dfrac{\big\|\tilde{\lambda}(\mathbf{x})\big\|_1^2}{\big\|\tilde{\lambda}(\mathbf{x})\big\|_2^2}\right), & 0 \leq t \leq \dfrac{\tilde{c}\big\|\tilde{\lambda}(\mathbf{x})\big\|_2^2}{\big\|\tilde{\lambda}(\mathbf{x})\big\|_1 \big\|\tilde{\lambda}(\mathbf{x})\big\|_\infty} \\ \exp\left(-c' t M \dfrac{\big\|\tilde{\lambda}(\mathbf{x})\big\|_1}{\big\|\tilde{\lambda}(\mathbf{x})\big\|_\infty}\right), & t \geq \dfrac{\tilde{c}\big\|\tilde{\lambda}(\mathbf{x})\big\|_2^2}{\big\|\tilde{\lambda}(\mathbf{x})\big\|_1 \big\|\tilde{\lambda}(\mathbf{x})\big\|_\infty} \end{cases} \end{aligned}$$

where the term $\tilde{\lambda}(\mathbf{x}) \in \mathbb{R}^d$ corresponds to the vector of eigenvalues of the matrix $\mathbf{X}^\top \mathbf{X} = \sum_{l=1}^L \mathbf{x}_l \mathbf{x}_l^\top \in \mathbb{R}^{d \times d}$ with

$$\mathbf{X} := \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_L^\top \end{pmatrix} \in \mathbb{R}^{L \times d}.$$

Similar to before, we consider the scenario in which t is small so that the tail decay is controlled by $\tilde{\Lambda}(\mathbf{x}) := \|\tilde{\lambda}(\mathbf{x})\|_1^2 / \|\tilde{\lambda}(\mathbf{x})\|_2^2$. The least favorable scenario corresponds to the case where each $\mathbf{x}_l$ is identical up to scaling, *i.e.*, $\mathbf{x}_l = \alpha_l \mathbf{z}$ with $\mathbf{0} \neq \mathbf{z} \in \mathbb{R}^d$ and $\alpha_l \in \mathbb{R}$. This implies that the matrix $\mathbf{X}^\top \mathbf{X} = \sum_{l=1}^L \mathbf{x}_l \mathbf{x}_l^\top = \mathbf{z}\mathbf{z}^\top \sum_{l=1}^L \alpha_l^2$ has rank 1 and thus $\tilde{\Lambda}(\mathbf{x}) = 1$ since the vector $\tilde{\lambda}(\mathbf{x})$ is 1-sparse. On the other hand, if $L \leq d$ and the L leading eigenvalues of $\mathbf{X}$ are nonzero and identical, then $\tilde{\Lambda}(\mathbf{x}) = L$. Since the nonzero eigenvalues of $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{X}\mathbf{X}^\top$ coincide and $\operatorname{tr}(\mathbf{X}^\top \mathbf{X}) = \operatorname{tr}(\mathbf{X}\mathbf{X}^\top) = \sum_{l=1}^L \|\mathbf{x}_l\|_2^2$, this requires $\mathbf{x}_l \perp \mathbf{x}_k \; \forall k \neq l$ with $\|\mathbf{x}_l\| = \|\mathbf{x}_k\| \; \forall k, l \in [L]$. Barring the orthogonality

condition, this is the same requirement leading to the most favorable tail decay behavior as in the previous setting where each sensor was equipped with an independent copy of a subgaussian random matrix. Unsurprisingly, the restrictions on the signal ensemble yielding the most favorable tail decay are more demanding when each sensor shares the same measurement matrix since each measurement vector $\mathbf{y}_l$ potentially yields less diversity if the underlying component signals are too similar. Most importantly, the numerical experiments confirm that the parameters $\Lambda(\mathbf{x})$ and $\tilde{\Lambda}(\mathbf{x})$ capture precisely the required "oversampling" rate required to match the concentration behavior between the most and least favorable scenarios. These phenomena were further investigated in [Par+11] where it was also pointed out that the obtained results are not strong enough to provide RIP-type results for block diagonal matrices by appealing to covering arguments as employed in [Bar+08] or [MPT08].

While [Wak+10; Roz+10; Par+11] already give some indication about the expected reconstruction behavior based on the energy distribution of the types of signal ensembles one aims to recover, the work did ultimately not result in a proof of the restricted isometry property for block diagonal random matrices. Instead, such a result, which was obtained via independent methods, was first reported for block diagonal matrices $\mathbf{A} = \operatorname{diag}\{\mathbf{\Phi}_l\}_{l=1}^{L}$ populated by independent copies $\mathbf{\Phi}_l$ of a standard Gaussian random matrix $\mathbf{\Phi} \in \mathbb{R}^{m\times d}$ in [Yap+11]. In particular, given a sparsity basis $\mathbf{\Psi} \in \mathrm{U}(D)$, the authors reduce the problem of establishing the restricted isometry property of $\mathbf{A}$ to the problem of asserting the concentration behavior of the random variable $|||(\mathbf{A}\mathbf{\Psi})^*\mathbf{A}\mathbf{\Psi} - \mathrm{Id}_D|||$ where $|||\cdot|||\colon \mathbb{C}^{D\times D} \to \mathbb{R}$ denotes the matrix norm defined as

$$|||\mathbf{B}||| := \sup_{\mathbf{x}\in\Sigma_s\cap\mathbb{B}_2^D} |\langle \mathbf{x}, \mathbf{B}\mathbf{x}\rangle_{\mathbb{C}}|$$

with $\langle\cdot,\cdot\rangle_{\mathbb{C}}$ corresponding to the standard sesquilinear inner product on $\mathbb{C}^D$. Using this formulation, they show that s-sparse vectors can be stably and robustly recovered from $M = mL = \Omega(s\tilde{\mu}(\mathbf{\Psi})^2 \log(s)^2 \log(D)^4)$ measurements with probability at least $1 - 8D^{-1}$ where

$$\tilde{\mu}(\mathbf{\Psi}) := \min\left\{\sqrt{D}\max_{i,j\in[D]}|\psi_{ij}|, \sqrt{L}\right\} = \sqrt{L}\min\left\{\sqrt{d}\max_{i,j\in[D]}|\psi_{ij}|, 1\right\} \tag{5.4}$$

measures the coherence of the sparsity basis with the canonical basis for $\mathbb{C}^D$. This coherence parameter reduces to 1 in case $\mathbf{\Psi}$ corresponds to the orthogonal DFT matrix and to $\sqrt{L}$ if $\mathbf{\Psi} = \mathrm{Id}_D$. This corresponds to the situation previously discussed where the energy of the target signal $\mathbf{x}$ is either uniformly spread across the index set $[D]$, or (potentially) completely concentrated in a single component vector $\mathbf{x}_l$, respectively (see also Section 5.5). Unfortunately, this probability bound does not exhibit the desired exponential decay in the failure probability one usually aims for in compressed sensing to assert that the RIP holds with high probability on the draw of $\mathbf{A}$.

Inspired by a novel powerful technique for deriving probability bounds on the RIP constants of sensing matrices by means of concentration results for the suprema of chaos processes established in [KMR14], Eftekhari *et al.* eventually managed to remove the aforementioned drawbacks in [Eft+15]. In addition to establishing the RIP for subgaussian block diagonal matrices, the work improves the failure probability from $1 - 8D^{-1}$ to $1 - C\exp(-\log(D)^2\log(s)^2)$, provided that the number of measurements is

chosen according to

$$M \gtrsim s\tilde{\mu}(\mathbf{\Psi}) \log(s)^2 \log(D)^2, \tag{5.5}$$

which also improves upon the earlier result by removing the polylogarithmic factor $\log(D)^2$. Moreover, by replacing the coherence parameter $\tilde{\mu}$ with an alternative quantity measuring roughly the orthogonality between partial basis expansion matrices $\mathbf{\Psi}_1, \dots, \mathbf{\Psi}_L \in \mathbb{C}^{d \times dL}$ with

$$\mathbf{\Psi} = \begin{pmatrix} \mathbf{\Psi}_1 \\ \vdots \\ \mathbf{\Psi}_L \end{pmatrix},$$

a similar result is established in the setting where $\mathbf{A}$ is a subgaussian block diagonal random matrix with constant block diagonal, *i.e.*, every sensor is assumed to be equipped with the same random matrix.

A similar result was later obtained in [CA18] by Chun and Adcock who recover the block diagonal sensing model as a special case of so-called *parallel acquisition systems*. This model assumes that L sensors acquire L different snapshots of a target signal $\mathbf{x} \in \mathbb{C}^D$ of the form

$$\mathbf{y}_l = \mathbf{B}_l \mathbf{x} \in \mathbb{C}^m$$

with $\mathbf{B}_l := \mathbf{A}_l \mathbf{H}_l \mathbf{\Psi}$ where $\mathbf{A}_l \in \mathbb{C}^{m \times D}$ are densely populated subgaussian measurement matrices and $\mathbf{H}_l \in \mathbb{C}^{D \times D}$ denote so-called *sensor profile matrices* which model environmental properties of the sensing problem. Appealing to the same proof technique employed in [Eft+15], they establish the so-called *asymmetric restricted isometry property* (ARIP) of the compound sensing matrix

$$\mathbf{B} := \begin{pmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_L \end{pmatrix},$$

codified as

$$\alpha \|\mathbf{x}\|_2^2 \leq \|\mathbf{B}\mathbf{x}\|_2^2 \leq \beta \|\mathbf{x}\|_2^2 \quad \forall \mathbf{x} \in \Sigma_s$$

as initially proposed by Foucart and Lai in [FL09]. This generalization reduces to the canonical RIP for the choice $\alpha = 1 - \delta$ and $\beta = 1 + \delta$, while allowing for more flexibility in the choice of sensor profile matrices. In particular, the classical RIP requires that $\mathbb{E}(m^{-1}\mathbf{A}^*\mathbf{A}) = \mathrm{Id}_D$. For the measurement matrix $\mathbf{B}$ of the parallel acquisition system, this implies (after rescaling) that

$$\frac{1}{L} \sum_{l=1}^{L} \mathbf{H}_l^* \mathbf{H}_l = \mathrm{Id}_D,$$

which represents a rather stringent condition on the profile matrices $\mathbf{H}_l$. Instead, under the asymmetric restricted isometry property, it suffices that the sensor profile matrices satisfy the so-called *joint near-isometry condition*:

$$\alpha \mathrm{Id}_D \preceq \frac{1}{L} \sum_{l=1}^{L} \mathbf{H}_l^* \mathbf{H}_l \preceq \beta \mathrm{Id}_D. \tag{5.6}$$

The notation $\mathbf{A} \preceq \mathbf{B}$ for two self-adjoint matrices $\mathbf{A}, \mathbf{B}$ denotes a generalized matrix inequality which signifies that the matrix $\mathbf{B} - \mathbf{A}$ belongs to the cone of positive semidefinite matrices. If $\mathbf{P} := L^{-1} \sum_{l=1}^{L} \mathbf{H}_l^* \mathbf{H}_l$ is nondegenerate, then condition (5.6) always holds with $\alpha = \lambda_{\min}(\mathbf{P})$ and $\beta = \lambda_{\max}(\mathbf{P})$ where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalue of a quadratic matrix, respectively (see also the discussion in [CA18, Section 2.2 and 2.3]). In the block diagonal setting where the l-th sensor profile matrix corresponds to a block diagonal matrix with Id_d as its l-th block and $\mathbf{0}_d$ otherwise, Chun and Adcock obtain the condition

$$M \gtrsim \delta^{-2} s \Gamma(\mathbf{\Psi})^2 \log(s)^2 \log(D) \log(m)$$

for the RIP of a general subgaussian block diagonal matrix to hold with high probability. This improves upon the work by Eftekhari *et al.* by replacing the coherence parameter $\tilde{\mu}(\mathbf{\Psi})$ with the smaller parameter

$$\Gamma(\mathbf{\Psi}) := \sqrt{L} \max_{\substack{l \in [L], \\ i \in [D]}} \|\mathbf{\Psi}_l \mathbf{e}_i\|_2 \leq \tilde{\mu}(\mathbf{\Psi})$$

with $\mathbf{e}_i \in \mathbb{C}^D$ denoting the i-th canonical basis vector. Moreover, the bound replaces the factor $\log(D)$ in (5.5) by $\log(m)$. Note, however, that this only constitutes a minor improvement since Eftekhari *et al.* technically establish the condition

$$M \gtrsim \delta^{-2} s \tilde{\mu}(\mathbf{\Psi})^2 \log(s)^2 \log(D) \log(M),$$

which they subsequently simplify to

$$M \gtrsim \delta^{-2} s \tilde{\mu}(\mathbf{\Psi})^2 \log(s)^2 \log(D)^2$$

to remove the dependence of M on $\log(mL) = \log(M)$.

More recently, the block diagonal measurement model was analyzed by Maly and Palzer in the context of distributed compressed sensing from 1-bit observations using a back-projected hard thresholding strategy [MP19]. Given the issues related to subgaussian observations in the 1-bit acquisition model (cf. Section 4.4), they consider Gaussian block diagonal measurement matrices $\mathbf{A} = \operatorname{diag}\{\mathbf{G}_l\}_{l=1}^{L} \in \mathbb{R}^{mL \times dL}$ where each $\mathbf{G}_l \in \mathbb{R}^{m \times d}$ denotes an independent copy of a standard Gaussian random matrix. As outlined in the introduction of this chapter, the distributed compressed sensing model assumes that each sensor observes a different signal $\mathbf{x}_l \in \mathbb{R}^d$ supported on the same index set $S \subset [d]$ of size s, giving rise to the set of s-row-sparse matrices

$$\Theta_s := \left\{ \mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_L \end{pmatrix} \in \mathbb{R}^{d \times L} : |\operatorname{rowsupp}(\mathbf{X})| \leq s \right\}$$

where $\operatorname{rowsupp}\colon \mathbb{R}^{d \times L} \to [d]$ denotes the index set of nonzero rows of a matrix. To avoid adversarial row-sparse matrices, the authors of [MP19] additionally assume that each signal contains the same energy, modeled by the signal set

$$\mathcal{K}_s := \left\{ \mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_L \end{pmatrix} \in \Theta_s : \|\mathbf{x}_l\|_2 = \|\mathbf{X}\|_{\mathrm{F}} / \sqrt{L} \; \forall l \in [L] \right\}.$$

Given the fact that their work addresses the recovery of jointly-sparse vectors from binary observations of the form $\mathbf{y}_l = \operatorname{sgn}(\mathbf{G}_l \mathbf{x}_l) \in \{\pm 1\}^m$, this requirement is rather mild due to

the scale invariance of the sgn-operator. In order to establish the recovery guarantee for their proposed row hard thresholding algorithm, they introduce the following variant of the restricted isometry property:

$$\frac{\|\mathbf{X}\|_{2,1}}{\sqrt{L}} - \delta\|\mathbf{X}\|_{\mathrm{F}} \leq \|\mathbf{A}\operatorname{vec}(\mathbf{X})\|_1 \leq \frac{\|\mathbf{X}\|_{2,1}}{\sqrt{L}} + \delta\|\mathbf{X}\|_{\mathrm{F}} \quad \forall \mathbf{X} \in \Theta_s \tag{5.7}$$

with $\operatorname{vec}\colon \mathbb{R}^{d\times L} \to \mathbb{R}^{dL}$ denoting the isomorphism which stacks the columns of a matrix into a single column vector and $\|\cdot\|_{2,1}$ corresponding to the matrix norm which sums up the ℓ_2-norms of the individual rows of a matrix. Equipped with this property, they show that every matrix $\mathring{\mathbf{X}} \in \mathcal{K}_s$ can be approximated from its quantized measurements[2]

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_L \end{pmatrix} = \operatorname{sgn}(\mathbf{A}\operatorname{vec}(\mathring{\mathbf{X}}))$$

by the matrix

$$\hat{\mathbf{X}} = \sqrt{\frac{\pi}{2m^2L}}\mathcal{H}_s^{\mathrm{row}}\Big(\begin{pmatrix} \mathbf{G}_1^\top\mathbf{y}_1 & \dots & \mathbf{G}_L^\top\mathbf{y}_L \end{pmatrix}\Big)$$

where $\mathcal{H}_s^{\mathrm{row}}\colon \mathbb{R}^{d\times L} \to \Theta_s$ denotes the row hard thresholding operator which only retains the s rows with largest ℓ_2-norm. In particular, they establish that

$$\big\|\mathring{\mathbf{X}} - \hat{\mathbf{X}}\big\|_{\mathrm{F}} \lesssim \sqrt{\delta}$$

holds with probability at least $1 - 2\exp(-c\delta^2 mL)$, provided that

$$mL \gtrsim \delta^{-2}s(\log(d/s) + L).$$

We point out that $\hat{\mathbf{X}}$ does not result from a projection of $\operatorname{vec}^{-1}(\mathbf{A}^\top\operatorname{vec}(\mathbf{Y}))$ on the set $\mathcal{K}_s$ but rather on Θ_s. A projection on $\mathcal{K}_s$ would substantially complicate the proposed algorithm since it is not clear whether fast projections on $\mathcal{K}_s$ are possible. However, using similar arguments as in the proof of Theorem 8 in [Fou16], the authors show that for any matrix $\mathring{\mathbf{X}} \in \mathcal{K}_s$, the back-projected vector $\hat{\mathbf{X}}$ is at most a constant multiple of the RIP constant δ apart from $\mathring{\mathbf{X}}$ (w. r. t. the squared Frobenius norm). The main effort of the work in [MP19] is therefore concerned with establishing that Gaussian block diagonal random matrices satisfy condition (5.7) with high probability.

5.3 The Group-RIP for General Block Diagonal Matrices

In this section, we establish the group-RIP for general subgaussian block diagonal matrices. We will make use of a powerful bound on the suprema of chaos processes first established in [KMR14, Theorem 3.1] to demonstrate that the block diagonal matrix $\mathbf{A}\boldsymbol{\Psi} \in \mathbb{C}^{M\times D}$ satisfies the group restricted isometry property with high probability on the draw of $\mathbf{A}$. The same technique was also employed in [Eft+15] to prove the canonical restricted isometry property for block diagonal matrices consisting of subgaussian blocks. In the

[2] As usual, the operator sgn is assumed to act element-wise on vectors and matrices.

present work, we make use of an improved version of the bound due to Dirksen [Dir15]. Before stating the result, we first define the following objects. Let $\mathcal{M} \subset \mathbb{C}^{m\times n}$ be a bounded set. Then the radii of $\mathcal{M}$ w. r. t. the Frobenius and operator norm are defined as

$$\rho_{\mathrm{F}}(\mathcal{M}) = \sup_{\boldsymbol{\Gamma}\in\mathcal{M}} \|\boldsymbol{\Gamma}\|_{\mathrm{F}} \quad \text{and} \quad \rho_{2\to 2}(\mathcal{M}) = \sup_{\boldsymbol{\Gamma}\in\mathcal{M}} \|\boldsymbol{\Gamma}\|_{2\to 2},$$

respectively. Lastly, we require the so-called γ_2-functional of $\mathcal{M}$ w. r. t. the operator norm.

Definition 5.4. *An* admissible sequence *of a metric space (T, Δ) is a collection of subsets $\{T_r \subset T : r \geq 0\}$ where $|T_r| \leq 2^{2^r}$ for every $r \geq 1$ and $|T_0| = 1$. The γ_2-functional is then defined by*

$$\gamma_2(T, \Delta) = \inf \sup_{t\in T} \sum_{r=0}^{\infty} 2^{r/2} \Delta(t, T_r)$$

where the infimum is taken over all admissible sequences.

Characterizing the γ_2-functional directly is generally a difficult undertaking. It is therefore customary to appeal to a classical result due to Talagrand which bounds $\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2})$ in terms of the following entropy integral of the metric space[3] $(\mathcal{M}, \|\cdot\|_{2\to 2})$ [Tal10]:

$$\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \lesssim \int_0^{\infty} \sqrt{\log \mathfrak{N}(\mathcal{M}, \|\cdot\|_{2\to 2}, \varepsilon)} \mathrm{d}\varepsilon \tag{5.8}$$

where $\mathfrak{N}$ denotes the internal covering number, *i.e.*, the cardinality of the smallest subset $\mathcal{N} \subset \mathcal{M}$ such that every point in $\mathcal{M}$ is at most ε apart from $\mathcal{N}$ w. r. t. the operator norm $\|\cdot\|_{2\to 2}$ (cf. Definition A.15). Mathematically, $\mathcal{N} \subset \mathcal{M}$ is called an ε-net of $\mathcal{M}$ if $\forall \boldsymbol{\Gamma} \in \mathcal{M}$ $\exists \boldsymbol{\Gamma}_0 \in \mathcal{N} : \|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}_0\|_{2\to 2} \leq \varepsilon$ with $\mathfrak{N}(\mathcal{M}, \|\cdot\|_{2\to 2}, \varepsilon) = |\mathcal{N}|$ if $\mathcal{N}$ is the smallest such net. Note that the integrand of the entropy integral (5.8) vanishes as soon as $\varepsilon \geq \rho_{2\to 2}(\mathcal{M})$ since $\mathcal{M}$ can then be covered by a single ball $\mathbb{B}_{2\to 2}^{m\times n}$ centered at an (arbitrary) element of $\mathcal{M}$.

Theorem 5.5 ([Dir15, Theorem 6.5])**.** *Let $\mathcal{M}$ be a matrix set, and denote by $\boldsymbol{\xi}$ a zero-mean, unit-variance subgaussian random vector with independent entries and subgaussian norm $\tau = \|\boldsymbol{\xi}\|_{\psi_2}$. Then, for $u \geq 1$,*

$$\mathbb{P}\left(\sup_{\boldsymbol{\Gamma}\in\mathcal{M}} \left| \|\boldsymbol{\Gamma}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|\boldsymbol{\Gamma}\boldsymbol{\xi}\|_2^2 \right| \geq c_\tau E_u \right) \leq e^{-u}$$

where

$$\begin{aligned} E_u := {} & \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2})^2 + \rho_{\mathrm{F}}(\mathcal{M})\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \\ & + \sqrt{u}\rho_{\mathrm{F}}(\mathcal{M})\rho_{2\to 2}(\mathcal{M}) + u\rho_{2\to 2}(\mathcal{M})^2, \end{aligned}$$

and $c_\tau > 0$ is a constant that only depends on τ.

[3] The metric on $\mathcal{M}$ is the one canonically induced by the norm $\|\cdot\|_{2\to 2}$.

5.3.1 Chaos Process for Block-Diagonal Group-RIP Matrices

In order to apply Theorem 5.5 to estimate the probability that $\mathbf{A}\boldsymbol{\Psi}$ as in (5.1) satisfies the group restricted isometry property, first note that we can equivalently express the group-RIP condition in (5.2) for $\mathbf{x} \in \Sigma_{\mathcal{I},s} \setminus \{\mathbf{0}\}$ as

$$\left| \frac{\|\mathbf{A}\boldsymbol{\Psi}\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} - 1 \right| \leq \delta.$$

With the definition of the set

$$\Omega := \Sigma_{\mathcal{I},s} \cap \mathbb{S}^{D-1} = \left\{ \mathbf{x} \in \mathbb{S}^{D-1} : \|\mathbf{x}\|_{\mathcal{I},0} \leq s \right\}$$

of s-group-sparse vectors on the unit Euclidean sphere, we may therefore express the group restricted isometry constant of $\mathbf{A}\boldsymbol{\Psi}$ as

$$\delta_s = \sup_{\mathbf{x} \in \Omega} \left| \|\mathbf{A}\boldsymbol{\Psi}\mathbf{x}\|_2^2 - 1 \right|. \tag{5.9}$$

Next, we transform the above expression into the form required by Theorem 5.5 following the ideas in [**Eft+15**], *i.e.*, we rewrite the equation so that the supremum is taken over a matrix set. To that end, recall the definition of the partial basis expansion matrices $\boldsymbol{\Psi}_l \in \mathbb{C}^{d \times dL}$ with $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1^\top \ \ldots \ \boldsymbol{\Psi}_L^\top)^\top$. In light of (5.1), we may now express the l-th measurement vector $\mathbf{y}_l \in \mathbb{C}^m$ of $\mathbf{y} \in \mathbb{C}^{mL}$ as

$$\begin{aligned} \mathbf{y}_l = \boldsymbol{\Phi}_l \boldsymbol{\Psi}_l \mathbf{x} &= \begin{pmatrix} \langle (\boldsymbol{\Phi}_l)_1, \boldsymbol{\Psi}_l \mathbf{x} \rangle \\ \vdots \\ \langle (\boldsymbol{\Phi}_l)_m, \boldsymbol{\Psi}_l \mathbf{x} \rangle \end{pmatrix} \\ &= \underbrace{\begin{pmatrix} (\boldsymbol{\Psi}_l \mathbf{x})^\top & & \\ & \ddots & \\ & & (\boldsymbol{\Psi}_l \mathbf{x})^\top \end{pmatrix}}_{=: V_l(\mathbf{x}) \in \mathbb{C}^{m \times md}} \cdot \underbrace{\begin{pmatrix} (\boldsymbol{\Phi}_l)_1 \\ \vdots \\ (\boldsymbol{\Phi}_l)_m \end{pmatrix}}_{=: \boldsymbol{\xi}_l \in \mathbb{R}^{md}} \end{aligned} \tag{5.10}$$

where $(\boldsymbol{\Phi}_l)_i \in \mathbb{C}^d$ denotes the i-th row of the matrix $\boldsymbol{\Phi}_l$. If the blocks $\boldsymbol{\Phi}_l$ are populated by independent copies of a τ-subgaussian random variable with unit-variance, then the vector $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^\top \ \ldots \ \boldsymbol{\xi}_L^\top)^\top$ is a unit-variance τ-subgaussian random vector. Defining the operator $V \colon \mathbb{C}^{dL} \to \mathbb{C}^{mL \times mdL}$ with

$$\mathbf{x} \mapsto V(\mathbf{x}) = \operatorname{diag}\left\{ V_l(\mathbf{x}) \right\}_{l=1}^L, \tag{5.11}$$

we therefore have $\mathbf{A}\boldsymbol{\Psi}\mathbf{x} \overset{\mathrm{d}}{=} V(\mathbf{x})\boldsymbol{\xi}$ where $\overset{\mathrm{d}}{=}$ denotes equality in distribution. Now note that

$$\mathbb{E}\|\mathbf{A}\boldsymbol{\Psi}\mathbf{x}\|_2^2 = \mathbf{x}^* \boldsymbol{\Psi}^* \mathbb{E}\left[\mathbf{A}^\top \mathbf{A}\right] \boldsymbol{\Psi}\mathbf{x} = m\|\mathbf{x}\|_2^2,$$

which follows from the fact that the rows of the matrices $\mathbf{A}_l$ are independent unit-variance random m-vectors with independent entries, as well as from unitarity of $\boldsymbol{\Psi}$. With (5.9), the group restricted isometry property of the matrix $1/\sqrt{m}\mathbf{A}\boldsymbol{\Psi}$ can therefore be expressed

as

$$\begin{aligned}\delta_s\left(\frac{1}{\sqrt{m}}\mathbf{A}\boldsymbol{\Psi}\right) &= \sup_{\mathbf{x}\in\Omega}\left|\left\|\frac{1}{\sqrt{m}}\mathbf{A}\boldsymbol{\Psi}\mathbf{x}\right\|_2^2 - 1\right| \\ &= \sup_{\mathbf{x}\in\Omega}\left|\frac{1}{m}\|\mathbf{A}\boldsymbol{\Psi}\mathbf{x}\|_2^2 - \frac{1}{m}m\|\mathbf{x}\|_2^2\right| \\ &= \frac{1}{m}\sup_{\mathbf{x}\in\Omega}\left|\|\mathbf{A}\boldsymbol{\Psi}\mathbf{x}\|_2^2 - \mathbb{E}\|\mathbf{A}\boldsymbol{\Psi}\mathbf{x}\|_2^2\right| \\ &\overset{\mathrm{d}}{=} \frac{1}{m}\sup_{\mathbf{x}\in\Omega}\left|\|V(\mathbf{x})\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|V(\mathbf{x})\boldsymbol{\xi}\|_2^2\right| \\ &= \frac{1}{m}\sup_{\boldsymbol{\Gamma}\in\mathcal{M}}\left|\|\boldsymbol{\Gamma}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|\boldsymbol{\Gamma}\boldsymbol{\xi}\|_2^2\right| \end{aligned} \tag{5.12}$$

where we set $\mathcal{M} := V(\Omega) = \{V(\mathbf{x}) : \mathbf{x} \in \Omega\}$. In order to apply Theorem 5.5, it remains to estimate the radii of $\mathcal{M}$ w. r. t. the Frobenius and operator norm, respectively, as well as to compute the γ_2-functional of $\mathcal{M}$ w. r. t. $\|\cdot\|_{2\to 2}$. These issues are addressed in the next two sections.

5.3.2 Radii Estimates

We begin with the estimation of $\rho_{\mathrm{F}}(\mathcal{M})$. To that end, first note that

$$\begin{aligned}\|V(\mathbf{x})\|_{\mathrm{F}}^2 &= \left\|\operatorname{diag}\left\{V_l(\mathbf{x})\right\}_{l=1}^L\right\|_{\mathrm{F}}^2 = \sum_{l=1}^L \|V_l(\mathbf{x})\|_{\mathrm{F}}^2 \\ &= \sum_{l=1}^L m\|\boldsymbol{\Psi}_l\mathbf{x}\|_2^2 = m\|\boldsymbol{\Psi}\mathbf{x}\|_2^2 = m\|\mathbf{x}\|_2^2.\end{aligned}$$

Since $\Omega \subset \mathbb{S}^{D-1}$, this immediately implies

$$\rho_{\mathrm{F}}(\mathcal{M}) = \sup_{\boldsymbol{\Gamma}\in\mathcal{M}}\|\boldsymbol{\Gamma}\|_{\mathrm{F}} = \sup_{\mathbf{x}\in\Omega}\|V(\mathbf{x})\|_{\mathrm{F}} = \sqrt{m}\sup_{\mathbf{x}\in\Omega}\|\mathbf{x}\|_2 = \sqrt{m}.$$

In order to estimate the radius $\rho_{2\to 2}(\mathcal{M})$, we require a simple generalization of Hölder's inequality to group ℓ_p-norms on $\mathbb{C}^D$ as defined in Definition 4.2. We state here a specialization to the conjugate pair $p = 1$, $q = \infty$.

Lemma 5.6. *Let $\mathbf{a}, \mathbf{b} \in \mathbb{C}^D$, and let $\mathcal{I}$ be a group partition of $[D]$. Then*

$$|\langle \mathbf{a}, \mathbf{b}\rangle| \le \|\mathbf{a}\|_{\mathcal{I},1} \cdot \|\mathbf{b}\|_{\mathcal{I},\infty}$$

where $\langle\cdot,\cdot\rangle$ denotes the bilinear form $\langle \mathbf{a}, \mathbf{b}\rangle = \sum_{i=1}^D a_i b_i$ on $\mathbb{C}^D$.

Proof. By the triangle and Hölder's inequality, we have

$$\begin{aligned}|\langle \mathbf{a}, \mathbf{b}\rangle| &= \left|\sum_{i=1}^G \langle \mathbf{a}_{\mathcal{I}_i}, \mathbf{b}_{\mathcal{I}_i}\rangle\right| \le \sum_{i=1}^G |\langle \mathbf{a}_{\mathcal{I}_i}, \mathbf{b}_{\mathcal{I}_i}\rangle| \le \sum_{i=1}^G \|\mathbf{a}_{\mathcal{I}_i}\|_2 \cdot \|\mathbf{b}_{\mathcal{I}_i}\|_2 \\ &\le \sum_{i=1}^G \|\mathbf{a}_{\mathcal{I}_i}\|_2 \cdot \max_{j\in[G]}\left\|\mathbf{b}_{\mathcal{I}_j}\right\|_2 = \|\mathbf{a}\|_{\mathcal{I},1} \cdot \|\mathbf{b}\|_{\mathcal{I},\infty}.\end{aligned}$$

□

We proceed as before and calculate

$$\begin{aligned}\|V(\mathbf{x})\|_{2\to 2} &= \left\|\operatorname{diag}\left\{V_l(\mathbf{x})\right\}_{l=1}^{L}\right\|_{2\to 2} = \max_{l\in[L]} \|V_l(\mathbf{x})\|_{2\to 2} \\ &= \max_{l\in[L]} \|V_l(\mathbf{x})V_l(\mathbf{x})^*\|_{2\to 2}^{1/2} = \max_{l\in[L]} \|\boldsymbol{\Psi}_l\mathbf{x}\|_2 \end{aligned} \tag{5.13}$$

where the second step follows from the fact that the operator norm of a block diagonal matrix corresponds to the maximum operator norm of the individual blocks. The last step follows because $V_l(\mathbf{x})V_l(\mathbf{x})^*$ is a diagonal matrix with m copies of $\|\boldsymbol{\Psi}_l\mathbf{x}\|_2^2$ on its diagonal whose largest singular value is simply $\|\boldsymbol{\Psi}_l\mathbf{x}\|_2^2$. Next, we invoke the bound $\|\mathbf{u}\|_2 \leq \sqrt{n}\|\mathbf{u}\|_\infty$ for $\mathbf{u}\in\mathbb{C}^n$, followed by an application of Lemma 5.6. This yields

$$\begin{aligned}\|\boldsymbol{\Psi}_l\mathbf{x}\|_2 &\leq \sqrt{d}\|\boldsymbol{\Psi}_l\mathbf{x}\|_\infty = \sqrt{d}\max_{i\in[d]} |\langle (\boldsymbol{\Psi}_l)_i, \mathbf{x}\rangle| \\ &\leq \sqrt{d}\max_{i\in[d]} \|(\boldsymbol{\Psi}_l)_i\|_{\mathcal{I},\infty}\cdot\|\mathbf{x}\|_{\mathcal{I},1}\end{aligned}$$

where $(\boldsymbol{\Psi}_l)_i$ denotes the i-th row of $\boldsymbol{\Psi}_l$. Overall, we find

$$\begin{aligned}\|V(\mathbf{x})\|_{2\to 2} &\leq \sqrt{d}\|\mathbf{x}\|_{\mathcal{I},1}\max_{\substack{l\in[L],\\ i\in[d]}} \|(\boldsymbol{\Psi}_l)_i\|_{\mathcal{I},\infty} \\ &= \sqrt{d}\|\mathbf{x}\|_{\mathcal{I},1}\max_{i\in[D]} \|\boldsymbol{\psi}_i\|_{\mathcal{I},\infty}\end{aligned}$$

where $\boldsymbol{\psi}_i\in\mathbb{C}^D$ denotes the i-th row of $\boldsymbol{\Psi}$. This bound turns out to be too loose when $\boldsymbol{\Psi}=\mathrm{Id}_D$ in which case we have $\sqrt{d}\max_{i\in[D]}\|\boldsymbol{\psi}_i\|_{\mathcal{I},\infty} = \sqrt{d}$. To obtain a more effective bound, we therefore also consider the simple bound

$$\begin{aligned}\|V(\mathbf{x})\|_{2\to 2} &= \max_{l\in[L]} \|\boldsymbol{\Psi}_l\mathbf{x}\|_2 \leq \|\boldsymbol{\Psi}\mathbf{x}\|_2 \\ &= \|\mathbf{x}\|_2 = \|\mathbf{x}\|_{\mathcal{I},2} \leq \|\mathbf{x}\|_{\mathcal{I},1},\end{aligned} \tag{5.14}$$

which follows from $\|\cdot\|_p \leq \|\cdot\|_q$ for $p\geq q\geq 1$. By defining the parameter

$$\mu_{\mathcal{I}}(\boldsymbol{\Psi}) := \min\left\{\sqrt{d}\max_{i\in[D]}\|\boldsymbol{\psi}_i\|_{\mathcal{I},\infty}, 1\right\},$$

we therefore arrive at

$$\|V(\mathbf{x})\|_{2\to 2} \leq \mu_{\mathcal{I}}(\boldsymbol{\Psi})\|\mathbf{x}\|_{\mathcal{I},1} \tag{5.15}$$

after combining both estimates for $\|V(\mathbf{x})\|_{2\to 2}$ which in turn yields

$$\begin{aligned}\rho_{2\to 2}(\mathcal{M}) &= \sup_{\mathbf{x}\in\Omega}\|V(\mathbf{x})\|_{2\to 2} \\ &\leq \mu_{\mathcal{I}}(\boldsymbol{\Psi})\sup_{\mathbf{x}\in\Omega}\|\mathbf{x}\|_{\mathcal{I},1} \\ &\leq \mu_{\mathcal{I}}(\boldsymbol{\Psi})\sqrt{s}.\end{aligned}$$

The last inequality holds since for $\mathbf{x}\in\Omega = \Sigma_{\mathcal{I},s}\cap\mathbb{S}^{D-1}$, we have

$$\begin{aligned}\|\mathbf{x}\|_{\mathcal{I},1} = \sum_{i=1}^{G}\|\mathbf{x}_{\mathcal{I}_i}\|_2 &\leq \left(\sum_{i=1}^{G}\|\mathbf{x}_{\mathcal{I}_i}\|_2^2\right)^{1/2}\left(\sum_{i=1}^{G}\mathbb{1}_{\{\mathbf{x}_{\mathcal{I}_i}\neq\mathbf{0}\}}\right)^{1/2} \\ &\leq \|\mathbf{x}\|_{\mathcal{I},2}\sqrt{s} = \sqrt{s}\|\mathbf{x}\|_2 = \sqrt{s}\end{aligned}$$

by the Cauchy-Schwarz inequality.

Since the parameter $\mu_{\mathcal{I}}(\mathbf{\Psi})$ will play a central role later on, some comments are in order. First, let us point out that with the trivial group partition $\widetilde{\mathcal{I}} = \{\{1\}, \dots, \{D\}\}$, the parameter $\mu_{\widetilde{\mathcal{I}}}(\mathbf{\Psi})$ reduces to the coherence parameter (up to scaling by $\sqrt{L}$) considered in [Eft+15]. In that case, the term $\max_{i\in[D]} \|\boldsymbol{\psi}_i\|_\infty$ corresponds to the constant associated with the bounded orthonormal system generated by the columns of the unitary matrix $\mathbf{\Psi}$ as defined in [FR13, Chapter 12]. In general, the term $\mu_{\mathcal{I}}(\mathbf{\Psi})$ measures how coherent the sparsity basis is with the canonical basis for $\mathbb{C}^D$. For instance, we clearly have $\mu_{\mathcal{I}}(\mathrm{Id}_D) = 1$. At the other end of the spectrum, we have for the orthogonal DFT matrix

$$\mathbf{F}_D := \frac{1}{\sqrt{D}}\left(e^{i2\pi\mu\nu/D}\right)_{0\le\mu,\nu\le D-1}$$

that $\mu_{\mathcal{I}}(\mathbf{F}_D) = \min\{\sqrt{g/L}, 1\}$ since every entry of $\mathbf{F}_D$ has constant modulus and hence $\|\boldsymbol{\psi}_i\|_{\mathcal{I},\infty} = \sqrt{g/D}\ \forall i \in [D]$. This implies that the bound on $\rho_{2\to2}(\mathcal{M})$ becomes more effective the more sensors one considers.

5.3.3 Metric Entropy Bound

Establishing a bound on the γ_2-functional via (5.8) will proceed in two steps. At small scales, we estimate the covering number by means of a standard volume comparison argument for norm balls covered in their respective metric. At larger scales, however, this bound is not be effective enough to yield optimal scaling behavior in s. To circumvent the problem, we employ a variation of Maurey's empirical method, which we develop below.

To start with, note that with $\|\mathbf{x}\|_V := \|V(\mathbf{x})\|_{2\to2}$, we have for $u \ge 0$ that

$$\mathfrak{N}(\mathcal{M}, \|\cdot\|_{2\to2}, u) = \mathfrak{N}(\Omega, \|\cdot\|_V, u)$$

since $\mathcal{M} = V(\Omega)$ by definition. With this we decompose the metric entropy integral as

$$\begin{aligned}&\int_0^{\rho_{2\to2}(\mathcal{M})} \sqrt{\log \mathfrak{N}(\mathcal{M}, \|\cdot\|_{2\to2}, \varepsilon)}\mathrm{d}\varepsilon \\ &= \int_0^{\lambda} \sqrt{\log \mathfrak{N}(\Omega, \|\cdot\|_V, \varepsilon)}\mathrm{d}\varepsilon + \int_\lambda^{\sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})} \sqrt{\log \mathfrak{N}(\Omega, \|\cdot\|_V, \varepsilon)}\mathrm{d}\varepsilon \end{aligned} \tag{5.16}$$

where the parameter $\lambda \in [0, \sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})]$ will be chosen later.

Estimation at Small Scales

At smaller covering radii, we use a common technique to estimate the covering number of a set which can be expressed as the union of simpler sets restricted to lower-dimensional coordinate subspaces. In particular, we express the set $\Omega = \Sigma_{\mathcal{I},s} \cap \mathbb{S}^{D-1}$ of s-group-sparse signals on the unit sphere as the union of $\binom{G}{s}$ unit Euclidean spheres supported on s groups of a group partition $\mathcal{I}$. Denote for $\mathcal{T} \subset \mathcal{I}$ the coordinate subspace of $\mathbb{C}^D$ supported on the index set $\bigcup_{S\in\mathcal{T}} S \subset [D]$ by $\mathbb{C}^D_{\mathcal{T}}$, *i.e.*,

$$\mathbb{C}^D_{\mathcal{T}} = \left\{\mathbf{x} \in \mathbb{C}^D : \mathbf{x}_S = \mathbf{0}\ \forall S \notin \mathcal{T}\right\}.$$

Then we can write

$$\Omega = \bigcup_{\substack{\mathcal{T}\subset\mathcal{I},\\ |\mathcal{T}|=s}} (\mathbb{S}^{D-1}\cap\mathbb{C}^D_{\mathcal{T}}) \subset \bigcup_{\substack{\mathcal{T}\subset\mathcal{I},\\ |\mathcal{T}|=s}} (\mathbb{B}^D_2\cap\mathbb{C}^D_{\mathcal{T}}).$$

The linear-algebraic dimension of the sets in this union is at most sg where g denotes the largest group of the partition $\mathcal{I}$ considered in $\mathcal{T}$. From the volume comparison argument for norm balls covered in their associated metrics (see Lemma A.19), one has that $\mathfrak{N}(\mathbb{B}^n_{\|\cdot\|},\|\cdot\|,t)\leq(1+2/t)^n$. With $\|\cdot\|_V\leq\|\cdot\|_2$ (cf. (5.14)), this yields for an arbitrary group index set $\mathcal{T}$ as above that

$$\begin{aligned}\mathfrak{N}(\Omega,\|\cdot\|_V,u) &\leq \binom{G}{s}\mathfrak{N}(\mathbb{B}^D_2\cap\mathbb{C}^D_{\mathcal{T}},\|\cdot\|_2,u/2)\\ &\leq \left(\frac{eG}{s}\right)^s\left(1+\frac{4}{u}\right)^{2sg}\end{aligned} \tag{5.17}$$

where the factor $1/2$ in the covering radius of the first estimate is due to the fact that the internal covering numbers are only almost increasing by inclusion, *i.e.*, if $U\subset W$, then $\mathfrak{N}(U,\cdot,t)\leq\mathfrak{N}(W,\cdot,t/2)$ (cf. Proposition A.18). The factor 2 in the exponent of the last estimate is due to the isomorphic identification of $\mathbb{C}^n$ with $\mathbb{R}^{2n}$. Finally, we invoked the standard bound $\binom{n}{k}\leq(en/k)^k$ for binomial coefficients.

Estimation at Higher Scales

To estimate $\mathfrak{N}(\Omega,\|\cdot\|_V,u)$ at higher scales, one possible strategy might be to appeal to Sudakov's inequality, relating the covering number of a superset of Ω to its mean width. As it turns out, this will yield the correct dependence on s albeit with a significant caveat, namely that $\gamma_2(\mathcal{M},\|\cdot\|_{2\to2})$ will not depend on $\mu_{\mathcal{I}}(\mathbf{\Psi})$ in a linear fashion but only logarithmically. This in turn means that we do not profit from nonlocalized unitary bases as we will see later.

To make matters concrete, we assume for the moment that we are working in $\mathbb{R}^D$ rather than $\mathbb{C}^D$. Next, note that we have by the Cauchy-Schwarz inequality that

$$\frac{\Omega}{\sqrt{s}}\subset\mathbb{B}^D_{\mathcal{I},1}$$

as argued before, and recall from (5.14) that $\|\cdot\|_V\leq\|\cdot\|_2$. By a change of variable, this yields for the entropy integral that

$$\begin{aligned}\int_\lambda^{\rho_{2\to2}(\mathcal{M})}\sqrt{\log\mathfrak{N}(\Omega,\|\cdot\|_V,u)}\mathrm{d}u &\leq \int_\lambda^{\sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})}\sqrt{\log\mathfrak{N}\left(\frac{\Omega}{\sqrt{s}},\|\cdot\|_2,\frac{u}{\sqrt{s}}\right)}\mathrm{d}u\\ &\leq \sqrt{s}\int_{\lambda/\sqrt{s}}^{\mu_{\mathcal{I}}(\mathbf{\Psi})}\sqrt{\log\mathfrak{N}\left(\mathbb{B}^D_{\mathcal{I},1},\|\cdot\|_2,\frac{u}{2}\right)}\mathrm{d}u \qquad (5.18)\\ &\lesssim \sqrt{s}\int_{\lambda/\sqrt{s}}^{\mu_{\mathcal{I}}(\mathbf{\Psi})}u^{-1}w(\mathbb{B}^D_{\mathcal{I},1})\mathrm{d}u \qquad (5.19)\end{aligned}$$

where the last step follows due to Sudakov minoration[4] (cf. Lemma A.21). To estimate the mean width of the norm ball $\mathbb{B}^D_{\mathcal{I},1}$, note that we have with $\mathbf{g} \in \mathbb{R}^D$ denoting a standard Gaussian random vector as usual that

$$\begin{aligned} w(\mathbb{B}^D_{\mathcal{I},1}) &= \mathbb{E} \sup_{\|\mathbf{x}\|_{\mathcal{I},1} \leq 1} \langle \mathbf{x}, \mathbf{g} \rangle = \mathbb{E} \|\mathbf{g}\|^*_{\mathcal{I},1} \\ &= \mathbb{E} \|\mathbf{g}\|_{\mathcal{I},\infty} = \mathbb{E} \max_{i \in [G]} \left\| \mathbf{g}_{\mathcal{I}_i} \right\|_2 \end{aligned}$$

since $\|\cdot\|_{\mathcal{I},\infty}$ is the dual norm of $\|\cdot\|_{\mathcal{I},1}$ (see, *e.g.*, [**Sra12**, Lemma 2]). By the same arguments as in the proof of Lemma 4.10, we further have

$$\begin{aligned} \mathbb{E} \max_{i \in [G]} \left\| \mathbf{g}_{\mathcal{I}_i} \right\|_2 &\leq \max_{i \in [G]} \mathbb{E} \left\| \mathbf{g}_{\mathcal{I}_i} \right\|_2 + \mathbb{E} \max_{i \in [G]} \left| \left\| \mathbf{g}_{\mathcal{I}_i} \right\|_2 - \mathbb{E} \left\| \mathbf{g}_{\mathcal{I}_i} \right\|_2 \right| \\ &\leq \sqrt{g} + \sqrt{2 \log(2G)} \end{aligned}$$

where the first term in the last estimate is due to Jensen's inequality, and the second one follows by Theorem A.5 and Proposition A.7. Invoking this estimate in (5.19) therefore yields

$$\int_{\lambda}^{\sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})} \sqrt{\log \mathfrak{N}(\Omega, \|\cdot\|_V, u)} \mathrm{d}u \lesssim \left(\sqrt{sg} + \sqrt{s \log(G)} \right) \log \left(\frac{\sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})}{\lambda} \right).$$

As we will see in Section 5.3.4, the term E_u in Theorem 5.5 is dominated by the γ_2-functional $\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2})^2$, which implies with the above bound that m depends linearly on s. This is known to be optimal to yield stable recovery guarantees in the block-sparse setting (see Section 5.5.3). This bound is still not effective enough, however, since we cannot capitalize on the effect of $\mu_{\mathcal{I}}(\mathbf{\Psi})$ in the acquisition model.

Apart from the volume comparison argument discussed above and Sudakov's inequality, a third commonly used technique to estimate covering numbers is Maurey's lemma, also known as Maurey's empirical method. In general, Maurey's lemma is concerned with the following question. Given a vector $\mathbf{x}$ in the convex hull of a finite set $U \subset \mathbb{R}^n$, how many elements of U are needed to approximate $\mathbf{x}$ within a desired level of accuracy? Maurey's empirical method answers this question by constructing a sequence of random vectors and estimating the number of elements required for the expected average to fall below a specific distance to $\mathbf{x}$. Unfortunately, unless the number of groups in the partition $\mathcal{I}$ is identical to the ambient dimension D, the group ℓ_1 unit ball can not be expressed as the convex hull of a finite set.[5] We will circumvent this problem by an additional covering argument.

Let $\mathbf{x} \in \mathbb{B}^D_{\mathcal{I},1}$ such that $\sum_{i=1}^G \|\mathbf{x}_{\mathcal{I}_i}\|_2 \leq 1$, and denote by $S \subset [G]$ the index set of nonzero groups of $\mathbf{x}$. Then we can express $\mathbf{x}$ as

$$\mathbf{x} = \sum_{j \in S} \mathbf{x}_{\mathcal{I}_j} = \sum_{j \in S} \left\| \mathbf{x}_{\mathcal{I}_j} \right\|_2 \underbrace{\frac{\mathbf{x}_{\mathcal{I}_j}}{\left\| \mathbf{x}_{\mathcal{I}_j} \right\|_2}}_{\in \mathbb{S}^{D-1}_{\mathcal{I}_j}} \tag{5.20}$$

[4] The covering radius of $u/2$ in (5.18) rather than u as assumed in Sudakov's inequality merely amounts to a multiplicative constant in (5.19), which we absorb in the notation.

[5] For instance, the group ℓ_1-ball in $\mathbb{R}^2$ for $G = 1$ (and therefore $g = 2$) corresponds to the ℓ_2-ball $\mathbb{B}^2_2$.

where $\mathbb{S}^{D-1}_{\mathcal{I}_j}$ denotes the subset of the complex unit sphere in $\mathbb{C}^D$ supported on an index set $\mathcal{I}_j$. Since Maurey's lemma is concerned with the estimation of the covering number of the convex hull of a finite point cloud w. r. t. an arbitrary metric, the argument does not immediately extend to the current setting. This is due to fact for every $\mathbf{x} \in \mathbb{B}^D_{\mathcal{I},1}$, the dictionary

$$U_{\mathbf{x}} := \left\{ \frac{\mathbf{x}_{\mathcal{I}_j}}{\left\|\mathbf{x}_{\mathcal{I}_j}\right\|_2} : j \in S \right\}$$

such that $\mathbf{x} \in \operatorname{conv}(U_{\mathbf{x}})$ depends on the particular choice of $\mathbf{x}$. In other words, since $\mathbb{B}^D_{\mathcal{I},1}$ does not generally admit a polytope representation, there exists no finite set $U \subset \mathbb{C}^D$ such that $\mathbb{B}^D_{\mathcal{I},1} = \operatorname{conv}(U)$.

To deal with the issue outlined above, we establish the following result, which generalizes Maurey's lemma to more complicated sets. With some abuse of notation, we first introduce the following generalization of the convex hull of a set. Let $\{\mathcal{U}_i\}_{i=1}^B$ be a collection of compact subsets in a normed space. Then we denote by $\operatorname{conv}_B(\mathcal{U}_1, \dots, \mathcal{U}_B)$ the set of convex combinations with each $\mathcal{U}_i$ contributing exactly one element to each vector $\mathbf{x} \in \operatorname{conv}_B(\mathcal{U}_1, \dots, \mathcal{U}_B)$. More precisely, we set

$$\operatorname{conv}_B(\mathcal{U}_1, \dots, \mathcal{U}_B) := \left\{ \sum_{i=1}^B \alpha_i \mathbf{u}_i : \sum_{i=1}^B \alpha_i = 1, \alpha_i \geq 0, \mathbf{u}_i \in \mathcal{U}_i \; \forall i \in [B] \right\}$$

where we use the index B in the notation conv_B to emphasize the fact that each element of $\operatorname{conv}_B(\mathcal{U}_1, \dots, \mathcal{U}_B)$ consists of exactly B vectors drawn from a different set $\mathcal{U}_i$. If $\mathcal{U} \subset \mathbb{R}^D$ is a compact subset, then by the Carathéodory theorem, we recover the usual notion of the convex hull of $\mathcal{U}$ as

$$\operatorname{conv}(\mathcal{U}) = \operatorname{conv}_{D+1}(\underbrace{\mathcal{U}, \dots, \mathcal{U}}_{\substack{D+1 \\ \text{copies of } \mathcal{U}}}).$$

We point out that the result below also holds if we assume the sets $\mathcal{U}_i$ to be both compact and convex in which case we may replace $\operatorname{conv}_B(\mathcal{U}_1, \dots, \mathcal{U}_B)$ by $\operatorname{conv}(\bigcup_{i=1}^B \mathcal{U}_i)$.

Proposition 5.7 (Maurey's extended lemma)**.** *Let $(X, \|\cdot\|_X)$ be a normed space, and let $\mathcal{U}_1, \dots, \mathcal{U}_B \subset X$ be compact sets. Assume that for every $K \in \mathbb{N}$ and $\mathbf{z}_i \in \bigcup_{j=1}^B \mathcal{U}_j$ with $i = 1, \dots, K$ the following holds:*

$$\mathbb{E} \left\| \sum_{i=1}^K \epsilon_i \mathbf{z}_i \right\|_X \leq A\sqrt{K}$$

where $(\epsilon_i)_{i=1}^K$ is an independent Rademacher sequence, and $A > 0$ is a constant. Then for every $u > 0$,

$$\log \mathfrak{N}(\operatorname{conv}_B(\mathcal{U}_1, \dots, \mathcal{U}_B), \|\cdot\|_X, u) \lesssim (A/u)^2 \log \left(\sum_{i=1}^B \mathfrak{N}(\mathcal{U}_i, \|\cdot\|_X, u/2) \right).$$

Proof. We first equip each set $\mathcal{U}_i$ with its own net $\mathcal{N}_i$ with covering radius $u/2$ w. r. t. the canonical metric induced by $\|\cdot\|_X$. Next, denote by

$$\pi_i \colon X \to \mathcal{N}_i \colon \mathbf{x} \mapsto \operatorname*{argmin}_{\mathbf{z} \in \mathcal{N}_i} \|\mathbf{x} - \mathbf{z}\|_X$$

the projection on $\mathcal{N}_i$ in terms of $\|\cdot\|_X$, and set for $\mathbf{x} \in X$,

$$\pi(\mathbf{x}) := \operatorname*{argmin}_{\mathbf{x}_0 \in \{\pi_i(\mathbf{x}) : i \in [B]\}} \|\mathbf{x} - \mathbf{x}_0\|_X \in \bigcup_{i=1}^{B} \mathcal{N}_i.$$

Consider now a vector $\mathbf{x} \in \operatorname{conv}_B(\mathcal{U}_1, \dots, \mathcal{U}_B)$ such that

$$\mathbf{x} = \alpha_1 \mathbf{u}_1 + \ldots + \alpha_B \mathbf{u}_B$$

with $\mathbf{u}_i \in \mathcal{U}_i$ and $\alpha_i \in [0,1]$ for $i \in [B]$ with $\sum_{i=1}^{B} \alpha_i = 1$. Since the convex multipliers $\{\alpha_i\}_i$ define a discrete probability distribution on $[B]$, this allows us to construct a random vector $\mathbf{z} \in X$ with

$$\mathbb{P}(\mathbf{z} = \mathbf{u}_i) = \alpha_i,$$

such that $\mathbb{E}\mathbf{z} = \mathbf{x}$. Consider now K independent copies $\mathbf{z}_1, \dots, \mathbf{z}_K$ of $\mathbf{z}$. Then we have by the triangle inequality that

$$\begin{aligned} \mathbb{E}\left\|\mathbf{x} - \frac{1}{K}\sum_{i=1}^{K} \pi(\mathbf{z}_i)\right\|_X &\leq \mathbb{E}\left\|\mathbf{x} - \frac{1}{K}\sum_{i=1}^{K} \mathbf{z}_i\right\|_X + \mathbb{E}\left\|\frac{1}{K}\sum_{i=1}^{K} (\mathbf{z}_i - \pi(\mathbf{z}_i))\right\|_X \\ &\leq \mathbb{E}\left\|\mathbf{x} - \frac{1}{K}\sum_{i=1}^{K} \mathbf{z}_i\right\|_X + \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}\|\mathbf{z}_i - \pi(\mathbf{z}_i)\|_X. \end{aligned} \tag{5.21}$$

For the summands of the second term we find

$$\mathbb{E}\|\mathbf{z}_i - \pi(\mathbf{z}_i)\|_X = \sum_{j=1}^{B} \alpha_j \|\mathbf{u}_j - \pi(\mathbf{u}_j)\|_X \leq \sum_{j=1}^{B} \alpha_j u/2 = u/2$$

since π maps every vector $\mathbf{u}_j \in \mathcal{U}_j$ to its respective $(u/2)$-net $\mathcal{N}_j$. Next, we focus on the first term in (5.21) for which we find

$$\mathbb{E}\left\|\mathbf{x} - \frac{1}{K}\sum_{i=1}^{K} \mathbf{z}_i\right\|_X = \frac{1}{K}\mathbb{E}\left\|\sum_{i=1}^{K} (\mathbf{z}_i - \mathbb{E}\mathbf{z}_i)\right\|_X$$

since $\mathbb{E}\mathbf{z}_i = \mathbf{x}$ for all $\mathbf{z}_i$. Fixing randomness by conditioning on $\{\mathbf{z}_i\}_i \subset \bigcup_{j=1}^{B} \mathcal{U}_j$ and invoking the Giné-Zinn symmetrization principle [**GZ84**] then yields

$$\mathbb{E}\left\|\mathbf{x} - \frac{1}{K}\sum_{i=1}^{K} \mathbf{z}_i\right\|_X \leq \frac{2}{K}\mathbb{E}\left\|\sum_{i=1}^{K} \epsilon_i \mathbf{z}_i\right\|_X \leq \frac{2}{K} A\sqrt{K} = \frac{2A}{\sqrt{K}}$$

where $(\epsilon_i)_i$ is an independent Rademacher sequence, and the last step follows by the assumption of Proposition 5.7. We therefore find by collecting our estimates that

$$\mathbb{E}\left\|\mathbf{x} - \frac{1}{K}\sum_{i=1}^{K} \mathbf{z}_i\right\|_X = \frac{2A}{\sqrt{K}} + \frac{u}{2},$$

which implies for

$$K \geq 16\frac{A^2}{u^2}$$

that there exists at least one realization of the random vector

$$\hat{\mathbf{z}} := \frac{1}{K}\sum_{i=1}^{K}\pi(\mathbf{z}_i)$$

such that $\|\mathbf{x}-\hat{\mathbf{z}}\|_X \leq u$. To complete the proof, it remains to count the number of possible realizations of $\hat{\mathbf{z}}$. Choosing the nets $\mathcal{N}_i$ as the smallest $(u/2)$-nets, we have $|\mathcal{N}_i| = \mathfrak{N}(\mathcal{U}_i, \|\cdot\|_X, u/2)$. Since π maps any element of X on one of the B nets $\mathcal{N}_i$, there are exactly

$$\left(\sum_{i=1}^{B}\mathfrak{N}(\mathcal{U}_i, \|\cdot\|_X, u/2)\right)^K$$

realizations of $\hat{\mathbf{z}}$. Since the above argument holds for any $\mathbf{x} \in \operatorname{conv}_B(\mathcal{U}_1, \dots, \mathcal{U}_B)$, we conclude that

$$\log \mathfrak{N}(\operatorname{conv}_B(\mathcal{U}_1, \dots, \mathcal{U}_B), \|\cdot\|_X, u) \leq 16\frac{A^2}{u^2}\log\left(\sum_{i=1}^{B}\mathfrak{N}(\mathcal{U}_i, \|\cdot\|_X, u/2)\right)$$

as claimed. □

From our previous discussion, we have that every vector $\mathbf{x} \in \mathbb{B}^D_{\mathcal{I},1}$ can be decomposed for $S = \operatorname{supp}_{\mathcal{I}}(\mathbf{x}) = \{i \in [G] : \mathbf{x}_{\mathcal{I}_i} \neq \mathbf{0}\}$ as

$$\mathbf{x} = \sum_{i\in S}\|\mathbf{x}_{\mathcal{I}_i}\|_2 \frac{\mathbf{x}_{\mathcal{I}_i}}{\|\mathbf{x}_{\mathcal{I}_i}\|_2}$$

where each vector $\mathbf{u}_i := \mathbf{x}_{\mathcal{I}_i}/\|\mathbf{x}_{\mathcal{I}_i}\|_2$ is 1-group-sparse w. r. t. the group partition $\mathcal{I}$ with $\|\mathbf{u}_i\|_2 = 1$ and therefore $\mathbf{u}_i \in \mathbb{S}^{D-1}_{\mathcal{I}_i}$. Note, however, that the choice $\mathcal{U}_i = \mathbb{S}^{D-1}_{\mathcal{I}_i}$ in Proposition 5.7 does not work since for points $\mathbf{x} \in \operatorname{int}(\mathbb{B}^D_{\mathcal{I},1})$, we have

$$\sum_{i=1}^{G}\|\mathbf{x}_{\mathcal{I}_i}\|_2 = \sum_{i=1}^{G}\alpha_i = \|\mathbf{x}\|_{\mathcal{I},1} < 1$$

and hence $\mathbf{x} \notin \operatorname{conv}_G(\mathbb{S}^{D-1}_{\mathcal{I}_1}, \dots, \mathbb{S}^{D-1}_{\mathcal{I}_G})$ since the definition of $\operatorname{conv}_G(\mathbb{S}^{D-1}_{\mathcal{I}_1}, \dots, \mathbb{S}^{D-1}_{\mathcal{I}_G})$ assumes that its elements consist of convex combinations of exactly G elements with $\sum_{i=1}^{G}\alpha_i = 1$. Instead, we may either choose $\mathcal{U}_i = \mathbb{S}^{D-1}_{\mathcal{I}_i} \cup \{\mathbf{0}\}$ or $\mathcal{U}_i = \mathbb{B}^D_2 \cap \mathbb{C}^D_{\mathcal{I}_i} = \mathbb{B}^D_{\{\mathcal{I}_i\},1}$. We choose the latter option here since the volume comparison argument we will use below to bound the covering number of each $\mathcal{U}_i$ (w. r. t. $\|\cdot\|_2$) yields the same bound for both $\mathbb{S}^{D-1}_{\mathcal{I}_i} \cup \{\mathbf{0}\}$ and $\mathbb{B}^D_2 \cap \mathbb{C}^D_{\mathcal{I}_i}$ since $\mathbb{S}^{D-1}_{\mathcal{I}_i} \cup \{\mathbf{0}\} \subset \mathbb{B}^D_2 \cap \mathbb{C}^D_{\mathcal{I}_i}$.

Proposition 5.8. *The covering number of the group ℓ_1 unit ball w. r. t. the canonical metric induced by $\|\cdot\|_V$ is bounded according to*

$$\sqrt{\log \mathfrak{N}(\mathbb{B}^D_{\mathcal{I},1}, \|\cdot\|_V, u)} \lesssim u^{-1}\mu_{\mathcal{I}}(\mathbf{\Psi})\sqrt{\log(D)}\left(\sqrt{\log(G)} + \sqrt{2g\log(1+4/u)}\right).$$

Proof. As discussed above, we choose $\mathcal{U}_i = \mathbb{B}^D_2 \cap \mathbb{C}^D_{\mathcal{I}_i}$ in Proposition 5.7 and equip each unit ball in the coordinate subspace $\mathbb{C}^D_{\mathcal{I}_i}$ supported on $\mathcal{I}_i$ with a net $\mathcal{N}_i$ of covering radius $u/2$. Given a vector $\mathbf{x} \in \mathbb{B}^D_{\mathcal{I},1} = \operatorname{conv}_G(\mathcal{U}_1, \dots, \mathcal{U}_G)$, it merely remains to find an

appropriate bound on the expected norm of the Rademacher sum $\mathbb{E}\|\sum_{i=1}^{K}\epsilon_i\mathbf{z}_i\|_V$ for K vectors $\mathbf{z}_1,\dots,\mathbf{z}_K \in \bigcup_{i=1}^{G}\mathcal{U}_i$. To that end, first note that we have by the definition of $\|\cdot\|_V = \|V(\cdot)\|_{2\to 2}$ and linearity of the operator V (cf. Equation (5.11)) that

$$\mathbb{E}\left\|\sum_{i=1}^{K}\epsilon_i\mathbf{z}_i\right\|_V = \mathbb{E}\left\|\sum_{i=1}^{K}\epsilon_i V(\mathbf{z}_i)\right\|_{2\to 2}.$$

Next, we invoke the following noncommutative Khintchine inequality for operator norms due to Eftekhari *et al.*

Lemma 5.9 ([Eft+15, Lemma 9])**.** *Let $\{\mathbf{V}_i\}_{i=1}^{K}$ be a collection of matrices with the same dimension and rank at most r. Denote by $(\epsilon_i)_{i=1}^{K}$ an independent Rademacher sequence. Then*

$$\mathbb{E}\left\|\sum_{i=1}^{K}\epsilon_i\mathbf{V}_i\right\|_{2\to 2} \lesssim \sqrt{\log(r)}\left(\sum_{i=1}^{K}\|\mathbf{V}_i\|_{2\to 2}^2\right)^{1/2}.$$

Since the operator V yields for any $\mathbf{x}\in\mathbb{C}^D$ a matrix of size $mL\times mdL$, we have $\operatorname{rank} V(\mathbf{z}_i) \le mL = M$. An application of Lemma 5.9 therefore yields

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^{K}\epsilon_i\mathbf{z}_i\right\|_V &\lesssim \sqrt{\log(M)}\left(\sum_{i=1}^{K}\|V(\mathbf{z}_i)\|_{2\to 2}^2\right)^{1/2}\\ &\le \sqrt{\log(M)}\left(\sum_{i=1}^{K}\mu_{\mathcal{I}}(\mathbf{\Psi})^2\|\mathbf{z}_i\|_{\mathcal{I},1}^2\right)^{1/2}\\ &\le \mu_{\mathcal{I}}(\mathbf{\Psi})\sqrt{\log(M)}\sqrt{K}\end{aligned}$$

where the second step is due to (5.15), and the last step follows since each vector $\mathbf{z}_i\in\mathcal{U}_i$ is 1-group-sparse w. r. t. $\mathcal{I}$ by construction.

To complete the proof, we need to bound the covering numbers of the coordinate-restricted unit balls $\mathbb{B}_2^D\cap\mathbb{C}_{\mathcal{I}_i}^D$. Assuming that we have for each net $\mathcal{N}_i$ that $|\mathcal{N}_i| = \mathfrak{N}(\mathbb{B}_2^D\cap\mathbb{C}_{\mathcal{I}_i}^D, \|\cdot\|_V, u/2)$, we denote by $\nu := \max_{i\in[G]}|\mathcal{N}_i|$ the cardinality of the biggest net. To estimate $|\mathcal{N}_i|$, we return to the volume comparison argument (Lemma A.19) and find with (5.14) that

$$\begin{aligned}|\mathcal{N}_i| &= \mathfrak{N}(\mathbb{B}_2^D\cap\mathbb{C}_{\mathcal{I}_i}^D, \|\cdot\|_V, u/2)\\ &\le \mathfrak{N}(\mathbb{B}_2^{g_i}, \|\cdot\|_2, u/2)\\ &\le \left(1+\frac{2}{u/2}\right)^{2g_i}\end{aligned}$$

and therefore

$$\nu \le \left(1+\frac{4}{u}\right)^{2g}$$

with $g = \max_{i\in[G]} g_i$ as usual. The factor 2 in the exponent is again due to isomorphic identification of $\mathbb{C}^{g_i}$ with $\mathbb{R}^{2g_i}$. Combining this estimate with $A \lesssim \mu_{\mathcal{I}}(\mathbf{\Psi})\sqrt{\log(M)} \le$

$\mu_{\mathcal{I}}(\mathbf{\Psi})\sqrt{\log(D)}$, we finally find by invoking Proposition 5.7 that

$$\begin{aligned}\sqrt{\log \mathfrak{N}(\mathbb{B}_{\mathcal{I},1}^D, \|\cdot\|_V, u)} &= \sqrt{\log \mathfrak{N}(\mathrm{conv}_G(\mathbb{B}_2^D \cap \mathbb{C}_{\mathcal{I}_1}^D, \dots, \mathbb{B}_2^D \cap \mathbb{C}_{\mathcal{I}_G}^D), \|\cdot\|_V, u)} \\ &\lesssim \frac{A}{u}\sqrt{\log\left(\sum_{i=1}^G |\mathcal{N}_i|\right)} \\ &\leq \frac{\mu_{\mathcal{I}}(\mathbf{\Psi})\sqrt{\log(D)}}{u}\left(\sqrt{\log(G\nu)}\right) \\ &\leq \frac{\mu_{\mathcal{I}}(\mathbf{\Psi})\sqrt{\log(D)}}{u}\left(\sqrt{\log(G)} + \sqrt{2g\log\left(1+\frac{4}{u}\right)}\right). \end{aligned} \tag{5.22}$$

This completes the proof. □

To establish our final bound on the γ_2-functional of $\mathcal{M}$, we split the entropy integral in two parts according to (5.16). We then control the first part via the volume comparison estimate (5.17) and bound the second integral via (5.18), followed by an application of Proposition 5.8. For the first integral, this yields[6]

$$\begin{aligned}\int_0^\lambda \sqrt{\log \mathfrak{N}(\Omega, \|\cdot\|_V, \varepsilon)}\mathrm{d}\varepsilon &\leq \int_0^\lambda \sqrt{s\log(eG/s) + 2sg\log(1+4/\varepsilon)}\mathrm{d}\varepsilon \\ &\leq \lambda\sqrt{s\log(eG/s)} + \lambda\sqrt{2sg\log(5e/\lambda)}. \end{aligned} \tag{5.23}$$

For the second integral, we find

$$\begin{aligned}&\int_\lambda^{\sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})} \sqrt{\log \mathfrak{N}(\Omega, \|\cdot\|_V, \varepsilon)}\mathrm{d}\varepsilon \\ &\lesssim 2\sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})\sqrt{\log(D)}\left(\int_{\lambda/(2\sqrt{s})}^{\mu_{\mathcal{I}}(\mathbf{\Psi})/2} \varepsilon^{-1}\sqrt{\log(G)}\mathrm{d}\varepsilon + \int_{\lambda/(2\sqrt{s})}^{\mu_{\mathcal{I}}(\mathbf{\Psi})/2} \varepsilon^{-1}\sqrt{g\log(1+8/\varepsilon)}\mathrm{d}\varepsilon\right)\end{aligned}$$

For the last integral, note that $\sqrt{\log(1+t^{-1})}$ is monotonically decreasing in t. Hence, we have that

$$\int_a^b t^{-1}\sqrt{\log(1+t^{-1})}\mathrm{d}t \leq \log(b/a)\sqrt{\log(1+a^{-1})}.$$

This yields

$$\begin{aligned}&\int_\lambda^{\sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})} \sqrt{\log \mathfrak{N}(\Omega, \|\cdot\|_V, \varepsilon)}\mathrm{d}\varepsilon \\ &\lesssim \sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})\sqrt{\log(D)}\log(\sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})/\lambda)\left(\sqrt{\log(G)} + \sqrt{g\log(1+16\sqrt{s}/\lambda)}\right).\end{aligned} \tag{5.24}$$

Compared with our previous estimate based on Sudakov's inequality for which we found

$$\int_\lambda^{\sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})} \sqrt{\log \mathfrak{N}(\Omega, \|\cdot\|_V, \varepsilon)}\mathrm{d}\varepsilon \lesssim \sqrt{s}\log(\sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})/\lambda)\left(\sqrt{\log(G)} + \sqrt{g}\right),$$

[6] The last estimate follows from the bound $\int_0^\alpha \sqrt{\log(1+t^{-1})}\mathrm{d}t \leq \alpha\sqrt{\log(e(1+\alpha^{-1}))}$ for $\alpha > 0$ (see, *e.g.*, [**FR13**, Lemma C.9]).

our new bound differs by an additional log-factor in D, as well as another logarithmic factor depending on λ. However, we also obtain the desired linear dependence on $\mu_{\mathcal{I}}(\mathbf{\Psi})$. Simplifying (5.23) and (5.24) by absorbing numerical constants into the implicit constant in the notation and collecting both estimates, we eventually find

$$\begin{aligned}\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \lesssim\ & \lambda\sqrt{s\log(G/s)} + \lambda\sqrt{sg\log(1/\lambda)} \\ & + \sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})\sqrt{\log(D)}\log(s\mu_{\mathcal{I}}(\mathbf{\Psi})/\lambda)\Big(\sqrt{\log(G)} + \sqrt{g\log(s/\lambda)}\Big),\end{aligned}$$

which, for the choice $\lambda = \mu_{\mathcal{I}}(\mathbf{\Psi})$, ultimately results in

$$\begin{aligned}\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) &\lesssim \mu_{\mathcal{I}}(\mathbf{\Psi})\sqrt{s\log(G/s)} + \mu_{\mathcal{I}}(\mathbf{\Psi})\sqrt{sg\log(1/\mu_{\mathcal{I}}(\mathbf{\Psi}))} \\ &\quad + \sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})\sqrt{\log(D)}\log(s)\Big(\sqrt{\log(G)} + \sqrt{g\log(s/\mu_{\mathcal{I}}(\mathbf{\Psi}))}\Big) \\ &\lesssim \sqrt{s}\mu_{\mathcal{I}}(\mathbf{\Psi})\sqrt{\log(D)}\log(s)\Big(\sqrt{\log(G)} + \sqrt{g\log(s/\mu_{\mathcal{I}}(\mathbf{\Psi}))}\Big). \qquad (5.25)\end{aligned}$$

5.3.4 Stable and Robust Group-Sparse Recovery with General Block Diagonal Operators

At this point, we are prepared to derive our main result by invoking Theorem 5.5 and collecting our obtained estimates for $\rho_{\mathrm{F}}(\mathcal{M}), \rho_{2\to 2}(\mathcal{M})$ and $\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2})$.

Theorem 5.10. *Let $\mathbf{A} = \operatorname{diag}\{\mathbf{\Phi}_l\}_{l=1}^{L} \in \mathbb{R}^{mL\times dL}$ be a block diagonal random matrix with subgaussian blocks $\mathbf{\Phi}_l$ whose entries are independent subgaussian zero-mean, unit-variance random variables $(\xi_i)_i$ with subgaussian norm $\tau = \max_i \|\xi_i\|_{\psi_2}$. Let further $\mathbf{\Psi} \in \mathrm{U}(dL)$ be a unitary matrix, and assume that*

$$m \gtrsim_\tau \delta^{-2} s\mu_{\mathcal{I}}(\mathbf{\Psi})^2\Big[\log(D)\log(s)^2[\log(G) + g\log(s/\mu_{\mathcal{I}}(\mathbf{\Psi}))] + \log(\eta^{-1})\Big]$$

with

$$\mu_{\mathcal{I}}(\mathbf{\Psi}) := \min\left\{\sqrt{d}\max_{i\in[D]}\|\boldsymbol{\psi}_i\|_{\mathcal{I},\infty}, 1\right\}$$

and $\boldsymbol{\psi}_i \in \mathbb{C}^D$ denoting the i-th row of $\mathbf{\Psi}$. Set $\tilde{\mathbf{A}} := m^{-1/2}\mathbf{A}\mathbf{\Psi}$. Then with probability at least $1-\eta$, every vector $\mathring{\mathbf{x}} \in \mathbb{C}^D$ acquired as $\mathbf{y} = \tilde{\mathbf{A}}\mathring{\mathbf{x}} + \mathbf{e}$ with $\|\mathbf{e}\|_2 \le \nu$ is approximated by a minimizer $\mathbf{x}^\star$ of

$$\begin{array}{ll} \underset{\mathbf{x}}{\textit{minimize}} & \|\mathbf{x}\|_{\mathcal{I},1} \\ \textit{s.t.} & \left\|\tilde{\mathbf{A}}\mathbf{x} - \mathbf{y}\right\|_2 \le \nu \end{array} \qquad (\mathrm{P}_{\mathcal{I},1})$$

with

$$\|\mathring{\mathbf{x}} - \mathbf{x}^\star\|_2 \le C_0\frac{\sigma_s(\mathring{\mathbf{x}})_{\mathcal{I},1}}{\sqrt{s}} + C_1\nu,$$

where $C_0, C_1 > 0$ are constants which only depend on δ.

Proof. Invoking Theorem 5.5 in combination with Equation (5.12), we have that for $u \geq 1$,

$$\begin{aligned}\mathbb{P}\left(\sup_{\boldsymbol{\Gamma}\in\mathcal{M}}\left|\|\boldsymbol{\Gamma}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|\boldsymbol{\Gamma}\boldsymbol{\xi}\|_2^2\right| \geq c_\tau E_u\right) &= \mathbb{P}\left(\frac{1}{m}\sup_{\boldsymbol{\Gamma}\in\mathcal{M}}\left|\|\boldsymbol{\Gamma}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|\boldsymbol{\Gamma}\boldsymbol{\xi}\|_2^2\right| \geq \frac{c_\tau E_u}{m}\right)\\ &= \mathbb{P}\left(\delta_s\left(\frac{1}{\sqrt{m}}\mathbf{A}\boldsymbol{\Psi}\right) \geq \frac{c_\tau E_u}{m}\right)\\ &\leq e^{-u}\end{aligned}$$

and hence

$$\mathbb{P}(\delta_s \geq \delta) \leq \mathbb{P}\left(\delta_s \geq \frac{c_\tau E_u}{m}\right) \leq e^{-u}$$

if $\delta \geq c_\tau E_u/m$. To that end, we bound the term E_u with $\rho_{\mathrm{F}} = \rho_{\mathrm{F}}(\mathcal{M}) = \sqrt{m}, \rho_{2\to2} = \rho_{2\to2}(\mathcal{M}) \leq \sqrt{s}\mu_{\mathcal{I}}(\boldsymbol{\Psi})$ and $\gamma_2 = \gamma_2(\mathcal{M}, \|\cdot\|_{2\to2})$:

$$\begin{aligned}c_\tau\frac{E_u}{m} &= c_\tau\frac{\gamma_2^2 + \rho_{\mathrm{F}}\gamma_2 + \sqrt{u}\rho_{\mathrm{F}}\rho_{2\to2} + u\rho_{2\to2}^2}{m}\\ &\leq c_\tau\frac{\gamma_2^2 + \sqrt{m}\gamma_2 + \sqrt{u}\sqrt{m}\sqrt{s}\mu_{\mathcal{I}}(\boldsymbol{\Psi}) + us\mu_{\mathcal{I}}(\boldsymbol{\Psi})^2}{m}\\ &= c_\tau\left(\frac{\gamma_2}{\sqrt{m}} + \left(\frac{\gamma_2}{\sqrt{m}}\right)^2 + \frac{\sqrt{u}\sqrt{s}\mu_{\mathcal{I}}(\boldsymbol{\Psi})}{\sqrt{m}} + \left(\frac{\sqrt{u}\sqrt{s}\mu_{\mathcal{I}}(\boldsymbol{\Psi})}{\sqrt{m}}\right)^2\right).\end{aligned}$$

Now choose

$$m \geq \tilde{\delta}^{-2}\max\left\{\gamma_2^2, us\mu_{\mathcal{I}}(\boldsymbol{\Psi})^2\right\} \geq \frac{1}{2}\tilde{\delta}^{-2}\left[\gamma_2^2 + us\mu_{\mathcal{I}}(\boldsymbol{\Psi})^2\right]$$

such that $\gamma_2/\sqrt{m} \leq \tilde{\delta}$ and $\sqrt{u}\sqrt{s}\mu_{\mathcal{I}}(\boldsymbol{\Psi})/\sqrt{m} \leq \tilde{\delta}$ with $\tilde{\delta} \leq 1$. Then

$$\begin{aligned}c_\tau\frac{E_u}{m} &= c_\tau\left(\frac{\gamma_2}{\sqrt{m}} + \left(\frac{\gamma_2}{\sqrt{m}}\right)^2 + \frac{\sqrt{u}\sqrt{s}\mu_{\mathcal{I}}(\boldsymbol{\Psi})}{\sqrt{m}} + \left(\frac{\sqrt{u}\sqrt{s}\mu_{\mathcal{I}}(\boldsymbol{\Psi})}{\sqrt{m}}\right)^2\right)\\ &\leq 2c_\tau\left(\frac{\gamma_2}{\sqrt{m}} + \frac{\sqrt{u}\sqrt{s}\mu_{\mathcal{I}}(\boldsymbol{\Psi})}{\sqrt{m}}\right)\\ &\leq \tilde{c}_\tau\tilde{\delta}\\ &=: \delta\end{aligned}$$

with $\tilde{c}_\tau := 4c_\tau$. For $u \geq 1$, we conclude that with probability at least $1 - e^{-u} =: 1 - \eta$, the matrix $\tilde{\mathbf{A}}$ satisfies the group-RIP of order $2s$ with constant δ, provided that

$$\begin{aligned}m &\geq C_\tau\delta^{-2}s\mu_{\mathcal{I}}(\boldsymbol{\Psi})^2\left[\log(D)\log(s)^2[\log(G) + g\log(s/\mu_{\mathcal{I}}(\boldsymbol{\Psi}))] + \log(\eta^{-1})\right]\\ &\geq \frac{1}{2}C_\tau\delta^{-2}\left(\left[\sqrt{s}\mu_{\mathcal{I}}(\boldsymbol{\Psi})\sqrt{\log(D)}\log(s)\left(\sqrt{\log(G)} + \sqrt{g\log(s/\mu_{\mathcal{I}}(\boldsymbol{\Psi}))}\right)\right]^2 + us\mu_{\mathcal{I}}(\boldsymbol{\Psi})^2\right)\\ &\geq \frac{1}{2}\tilde{c}_\tau^2\delta^{-2}\left[\gamma_2(\mathcal{M}, \|\cdot\|_{2\to2})^2 + us\mu_{\mathcal{I}}(\boldsymbol{\Psi})^2\right]\end{aligned}$$

with $C_\tau := C\tilde{c}_\tau^2 = 16Cc_\tau$ and $C > 0$ denoting the implicit constant in (5.25). The claim then follows by Theorem 5.2. □

Before discussing the attained bound, we first turn to the second acquisition model considered in this chapter in which we assume that each sensor is equipped with an identical copy of the *same* random matrix drawn once.

5.4 The Group-RIP for Block Diagonal Matrices with Repeated Blocks

5.4.1 Formulation as a Chaos Process

Intuitively, sharing the same measurement matrix between individual sensors should not to lead to wildly different conditions for the resulting block diagonal measurement matrix to satisfy the group restricted isometry property. As we will establish in this section, this intuition is confirmed up to substitution of the coherence parameter $\mu_{\mathcal{I}}$ of the sparsity basis by a more involved complexity parameter, which we derive below. To make matters precise, we now consider measurements of the form

$$\mathbf{y} = \mathbf{A}\boldsymbol{\Psi}\mathbf{x} = \begin{pmatrix} \boldsymbol{\Phi} & & \\ & \ddots & \\ & & \boldsymbol{\Phi} \end{pmatrix} \boldsymbol{\Psi}\mathbf{x} = \begin{pmatrix} \boldsymbol{\Phi}\boldsymbol{\Psi}_1\mathbf{x} \\ \vdots \\ \boldsymbol{\Phi}\boldsymbol{\Psi}_L\mathbf{x} \end{pmatrix}$$

where $\boldsymbol{\Psi}_l \in \mathbb{C}^{d\times dL}$ are the partial basis expansion matrices as before. Once again our goal is to reformulate the group-RIP constant of a measurement matrix of this form to be amenable to the probabilistic analysis of Theorem 5.5. While we could use the same transformations V_l as in the case of unique per-sensor matrices (cf. Equation (5.10)) and set

$$\mathbf{A}\boldsymbol{\Psi}\mathbf{x} \overset{\mathrm{d}}{=} \begin{pmatrix} V_1(\mathbf{x}) \\ \vdots \\ V_L(\mathbf{x}) \end{pmatrix} \boldsymbol{\xi} =: V'(\mathbf{x})\boldsymbol{\xi}$$

with $\boldsymbol{\xi} \in \mathbb{R}^{md}$ now denoting a unit-variance τ-subgaussian random vector of size md instead of mdL, the lack of a (block) diagonal structure in the elements of the image of V' complicates the calculation of both $\rho_{2\to 2}$ and γ_2 as we cannot concisely express the operator norm in terms of a mixed (ℓ_∞, ℓ_2) vector norm as in Equation (5.13). However, since we only require $\|\mathbf{A}\boldsymbol{\Psi}\mathbf{x}\|_2^2$ and $\|V'(\mathbf{x})\boldsymbol{\xi}\|_2^2$ to be identical in distribution to apply Theorem 5.5, we are free to reorder the rows of $V'(\mathbf{x})$ as we choose. To that end, we define the alternative operator $\widetilde{V}\colon \mathbb{C}^D \to \mathbb{C}^{L\times d}$ with

$$\mathbf{x} \mapsto \widetilde{V}(\mathbf{x}) := \begin{pmatrix} (\boldsymbol{\Psi}_1\mathbf{x})^\top \\ \vdots \\ (\boldsymbol{\Psi}_L\mathbf{x})^\top \end{pmatrix} \in \mathbb{C}^{L\times d}.$$

With this, we define the map

$$\widehat{V}(\mathbf{x}) := \begin{pmatrix} \widetilde{V}(\mathbf{x}) & & \\ & \ddots & \\ & & \widetilde{V}(\mathbf{x}) \end{pmatrix} \in \mathbb{C}^{mL\times md},$$

which maps a vector $\mathbf{x} \in \mathbb{C}^D$ to a block diagonal matrix with m copies of $\widetilde{V}(\mathbf{x})$ on its diagonal such that $\|\mathbf{A}\boldsymbol{\Psi}\mathbf{x}\|_2^2 \overset{\mathrm{d}}{=} \|\widehat{V}(\mathbf{x})\boldsymbol{\xi}\|_2^2$ as desired. Similar to Section 5.3, we also define

the set $\widehat{\mathcal{M}} := \widehat{V}(\Omega)$ as the image of the set $\Omega = \Sigma_{\mathcal{I},s} \cap \mathbb{S}^{D-1}$ of s-group-sparse vectors on the unit sphere so that

$$\mathbb{P}\left(\sup_{\mathbf{x}\in\Omega}\left|\left\|\frac{1}{\sqrt{m}}\mathbf{A}\boldsymbol{\Psi}\mathbf{x}\right\|_2^2 - 1\right| \geq \delta\right) = \left(\frac{1}{m}\sup_{\boldsymbol{\Gamma}\in\widehat{\mathcal{M}}}\left|\|\boldsymbol{\Gamma}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|\boldsymbol{\Gamma}\boldsymbol{\xi}\|_2^2\right| \geq \delta\right).$$

It remains to estimate the radii of $\widehat{\mathcal{M}}$, as well as its metric entropy integral. Unsurprisingly, we mostly proceed in the same way as before. For convenience of notation, we associate with $\widehat{V}$ the norm $\|\cdot\|_{\widehat{V}}$ on $\mathbb{C}^D$ induced by $\|\cdot\|_{\widehat{V}} := \|\widehat{V}(\cdot)\|_{2\to 2}$.

5.4.2 Estimation of Radii and the Metric Entropy

We begin by calculating the radius of $\widehat{\mathcal{M}}$ w. r. t. the Frobenius norm. First, note that

$$\begin{aligned}\left\|\widehat{V}(\mathbf{x})\right\|_{\mathrm{F}}^2 &= \sum_{i=1}^{m}\left\|\widetilde{V}(\mathbf{x})\right\|_{\mathrm{F}}^2 = m\,\mathrm{tr}(\widetilde{V}(\mathbf{x})\widetilde{V}(\mathbf{x})^*)\\ &= m\sum_{l=1}^{L}\|\boldsymbol{\Psi}_l\mathbf{x}\|_2^2 = m\|\mathbf{x}\|_2^2\end{aligned}$$

and therefore

$$\rho_{\mathrm{F}}(\widehat{\mathcal{M}}) = \sup_{\boldsymbol{\Gamma}\in\widehat{\mathcal{M}}}\|\boldsymbol{\Gamma}\|_{\mathrm{F}} = \sup_{\mathbf{x}\in\Omega}\left\|\widehat{V}(\mathbf{x})\right\|_{\mathrm{F}} = \sup_{\mathbf{x}\in\Omega}\sqrt{m}\|\mathbf{x}\|_2 = \sqrt{m}.$$

Next, denote as before by $S \subset [G]$ the index set of nonzero groups of $\mathbf{x} \in \mathbb{C}^D$ w. r. t. $\mathcal{I}$. Then we have due to linearity of $\widetilde{V}$ and consequently linearity of $\widehat{V}$ that

$$\begin{aligned}\left\|\widehat{V}(\mathbf{x})\right\|_{2\to 2} = \left\|\widetilde{V}(\mathbf{x})\right\|_{2\to 2} &= \left\|\sum_{i=1}^{G}\widetilde{V}(\mathbf{x}_{\mathcal{I}_i})\right\|_{2\to 2}\\ &\leq \sum_{i\in S}\|\mathbf{x}_{\mathcal{I}_i}\|_2 \cdot \left\|\widetilde{V}\left(\frac{\mathbf{x}_{\mathcal{I}_i}}{\|\mathbf{x}_{\mathcal{I}_i}\|_2}\right)\right\|_{2\to 2}\\ &\leq \|\mathbf{x}\|_{\mathcal{I},1}\max_{i\in S}\left\|\widetilde{V}\left(\frac{\mathbf{x}_{\mathcal{I}_i}}{\|\mathbf{x}_{\mathcal{I}_i}\|_2}\right)\right\|_{2\to 2}\\ &\leq \|\mathbf{x}\|_{\mathcal{I},1}\max_{i\in[G]}\sup_{\mathbf{u}\in\mathbb{S}^{D-1}_{\mathcal{I}_i}}\left\|\widetilde{V}(\mathbf{u})\right\|_{2\to 2}.\end{aligned} \tag{5.26}$$

In the edge case where the number of groups G coincides with the ambient dimension D (*i.e.*, in case of regular sparsity rather than group-sparsity), the supremum in (5.26) can be easily computed as each coordinate-restricted unit sphere $\mathbb{S}^{D-1}_{\mathcal{I}_i}$ reduces w. l. o. g. to a two-element[7] set $\{\pm\mathbf{e}_i\}$ where $\mathbf{e}_i \in \mathbb{R}^D$ denotes the i-th canonical unit vector. In other words, if the vectors $\mathbf{u}$ in the supremum in the last line only take finitely many values, estimating the bound amounts to taking the maximum of the operator norms of the matrices produced by $\widetilde{V}$. However, the same does not hold for $G < D$, which does not allow one to evaluate (5.26) numerically. To circumvent this computability issue, we estimate the supremum as follows.

[7] In light of the linearity of $\widetilde{V}$, this in turn implies the supremum in (5.26) is taken over a singleton set.

Denote by $\mathbf{u}$ an arbitrary unit-normalized 1-group-sparse vector w. r. t. the group partition $\mathcal{I}$. Then

$$\begin{aligned}\left\|\widetilde{V}(\mathbf{u})\right\|_{2\to 2} &= \sup_{\mathbf{z}\in\mathbb{B}_2^d}\left\|\widetilde{V}(\mathbf{u})\mathbf{z}\right\|_2 \leq \sqrt{L}\sup_{\mathbf{z}\in\mathbb{B}_2^d}\left\|\widetilde{V}(\mathbf{u})\mathbf{z}\right\|_\infty \\ &= \sqrt{L}\sup_{\mathbf{z}\in\mathbb{B}_2^d}\max_{l\in[L]}|\langle\boldsymbol{\Psi}_l\mathbf{u},\mathbf{z}\rangle| = \sqrt{L}\sup_{\mathbf{z}\in\mathbb{B}_2^d}\max_{l\in[L]}\left|\left\langle\mathbf{u},\boldsymbol{\Psi}_l^\top\mathbf{z}\right\rangle\right| \\ &\leq \sqrt{L}\|\mathbf{u}\|_{\mathcal{I},1}\sup_{\mathbf{z}\in\mathbb{B}_2^d}\max_{l\in[L]}\left\|\boldsymbol{\Psi}_l^\top\mathbf{z}\right\|_{\mathcal{I},\infty} = \sqrt{L}\max_{l\in[L]}\sup_{\mathbf{z}\in\mathbb{B}_2^d}\left\|\boldsymbol{\Psi}_l^\top\mathbf{z}\right\|_{\mathcal{I},\infty}\end{aligned}$$

where in the last inequality we invoked Lemma 5.6 and used the fact that $\|\mathbf{u}\|_{\mathcal{I},1} = 1$ since $\mathbf{u}$ is a unit-norm vector supported on a single group of $\mathcal{I}$. Expanding the supremum, we find

$$\begin{aligned}\sup_{\mathbf{z}\in\mathbb{B}_2^d}\left\|\boldsymbol{\Psi}_l^\top\mathbf{z}\right\|_{\mathcal{I},\infty} &= \sup_{\mathbf{z}\in\mathbb{B}_2^d}\max_{i\in[G]}\left\|(\boldsymbol{\Psi}_l^\top\mathbf{z})_{\mathcal{I}_i}\right\|_2 = \max_{i\in[G]}\sup_{\mathbf{z}\in\mathbb{B}_2^{|\mathcal{I}_i|}}\left\|((\boldsymbol{\Psi}_l)_{\mathcal{I}_i})^\top\mathbf{z}\right\|_2 \\ &= \max_{i\in[G]}\left\|((\boldsymbol{\Psi}_l)_{\mathcal{I}_i})^\top\right\|_{2\to 2} = \max_{i\in[G]}\|(\boldsymbol{\Psi}_l)_{\mathcal{I}_i}\|_{2\to 2},\end{aligned}$$

where $(\boldsymbol{\Psi}_l)_{\mathcal{I}_i}\in\mathbb{C}^{d\times|\mathcal{I}_i|}$ denotes the submatrix of $\boldsymbol{\Psi}_l$ restricted to the columns indexed by $\mathcal{I}_i$, and we used the fact that

$$\begin{aligned}\|\mathbf{A}\|_{p\to q} &:= \sup_{\|\mathbf{x}\|_p\leq 1}\|\mathbf{A}\mathbf{x}\|_q = \sup_{\|\mathbf{x}\|_p\leq 1}\sup_{\|\mathbf{z}\|_{q'}\leq 1}|\langle\mathbf{A}\mathbf{x},\mathbf{z}\rangle_{\mathbb{C}}| = \sup_{\|\mathbf{z}\|_{q'}\leq 1}\sup_{\|\mathbf{x}\|_p\leq 1}|\langle\mathbf{x},\mathbf{A}^*\mathbf{z}\rangle_{\mathbb{C}}| \\ &= \sup_{\|\mathbf{z}\|_{q'}\leq 1}\sup_{\|\mathbf{x}\|_p\leq 1}\left|\overline{\langle\mathbf{x},\mathbf{A}^*\mathbf{z}\rangle_{\mathbb{C}}}\right| = \sup_{\|\mathbf{z}\|_{q'}\leq 1}\sup_{\|\mathbf{x}\|_p\leq 1}\left|\left\langle\overline{\mathbf{x}},\overline{\mathbf{A}^*}\overline{\mathbf{z}}\right\rangle_{\mathbb{C}}\right| \\ &= \sup_{\|\mathbf{z}\|_{q'}\leq 1}\sup_{\|\mathbf{x}\|_p\leq 1}\left|\left\langle\mathbf{x},\mathbf{A}^\top\mathbf{z}\right\rangle_{\mathbb{C}}\right| = \sup_{\|\mathbf{z}\|_{q'}\leq 1}\left\|\mathbf{A}^\top\mathbf{z}\right\|_{p'} = \left\|\mathbf{A}^\top\right\|_{q'\to p'}\end{aligned}$$

for $p,p',q,q'\geq 1$ with $1/p+1/p'=1$ and $1/q+1/q'=1$ (see also Example A.14). This chain of estimates therefore yields

$$\left\|\widehat{V}(\mathbf{x})\right\|_{2\to 2} \leq \|\mathbf{x}\|_{\mathcal{I},1}\sqrt{L}\max_{\substack{i\in[G],\\ l\in[L]}}\|(\boldsymbol{\Psi}_l)_{\mathcal{I}_i}\|_{2\to 2}. \tag{5.27}$$

Unfortunately, this bound is too loose in the previously discussed edge case where $G=D$ with vectors $\mathbf{x}\in\mathbb{C}^D$ being s-sparse as it does not reduce to

$$\left\|\widehat{V}(\mathbf{x})\right\|_{2\to 2} \leq \|\mathbf{x}\|_1\max_{i\in[D]}\left\|\widetilde{V}(\mathbf{e}_i)\right\|_{2\to 2},$$

which we would obtain from (5.26). In other words, the estimate does not reduce to the natural bound in the sparse setting. To remedy the situation, we also consider the following simpler bound. Note that for $i\in S=\operatorname{supp}_{\mathcal{I}}(\mathbf{x})=\{i\in[G]:\mathbf{x}_{\mathcal{I}_i}\neq\mathbf{0}\}$, we have

$$\begin{aligned}\left\|\widetilde{V}\left(\frac{\mathbf{x}_{\mathcal{I}_i}}{\|\mathbf{x}_{\mathcal{I}_i}\|_2}\right)\right\|_{2\to 2} &\leq \sup_{\mathbf{u}\in\mathbb{S}_{\mathcal{I}_i}^{D-1}}\left\|\widetilde{V}(\mathbf{u})\right\|_{2\to 2} = \sup_{\mathbf{u}\in\mathbb{S}_{\mathcal{I}_i}^{D-1}}\left\|\sum_{j\in\mathcal{I}_i}|u_j|\widetilde{V}(\mathbf{e}_j)\right\|_{2\to 2} \\ &\leq \sup_{\mathbf{u}\in\mathbb{S}_{\mathcal{I}_i}^{D-1}}\sum_{j\in\mathcal{I}_i}|u_j|\cdot\left\|\widetilde{V}(\mathbf{e}_j)\right\|_{2\to 2} \leq \sup_{\mathbf{u}\in\mathbb{S}_{\mathcal{I}_i}^{D-1}}\|\mathbf{u}\|_1\max_{j\in\mathcal{I}_i}\left\|\widetilde{V}(\mathbf{e}_j)\right\|_{2\to 2} \\ &\leq \sqrt{|\mathcal{I}_i|}\max_{j\in\mathcal{I}_i}\left\|\widetilde{V}(\mathbf{e}_j)\right\|_{2\to 2} \leq \sqrt{g}\max_{j\in\mathcal{I}_i}\left\|\widetilde{V}(\mathbf{e}_j)\right\|_{2\to 2}\end{aligned}$$

where the second to last estimate follows by Cauchy-Schwarz. Substituting this estimate into (5.26) and defining the parameter

$$\begin{aligned}\omega_{\mathcal{I}}(\mathbf{\Psi}) &:= \min\left\{\sqrt{g}\max_{i\in[D]}\left\|\widetilde{V}(\mathbf{e}_i)\right\|_{2\to 2}, \sqrt{L}\max_{\substack{l\in[L],\\ i\in[G]}}\|(\mathbf{\Psi}_l)_{\mathcal{I}_i}\|_{2\to 2}\right\}\\ &= \min\left\{\sqrt{g}\max_{i\in[D]}\left\|\begin{pmatrix}(\mathbf{\Psi}_1)_{\{i\}} & \dots & (\mathbf{\Psi}_L)_{\{i\}}\end{pmatrix}\right\|_{2\to 2}, \sqrt{L}\max_{\substack{l\in[L],\\ i\in[G]}}\|(\mathbf{\Psi}_l)_{\mathcal{I}_i}\|_{2\to 2}\right\},\end{aligned}$$

we find

$$\left\|\widehat{V}(\mathbf{x})\right\|_{2\to 2} \leq \omega_{\mathcal{I}}(\mathbf{\Psi})\|\mathbf{x}\|_{\mathcal{I},1},$$

which yields

$$\rho_{2\to 2}(\widehat{\mathcal{M}}) = \sup_{\mathbf{\Gamma}\in\widehat{\mathcal{M}}}\|\mathbf{\Gamma}\|_{2\to 2} = \sup_{\mathbf{x}\in\Omega}\left\|\widehat{V}(\mathbf{x})\right\|_{2\to 2} \leq \omega_{\mathcal{I}}(\mathbf{\Psi})\sup_{\mathbf{x}\in\Omega}\|\mathbf{x}\|_{\mathcal{I},1} \leq \sqrt{s}\omega_{\mathcal{I}}(\mathbf{\Psi}).$$

Lastly, we need to estimate the γ_2-functional of the set $\widehat{\mathcal{M}}$ w. r. t. the operator norm. To that end, we point out that

$$\begin{aligned}\left\|\widehat{V}(\mathbf{x})\right\|_{2\to 2}^2 = \left\|\widetilde{V}(\mathbf{x})\right\|_{2\to 2}^2 = \left\|\widetilde{V}(\mathbf{x})^*\right\|_{2\to 2}^2 &\leq \left\|\widetilde{V}(\mathbf{x})^*\right\|_{\mathrm{F}}^2\\ = \left\|\begin{pmatrix}\mathbf{\Psi}_1\mathbf{x} & \dots & \mathbf{\Psi}_L\mathbf{x}\end{pmatrix}\right\|_{\mathrm{F}}^2 &= \sum_{l=1}^{L}\|\mathbf{\Psi}_l\mathbf{x}\|_2^2\\ = \|\mathbf{\Psi}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2 &\leq \|\mathbf{x}\|_{\mathcal{I},1}^2.\end{aligned}$$

A careful review of the arguments presented in Section 5.3.3 then reveals that the difference in derivation amounts to replacing the induced norm $\|\cdot\|_V$ with $\|\cdot\|_{\widehat{V}}$. In particular, since we established that $\|\cdot\|_{\widehat{V}} \leq \omega_{\mathcal{I}}(\mathbf{\Psi})\|\cdot\|_{\mathcal{I},1}$ and $\|\cdot\|_{\widehat{V}} \leq \|\cdot\|_2$, estimating the γ_2-functional of $\widehat{\mathcal{M}}$ by means of the metric entropy integral

$$\begin{aligned}\gamma_2(\widehat{\mathcal{M}}, \|\cdot\|_{2\to 2}) &\lesssim \int_0^{\rho_{2\to 2}(\widehat{\mathcal{M}})}\sqrt{\log \mathfrak{N}(\widehat{\mathcal{M}}, \|\cdot\|_{2\to 2}, \varepsilon)}\mathrm{d}\varepsilon\\ &= \int_0^{\sqrt{s}\omega_{\mathcal{I}}(\mathbf{\Psi})}\sqrt{\log \mathfrak{N}(\Omega, \|\cdot\|_{\widehat{V}}, \varepsilon)}\mathrm{d}\varepsilon\end{aligned}$$

proceeds identically to the derivation in Section 5.3.3. We therefore immediately conclude

$$\gamma_2(\widehat{\mathcal{M}}, \|\cdot\|_{2\to 2}) \lesssim \sqrt{s}\omega_{\mathcal{I}}(\mathbf{\Psi})\sqrt{\log(D)}\log(s)\left(\sqrt{\log(G)} + \sqrt{g\log(s/\omega_{\mathcal{I}}(\mathbf{\Psi}))}\right).$$

As in the case of Theorem 5.10, Theorem 5.11 below now follows by invoking Theorem 5.5 with the respective estimates for $\rho_{\mathrm{F}}(\widehat{\mathcal{M}}), \rho_{2\to 2}(\widehat{\mathcal{M}})$ and $\gamma_2(\widehat{\mathcal{M}}, \|\cdot\|_{2\to 2})$.

Theorem 5.11. *Let* $\mathbf{A} = \operatorname{diag}\{\mathbf{\Phi}\}_{l=1}^{L} \in \mathbb{R}^{mL\times dL}$ *be a block diagonal random matrix generated by the random matrix* $\mathbf{\Phi} \in \mathbb{R}^{m\times d}$ *whose entries are independent subgaussian zero-mean, unit-variance random variables with subgaussian norm* τ. *Let* $\mathbf{\Psi} \in \mathrm{U}(dL)$ *be a unitary matrix, and assume that*

$$m \gtrsim_{\tau} \delta^{-2}s\omega_{\mathcal{I}}(\mathbf{\Psi})^2\Big[\log(D)\log(s)^2[\log(G) + g\log(s/\omega_{\mathcal{I}}(\mathbf{\Psi}))] + \log(\eta^{-1})\Big].$$

Define the scaled measurement matrix $\tilde{\mathbf{A}} := m^{-1/2}\mathbf{A}\boldsymbol{\Psi}$. *Then with probability at least* $1-\eta$, *every vector* $\mathring{\mathbf{x}} \in \mathbb{C}^D$ *acquired as* $\mathbf{y} = \tilde{\mathbf{A}}\mathring{\mathbf{x}} + \mathbf{e}$ *with* $\|\mathbf{e}\|_2 \leq \nu$ *is approximated by a minimizer* $\mathbf{x}^\star$ *of Problem* ($\mathrm{P}_{\mathcal{I},1}$) *with*

$$\|\mathring{\mathbf{x}} - \mathbf{x}^\star\|_2 \leq C_0 \frac{\sigma_s(\mathring{\mathbf{x}})_{\mathcal{I},1}}{\sqrt{s}} + C_1\nu$$

where the constants $C_0, C_1 > 0$ *only depend on* δ.

5.5 Discussion

In this section, we comment on a few connections of our attained bounds by putting them in context with related signal and acquisition models.

5.5.1 Influence of the Coherence Parameter

We focus on the general block diagonal setup first in which every sensor is equipped with an independent copy of a subgaussian random matrix. For simplicity, we choose the failure probability η in Theorem 5.10 such that the condition on m simplifies to

$$m \gtrsim_\tau \delta^{-2} s\mu_{\mathcal{I}}(\boldsymbol{\Psi})^2 \log(D)\log(s)^2[\log(G) + g\log(s/\mu_{\mathcal{I}}(\boldsymbol{\Psi}))]. \tag{5.28}$$

Moreover, we assume that the number of sensors L exceeds the group size g. As discussed in Section 5.3.2, the parameter $\mu_{\mathcal{I}}(\boldsymbol{\Psi})$ ranges between the extreme points $\sqrt{g/L}$ and 1 corresponding to the choices $\boldsymbol{\Psi} = \mathbf{F}_D$ and $\boldsymbol{\Psi} = \mathrm{Id}_D$, respectively. We may therefore also lower bound $\mu_{\mathcal{I}}(\boldsymbol{\Psi})$ by $\sqrt{g/L}$ in the last log-factor of (5.28), which yields

$$m \gtrsim_\tau \delta^{-2} s\mu_{\mathcal{I}}(\boldsymbol{\Psi})^2 \log(D)\log(s)^2\left[\log(G) + g\log\left(\frac{sL}{g}\right)\right]. \tag{5.29}$$

For $\boldsymbol{\Psi} = \mathbf{F}_D$, this shows that the number of measurements per sensor decreases almost linearly in L. This in turn implies that roughly the same recovery fidelity can be maintained if the number of measurements per sensor is reduced by adding more sensors to the acquisition system. However, since each sensor takes fewer samples in this scenario, this ultimately results in a net gain since the energy consumption per sensor is reduced. On the other hand, if target signals are group-sparse w. r. t. the canonical basis, such a reduction does not seem possible. This is due to the fact that in the worst case scenario, all active groups might be restricted to a single chunk $\mathbf{x}_l$. In this case, each measurement operator $\boldsymbol{\Phi}_l$ has to act as a group-RIP matrix. This drawback is fundamental to the acquisition model and cannot be overcome by a refined proof technique.

5.5.2 Reduction to Sparse Vector Recovery

We now consider the edge case of group-sparse recovery in which the size of each group tends to 1 and therefore $G = D$. This corresponds to the setting of sparse recovery via block diagonal operators addressed in [**Eft+15**] and [**CA18**]. With (5.29), the required number of measurements for $G = D$ reduces to

$$m \gtrsim_\tau \delta^{-2} s\mu_{\mathcal{I}}(\boldsymbol{\Psi})^2 \log(D)\log(s)^2[\log(D) + \log(sL)].$$

Since $s \leq D$ and $L = D/d \leq D$, it consequently suffices to choose

$$m \gtrsim_\tau \delta^{-2} s \mu_{\mathcal{I}}(\boldsymbol{\Psi})^2 \log(D)^2 \log(s)^2.$$

Recalling the definition of the coherence parameter

$$\mu_{\mathcal{I}}(\boldsymbol{\Psi}) = \min\left\{\sqrt{d} \max_{i \in [D]} \|\boldsymbol{\psi}_i\|_{\mathcal{I},\infty}, 1\right\},$$

we have for $\mathcal{I} = \{\{1\}, \dots, \{D\}\}$ that $\|\cdot\|_{\mathcal{I},\infty} = \|\cdot\|_\infty$ and therefore

$$\mu_{\mathcal{I}}(\boldsymbol{\Psi}) = \frac{1}{\sqrt{L}} \min\left\{\sqrt{D} \max_{i \in [D]} \|\boldsymbol{\psi}_i\|_\infty, \sqrt{L}\right\} =: \frac{1}{\sqrt{L}} \tilde{\mu}(\boldsymbol{\Psi})$$

where $\tilde{\mu}(\boldsymbol{\Psi})$ denotes a rescaled coherence parameter in accordance with the definition used by Eftekhari *et al.* (cf. Equation (5.4) and Equation (5) in [Eft+15]). This now implies that the conclusion of Theorem 5.10 holds if

$$mL \gtrsim_\tau \delta^{-2} s \tilde{\mu}(\boldsymbol{\Psi})^2 \log(D)^2 \log(s)^2,$$

which is precisely the statement of Theorem 1 in [Eft+15]. A similar argument yields the specialization to the situation in which each sensor is equipped with the same random matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m \times d}$. As discussed in Section 5.4, the parameter $\omega_{\mathcal{I}}(\boldsymbol{\Psi})$ w. r. t. the trivial group partition $\mathcal{I} = \{\{1\}, \dots, \{D\}\}$ reduces to

$$\omega_{\mathcal{I}}(\boldsymbol{\Psi}) = \max_{i \in [D]} \left\|\tilde{V}(\mathbf{e}_i)\right\|_{2 \to 2}.$$

In this case, one has that $1/\sqrt{L} \leq \omega_{\mathcal{I}}(\boldsymbol{\Psi}) \leq 1$ [Eft+15]. Defining the so-called *block-coherence* parameter $\tilde{\omega}(\boldsymbol{\Psi}) := \sqrt{L}\omega_{\mathcal{I}}(\boldsymbol{\Psi})$ to borrow terminology from Eftekhari *et al.* (cf. [Eft+15, Equation (9)]), this yields the condition

$$mL \gtrsim_\tau \delta^{-2} s \tilde{\omega}(\boldsymbol{\Psi})^2 \log(D)^2 \log(s)^2,$$

which reproduces the statement of Theorem 2 in [Eft+15].

5.5.3 Comparison to Dense Measurement Matrices

As alluded to several times throughout this thesis, it is by now a well-established fact that $\mathcal{O}(s \log(D/s))$ nonadaptive measurements based on subgaussian random ensembles are sufficient to stably reconstruct sparse or compressible vectors from their linear projections. Moreover, this bound is fundamental in that it is known to be optimal among all encoder-decoder pairs $(\mathbf{A}, \Delta)$ with measurement matrix $\mathbf{A} \in \mathbb{C}^{M \times D}$ and decoding map $\Delta \colon \mathbb{C}^M \to \mathbb{C}^D$ such that

$$\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_2 \leq \frac{C}{\sqrt{s}} \sigma_s(\mathbf{x})_1 \quad \forall \mathbf{x} \in \mathbb{C}^D$$

for $C > 0$ [FR13, Chapter 10]. Such a fundamental lower bound on the required number of measurements was recently also established for the case of block-sparse vectors by

Dirksen and Ullrich [DU18] (see also [ADR16, Theorem 2.4]). In particular, using new results on Gelfand numbers, the authors show that stability results of the form

$$\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_2 \leq \frac{C}{\sqrt{s}} \sigma_s(\mathbf{x})_{\mathcal{I},1} \quad \forall \mathbf{x} \in \mathbb{C}^D$$

for arbitrary encoder-decoder pairs $(\mathbf{A}, \Delta)$ require at least

$$M \geq c_1(s \log(eG/s) + sg) \quad \text{with} \quad s > c_2$$

measurements where the constants c_1 and c_2 only depend on $C > 0$ (cf. [DU18, Corollary 1.2]). Perhaps most surprisingly about this result is the linear dependence on the total number of nonzero coefficients sg. In light of (5.29), we almost recover this scaling behavior in the total number of measurements M for the block diagonal measurement setup, albeit with an additional logarithmic factor $\log(sL/g)$, which is an artifact of the proof technique employed in Section 5.3.3. The other polylogarithmic factors, as well as the dependence on $\mu_{\mathcal{I}}(\mathbf{\Psi})$, on the other hand, are due to the particulars of the measurement setup compared to the situation in which we employ one densely populated measurement matrix to observe the entire signal. Note, however, that by the discussion in Section 5.5.1, the total number of measurements becomes

$$mL \gtrsim_\tau \delta^{-2} g \log(D) \log(s)^2 \left[s \log(G) + sg \log\left(\frac{sL}{g}\right) \right]$$

if $\mathbf{\Psi}$ is chosen as the DFT matrix. This means that the resulting bound scales quadratically in g. We point out that the multiplicative dependence on g in (5.25) originates from the volume comparison argument invoked in the context of our extension of Maurey's lemma (see Section 5.3.3). We conjecture that this dependence on g is suboptimal and should be improvable by a more sophisticated proof technique. Whether the dependence on the additional log-factors can be improved any further remains an open problem.

5.5.4 Distributed Compressed Sensing

As mentioned in the introduction, the measurement model (5.1) frequently appears in the context of recovering multiple versions of a vector sharing a common low-complexity structure. This model appears for instance in the context of distributed sensing where one aims to estimate the structure of a ground truth signal observed by spatially distributed sensors which each observe a slightly different version of the signal due to channel propagation effects. Another classic example is that of the so-called MMV model in which a single sensor acquires various temporal snapshots of a signal whose low-complexity structure is assumed to be stationary[8] with the intent of reducing the influence of measurement noise in a single-snapshot model. This particular model can be cast in the setting of Section 5.4 where we interpret each observation in the MMV model as an independent observation by a distinct sensor equipped with the same measurement matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times d}$. Assuming that the ground truth signal is s-sparse, we can interpret both situations as trying to recover an s-group-sparse vector w. r. t. the group partition $\mathcal{I} = \{\mathcal{I}_1, \dots, \mathcal{I}_d\}$ with

$$\mathcal{I}_i = \{i, d + i, \dots, (L-1)d + i\}. \tag{5.30}$$

[8] In particular, this model assumes the support set to be constant, while amplitudes and phases of the coefficients of each vector are allowed to change between different observations.

In both situations, we assume that each signal $\mathbf{z}_l = \widetilde{\mathbf{\Psi}}\mathbf{x}_l \in \mathbb{C}^d$ is sparse in the same basis $\widetilde{\mathbf{\Psi}} \in \mathrm{U}(d)$. We can therefore choose $\mathbf{\Psi} = \operatorname{diag}\{\widetilde{\mathbf{\Psi}}\}_{l=1}^L \in \mathrm{U}(D)$ in Theorem 5.10. This setup, however, is not able to cope with certain adversarial vectors. More precisely, due to the particular group partition structure, the knowledge about the periodicity in the support structure can not necessarily be exploited in all recovery scenarios. To see this, consider the situation in which only a single vector $\mathbf{x}_l$ is different from $\mathbf{0}$. The vector $\mathbf{x} = (\mathbf{0}^\top \ \dots \ \mathbf{0}^\top \ \mathbf{x}_l^\top \ \mathbf{0}^\top \ \dots \ \mathbf{0}^\top)^\top$ is then by definition s-group-sparse (w. r. t. the group partition $\mathcal{I}$) if $\mathbf{x}_l$ is s-sparse. Regardless of the sparsity basis $\widetilde{\mathbf{\Psi}} \in \mathrm{U}(d)$, only the vector $\mathbf{y}_l$ carries information about $\mathbf{x}_l$, which implies that each matrix $\mathbf{\Phi}_l$ should satisfy the classical restricted isometry property to recover $\mathbf{x}$. This happens with high probability as soon as $m = \Omega(s \log(d/s))$. In this case, instead of solving Problem ($\mathrm{P}_{\mathcal{I},1}$) directly, it is more favorable to solve for each $l \in [L]$ the problem

$$\begin{aligned} \underset{\mathbf{u}}{\text{minimize}} \quad & \|\mathbf{u}\|_1 \\ \text{s.t.} \quad & \mathbf{y}_l = \mathbf{\Phi}_l \widetilde{\mathbf{\Psi}} \mathbf{u}. \end{aligned}$$

Unfortunately, this behavior is not accurately captured by Theorem 5.10 since we have by (5.28) with $G = d$ and $g = L$ that

$$m \gtrsim_\tau \delta^{-2} \mu_{\mathcal{I}}(\mathbf{\Psi})^2 s \log(D) \log(s)^2 [\log(d) + L \log(s/\mu_{\mathcal{I}}(\mathbf{\Psi}))].$$

This predicts a much worse scaling behavior than what is required for each matrix $\mathbf{\Phi}_l \widetilde{\mathbf{\Psi}}$ to satisfy the canonical RIP. The problem is ultimately rooted in the fact that independent of $\widetilde{\mathbf{\Psi}} \in \mathrm{U}(d)$, only the measurements $\mathbf{y}_l$ carry information about $\mathbf{x}_l$. Since uniform recovery guarantees based on the restricted isometry property or variants thereof represent worst-case analyses, such adversarial examples are unavoidable in general. It would therefore be of interest to establish an average case error bound in the spirit of [**BKR10**] since such bounds are often more relevant from a practical perspective.

Adversarial situations in joint-sparse recovery had previously been discussed by van den Berg and Friedlander [**BF09**] who consider sufficiency conditions for noiseless joint-sparse recovery based on dual certificates. Instead of considering signals with only one s-sparse nonzero signal $\mathbf{x}_l$, they consider signals $\mathbf{x}$ in which every $\mathbf{x}_l$ is at most 1-sparse with $\operatorname{supp}(\mathbf{x}_l) \neq \operatorname{supp}(\mathbf{x}_{l'})$ for any $l \neq l'$. In this setting, they show that there are signals $\mathring{\mathbf{x}} \in \mathbb{R}^D$ which—given the linear measurements $\mathbf{y} = \operatorname{diag}\{\mathbf{\Phi}\}_{l=1}^L \mathring{\mathbf{x}}$—can provably be recovered by the program

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \mathbf{y} = \operatorname{diag}\{\mathbf{\Phi}\}_{l=1}^L \mathbf{x} \end{aligned}$$

but not via group ℓ_1-minimization, *i.e.*, as solutions of Problem ($\mathrm{P}_{\mathcal{I},1}$) with $\mathbf{\Psi} = \mathrm{Id}_D$ and $\nu = 0$.

As briefly commented on in Section 5.2, the problem of distributed compressed sensing was also recently addressed in the context of quantized compressed sensing with binary observations by Maly and Palzer [**MP19**] who impose an additional norm constraint on each signal to avoid adversarial scenarios as outlined above. However, even with this modified signal model, the adversarial example discussed above still applies if one signal $\mathbf{x}_l$ is exactly s-sparse, while any other signal $\mathbf{x}_{l'}$ with $l' \neq l$ is 1-sparse with the entire

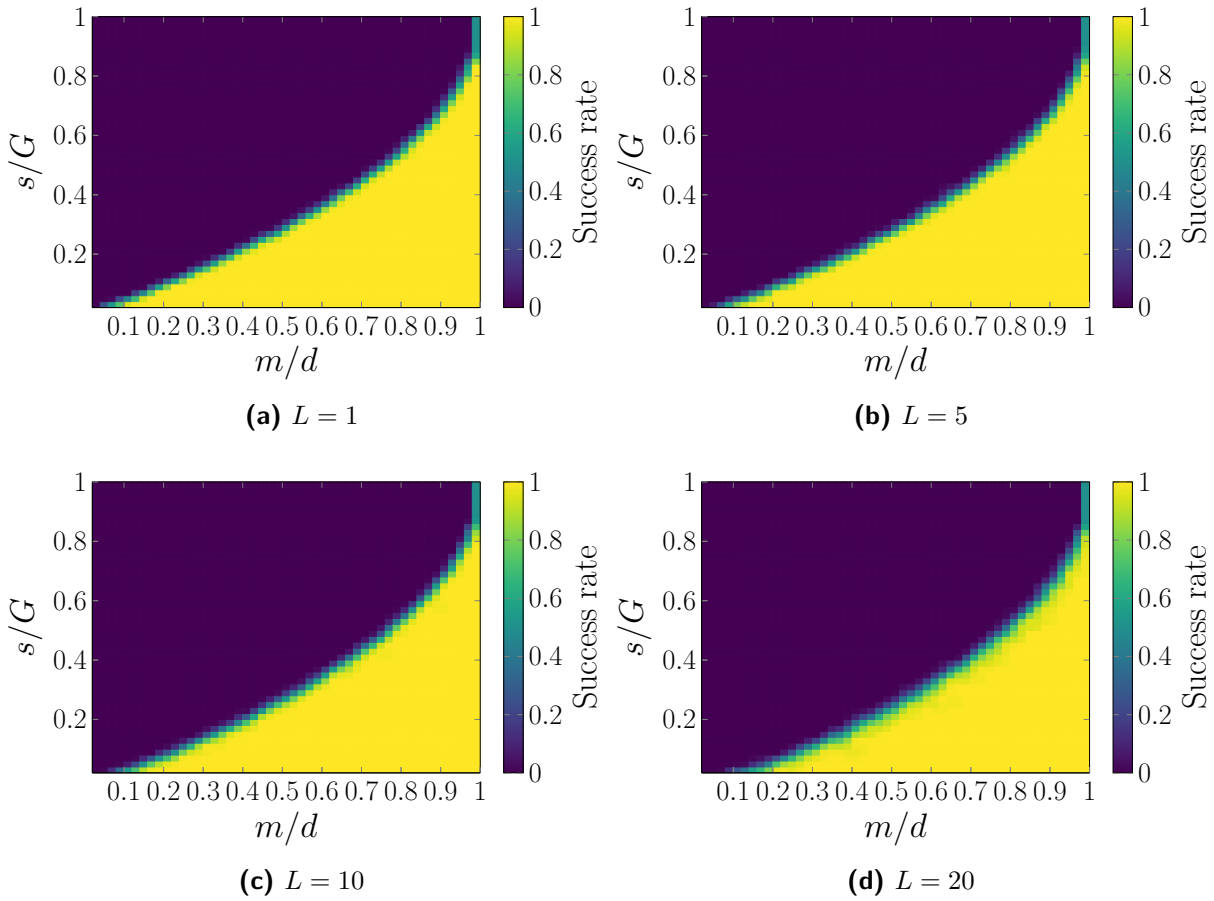
signal energy concentrated on the same coordinate in each vector $\mathbf{x}_{l'}$. The resulting signal is therefore s-group-sparse as in the previous example. In that case, each measurement vector $\mathbf{y}_{l'}$ only carries information about a single nonzero coordinate of $\mathbf{x}_l$, which implies that each $\mathbf{\Phi}_l$ must itself be able to recover every $(s-1)$-sparse vector for the entire vector $\mathbf{x}$ to be recovered as desired.

To summarize, without further restrictions on the particular signal model, it is not clear how adversarial examples as discussed above can be dealt with in order to obtain nontrivial uniform recovery guarantees. However, the conclusion of the work in [Eft+15] and our results is that sparsity or group-sparsity in a nonlocalized unitary basis such as the DFT basis bears the potential to reduce the number of measurements required for stable and robust signal recovery by distributing the energy of nonzero coefficients across the entire signal support. As pointed out above, however, this requires that the unitary matrix corresponding to the sparsity basis of the signal class does *not* itself exhibit a block diagonal structure.

5.6 Empirical Phase Transition Evaluation

We now turn to an empirical investigation of the group-sparse recovery problem from block diagonal observations in terms of the so-called *phase transition* phenomenon. Such phenomena collectively describe the sudden change in behavior of a system when certain parameters cross a critical threshold. In the compressed sensing literature, it has been observed early on that such a critical line exists where recovery of s-sparse vectors in $\mathbb{R}^d$ from m measurements changes from almost certain success to almost certain failure when the number of measurements and the sparsity level varies over the half-open unit square $(m/d, s/m) \in (0,1]^2$. A substantial body of research has since been dedicated to explain, predict and quantify both the position, as well as the width of the transition region [DT05; Don06b; DT09b; DT09a; DT10a; DT10b]. The first result to rigorously ascertain the phase transition behavior in the nonasymptotic regime for Gaussian measurement ensembles was reported by Amelunxen *et al.* in [Ame+14]. Their work, which exposes a deep connection between successful recovery via ℓ_1-minimization and the concentration behavior of so-called *intrinsic volumes* in the theory of conic integral geometry, first managed to not only establish that recovery succeeds in one region, but also that recovery will fail with high probability in the other. This is in stark contrast to previous results which were only able to predict the position of the success region but otherwise could not assess whether recovery would succeed or fail in the other.

Throughout our experiments, we consider vectors $\mathring{\mathbf{x}} \in \mathbb{C}^D$ with $D = 1000$. For a fixed number of L sensors, we draw L random matrices $\tilde{\mathbf{\Phi}}_l \in \mathbb{R}^{d\times d}$ populated by independent standard Gaussian random variables. These matrices are then fixed throughout the process of generating one phase transition diagram. Given a pair (m, s), we construct the individual sensing matrices $\mathbf{\Phi}_l \in \mathbb{R}^{m\times D}$ by retaining the first m rows of each square matrix $\tilde{\mathbf{\Phi}}_l$ to form the compound block diagonal sensing matrix $\mathbf{A} = \operatorname{diag}\{m^{-1}\mathbf{\Phi}_l\}_{l=1}^{L}$. We partition the index set $[D]$ into $G = 100$ nonoverlapping groups $\mathcal{I} = \{\mathcal{I}_1, \ldots, \mathcal{I}_{100}\}$ such that every group $\mathcal{I}_i$ contains $g = 10$ elements. To that end, we shuffle the elements of the set $[D]$ and split them into G groups, which we fix throughout all experiments. For each of the 50×50 parameter combinations (m, s), we draw 20 s-group-sparse vectors $\mathring{\mathbf{x}} \in \mathbb{C}^D$, which we recover via Problem ($\mathrm{P}_{\mathcal{I},1}$). Given the group partition $\mathcal{I}$, we draw the



(a) $L = 1$ **(b)** $L = 5$

(c) $L = 10$ **(d)** $L = 20$

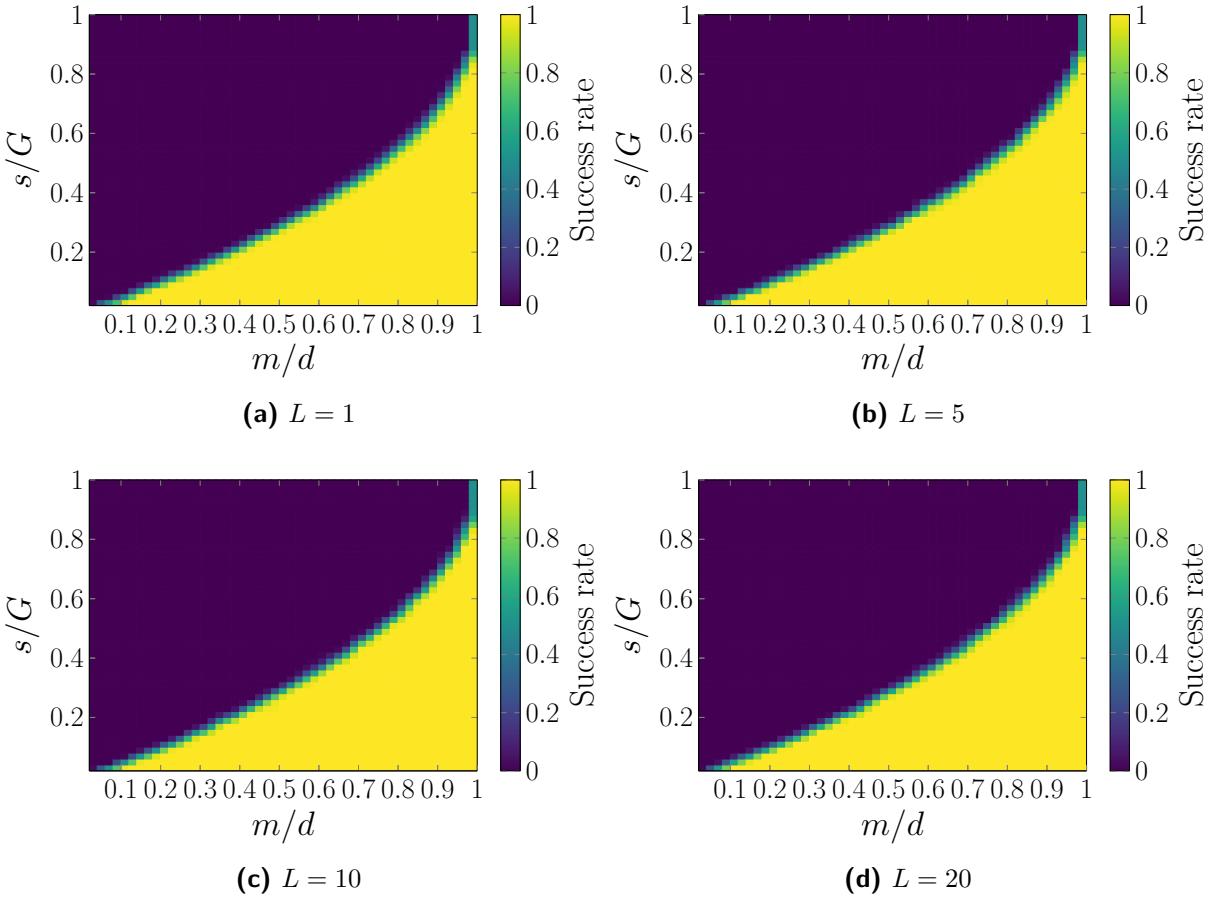
Figure 5.1: Phase transition diagrams for different numbers of sensors (L) with $\mathbf{\Psi} = \mathrm{Id}_D$ when the group-sparsity level s and the number of measurements m per sensor vary, and the number of groups G and the signal dimension per block d is fixed

set of active groups uniformly at random from $[G]$. The nonzero entries in each group are then populated by circularly symmetric Gaussian random variables. In other words, given an active group index $k \in S = \mathrm{supp}_{\mathcal{I}}(\mathring{\mathbf{x}})$, we set $\mathring{\mathbf{x}}_{\overline{\mathcal{I}_k}} = \mathbf{0}$ and $\mathring{\mathbf{x}}_{\mathcal{I}_k} = 2^{-1/2}(\mathbf{g}_k + i\mathbf{h}_k)$ where $\mathbf{g}_k, \mathbf{h}_k \in \mathbb{R}^g$ denote two independent standard Gaussian random vectors and $i = \sqrt{-1}$. We then measure how many vectors are successfully recovered according to the success criterion

$$\frac{\|\mathring{\mathbf{x}} - \mathbf{x}^\star\|_2}{\|\mathring{\mathbf{x}}\|_2} \leq 10^{-3}$$

with $\mathbf{x}^\star$ denoting the optimal solution of Problem ($\mathrm{P}_{\mathcal{I},1}$) for $\mathbf{y} = \mathrm{diag}\{\mathbf{\Phi}_l\}_{l=1}^L \mathbf{\Psi}\mathring{\mathbf{x}}$. We repeat this experiment for two different sparsity bases $\mathbf{\Psi} \in \mathrm{U}(D)$ at the low and high end of the coherence spectrum, namely the DFT and the canonical basis.

The results of the first set of experiments in which we investigate the recovery of



(a) $L = 1$ **(b)** $L = 5$

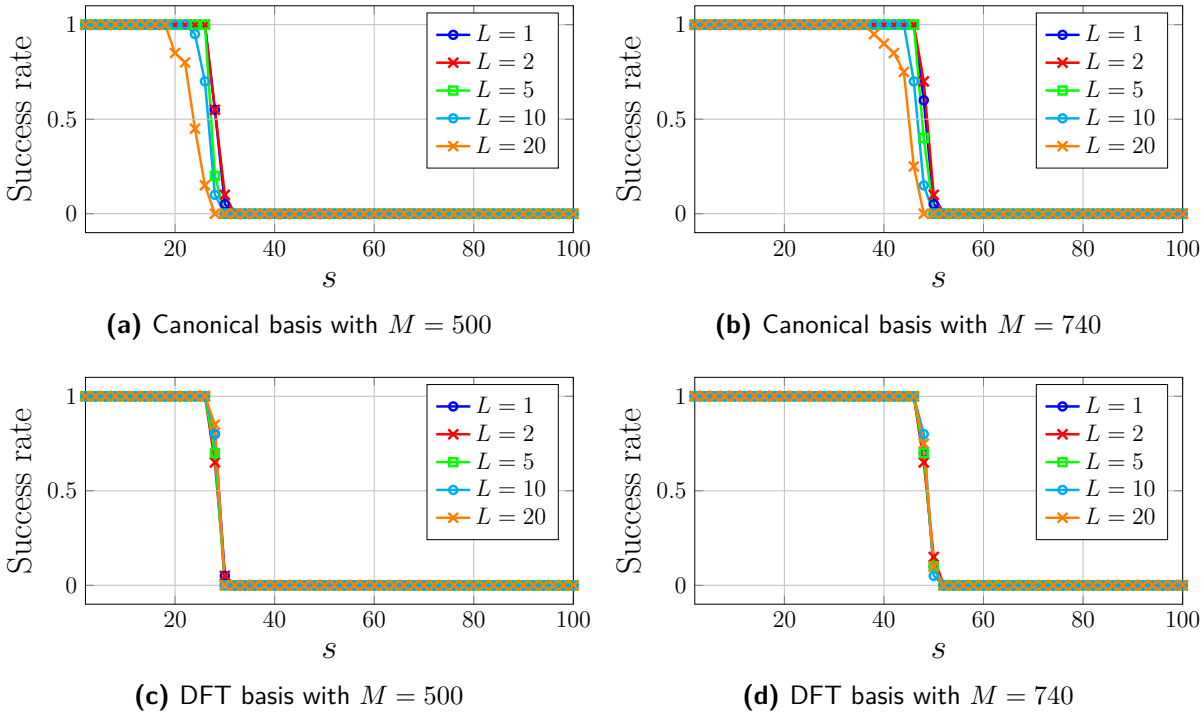
(c) $L = 10$ **(d)** $L = 20$

Figure 5.2: Phase transition diagrams for different numbers of sensors with $\mathbf{\Psi} = \mathbf{F}_D$

group-sparse vectors w. r. t. the canonical basis are shown in Figure 5.1.[9] Despite the fact that our bound does not predict that the number of measurements required per sensor for $\mathbf{A}$ to satisfy the group-RIP decreases linearly with L, the differences in performance are much less dramatic than one might anticipate. The biggest differences are observed for small values of m. More precisely, for $L = 1$, the transition line tapers off slightly more for $m \to 0$ compared to the scenario where $\mathbf{A}$ contains $L = 20$ blocks. Additionally, it appears that the transition zone where the empirical recovery rate changes from successful recovery with probability 1 to 0 slightly widens as L increases.

We repeat the same experiment for group-sparse signals in the frequency domain, *i.e.*, we set $\mathbf{\Psi} = \mathbf{F}_D$. The results are shown in Figure 5.2. As predicted by Theorem 5.10, the effects of varying L are even less pronounced than in case of the canonical basis since neither the previous behavior around $m = 0.1d$, nor the widening of the transition zone can

[9]Note that we normalize abscissa and ordinate by D and G, respectively. In phase transition diagrams for sparse recovery, it is often more desirable to normalize the ordinate by M to magnify the transition behavior at lower values of M. This is motivated by the fact that there is no hope to recover an s-sparse vector in $\mathbb{C}^D$ from fewer than s observations. In other words, for a fixed M, it suffices to consider the range $s \in (0, M]$. In our case, however, this would severely limit resolution since an s-group-sparse vector has $s \cdot g$ rather than s nonzero entries. By considering 50 uniformly spaced values for s, this implies for $g = 10$ that the lowest value we can consider on the abscissa would be $M/D = m/d = 0.5$. Considering that this excludes half the range for m, we therefore opt to consider the full range of values for s between 1 and G for every fixed m.



(a) Canonical basis with $M = 500$

(b) Canonical basis with $M = 740$

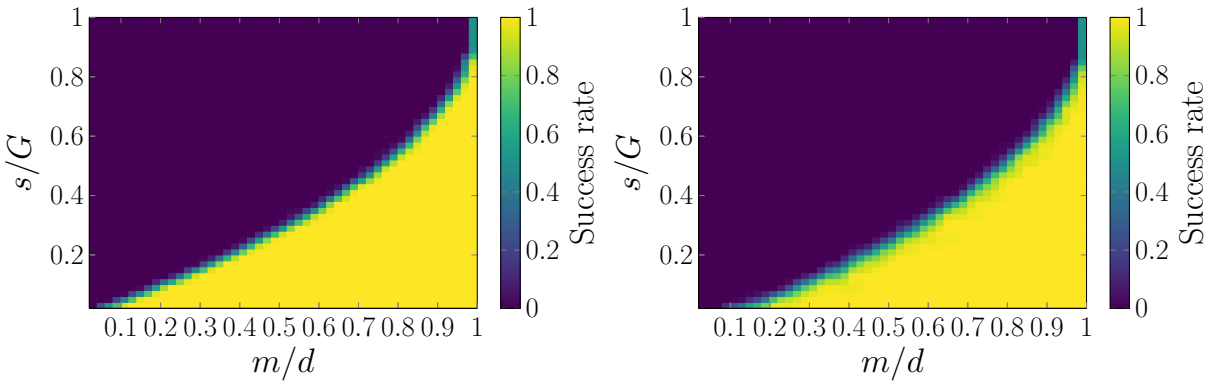
(c) DFT basis with $M = 500$

(d) DFT basis with $M = 740$

Figure 5.3: Sectional cuts through the phase transition diagrams in Figure 5.1 and 5.2 demonstrating the effects of varying numbers of sensors on the recovery performance for different sparsity bases

be observed. This confirms the intuition that the incoherence of the Fourier basis with the canonical basis allows for a reduction in the number of measurements per sensor without affecting the overall reconstruction performance. To inspect this behavior a little closer, we additionally plot two sections through each phase transition diagram for $M = 500$ and $M = 740$ in Figure 5.3. This representation clearly demonstrates the diminishing performance with an increased number of measurements for canonically group-sparse vectors. For frequency group-sparse vectors, however, the performance is invariant under the choice of L.

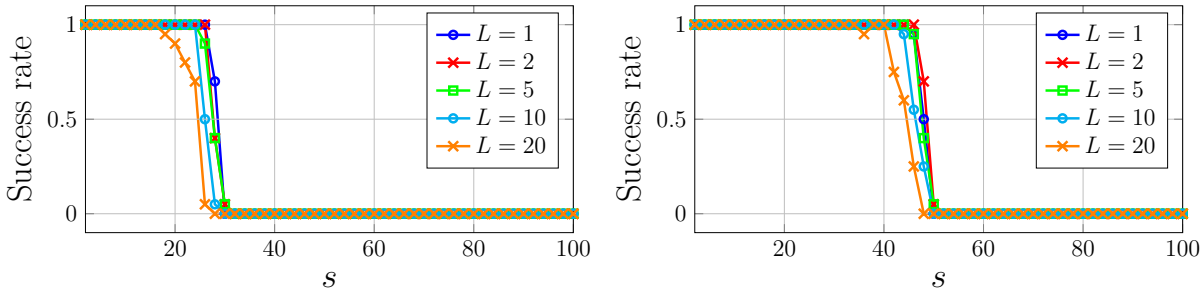
Finally, we conduct the same experiments as before for the scenario in which each sensor is equipped with a copy of the same random matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times d}$ which is drawn once and then fixed throughout all subsequent experiments. The results are shown in Figure 5.4. The phase transition diagrams confirm the assumption that the general recovery behavior is comparable to the previous setting given the identical dependence of m on s, g, D and G predicted by Theorem 5.11. More precisely, we observe a similar widening of the transition zone as the number of sensors increases both for the canonical and the Fourier basis, as well as a reduced tapering of the phase transition diagrams for small m. In contrast to the scenario in which we equip each sensor with an independent sensing matrix, the sectional cuts through the individual diagrams depicted in Figure 5.5 further reveal a slight drop in recovery performance for the DFT basis as the number of sensors L increases. This effect is likely captured by the parameter $\omega_{\mathcal{I}}(\mathbf{\Psi})$, which—due to its complicated nature—does not admit a straightforward calculation and interpretation for $\mathbf{\Psi} = \mathbf{F}_D$ as the coherence parameter $\mu_{\mathcal{I}}(\mathbf{\Psi})$ in the previous setting. Finding a more meaningful bound for $\omega_{\mathcal{I}}(\mathbf{\Psi})$ therefore remains an interesting open problem in this context.



(a) $L = 1,\ \mathbf{\Psi} = \mathrm{Id}_D$ **(b)** $L = 20,\ \mathbf{\Psi} = \mathrm{Id}_D$

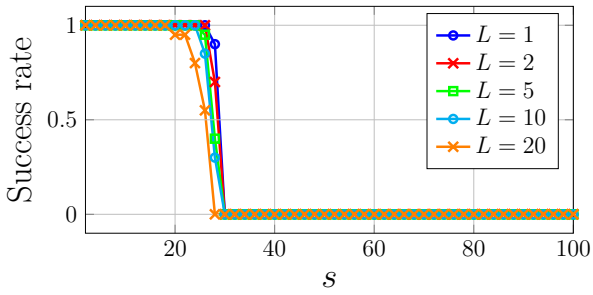
(c) $L = 1,\ \mathbf{\Psi} = \mathbf{F}_D$ **(d)** $L = 20,\ \mathbf{\Psi} = \mathbf{F}_D$

Figure 5.4: Phase transition diagrams for different numbers of sensors and sparsity bases when each sensor is equipped with an identical copy of the prototype subgaussian measurement matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times d}$

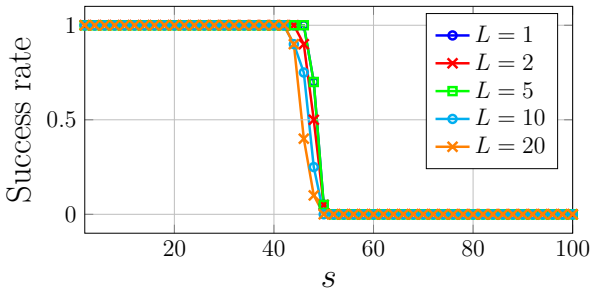


(a) Canonical basis with $M = 500$

(b) Canonical basis with $M = 740$



(c) DFT basis with $M = 500$



(d) DFT basis with $M = 740$

Figure 5.5: Sectional cuts through the phase transition diagrams in Figure 5.4 demonstrating the effects of varying numbers of sensors on the recovery performance for different sparsity bases

5.7 Conclusion

In this chapter, we established conditions on the number of measurements required to stably and robustly recover group-sparse vectors by means of block diagonal measurement matrices whose blocks either consist of independent or identical copies of a subgaussian random matrix. Appealing to a powerful concentration bound on the suprema of chaos processes, we derived conditions on the number of measurements required for subgaussian block diagonal random matrices to satisfy the so-called group restricted isometry property. This generalizes an earlier result due to Eftekhari *et al.* who first established a similar result for the canonical sparsity model. As a side effect of our proof, we established a generalization of Maurey's lemma, which allows estimating the covering number of sets which can be expressed as the convex combination of elements drawn from compact sets.

Although certain adversarial group partitions, which includes the distributed sensing model, may lead to suboptimal predictions on the number of measurements for stable and robust recovery, such cases are generally avoided if signals are group-sparse in nonlocalized sparsity bases, whose basis matrices are in turn not block diagonal. In this case, our results predict almost optimal scaling behavior up to logarithmic factors. If target signals admit a group-sparse representation in terms of the discrete Fourier transform basis, the number of measurements per sensor decreases linearly in L. This implies that reducing the number of measurements in each sensor can be compensated for by adding more sensors to the acquisition system without affecting the reconstruction quality too much. This in turn reduces both the storage requirements and energy consumption of each sensor, allowing for the deployment of less sophisticated and hence cheaper sampling devices.

Open Problems

As alluded to in the discussions in Section 5.5, there remain several open problems, which deserve further attention in order to provide a more comprehensive theoretical understanding of the block diagonal measurement model for group-sparse signal recovery. Most prominently, these issues include answering the question whether the dependence on the group size g can be improved to match more closely the fundamental lower bound on the number of measurements established for general sensing matrices. We point out that one potential avenue might be the consideration of a group-sparse variant of the asymmetric restricted isometry property considered by Chun and Adcock in [**CA18**]. However, the general analysis follows the same strategy employed in [**Eft+15**] and our work. The improvements we expect from such an undertaking are therefore in terms of the coherence parameter $\mu_{\mathcal{I}}(\boldsymbol{\Psi})$ as discussed in Section 5.2, not in an improved estimate of the γ_2-functional since the problem of bounding the covering number of the matrix set $\mathcal{M}$ remains.

Given the pessimistic nature of RIP-based recovery guarantees, it is highly desirable to complement our findings with an average case analysis along the lines of [**BKR10**], which can be expected to paint a much more optimistic picture. This assessment is also reinforced by our numerical experiments, which empirically demonstrate that the gap between recovery of canonically and Fourier group-sparse signals is much less pronounced than what might be expected from the recovery bounds established in this chapter.

Similar to the previous chapter, we emphasize again that in certain applications, it is of vital importance to allow for overlapping group structures to facilitate satisfactory signal modeling. So far, we categorically excluded such models. In the future, however, it would be desirable to examine whether our results can also be extended to groups with overlapping coefficients. The main issue in this context is finding a suitable formulation of a recovery problem in the spirit of Problem ($\mathrm{P}_{\mathcal{I},1}$). For instance, naively minimizing the sum of ℓ_2-norms of each group to generalize the group ℓ_1-norm leads to unwanted side effects such as selecting all groups which contain one index if it is present in one of the groups [**HBM12**]. This issue is rooted in the fact that with overlapping groups, there no longer exists a unique decomposition of each vector into individual 1-group-sparse vectors. It is therefore necessary to consider alternative formulations of group-sparsity priors, which are much less studied than formulations based on the group ℓ_1-norm in the nonoverlapping case. A starting point might be the recent work by Ahsen and Vidyasagar [**AV17**], which provides a general framework to study the recovery performance of various group-sparsity priors.

Finally, despite the fact that the block diagonal acquisition system allows for a reduction in computational complexity by reducing the number of measurements per sensor, it does not address the issue that implementing purely random measurement operators in a practical system is difficult in general. To overcome such issues, it would therefore be desirable to combine the block diagonal measurement model with other more structured designs such as random Toeplitz matrices [**Hau+10**] or circulant matrices generated by subgaussian random vectors [**RRT12**], giving rise to random convolution operators with efficient hardware implementations. In the latter case, one might be able to reuse the analysis presented in [**KMR14**], which also introduced the chaos bound in the spirit of Theorem 5.5 used to establish the recovery guarantees in this chapter.

6

Conclusion

The vast majority of theoretical works on compressed sensing operate under the assumption that measurements are essentially available during recovery at infinite precision. Given the importance of digital signal processing, however, practical CS-based measurement devices are required to transfer analog signals into the finite-valued digital domain for subsequent processing, analysis, storage or transmission. This step generally introduces a nonnegligible amount of signal-dependent quantization noise in the measurements, which can only be appropriately accounted for within the classical theory under a high-resolution assumption. Moreover, in certain areas such as wireless communication or image processing, which commonly involve exceedingly high-dimensional signals, acquiring compressive measurements might still be prohibitive if ADCs of an acquisition system are required to record high-resolution measurements. Since there exists an inverse relationship between the bit depth of an ADC and its attainable sampling rate, acquiring high-precision measurements at exceedingly high sampling rates therefore often necessitates the use of costly specialized and energy-demanding hardware circuitry.

In this thesis, we explored several different avenues to recover low-complexity signals from excessively quantized measurements or partial compressive observations acquired by one or more distinct sensors. These problems are motivated by technological advances

towards cheap and energy-efficient sensing devices in emerging communication and monitoring methodologies such as the internet of things, Industry 4.0, as well as more general wireless sensor networks. The main focus in this context was on the 1-bit compressed sensing acquisition model, in which each component of a measurement vector is represented by a single information bit representing the sign of the associated linear measurement. This simplistic, memoryless quantization scheme is particularly well-suited for use in low-cost sensing devices due to the ability to implement each quantizer in the system by means of simple comparators operating at a fixed voltage level.

Our second main focus was on a measurement paradigm in which either a single or multiple sensors observe distinct portions of a particular target vector. In the first case, this measurement model is motivated by applications such as video streaming or imaging, which commonly necessitate the acquisition of low-complexity signals in *chunks* due to exceedingly high dimensions of the respective signal space. Moreover, in multi-sensor networks, the acquisition model can be used to reduce the amount of data that individual sensors have to acquire and transmit to a dedicated fusion center to reconstruct low-complexity target signals. Assuming that it is possible to compensate a potential loss in reconstruction fidelity due to a reduction in the number of measurements per sensor by adding more sensors to the network, this model allows to trade off reconstruction fidelity for cost reduction of the sensor nodes, as well as transmission load between individual sensors and the fusion center.

In the following, we summarize the results of the three main chapters of this thesis.

6.1 Summary

Chapter 3: Estimation of Frequency-Sparse Signals from Binary Measurements

While sparsity represents a ubiquitous signal characteristic in various domains of science and engineering, such low-complexity signal structures usually only reveal themselves after transforming elements of a signal class into a suitable representation in terms of an orthonormal basis or more generally an overcomplete dictionary or frame. Due to its central importance in fields like medical imaging, wireless communication, radar and seismology, which abound with periodic, oscillatory phenomena, the Fourier basis represents one of the most important sparsity bases in CS-based engineering applications.

In Chapter 3, we considered the acquisition of frequency-sparse signals from 1-bit quantized real-valued time domain measurements. Rather than considering random mixing of compressive measurements with purely random ensembles, which is generally difficult to realize in a physical system, we instead considered hardware-friendly subsampling or oversampling schemes. Adopting various reconstruction schemes proposed in the literature on 1-bit compressed sensing, we demonstrated empirically that faithful signal recovery is possible by exploiting the symmetric nature of the conjugate symmetric signal space. In particular, we developed a modification of the *binary iterative hard thresholding* algorithm geared towards sparse conjugate symmetric signal recovery. Most importantly, this modification requires a conjugate symmetric version of the *hard thresholding operator* to guarantee that the algorithm always yields solutions with a real-valued inverse discrete Fourier transform. Appealing to a noise-adaptive reconstruction scheme proposed in the literature, it was also shown that the proposed method can be hardened against noisy

1-bit observations, provided that an estimate of the number of bit flips is available.

While the concept of oversampling (*i.e.*, acquiring more measurements than the dimension of the signal space) is of no interest in the classical linear CS theory, the same does not apply in the nonlinear setting. This is rooted in the fact that an overdetermined system of linear equations always admits a unique solution, while a nonlinear system does not.[1] Considering the assumption of cheap and energy-efficient 1-bit quantizers, we therefore also considered super-Nyquist sampling of conjugate symmetric vectors. The measurement operator proposed in this context is based on the concept of *exact interpolation*, a simple frequency domain zero-padding scheme for interpolation in the time domain, which does not rely on random sampling. It was demonstrated numerically that the proposed oversampling scheme improves reconstruction fidelity beyond the level attainable for compressive 1-bit observations. Moreover, since the oversampling scheme does not rely on random sampling, the proposed method is particularly hardware-friendly.

Chapter 4: Single-Bit Group-Sparse Signal Recovery

In Chapter 4, we turned our attention to the recovery of group-sparse signals from 1-bit quantized measurements of Gaussian linear projections. This generalized sparsity model assumes that nonzero coefficients in a signal are always confined to nonoverlapping coefficient groups instead of appearing in isolated positions. Group-sparse modeling finds widespread adoption in areas such as wireless communication, facial recognition, speech detection and model selection in statistics.

Generalizing several recovery schemes proposed in the 1-bit CS literature to the group-sparse setting, we showed that group-sparse vectors can be estimated up to a desired fidelity. The required number of measurements depends optimally on the group-sparsity level, the group size and the ambient signal space dimension. In this context, we also established a novel noise robustness result for a simple noniterative group hard thresholding scheme. The correct behavior of the individual reconstruction methods was further confirmed in a series of numerical experiments.

While the sign operator used to model the 1-bit quantization step is invariant under positive scaling of its input, leading to an unresolvable global scale ambiguity during reconstruction, this ambiguity can be removed by adding a known pre-quantization dither to each coordinate of the linear part of the acquisition model. Under this model, it was also demonstrated in Chapter 4 that both direction and norm of group-sparse vectors can be estimated if target vectors are confined to scaled unit balls, whose radius is known a priori. In this context, we analyzed six different reconstruction strategies by relating them to results previously established in the undithered setting. Again, we complimented our theoretical analysis with various numerical experiments on synthetic data to confirm the behavior of each recovery scheme empirically.

[1] We abuse terminology here and refer to a nonlinear equation system of the form $\mathbf{y} = f(\mathbf{A}\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{A} \in \mathbb{R}^{m \times d}$ and $f \colon \mathbb{R}^m \to \mathbb{R}^m$ an arbitrary nonlinearity as *overdetermined* if $m \geq d$.

Chapter 5: Recovery of Group-Sparse Vectors with Block Diagonal Measurement Operators

In the last main chapter of this thesis, the problem of recovering group-sparse vectors from partial compressive observations was considered. This model gives rise to two distinct sensing scenarios. In the first scenario, it is assumed that a single sensor acquires measurements of a high-dimensional group-sparse vector in lower-dimensional chunks such that individual linear measurements only capture partial information about the target vector. The second scenario assumes that multiple sensors, which are each equipped with their own subgaussian random sensing matrix, observe distinct portions of a vector. In both cases, the resulting measurement operator is modeled as a block diagonal subgaussian random matrix whose block diagonal elements are either identical or independent copies of a prototype random matrix.

In order to establish conditions under which stable and robust recovery is possible, we appealed to a group-sparse version of the restricted isometry property. Our resulting bound depends linearly on the group-sparsity level (up to polylogarithmic factors). Due to the block diagonal structure of the measurement operator, it is to be expected that any bound on the number of measurements also depends locally on the considered sparsity basis. This had previously been demonstrated in the canonical sparsity setting and extends to the group-sparse case. In particular, the respective coherence parameter yields the most favorable scaling in case of the discrete Fourier basis in which case our bound shows that the number of measurements per sensor can be reduced if more sensors are added to the system. In the case of the canonical basis, this reduction does not seem possible. These observations were also confirmed during numerical experiments. More precisely, it was demonstrated that the phase transition behavior remains almost unchanged in case of the Fourier basis when the total number of measurements $M = mL$ is fixed and the ratio M/L varies. In case of the canonical basis on the other hand, the success region of the phase transition diagram visibly shrinks as the number of sensors increases when the total number of measurements is fixed.

6.2 Outlook and Future Work

In addition to the open problems discussed in the conclusions of the individual main chapters, we single out the following problems in particular.

The signal model considered in Chapter 3 is admittedly highly idealized in that frequencies making up the conjugate symmetric signals are assumed to be integer multiples of the frequency resolution. If this assumption is violated, target signals do not admit a sparse representation w. r. t. the discrete Fourier basis, nor are they compressible in the sense that nonzero coefficients decay as $i^{-1/p}$ for some $p \in (0, 1)$, where i denotes the i-th largest entry of a vector in absolute value. One approach to address this issue is by replacing the DFT basis by an overcomplete DFT frame as previously considered in [DB13] for the recovery of frequency-sparse signals from linear Gaussian observations. This approach should be contrasted with the construction of the measurement operator in the context of 1-bit recovery from oversampled time domain measurements, which, up to scaling of the columns, also corresponds to a DFT frame. A second option could be via the so-called *atomic norm minimization* framework, which allows to lift any integrality

constraint on the frequencies of target signals. This was recently considered in [FC18] for measurement matrices composed of circularly symmetric standard Gaussian random variables.

While the bound on the number of measurements to guarantee stable and robust recovery of group-sparse vectors from block diagonal observations established in Chapter 5 is optimal w. r. t. the sparsity level, it scales quadratically in the group size when the group-sparsity basis corresponds to the orthogonal DFT matrix. This artifact can be traced back to our bound on the covering number at higher scales, which we established by generalizing Maurey's empirical method to sets that do not admit a polytope representation. Since it is unlikely that this dependence is optimal, it is highly desirable to remove this additional factor. Moreover, it is unknown whether the logarithmic factors in our bound can be improved. In both cases, establishing a fundamental lower bound on the number of measurements required to guarantee the existence of stable encoder-decoder pairs bears the potential to resolve both questions. A natural starting point would be the work in [DU18], which establishes lower bounds on the Gelfand numbers of mixed-norm embeddings, leading to optimal lower bounds on the number of measurements to guarantee stable recovery of block-sparse signals.

A

Mathematical Preliminaries

In this section, we introduce some common definitions and properties of random variables. We also collect some useful concepts and results from convex analysis and geometric functional analysis used throughout this thesis.

A.1 Random Variables and Subgaussian Distributions

Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space consisting of the sample space Ω, the Borel measurable event space represented by the σ-algebra $\Sigma \subseteq 2^{\Omega}$ and a probability measure $\mathbb{P}\colon \Sigma \to [0,1]$. The elements of the space of matrix-valued Borel measurable functions from Ω to $\mathbb{R}^{m\times d}$ are called *random matrices*. This space inherits a probability measure as the pushforward of the measure $\mathbb{P}$, *i.e.*, given a Borel measurable set $A \in \mathcal{B} \subset \mathbb{R}^{m\times d}$, we have

$$\mu_X(A) := \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}) = \mathbb{P}(X^{-1}(A)) \triangleq \mathbb{P}(X \in A)$$

such that the triple $(\mathbb{R}^{m\times d}, \mathcal{B}, \mu_X)$ is again a probability space. For $d = 1$, we obtain the space of random vectors; the space of random variables corresponds to the choice $m = d = 1$. Given a scalar random variable X, the *expected value* of X is defined as

$$\mathbb{E}X := \int X \mathrm{d}\mathbb{P} \triangleq \int_{\Omega} X(\omega)\mathbb{P}(\mathrm{d}\omega)$$

if the integral exists. Moreover, if $\mathbb{E}e^{tX}$ exists for all $|t| < h$ for some $h \in \mathbb{R}$, then the map

$$M_X \colon \mathbb{R} \to \mathbb{R} \colon t \mapsto M_X(t) = \mathbb{E}e^{tX} = \int e^{tX} \mathrm{d}\mathbb{P},$$

known as the *moment generating function* (MGF), fully determines the distribution of X. Finally, the p-th absolute moment of X is defined as

$$\mathbb{E}|X|^p = \int_\Omega |X(\omega)|^p \mathbb{P}(\mathrm{d}\omega).$$

This leads to the notion of the so-called L^p-norm

$$\|X\|_{L^p} := (\mathbb{E}|X|^p)^{1/p}$$

for $p \geq 1$, which turns the space of random variables equipped with $\|\cdot\|_{L^p}$ into a normed vector space.

A particular class of random variables, which finds widespread use throughout the theory of compressed sensing, are so-called *subgaussian* random variables whose L^p norm increases at most as $\sqrt{p}$. The name subgaussian is owed to the fact that subgaussian random variables have tail probabilities which decay at least as fast as the tails of the Gaussian distribution [**Ver18**, Section 2.5]. This leads to the following definition.

Definition A.1 (Subgaussian random variables)**.** *A random variable X is called* subgaussian *if it satisfies one of the following equivalent properties.*

(i) The tails of X satisfy

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/K_1^2) \quad \forall t \geq 0.$$

(ii) The absolute moments of X satisfy

$$\|X\|_{L^p} \leq K_2\sqrt{p} \quad \forall p \geq 1.$$

(iii) The super-exponential moment of X satisfies

$$\mathbb{E}\exp(X^2/K_3^2) \leq 2.$$

(iv) If $\mathbb{E}X = 0$, then the MGF of X satisfies

$$\mathbb{E}\exp(tX) \leq \exp(K_4^2 t^2) \quad \forall t \in \mathbb{R}.$$

The constants $K_1, \ldots, K_4$ are universal.

Note that the constants $K_i > 0$ for $i = 1, 2, 3, 4$ differ from each other at most by a constant factor, which in turn deviate only by a constant factor from the so-called *subgaussian norm* $\|\cdot\|_{\psi_2}$ introduced next.

Definition A.2 (Subgaussian norm)**.** *Given a random variable X, the subgaussian norm of X is defined as*

$$\|X\|_{\psi_2} := \inf\{s > 0 : \mathbb{E}\psi_2(X/s) \leq 1\},$$

where $\psi_2(t) := \exp(t^2) - 1$ is called an Orlicz function*.*

The set of subgaussian random variables defined on a common probability space equipped with the norm $\|\cdot\|_{\psi_2}$ forms a normed space known as *Orlicz space*. Note that some authors instead define the subgaussian norm as

$$\|X\|_{\psi_2} := \sup_{p \geq 1} \frac{1}{\sqrt{p}} \|X\|_{L^p}. \tag{A.1}$$

In light of Definition A.1, these definitions are—up to multiplicative constants—equivalent. As a consequence of (A.1) and Definition A.1(ii) above, a random variable is subgaussian if its subgaussian norm is finite. For instance, the subgaussian norm of a Gaussian random variable $g \sim \mathsf{N}(0, \sigma^2)$ is multiplicatively bounded above by σ times a constant: $\|g\|_{\psi_2} \lesssim \sigma$. On the other hand, the subgaussian norm of a Rademacher[1] random variable ϵ is given by $\|\epsilon\|_{\psi_2} = 1/\sqrt{\log(2)}$. Gaussian and Bernoulli random variables are therefore typical instances of subgaussian random variables. Other examples include random variables following the Steinhaus[2] distribution, as well as any bounded random variables in general, which includes all discrete distributions.

It is oftentimes convenient to extend the notion of subgaussianity from random variables to random vectors.

Definition A.3. *Let* $\mathbf{X}$ *be a random vector on* $\mathbb{R}^d$.

(i) The random vector $\mathbf{X}$ *is said to be subgaussian if the random variable* $\langle \mathbf{X}, \boldsymbol{\theta} \rangle$ *is subgaussian for all* $\boldsymbol{\theta} \in \mathbb{R}^d$.

(ii) The vector $\mathbf{X}$ *is called* isotropic *if* $\mathbb{E}|\langle \mathbf{X}, \boldsymbol{\theta} \rangle|^2 = \|\boldsymbol{\theta}\|_2^2$ *for all* $\boldsymbol{\theta} \in \mathbb{R}^d$.

Taking the supremum of the subgaussian norm of $\langle \mathbf{X}, \boldsymbol{\theta} \rangle$ in the previous definition over all unit directions $\boldsymbol{\theta}$ then leads to the definition of the subgaussian norm for random vectors.

Definition A.4 (Subgaussian vector norm)**.** *The subgaussian norm of a* d*-dimensional random vector* $\mathbf{X}$ *is*

$$\|\mathbf{X}\|_{\psi_2} := \sup_{\boldsymbol{\theta} \in \mathbb{S}^{d-1}} \|\langle \mathbf{X}, \boldsymbol{\theta} \rangle\|_{\psi_2}.$$

Next, we introduce an important result about the concentration behavior of Lipschitz continuous functions acting on Gaussian random vectors. In light of Definition A.1(iv), the following result establishes that random variables defined by such Lipschitz mappings are subgaussian.

Theorem A.5 ([BLM13, Theorem 5.5])**.** *Let* $\mathbf{g} \in \mathbb{R}^d$ *be a standard Gaussian random vector, and denote by* $f \colon \mathbb{R}^d \to \mathbb{R}$ *an* L*-Lipschitz function. Then, for all* $\theta \in \mathbb{R}$*, it holds that*

$$\mathbb{E} \exp\left(\theta(f(\mathbf{g}) - \mathbb{E} f(\mathbf{g}))\right) \leq \exp\left(\frac{\theta^2}{2} L^2\right).$$

[1] A Rademacher random variable takes on values in $\{\pm 1\}$ with equal probability.

[2] A Steinhaus random variable is a complex random variable distributed uniformly on the complex unit circle.

Remark A.6. *With* $X := f(\mathbf{g}) - \mathbb{E}f(\mathbf{g})$*, Markov's inequality* $t\mathbb{P}(X \geq t) \leq \mathbb{E}X$ *immediately yields*

$$\begin{aligned}\mathbb{P}(X \geq t) &= \mathbb{P}(\exp(\theta X) \geq \exp(\theta t)) \\ &\leq \frac{\mathbb{E}\exp(\theta X)}{\exp(\theta t)} \\ &\leq \exp\left(\frac{\theta^2}{2}L^2 - \theta t\right).\end{aligned}$$

Optimizing w. r. t. θ *therefore yields with* $\theta = t/L^2$ *that*

$$\mathbb{P}(f(\mathbf{g}) - \mathbb{E}f(\mathbf{g}) \geq t) \leq \exp\left(-\frac{t^2}{2L^2}\right)$$

for $t > 0$*. This result, which establishes that Lipschitz functions acting on Gaussian random vectors concentrate sharply around their mean, is known as the* concentration of measure inequality*.*

Lastly, we will sometimes require a bound on the maximum of subgaussian random variables. The following result shows that, in expectation, this maximum depends logarithmically on the number of variables.

Proposition A.7 ([FR13, Proposition 7.29])**.** *Let* $X_1, \dots X_n$ *be a sequence of zero-mean subgaussian random variables with* $\mathbb{E}\exp(\theta X_i) \leq \exp(c_i\theta^2)\ \forall \theta \in \mathbb{R}, i \in [n]$*. Then*

$$\mathbb{E}\max_{i\in[n]}|X_i| \leq \sqrt{4c\log(2n)}$$

for $c := \max_{i\in[n]} c_i$*.*

A.2 Convex Analysis and Geometric Functional Analysis

In this section, we collect a few standard definitions in convex analysis and geometric functional analysis. We begin with the definition of convex sets and convex functions.

Definition A.8 (Convex set)**.** *A set* $C \subset \mathbb{R}^d$ *is* called *convex if every line segment between two points* $\mathbf{x}, \mathbf{y} \in C$ *of the form* $\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$ *with* $\lambda \in [0,1]$ *is contained in* C*.*

Definition A.9 (Convex function)**.** *A function* $f\colon \mathbb{R}^d \to \mathbb{R}$ *is called* convex *if, for* $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(f) := \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) < \infty\}$ *and* $\lambda \in [0,1]$*,*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}).$$

A convex function is called closed *if its* epigraph

$$\operatorname{epi}(f) := \{(\mathbf{x}, t) \in \mathbb{R}^{d+1} : \mathbf{x} \in \operatorname{dom}(f), f(\mathbf{x}) \leq t\}$$

is a closed set.

Oftentimes, it will be convenient to "convexify" a nonconvex set $C \subset \mathbb{R}^d$. A natural way to do this is by considering the smallest convex set in $\mathbb{R}^d$ which contains C. This gives rise to the definition of the convex hull.

Definition A.10 (Convex hull)**.** *The* convex hull *of a set $C \subset \mathbb{R}^d$, denoted by* $\operatorname{conv}(C)$*, is the set of all convex combinations*

$$\operatorname{conv}(C) := \left\{ \sum_i \lambda_i \mathbf{x}_i : \mathbf{x}_i \in C, \lambda_i \geq 0, \sum_i \lambda_i = 1 \right\}.$$

Equipped with this concept, we will make frequent use of the following well-known result about the supremum of linear functions over compact sets.

Proposition A.11. *Let $C \subset \mathbb{R}^d$ be a compact set and $\mathbf{u} \in \mathbb{R}^d$. Then*

$$\sup_{\mathbf{x} \in C} \langle \mathbf{u}, \mathbf{x} \rangle = \sup_{\mathbf{x} \in \operatorname{conv}(C)} \langle \mathbf{u}, \mathbf{x} \rangle.$$

Proof. Denote by $\mathbf{x}^\star$ a vector in $\operatorname{conv}(C)$ which attains the supremum on the right-hand side. Since $\mathbf{x}^\star$ is a convex combination of elements in C, we have

$$\langle \mathbf{u}, \mathbf{x}^\star \rangle = \left\langle \mathbf{u}, \sum_{i=1}^k \lambda_i \mathbf{x}_i \right\rangle = \sum_{i=1}^k \lambda_i \langle \mathbf{u}, \mathbf{x}_i \rangle \leq \sum_{i=1}^k \lambda_i \langle \mathbf{u}, \mathbf{x}^\star \rangle = \langle \mathbf{u}, \mathbf{x}^\star \rangle$$

with $\mathbf{x}_i \in C$ and $\lambda_i \geq 0$, $\sum_{i=1}^k \lambda_i = 1$ for some $k \in \mathbb{N}$. For the above inequality to hold with equality, we therefore must have that every $\mathbf{x}_i \in C$ is optimal as claimed. □

Remark A.12. *By the same argument, Proposition A.11 also holds for the infimum of a linear function on a compact set.*

By the triangle inequality and homogeneity of norms, it immediately follows that the norm ball $\mathbb{B}^d_{\|\cdot\|} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1\}$ of an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$ is a convex set. The supremum of a linear function over a norm ball has a special meaning in convex analysis. This is the content of the following definition.

Definition A.13 (Dual norm)**.** *Let $\|\cdot\|$ be a norm. Then its dual norm $\|\cdot\|_*$ is defined as*

$$\|\mathbf{x}\|_* := \sup_{\|\mathbf{z}\| \leq 1} \langle \mathbf{z}, \mathbf{x} \rangle.$$

Example A.14. *For $p \in [1, \infty)$, the dual norm of the ℓ_p-norm on $\mathbb{R}^d$ is $\|\cdot\|_{p'}$ where $1/p + 1/p' = 1$. For $p = \infty$, one has the special pair $p = \infty$, $p' = 1$, i.e., $\|\cdot\|_1$ is the dual norm of $\|\cdot\|_\infty$. The tuple (p, p') is sometimes called a* conjugate pair*.*

Next, we introduce a geometric parameter which measures the complexity of a bounded set in a metric space. This complexity parameter is based on the idea of discretizing a set by approximating it with finitely many balls of a given radius whose union covers the original set.

Definition A.15 (Nets and covering numbers)**.** *Let (X, Δ) be a metric space, and let $U \subset X$ be bounded. Then a set $\mathcal{N} \subset U$ is called an* ε-net *or* ε-cover *of U w. r. t. Δ if for every $x \in X$ there exists at least one point $x_0 \in \mathcal{N}$ such that $\Delta(x, x_0) \leq \varepsilon$. The cardinality of the smallest such net, denoted by $\mathfrak{N}(U, \Delta, \varepsilon)$, is called the* covering number *of U. If $(X, \|\cdot\|)$ is a normed space, and Δ is the canonical metric induced by $\|\cdot\|$, we also write $\mathfrak{N}(U, \|\cdot\|, \varepsilon)$. Moreover, if $\|\cdot\| = \|\cdot\|_2$, we sometimes simply write $\mathfrak{N}(U, \varepsilon)$ for short.*

A concept closely related to the covering numbers are the so-called *packing numbers*.

Definition A.16 (Packing numbers)**.** *Let (X, Δ) be a metric space and $U \subset X$. The cardinality of the largest subset $\mathcal{P} \subset U$ such that all distinct points $\mathbf{x}, \mathbf{y} \in \mathcal{P}$ are ε-separated w. r. t. the metric Δ is called the* packing number *of U. It is denoted by $\mathfrak{P}(U, \Delta, \varepsilon)$.*

The following useful relation exists between covering and packing numbers.

Lemma A.17 ([Ver18, Lemma 4.2.8])**.** *For any subset U of a metric space (X, Δ), the following relation holds:*

$$\mathfrak{P}(U, \Delta, 2\varepsilon) \leq \mathfrak{N}(U, \Delta, \varepsilon) \leq \mathfrak{P}(U, \Delta, \varepsilon) \quad \forall \varepsilon > 0.$$

The following properties of covering numbers will prove useful throughout this thesis.

Proposition A.18. *The following properties hold for the covering number of a set $V \subset X$ in a metric space (X, Δ).*

(i) It holds that $\mathfrak{N}(V, \Delta, \varepsilon) = \mathfrak{N}(V/\alpha, \Delta, \varepsilon/\alpha)$ for any $\alpha > 0$.

(ii) For $\alpha > 0$, it holds that $\mathfrak{N}(\alpha V, \Delta, \varepsilon) = \mathfrak{N}(V, \Delta, \varepsilon/\alpha)$.

(iii) If Δ' is another metric on X with $\Delta(\mathbf{x}, \mathbf{y}) \leq \Delta'(\mathbf{x}, \mathbf{y})\ \forall \mathbf{x}, \mathbf{y} \in X$, then $\mathfrak{N}(V, \Delta, \varepsilon) \leq \mathfrak{N}(V, \Delta', \varepsilon)$.

(iv) If $U \subset V$, then $\mathfrak{N}(U, \Delta, \varepsilon) \leq \mathfrak{N}(V, \Delta, \varepsilon/2)$.

Proof. The first three properties are immediate.

To establish A.18(iv), we use Lemma A.17. To that end, first note that if $\mathcal{P} \subset U$ is a maximal ε-packing of U, then it is also an ε-packing of V. However, since V is a bigger set, the packing number can only increase, hence $\mathfrak{P}(U, \Delta, \varepsilon) \leq \mathfrak{P}(V, \Delta, \varepsilon)$. By Lemma A.17, we therefore have

$$\mathfrak{N}(U, \Delta, \varepsilon) \leq \mathfrak{P}(U, \Delta, \varepsilon) \leq \mathfrak{P}(V, \Delta, \varepsilon) \leq \mathfrak{N}(V, \Delta, \varepsilon/2),$$

which completes the proof. □

Calculating the covering number of a set explicitly is generally a difficult task. A classical bound on the covering number of norm balls w. r. t. their associated metrics is the following estimate based on the comparison of volumes.

Lemma A.19 (Volume comparison, [FR13, Proposition C.3])**.** *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ with $\mathbb{B}^d_{\|\cdot\|}$ denoting its associated unit ball. Denote by Δ the metric induced by $\|\cdot\|$. Then for $t > 0$,*

$$\mathfrak{N}\left(\mathbb{B}^d_{\|\cdot\|}, \Delta, t\right) \leq \left(1 + \frac{2}{t}\right)^d.$$

While the covering number is a purely geometric complexity parameter of a set, the following quantity represents a stochastic measure of complexity. Intuitively speaking, it is roughly equivalent to the average width of a set between two parallel supporting hyperplanes whose normals are drawn uniformly from the Haar measure on the unit sphere $\mathbb{S}^{d-1}$.

Definition A.20. *The* Gaussian mean width *(or simply* mean width *for short) of a set* $\mathcal{K} \subset \mathbb{R}^d$ *is defined as*

$$w(\mathcal{K}) := \mathbb{E}_{\mathbf{g}} \sup_{\mathbf{u} \in \mathcal{K}} \langle \mathbf{g}, \mathbf{u} \rangle$$

for $\mathbf{g} \sim \mathsf{N}(\mathbf{0}, \mathrm{Id})$.

The mean width of a set and its covering number are closely related by the following convenient result due to Sudakov. Considering that the mean width is oftentimes easily bounded, it provides us with an easy way to bound the covering number w. r. t. the Euclidean norm.

Lemma A.21 (Sudakov minoration, [Ver18, Theorem 8.1.13])**.** *Let* $U \subset \mathbb{R}^d$ *be a bounded set. Then*

$$\sup_{\varepsilon > 0} \varepsilon \sqrt{\log \mathfrak{N}(U, \|\cdot\|_2, \varepsilon)} \lesssim w(U).$$

B

Stable and Robust Recovery via Group-RIP Matrices

In general, necessary and sufficient conditions for stable and robust recovery depend on the null space property as discussed in Section 2.2. In its most basic form, the NSP of a matrix $\mathbf{A}$ ensures that the null space of $\mathbf{A}$ does not contain any sparse vectors of a certain order besides the zero vector, implying uniqueness of sparse vectors under the linear map defined by $\mathbf{A}$. In this section, we provide a similar condition for group-sparse recovery. The particular group-sparse NSP is a natural generalization of the block-sparse NSP originally introduced in [**GM17**]. Similar to the proof in the block-sparse case, the structure of our proof follows the example of the proof in the canonical sparsity setting presented in Chapter 4 and 6 of [**FR13**].

B.1 Robust Group-NSP

We first introduce the group null space property and show how it implies Theorem 5.2. We then show that the group-RIP implies the group null space property. We start by fixing some notation for the remainder of this appendix. Given a group partition $\mathcal{I} = \{\mathcal{I}_1, \dots, \mathcal{I}_G\}$

and a group index set $S \subset [G]$, we denote by $\mathcal{I}_S$ the subpartition $\{\mathcal{I}_i : i \in S\}$. Moreover, we denote by $\mathcal{I}_{\overline{S}}$ the partition consisting of the groups indexed by $\overline{S} = [G] \setminus S$. Finally, with slight abuse of notation, we write $\mathbf{x}_{\mathcal{I}_S}$ for the vector $\mathbf{x} \in \mathbb{C}^D$ restricted to the index set $\bigcup_{i \in S} \mathcal{I}_i$, *i.e.*, $\mathbf{x}_{\mathcal{I}_S} \triangleq \sum_{i \in S} \mathbf{x}_{\mathcal{I}_i}$.

Definition B.1 (ℓ_2-robust group-NSP)**.** *A matrix $\mathbf{A} \in \mathbb{C}^{M \times D}$ is said to satisfy the ℓ_2-robust* group null space property *(group-NSP) of order s w. r. t. an arbitrary norm $\|\cdot\|$ and constants $\rho \in (0,1)$ and $\tau > 0$ if for all $\mathbf{v} \in \mathbb{C}^D$ and for all $S \subset \{1, \dots, G\}$ with $|S| = s$,*

$$\|\mathbf{v}_{\mathcal{I}_S}\|_2 \leq \frac{\rho}{\sqrt{s}} \left\|\mathbf{v}_{\mathcal{I}_{\overline{S}}}\right\|_{\mathcal{I},1} + \tau \|\mathbf{A}\mathbf{v}\|.$$

Given a matrix which satisfies the ℓ_2-robust group-NSP, one may now establish the following result, which immediately implies Theorem 5.2.

Theorem B.2. *Suppose that the matrix $\mathbf{A} \in \mathbb{C}^{M \times D}$ satisfies the ℓ_2-robust group null space property of order s w. r. t. $\|\cdot\|$, and constants $\rho \in (0,1)$ and $\tau > 0$. Then for any $\mathbf{x}, \mathbf{z} \in \mathbb{C}^D$,*

$$\|\mathbf{z} - \mathbf{x}\|_2 \leq \frac{C_0}{\sqrt{s}} \Big(\|\mathbf{z}\|_{\mathcal{I},1} - \|\mathbf{x}\|_{\mathcal{I},1} + 2\sigma_s(\mathbf{x})_{\mathcal{I},1} \Big) + C_1 \|\mathbf{A}(\mathbf{z} - \mathbf{x})\|$$

where

$$C_0 = \frac{(1+\rho)^2}{1-\rho} \quad \text{and} \quad C_1 = \frac{(3+\rho)\tau}{1-\rho}.$$

Proof. The ℓ_2-robust group-NSP directly implies that for any $\mathbf{x}, \mathbf{z} \in \mathbb{C}^D$ and $\mathbf{v} := \mathbf{z} - \mathbf{x}$, we have

$$\begin{aligned} \|\mathbf{v}\|_2 &\leq \|\mathbf{v}_{\mathcal{I}_S}\|_2 + \left\|\mathbf{v}_{\mathcal{I}_{\overline{S}}}\right\|_2 \\ &\leq \frac{\rho}{\sqrt{s}} \left\|\mathbf{v}_{\mathcal{I}_{\overline{S}}}\right\|_{\mathcal{I},1} + \tau \|\mathbf{A}\mathbf{v}\| + \left\|\mathbf{v}_{\mathcal{I}_{\overline{S}}}\right\|_2 \end{aligned} \tag{B.1}$$

for an arbitrary index set $S \subset [G]$ with $|S| \leq s$. We first provide a bound for $\|\mathbf{v}_{\mathcal{I}_{\overline{S}}}\|_2$ in terms of $\|\cdot\|_{\mathcal{I},1}$. Denote by $\{\check{\mathcal{I}}_1, \dots, \check{\mathcal{I}}_G\}$ the nonincreasing group rearrangement of $\mathcal{I}$ such that

$$\left\|\mathbf{v}_{\check{\mathcal{I}}_1}\right\|_2 \geq \left\|\mathbf{v}_{\check{\mathcal{I}}_2}\right\|_2 \geq \cdots \geq \left\|\mathbf{v}_{\check{\mathcal{I}}_G}\right\|_2.$$

We now choose S as the index set of the best s-term group approximation of $\mathbf{v}$, which implies that

$$\begin{aligned} \left\|\mathbf{v}_{\mathcal{I}_{\overline{S}}}\right\|_2^2 &\leq \sum_{j=s+1}^{G} \left\|\mathbf{v}_{\check{\mathcal{I}}_j}\right\|_2^2 \\ &\leq \left(\frac{1}{s} \sum_{j=1}^{s} \left\|\mathbf{v}_{\check{\mathcal{I}}_j}\right\|_2 \right) \left(\sum_{j=s+1}^{G} \left\|\mathbf{v}_{\check{\mathcal{I}}_j}\right\|_2 \right) \\ &\leq \frac{1}{s} \|\mathbf{v}\|_{\mathcal{I},1}^2. \end{aligned}$$

Applying this inequality in (B.1) therefore yields

$$\|\mathbf{v}\|_2 \leq \frac{1+\rho}{\sqrt{s}}\|\mathbf{v}\|_{\mathcal{I},1} + \tau\|\mathbf{A}\mathbf{v}\|. \tag{B.2}$$

Next we bound the term $\|\mathbf{v}\|_{\mathcal{I},1}$. First note that if the ℓ_2-robust group-NSP holds, the Cauchy-Schwarz inequality implies the following bound on the group ℓ_1-norm:

$$\|\mathbf{v}_{\mathcal{I}_S}\|_{\mathcal{I},1} \leq \sqrt{s}\|\mathbf{v}_{\mathcal{I}_S}\|_2 \leq \rho\left\|\mathbf{v}_{\mathcal{I}_{\overline{S}}}\right\|_{\mathcal{I},1} + \tau\sqrt{s}\|\mathbf{A}\mathbf{v}\|. \tag{B.3}$$

This immediately implies that

$$\begin{aligned}\|\mathbf{v}\|_{\mathcal{I},1} &= \|\mathbf{v}_{\mathcal{I}_S}\|_{\mathcal{I},1} + \left\|\mathbf{v}_{\mathcal{I}_{\overline{S}}}\right\|_{\mathcal{I},1} \\ &\leq (1+\rho)\left\|\mathbf{v}_{\mathcal{I}_{\overline{S}}}\right\|_{\mathcal{I},1} + \tau\sqrt{s}\|\mathbf{A}\mathbf{v}\|. \end{aligned} \tag{B.4}$$

In order to bound $\|\mathbf{v}_{\mathcal{I}_{\overline{S}}}\|_{\mathcal{I},1}$, we apply (B.3) once again in combination with the following result, which is easily adopted to the group-sparse setting from [**FR13**, Lemma 4.15].

Lemma B.3. *Consider a group partition $\mathcal{I} = \{\mathcal{I}_1, \dots, \mathcal{I}_G\}$. Then for $T \subset [G]$ and any two vectors $\mathbf{x}, \mathbf{z} \in \mathbb{C}^D$, we have for $\mathbf{v} := \mathbf{z} - \mathbf{x}$ that*

$$\left\|\mathbf{v}_{\mathcal{I}_{\overline{T}}}\right\|_{\mathcal{I},1} \leq \|\mathbf{z}\|_{\mathcal{I},1} - \|\mathbf{x}\|_{\mathcal{I},1} + \|\mathbf{v}_{\mathcal{I}_T}\|_{\mathcal{I},1} + 2\left\|\mathbf{x}_{\mathcal{I}_{\overline{T}}}\right\|_{\mathcal{I},1}.$$

Invoking (B.3) in Lemma B.3 with $T = S$, we obtain

$$\begin{aligned}\left\|\mathbf{v}_{\mathcal{I}_{\overline{S}}}\right\|_{\mathcal{I},1} &\leq \|\mathbf{z}\|_{\mathcal{I},1} - \|\mathbf{x}\|_{\mathcal{I},1} + \|\mathbf{v}_{\mathcal{I}_S}\|_{\mathcal{I},1} + 2\left\|\mathbf{x}_{\mathcal{I}_{\overline{S}}}\right\|_{\mathcal{I},1} \\ &\leq \|\mathbf{z}\|_{\mathcal{I},1} - \|\mathbf{x}\|_{\mathcal{I},1} + \rho\left\|\mathbf{v}_{\mathcal{I}_{\overline{S}}}\right\|_{\mathcal{I},1} + \tau\sqrt{s}\|\mathbf{A}\mathbf{v}\| + 2\left\|\mathbf{x}_{\mathcal{I}_{\overline{S}}}\right\|_{\mathcal{I},1},\end{aligned}$$

which implies that

$$\left\|\mathbf{v}_{\mathcal{I}_{\overline{S}}}\right\|_{\mathcal{I},1} \leq \frac{1}{1-\rho}\left(\|\mathbf{z}\|_{\mathcal{I},1} - \|\mathbf{x}\|_{\mathcal{I},1} + 2\left\|\mathbf{x}_{\mathcal{I}_{\overline{S}}}\right\|_{\mathcal{I},1}\right) + \frac{\tau\sqrt{s}}{1-\rho}\|\mathbf{A}\mathbf{v}\|,$$

and consequently from (B.4) that

$$\|\mathbf{v}\|_{\mathcal{I},1} \leq \frac{1+\rho}{1-\rho}\left(\|\mathbf{z}\|_{\mathcal{I},1} - \|\mathbf{x}\|_{\mathcal{I},1} + 2\left\|\mathbf{x}_{\mathcal{I}_{\overline{S}}}\right\|_{\mathcal{I},1}\right) + \frac{2\tau\sqrt{s}}{1-\rho}\|\mathbf{A}\mathbf{v}\|.$$

Since the group support S corresponds to the s groups with largest ℓ_2-norm, it minimizes the term $\|\mathbf{x}_{\mathcal{I}_{\overline{S}}}\|_{\mathcal{I},1}$ on the right-hand side and therefore $\|\mathbf{x}_{\mathcal{I}_{\overline{S}}}\|_{\mathcal{I},1} = \sigma_s(\mathbf{x})_{\mathcal{I},1}$. Combined with (B.2), the claim follows. □

Since the result above holds for any $\mathbf{x}$ and $\mathbf{z}$, choosing $\mathbf{x} = \mathring{\mathbf{x}}$ and $\mathbf{z} = \mathbf{x}^\star$ with $\mathbf{x}^\star$ denoting a minimizer of Problem ($\mathrm{P}_{\mathcal{I},1}$) immediately implies the following theorem. More precisely, by optimality of $\mathbf{x}^\star$ for Problem ($\mathrm{P}_{\mathcal{I},1}$), we have $\|\mathbf{x}^\star\|_{\mathcal{I},1} - \|\mathring{\mathbf{x}}\|_{\mathcal{I},1} \leq 0$ and $\|\mathbf{A}(\mathring{\mathbf{x}} - \mathbf{x}^\star)\|_2 = \|\mathbf{y} - \mathbf{A}\mathbf{x}^\star\|_2 \leq \nu$ by feasibility of $\mathbf{x}^\star$.

Theorem B.4. *Suppose that $\mathbf{A} \in \mathbb{C}^{M\times D}$ satisfies the ℓ_2-robust group-NSP of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then for all $\mathring{\mathbf{x}} \in \mathbb{C}^D$ and $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} + \mathbf{e}$ with $\|\mathbf{e}\|_2 \leq \nu$, any solution $\mathbf{x}^\star$ of Problem $(\mathrm{P}_{\mathcal{I},1})$ approximates $\mathring{\mathbf{x}}$ with error*

$$\|\mathring{\mathbf{x}} - \mathbf{x}^\star\|_2 \leq 2C_0 \frac{\sigma_s(\mathring{\mathbf{x}})_{\mathcal{I},1}}{\sqrt{s}} + C_1\nu$$

with $C_0, C_1 > 0$ as in Theorem B.2.

The ℓ_2-robust group-NSP provides a necessary and sufficient condition for recovery of group-sparse vectors. In the next section, we establish that the group-RIP implies the ℓ_2-robust group-NSP and therefore yields a sufficient condition for stable and robust recovery of group-sparse vectors.

B.2 The Group-RIP Implies the ℓ_2-Robust Group-NSP

In light of the previous section, it suffices to prove that the group-RIP of order $2s$ with constant δ implies the ℓ_2-robust group-NSP to establish Theorem 5.2 according to the following result.

Proposition B.5. *Assume that the matrix $\mathbf{A} \in \mathbb{C}^{M\times D}$ satisfies the group-RIP of order s with constant $\delta < 4/\sqrt{41}$. Then $\mathbf{A}$ satisfies the ℓ_2-robust group-NSP with constants*

$$\rho = \frac{\delta}{\sqrt{1-\delta^2} - \delta/4} \quad \text{and} \quad \tau = \frac{\sqrt{1+\delta}}{\sqrt{1-\delta^2} - \delta/4}. \tag{B.5}$$

We follow the general proof strategy in [**FR13**, Chapter 6], which employs a common splitting technique prominently used throughout the compressed sensing literature.

Proof of Proposition B.5. Consider the group index sets $S_0, S_1, \ldots$ such that $S_i \subset [G]$ corresponds to the group indices of the s groups with largest ℓ_2-norms in $\overline{\bigcup_{j<i} S_j}$. If the group-NSP is established for S_0, yielding the largest possible $\|\mathbf{v}_{\mathcal{I}_{S_0}}\|_2$, then it also holds for any other S_i with $i > 0$. Assuming the group-RIP holds, we can write $\|\mathbf{A}\mathbf{v}_{\mathcal{I}_{S_0}}\|_2^2 = (1+t)\|\mathbf{v}_{\mathcal{I}_{S_0}}\|_2^2$ with $|t| \leq \delta$. We can therefore bound the term $\|\mathbf{A}\mathbf{v}_{\mathcal{I}_{S_0}}\|_2^2$ by

$$\begin{aligned}
\left\|\mathbf{A}\mathbf{v}_{\mathcal{I}_{S_0}}\right\|_2^2 &= \left\langle \mathbf{A}\mathbf{v}_{\mathcal{I}_{S_0}}, \mathbf{A}\left(\mathbf{v} - \sum_{k\geq 1} \mathbf{v}_{\mathcal{I}_{S_k}}\right)\right\rangle_{\mathbb{C}} \\
&= \left\langle \mathbf{A}\mathbf{v}_{\mathcal{I}_{S_0}}, \mathbf{A}\mathbf{v}\right\rangle_{\mathbb{C}} - \sum_{k\geq 1}\left\langle \mathbf{A}\mathbf{v}_{\mathcal{I}_{S_0}}, \mathbf{A}\mathbf{v}_{\mathcal{I}_{S_k}}\right\rangle_{\mathbb{C}} \\
&\leq \left\|\mathbf{A}\mathbf{v}_{\mathcal{I}_{S_0}}\right\|_2 \|\mathbf{A}\mathbf{v}\|_2 + C_t \sum_{k\geq 1} \left\|\mathbf{v}_{\mathcal{I}_{S_0}}\right\|_2 \left\|\mathbf{v}_{\mathcal{I}_{S_k}}\right\|_2
\end{aligned}$$

where the last inequality follows from Lemma B.6 given at the end of this section with $C_t := \sqrt{\delta^2 - t^2}$. Using $\|\mathbf{A}\mathbf{v}_{\mathcal{I}_{S_0}}\|_2 = \sqrt{1+t}\|\mathbf{v}_{\mathcal{I}_{S_0}}\|_2$, we arrive at an expression similar to the group-NSP, namely

$$(1+t)\left\|\mathbf{v}_{\mathcal{I}_{S_0}}\right\|_2 \leq C_t \sum_{k\geq 1} \left\|\mathbf{v}_{\mathcal{I}_{S_k}}\right\|_2 + \sqrt{1+t}\|\mathbf{A}\mathbf{v}\|_2. \tag{B.6}$$

Although we trivially have

$$\sum_{k\geq 1}\big\|\mathbf{v}_{\mathcal{I}_{S_k}}\big\|_2 = \big\|\mathbf{v}_{\mathcal{I}_{\overline{S_0}}}\big\|_{\mathcal{I},1},$$

we require an additional factor of $1/\sqrt{s}$ to obtain the ℓ_2-robust group-NSP according to Definition B.1. To that end, we invoke [**FR13**, Lemma 6.14] and find

$$\sum_{k\geq 1}\big\|\mathbf{v}_{\mathcal{I}_{S_k}}\big\|_2 \leq \frac{1}{\sqrt{s}}\big\|\mathbf{v}_{\mathcal{I}_{\overline{S_0}}}\big\|_{\mathcal{I},1} + \frac{1}{4}\big\|\mathbf{v}_{\mathcal{I}_{S_0}}\big\|_2.$$

Applying this inequality in (B.6) now yields

$$\begin{aligned}\big\|\mathbf{v}_{\mathcal{I}_{S_0}}\big\|_2 &\leq \frac{1}{\sqrt{s}}\frac{\sqrt{\delta^2-t^2}}{1+t-\frac{\sqrt{\delta^2-t^2}}{4}}\big\|\mathbf{v}_{\mathcal{I}_{\overline{S_0}}}\big\|_{\mathcal{I},1} + \frac{\sqrt{1+t}}{1+t-\frac{\sqrt{\delta^2-t^2}}{4}}\|\mathbf{A}\mathbf{v}\|_2\\ &\leq \frac{1}{\sqrt{s}}\frac{\frac{\sqrt{\delta^2-t^2}}{1+t}}{1-\frac{1}{4}\frac{\sqrt{\delta^2-t^2}}{1+t}}\big\|\mathbf{v}_{\mathcal{I}_{\overline{S_0}}}\big\|_{\mathcal{I},1} + \frac{1}{\sqrt{1+t}}\frac{1}{1-\frac{1}{4}\frac{\sqrt{\delta^2-t^2}}{1+t}}\|\mathbf{A}\mathbf{v}\|_2.\end{aligned}$$

Next, note that since $|t|\leq\delta$, one has $1/\sqrt{1+t}\leq 1/\sqrt{1-\delta}$. Moreover, it holds that $\sqrt{\delta^2-t^2}/(1+t)\leq\delta/\sqrt{1-\delta^2}$. Invoking these estimates, we therefore find

$$\big\|\mathbf{v}_{\mathcal{I}_{S_0}}\big\|_2 \leq \frac{1}{\sqrt{s}}\frac{\delta}{\sqrt{1-\delta^2}-\delta/4}\big\|\mathbf{v}_{\mathcal{I}_{\overline{S_0}}}\big\|_{\mathcal{I},1} + \frac{\sqrt{1+\delta}}{\sqrt{1-\delta^2}-\delta/4}\|\mathbf{A}\mathbf{v}\|_2.$$

This means that $\mathbf{A}$ satisfies the ℓ_2-robust group-NSP with constants ρ and τ as in (B.5). Since we require $\rho<1$, this in turn implies $\delta<4/\sqrt{41}$ as claimed. □

It remains to establish the following lemma used in the proof of Proposition B.5, which we extract from the proof of Theorem 6.13 in [**FR13**].

Lemma B.6. *Suppose that the matrix $\mathbf{A}\in\mathbb{C}^{M\times D}$ satisfies the group-RIP of order $2s$ with constant $\delta\in(0,1)$. Then for any two disjoint sets $S_0,S_1\subset[G]$ with cardinality s, it holds for $|t|\leq\delta$ that*

$$\big|\big\langle\mathbf{A}\mathbf{v}_{\mathcal{I}_{S_0}},\mathbf{A}\mathbf{v}_{\mathcal{I}_{S_1}}\big\rangle_{\mathbb{C}}\big| \leq \sqrt{\delta^2-t^2}\big\|\mathbf{v}_{\mathcal{I}_{S_0}}\big\|_2\big\|\mathbf{v}_{\mathcal{I}_{S_1}}\big\|_2.$$

Proof. To start with, we normalize the two vectors to have unit ℓ_2-norm by defining the auxiliary vectors $\mathbf{u}:=\mathbf{v}_{\mathcal{I}_{S_0}}/\|\mathbf{v}_{\mathcal{I}_{S_0}}\|_2$ and $\mathbf{w}:=\theta\mathbf{v}_{\mathcal{I}_{S_1}}/\|\mathbf{v}_{\mathcal{I}_{S_1}}\|_2$ where $\theta\in\mathbb{C}$ with $|\theta|=1$ is chosen such that $\mathfrak{Re}\langle\mathbf{A}\mathbf{u},\mathbf{w}\rangle_{\mathbb{C}}=|\langle\mathbf{A}\mathbf{u},\mathbf{w}\rangle_{\mathbb{C}}|$. Denote further by $\alpha,\beta>0$ two parameters to be chosen later. Then

$$\begin{aligned}2|\langle\mathbf{A}\mathbf{u},\mathbf{A}\mathbf{w}\rangle_{\mathbb{C}}| &= \frac{1}{\alpha+\beta}\Big(\|\mathbf{A}(\alpha\mathbf{u}+\mathbf{w})\|_2^2 - \|\mathbf{A}(\beta\mathbf{u}-\mathbf{w})\|_2^2 - (\alpha^2-\beta^2)\|\mathbf{A}\mathbf{u}\|_2^2\Big)\\ &\leq \frac{1}{\alpha+\beta}\Big[(1+\delta)\|\alpha\mathbf{u}+\mathbf{w}\|_2^2 - (1-\delta)\|\beta\mathbf{u}-\mathbf{w}\|_2^2 - (\alpha^2-\beta^2)(1+t)\|\mathbf{u}\|_2^2\Big]\\ &\leq \frac{1}{\alpha+\beta}\Big[(1+\delta)(\alpha^2+1)^2 - (1-\delta)(\beta^2+1)^2 - (\alpha^2-\beta^2)(1+t)\Big]\\ &\leq \frac{1}{\alpha+\beta}\Big[\alpha^2(\delta-1)+\beta^2(\delta+1)+2\delta\Big].\end{aligned}$$

Choosing $\alpha=(\delta+t)/\sqrt{\delta^2-t^2}$ and $\beta=(\delta-t)/\sqrt{\delta^2-t^2}$ completes the proof. □

List of Abbreviations

ADC	analog-to-digital converter	2, 20
AOA	angle of arrival	21
AOP	adaptive outlier pursuit	27
ARIP	asymmetric restricted isometry property	123
BCS	block-based compressed sensing	117
BFCS	binary fused compressive sensing	49
BIHT	binary iterative hard thresholding	iii, 3, 17, 21, 26
block-RIP	block restricted isometry property	119
BOS	bounded orthonormal system	11
BP	basis pursuit	9
CoSaMP	compressive sampling matching pursuit	5, 30
CS	compressed sensing	iii, 1, 7, 8
CS-AOP	conjugate symmetric adaptive outlier pursuit	39, 40
CS-BIHT	conjugate symmetric binary iterative hard thresholding	30
DCS	distributed compressed sensing	20, 116
DCT	discrete cosine transform	11
DFT	discrete Fourier transform	4, 5, 11
DOA	direction of arrival	20
FDMA	frequency-division multiple access	47
FFT	fast Fourier transform	32
fMRI	functional magnetic resonance imaging	47
group-NSP	group null space property	169, 170
group-RIC	group restricted isometry constant	53
group-RIP	group restricted isometry property	6, 53
HRA	high-resolution assumption	13
HTP	hard thresholding pursuit	5, 30
IHT	iterative hard thresholding	5, 24
IoT	internet of things	2, 156

LASSO	least-absolute shrinkage selection operator	23, 113
MC	Monte Carlo	33
MGF	moment generating function	162
MMV	multiple measurement vector	116
MRI	magnetic resonance imaging	2, 19, 46, 47, 175
MSP	matched sign pursuit	17
NSP	null space property	9
QCBP	quadratically-constrained basis pursuit	14
QCS	quantized compressed sensing	3, 12
QIHT	quantized iterative hard thresholding	30
RFPI	renormalized fixed-point iteration	17
RIC	restricted isometry constant	178
RIP	restricted isometry property	4, 10
RoBFCS	robust binary fused compressive sensing	50
RSS	restricted-step shrinkage	17
SNR	signal-to-noise ratio	16
SOCP	second-order cone program	82
STFT	short-time Fourier transform	20
STrMP	sign-truncated matching pursuit	17
TV	total variation	49
WSN	wireless sensor network	116

List of Symbols

$\subset$	Proper subset, *i.e.*, $A \subset B$ if and only if $A \neq \emptyset$, $A \neq B$ and $a \in A \implies a \in B$
$\subseteq$	Regular subset
$\mathbb{N}$	The set $\{1, 2, \ldots\}$ of natural numbers
$[d]$	The set of integers from 1 to $d \in \mathbb{N}$, *i.e.*, $[d] = [1, d] \cap \mathbb{N} = \{1, \ldots, d\}$
$\mathbb{R}$	The set of real numbers
$\mathbb{C}$	The set of complex numbers
$\mathbb{K}$	Either the field $\mathbb{R}$ or $\mathbb{C}$
$\mathbb{K}^d$	The vector space $\mathbb{K}^d$ over $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$ of dimension d
$\mathbb{K}^d_I$	The coordinate subspace of $\mathbb{K}^d$ restricted to coordinates indexed by $I \subseteq [d]$
$\mathbb{X}_d$	The subspace of conjugate symmetric vectors of $\mathbb{C}^d$
$\mathrm{O}(d)$	The orthogonal group
$\mathrm{U}(d)$	The unitary group
$\lvert S \rvert$	The cardinality of a discrete set S
$\overline{S}$	The complement of a subset, *e.g.*, $\overline{S} = [d] \setminus S$ for $S \subset [d]$
Id_d	The identity matrix of size $d \times d$
$\mathbf{F}_d$	The orthogonal discrete Fourier transform matrix of size $d \times d$
$\mathbf{\Psi}$	A change of basis matrix associated with a sparsity basis
$\mathbf{1}$	The all ones vector
$\mathbf{0}$	The zero vector
$\mathbf{y}$	Usually the measurement vector
$\mathbf{A}$	A measurement matrix of size $m \times d$ (Chapter 3 and 4) or a block diagonal measurement matrix of size $M \times D$ (Chapter 5)
$\mathbf{A}^\top$	The transpose of a matrix
$\mathbf{A}^*$	The adjoint operator of the linear map induced by a matrix
$\ker(\mathbf{A})$	The null space of a matrix
$\overline{x}$	The complex conjugate of a complex number $x \in \mathbb{C}$ (applies element-wise to vectors and matrices)
$\mathcal{F}$	The Fourier transform operator
f_b	The Nyquist frequency, *i.e.*, the highest frequency of a band-limited signal
f_r	The sampling rate of an ADC
sgn	The sign function of a real number (acts element-wise on vectors)
$\mathfrak{Re}, \mathfrak{Im}$	The real and imaginary parts of a matrix, vector or scalar

$\check{\mathbf{x}}$	The nonincreasing rearrangement of a vector
$\mathbf{a} \circ \mathbf{b}$	The element-wise product (Hadamard product) of two vectors
$\|\cdot\|_p$	The ℓ_p-norm
$\|\cdot\|_0$	The ℓ_0-pseudonorm, *i.e.*, the number of nonzero entries of a vector
$\|\cdot\|_{p\to q}$	The operator norm between the normed spaces $(\mathbb{K}^d, \|\cdot\|_p)$ and $(\mathbb{K}^m, \|\cdot\|_q)$
$\mathbb{B}_p^d$	The unit ball of $\|\cdot\|_p$ in $\mathbb{K}^d$
$\mathbb{S}^{d-1}$	The unit Euclidean sphere in $\mathbb{K}^d$, *i.e.*, the boundary of $\mathbb{B}_2^d$
$\mathbb{S}_I^{d-1}$	The unit sphere in the coordinate subspace $\mathbb{K}_I^d$
$\operatorname{supp}(\mathbf{x})$	The support of a vector $\mathbf{x}$
$\langle \cdot, \cdot \rangle$	The canonical inner product on $\mathbb{R}^d$
$\langle \cdot, \cdot \rangle_{\mathbb{C}}$	The canonical inner product on $\mathbb{C}^d$
Δ_{H}	The normalized Hamming distance
Δ_{supp}	The support error, *i.e.*, the cardinality of the symmetric set difference between the support sets of two vectors
$\mathring{\mathbf{x}}$	A sparse or group-sparse target vector to be recovered
$\mathbf{x}^\star$	The optimal solution of an optimization problem
$\hat{\mathbf{x}}$	Usually the vector produced by a reconstruction map
s	The sparsity or group-sparsity level, *i.e.*, the number of nonzero entries or groups of a vector
L	Usually the number of sensors in Chapter 5
δ_s	The restricted isometry constant associated with a particular type of restricted isometry property
G	The number of groups
$\mathcal{I}$	A group partition, *i.e.*, a set of index sets partitioning a set into G nonoverlapping groups
$\widetilde{\mathcal{I}}$	Usually the trivial partition $\widetilde{\mathcal{I}} = \{\{1\}, \dots, \{d\}\}$ of $[d]$
g	The size of the largest group in a group partition $\mathcal{I}$
$\|\cdot\|_{\mathcal{I},p}$	The group ℓ_p-norm
$\Sigma_s(\mathcal{V})$	The set of s-sparse vectors in a vector space $\mathcal{V}$
$\widetilde{\Sigma}_s$	The sparse vectors with unit Euclidean norm
$\Sigma_{\mathcal{I},s}$	The set of s-group-sparse vectors w. r. t. $\mathcal{I}$: $\Sigma_{\mathcal{I},s} = \Sigma_{\mathcal{I},s}(\mathbb{R}^d)$ (Chapter 4) or $\Sigma_{\mathcal{I},s} = \Sigma_{\mathcal{I},s}(\mathbb{C}^D)$ (Chapter 5)
$\widetilde{\Sigma}_{\mathcal{I},s}$	The set of group-sparse vectors on the unit sphere
$\mathcal{E}_{\mathcal{I},s}$	The set of effectively group-sparse vectors w. r. t. $\mathcal{I}$
$\widetilde{\mathcal{E}}_{\mathcal{I},s}$	The set of effectively group-sparse vectors on the unit sphere
$\sigma_s(\cdot)_p$	The best s-term approximation error w. r. t. $\|\cdot\|_p$
$\sigma_s^{\mathbb{X}}(\cdot)_p$	The best conjugate symmetric s-term approximation error w. r. t. $\|\cdot\|_p$
$\sigma_{\mathcal{I},s}(\cdot)_p$	The best s-group approximation error w. r. t. $\mathcal{I}$ and $\|\cdot\|_p$
$\mathcal{H}_s$	The hard thresholding operator, *i.e.*, the projection operator on Σ_s
$\mathcal{H}_{s,p}^{\mathbb{X}}$	The conjugate symmetric hard thresholding operator, *i.e.*, the projector on $\Sigma_s(\mathbb{X}_d)$ w. r. t. $\|\cdot\|_p$
$\mathcal{H}_{\mathcal{I},s}$	The group sparse hard thresholding operator, *i.e.*, the projection operator on $\Sigma_{\mathcal{I},s}$
$\mathbb{1}_E$	The indicator function of an event E
$\mathbb{P}$	The probability measure of a probability space
$\mathbb{E}X$	The expectation of a random variable X

$\mathsf{N}(\boldsymbol{\mu}, \mathbf{C})$	The multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$
$\mathsf{U}(I)$	The uniform distribution over an interval I
$\mathsf{B}_m(p)$	The m-dimensional Bernoulli distribution with independent entries
$\mathbf{g}$	Usually a standard Gaussian random vector
$\operatorname{conv}(\mathcal{K})$	The convex hull of a set $\mathcal{K}$
γ_U	The Minkowski gauge function of a set U
$\mathcal{N}_\varepsilon$	An ε-net of a set
$\mathfrak{N}$	The covering number
$\mathfrak{P}$	The packing number
w	The Gaussian mean width
η	Usually a failure probability
ε	Usually the reconstruction error of a recovery scheme
θ^2	Usually the variance of a random quantization dither
σ^2	Usually the variance of an additive noise term
$\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{PV}}$	The group-sparse direction recovery map associated with Problem ($\mathrm{P}_{4.2}$)
$\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{corr}}$	The group-sparse direction recovery map associated with Problem ($\mathrm{P}_{4.5}$)
$\widetilde{\Delta}_{\mathcal{I}}^{\mathrm{ht}}$	The group-sparse hard thresholding direction reconstruction map (cf. Equation (4.18))
$\Delta_{\mathcal{I}}^{\mathrm{nc}}$	The norm-constrained group-sparse reconstruction map associated with Problem ($\mathrm{P}_{4.8}$)
$\Delta_{\mathcal{I}}^{\mathrm{PV}}$	The group-sparse reconstruction map associated with Problem ($\mathrm{P}_{4.9}$) (cf. Equation (4.26))
$\Delta_{\mathcal{I}}^{\Pi}$	The group-sparse recovery map associated with Problem ($\mathrm{P}_{4.13}$)
$\Delta_{\mathcal{I}}^{\mathrm{corr}}$	The projection-based group-sparse reconstruction map associated with Problem ($\mathrm{P}_{4.10}$) (cf. Equation (4.27))
$\Delta_{\mathcal{I}}^{\mathrm{ht}}$	The group-sparse hard thresholding recovery map (cf. Equation (4.31))
$\Delta_{\mathcal{I}}^{\Pi\text{-ht}}$	The projection-based group-sparse hard thresholding recovery map (cf. Equation (4.35))

References

[**ABK17**] J. Acharya, A. Bhattacharyya, and P. Kamath. "Improved bounds for universal one-bit compressive sensing". In: *IEEE International Symposium on Information Theory (ISIT)*. June 2017, pp. 2353–2357.

[**Ada+11**] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. "Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling". In: *Constructive Approximation* 34.1 (2011), pp. 61–88.

[**Adl+16**] A. Adler, D. Boublil, M. Elad, and M. Zibulevsky. "A deep learning approach to block-based compressed sensing of images". In: (2016). arXiv: 1606.01519.

[**ADR16**] U. Ayaz, S. Dirksen, and H. Rauhut. "Uniform recovery of fusion frame structured sparse signals". In: *Applied and Computational Harmonic Analysis* 2 (2016), pp. 341–361.

[**AFN12**] A. Argyriou, R. Foygel, and S. Nathan. "Sparse Prediction with the k-Support Norm". In: *Advances in Neural Information Processing Systems (NIPS)*. 2012, pp. 1457–1465.

[**Aky+02**] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. "Wireless sensor networks: A survey". In: *Computer Networks* 38.4 (2002), pp. 393–422.

[**AL12**] D. D. Ariananda and G. Leus. "Compressive wideband power spectrum estimation". In: *IEEE Transactions on Signal Processing* 60.9 (2012), pp. 4775–4789.

[**Ame+14**] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. "Living on the edge: Phase transitions in convex programs with random data". en. In: *Information and Inference* 3.3 (Sept. 2014), pp. 224–294. ISSN: 2049-8764, 2049-8772.

[**Asi+10**] M. S. Asif, D. Reddy, P. T. Boufounos, and A. Veeraraghavan. "Streaming compressive sensing for high-speed periodic videos". In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2010, pp. 3373–3376.

[**AV17**] M. E. Ahsen and M. Vidyasagar. "Error bounds for compressed sensing algorithms with group sparsity: A unified approach". In: *Applied and Computational Harmonic Analysis* 43.2 (2017), pp. 212–232.

[**Aya18**] U. Ayaz. *Sparse recovery of fusion frame structured signals*. 2018. arXiv: 1804.02079.

[**BA10a**] P. T. Boufounos and M. S. Asif. "Compressive sampling for streaming signals with sparse frequency content". In: *Conference on Information Sciences and Systems (CISS)*. IEEE. 2010, pp. 1–6.

[**BA10b**] P. T. Boufounos and M. S. Asif. "Compressive sensing for streaming signals using the streaming greedy pursuit". In: *Military Communications Conference (MILCOM)*. IEEE. 2010, pp. 1205–1210.

[**Baj+10**] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak. "Compressed channel sensing: A new approach to estimating sparse multipath channels". In: *Proceedings of the IEEE* 98.6 (2010), pp. 1058–1076.

[**Bal+16**] L. Baldassarre, N. Bhan, V. Cevher, A. Kyrillidis, and S. Satpathi. "Group-sparse model selection: Hardness and relaxations". In: *IEEE Transactions on Information Theory* 62.11 (2016), pp. 6508–6534.

[**Bar+08**] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. "A simple proof of the restricted isometry property for random matrices". In: *Constructive Approximation* 28.3 (2008), pp. 253–263.

[**Bar+17a**] R. Baraniuk, S. Foucart, D. Needell, Y. Plan, and M. Wootters. "One-bit compressive sensing of dictionary-sparse signals". In: *Information and Inference* 7.1 (2017), pp. 83–104.

[**Bar+17b**] R. G. Baraniuk, S. Foucart, D. Needell, Y. Plan, and M. Wootters. "Exponential decay of reconstruction error from binary measurements of sparse signals". In: *IEEE Transactions on Information Theory* 63.6 (2017), pp. 3368–3385.

[**BB08**] P. T. Boufounos and R. G. Baraniuk. "1-bit compressive sensing". In: *Conference on Information Sciences and Systems (CISS)*. IEEE. 2008, pp. 16–21.

[**BD09**] T. Blumensath and M. E. Davies. "Iterative hard thresholding for compressed sensing". In: *Applied and Computational Harmonic Analysis* 27.3 (2009), pp. 265–274. ISSN: 1063-5203.

[**Ber+09**] C. R. Berger, S. Zhou, J. C. Preisig, and P. K. Willett. "Sparse channel estimation for multicarrier underwater acoustic communication: From subspace methods to compressed sensing". In: *IEEE Transactions on Signal Processing* 58 (2009), pp. 1708–1721.

[**Ber+10**] C. R. Berger, Z. Wang, J. Huang, and S. Zhou. "Application of compressive sensing to sparse channel estimation". In: *IEEE Communications Magazine* 48.11 (2010), pp. 164–174.

[**BF09**] E. van den Berg and M. P. Friedlander. "Joint-sparse recovery from multiple measurements". In: (2009). arXiv: 0904.2051.

[**BKR10**] P. T. Boufounos, G. Kutyniok, and H. Rauhut. "Average case analysis of sparse recovery from combined fusion frame measurements". In: *Conference on Information Sciences and Systems (CISS)*. Mar. 2010, pp. 1–6.

[**BL15**] D. Bilyk and M. T. Lacey. "Random tessellations, restricted isometric embeddings, and one bit sensing". In: (2015). arXiv: 1512.06697.

[**BLM13**] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN: 9780199535255.

[**Blu12**] T. Blumensath. "Accelerated iterative hard thresholding". In: *Signal Processing* 92.3 (2012), pp. 752–756.

[**Bou+15**] P. T. Boufounos, L. Jacques, F. Krahmer, and R. Saab. "Quantization and compressive sensing". In: *Compressed Sensing and Its Applications*. Springer, 2015, pp. 193–237.

[**Bou09**] P. T. Boufounos. "Greedy sparse signal reconstruction from sign measurements". In: *Asilomar Conference on Signals, Systems and Computers*. IEEE. 2009, pp. 1305–1309.

[**Bou10**] P. T. Boufounos. "Reconstruction of sparse signals from distorted randomized measurements". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2010, pp. 3998–4001.

[**CA18**] I. Y. Chun and B. Adcock. "Uniform recovery from subgaussian multi-sensor measurements". In: *Applied and Computational Harmonic Analysis* (2018).

[**Can08**] E. J. Candès. "The restricted isometry property and its implications for compressed sensing". In: *Comptes Rendus Mathematique* 346.9-10 (2008), pp. 589–592.

[**CB15**] S. Chen and A. Banerjee. "One-bit compressed sensing with the k-support norm". In: *Artificial Intelligence and Statistics*. 2015, pp. 138–146.

[**CD94**] S. Chen and D. Donoho. "Basis pursuit". In: *Asilomar Conference on Signals, Systems and Computers*. Vol. 1. IEEE. 1994, pp. 41–44.

[**CD99**] E. J. Candès and D. L. Donoho. "Curvelets – a surprisingly effective nonadaptive representation for objects with edges". en. In: *Curves and Surfaces Fitting*. Ed. by L. L. Schumaker, A. Cohen, and C. Rabut. Vanderbilt University Press, 1999, p. 16.

[**CDD09**] A. Cohen, W. Dahmen, and R. DeVore. "Compressed sensing and best k-term approximation". In: *Journal of the American Mathematical Society* 22.1 (2009), pp. 211–231.

[**CGH18**] T. Chen, M. Guo, and X. Huang. "Direction finding using compressive one-bit measurements". In: *IEEE Access* 6 (2018), pp. 41201–41211.

[**CGM01**] R. Coifman, F. Geshwind, and Y. Meyer. "Noiselets". In: *Applied and Computational Harmonic Analysis* 10.1 (2001), pp. 27–44. ISSN: 1063-5203.

[**CH06**] J. Chen and X. Huo. "Theoretical results on sparse representations of multiple-measurement vectors". In: *IEEE Transactions on Signal Processing* 54 (2006), pp. 4634–4643.

[**CH12**] C. Chen and J. Huang. "Compressive sensing MRI with wavelet tree sparsity". In: *Advances in Neural Information Processing Systems (NIPS)*. 2012, pp. 1115–1123.

[**CK12**] P. G. Casazza and G. Kutyniok. *Finite Frames: Theory and Applications*. Springer, 2012.

[**CRT06a**] E. J. Candès, J. K. Romberg, and T. Tao. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on Information Theory* 52 (2006), pp. 489–509.

[**CRT06b**] E. J. Candès, J. K. Romberg, and T. Tao. "Stable signal recovery from incomplete and inaccurate measurements". In: *Communications on Pure and Applied Mathematics* 59.8 (2006), pp. 1207–1223. ISSN: 1097-0312.

[**CT05**] E. J. Candès and T. Tao. "Decoding by linear programming". In: *IEEE Transactions on Information Theory* 51.12 (Dec. 2005), pp. 4203–4215. ISSN: 0018-9448.

[**CT06a**] E. J. Candès and T. Tao. "Near-optimal signal recovery from random projections: Universal encoding strategies?" In: *IEEE Transactions on Information Theory* 52 (2006), pp. 5406–5425.

[**CT06b**] E. J. Candès and T. Tao. "Near-optimal signal recovery from random projections: Universal encoding strategies?" In: *IEEE Transactions on Information Theory* 52 (2006), pp. 5406–5425.

[**Cui+18**] W. Cui, F. Jiang, X. Gao, S. Zhang, and D. Zhao. "An efficient deep quantized compressed sensing coding framework of natural images". In: *ACM Multimedia Conference on Multimedia Conference*. ACM. 2018, pp. 1777–1785.

[**Dav+12**] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok. "Introduction to compressed sensing". In: *Compressed Sensing: Theory and Applications* 105 (2012), p. 106.

[**DB13**] M. F. Duarte and R. G. Baraniuk. "Spectral compressive sensing". In: *Applied and Computational Harmonic Analysis* 35.1 (2013), pp. 111–129.

[**Dir15**] S. Dirksen. "Tail bounds via generic chaining". In: *Electronic Journal of Probability* 20 (2015).

[**DJR17**] S. Dirksen, H. C. Jung, and H. Rauhut. "One-bit compressed sensing with partial Gaussian circulant matrices". In: (2017). arXiv: 1710.03287.

[**DLR18**] S. Dirksen, G. Lecué, and H. Rauhut. "On the gap between restricted isometry properties and sparse recovery conditions". In: *IEEE Transactions on Information Theory* 64 (2018), pp. 5478–5487.

[**DM18a**] S. Dirksen and S. Mendelson. "Non-Gaussian hyperplane tessellations and robust one-bit compressed sensing". In: 2018. arXiv: 1805.09409.

[**DM18b**] S. Dirksen and S. Mendelson. "Robust one-bit compressed sensing with partial circulant matrices". In: (2018). arXiv: 1812.06719.

[**Don06a**] D. L. Donoho. "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution". In: *Communications on Pure and Applied Mathematics* 59.6 (2006), pp. 797–829.

[**Don06b**] D. L. Donoho. "High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension". In: *Discrete & Computational Geometry* 35.4 (2006), pp. 617–652.

[**Don06c**] D. L. Donoho. "Compressed sensing". In: *IEEE Transactions on Information Theory* 52 (2006), pp. 1289–1306.

[**DT05**] D. L. Donoho and J. Tanner. "Neighborliness of randomly projected simplices in high dimensions". In: *Proceedings of the National Academy of Sciences* 102.27 (2005), pp. 9452–9457.

[**DT09a**] D. L. Donoho and J. Tanner. "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing". In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 367.1906 (2009), pp. 4273–4293.

[**DT09b**] D. Donoho and J. Tanner. "Counting faces of randomly projected polytopes when the projection radically lowers dimension". In: *Journal of the American Mathematical Society* 22.1 (2009), pp. 1–53.

[**DT10a**] D. L. Donoho and J. Tanner. "Counting the faces of randomly-projected hypercubes and orthants, with applications". In: *Discrete & Computational Geometry* 43.3 (2010), pp. 522–541.

[**DT10b**] D. L. Donoho and J. Tanner. "Exponential bounds implying construction of compressed sensing matrices, error-correcting codes, and neighborly polytopes by random sampling". In: *IEEE Transactions on Information Theory* 56.4 (2010), pp. 2002–2016.

[**DU18**] S. Dirksen and T. Ullrich. "Gelfand numbers related to structured sparsity and Besov space embeddings with small mixed smoothness". In: *Journal of Complexity* (2018).

[**Eft$^+$15**] A. Eftekhari, H. L. Yap, C. J. Rozell, and M. B. Wakin. "The restricted isometry property for random block diagonal matrices". In: *Applied and Computational Harmonic Analysis* 38.1 (2015), pp. 1–31. ISSN: 1063-5203.

[**EM09**] Y. C. Eldar and M. Mishali. "Robust recovery of signals from a structured union of subspaces". In: *IEEE Transactions on Information Theory* 55.11 (Nov. 2009), pp. 5302–5316. ISSN: 0018-9448.

[**End10**] J. H. G. Ender. "On compressive sensing applied to radar". In: *Signal Processing* 90.5 (2010), pp. 1402–1414.

[**EV09**] E. Elhamifar and R. Vidal. "Sparse subspace clustering". In: *IEEE Conference on Computer Vision and Pattern Recognition*. June 2009, pp. 2790–2797.

[**FC18**] H. Fu and Y. Chi. "Quantized spectral compressed sensing: Cramer-Rao bounds and recovery algorithms". In: *IEEE Transactions on Signal Processing* 66.12 (2018), pp. 3268–3279.

[**FL09**] S. Foucart and M.-J. Lai. "Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$". In: *Applied and Computational Harmonic Analysis* 26.3 (2009), pp. 395–407.

[**FMT12**] J. E. Fowler, S. Mun, and E. W. Tramel. "Block-based compressed sensing of images and video". In: *Foundations and Trends in Signal Processing* 4.4 (2012), pp. 297–416.

[**Fou11**] S. Foucart. "Hard thresholding pursuit: An algorithm for compressive sensing". In: *SIAM Journal on Numerical Analysis* 49 (2011), pp. 2543–2563.

[**Fou16**] S. Foucart. "Flavors of compressive sensing". In: *International Conference on Approximation Theory*. Springer. 2016, pp. 61–104.

[**FR13**] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Vol. 1. 3. Birkhäuser, Basel, 2013. ISBN: 9780817649470.

[**Fra89**] D. Fraser. "Interpolation by the FFT revisited-an experimental investigation". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37.5 (1989), pp. 665–675.

[**FSY10**] A. C. Fannjiang, T. Strohmer, and P. Yan. "Compressed remote sensing of sparse objects". In: *SIAM Journal on Imaging Sciences* 3.3 (2010), pp. 595–618.

[**Gan07**] L. Gan. "Block compressed sensing of natural images". In: *International Conference on Digital Signal Processing* (2007), pp. 403–406.

[**Gao+17**] Y. Gao, D. Hu, Y. Chen, and Y. Ma. "Gridless 1-b DOA estimation exploiting SVM approach". In: *IEEE Communications Letters* 21.10 (2017), pp. 2210–2213.

[**GBK08**] U. Gamper, P. Boesiger, and S. Kozerke. "Compressed sensing in dynamic MRI". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 59.2 (2008), pp. 365–373.

[**Gen17**] General Electric. *Overcoming one of the biggest MR imaging challenges... scan time*. 2017. URL: http://newsroom.gehealthcare.com/overcoming-one-of-the-biggest-mr-imaging-challenges-scan-time-rsna/ (visited on 02/19/2019).

[**GM17**] Y. Gao and M. Ma. "A new bound on the block restricted isometry constant in compressed sensing". In: *Journal of Inequalities and Applications*. 2017.

[**GNR10**] A. Gupta, R. Nowak, and B. Recht. "Sample complexity for 1-bit compressed sensing and sparse classification". In: *IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2010, pp. 1553–1557.

[**Gop+13**] S. Gopi, P. Netrapalli, P. Jain, and A. Nori. "One-bit compressed sensing: Provable support and vector recovery". In: *International Conference on Machine Learning*. 2013, pp. 154–162.

[**Gor88**] Y. Gordon. "On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$". In: *Geometric Aspects of Functional Analysis*. Ed. by J. Lindenstrauss and V. D. Milman. Berlin, Heidelberg: Springer Berlin Heidelberg, 1988, pp. 84–106. ISBN: 978-3-540-39235-4.

[**GZ84**] E. Gine and J. Zinn. "Some limit theorems for empirical processes". In: *The Annals of Probability* 12.4 (Nov. 1984), pp. 929–989.

[**Hau+10**] J. Haupt, W. U. Bajwa, G. Raz, and R. Nowak. "Toeplitz compressed sensing matrices with applications to sparse channel estimation". In: *IEEE Transactions on Information Theory* 56.11 (2010), pp. 5862–5875.

[**HB11**] J. Haupt and R. Baraniuk. "Robust support recovery using sparse compressive sensing matrices". In: *Conference on Information Sciences and Systems (CISS)*. IEEE. 2011, pp. 1–6.

[**HBM12**] J. Huang, P. Breheny, and S. Ma. "A selective review of group selection in high-dimensional models". In: *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 27.4 (2012).

[**HS09**] M. A. Herman and T. Strohmer. "High-resolution radar via compressed sensing". In: *IEEE Transactions on Signal Processing* 57 (2009), pp. 2275–2284.

[**HS18**] T. Huynh and R. Saab. "Fast binary embeddings, and quantized compressed sensing with structured matrices". In: (2018). arXiv: 1801.08639.

[**HXL18**] X. Huang, P. Xiao, and B. Liao. "One-bit direction of arrival estimation with an improved fixed-point continuation algorithm". In: *International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE. 2018, pp. 1–4.

[**Jac+13**] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk. "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors". In: *IEEE Transactions on Information Theory* 59.4 (2013), pp. 2082–2102.

[**Jac16**] L. Jacques. "Error decay of (almost) consistent signal estimations from quantized Gaussian random projections". In: *IEEE Transactions on Information Theory* 62.8 (2016), pp. 4696–4709.

[**JCM12**] K. Jia, T.-H. Chan, and Y. Ma. "Robust and practical face recognition via structured sparsity". In: *European Conference on Computer Vision*. Springer. 2012, pp. 331–344.

[**JDV13**] L. Jacques, K. Degraux, and C. D. Vleeschouwer. "Quantized iterative hard thresholding: Bridging 1-bit and high-resolution quantized compressed sensing". In: (2013). arXiv: 1305.1786.

[**JHF11**] L. Jacques, D. K. Hammond, and J. M. Fadili. "Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine". In: *IEEE Transactions on Information Theory* 57.1 (2011), pp. 559–571.

[**JOV09**] L. Jacob, G. Obozinski, and J.-P. Vert. "Group lasso with overlap and graph lasso". In: *International Conference on Machine Learning*. ACM. 2009, pp. 433–440.

[**KBM19a**] N. Koep, A. Behboodi, and R. Mathar. "Performance analysis of one-bit group-sparse signal reconstruction". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2019, pp. 5272–5276.

[**KBM19b**] N. Koep, A. Behboodi, and R. Mathar. "The group restricted isometry property for subgaussian block diagonal matrices". In: *IEEE International Symposium on Information Theory (ISIT)*. to appear. 2019.

[**KL12**] G. Kutyniok and D. Labate. *Shearlets: Multiscale Analysis for Multivariate Data*. Ed. by G. Kutyniok and D. Labate. Applied and Numerical Harmonic Analysis. OCLC: ocn794844320. New York: Birkhäuser, 2012. ISBN: 978-0-8176-8315-3.

[**KM17**] N. Koep and R. Mathar. "Binary iterative hard thresholding for frequency-sparse signal recovery". In: *International ITG Workshop on Smart Antennas (WSA)*. Mar. 2017, pp. 1–7.

[**KM18**] N. Koep and R. Mathar. "Block-sparse signal recovery from binary measurements". In: *IEEE Statistical Signal Processing Workshop (SSP)*. June 2018, pp. 293–297.

[**KMR14**] F. Krahmer, S. Mendelson, and H. Rauhut. "Suprema of chaos processes and the restricted isometry property". In: *Communications on Pure and Applied Mathematics* 67.11 (2014), pp. 1877–1904.

[**KSW16**] K. Knudson, R. Saab, and R. Ward. "One-bit compressive sensing with norm estimation". In: *IEEE Transactions on Information Theory* 62 (2016), pp. 2748–2758.

[**Kut13**] G. Kutyniok. "Theory and applications of compressed sensing". In: *GAMM-Mitteilungen* 36.1 (2013), pp. 79–101.

[**Las+11**] J. N. Laska, Z. Wen, W. Yin, and R. G. Baraniuk. "Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements". In: *IEEE Transactions on Signal Processing* 59 (2011), pp. 5289–5301.

[**LDP07**] M. Lustig, D. Donoho, and J. M. Pauly. "Sparse MRI: The application of compressed sensing for rapid MR imaging". In: *Magnetic Resonance in Medicine* 58.6 (2007), pp. 1182–1195.

[**Le+05**] B. Le, T. W. Rondeau, J. H. Reed, and C. W. Bostian. "Analog-to-digital converters". In: *IEEE Signal Processing Magazine* 22.6 (2005), pp. 69–77.

[**LGX16**] W. Liu, D. Gong, and Z. Xu. "One-bit compressed sensing by greedy algorithms". In: *Numerical Mathematics: Theory, Methods and Applications* 9.2 (2016), pp. 169–184.

[**Luo+09**] X. Luo, J. Zhang, J. Yang, and Q. Dai. "Image fusion in compressed sensing". In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2009, pp. 2205–2208.

[**Lus+08**] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly. "Compressed sensing MRI". In: *IEEE Signal Processing Magazine* 25.2 (2008), p. 72.

[**LV17**] C.-L. Liu and P. Vaidyanathan. "One-bit sparse array DOA estimation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 3126–3130.

[**Lyo04**] R. G. Lyons. *Understanding Digital Signal Processing*. Upper Saddle River, NJ, USA: Prentice Hall, 2004. ISBN: 0131089897.

[**Men14**] S. Mendelson. "Learning without concentration". In: *Conference on Learning Theory*. 2014, pp. 25–39.

[**MF09**] S. Mun and J. E. Fowler. "Block compressed sensing of images using directional transforms". In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2009, pp. 3021–3024.

[**MF11**] S. Mun and J. E. Fowler. "Residual reconstruction for block-based compressed sensing of video". In: *Data Compression Conference (DCC)*. IEEE. 2011, pp. 183–192.

[**MF12**] S. Mun and J. E. Fowler. "DPCM for quantized block-based compressed sensing of images". In: *European Signal Processing Conference (EUSIPCO)*. IEEE. 2012, pp. 1424–1428.

[**Mos+16**] A. Moshtaghpour, L. Jacques, V. Cambareri, K. Degraux, and C. De Vleeschouwer. "Consistent basis pursuit for signal and matrix estimates in quantized compressed sensing". In: *IEEE Signal Processing Letters* 23.1 (2016), pp. 25–29.

[**Mos10**] A. Mosek. "The MOSEK optimization software". In: 54.2-1 (2010), p. 5.

[**MP19**] J. Maly and L. Palzer. "Analysis of hard-thresholding for distributed compressed sensing with one-bit measurements". In: *Information and Inference* (Apr. 2019). ISSN: 2049-8772.

[**MPT08**] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. "Uniform uncertainty principle for Bernoulli and subgaussian ensembles". In: *Constructive Approximation* 28.3 (2008), pp. 277–289.

[**NT09**] D. Needell and J. A. Tropp. "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples". In: *Applied and Computational Harmonic Analysis* 26.3 (2009), pp. 301–321.

[**OJV11**] G. Obozinski, L. Jacob, and J.-P. Vert. "Group lasso with overlaps: The latent group lasso approach". In: (2011). arXiv: 1110.0413.

[**Par+11**] J. Y. Park, H. L. Yap, C. J. Rozell, and M. B. Wakin. "Concentration of measure for block diagonal matrices with applications to compressive signal processing". In: *IEEE Transactions on Signal Processing* 59.12 (2011), pp. 5859–5875.

[**PB14**] N. Parikh and S. P. Boyd. "Proximal algorithms". In: *Foundations and Trends in Optimization* 1 (2014), pp. 127–239.

[**Phi18**] Philips. *Philips showcases portfolio of innovative solutions at RSNA 2018, comprising imaging systems, intelligent software applications and services*. 2018. URL: https://www.philips.com/a-w/about/news/archive/standard/news/press/2018/20181125-philips-showcases-portfolio-of-innovative-solutions-at-rsna-2018.html (visited on 02/19/2019).

[**Pol+09**] Y. L. Polo, Y. Wang, A. Pandharipande, and G. Leus. "Compressive wide-band spectrum sensing". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2009, pp. 2337–2340.

[**PV13a**] Y. Plan and R. Vershynin. "One-bit compressed sensing by linear programming". In: *Communications on Pure and Applied Mathematics* 66.8 (Feb. 2013), pp. 1275–1297.

[**PV13b**] Y. Plan and R. Vershynin. "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach". In: *IEEE Transactions on Information Theory* 59.1 (2013), pp. 482–494.

[**PV14**] Y. Plan and R. Vershynin. "Dimension reduction by random hyperplane tessellations". In: *Discrete & Computational Geometry* 51.2 (2014), pp. 438–461.

[**PV16**] Y. Plan and R. Vershynin. "The generalized lasso with non-linear observations". In: *IEEE Transactions on Information Theory* 62.3 (2016), pp. 1528–1537.

[**PVY17**] Y. Plan, R. Vershynin, and E. Yudovina. "High-dimensional estimation with geometric constraints". In: *Information and Inference* 6.1 (2017), pp. 1–40.

[**PW09**] J. Y. Park and M. B. Wakin. "A multiscale framework for compressive sensing of video". In: *Picture Coding Symposium (PCS)*. IEEE. 2009, pp. 1–4.

[**Qin+18**] Z. Qin, J. Fan, Y. Liu, Y. Gao, and G. Y. Li. "Sparse representation for wireless communications: A compressive sensing approach". In: *IEEE Signal Processing Magazine* 35.3 (2018), pp. 40–58.

[**Rao+13**] N. Rao, C. Cox, R. Nowak, and T. T. Rogers. "Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis". In: *Advances in Neural Information Processing Systems (NIPS)*. 2013, pp. 2202–2210.

[**Rao+14**] N. Rao, R. Nowak, C. Cox, and T. Rogers. "Classification with sparse overlapping groups". In: (2014). arXiv: 1402.4512.

[**Rau10**] H. Rauhut. "Compressive sensing and structured random matrices". In: *Theoretical Foundations and Numerical Methods for Sparse Recovery* 9 (2010), pp. 1–92.

[**REC04**] B. D. Rao, K. Engan, and S. F. Cotter. "Sparse solutions to linear inverse problems with multiple measurement vectors". In: *IEEE Transactions on Signal Processing* 53 (2004), pp. 2477–2488.

[**Roc15**] R. T. Rockafellar. *Convex analysis*. Princeton University Press, 2015.

[**Roz+10**] C. J. Rozell, H. L. Yap, J. Y. Park, and M. B. Wakin. "Concentration of measure for block diagonal matrices with repeated blocks". In: *Conference on Information Sciences and Systems (CISS)*. IEEE. 2010, pp. 1–6.

[**RRT12**] H. Rauhut, J. Romberg, and J. A. Tropp. "Restricted isometries for partial random circulant matrices". In: *Applied and Computational Harmonic Analysis* 32.2 (2012), pp. 242–254.

[**RSV08**] H. Rauhut, K. Schnass, and P. Vandergheynst. "Compressed sensing and redundant dictionaries". In: *IEEE Transactions on Information Theory* 54.5 (2008), pp. 2210–2219.

[**Sar+05**] S. Sarvotham, D. Baron, M. Wakin, M. F. Duarte, and R. G. Baraniuk. "Distributed compressed sensing of jointly sparse signals". In: *Asilomar Conference on Signals, Systems and Computers*. 2005, pp. 1537–1541.

[**Sie16**] Siemens Healthcare GmbH. *Faster MRI scans with compressed sensing from Siemens Healthineers*. 2016. URL: https://www.healthcare.siemens.com/press-room/press-releases/pr-2016110086hcen.html (visited on 02/19/2019).

[**Sim+13**] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. "A sparse-group lasso". In: *Journal of Computational and Graphical Statistics* 22.2 (2013), pp. 231–245.

[**Sra12**] S. Sra. "Fast projections onto mixed-norm balls with applications". In: *Data Mining and Knowledge Discovery* 25.2 (2012), pp. 358–377.

[**Stö+15**] C. Stöckle, J. Munir, A. Mezghani, and J. A. Nossek. "1-bit direction of arrival estimation based on compressed sensing". In: *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. June 2015, pp. 246–250.

[**Sul+12**] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand. "Overview of the high efficiency video coding (HEVC) standard". In: *IEEE Transactions on Circuits and Systems for Video Technology* 22.12 (2012), pp. 1649–1668.

[**Tal10**] M. Talagrand. *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. Springer, 2010. ISBN: 3642063861.

[**TG07**] Z. Tian and G. B. Giannakis. "Compressed sensing for wideband cognitive radios". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 4. 2007, pp. IV–1357.

[**TH08**] G. Tauböck and F. Hlawatsch. "A compressed sensing technique for OFDM channel estimation in mobile environments: Exploiting channel sparsity for reducing pilots". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 8. 2008, pp. 2885–2888.

[**TP14**] A. M. Tillmann and M. E. Pfetsch. “The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing”. In: *IEEE Transactions on Information Theory* 60.2 (2014), pp. 1248–1259.

[**Van12**] R. J. Vanderbei. “Fast Fourier optimization”. In: *Mathematical Programming Computation* 4.1 (2012), pp. 53–69.

[**Ver12**] R. Vershynin. “Lectures in Geometric Functional Analysis”. In: 2012.

[**Ver18**] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[**Wak+10**] M. B. Wakin, J. Y. Park, H. L. Yap, and C. J. Rozell. “Concentration of measure for block diagonal measurement matrices”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2010, pp. 3614–3617.

[**Wal99**] R. H. Walden. “Analog-to-digital converter survey and analysis”. In: *IEEE Journal on Selected Areas in Communications* 17.4 (1999), pp. 539–550.

[**Wan+09**] Y. Wang, A. Pandharipande, Y. L. Polo, and G. Leus. “Distributed compressive wide-band spectrum sensing”. In: *Information Theory and Applications Workshop*. IEEE. 2009, pp. 178–183.

[**WC17**] U. L. Wijewardhana and M. Codreanu. “A bayesian approach for online recovery of streaming signals from compressive measurements”. In: *IEEE Transactions on Signal Processing* 65.1 (2017), pp. 184–199.

[**XJ18**] C. Xu and L. Jacques. “Quantized compressive sensing with RIP matrices: The benefit of dithering”. In: (2018). arXiv: 1801.05870.

[**YA09**] T. Yucek and H. Arslan. “A survey of spectrum sensing algorithms for cognitive radio applications”. In: *IEEE Communications Surveys & Tutorials* 11.1 (2009), pp. 116–130.

[**Yan+09**] Y. Yang, O. C. Au, L. Fang, X. Wen, and W. Tang. “Perceptual compressive sensing for image signals”. In: *IEEE International Conference on Multimedia and Expo*. IEEE. 2009, pp. 89–92.

[**Yap+11**] H. L. Yap, A. Eftekhari, M. B. Wakin, and C. J. Rozell. “The restricted isometry property for block diagonal matrices”. In: *Conference on Information Sciences and Systems (CISS)*. IEEE. 2011, pp. 1–6.

[**Yu+16**] K. Yu, Y. D. Zhang, M. Bao, Y.-H. Hu, and Z. Wang. “DOA estimation from one-bit compressed array data via joint sparse representation”. In: *IEEE Signal Processing Letters* 23.9 (2016), pp. 1279–1283.

[**YYO12**] M. Yan, Y. Yang, and S. Osher. “Robust 1-bit compressive sensing using adaptive outlier pursuit”. In: *IEEE Transactions on Signal Processing* 60.7 (2012), pp. 3868–3875.

[**ZF14a**] X. Zeng and M. A. T. Figueiredo. “Robust binary fused compressive sensing using adaptive outlier pursuit”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, pp. 7674–7678.

[**ZF14b**] X. Zeng and M. A. T. Figueiredo. “Binary fused compressive sensing: 1-bit compressive sensing meets group sparsity”. In: (2014). arXiv: 1402.5074.

[**ZLT11**] F. Zeng, C. Li, and Z. Tian. “Distributed compressive spectrum sensing in cooperative multihop cognitive networks”. In: *IEEE Journal of Selected Topics in Signal Processing* 5.1 (2011), pp. 37–48.

Curriculum Vitae

Niklas Koep, M.Sc.	Born on January 10, 1988 in Birkesdorf, Germany
1994 – 1998	Attendance of elementary school, Katholische Grundschule Merzenich "Am Weinberg", Germany
1998 – 2007	Attendance of high school, Städtisches Gymnasium am Wirteltor, Düren, Germany
2007 – 2013	Study of Electrical Engineering, Information Technology and Computer Engineering, RWTH Aachen University, Germany
2013 – 2014	Research assistant at the Institute of Communication Systems and Data Processing (IND), RWTH Aachen University, Germany
Since 2014	Research assistant at the Institute for Theoretical Information Technology (TI), RWTH Aachen University, Germany