Technische Universität München

Fakultät für Mathematik
Lehrstuhl für Angewandte Numerische Analysis

Recovery Algorithms for Quantized Compressed Sensing

Johannes Benjamin Maly

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Michael Wolf

Prüfer der Dissertation:
1. Prof. Dr. Massimo Fornasier
2. Prof. Dr. Holger Rauhut
   Rheinisch-Westfälische Technische Hochschule Aachen
   Prof. Dr. Laurent Jacques
   Université Catholique de Louvain
3. Prof. Dr. Gerhard Kramer
   Prof. Dr. Kiryung Lee
   Ohio State University

Die Dissertation wurde am 17.01.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 29.04.2019 angenommen.

Acknowledgements

There are two persons I especially wish to thank at this point. First, my supervisor Massimo Fornasier for his constant guidance and support throughout those three years. I learned many more things than only math from him. And second, Carl-Friedrich Kreiner who steadily accompanied me on my academic journey at TUM, be it as an exercise instructor, Bachelor's thesis advisor, or Ph.D. mentor.

Moreover, I would like to mention those I was allowed to closely learn and work with: Mark Iwen, Valeriya Naumova, Lars Palzer, and all my colleagues at the unit M-15 of TUM. I truly enjoyed the time we shared.

Contents

Chapter 1

Introduction

Many practical problems of science and technology boil down to determining quantities of interest from measured information. Be it in Physics, Engineering, Chemistry or Biology, sensors are everywhere and measure signals (like sound waves or images), structure of materials, and concentration of substances. Since measurements cost time and money, one is in general interested in reducing the amount necessary to identify the quantity of interest.

In principle, a measurement process can be regarded as a set of questions whose answers are used to guess the unknown quantity. The number of questions one has to ask may be considerably decreased by a smart design which uses preliminary knowledge on the unknown quantity. Have a look at the old family picture in Figure 1.1 showing five children, two boys and three girls. I select one of them and ask you to guess my choice by posing yes/no questions. A possible question of yours might be

Have you selected the boy who sits left? (Q1)

But (Q1) is not an optimal choice in general. If by chance you picked the correct child, you would win. But in the more probable case of making the wrong choice (I selected the other boy), you gained almost no information. The number of possible correct answers has just been reduced by one. A better question to ask would be

Have you selected a child who is sitting? (Q2)

In this case you eliminate at least two wrong possibilities independent of my concrete choice. Let us assume now that you are given some preliminary information before asking any questions, namely, that I selected a boy. The situation drastically changes as both questions, (Q1) and (Q2), consequently yield the same amount of information; they determine the correct solution.

This simple example illustrates two important aspects. First, if one wishes to use, independent of the unknown quantity, as few questions as possible, it is crucial to design the questions such that the gain of information does not depend on the answer. Second, any preliminary information on the unknown quantity influences effectiveness of questions and may allow to obtain a solution from considerably less inquiries.

Figure 1.1: Family picture

Measurement processes are usually modeled by a function f mapping the unknown quantity $\mathbf{x}$ to observed measurements $\mathbf{y}$. If f is linear, $\mathbf{x} \in \mathbb{R}^N$ a real-valued vector, and there are m measurements, we can write $\mathbf{y} = \mathbf{A}\mathbf{x}$, for some matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$. One may think of posing m questions of the form $\langle \mathbf{a}_i, \mathbf{x}\rangle$, $i = 1, ..., m$, and guessing $\mathbf{x}$ from the answers y_i. On the one hand, it is in general not possible to recover $\mathbf{x}$ from $\mathbf{A}$ and $\mathbf{y}$ if $m < N$. On the other hand, in applications N might be exceedingly large while m is often restricted due to budget or technical feasibility. To solve the dilemma one must use preliminary knowledge on $\mathbf{x}$ meaning that $\mathbf{x}$ belongs to some relatively small subset K of $\mathbb{R}^N$ and m measurements suffice to distinguish points in K. A popular choice for K is the set of sparse vectors, i.e., vectors which have only few non-zero entries. Indeed, in many applications this is a realistic assumption.

A demonstrative example from everyday life are digital cameras. Let us consider a ten megapixel camera which nowadays can be found in every smartphone. "Ten megapixel" means that in order to capture the picture (here the unknown quantity is an image), the camera takes ten million measurements corresponding to one color value per image point. In raw format the digital picture uses around 30MB. However, widespread compression techniques like JPEG allow to reduce the size to less than 4MB keeping the relevant information (the picture in Figure 1.1, which has been compressed to JPEG, substantiates the point). Image compression like JPEG is based on the following idea. A ten megapixel picture can be stored as a ten million-dimensional vector $\bar{\mathbf{x}}$. Each entry of $\bar{\mathbf{x}}$ corresponds to the color value of one pixel. In general, $\bar{\mathbf{x}}$ is dense, i.e., almost all entries are non-zero. Yet, vectors corresponding to pictures are not spread in the whole space. They mainly concentrate around a union of low-dimensional subspaces. Rotating the coordinate system by applying a suitable Wavelet transformation leads to a sparsified representation of $\bar{\mathbf{x}}$ such that most of the entries are close to zero and thus negligible. Setting those entries to zero one obtains a good approximation $\mathbf{x}$ of $\bar{\mathbf{x}}$ which is sparse. Instead of storing $\bar{\mathbf{x}}$, JPEG stores the non-zero entries of $\mathbf{x}$ and their locations. Quality of the JPEG image depends on the required compression rate which regulates the number of thresholded entries. Since $80 - 90\%$ of the measurements are cast away in the end, one wonders if the measurement process could be optimized such that a camera with one megapixel outputs the same picture. While maybe of moderate interest for hobby photographers, in *magnetic resonance imaging* (MRI) a measurement reduction of 90% massively decreases the duration of examinations reducing costs and supporting the well-being of patients.

And not only images but also sound waves and electromagnetic waves become sparse under suitable orthogonal transformations like Wavelet or Fourier transform. Early works on sparse frequency estimation [122], signal recovery [51], and image processing [146] considered sparsity promoting regularization to de-noise signals and to recover sparse signals from underdetermined measurements. In statistics, the Least Absolute Shrinkage and Selection Operator (LASSO) became popular for obtaining sparse and thus interpretable solutions to regression problems [160]. While already in the 1990s sparse approximation algorithms have been analyzed [33, 124] and conditions for successful recovery of sparse signals have followed in the early 2000s [50, 70], only the seminal works [30] and [48] carved out effectiveness of sparse regularization in combination with random measurement matrices to solve underdetermined systems of equations. By stressing the importance for signal processing applications they have founded the field of compressed sensing resulting in the improved design of measurement devices in, e.g., MRI [82, 129] and radar [85].

Though the above described linear measurement model, on which classical compressed sensing theory relies, has proven useful in various applications, it omits a peculiarity all digital measurement processes have in common. Single measurements y_i are not real numbers but only elements of a finite alphabet. A computer could not even store one single real number with infinite precision and thus quantizes, i.e., it projects real numbers to a finite subset of $\mathbb{R}$ and just stores the approximation. (Quantization theory dates back to Shannon's fundamental work [150, 151, 134] on information theory in 1948 and examines the influence of quantizers on signal distortion.) Even if the underlying measurement model behaves linear, additional quantization leads to non-linearity and discontinuity. As long as the quantization has a high resolution, it may be modeled in compressed sensing as additive noise on the perfect measurements $\mathbf{y}$, a situation which is handled anyway by classical results. However, in the case of rough quantization – a popular example are user recommendation systems like in the Netflix Prize [19] where each measurement corresponds to a natural number between one and five – it is crucial to exploit knowledge on the quantization process for obtaining optimal performance when recovering from compressed measurements. Especially one-bit quantization rouse a lot of interest in the past few years. In this extreme setting, each linear measurement is quantized to one bit, namely, its sign. One-bit compressed sensing is attractive as sensors measuring only one bit of information are cheap to produce and quite robust to pre-quantization noise. However, compressed sensing algorithms need to be adapted/extended to exploit knowledge on the quantization process and to work properly in this highly non-linear setting.

Contribution In this self-contained thesis we present novel algorithmic approaches to quantized compressed sensing and matrix recovery, their theoretical analysis, and empirical experiments demonstrating their performance as well as the validity of theoretical results. *The essential theoretical contribution is the generalization and non-trivial adaption of restricted isometry properties to treat nonlinear extensions of compressed sensing (dis-*

tributed quantization and manifold quantization) and non-convex programs (A-T-LAS$_{2,1}$). Starting from a detailed introduction into compressed sensing and quantization, the thesis embeds [63, 64, 108, 91, 125], extracts from the Master's thesis [169], and some yet unpublished insights into recent research. While Chapter 2 and 3 are a review of established results, Chapter 4-7 and the Appendix consist of original work.

In **Chapter 2**, we provide a brief overview over the fundamental concepts of compressed sensing. We explain how the problem of sparse reconstruction provoked definition of null space properties and restricted isometry properties, how those properties are connected to stable and robust recovery, and present different popular recovery strategies of compressed sensing.

In **Chapter 3**, we introduce the concept of quantization and discuss the challenges it imposes on classical compressed sensing theory. We present generalized notions of dimensionality to capture intrinsic complexity of signal sets and report on recent progress in recovery from one-bit and multi-bit quantized compressed sensing measurements.

In **Chapter 4**, we propose a suitable signal class for distributed compressed sensing and provide the first theoretical justification for numerically observed performance of joint signal recovery from one-bit measurements. The results of this chapter have been published in [125].

In **Chapter 5**, we propose the first tractable algorithm for recovering signals living on low-dimensional manifolds from one-bit measurements. We proof approximation guarantees which resemble recent theoretical but algorithmicly intractable results and apply them to manifolds learned from samples. The results of this chapter have been published in [108, 91].

In **Chapters 6 & 7**, we explain how multi-bit compressed sensing relates to classification problems in machine learning and how this relation led to the problem of recovering matrices having multiple structures. We propose a novel algorithm for matrix sensing which profits from sparsity and low-rankness simultaneously and introduce signal sets of matrices which have effectively sparse and non-orthogonal low-rank decompositions to analyze its performance. The results of Chapter 7 have been published in [63, 64].

In the **Appendix**, we discuss a parameter choice strategy for the Least Absolute Shrinkage and Selection Operator (LASSO) by using an elementary sparse recovery algorithm, work that has been published in [169], and provide some technical and straightforward proofs from Chapter 7.

Notation Let us fix some notational conventions for the rest of the thesis. We abbreviate $[n] = \{1, ..., n\}$. We use lowercase and uppercase bold letters for vectors and matrices, respectively. Hence, a matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ can be clearly distinguished from its rows $\mathbf{a}_i$ and its entries $a_{i,j}$, for $i \in [n_1], j \in [n_2]$. We denote *rank* and *kernel* of $\mathbf{A}$ by $\operatorname{rank}(\mathbf{A})$ and

$\ker(\mathbf{A})$.

By *singular value decomposition* (SVD), any matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ can be decomposed into the product $\mathbf{U\Sigma V}^T$ where $\mathbf{U} \in \mathbb{R}^{n_1 \times n_1}, \mathbf{V} \in \mathbb{R}^{n_2 \times n_2}$ are orthogonal matrices and $\mathbf{\Sigma} \in \mathbb{R}^{n_1 \times n_2}$ is diagonal. The diagonal entries $\sigma_i \geq 0$, for $i \in [\min\{n_1, n_2\}]$, are called *singular values* and the columns of $\mathbf{U}$ and $\mathbf{V}$ are called *singular vectors*. If $\operatorname{rank}(\mathbf{A}) = R$, we can use the *rank-reduced* SVD $\mathbf{U\Sigma V}^T$ where $\mathbf{U} \in \mathbb{R}^{n_1 \times R}, \mathbf{V} \in \mathbb{R}^{n_2 \times R}$ have orthonormal columns and $\mathbf{\Sigma} \in \mathbb{R}^{R \times R}$ with $\sigma_i > 0$, for all $i \in [R]$.

We use a variety of norms in this work: let $\|\cdot\|_p$, for $p > 0$, denote the *ℓ_p-(quasi)-norm* of a vector. For $p = \infty$ we get the *supremum norm* and by abuse of notation, for $p = 0$, the *ℓ_0-norm* which is not a norm but counts the number of non-zero entries of a vector (it can be interpreted as the limit $p \to 0$). All ℓ_p-(quasi)-norms are transferred to matrices by applying them to the vector of singular values. In this case, they are called *Schatten-ℓ_p-(quasi)-norms*. Moreover, the ℓ_∞-norm becomes the *operator norm* and is denoted by $\|\cdot\|_{2\to 2}$ while the ℓ_1-norm becomes the *nuclear norm* and is denoted by $\|\cdot\|_*$. The Schatten-ℓ_2-norm is also called *Frobenius norm* and written as $\|\cdot\|_F$. In addition to those, we will use the mixed matrix norms $\|\cdot\|_{2,1}$ and $\|\cdot\|_{1,2}$ which are defined as the sum of the ℓ_2-norms of the columns resp. rows.

We write $\mathcal{B}_p(\mathbf{z}, r)$ to denote the ℓ_p-ball of radius $r > 0$ at $\mathbf{z} \in \mathbb{R}^N$ and $\mathbb{S}^{N-1}$ to denote the $(N-1)$-dimensional Euclidean unit sphere.

An *$(\varepsilon, \|\cdot\|)$-net* $\tilde{K}$ of a set $K \subset \mathbb{R}^N$ is a subset $\tilde{K} \subset K$ such that for any $\mathbf{z} \in K$ there exists $\tilde{\mathbf{z}} \in \tilde{K}$ with $\|\mathbf{z} - \tilde{\mathbf{z}}\| \leq \varepsilon$. The *covering number* $N(K, \|\cdot\|, \varepsilon)$ denotes the cardinality of a minimal $(\varepsilon, \|\cdot\|)$-net of K. If the underlying metric is clear, we write ε-net and $N(K, \varepsilon)$.

We denote the *support* of a vector $\mathbf{z} \in \mathbb{R}^N$ by $\operatorname{supp}(\mathbf{z}) = \{i \in [N] \colon z_i \neq 0\}$ and define the set of s-sparse vectors $\Sigma_s^N = \{\mathbf{z} \in \mathbb{R}^N \colon |\operatorname{supp}(\mathbf{z})| \leq s\}$. For any set $S \subset [N]$, $\mathbf{z}_S \in \mathbb{R}^N$ is the *restriction* of $\mathbf{z}$ to S, i.e., all entries not in S are set to zero. The *best s-term approximation* of $\mathbf{x}$ in ℓ_p is defined as $\sigma_s(\mathbf{x})_p \coloneqq \inf\{\|\mathbf{x} - \mathbf{z}\|_p : \mathbf{z} \in \Sigma_s^N\}$

The *probability* of an event A and the *expectation* of a random variable W are written as $\Pr[A]$ and $\mathsf{E}[W]$. For $K \subset \mathbb{R}^N$, let $\mathcal{U}(K)$ be the *uniform distribution* on K and denote the *normal distribution* with expectation $\boldsymbol{\mu} \in \mathbb{R}^N$ and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{N \times N}$ by $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$.

We use $a \gtrsim b$ and $a \lesssim b$ to express $a \gtrsim Cb$ and $a \lesssim Cb$, for some absolute constant $C > 0$. If $a \gtrsim b$ and $a \lesssim b$, we write $a \simeq b$.

We write $\operatorname{dist}(K, K') = \inf_{\mathbf{z} \in K, \mathbf{z}' \in K'} \|\mathbf{z} - \mathbf{z}'\|_2$ for the *distance* of two sets $K, K' \subset \mathbb{R}^N$ and by abuse of notation $\operatorname{dist}(\mathbf{0}, K) = \inf_{\mathbf{z} \in K} \|\mathbf{z}\|_2$. We denote the *diameter* of a set by $\operatorname{diam}(K) = \sup_{\mathbf{z} \in K - K} \|\mathbf{z}\|_2$.

We use $\lceil r \rceil$ and $\lfloor r \rfloor$ to denote the smallest $z \in \mathbb{Z}$ with $r \leq z$ and the largest $z \in \mathbb{Z}$ with $z \leq r$.

Let $\operatorname{Vol}_d(K)$ be the *d-dimensional volume* of a set $K \subset \mathbb{R}^N$ and write $\operatorname{Vol}(K)$ if $d = N$.

For $K \subset \mathbb{R}^N$, $\mathbb{P}_K$ denotes the *projection* onto K whenever it is uniquely defined. If $K = \mathbb{S}^{N-1}$, we simply write $\mathbb{P}_{\mathbb{S}}$.

We work with various distance measures: for $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^N$, the *Hamming distance* is defined as $d_H(\mathbf{z}, \mathbf{z}') = |\{i \in [N] \colon z_i \neq z_i'\}|$. The *normalized geodesic distance* is denoted by $d_G(\mathbf{z}, \mathbf{z}') = \frac{1}{\pi} \arccos(\langle \mathbf{z}, \mathbf{z}' \rangle)$, for any $\mathbf{z}, \mathbf{z}' \in \mathbb{S}^{N-1}$, and fulfills $d_G(\mathbf{z}, -\mathbf{z}) = 1$. It can be extended to $\mathbb{R}^N \setminus \{\mathbf{0}\}$ by defining $d_G(\mathbf{z}, \mathbf{z}') \coloneqq d_G(\mathbb{P}_{\mathbb{S}}(\mathbf{z}), \mathbb{P}_{\mathbb{S}}(\mathbf{z}'))$. The distance d_A is more involved and can be found in Definition 3.3.4.

We denote the *vectorization* of a matrix $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$ by $\mathrm{vec}(\mathbf{Z}) \in \mathbb{R}^{n_1 n_2}$.
The *indicator function* $\mathbb{1}_K(\mathbf{z})$ of a set $K \subset \mathbb{R}^N$ is one if $\mathbf{z} \in K$ and zero if $\mathbf{z} \notin K$.

Chapter 2

Compressed Sensing

In this chapter we introduce in detail the basic concepts of compressed sensing and provide a collection of well-known definitions and results. Starting with theoretical lower bounds on the minimal number of measurements necessary to identify sparse signals from their linear measurements, the chapter leads the way to efficient algorithms which are guaranteed to approximate sparse signals from noisy measurements.

2.1 Sparsity and Under-determined Linear Systems

As already mentioned in Chapter 1, we are interested in recovering high-dimensional signals $\mathbf{x} \in \mathbb{R}^N$ from few linear measurements of the form

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{2.1}$$

where $\mathbf{A} \in \mathbb{R}^{m \times N}$ models the measurement process and $\mathbf{y} \in \mathbb{R}^m$ is a vector containing the measurements. In general, the problem is ill-posed as soon as $m < N$. To allow reconstruction one needs additional assumptions. Though living in a high-dimensional space, many signals in real-world applications are concentrated on lower dimensional manifolds, i.e., their intrinsic dimension is small in comparison to the ambient dimension N, cf. Section 1. (We will discuss more general measures of dimensionality than the linear dimension in Section 3.2.) In the simplest case we may assume $\mathbf{x}$ to be s-sparse which means that the support $\text{supp}(\mathbf{x}) := |\{i \in [N] \colon x_i \neq 0\}|$ of $\mathbf{x}$ is of size s at most. Under knowledge of the support this corresponds to an intrinsic signal dimension s which is independent of N. We will see later that even without any knowledge of the support the ambient dimension N has just a mild influence on this. From now on we denote the set of s-sparse vectors in $\mathbb{R}^N$ by Σ_s^N.

The first interesting question is, how many measurements are necessary to uniquely identify each s-sparse signal. To be more precise, what is the minimal m, such that $\mathbf{A}\mathbf{z} \neq \mathbf{A}\mathbf{z}'$ implies $\mathbf{z} \neq \mathbf{z}'$, for all $\mathbf{z}, \mathbf{z}' \in \Sigma_s^N$? The following observation, which relies on basic linear algebra, provides an answer.

Lemma 2.1.1 ([34, Lemma 3.1])**.** *Given $\mathbf{A} \in \mathbb{R}^{m \times N}$, the following properties are equivalent:*

(i) *Every s-sparse vector $\mathbf{x} \in \mathbb{R}^N$ is the unique s-sparse solution of $\mathbf{Az} = \mathbf{Ax}$, that is, if $\mathbf{Ax} = \mathbf{Az}$ and both $\mathbf{x}$ and $\mathbf{z}$ are s-sparse, then $\mathbf{x} = \mathbf{z}$.*

(ii) *The nullspace $\ker(\mathbf{A})$ does not contain any $2s$-sparse vector other than the zero vector, i.e., $\ker(\mathbf{A}) \cap \Sigma_{2s}^N = \{\mathbf{0}\}$*

(iii) *Every set of $2s$ columns of $\mathbf{A}$ is linearly independent.*

Proof: *(i)* $\Rightarrow$ *(ii)*: Assume *(i)* holds and $\mathbf{z} \in \ker(\mathbf{A}) \cap \Sigma_{2s}^N$. Then $\mathbf{z}$ can be written as $\mathbf{z} = \mathbf{z}_1 - \mathbf{z}_2$ where $\mathbf{z}_1, \mathbf{z}_2 \in \Sigma_s^N$ and $\operatorname{supp}(\mathbf{z}_1) \cap \operatorname{supp}(\mathbf{z}_2) = \emptyset$. As $\mathbf{Az}_1 - \mathbf{Az}_2 = \mathbf{Az} = \mathbf{0}$, *(a)* implies $\mathbf{z}_1 = \mathbf{z}_2$. But as $\mathbf{z}_1$ and $\mathbf{z}_2$ have disjoint supports we get $\mathbf{z}_1 = \mathbf{z}_2 = \mathbf{z} = \mathbf{0}$.
(ii) $\Rightarrow$ *(iii)*: Assume *(ii)* and that $\mathbf{z} \in \mathbb{R}^N$ encodes a linear combination of $2s$ columns of $\mathbf{A}$ which yields zero, i.e., $\mathbf{z} \in \Sigma_{2s}^N$ with $\mathbf{Az} = \mathbf{0}$. Property *(ii)* implies $\mathbf{z} = \mathbf{0}$.
(iii) $\Rightarrow$ *(i)*: Assume *(iii)* and that $\mathbf{x}, \mathbf{z} \in \Sigma_s^N$ yield the same measurements $\mathbf{Ax} = \mathbf{Az}$. As $\mathbf{x} - \mathbf{z} \in \Sigma_{2s}^N$ and $\mathbf{A}(\mathbf{x} - \mathbf{z}) = \mathbf{0}$, property *(iii)* implies $\mathbf{x} = \mathbf{z}$. ■

The third property of Lemma 2.1.1 shows that at least $m \geq 2s$ linear measurements are necessary to uniquely recover s-sparse signals from their measurements. In fact, one can construct matrices $\mathbf{A} \in \mathbb{R}^{m \times N}$ for $m = 2s$ and arbitrary $N \geq 2s$ such that *(iii)* is fulfilled (see [66, Theorem 2.14]). It is even possible to provide a program which recovers any s-sparse signal $\mathbf{x}$ from its measurements $\mathbf{y}$ in that case. One just asks for the sparsest signal fulfilling the measurements, that is

$$\min_{\mathbf{z} \in \mathbb{R}^N} \|\mathbf{z}\|_0, \quad \text{subject to } \mathbf{Az} = \mathbf{y}, \tag{2.2}$$

where the ℓ_0-norm is defined as $\|\mathbf{z}\|_0 = |\operatorname{supp}(\mathbf{z})|$ (by abuse of notation we write and call $\|\cdot\|_0$ a norm). It is straight-forward to check that if we assume *(i)* in Lemma 2.1.1 for the measurement operator $\mathbf{A}$, the signal $\mathbf{x}$ is the unique solution to (2.2).

This could be the end of the story as we obtained a procedure for recovering s-sparse signals in arbitrarily high dimensions from $m = 2s$ linear measurements. However, the program (2.2) has several drawbacks. First, it is in general NP-hard to solve (cf. [130]). To find a solution one has to solve a linear system for all possible support combinations, the number of which grows like $\binom{N}{s} \approx (N/s)^s$. Second, the program is neither stable nor robust. Being stable means that under small sparsity defects the approximation still works, i.e., if for any $\mathbf{x} \in \mathbb{R}^N$ the *best s-term approximation* is denoted by $\sigma_s(\mathbf{x})_1 := \inf\{\|\mathbf{x} - \mathbf{z}\|_1 : \mathbf{z} \in \Sigma_s^N\}$ then the error in recovery of $\mathbf{x}$ is at most $\mathcal{O}(\sigma_s(\mathbf{x})_1)$. Vectors $\mathbf{x}$ for which $\sigma_s(\mathbf{x})_1$ is small are often called *compressible*. Being robust means that under unknown noise $\boldsymbol{\eta} \in \mathbb{R}^m$ on the measurements, i.e., $\mathbf{y} = \mathbf{Ax} + \boldsymbol{\eta}$, the error in recovery of $\mathbf{x}$ is at most $\mathcal{O}(\|\boldsymbol{\eta}\|_2)$. In case of Fourier measurements (if $\mathbf{A}$ computes a subset of coefficients of the discrete Fourier transform) the first issue can be avoided by using Prony's method which identifies $\operatorname{supp}(\mathbf{x})$ by identifying the zeros of a well-chosen polynomial (see [66, Theorem 2.15] and references therein). Nonetheless, Prony's method suffers by construction the same lack of stability and robustness as (2.2) provoking the

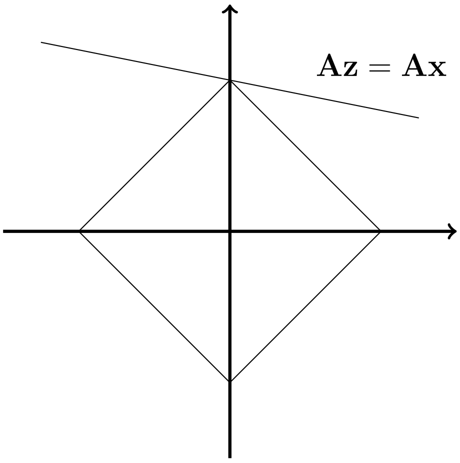


Figure 2.1: Interplay between ℓ_1-ball and $\ker(\mathbf{A})$.

question if one can recover all s-sparse signals in polynomial time in a stable and robust way from $m = 2s$ measurements or, if this is not possible, how many measurements will suffice.

Remark 2.1.2. *We restrict ourselves in this chapter to uniform recovery of signals, that is we are interested in measurement operators* $\mathbf{A}$ *which work for all s-sparse signals at once. There has also been research on non-uniform recovery where one specific but unknown s-sparse signal* $\mathbf{x}$ *is fixed and* $\mathbf{A}$ *may be designed in a way to perform especially well for the recovery of* $\mathbf{x}$*. In this setting it is possible to guarantee unique recovery by* (2.2) *for* $m = s + 1$ *as shown in [168, Theorem 2.1].*

2.2 NSP and RIP - Stability and Robustness

The last section showed that recovery of sparse signals from few measurements is in theory possible but also that it is not obvious how to recover in practice. A possible approach is to relax (2.2) to a tractable program by noting that $\|\mathbf{z}\|_p^p \to \|\mathbf{z}\|_0$ for $p \to 0$ with $p > 0$. However, using $\|\cdot\|_p$, for $0 < p < 1$, yields a non-convex optimization problem which is hard to solve. Choosing $p = 1$ leads to

$$\min_{\mathbf{z} \in \mathbb{R}^N} \|\mathbf{z}\|_1, \quad \text{subject to } \mathbf{Az} = \mathbf{y}, \tag{2.3}$$

which is often called *basis pursuit* and can be viewed as convex relaxation of (2.2). One might wonder if solutions to (2.2) and (2.3) agree. The sketch in Figure 2.1 suggests that the solutions to both problems are identical if $\ker(\mathbf{A})$ is not aligned with high dimensional faces (here the one-dimensional faces) of the ℓ_1-ball. In this case the inflated ℓ_1-ball hits the affine space $\mathbf{Az} = \mathbf{y}$ in exactly one point which corresponds to a sparse solution.

The observation that the kernel geometry plays a key role (recall (*ii*) in Lemma 2.1.1) motivates the following definition of *null space property*. In a more general form it was introduced under this name in [34] but we state the commonly known version in [66]. For any $\mathbf{z} \in \mathbb{R}^N$ and $S \subset [N]$ we denote by $\mathbf{z}_S \in \mathbb{R}^N$ the vector which has set to zero all entries not contained in S.

Definition 2.2.1 (NSP, [66, Definition 4.1]). *A matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$ is said to satisfy the* null space property *relative to a set $S \subset [N]$ if*

$$\|\mathbf{z}_S\|_1 < \|\mathbf{z}_{S^c}\|_1 \quad \text{for all } \mathbf{z} \in \ker(\mathbf{A}) \setminus \{\mathbf{0}\}.$$

It is said to satisfy the null space property of order s if it satisfies the null space property relative to any set $S \subset [N]$ with $|S| \le s$.

Remark 2.2.2. *The NSP of order s implies property (ii) in Lemma 2.1.1.*

It turns out that $\mathbf{A}$ satisfies the NSP if and only if (2.2) and (2.3) are equivalent, i.e., the NSP fully characterizes when (2.2) can be solved by convex relaxation.

Theorem 2.2.3 ([66, Theorem 4.5]). *Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$, every s-sparse vector $\mathbf{x} \in \mathbb{R}^N$ is the unique solution of ℓ_1-minimization subject to $\mathbf{y} = \mathbf{A}\mathbf{z}$ if and only if $\mathbf{A}$ satisfies the null space property of order s.*

Proof: Consider first one fixed support set $S \subset [N]$ with $|S| \le s$. Assume that every $\mathbf{x} \in \mathbb{R}^N$ with $\operatorname{supp}(\mathbf{x}) \subset S$ is the unique minimizer of (2.3). Hence, for any $\mathbf{v} \in \ker(\mathbf{A}) \setminus \{\mathbf{0}\}$, the vector $\mathbf{v}_S$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{v}_S$. This implies $\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{S^c}\|_1$ as $-\mathbf{v}_{S^c} \neq \mathbf{v}_S$ and by $\mathbf{A}(\mathbf{v}_S + \mathbf{v}_{S^c}) = \mathbf{A}\mathbf{v} = \mathbf{0}$ one has $\mathbf{A}(-\mathbf{v}_{S^c}) = \mathbf{A}\mathbf{v}_S$.

Conversely, let us assume that the NSP relative to S holds. Given $\mathbf{x} \in \mathbb{R}^N$ with $\operatorname{supp}(\mathbf{x}) \subset S$ and $\mathbf{z} \in \mathbb{R}^N$ such that $\mathbf{x} \neq \mathbf{z}$ and $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{z}$, define $\mathbf{v} = \mathbf{x} - \mathbf{z} \in \ker(\mathbf{A})$. We get

$$\begin{aligned}\|\mathbf{x}\|_1 &\le \|\mathbf{x} - \mathbf{z}_S\|_1 + \|\mathbf{z}_S\|_1 = \|\mathbf{v}_S\|_1 + \|\mathbf{z}_S\|_1 \\ &< \|\mathbf{v}_{S^c}\|_1 + \|\mathbf{z}_S\|_1 = \|\mathbf{z}\|_1 ,\end{aligned}$$

which shows that $\mathbf{x}$ is the unique minimizer of (2.3). The claim follows by varying over all possible support sets. ■

Theorem 2.2.3 implies the equivalence of (2.2) and (2.3) as, for $\mathbf{x} \in \Sigma_s^N$ minimizing (2.3), any minimizer $\mathbf{x}'$ of (2.2) fulfills $\|\mathbf{x}'\|_0 \le \|\mathbf{x}\|_0$ and hence $\mathbf{x}' = \mathbf{x}$. (The proof illustrates why linearity of $\mathbf{A}$ is a crucial assumption for classical compressed sensing and indicates that considering non-linear measurement processes poses a non-trivial challenge.) Let us defer the question of how many measurements m are necessary to guarantee the NSP for $\mathbf{A}$ and first concentrate on the second issue of (2.2), namely stability and robustness of reconstruction. The NSP as defined in Definition 2.2.1 does not suffice to ensure stability of (2.3) and the linear constraints of (2.3) are too strict to allow noisy measurements of type

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\eta}, \tag{2.4}$$

where $\boldsymbol{\eta} \in \mathbb{R}^m$ is the unknown noise. We hence introduce for $\eta \ge 0$ the convex program

$$\min_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{z}\|_1 , \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \le \eta. \tag{2.5}$$

The minimization in (2.5) is a relaxation of (2.3) and commonly known as *basis pursuit denoising.* In combination with the following stronger NSP it guarantees stable and robust recovery in polynomial time if η is suitably chosen in dependence on the noise level $\|\boldsymbol{\eta}\|_2$.

Definition 2.2.4 (Stable and robust NSP, [66, Definition 4.17]). *A matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$ is said to satisfy the* stable and robust null space property *with constants $0 < \rho < 1$ and $\tau > 0$ relative to a set $S \subset [N]$ if*

$$\|\mathbf{z}_S\|_1 < \rho\, \|\mathbf{z}_{S^c}\|_1 + \tau\, \|\mathbf{A}\mathbf{z}\|_2 \quad \textit{for all } \mathbf{z} \in \mathbb{R}^N.$$

It is said to satisfy the stable and robust null space property of order s if it satisfies the stable and robust null space property with constants $0 < \rho < 1$ and $\tau > 0$ relative to any set $S \subset [N]$ with $|S| \leq s$.

Theorem 2.2.5 ([66, Theorem 4.19]). *Suppose that $\mathbf{A} \in \mathbb{R}^{m\times N}$ satisfies the stable and robust NSP of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $\mathbf{x} \in \mathbb{R}^N$ with measurements* (2.4)*, a solution $\hat{\mathbf{x}}$ of* (2.5) *with $\eta \geq \|\boldsymbol{\eta}\|_2$ fulfills*

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_1 \leq \frac{2(1+\rho)}{1-\rho}\sigma_s(\mathbf{x})_1 + \frac{4\tau}{1-\rho}\eta. \tag{2.6}$$

Theorem 2.2.5 is the consequence of the stronger statement [66, Theorem 4.20] which provides an equivalence relation, not only an implication of the stable and robust NSP. The parameters ρ and τ in Definition 2.2.4 control stability and robustness as can be seen from Theorem 2.2.5. For $\rho = 1$ and $\mathbf{z} \in \ker(\mathbf{A}) \setminus \{\mathbf{0}\}$ one recovers the NSP.

Though null space properties yield valuable insights into the solvability of (2.2) it is in many situations convenient to work with a stronger condition, the so-called *restricted isometry property.* This property of a matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$ has been introduced in [31] under the name "uniform uncertainty principle" and implies the stable and robust NSP.

Definition 2.2.6 (RIP, [66, Definition 6.1]). *A matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$ satisfies the restricted isometry property of order s (s-RIP) with RIP constant $0 < \delta < 1$ if, for all $\mathbf{z} \in \Sigma_s^N$,*

$$(1-\delta)\|\mathbf{z}\|_2^2 \leq \|\mathbf{A}\mathbf{z}\|_2^2 \leq (1+\delta)\|\mathbf{z}\|_2^2. \tag{2.7}$$

Remark 2.2.7. *There exists an alternative definition of s-RIP without squares in which* (2.7) *is replaced by*

$$(1-\delta)\|\mathbf{z}\|_2 \leq \|\mathbf{A}\mathbf{z}\|_2 \leq (1+\delta)\|\mathbf{z}\|_2. \tag{2.8}$$

Both definitions are equivalent up to a slight modification of the RIP constant as, for $0 < \delta < 1$, (2.7) *implies* (2.8) *and* (2.8) *with RIP constant $\delta/3$ implies* (2.7)*.*

Theorem 2.2.8. *If $\mathbf{A} \in \mathbb{R}^{m\times N}$ satisfies the $2s$-RIP with $\delta < 4/\sqrt{41} \approx 0.6$, $\mathbf{A}$ also satisfies the stable and robust NSP with constants $\rho = c_\delta$ and $\tau = \sqrt{s}c'_\delta$ where $0 < c_\delta < 1$ and $c'_\delta > 0$ only depend on δ.*

Theorem 2.2.8 is a consequence of [66, Theorem 6.13]. The idea behind Definition 2.2.6 is that $\mathbf{A}$ cannot preserve the whole geometry (distances and angles) of $\mathbb{R}^N$ if $m \ll N$.

Distinct points will fall together. If $\mathbf{A}$ satisfies an RIP, however, it acts almost like an isometry when restricted to sparse vectors. In other words, $\mathbf{A}$ preserves the geometry of Σ_s^N. It is straight-forward to verify that two different s-sparse vectors can be distinguished by their measurements if $\mathbf{A}$ has the $2s$-RIP.
In Theorem 2.2.8 a $2s$-RIP is used. There exist various other stability and robustness guarantees under assumption of ts-RIPs, for $t \geq 4/3$. Cai and Zhang showed in [26] that $\delta < \sqrt{(t-1)/t}$ is a sharp upper bound on the ts-RIP constant of $\mathbf{A}$ to guarantee exact recovery of all s-sparse signals in the noiseless case.

We have characterized matrices which will allow stable and robust recovery of signals in $\mathbb{R}^N$ in polynomial time. So far we ignored how many measurements are necessary to construct such matrices. Is it sufficient to have $m = 2s$ as in Lemma 2.1.1? Unfortunately, the answer is no. By relating the problem of stable recovery to Gelfand widths of ℓ_p-balls it has been shown that (cf. [34] and references therein), for some absolute constant $C > 0$,

$$m \geq Cs \log\left(\frac{eN}{s}\right) \tag{2.9}$$

linear measurements are necessary to guarantee stable estimates as (2.6). In contrast to Lemma 2.1.1 the ambient dimension N has now a mild influence. However, up to the log-factor the necessary number of measurements still scales linear in s. The condition in (2.9) is not only necessary but also sufficient for the existence of s-RIP matrices. The following result has been first derived in [31]. We report the proof presented in [13] as it is elementary and illustrates the main tools for deriving RIPs of random matrices.

Theorem 2.2.9. *Let $\mathbf{A} \in \mathbb{R}^{m\times N}$ have standard Gaussian iid entries $a_{i,j} \sim \mathcal{N}(0,1)$ and assume that (2.9) holds for $C = C'\delta^{-2}$ with $0 < \delta < 1$ and an absolute constant C'. Then with probability at least $1 - \exp(-c\delta^2 m)$, for some absolute constant $c > 0$, the matrix $\frac{1}{\sqrt{m}}\mathbf{A}$ satisfies the s-RIP with RIP constant δ.*

Proof: We show the alternative RIP in Remark 2.2.7. The claim follows by equivalence of both definitions. As (2.8) is invariant under ℓ_2-norm scaling we restrict the proof to $\|\mathbf{x}\|_2 = 1$. Note that, for all $\mathbf{z} \in \mathbb{R}^N$, the matrix $\frac{1}{\sqrt{m}}\mathbf{A}$ fulfills

$$\mathsf{Pr}\left[\left|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{z}\|_2^2\right| \geq \varepsilon \|\mathbf{z}\|_2^2\right] \leq 2e^{-cm\varepsilon^2}, \tag{2.10}$$

for some absolute constant $c > 0$ (see, e.g., [164, Proposition 5.16]). Let us fix a support set $S \subset [N]$ with $|S| = s$ and define the set $\Sigma_S := \{\mathbf{x} \in \mathbb{S}^{N-1} \colon \operatorname{supp}(\mathbf{x}) = S\}$. Choose a minimal $\delta/4$-cover $Q_S \subset \Sigma_S$, i.e., for all $\mathbf{x} \in \Sigma_S$ there exists $\mathbf{q} \in Q_S$ with $\|\mathbf{q}\|_2 = 1$ and $\|\mathbf{x} - \mathbf{q}\|_2 \leq \delta/4$. One can find such a cover of cardinality $|Q_S| \leq (12/\delta)^s$ (see [28, Section 3]). By setting $\varepsilon = \delta/2$ and applying a union bound to (2.10) we get with probability at least $1 - 2(12/\delta)^s \exp(-cm\delta^2/4)$ that

$$\left(1 - \frac{\delta}{2}\right)\|\mathbf{q}\|_2^2 \leq \left\|\frac{1}{\sqrt{m}}\mathbf{A}\mathbf{q}\right\|_2^2 \leq \left(1 + \frac{\delta}{2}\right)\|\mathbf{q}\|_2^2,$$

for all $\mathbf{q} \in Q_S$ which implies

$$\left(1 - \frac{\delta}{2}\right) \|\mathbf{q}\|_2 \leq \left\| \frac{1}{\sqrt{m}} \mathbf{A}\mathbf{q} \right\|_2 \leq \left(1 + \frac{\delta}{2}\right) \|\mathbf{q}\|_2 .$$

Define now $A \geq 0$ to be the smallest number such that

$$\left\| \frac{1}{\sqrt{m}} \mathbf{A}\mathbf{x} \right\|_2 \leq (1 + A) \|\mathbf{x}\|_2 ,$$

for all $\mathbf{x} \in \Sigma_S$. For any $\mathbf{x} \in \Sigma_S$, we can choose a $\mathbf{q} \in Q_S$ of minimal distance. As $\mathbf{x} - \mathbf{q} \in \Sigma_S$, we get

$$\left\| \frac{1}{\sqrt{m}} \mathbf{A}\mathbf{x} \right\|_2 \leq \left\| \frac{1}{\sqrt{m}} \mathbf{A}\mathbf{q} \right\|_2 + \left\| \frac{1}{\sqrt{m}} \mathbf{A}(\mathbf{x} - \mathbf{q}) \right\|_2 \leq 1 + \frac{\delta}{2} + (1 + A)\frac{\delta}{4}.$$

By minimality of A this implies $A \leq \delta/2 + (1 + A)\delta/4$ and, hence, $A \leq \delta$. The lower bound can be obtained in a similar way.
We have now proven the RIP on a fixed support set $S \subset [N]$ with probability at least $1 - 2(12/\delta)^s \exp(-cm\delta^2/4)$. To conclude it suffices to bound the number of possible supports by $\binom{N}{s} \leq (eN/s)^s$, apply a union bound, and use the assumption (2.9). ■

Theorem 2.2.8 shows that there exist RIP matrices for $m = \mathcal{O}(s \log(eN/s))$ and that they can be obtained with exceedingly high probability by drawing the matrix entries at random. Similar results have been derived as well for other distributions which follow concentration laws like (2.10), e.g., all sub-gaussian distributions (cf. [13]). As the $2s$-RIP implies stable and robust recovery of s-sparse signals in polynomial time, we have resolved all aforementioned issues of the ℓ_0-minimization in (2.2).
The proof of Theorem 2.2.8 reveals the main ingredients of proving an RIP for random matrices. One starts with a concentration inequality which is extended to all points in the signal set. In the above proof this happened by combining a elementary calculation with union bounds. The reader will rediscover this basic strategy in the proofs of the generalized RIPs of Chapter 4 and 7 as both rely on a covering of the signal set. Only the extension to the whole set is done by more sophisticated methods.
It is still an open problem to prove RIPs for deterministic matrices in the optimal measurement regime $m = \mathcal{O}(s \log(eN/s))$. The best obtained results need $m \geq Cs^2$ which is in terms of s substantially worse than (2.9).

As we initially were interested in practical signal recovery from compressed sensing measurements, Theorem 2.2.8 is still unsatisfactory. Matrices with subgaussian entries are almost surely dense. Consequently, they are expensive to store and matrix-vector calculations become time-consuming in high dimensions. Moreover, it is hard to construct measurement devices which correspond to such random matrices. Because of these shortcomings other types of measurement operators have been examined. Under slightly worse log-factors the RIP has been proven for randomly subsampled Fourier matrices in [31, 145] and partial random circulant matrices in [107]. These matrices can be efficiently stored and allow fast matrix-vector multiplications. Moreover, Fourier measurements naturally appear in many imaging applications.

All above results remain valid if stated for complex-valued signal vectors. As we restrict ourselves in the following to the real-valued case, we refrain from presenting them in full generality.

2.3 Recovery Methods

Basis pursuit de-noising is not the sole polynomial time method to solve (2.2) in a stable and robust way. Over the years, a vast amount of algorithms has been proposed to efficiently find sparse solutions to (2.1) and (2.4). They can be split into three main groups: greedy methods, e.g., the orthogonal matching pursuit (OMP) [44, 163], iterative thresholding algorithms, e.g., the iterative soft thresholding algorithm [39] and convex optimization approaches as basis pursuit. All groups exhibit different advantages and drawbacks. Roughly spoken, greedy methods, which search for global solutions by successively making locally optimal choices, are in general simple to implement and extremely fast but need a good stopping criterion to prevent them from overshooting. Iterative thresholding algorithms, which alternate between gradient descent and projection steps, are also simple to implement but their convergence rate might depend heavily on a suitable parameter choice. Optimization approaches, which combine data fidelity with regularization terms, are convenient to analyze but can become hard to solve efficiently for high-dimensional problems. In the rest of this chapter we present three popular methods which belong to the first two groups and will reappear in later chapters.

2.3.1 Orthogonal Matching Pursuit

As already mentioned, *orthogonal matching pursuit* (OMP) belongs to the group of greedy algorithms. It is popular for its speed, performance and simplicity and aims at approximating $\mathbf{x}$ in (2.4) by some s-sparse $\mathbf{x}_{\text{OMP}} \in \mathbb{R}^N$ where the desired support size s has to be known in advance as it defines the stopping criterion of OMP (cf. Algorithm 1).
The algorithm starts with the residuum $\mathbf{e} := \mathbf{y}$, the solution vector $\mathbf{x}_{\text{OMP}}^0 = \mathbf{0}$, and the support set $\Lambda^0 := \emptyset$ which will be greedily built in s steps. In each iteration step l one index is added to Λ^{l-1}, namely, the index maximizing the scalar product between the columns of $\mathbf{A}$ and the current residuum. After having enlarged the support, the target vector $\mathbf{x}_{\text{OMP}}^{l+1}$ is updated by a least squares fit and the new residuum is given by $\mathbf{e}^{l+1} = \mathbf{y} - \mathbf{A}\mathbf{x}_{\text{OMP}}^{l+1}$. After s steps OMP terminates.
If $\mathbf{A}$ satisfies the RIP, the following statements hold for OMP. Note that Theorem 2.3.1 requires more iterations for recovery than Theorem 2.3.2 but is based on less restrictive assumptions.

Theorem 2.3.1 ([66, Proposition 6.24]). *Let $\mathbf{A} \in \mathbb{R}^{m \times N}$ fulfill the RIP of order $13s$ and $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\eta}$, for some s-sparse $\mathbf{x} \in \mathbb{R}^N$ with $\Lambda = \operatorname{supp}(\mathbf{x})$ and $\boldsymbol{\eta} \in \mathbb{R}^m$. Let $(\mathbf{x}_{OMP}^l)_l$ denote the sequence defined by Algorithm 1 started at the index set $\Lambda^0 = \emptyset$. If the $13s$-RIP*

Algorithm 1 : OMP$(\mathbf{y}, \mathbf{A}, s)$

```
Require: $\mathbf{A} \in \mathbb{R}^{m \times N}$, $\mathbf{y} \in \mathbb{R}^m$, $s \in \mathbb{N}$
$\mathbf{e}^0 = \mathbf{y}$, $\mathbf{x}^0_{\text{OMP}} = \mathbf{0} \in \mathbb{R}^N$, $\Lambda^0 = \emptyset$, $l = 0$        ▷ initialize

while $l < s$ do
    $\mathbf{h}^l = \mathbf{A}^T \mathbf{e}^l$        ▷ match
    $\Lambda^{l+1} = \Lambda^l \cup \{\arg\max_j \mid \mathbf{h}^l(j) \mid\}$        ▷ identify
    $\mathbf{x}^{l+1}_{\text{OMP}} = \arg\min_{\mathbf{z}:\text{supp}(\mathbf{z}) \subset \Lambda^{l+1}} \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2$,    $\mathbf{e}^{l+1} = \mathbf{y} - \mathbf{A}\mathbf{x}^{l+1}_{\text{OMP}}$        ▷ update
    $l = l + 1$
end while

return $\mathbf{x}_{\text{OMP}} = \mathbf{x}^l_{\text{OMP}} = \arg\min_{\mathbf{z}:\text{supp}(\mathbf{z}) \subset \Lambda^l} \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2$
```

constant satisfies $\delta < 1/6$*, there is a constant* $C > 0$ *depending only on* δ*, such that*

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}^{12s}_{OMP}\|_2 \leq C\|\boldsymbol{\eta}\|_2.$$

Note that, if $\boldsymbol{\eta} = \mathbf{0}$*, this implies exact* s*-sparse recovery via OMP in* $12s$ *iterations.*

If the signal fulfills an additional decay condition and there is no noise on the measurements, one can guarantee recovery in s iterations.

Theorem 2.3.2 ([41, Theorem 4.1])**.** *Suppose that the matrix* $\mathbf{A} \in \mathbb{R}^{m \times N}$ *satisfies the RIP of order* $s + 1$ *with RIP constant* $\delta < \frac{1}{3}$*. Suppose* $\mathbf{x} \in \mathbb{R}^N$ *is* s*-sparse and for all* $j \in \{1, 2, ..., s - 1\}$ *it holds that*

$$\frac{r_j(\mathbf{x})}{r_{j+1}(\mathbf{x})} \geq \alpha,$$

where $r_j(\mathbf{x})$ *denotes the* j*-th largest entry of* $\mathbf{x}$ *in absolute value. If*

$$\alpha > \frac{1 + 2\frac{\delta}{1-\delta}\sqrt{k-1}}{1 - 2\frac{\delta}{1-\delta}},$$

then OMP recovers $\mathbf{x}$ *exactly from* $\mathbf{y} = \mathbf{A}\mathbf{x}$ *in* s *steps.*

In both cases it is important to stop OMP after a specific number of iterations which depends on the exact and, in general, unknown support size of $\mathbf{x}$. Moreover, without special structure of the signal, OMP is not able to provide guaranteed recovery in s steps (cf. [49, Theorem 7.3]).

2.3.2 Iterative Hard-Thresholding

By its name it is obvious that *iterative hard-thresholding* (IHT) belongs to the group of iterative thresholding algorithms. Similar to OMP it is a simple and efficient method which requires knowledge on the sparsity s of the signal to be recovered (cf. Algorithm 2).

Algorithm 2 : IHT$(\mathbf{y}, \mathbf{A}, s)$

```
Require: $\mathbf{A} \in \mathbb{R}^{m \times N}$, $\mathbf{y} \in \mathbb{R}^m$, $s \in \mathbb{N}$, number of iterations $L$
$\mathbf{x}^0 = \mathbf{0} \in \mathbb{R}^N$, $l = 0$                                   ▷ initialize

while $l < L$ do
    $\mathbf{x}^{l+1} = \mathbb{H}_s(\mathbf{x}^l + \mathbf{A}^T(\mathbf{A}\mathbf{x}^l - \mathbf{y}))$          ▷ gradient and hard thresholding step
    $l = l + 1$
end while

return $\mathbf{x}_{\mathrm{IHT}} = \mathbf{x}^L$
```

Algorithm 3 : ISTA$(\mathbf{y}, \mathbf{A}, \alpha, L)$

```
Require: $\mathbf{A} \in \mathbb{R}^{m \times N}$, $\mathbf{y} \in \mathbb{R}^m$, $\alpha > 0$, number of iterations $L$
$\mathbf{x}^0 = \mathbf{0} \in \mathbb{R}^N$, $l = 0$                                   ▷ initialize

while $l < L$ do
    $\mathbf{x}^{l+1} = \mathbb{S}_{\alpha/2}(\mathbf{x}^l + \mathbf{A}^T(\mathbf{A}\mathbf{x}^l - \mathbf{y}))$          ▷ gradient and soft thresholding step
    $l = l + 1$
end while

return $\mathbf{x}_{\mathrm{ISTA}} = \mathbf{x}^L$
```

The algorithm alternates between gradient descent steps of the least squared error function $\mathbf{z} \mapsto \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2$ and back-projections to the set Σ_s^N. The name IHT comes from the projection operator $\mathbb{H}_s$ which is called *hard-thresholding operator*. It only keeps the s in magnitude largest entries of a vector and sets the rest to zero.
As before the RIP of $\mathbf{A}$ can be used to deduce stable and robust recovery.

Theorem 2.3.3 ([66, Theorem 6.21])**.** *Suppose that $\mathbf{A} \in \mathbb{R}^{m \times N}$ satisfies the RIP of order $6s$ with RIP constant $\delta < 1/\sqrt{3} \approx 0.6$. Then, for $\mathbf{x} \in \mathbb{R}^N$ and $\boldsymbol{\eta} \in \mathbb{R}^m$, the sequence $\mathbf{x}^l$ defined by Algorithm 2 with $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\eta}$, $\mathbf{x}^0 = \mathbf{0}$, and s replaced by $2s$ satisfies, for any $l \geq 0$,*

$$\left\|\mathbf{x} - \mathbf{x}^l\right\|_1 \leq C\sigma_s(\mathbf{x})_1 + D\sqrt{s}\,\|\boldsymbol{\eta}\|_2 + 2\rho^l\sqrt{s}\,\|\mathbf{x}\|_2\,,$$

where the constants $C, D > 0$ and $0 < \rho < 1$ depend only on δ.

The error bound in Theorem 2.3.3 is similar to (2.6) if one replaces the NSP assumption by an RIP assumption and uses Theorem 2.2.8. The difference lies within the additional error term $2\rho^l\sqrt{s}\,\|\mathbf{x}\|_2$ which linearly converges to 0. Moreover, the result does not guarantee convergence of the sequence $\mathbf{x}^l$.

2.3.3 Iterative Soft-Thresholding and LASSO

With the *iterative soft-thresholding algorithm* (ISTA) we have another representative of the thresholding algorithms. We will see, however, that ISTA is also closely related to a

convex relaxation approach which provides additional tools for analysis. In contrast to IHT, ISTA does not require the signal's sparsity as input. Instead there is a free parameter $\alpha > 0$ to choose, which controls sparsity of the approximation (cf. Algorithm 3).
ISTA is a proximal gradient descent algorithm, i.e., it approximates the minimizer of a convex but partially non-smooth functional by alternating between gradient descent steps of the smooth component and proximal mappings of the non-smooth component (for further details on proximal mappings and proximal gradient descent refer to [136]). Similar to IHT the gradient descent step is computed for $\mathbf{z} \mapsto \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2$. Proximal mapping of ISTA is the *soft-thresholding operator*

$$(\mathbb{S}_\alpha(\mathbf{z}))_i = \begin{cases} z_i - \alpha & z_i > \alpha \\ 0 & |z_i| \leq \alpha \\ z_i + \alpha & z_i < -\alpha, \end{cases} \tag{2.11}$$

which acts component-wise on vectors and motivates the algorithm's name. The soft-thresholding operator iteratively shrinks all components to zero. One might view it as a smoothened version of the hard-thresholding operator. As $\mathbb{S}_{\alpha/2}$ is the proximal mapping of $\mathbf{z} \mapsto \alpha \|\mathbf{z}\|_1$, the underlying optimization problem of ISTA is given by

$$\min_{\mathbf{z} \in \mathbb{R}^N} \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2 + \alpha \|\mathbf{z}\|_1 , \tag{2.12}$$

which is commonly known as *least absolute shrinkage and selection operator* (LASSO) in statistics and is, for suitable choice of $\alpha > 0$, equivalent to the basis pursuit denoising in (2.5), see [66, Proposition 3.2]. This can be used to transfer theoretical results obtained for (2.5) to ISTA. (As proximal mappings are generalized projections, IHT may also be interpreted as proximal gradient descent algorithm but for a non-convex problem.)
The analysis of ISTA can be splitted in two independent parts. First, convergence of the iterates is established under mild assumptions on $\mathbf{A}$. Second, by equivalence of (2.12) and (2.5) we can use Theorem 2.2.5 to bound the worst case distance between minimizers of (2.12) and the original signal $\mathbf{x}$ depending on sparsity defect and noise level.

Theorem 2.3.4. *Let $\mathbf{A} \in \mathbb{R}^{m \times N}$ and $\mathbf{y} \in \mathbb{R}^m$. If $\|\mathbf{A}\|_{2\to 2} < \sqrt{2}$, the sequence of iterates $\mathbf{x}^l$ in Algorithm 3 converges to a minimizer of* (2.12).

The convergence of Algorithm 3 has been established in [39] for $\|\mathbf{A}\|_{2\to 2} < 1$. In [36] the authors could relax the condition $\|\mathbf{A}\|_{2\to 2} < 1$ to $\|\mathbf{A}\|_{2\to 2} < \sqrt{2}$. In the end, the assumption on $\|\mathbf{A}\|_{2\to 2}$ is always fulfilled by a proper rescaling of (2.12). If one replaces $\mathbf{A}$, $\mathbf{y}$, and α by $\mathbf{A}/\|\mathbf{A}\|_{2\to 2}$, $\mathbf{y}/\|\mathbf{A}\|_{2\to 2}$, and $\alpha/\|\mathbf{A}\|_{2\to 2}^2$, the minimizers of (2.12) do not change but Theorem 2.3.4 applies.

Theorem 2.3.5. *Suppose that the $2s$-th RIP constant of $\mathbf{A} \in \mathbb{R}^{m \times N}$ satisfies $\delta < 4/\sqrt{41} \approx 0.6$. Then, for any $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^m$ with $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \eta$ the following holds. Denote by $\mathbf{x}_\alpha$ a minimizer of* (2.12). *If $\alpha > 0$ such that $\eta_\alpha := \|\mathbf{y} - \mathbf{A}\mathbf{x}_\alpha\|_2 \geq \eta$, one has*

$$\|\mathbf{x} - \mathbf{x}_\alpha\|_1 \leq C\sigma_s(\mathbf{x})_1 + D\sqrt{s}\eta_\alpha,$$

for constants C and D depending on the RIP constant δ of $\mathbf{A}$.

Proof: Theorem 2.3.5 is a re-statement of Theorem 2.2.5 where we use Theorem 2.2.8 and that by [66, Proposition 3.2] the minimizer $\mathbf{x}_\alpha$ is a solution to (2.5) with η replaced by η_α. ■

Under suitable choice of α (such that $\eta_\alpha = \eta$), Theorem 2.3.5 provides stable and robust approximation guarantees which are similar to the ones in Theorem 2.2.5. In combination with Theorem 2.3.4 they apply to ISTA.

A notable property of ISTA is that it does not require an RIP of $\mathbf{A}$ to converge. Even if the measurements do not allow unique identification of sparse signals, ISTA still produces solutions which have a small error in measurements and small ℓ_1-norm (we further elaborate on the connection of small ℓ_1-norm and sparsity in Section 3.2).

We conclude by illustrating how α regulates the trade-off between data fidelity and ℓ_1-norm/sparsity. Let $\mathbf{x} \in \mathbb{R}^N$ with noisy measurements $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\eta}$ be the signal and let $\mathbf{x}_\alpha$ be the minimizer of (2.12). Then,

$$\begin{aligned}\|\mathbf{y} - \mathbf{A}\mathbf{x}_\alpha\|_2^2 &\le \|\mathbf{y} - \mathbf{A}\mathbf{x}_\alpha\|_2^2 + \alpha\,\|\mathbf{x}_\alpha\|_1 \\ &\le \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \alpha\,\|\mathbf{x}\|_1 = \|\boldsymbol{\eta}\|_2^2 + \alpha\,\|\mathbf{x}\|_1\end{aligned}$$

and

$$\begin{aligned}\|\mathbf{x}_\alpha\|_1 &\le \frac{1}{\alpha}\left(\|\mathbf{y} - \mathbf{A}\mathbf{x}_\alpha\|_2^2 + \alpha\,\|\mathbf{x}_\alpha\|_1\right) \\ &\le \frac{1}{\alpha}\left(\|\mathbf{y} - \mathbf{A}\cdot\mathbf{0}\|_2^2 + \alpha\,\|\mathbf{0}\|_1\right) = \frac{\|\mathbf{y}\|_2^2}{\alpha}\end{aligned}$$

by using $\mathbf{x}$ and $\mathbf{0}$ as competitors in (2.12). Hence, a small α promotes data fidelity and a large α small ℓ_1-norm/sparsity. We discuss parameter choice strategies for ISTA in Appendix A.

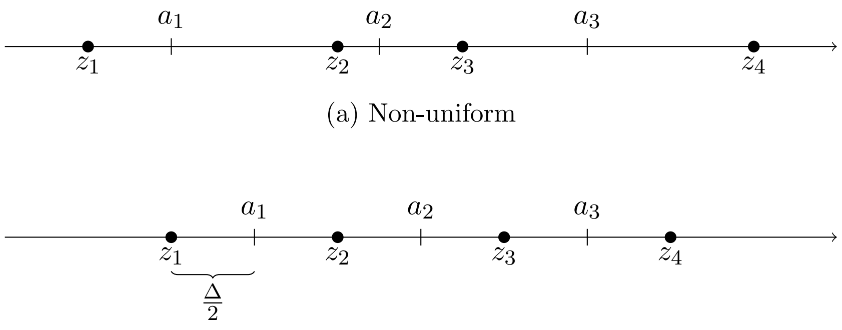
Chapter 3

Quantized Compressed Sensing

In this chapter we explore limitations of the classical compressed sensing model which has been presented in the last chapter. We start with a short introduction on quantization and modify the measurement model to better reflect reality. By generalizing the notion of sparsity we build a bridge to more general signal sets and probabilistic definitions of dimension and complexity. The last two sections detail one-bit and multi-bit quantization in compressed sensing and discuss several important recent results.

3.1 A Measurement Model Meets the Real World

The last chapter showed that s-sparse signals can be recovered in a stable and robust way from few linear measurements. As already indicated in Chapter 1, the measurement model in (2.1) does not directly apply to real-world applications. The assumption of having real (or complex) valued measurements does not incorporate the limitation of finite data storage. To illustrate this point consider a single real number $r \in \mathbb{R}$ which shall be saved to a computer. The computer does not memorize r but rounds to a fixed number of digits, i.e., it quantizes r to a finite subset of $\mathbb{Q}$. By dictionary definition, quantization is the division of a quantity into small but measurable increments (Merriam-Webster). We define a quantizer $Q\colon \mathcal{Z} \to \mathcal{F}$ to be a function mapping a continuous and infinite set $\mathcal{Z}$ to a discrete and finite alphabet $\mathcal{F} = \{\mathbf{z}_i \colon i \in [n]\} \subset \mathcal{Z}$, where $n \in \mathbb{N}$ and the $\mathbf{z}_i$ are called *quantization values.* We call $R = \log_2(n)$ the *rate* of Q such that an R-bit quantizer has 2^R quantization values.
If $\mathcal{Z} \subset \mathbb{R}$ the quantizer Q is normally defined by a set of intervals $S_i = (a_{i-1}, a_i] \subset \mathbb{R}$ such that $Q(r) = z_i$ if and only if $r \in S_i$. The interval limits $a_0 < ... < a_n$ are often called *quantization thresholds.* A quantizer is called *uniform* if all finite S_i have the same length and $z_i = (a_i - a_{i-1})/2$. The thresholds a_1 and a_{n-1} define the quantizer's *range.* We say that Q *saturates* outside its range, i.e., if $Q(r) \in \{z_1, z_n\}$ and $\mathcal{Z}$ is unbounded we have no information on the quantization error $|Q(r) - r|$. A uniform quantizer is fully determined by its range and $\Delta := |S_i|$. Figure 3.1 shows two examples of one-dimensional quantizers. It is straight-forward to generalize those concepts to the higher dimensional case $\mathcal{Z} \subset \mathbb{R}^D$ by replacing the intervals S_i by connected subsets of $\mathbb{R}^D$.



(a) Non-uniform

(b) Uniform

Figure 3.1: One-dimensional quantizers with $a_0 = -\infty$ and $a_4 = \infty$

A key quantity of a quantizer Q is the *worst-case distortion*

$$\epsilon := \sup_{\mathbf{z} \in \mathcal{Z}} \|\mathbf{z} - Q(\mathbf{z})\|_2 \tag{3.1}$$

which measures the worst-case error caused by applying Q. For given rate R a quantizer is optimal if the choice of $\mathcal{F} \subset \mathcal{Z}$ and the mapping Q are minimizing ϵ in (3.1). Obviously, $\epsilon = \infty$ for all finite quantizer if $\mathcal{Z}$ is unbounded. If $\mathcal{Z} \subset \mathbb{R}^D$, a simple volume argument (cf. [22]) shows that

$$\epsilon \gtrsim 2^{-\frac{R}{D}}. \tag{3.2}$$

The best one can hope for is hence an exponential decay of quantization error in the rate. For a detailed survey on the history of quantization, which dates back to Shannon's fundamental work [150, 151, 134] on information theory in 1948 and the following years, and a collection of important results on quantizer design and quantization methods refer to [77].

We will concentrate in this work on *uniform scalar quantizers*. Scalar quantization is a straight-forward approach to design quantizers in $\mathbb{R}^D$. Instead of quantizing the whole vector $\mathbf{z} \in \mathbb{R}^D$, one defines a one-dimensional quantizer Q and applies it independently to each component of $\mathbf{z}$. We call Q an B-bit quantizer if its rate per entry is B which corresponds to an overall rate $R = DB$. If Q is uniform, we speak of *uniform* scalar quantization. It has been shown that uniform scalar quantizers approach optimality when the bit-rate increases (see [77]).

Let us modify (2.1) by introducing a uniform scalar quantizer $Q \colon \mathbb{R}^m \to \mathbb{R}^m$. The quantized compressed sensing model hence reads

$$\mathbf{y} = Q(\mathbf{A}\mathbf{x}). \tag{3.3}$$

As Q neither has to be linear nor even continuous, we loose several favorable properties of the measurement process. Moreover, in addition to the number of measurements m we have the bit-rate per entry B of Q as a free parameter which influences reconstruction quality. At first sight, the role of both parameters seems clear. Increasing m leads

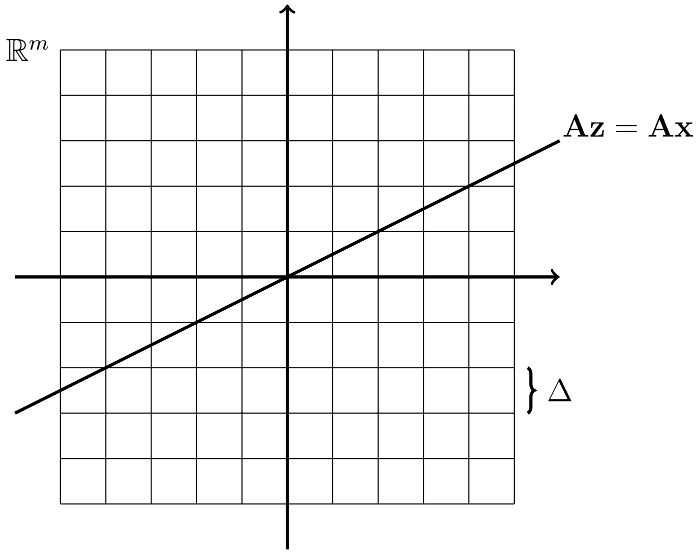


Figure 3.2: Only a small fraction of the quantization cells is used by the measurements.

to stronger compressed sensing guarantees while increasing B reduces measurement perturbations. But the interplay of both quantities has to be considered as well (cf. [24]). Assume for simplicity that $\mathbf{x}$, instead of being s-sparse, lies on an s-dimensional subspace U of $\mathbb{R}^N$. In this case $\mathbf{A}U$ is an s-dimensional subspace of $\mathbb{R}^m$. As m is typically a multiple of s due to oversampling of the intrinsic dimension of $\mathbf{x}$, the measurements of signals in U only use a small fraction of the quantization cells provided by Q (see Figure 3.2). When transferring this fact to the set Σ_s^N which consists of $\binom{n}{s}$ different s-dimensional subspaces, one can show (see [22]) that for an B-bit quantizer Q the number $I_{s,m,B}$ of quantization cells which are intersected by $\mathbf{A}\Sigma_s^N$ is bounded by

$$I_{s,m,B} \lesssim \left(\frac{2^B mn}{s^2} \right)^s .$$

Consequently, a volume argument similar to the one used for obtaining (3.2) shows (see [22]) that the worst-case distortion of $\mathcal{Z} = \mathbf{A}\Sigma_s^N$ can be lower bounded by

$$\epsilon \gtrsim \frac{2^{-B} s}{m} = \frac{2^{-\frac{R}{m}} s}{m}. \tag{3.4}$$

Enhancing stability and robustness of reconstruction by increasing the number of measurements hence degrades the exponential error decay in the overall bit-rate R of Q. Moreover, the lower bound in (3.4) shows that for a fixed one-dimensional bit-rate B the quantization resolution on $\mathbf{A}\Sigma_s^N$ (and with it the worst-case reconstruction error) decay at most linearly in m.

When it comes to reconstructing signals from their quantized measurements in (3.3), one could treat the quantization distortion as additive bounded noise. In this case (3.3) is replaced by (2.4) with $\boldsymbol{\eta} = Q(\mathbf{Ax}) - \mathbf{Ax}$. If we assume for simplicity that Q is uniform with quantization intervalls of length Δ and that the range of Q is sufficiently large, the noise is bounded by $\|\boldsymbol{\eta}\|_2 \leq \sqrt{m}\,\|Q(\mathbf{Ax}) - \mathbf{Ax}\|_\infty \leq \sqrt{m}\frac{\Delta}{2}$. To recover one applies (2.5) and uses the stability and robustness results presented in Chapter 2, cf. [22, 26]. This approach, however, suffers certain drawbacks as mentioned in [97]. Most important it does not guarantee measurement consistency, i.e., the ℓ_2-constraint of (2.5) does not guarantee that $\|Q(\mathbf{Az}) - \mathbf{Az}\|_\infty \leq \frac{\Delta}{2}$, which means a loss of information in the recovery process. Already the seminal work on compressed sensing [31] suggested to enforce measurement consistency when dealing with quantized measurements. Since then, it has been shown that measurement consistency helps in reaching the lower bound in (3.4), cf. [141, 73]. We will discuss in this chapter several recovery algorithms which are tailored to the measurement model (3.3).

There are two ways to deal with saturation in compressed sensing: one either assumes that the quantizer's range is sufficiently large to cover the measurements of all signals of interest or one mistrusts and dismisses any measurements which are quantized to the boundary of the range. In the first case, the quantizer can be assumed to be infinite and is, if working with uniform quantizers, only characterized by Δ, e.g., [94]. In the second case, one uses democracy of the measurement matrix $\mathbf{A}$ to rely only on parts of the measurements. Democracy is a property which guarantees RIP even for submatrices of $\mathbf{A}$ (cf. [111, 27, 79]) and is with high probability fulfilled by Gaussian random matrices, see [42]. Note that in the one-bit setting we mainly focus on in later chapters, saturation is not an issue by coarseness of the quantization (there are only two quantization intervalls in each measurement which are both of infinite length).

A more sophisticated approach to quantization in compressed sensing, we only like to mention here, are so-called feedback quantizers which recursively compute the bit sequence encoding the measurements. This line of work originated from Sigma-Delta modulation of bandlimited signals [76, 133] and frame expansions [18, 17] in the sparse recovery framework. In [80], they introduced and analyzed such an approach for Gaussian measurements and subsequent works generalized the results to subgaussian random measurements [106, 61]. Recovery guarantees for subgaussian measurements based on convex optimization were proven in [147] and extended to partial random circulant matrices in [62]. As can be seen in [15], feedback quantizers are able to exceed the linear decay in m of (3.4). For further details we refer the reader to the overview chapter [22].

3.2 General Signal Sets

Before discussing one-bit and multi-bit compressed sensing in detail we take a closer look at our signal model. So far we only considered the signal set $\Sigma_s^N \subset \mathbb{R}^N$ and used its intrinsic low-dimensionality to justify (almost) loss-less compression into lower dimensional spaces $\mathbb{R}^m$. We already saw how restrictive it is to assume sparsity of signals and extended recovery results to compressible vectors. Moreover, as mentioned in Chapter 1, one normally needs a suitable transform to sparsify the representation of signals. This transformation or sparsifying dictionary is in general hard to find; if existent at all. It

might even be possible that the signals of interest lie on low-dimensional sub-manifolds of the ambient space and do not fit the linear subspace framework of sparse vectors.
Several works [138, 21, 65] replaced Σ_s^N by the set of *effectively* sparse vectors containing vectors which are not necessarily sparse but lie close to sparse vectors [138, Lemma 3.2].

Definition 3.2.1 (Effectively Sparse Vectors). *Let*

$$K_{N,s} = \{\mathbf{z} \in \mathbb{R}^N \colon \; \|\mathbf{z}\|_2 \leq 1, \|\mathbf{z}\|_1 \leq \sqrt{s}\} = \mathcal{B}_2(\mathbf{0},1) \cap \mathcal{B}_1(\mathbf{0},\sqrt{s}).$$

We call all $\mathbf{z} \in \mathbb{R}^N$ *with* $\|\mathbf{z}\|_1 / \|\mathbf{z}\|_2 \in K_{N,s}$ *effectively* s*-sparse.*

Note that $\Sigma_s^N \cap \mathcal{B}_2(\mathbf{0},1) \subset K_{N,s}$ as $\|\mathbf{z}\|_1 \leq \sqrt{s}\, \|\mathbf{z}\|_2$ for all $\mathbf{z} \in \Sigma_s^N$ and that all s-sparse vectors are effectively s-sparse as well. By [138, Lemma 3.1] one has

$$\operatorname{conv}(\Sigma_s^N \cap \mathcal{B}_2(\mathbf{0},1)) \subset K_{N,s} \subset 2 \; \operatorname{conv}(\Sigma_s^N \cap \mathcal{B}_2(\mathbf{0},1)), \tag{3.5}$$

that is the set $K_{N,s}$ can be interpreted as the convex hull of $\Sigma_s^N \cap \mathcal{B}_2(\mathbf{0},1)$. Plan and Vershynin showed in [138] that the covering numbers of $\Sigma_s^N \cap \mathcal{B}_2(\mathbf{0},1)$ and $K_{N,s}$ are equivalent in dependence on s and N.

Lemma 3.2.2 ([138, Lemma 3.3 & 3.4]). *For* $\varepsilon \in (0,1)$ *and* $s \leq N$ *we have*

$$\log\left(\mathcal{N}(\Sigma_s^N \cap \mathcal{B}_2(\mathbf{0},1), \varepsilon)\right) \leq s \log\left(\frac{3N}{\varepsilon s}\right)$$

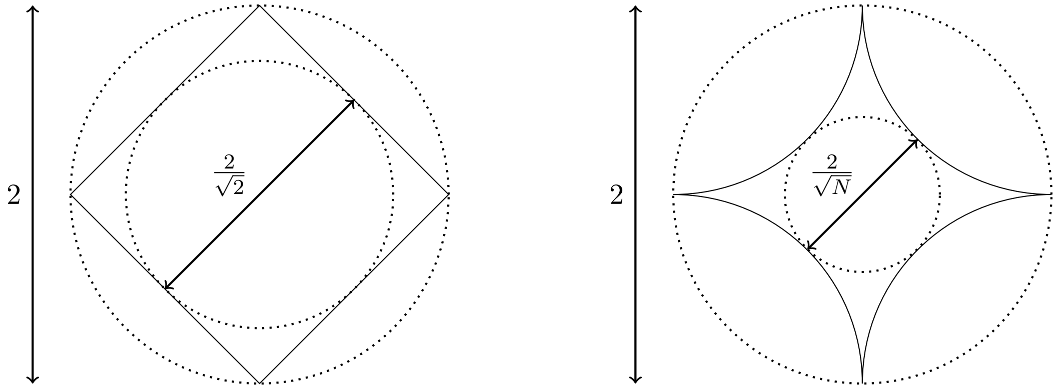
and

$$\begin{aligned} \log\left(\mathcal{N}(K_{N,s}, \varepsilon)\right) &\leq \begin{cases} N \log\left(\frac{6}{\varepsilon}\right), & 0 < \varepsilon < 2\sqrt{\frac{s}{N}}, \\ \frac{4s}{\varepsilon^2} \log\left(\frac{9\varepsilon N}{s}\right), & \textit{else} \end{cases} \\ &\lesssim \frac{s}{\varepsilon^2} \log\left(\frac{2N}{s}\right). \end{aligned}$$

They used this equivalence to extend their results to effectively sparse signals while keeping the sufficient number of measurements linear in s up to log-factors. It is a crucial observation that the logarithm of the covering number directly relates to the number of linear measurements needed for compressing the sets Σ_s^N and $K_{N,s}$ (and hence to their intrinsic dimension resp. complexity). To elaborate on this relation we have to understand the geometry of convex sets in high dimensions.

3.2.1 Convex Sets in High Dimensions and the Gaussian Mean Width

Phenomena which seem counter-intuitive at first sight are common in infinite-dimensional vector spaces. For instance, the existence of non-continuous linear functions. There are, however, similar counter-intuitive effects in high- but finite-dimensional spaces. A beautiful example is the geometry of convex sets which we illustrate by means of the ℓ_1-ball $\mathcal{B}_1(\mathbf{0},1) \subset \mathbb{R}^N$. Accustomed to three dimensions at most, our idea of $\mathcal{B}_1(\mathbf{0},1)$ is two-dimensional as depicted in Figure 3.3 (a).
When moving to higher dimensions the spread of volume in space changes. This can be seen by comparing (a) and (b) in Figure 3.3. While the outer ℓ_2-ball depending on



(a) $\mathbb{R}^2$

(b) $\mathbb{R}^N$

Figure 3.3: Geometry of $\mathcal{B}_1(\mathbf{0},1)$ in different dimensions

the maximum width of $\mathcal{B}_1(\mathbf{0},1)$ remains unchanged, the inner ℓ_2-ball carrying most of the volume mass of $\mathcal{B}_1(\mathbf{0},1)$ shrinks to zero (as noted in [165], one can easily check that $\mathrm{Vol}(\mathcal{B}_1(\mathbf{0},1))^{\frac{1}{N}} \simeq \mathrm{Vol}(\mathcal{B}_2(\mathbf{0},1/\sqrt{N}))^{\frac{1}{N}} \simeq \frac{1}{N}$). This observation can be transferred to general convex sets and heuristically stated as follows: *in high-dimensional spaces, convex sets consist of bulks, which are small in diameter but contain most of the volume, and outliers, which have almost no volume but reach out far into space*, cf. [127, 165]. Note that the non-convex looking shape of Figure 3.3 (b) is not contradicting convexity of $\mathcal{B}_1(\mathbf{0},1)$. It just emphasizes the special topological structure of high-dimensional spaces which approaches as a limit the structure of sequence spaces. (In the space $\mathbb{R}^{|\mathbb{N}|}$ of real-valued sequences, any scaled unit vector $\varepsilon\mathbf{e}_i$, for $\varepsilon \in (0,1)$ and $i \in \mathbb{N}$, lies within the unit ℓ_2-ball while any scaled vector pointing into space $(\varepsilon,\varepsilon,...)$, for $\varepsilon > 0$, lies outside of any ℓ_2-ball.)
Those rather informal considerations are supported by rigourous mathematical results. In [78] the authors show that the volume of isotropic convex bodies concentrates around the sphere of an ℓ_2-ball. As all convex sets are isotropic up to an invertible linear transformation, this implies for general convex sets that the volume is mainly concentrated around the boundary of an ellipsoid. Dvoretzky's theorem [53, 52] characterizes the shape of random cuts of low-dimensional subspaces through certain convex bodies like the ℓ_1-ball with high probability as ℓ_2-balls. The intuition behind Dvoretzky's theorem is that low-dimensional random subspaces miss with high-probability the outliers of convex sets and only detect the bulk which carries the mass.

In order to capture the complexity of a general set $K \subset \mathbb{R}^N$ it, hence, might be interesting to look at intersections of K with randomly oriented low-dimensional subspaces. We define for $\mathbf{u} \sim \mathcal{U}(\mathbb{S}^{N-1})$ the *spherical mean width* of a set $K \subset \mathbb{R}^N$

$$\tilde{w}(K-K) := \mathsf{E}\left[\sup_{\mathbf{z}\in K-K}\langle \mathbf{u},\mathbf{z}\rangle\right].$$

As illustrated in Figure 3.4, the inner term $\sup_{\mathbf{z}\in K-K}\langle \mathbf{u},\mathbf{z}\rangle = \sup_{\mathbf{z}_1,\mathbf{z}_2\in K}\langle \mathbf{u},\mathbf{z}_1-\mathbf{z}_2\rangle$ measures the maximum width of K in direction $\mathbf{u}$. By picking $\mathbf{u} \in \mathbb{S}^{N-1}$ uniformly at

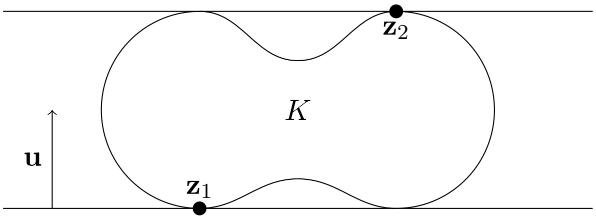


Figure 3.4: The width of a set K in direction $\mathbf{u}$.

random and averaging over all choices, the spherical mean width measures an averaged one-dimensional cut length through K. This definition has a drawback. The entries of $\mathbf{u}$ are not independent. It is convenient to replace $\mathbf{u}$ by a Gaussian vector $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{Id}_N)$.

Definition 3.2.3 (Gaussian mean width, [165])**.** *Let* $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{Id}_N)$ *be a standard Gaussian vector with iid entries. The* Gaussian mean width *of a set* $K \subset \mathbb{R}^N$ *is defined as*

$$w(K-K) := \mathsf{E}\left[\sup_{\mathbf{z} \in K-K} \langle \mathbf{g}, \mathbf{z} \rangle \right].$$

Remark 3.2.4. *If* $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{Id}_N)$*, the vector* $\mathbf{g}/\left\|\mathbf{g}\right\|_2$ *is uniformly distributed on the unit sphere. As* $\left\|\mathbf{g}\right\|_2$ *and* $\mathbf{g}/\left\|\mathbf{g}\right\|_2$ *are independent and* $\mathsf{E}[\left\|\mathbf{g}\right\|_2] \simeq \sqrt{N}$*, we obtain*

$$w(K-K) = \mathsf{E}[\left\|\mathbf{g}\right\|_2] \, \tilde{w}(K-K) \simeq \sqrt{N}\tilde{w}(K-K),$$

that is w *and* $\tilde{w}$ *are equivalent up to* $\sqrt{N}$*, cf. [165].*

In the following, we examine if $w(K-K)$ is a suitable measure of complexity for $K \in \mathbb{R}^N$, collect several properties of $w(K-K)$ including alternative but basically equivalent definitions, and analyze the relation between $w(K-K)$ and $N(K, \varepsilon)$. For a more detailed introduction on the geometry of convex sets and further discussion of the above mentioned results refer to [10, 69].

3.2.2 Properties of the Gaussian Mean Width

It is clear from Definition 3.2.3 that the Gaussian mean width $w(K-K)$ is invariant under translations and orthogonal transformations of K. In these points it behaves similar to the linear dimension of subspaces. In contrast to the linear dimension, the Gaussian mean width scales with the diameter of K, i.e., for $\alpha > 0$ one has $w(\alpha(K-K)) = \alpha w(K-K)$. When comparing the width of different sets we thus have to take their scaling into account. Let's have a look at some concrete examples, cf. [139].

Lemma 3.2.5. *The following bounds hold:*

(i) If $K = \mathcal{B}_2(\mathbf{0}, 1) \subset \mathbb{R}^N$*, we have that* $w(K-K) \leq 2\sqrt{N}$

(ii) If $L \subset \mathbb{R}^N$ *is a subspace with* $\dim(L) = d$ *and* $K = L \cap \mathcal{B}_2(\mathbf{0}, 1)$*, we have that* $w(K-K) \leq 2\sqrt{d}$

(iii) If $K \subset \mathbb{R}^N$ *is a finite set, we have that* $w(K-K) \lesssim \operatorname{diam}(K)\sqrt{\log(|K|)}$*.*

Proof: The first statement follows from $w(\mathcal{B}_2(\mathbf{0},1)-\mathcal{B}_2(\mathbf{0},1)) = 2\,\mathsf{E}[\|\mathbf{g}\|_2] \le 2\sqrt{N}$. To show the second statement first note that without loss of generality we can set L to be the subspace spanned by the first d canonical basis vectors which restricts the inner product $\langle \mathbf{g},\mathbf{z}\rangle$ to the first d entries of $\mathbf{g}$. Now use the argumentation of (i) where N is replaced by d. The third statement is by

$$w(K-K) \le \operatorname{diam}(K)\,\mathsf{E}\left[\max_{\mathbf{z}\in K-K}\langle \mathbf{g},\frac{\mathbf{z}}{\|\mathbf{z}\|_2}\rangle\right] \lesssim \operatorname{diam}(K)\sqrt{\log(|K|)}$$

the direct consequence of a standard bound [66, Proposition 8.1] on the expected maximum of $|K|(|K|-1)$ standard Gaussians. ■

Lemma 3.2.5 confirms that $w(K-K)^2$ is up to constants a valid generalization of the linear dimension to arbitrary sets. Point (iii) is especially interesting in view of the Johnson-Lindenstrauss lemma and its extensions [100, 1] which state that n points in $\mathbb{R}^N$ can be embedded linearly and almost isometrically into $\mathbb{R}^m$ if $m \gtrsim \log(n)$. Hence, up to a constant $w(K-K)^2$ exactly describes the number of linear measurements which are sufficient to guarantee compression of finite sets into low dimensions.
In the literature several variations of Gaussian width appeared which are basically equivalent. As the results we are going to use depend on different definitions, we present and relate them to each other in the following lemma.

Lemma 3.2.6 ([91, Definition 4.1])**.** *Let* $\mathbf{g} \sim \mathcal{N}(\mathbf{0},\mathbf{Id}_N)$. *For a subset* $K \subset \mathbb{R}^N$ *define*

(i) the Gaussian width*:* $w(K) := \mathsf{E}[\sup_{\mathbf{x}\in K}\langle \mathbf{g},\mathbf{x}\rangle]$

(ii) the Gaussian complexity*:* $\gamma(K) = \mathsf{E}[\sup_{\mathbf{x}\in K}|\langle \mathbf{g},\mathbf{x}\rangle|]$.

By combining Properties 5. and 6. of Proposition 2.1 in [139] on has

$$w(K-K) \le 2w(K) \le 2\gamma(K) \le 2\left(w(K-K)+\sqrt{\frac{2}{\pi}}\operatorname{dist}(\mathbf{0},K)\right), \tag{3.6}$$

where $\operatorname{dist}(\mathbf{0},K) = \inf_{\mathbf{z}\in K}\|\mathbf{z}\|_2$.

If restricted to a bounded region of $\mathbb{R}^N$, Gaussian width and Gaussian complexity are up to multiplicative and additive constants equivalent to the Gaussian mean width by (3.6). The Gaussian width has several useful properties in addition to ones named above.

Lemma 3.2.7 ([166, Proposition 7.5.2])**.** *Let* $K \subset \mathbb{R}^N$. *Then,*

(i) the Gaussian width is finite if and only if K is bounded and, for $\alpha > 0$,

$$w(\alpha K) = \alpha w(K),$$

(ii) the Gaussian width is invariant under affine orthogonal transformations, i.e., for any orthogonal matrix $\mathbf{U} \in \mathbb{R}^{N\times N}$ *and* $\mathbf{b} \in \mathbb{R}^N$,

$$w(\mathbf{U}K+\mathbf{b}) = w(K),$$

(iii) the Gaussian width is invariant under taking the convex hull, i.e.,

$$w(\operatorname{conv}(K)) = w(K),$$

(iv) the Gaussian width fulfills

$$0 \leq \frac{1}{\sqrt{2\pi}} \operatorname{diam}(K) \leq w(K) \leq \frac{\sqrt{N}}{2} \operatorname{diam}(K).$$

In the beginning of Section 3.2, we observed a connection between the sufficient number of linear measurements for almost lossless compression of the sets Σ_s^N and $K_{N,s}$ to lower dimensions and the covering numbers of Σ_s^N and $K_{N,s}$. Moreover, Lemma 3.2.5 showed a connection between the sufficient number of linear measurements for almost lossless compression of finite sets $K \subset \mathbb{R}^N$ to lower dimensions and the Gaussian mean width. There is a close relation between the covering number $N(K, \varepsilon)$ of a set K and its Gaussian width $w(K)$.

Theorem 3.2.8 (Sudakov's minoration and Dudley's inequality)**.** *Let $K \subset \mathbb{R}^N$ be bounded. Then,*

$$\sup_{\varepsilon > 0} \varepsilon \sqrt{\log\left(N(K,\varepsilon)\right)} \lesssim w(K) \lesssim \int_0^\infty \sqrt{\log\left(N(K,\varepsilon)\right)}\, \mathrm{d}\varepsilon. \tag{3.7}$$

The lower bound is called Sudakov minoration *[114, Theorem 3.18], the upper bound is called* Dudley's inequality *[114, Theorem 11.17].*

Remark 3.2.9. *As $N(K, \varepsilon) = 1$ for $\varepsilon \geq \operatorname{diam}(K)$, integral's upper boundary may be replaced by* $\operatorname{diam}(K)$*.*

Sudakov minoration and Dudley's inequality are in their original form more general bounds on the expected suprema of Gaussian processes. Theorem 3.2.8 states them for the special case of $w(K)$, a quite simple Gaussian process. The proofs are based on the theory of stochastic processes and are well-presented in [166]. For Sudakov's minoration inequality one compares $w(K)$ to an even simpler Gaussian process whose expected supremum can be controlled by the left hand side of (3.7). For Dudley's inequality one uses a chaining argument which controlls the supremum on K by a chain of coverings which refine in a dyadic way.

The bounds in (3.7) are not tight. Sophisticated methods like generic chaining [157, 158] produce tight but more complicated bounds. Comparison theorems [114, Section 3.3] like the one used to prove Sudakov's minoration inequality can be applied to get different estimates. As (3.7) suffices for our purpose, we do not detail the just mentioned tools any further.

The quantity $\log(N(K, \varepsilon))$, which appears in Theorem 3.2.8, is also called *metric entropy* of K. It characterizes the necessary number of bits to encode K in a way that any $\mathbf{z} \in K$ can be decoded with error at most ε in the Euclidean metric, see [166, Proposition 4.3.1]. Consequently, the expected diameter of random one-dimensional cuts through a set K given by $w(K)$ is equivalent to the intrinsic complexity of K in terms of coding.

There is even more evidence that one-dimensional random cuts provide a good measure of intrinsic complexity for signal sets in compressed sensing. The so-called M^*-bound [137, 114], an important result in asymptotic convex geometry, states that, for a set $K \subset \mathbb{R}^N$, the expected diameter of a random cut through K, by a hyperplane E of codimension m, is bounded from above by $w(K)/\sqrt{m}$ (up to a constant). If $\mathbf{x} \in K$ is an unknown signal in a signal set $K \subset \mathbb{R}^N$, the entries of $\mathbf{A} \in \mathbb{R}^{m\times N}$ are iid Gaussian, and $\mathbf{y} \in \mathbb{R}^m$ is obtained from (2.1), the M^*-bound implies that the expected worst-case error $\mathsf{E}[\sup_{\mathbf{x}\in K} \|\hat{\mathbf{x}} - \mathbf{x}\|_2]$ of

$$\hat{\mathbf{x}} \in K, \quad \text{subject to } \mathbf{A}\hat{\mathbf{x}} = \mathbf{y},$$

can be upper bounded by $w(K)/\sqrt{m}$ (choose $E = \ker(A)$, cf. [126]). Consequently, $m \gtrsim w(K)^2$ measurements suffice to obtain reasonable uniform approximation guarantees.

Let us come back to our initial observation, namely that the signal sets Σ_s^N and $K_{N,s}$ need in terms of s and N similar amounts of linear measurements for almost loss-less compression and that their covering numbers are of similar order. Combining Lemma 3.2.2 and Theorem 3.2.8, one obtains

$$w(\Sigma_s^N \cap \mathcal{B}_2(\mathbf{0},1)) \lesssim \sqrt{s\log\left(\frac{3eN}{s}\right)}.$$

By (3.5) and Lemma 3.2.7 (*iii*) the same bound holds for $w(K_{N,s})$. As Σ_s^N and $K_{N,s}$ share the same intrinsic complexity, they need a similar amount of linear measurements to be compressed. Having the Gaussian width as a complexity measure at hand, we can consider general sets $K \subset \mathbb{R}^N$ as signal sets from now on.

3.3 One-bit Quantization

In the thesis, we mainly concentrate on the *one-bit compressed sensing* model

$$\mathbf{y} = \operatorname{sign}(\mathbf{A}\mathbf{x}), \tag{3.8}$$

where $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{A} \in \mathbb{R}^{m\times N}$, and $\mathbf{y} \in \{-1,1\}^m$. This extreme case of uniform scalar quantization (3.3) was introduced to the compressed sensing framework in [25]. Each measurement is quantized to one single bit. Note that (3.8) looses any scaling information of $\mathbf{x}$ and $\mathbf{A}\mathbf{x}$, i.e., one can only hope to recover $\mathbf{x}$ up to its norm. Though the loss of information in (3.8) seems overwhelming, one-bit measurement devices are simple and cheap to produce and make (3.8) appealing in hardware implementations, cf. [25]. Moreover, numerical studies [25, 23, 92] showed successful approximation of sparse, unit norm signals from one-bit measurements via modified greedy compressed sensing algorithms and ℓ_1-minimization with non-convex constraints (restriction to $\mathbb{S}^{N-1}$). In [92] the authors provided near-optimal – with respect to (3.4) – approximation guarantees for sparse, unit norm vectors by *consistent reconstruction*

$$\hat{\mathbf{x}} \in \Sigma_s^N \cap \mathbb{S}^{N-1}, \quad \text{subject to } \operatorname{sign}(\mathbf{A}\hat{\mathbf{x}}) = \mathbf{y}, \tag{3.9}$$

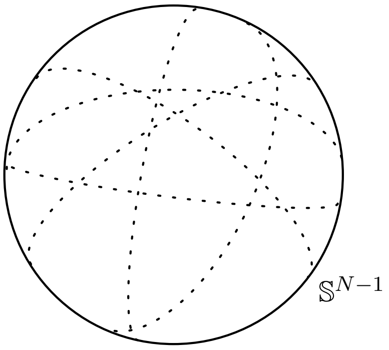


Figure 3.5: Tessellation of the sphere by random hyperplanes.

where the approximation of $\mathbf{x}$ is picked from all signals yielding the correct measurements, a non-tractable procedure in general. To get a intuition for those results one should see (3.8) from a geometric point of view. Each row $\mathbf{a}_i \in \mathbb{R}^N$ of $\mathbf{A}$ can be interpreted as normal vector of a hyperplane $H_{\mathbf{a}_i}$ splitting $\mathbb{R}^N$ into two half-spaces. The one bit measurements $y_i = \operatorname{sign}(\langle \mathbf{a}_i, \mathbf{x} \rangle)$ specify on which side of $H_{\mathbf{a}_i}$ the signal $\mathbf{x}$ is located. All hyperplanes together yield a tessellation of $\mathbb{S}^{N-1}$ as depicted in Figure 3.5, which is taken from [108], and $\mathbf{y}$ encodes in which of the tessellation cells $\mathbf{x}$ lies.

A *dither* $\boldsymbol{\tau} \in \mathbb{R}^N$ with uniformly distributed iid entries τ_i can be introduced to (3.8) to obtain the slightly modified model

$$\mathbf{y} = \operatorname{sign}(\mathbf{A}\mathbf{x} + \boldsymbol{\tau}). \tag{3.10}$$

Dithering is a common tool to improve the statistical properties of the quantization process by randomizing the unquantized input [77]. In the one-bit compressed sensing model it adds affine shifts to the hyperplanes $H_{\mathbf{a}_i}$ and thus enables reconstruction of the signal norm [104, 15].

3.3.1 Recovery via Linear Programming and Single Backprojection

To provide a tractable alternative to (3.9), Plan and Vershynin suggested in [138] to recover signals from their one-bit measurements (3.8) by the convex program

$$\min_{\mathbf{z} \in \mathbb{R}^N} \|\mathbf{z}\|_1, \quad \text{subject to } \operatorname{sign}(\mathbf{A}\mathbf{z}) = \mathbf{y} \text{ and } \|\mathbf{A}\mathbf{z}\|_1 = m. \tag{3.11}$$

The optimization in (3.11) is a modification of basis pursuit. The additional constraint on $\|\mathbf{A}\mathbf{z}\|_1 = m$ enforces a minimal distance to zero on the minimizer and replaces the non-convex constraint $\mathbf{z} \in \mathbb{S}^{N-1}$ in [25] (the right-hand side of the equation can be chosen as an arbitrary constant greater than zero). Plan and Vershynin proved the following result.

Theorem 3.3.1 ([138, Theorem 1.1]). *Let $\delta > 0$ and $\mathbf{A} \in \mathbb{R}^{m \times N}$ be a random matrix with iid Gaussian entries $a_{i,j} \sim \mathcal{N}(0,1)$ where*

$$m \gtrsim \varepsilon^{-5} s \log\left(\frac{2N}{s}\right) \log\left(\frac{2N}{m} + \frac{2m}{N}\right). \tag{3.12}$$

Then with probability at least $1 - C\exp(-c\delta m)$ *(*$c, C > 0$ *denote absolute constants) the following holds uniformly for all* $\mathbf{x} \in \mathbb{R}^N$ *with* $\left\| \mathbf{x} \right\|_1 / \left\| \mathbf{x} \right\|_2 \leq \sqrt{s}$*. If* $\mathbf{y}$ *is defined as in* (3.8)*, the approximation* $\hat{\mathbf{x}}$ *computed by* (3.11) *fulfills*

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

Theorem 3.3.1 not only holds for s-sparse vectors but all vectors which are effectively sparse (cf. Section 3.2) and thus guarantees uniform and stable approximation of sparse vectors from one-bit measurements.
In comparison to [92] the error decay is not near-optimal anymore. As up to log-factors $\varepsilon = \mathcal{O}\left(\sqrt[5]{\frac{s}{m}}\right)$, the error decays much slower than predicted by (3.4). In contrast to [92], Theorem 3.3.1 comes with a tractable approach.
The optimization in (3.11) is especially attractive because it can be stated as the linear program [138]

$$\min_{\mathbf{z},\mathbf{u}\in\mathbb{R}^N} \sum_{i=1}^{N} u_i, \quad \text{subject to} \begin{cases} -u_i \leq z_i \leq u_i, & \forall i \in [N], \\ y_i \langle \mathbf{a}_i, \mathbf{z} \rangle \geq 0, & \forall i \in [N], \\ \frac{1}{m} \sum_{i=1}^{N} y_i \langle \mathbf{a}_i, \mathbf{z} \rangle \geq 1, & \end{cases}$$

which is efficient to solve. Here $\mathbf{a}_i$, for $i \in [N]$, again denote the rows of $\mathbf{A}$ and $\mathbf{u}$ is a vector of dummy variables replacing the ℓ_1-norm.

Foucart provided in [65] a slightly improved version of Theorem 3.3.1. He could drastically shorten the proof by assuming an ℓ_1/ℓ_2-RIP for $\mathbf{A}$. Instead of being a near-isometry between $(\mathbb{R}^N, \left\|\cdot\right\|_2)$ and $(\mathbb{R}^m, \left\|\cdot\right\|_2)$, the matrix $\mathbf{A}$ is in this case a near-isometry between $(\mathbb{R}^N, \left\|\cdot\right\|_2)$ and $(\mathbb{R}^m, \left\|\cdot\right\|_1)$ when restricted to sparse signals, that is

$$(1-\delta) \left\| \mathbf{z} \right\|_2 \leq \left\| \mathbf{A}\mathbf{z} \right\|_1 \leq (1+\delta) \left\| \mathbf{z} \right\|_2, \tag{3.13}$$

for $\delta > 0$ and $\mathbf{z} \in \Sigma_s^N$. By combining (3.13) with a simple relation between $\mathbf{y} = \operatorname{sign}(\mathbf{A}\mathbf{x})$ and $\left\| \mathbf{A}\mathbf{x} \right\|_1$, Foucart was able reduce the number of sufficient measurements from (3.12) to

$$m \gtrsim \varepsilon^{-4} s \log(eN/s). \tag{3.14}$$

In the same work, he showed with a similar strategy the following result for the single backprojection algorithm

$$\hat{\mathbf{x}} = \mathbb{H}_s(\mathbf{A}^T \mathbf{y}), \tag{3.15}$$

leading to (3.14) as well.

Theorem 3.3.2 ([65, Theorem 8])**.** *If* $\mathbf{A}$ *satisfies* (3.13) *for all* $\mathbf{z} \in \Sigma_{2s}^N$ *with constant* $\varepsilon > 0$*, then every* $\mathbf{x} \in \Sigma_s^N \cap \mathbb{S}^{N-1}$ *with one-bit measurements* $\mathbf{y}$ *as in* (3.8) *is approximated by* (3.15) *with error*

$$\left\| \mathbf{x} - \hat{\mathbf{x}} \right\|_2 \leq 2\sqrt{5\varepsilon}.$$

Both results, Theorem 3.3.1 and Theorem 3.3.2, show the possibility of compressed sensing even under heavy quantization by very simple means. However, they do not consider robustness against measurement defects. The optimization (3.11) might even become infeasible if just a single one-bit measurement is flipped.

3.3.2 Uniform Tessellation by Random Hyperplanes

Apart from guaranteeing near-optimal recovery in the noiseless case by consistent reconstruction (as discussed in the beginning of Section 3.3), a core result of [92] was to show that under suitable assumptions, the map $\mathbf{z} \mapsto \operatorname{sign}(\mathbf{A}\mathbf{z})$ behaves almost like an isometry between $(\mathbb{S}^{N-1}, d_G)$ and $(\{-1,1\}^m, d_H)$ where $d_G(\mathbf{z},\mathbf{z}') = \frac{1}{\pi}\arccos(\langle \mathbf{z},\mathbf{z}'\rangle)$ denotes the normalized geodesic distance, i.e., poles have distance $d_G(\mathbf{z},-\mathbf{z}) = 1$, and $d_H(\mathbf{y},\mathbf{y}') = |\{i \in [m] \colon y_i \neq y_i'\}|$ the Hamming distance. To be more precise, they introduced the following concept.

Definition 3.3.3 ([92, Definition 1])**.** *Let $\varepsilon \in (0,1)$. A mapping $F \colon \mathbb{R}^N \to \{-1,1\}^m$ is a* binary ε-stable embedding *of order s if*

$$d_G(\mathbf{z},\mathbf{z}') - \varepsilon \leq \frac{1}{m} d_H(F(\mathbf{z}),F(\mathbf{z}')) \leq d_G(\mathbf{z},\mathbf{z}') + \varepsilon, \tag{3.16}$$

for all $\mathbf{z},\mathbf{z}' \in \mathbb{S}^{N-1}$ with $|\operatorname{supp}(\mathbf{z}) \cup \operatorname{supp}(\mathbf{z}')| \leq s$.

Definition 3.3.3 is closely related to the concept of RIPs. In contrast to a multiplicative influence of δ in the definition of the s-RIP, the influence of ε is additive. Consequently, $F(\mathbf{z}) = F(\mathbf{z}')$ does not imply $\mathbf{z} = \mathbf{z}'$, but only $d_G(\mathbf{z},\mathbf{z}') \leq \varepsilon$. For $F(\mathbf{z}) = \operatorname{sign}(\mathbf{A}\mathbf{z})$ and $\mathbf{A} \in \mathbb{R}^{m\times N}$ having iid Gaussian entries [92, Theorem 3] shows that

$$m \gtrsim \varepsilon^{-2} s \log(N) \tag{3.17}$$

measurement suffice to guarantee (3.16) with high probability.
The geometrical meaning of Definition 3.3.3 and [92, Theorem 3] is twofold. First, if m satisfies (3.17), then m randomly oriented hyperplanes provide with high probability a tessellation of $\Sigma_s^N \cap \mathbb{S}^{N-1}$ such that the diameter of tessellation cells is uniformly bounded by ε. This could be deduced for approximately sparse vectors in [138] from Theorem 3.3.1 as well (the result was improved and generalized in [21]). Second and more important, the distance of two points in $\mathbf{z},\mathbf{z}' \in \Sigma_s^N \cap \mathbb{S}^{N-1}$ is encoded in the number of hyperplanes which separate $\mathbf{z}$ and $\mathbf{z}'$. This allows robust recovery of signals $\mathbf{x} \in \Sigma_s^N \cap \mathbb{S}^{N-1}$ from their one-bit measurements $\mathbf{y}$ by the non-tractable program

$$\min_{\mathbf{z}\in\mathbb{R}^N} d_H(\mathbf{y}, \operatorname{sign}(\mathbf{A}\mathbf{z})), \quad \text{subject to } \mathbf{z} \in \Sigma_s^N \cap \mathbb{S}^{N-1}. \tag{3.18}$$

In [140] Definition 3.3.3 was generalized to arbitrary subsets of $\mathbb{S}^{N-1}$ and formulated from the geometric point of view.

Definition 3.3.4 ([140, Definition 1.1])**.** *Let $K \subset \mathbb{S}^{N-1}$ and an arrangement of m hyperplanes in $\mathbb{R}^N$ be given via a matrix A (i.e., the i-th row of A is the normal to the i-th hyperplane). Let $d_A(\mathbf{z},\mathbf{z}') = \frac{1}{m} d_H(\operatorname{sign}(\mathbf{A}\mathbf{z}), \operatorname{sign}(\mathbf{A}\mathbf{z}'))$ denote the fraction of hyperplanes separating $\mathbf{z}$ and $\mathbf{z}'$ in K. Given $\varepsilon > 0$, the hyperplanes provide an* ε-uniform tessellation *of K if*

$$|d_A(\mathbf{z},\mathbf{z}') - d_G(\mathbf{z},\mathbf{z}')| \leq \varepsilon$$

holds for all $\mathbf{z},\mathbf{z}' \in K$.

The authors could show that Gaussian matrices have with high probability this more general property (if the normal vector $\mathbf{a}$ of a hyperplane has iid entries $a_{i,j}$, the hyperplane is drawn from a uniform distribution according to the Haar measure). For later usage, we state it as in [91].

Theorem 3.3.5 ([140, Theorem 1.2]). *Consider a subset $K \subset \mathbb{S}^{N-1}$ and let $\varepsilon > 0$. Let*

$$m \geq \bar{C}\varepsilon^{-6} \max\{w(K)^2, 2/\pi\} \tag{3.19}$$

and consider an arrangement of m independent random hyperplanes in $\mathbb{R}^N$ uniformly distributed according to the Haar measure. Then with probability at least $1-2\exp(-c\delta^2 m)$, these hyperplanes provide an ε-uniform tessellation of K. Here, $\bar{C} > 0$ denotes an absolute constant.

Having the considerations on the Gaussian mean width and its equivalent variants in mind, Theorem 3.3.5 states that if the number of one-bit measurements (3.8) scale at least linearly in intrinsic dimension of a set $K \subset \mathbb{S}^{N-1}$ then with high probability the percentage of different measurements of two points $\mathbf{z}, \mathbf{z}' \in K$ is closely related to their distance on the sphere. Note the exceedingly worse dependence on ε in (3.19).
In its original form Theorem 3.3.5 uses $\gamma(K)$ instead of $w(K)$. However, note that by (3.6) we know that $\gamma(K) \leq w(K-K) + \sqrt{2/\pi} \leq 3w(K)$, for $K \subseteq \mathbb{S}^{N-1}$ and $w(K) \geq \sqrt{2/\pi}$ which is reasonable to assume. Changing $\bar{C}$ by a factor of 9, Theorem 3.3.5 can be stated as above.
As (3.8) is blind to scaling, Definition 3.3.3 and Definition 3.3.4 are restricted to $\mathbb{S}^{N-1}$. By considering the dithered measurement model (3.10) in [47], Dirksen and Mendelson recently removed this restriction, improved (3.19) and generalized the statement of Theorem 3.3.5 to subgaussian and heavy-tailed measurement matrices $\mathbf{A}$, i.e., the entries of $\mathbf{A}$ are iid subgaussian/heavy-tailed random variables. After providing a definition of subgaussian random variables we state the result in the subgaussian case.

Definition 3.3.6 (Subgaussian Random Variable). *A random variable $\xi \in \mathbb{R}$ is called* $\mathcal{K}$-subgaussian *if the tail bound $\mathsf{Pr}[|\xi| > t] \leq C\exp(-ct^2/\mathcal{K}^2)$ holds where $c, C > 0$ are absolute constants. The smallest possible number for $\mathcal{K} > 0$ is called* subgaussian norm *of ξ and denoted by $\|\xi\|_{\psi_2}$.*

Remark 3.3.7. *The class of subgaussian random variables covers many special cases as Gaussian, Bernoulli, and more generally all bounded random variables (see [164]).*

Theorem 3.3.8 ([47, Theorem 1.1]). *Let $\mathbf{A} \in \mathbb{R}^{m\times N}$ be a matrix with iid $\mathcal{K}$-subgaussian entries. There exist constants $c_0, ..., c_4$ depending only on $\mathcal{K}$ such that the following holds. Fix $R > 0$ and $\varepsilon \in (0, R)$. If $K \subset \mathcal{B}_2(\mathbf{0}, R)$, $\boldsymbol{\tau} \in \mathbb{R}^m$ with iid entries $\tau_i \sim \mathcal{U}([-\lambda, \lambda])$, for $\lambda = c_0 R$, and*

$$m \geq c_1 \frac{R \log\left(\frac{eR}{\varepsilon}\right)}{\varepsilon^3} \gamma(K)^2,$$

then with probability at least $1 - 8\exp(-c_2 m\varepsilon/R)$, for any $\mathbf{z}, \mathbf{z}' \in \operatorname{conv}(K)$ such that $\|\mathbf{z}-\mathbf{z}'\|_2 \geq \varepsilon$, one has

$$c_3 \frac{\|\mathbf{z}-\mathbf{z}'\|_2}{R} \leq \frac{1}{m} d_H(\operatorname{sign}(\mathbf{A}\mathbf{z}+\boldsymbol{\tau}), \operatorname{sign}(\mathbf{A}\mathbf{z}'+\boldsymbol{\tau})) \leq c_4 \sqrt{\log\left(\frac{eR}{\varepsilon}\right)} \frac{\|\mathbf{z}-\mathbf{z}'\|_2}{R}. \tag{3.20}$$

Apart from treating more general signal sets and measurement matrices, Theorem 3.3.8 reduces the sufficient number of measurements from $m \gtrsim \mathcal{O}(\varepsilon^{-6})$ to $\mathcal{O}(\varepsilon^{-3}\log(\varepsilon^{-1}))$. Note that by assuming $\|\mathbf{z}-\mathbf{z}'\|_2 \geq \varepsilon$ the isometric relation in (3.20) is just like (3.16) of additive nature and that there is a mild dependence on ε in the upper bound of (3.20).

3.3.3 Recovery from Noisy Measurements

While Section 3.3.2 shows the possibility of faithfully approximating signals from noisy one-bit measurements, it provides no efficient recovery strategy. As mentioned above, the convex one-bit basis pursuit in (3.11) cannot handle bit-flips. If $\mathbf{y}$ is obtained from a signal $\mathbf{x} \in K \subset \mathbb{R}^N$ via (3.8) where some measurements are corrupted, Plan and Vershynin proposed in [139] to approximate $\mathbf{x}$ by a solution of

$$\max_{\mathbf{z}\in\mathbb{R}^N} \sum_{i=1}^m y_i \langle \mathbf{a}_i, \mathbf{z}\rangle, \quad \text{subject to } \mathbf{z} \in K, \tag{3.21}$$

where $\mathbf{a}_i$ is the i-th row of $\mathbf{A}$. If K is convex, the program (3.21) is convex, too. Its main idea is to maximize the correlation between quantized measurements of the signal and unquantized measurements of the approximation. In [47] the authors elaborately describe how (3.21) can be interpreted as a convex relaxation of (3.18). A geometric interpretation is the following, cf. [91]. The optimization in (3.21) may be re-stated equivalently as

$$\begin{aligned}\min_{\mathbf{z}\in K} \Bigg(&\sum_{i\colon y_i \neq \operatorname{sign}(\langle \mathbf{a}_i,\mathbf{z}\rangle)} \|\mathbf{a}_i\|_2 \left\| \mathbf{z} - \mathbb{P}_{H_{\mathbf{a}_i}} \mathbf{z} \right\|_2 \\ &- \sum_{i\colon y_i = \operatorname{sign}(\langle \mathbf{a}_i,\mathbf{z}\rangle)} \|\mathbf{a}_i\|_2 \left\| \mathbf{z} - \mathbb{P}_{H_{\mathbf{a}_i}} \mathbf{z} \right\|_2 \Bigg),\end{aligned} \tag{3.22}$$

where $\mathbb{P}_{H_{\mathbf{a}_i}}$ denotes the orthogonal projection onto the $N-1$ dimensional subspace $H_{\mathbf{a}_i}$ perpendicular to $\mathbf{a}_i$. To see this note that $\langle \mathbf{a}_i, \mathbf{z}\rangle / \|\mathbf{a}_i\|_2 = \operatorname{sign}(\langle \mathbf{a}_i, \mathbf{z}\rangle)\|\mathbf{z} - \mathbb{P}_{H_{\mathbf{a}_i}}\|_2$. Hence, (3.21) punishes incorrect measurements of a feasible point $\mathbf{z} \in K$ by its distance to the 'measurements border' $H_{\mathbf{a}_i}$ while rewarding correct ones. Plan and Vershynin also provided robust approximation guarantees for Gaussian matrices $\mathbf{A}$, $K \subset \mathcal{B}_2(\mathbf{0},1)$, and $\mathbf{x} \subset K \cap \mathbb{S}^{N-1}$.

Theorem 3.3.9 ([139, Theorem 1.3])**.** *Let $\mathbf{A} \in \mathbb{R}^{m\times N}$ have iid Gaussian entries $a_i \sim \mathcal{N}(0,1)$ and let K be a subset of the Euclidean unit ball in $\mathbb{R}^N$. Let $\varepsilon > 0$ and suppose that*

$$m \geq C'\varepsilon^{-6} w(K)^2.$$

Then with probability at least $1 - 8\exp(-c\varepsilon^2 m)$, the following event occurs. Consider a signal $\mathbf{x} \in K$ satisfying $\|\mathbf{x}\|_2 = 1$ and its (unknown) uncorrupted one-bit measurements $\mathbf{y}$ as defined in (3.8). Let $\tilde{\mathbf{y}} = (\tilde{y}_1, ..., \tilde{y}_m) \in \{-1,1\}^m$ be any (corrupted) measurements satisfying $d_H(\tilde{\mathbf{y}}, \mathbf{y}) \leq \tau m$. Then

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{z}\in\mathbb{R}^N} \sum_{i=1}^m \tilde{y}_i \langle \mathbf{a}_i, \mathbf{z}\rangle, \qquad \textit{subject to } \mathbf{z} \in K,$$

with input $\tilde{\mathbf{y}}$ *satisfies*

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq \varepsilon\sqrt{\log\left(\frac{e}{\varepsilon}\right)} + 11\tau\sqrt{\log\left(\frac{e}{\tau}\right)}.$$

Similar to Theorem 3.3.5 the result shows an optimal dependence on the intrinsic dimension of K but a suboptimal dependence on ε. Note that in order to reach $\|\hat{\mathbf{x}} - \mathbf{x}\|_2 = \mathcal{O}(\varepsilon)$ in the noiseless case, the measurements have to be $\mathcal{O}(\varepsilon^{-12}w(K)^2)$.
Considering the dithered model (3.10) and adding a regularization term to (3.21), Dirksen and Mendelson were able to improve and generalize Theorem 3.3.9 in [47] as well. To be precise, they examined

$$\max_{\mathbf{z}\in\mathbb{R}^N} \sum_{i=1}^m y_i\langle \mathbf{a}_i, \mathbf{z}\rangle - \frac{1}{2\lambda}\|\mathbf{z}\|_2^2, \quad \text{subject to } \mathbf{z} \in K, \tag{3.23}$$

for $\lambda > 0$, and extended Theorem 3.3.9 to subgaussian and heavy-tailed measurement matrices. We again restrict ourselves to the subgaussian case. Moreover, we do not treat pre-quantization noise which is covered by their result.

Theorem 3.3.10 ([47, Theorem 1.1])**.** *Let* $\mathbf{A} \in \mathbb{R}^{m\times N}$ *be a matrix with iid* $\mathcal{K}$*-subgaussian entries. There exist constants* $c_0, ..., c_4$ *depending only on* $\mathcal{K}$ *such that the following holds. Fix* $R > 0$ *and* $\varepsilon \in (0, R)$*. Let* $K \subset R\mathcal{B}_2(\mathbf{0}, 1)$*,* $\boldsymbol{\tau} \in \mathbb{R}^m$ *with iid entries* $\tau_i \sim \mathcal{U}([-\lambda, \lambda])$*, for* $\lambda \geq c_0R + \varepsilon$*, and put* $r = c_1\varepsilon/\sqrt{\log(e\lambda/\varepsilon)}$*. Assume that*

$$m \geq c_2\lambda\left(\frac{w\left((K-K)\cap r\mathcal{B}_2(\mathbf{0},1)\right)^2}{\varepsilon^3} + \frac{\log(N(K,r))}{\varepsilon}\right)$$

and that the fraction of corrupted bits β *in* $\mathbf{y}$ *is bounded by* $\beta \leq c_3\varepsilon/\lambda$*.*
Then with probability at least $1 - 10\exp(-c_4m\varepsilon/\lambda)$*, for any* $\mathbf{x} \in K$*, any solution* $\hat{\mathbf{x}}$ *of* (3.23) *satisfies*

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

Remark 3.3.11. *As* $r = \mathcal{O}(\varepsilon/\log(\varepsilon^{-1}))$*, the quantity* $w((K-K)\cap\mathcal{B}(\mathbf{0}, r))$ *can be interpreted as a local Gaussian mean width (cf. [165]) which characterizes the intrinsic complexity of* K *on small balls. It is naturally upper-bounded by* $w(K-K)$*. By applying Sudakov's minoration in Theorem 3.2.8 to the metric entropy, we get*

$$\log(N(K,r)) \lesssim \frac{w(K)^2}{r^2} \leq c\frac{\log\left(\frac{e}{\varepsilon}\right)}{\varepsilon^2}w(K)^2,$$

for $c > 0$ *only depending on* $\mathcal{K}$*. One can thus replace the sufficient number of measurements in Theorem 3.3.10 by*

$$m \geq cc_2\lambda\frac{\log\left(\frac{e}{\varepsilon}\right)}{\varepsilon^3}w(K)^2.$$

Remark 3.3.11 shows that Theorem 3.3.10 massively improves on Theorem 3.3.9. It reduces the lower bound on m from $\mathcal{O}(\varepsilon^{-12}w(K)^2)$ to less than $\mathcal{O}(\varepsilon^{-4}w(K)^2)$.

The results presented so far assume $\mathbf{A}$ to be a dense random matrix. As already mentioned in Chapter 2, this assumption is not practical in view of storage and computation. There are only few results on structured measurement matrices in the uniform scalar one-bit setting we are interested in. Gaussian circulant matrices have been considered in [46] and the announced follow-up work of [47] shall extend Theorem 3.3.8 and Theorem 3.3.10 to this regime.

3.4 Multi-bit Quantization

As the last section showed it is expensive to obtain with one-bit measurements high precision in signal estimation. Letting ε to zero blows up the required number of measurements. The lower bound (3.4), however, suggests that a slight increase in the number of bits per measurement should cause a notable decrease of ε. We thus consider the multi-bit compressed sensing model

$$\mathbf{y} = U(\mathbf{Az}), \tag{3.24}$$

where $U \colon \mathbb{R} \to \mathbb{R}$ is a uniform scalar B-bit quantizer which is applied to $\mathbf{Az}$ componentwise, for $B \in \mathbb{N}$. Together with B the range of U determines the entrywise worst-case distortion Δ of U. The results we present in this section assume the range of U to be sufficiently large, i.e., they work with an infinite range and only depend on Δ. This assumption is reasonable as signals of interest and, hence, their measurements usually lie in a ball of finite radius (see also Section 3.1). As in the one-bit case, one may introduce a dither $\boldsymbol{\tau} \in \mathbb{R}^m$ to obtain

$$\mathbf{y} = U(\mathbf{Az} + \boldsymbol{\tau}). \tag{3.25}$$

Let us extend the geometric intuition of one-bit compressed sensing (3.8) illustrated in Figure 3.5 to our new setting. By replacing sign with U, each hyperplane $H_{\mathbf{a}_i}$, representing one single measurement, gets replaced by a bundle of parallel hyperplanes. Figure 3.6 depicts this situation for a 2-bit quantizer U which is centered at $\mathbf{0}$. Note that the distances between hyperplanes in different bundles vary depending on the norm of the corresponding measurement vector $\mathbf{a}_i$.

A first important step towards understanding the influence of the number of bits per measurement on the approximation quality in uniformly scalar quantized compressed sensing was done in [93]. The author transferred the Johnson-Lindenstrauss lemma into the uniformly quantized setting and characterized the relation between Δ and how close the Johnson-Lindenstrauss embedding is to an isometry.
Building upon this work, the author could show in [95] a multi-bit tessellation result in flavor of Theorem 3.3.5 for Gaussian measurement matrices and the signal set of sparse vectors (subgaussian measurements and general signal sets are also treated but require rather involved assumptions). We state it here in simplified form.

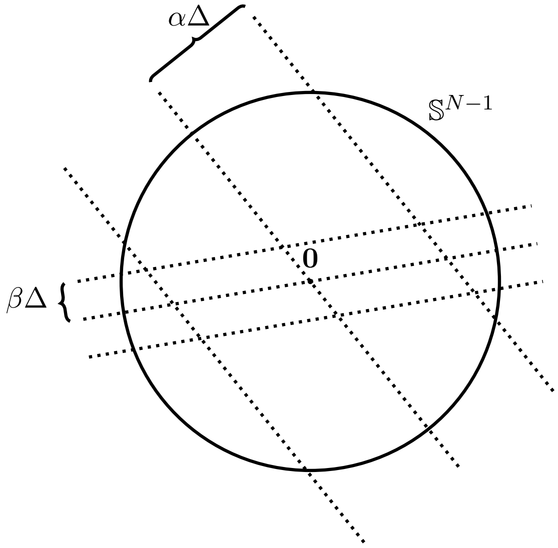


Figure 3.6: Tessellation of the sphere by parallel bundles of hyperplanes.

Theorem 3.4.1 ([95, Proposition 1]). *Fix $\delta \in (0,1)$, $\Delta > 0$, and $K = \Sigma_s^N \cap \mathcal{B}(\mathbf{0},1)$. Assume that $\mathbf{A} \in R^{m\times N}$ has iid Gaussian entries $a_{i,j} \sim \mathcal{N}(0,1)$ and that $\boldsymbol{\tau} \in \mathbb{R}^m$ has iid entries $\tau_i \sim \mathcal{U}([0,\Delta])$. If*

$$m \gtrsim \frac{1}{\delta^2} w(K)^2 \log\left(1 + \frac{1}{\Delta\sqrt{\delta^3}}\right),$$

one has with probability at least $1 - \exp(-c'\delta m)$ for all pairs $\mathbf{z}, \mathbf{z}' \in K$ that

$$(1-c\delta)\left\|\mathbf{z}-\mathbf{z}'\right\|_2 - c\delta\Delta \le \frac{1}{m}\sqrt{\frac{\pi}{2}}\left\|U(\mathbf{A}\mathbf{z}+\boldsymbol{\tau}) - U(\mathbf{A}\mathbf{z}'+\boldsymbol{\tau})\right\|_1 \le (1+c\delta)\left\|\mathbf{z}-\mathbf{z}'\right\|_2 + c\delta\Delta,$$

for absolute constants $c, c' > 0$.

It is striking that Theorem 3.4.1 bridges between uniform one-bit tessellations and ℓ_1/ℓ_2-RIPs. For $\Delta = 1$ the result looks similar to (3.16) while for $\Delta \ll 1$ we recover the RIP in (3.13) (the scaling $\sqrt{\pi/2}$ is necessary for Gaussian matrices $\mathbf{A}$ to have an ℓ_1/ℓ_2-RIP).

In addition, [95] deduced an approximation guarantee for consistent reconstruction from Theorem 3.4.1. We again refrain from presenting the result in full generality but restrict ourselves to the special case of Gaussian measurements and sparse vectors, a case for which the result already appeared in [94].

Theorem 3.4.2 ([94, Theorem 2]). *Fix $\delta \in (0,1)$, $\Delta > 0$, $\theta \in (0,1)$, and $K = \Sigma_s^N \cap \mathcal{B}(\mathbf{0},1)$. Assume that $\mathbf{A} \in R^{m\times N}$ has iid Gaussian entries $a_{i,j} \sim \mathcal{N}(0,1)$ and that $\boldsymbol{\tau} \in \mathbb{R}^m$ has iid entries $\tau_i \sim \mathcal{U}([0,\Delta])$. If*

$$m \ge \frac{4\Delta + 2\delta}{\delta}\left(2s\log\left(\frac{56N}{\sqrt{s}\delta}\right) + \log\left(\frac{1}{2\theta}\right)\right),$$

one has with probability at least $1-\theta$ *for all* $\mathbf{x} \in K$ *that any* $\hat{\mathbf{x}} \in K$ *with* $U(\mathbf{A}\hat{\mathbf{x}} + \boldsymbol{\tau}) = U(\mathbf{A}\mathbf{x} + \boldsymbol{\tau})$ *satisfies*

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

Consequently, one can recover up to quantizer refinement Δ with $m \gtrsim s\log(N/\sqrt{K}\Delta)$ measurements which is for small Δ way better than the one-bit recovery guarantees presented in Section 3.3.
The above results led the way to recovery guarantees for multi-bit basis pursuit [128] and single backprojection [171] which are straight-forward adaptions of the algorithms in Section 3.3.1. We will discuss multi-bit basis pursuit and its connections to support vector machines more detailed in Chapter 6.

Chapter 4

Joint Recovery with One-bit Measurements

As seen in Chapter 3, one-bit compressed sensing is relevant from a practical point of view but requires many measurements when aiming for precise approximation. In this chapter, we examine one-bit quantization in the framework of distributed compressed sensing, i.e., joint recovery of signals. After introducing the concept of joint recovery and providing a brief review of recent developments, we define our problem setting and present the main results. We conclude the chapter with numerical experiments which support the theoretical considerations. All results and all numerical experiments stated in this chapter are joint work with Lars Palzer and have been published in [125].

4.1 Distributed Compressed Sensing

We discussed in Chapter 2 that in order to reconstruct s-sparse signals $\mathbf{x} \in \mathbb{R}^N$ in a stable and robust way from linear measurements of type (2.1), we need at least $m \gtrsim s \log(eN/s)$ measurements. This lower bound consists of two parts. It requires s measurements to identify the entries under knowledge of $\operatorname{supp}(\mathbf{x})$ and an additional factor of $\log(eN/s)$ to find the support. Having knowledge of $\operatorname{supp}(\mathbf{x})$ hence would reduce the number of necessary measurements. However, this assumption is not practical. A more practical assumption, which appears naturally, is we do not only recover one signal $\mathbf{x}$ but several signals $\mathbf{x}_1, ..., \mathbf{x}_L \in \mathbb{R}^N$ that share a common support. For example, in MRI [170] a signal that is sparse in Fourier basis may be measured at different locations, which leads to different attenuations and phase shifts at every node. Another application is MIMO communications [143]. By exploiting the joint support structure one would hope to reduce the number of measurements per signal from $\mathcal{O}(s \log(N/s))$ to $\mathcal{O}(s)$.

Two main measurement models for joint recovery from compressed measurements have been established. The first one is commonly known as *Multiple Measurement Vectors (MMV).* All signals are measured by the same measurement matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ (resp. the same sensor) and the model in (2.1) becomes

$$\mathbf{Y} = \mathbf{A}\mathbf{X}, \tag{4.1}$$

where $\mathbf{X} \in \mathbb{R}^{N \times L}$ and $\mathbf{Y} \in \mathbb{R}^{m \times L}$ are matrices containing the signals and their corresponding measurement vectors as columns. For this model one can only improve the average performance when compared to single vector compressed sensing, while the worst-case analysis shows no improvement, see [58].

In the second model, one considers distinct measurement matrices $\mathbf{A}^{(1)}, ..., \mathbf{A}^{(L)} \in \mathbb{R}^{m \times N}$ (resp. distinct sensors) for each signal $\mathbf{x}_l \in \mathbb{R}^N$, $l \in [L]$. Hence, there are seperate measurement processes of type (2.1) for each $l \in [L]$ yielding L different measurement vectors $\mathbf{y}_l \in \mathbb{R}^m$. We may write

$$\operatorname{vec}(\mathbf{Y}) = \mathbf{A} \cdot \operatorname{vec}(\mathbf{X}), \tag{4.2}$$

where $\mathbf{A} \in \mathbb{R}^{mL \times NL}$ is block diagonal and built from the blocks $\mathbf{A}^{(l)}$, and $\operatorname{vec}(\cdot)$ denotes the vectorization of a matrix. Jointly sparse signal ensembles $\mathbf{X}$ can be recovered from measurements of type (4.2) via $\ell_{2,1}$-minimization if $\mathbf{A}$ satisfies a certain block RIP [57]. Moreover, the authors of [56] relate the number of measurements to guarantee block RIPs for random matrices to properties of the signal ensembles $\mathbf{X}$. They show that one can profit from joint structure if the information in $\mathbf{X}$ is spread among multiple signals $\mathbf{x}_l$. For instance, if all $\mathbf{x}_l$ but one are zero one will need $m = \mathcal{O}(s \log(eN/s))$ measurements per signal and joint recovery becomes useless. To obtain meaningful recovery guarantees for distributed compressed sensing one thus needs assumptions beyond a joint support set.

The idea to jointly recover several signals has been introduced to compressed sensing in [16] under the name *distributed compressed sensing.* We refer the reader to [16, 43] for a more detailed introduction to distributed compressed sensing. Joint recovery is closely related to model-based compressed sensing [14, 57] where one assumes certain structures of the signal support in addition to sparsity. Joint sparsity of several signals appears in this framework also under the name block sparsity of one signal.

4.2 A Distributed One-Bit Model

The papers [159, 103, 81] numerically exemplify increased performance of jointly sparse signal recovery from one-bit measurements as in the unquantized setting, but they do not provide theoretical justification for the improvements. To close this gap let us consider the following model which corresponds to (4.2) above. Suppose we are given one-bit measurements $\mathbf{Y} \in \mathbb{R}^{m \times L}$ obtained from L signals $\mathbf{x}_l \in \mathbb{R}^N$, $l \in [L]$, that form the columns of a matrix $\mathbf{X} \in \mathbb{R}^{N \times L}$. For simplicity we write $\mathbf{x} = \operatorname{vec}(\mathbf{X}) = (\mathbf{x}_1^T, ..., \mathbf{x}_L^T)^T$ and $\mathbf{y} = \operatorname{vec}(\mathbf{Y}) = (\mathbf{y}_1^T, ..., \mathbf{y}_L^T)^T$. The linear measurement process can then be described by

$$\mathbf{y} = \operatorname{sign}\big(\mathbf{A}\mathbf{x}\big), \tag{4.3}$$

where $\mathbf{A} \in \mathbb{R}^{mL \times NL}$ is a block diagonal measurement matrix built from the submatrices $\mathbf{A}^{(l)} \in \mathbb{R}^{m \times N}$, $l \in [L]$, which have iid Gaussian entries $A_{i,j}^{(l)} \sim \mathcal{N}(0,1)$, such that

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}^{(1)} & & \\ & \ddots & \\ & & \mathbf{A}^{(L)} \end{pmatrix}. \tag{4.4}$$

We denote the i-th column of $(\mathbf{A}^{(l)})^T$ by $\mathbf{a}_i^{(l)}$, i.e., $\mathbf{a}_i^{(l)}$ is the transposed i-th row of $\mathbf{A}^{(l)}$. Let $\theta > 0$ be an appropriate scaling to be determined later. We aim to approximate $\mathbf{x}$ by a single back-projected hard-thresholding step (cf. Section 3.3.1)

$$\hat{\mathbf{x}} = \tilde{\mathbb{H}}_s \left((\theta \mathbf{A})^T \mathbf{y} \right), \tag{4.5}$$

where the modified matrix hard-thresholding operator $\tilde{\mathbb{H}}_s(\mathbf{z}) = \tilde{\mathbb{H}}_s(\text{vec}(\mathbf{Z})) = \text{vec}(\mathbb{H}_s(\mathbf{Z}))$ keeps only the s rows of largest ℓ_2-norm, for all $\mathbf{z} = \text{vec}(\mathbf{Z}) \in \mathbb{R}^{NL}$ with $\mathbf{Z} \in \mathbb{R}^{N \times L}$. We denote the support of a signal ensemble $\mathbf{Z} \in \mathbb{R}^{N \times L}$, i.e., the set of non-zero rows of $\mathbf{Z}$, by $\text{supp}(\mathbf{Z}) \subset [N]$ and define the set $\mathcal{S}_{s,L}$ of admissible signal ensembles

$$\mathcal{S}_{s,L} = \left\{ \mathbf{z} = \text{vec}(\mathbf{Z}) \colon \mathbf{Z} = \begin{pmatrix} | & & | \\ \mathbf{z}_1 & \cdots & \mathbf{z}_L \\ | & & | \end{pmatrix} \in \mathbb{R}^{N \times L}, |\,\text{supp}(\mathbf{Z})| \leq s, \|\mathbf{z}_l\|_2 = \|\mathbf{z}\|_2 / \sqrt{L} \right\} \tag{4.6}$$

which contains all ensembles sharing a common support of size at most s and a common ℓ_2-norm. As the non-dithered sign-bit measurements (4.3) are invariant under scaling, we only ask for approximation of the directions of the individual signals. Consequently, whenever we use the terms "approximation of signals" or "recovery of signals" we implicitly mean "approximation/recovery of each signal up to the scaling" and restrict the results to signals of fixed norm.

4.3 An Appropriate RIP

In preparation, we show that Gaussian measurements of the form (4.4) fulfill under suitable scaling with high probability an $\ell_1/\ell_{2,1}$-*restricted isometry property* ($\ell_1/\ell_{2,1}$-RIP) on

$$\mathcal{K}_{s,L} = \left\{ \mathbf{z} = \text{vec}(\mathbf{Z}) \colon \mathbf{Z} \in \mathbb{R}^{N \times L}, |\,\text{supp}(\mathbf{Z})| \leq s \right\} \tag{4.7}$$

if $mL \gtrsim s(\log(eN/s) + L)$. Note that $\mathcal{K}_{s,L}$ is a relaxation of $\mathcal{S}_{s,L}$. Let us first define what we mean by $\ell_1/\ell_{2,1}$-RIP. Recall that, for $\mathbf{z} = \text{vec}(\mathbf{Z})$, $\|\mathbf{z}\|_{2,1} = \|\mathbf{Z}\|_{2,1} = \sum_l \|\mathbf{z}_l\|_2$.

Definition 4.3.1 ($\ell_1/\ell_{2,1}$-RIP)**.** *A matrix $\mathbf{B} \in \mathbb{R}^{mL \times NL}$ satisfies the $\ell_1/\ell_{2,1}$-RIP on $\mathcal{K}_{s,L}$ with RIP-constant $\delta \in (0,1)$ if*

$$\frac{\|\mathbf{z}\|_{2,1}}{\sqrt{L}} - \delta \|\mathbf{z}\|_2 \leq \|\mathbf{B}\mathbf{z}\|_1 \leq \frac{\|\mathbf{z}\|_{2,1}}{\sqrt{L}} + \delta \|\mathbf{z}\|_2, \tag{4.8}$$

for all $\mathbf{z} \in \mathcal{K}_{s,L}$.

As $\|\mathbf{z}\|_{2,1} \leq \sqrt{L}\,\|\mathbf{z}\|_2$, the upper bound in (4.8) can be replaced by $(1+\delta)\,\|\mathbf{z}\|_2$. Moreover, we know that $\|\mathbf{x}\|_{2,1} = \sqrt{L}\,\|\mathbf{x}\|_2$, for $\mathbf{x} \in \mathcal{S}_{s,L}$. Consequently, the $\ell_1/\ell_{2,1}$-RIP in (4.8) becomes a full ℓ_1/ℓ_2-RIP if restricted to $\mathcal{S}_{s,L}$, i.e.,

$$(1-\delta)\,\|\mathbf{x}\|_2 \leq \|\mathbf{B}\mathbf{x}\|_1 \leq (1+\delta)\,\|\mathbf{x}\|_2 \,.$$

The signal model $\mathcal{S}_{s,L}$ appears to be well-chosen as the ensembles in $\mathcal{S}_{s,L}$ when multiplied to block-diagonal Gaussian measurement matrices behave like single vectors multiplied to dense Gaussian measurement matrices. The following lemma characterizes a sufficient number of measurements for $\theta\mathbf{A}$, with $\mathbf{A}$ defined as in (4.4), to fulfill the above introduced $\ell_1/\ell_{2,1}$-RIP. Its proof is inspired by [140, Cor. 2.3].

Lemma 4.3.2 ($\ell_1/\ell_{2,1}$-RIP)**.** *For $\theta = \sqrt{\pi/(2Lm^2)}$ and $mL \gtrsim \delta^{-2}s(\log(eN/s)+L)$ the operator $\theta\mathbf{A}$, with $\mathbf{A}$ defined as in (4.4), has the $\ell_1/\ell_{2,1}$-RIP on $\mathcal{K}_{s,L}$ with RIP-constant δ with probability at least $1-2\exp\left(-\delta^2 mL/(4\pi)\right)$.*

For $L=1$ the above result resembles known bounds on the sufficient number of measurements to have ℓ_1/ℓ_2-RIPs for random Gaussian matrices with high probability as in this case $m \gtrsim s\log(eN/s)$ is required. If $L \geq \log(eN/s)$, we may estimate $(\log(eN/s)+L)/L \leq 2$ and Lemma 4.3.2 instead requires $m \gtrsim \delta^{-2}s$, i.e., only $\mathcal{O}(s)$ measurements per signal.
In [56] the authors examined classical ℓ_2-RIPs for random Gaussian block matrices $\mathbf{A}$ and showed that the sufficient number of measurements depends on how the information of sparse signals is distributed on the different blocks of $\mathbf{A}$. Lemma 4.3.2 extends their result to $\ell_1/\ell_{2,1}$-RIPs in the special case that all signals have the same support.

To prove Lemma 4.3.2, we have to control the Gaussian width of $\mathcal{K}_{s,L}$ when intersected with the unit ball $\mathcal{B}_2(\mathbf{0},1) \subset \mathbb{R}^{NL}$.

Lemma 4.3.3 (Metric entropy of $\mathcal{K}_{s,L} \cap \mathcal{B}_2(\mathbf{0},1)$)**.** *For $\varepsilon \in (0,1)$ we have*

$$\log\left(N(\mathcal{K}_{s,L} \cap \mathcal{B}_2(\mathbf{0},1),\varepsilon)\right) \leq s\log\left(\frac{eN}{s}\right) + sL\log\left(\frac{3}{\varepsilon}\right).$$

Proof: As $\mathcal{K}_{s,L} \cap \mathcal{B}_2(\mathbf{0},1)$ is the union of $\binom{N}{s}$ unit ℓ_2-balls in $\mathbb{R}^{sL}$ embedded into $\mathbb{R}^{NL}$ and each unit ball can be covered by an ε-net of cardinality at most $(3/\varepsilon)^{sL}$ (see [28, Section 3]), we know that

$$N\left(\mathcal{K}_{s,L} \cap \mathcal{B}_2(\mathbf{0},1),\varepsilon\right) \leq \binom{N}{s}\left(\frac{3}{\varepsilon}\right)^{sL} \leq \left(\frac{eN}{s}\right)^{s}\left(\frac{3}{\varepsilon}\right)^{sL}.$$

■

Lemma 4.3.3 leads to a direct bound for $w(\mathcal{K}_{s,L} \cap \mathcal{B}_2(\mathbf{0},1))$.

Lemma 4.3.4 (Gaussian width of $\mathcal{K}_{s,L} \cap \mathcal{B}_2(\mathbf{0},1)$)**.** *We have*

$$w(\mathcal{K}_{s,L} \cap \mathcal{B}_2(\mathbf{0},1)) \lesssim \sqrt{s\left(\log\left(\frac{eN}{s}\right)+L\right)}.$$

Proof: We obtain

$$\begin{aligned} w(\mathcal{K}_{s,L} \cap \mathcal{B}_2(\mathbf{0},1)) &\overset{\text{(i)}}{\leq} 24 \int_0^1 \sqrt{\log\left(N(\mathcal{K}_{s,L} \cap \mathcal{B}_2(\mathbf{0},1),\varepsilon)\right)}\, d\varepsilon \\ &\overset{\text{(ii)}}{\leq} 24 \sqrt{\int_0^1 1^2 \, d\varepsilon} \cdot \sqrt{\int_0^1 \log\left(N(\mathcal{S}_{s,L} \cap \mathcal{B}_2(\mathbf{0},1),\varepsilon)\right)\, d\varepsilon} \\ &\overset{\text{(iii)}}{\leq} 24 \sqrt{s\left(\log\left(\frac{eN}{s}\right) + L(1+\log 3)\right)}, \end{aligned}$$

where (i) follows from Theorem 3.2.8, (ii) from Hölder's inequality and (iii) from Lemma 3.2.2. ■

The main part of the proof of Lemma 4.3.2 is accomplished in the following technical result. It states a concentration inequality for bounded subsets of $\mathbb{R}^{NL}$ and is a slightly adapted version of [140, Lemma 2.1]. Recall that $\gamma(\mathcal{K})$ denotes the Gaussian complexity of $\mathcal{K}$.

Lemma 4.3.5. *Consider a bounded subset $\mathcal{K} \subset \mathbb{R}^{NL}$ and let $\mathbf{a}_i^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{Id}_N)$, $i \in [m]$, $l \in [L]$ be independent Gaussian vectors in $\mathbb{R}^N$. Define*

$$Z := \sup_{\mathbf{x}\in\mathcal{K}} \left| \sum_{i=1}^m \sum_{l=1}^L \sqrt{\frac{\pi}{2Lm^2}} \left| \left\langle \mathbf{a}_i^{(l)}, \mathbf{x}_l \right\rangle \right| - \frac{1}{\sqrt{L}} \|\mathbf{x}\|_{2,1} \right|. \tag{4.9}$$

Then we have

$$\mathsf{E}[Z] \leq \sqrt{8\pi} \frac{\gamma(\mathcal{K})}{\sqrt{mL}}$$

and

$$\Pr\left[Z > \frac{\sqrt{8\pi}\gamma(\mathcal{K})}{\sqrt{mL}} + u \right] \leq 2\exp\left(-\frac{mLu^2}{\pi d(\mathcal{K})^2} \right), \tag{4.10}$$

where $d(\mathcal{K}) := \max_{\mathbf{x}\in\mathcal{K}} \|\mathbf{x}\|_2$.

Proof: Let $g \sim \mathcal{N}(0,1)$ and note that $\mathsf{E}[|g|] = \sqrt{2/\pi}$. Then, we have

$$\mathsf{E}\left[\sum_{i=1}^m \sum_{l=1}^L \sqrt{\frac{\pi}{2Lm^2}} \left| \left\langle \mathbf{a}_i^{(l)}, \mathbf{x}_l \right\rangle \right| \right] = \sum_{i=1}^m \sum_{l=1}^L \sqrt{\frac{\pi}{2Lm^2}}\, \mathsf{E}[|g|] \, \|\mathbf{x}_l\|_2 = \frac{\|\mathbf{x}\|_{2,1}}{\sqrt{L}}.$$

Define now for $i \in [m], l \in [L]$ the random variables $\boldsymbol{\vartheta}_i^{(l)} = \sqrt{\pi/(2Lm^2)} \left| \left\langle \mathbf{a}_i^{(l)}, \mathbf{x}_l \right\rangle \right|$, identically distributed independent copies $\hat{\boldsymbol{\vartheta}}_i^{(l)}$, and independent Rademacher variables $\varepsilon_{i,l}$, i.e., $\mathsf{Pr}[\varepsilon_{i,l} = 1] = \mathsf{Pr}[\varepsilon_{i,l} = -1] = 1/2$. We obtain

$$\begin{aligned}
\mathsf{E}[Z] &= \mathsf{E}_{\boldsymbol{\vartheta}}\left[\sup_{\mathbf{x}\in\mathcal{K}}\left|\sum_{i=1}^{m}\sum_{l=1}^{L}(\boldsymbol{\vartheta}_i^{(l)} - \mathsf{E}_{\boldsymbol{\vartheta}}\left[\boldsymbol{\vartheta}_i^{(l)}\right])\right|\right] \\
&= \mathsf{E}_{\boldsymbol{\vartheta}}\left[\sup_{\mathbf{x}\in\mathcal{K}}\left|\sum_{i=1}^{m}\sum_{l=1}^{L}\left(\boldsymbol{\vartheta}_i^{(l)} - \mathsf{E}_{\boldsymbol{\vartheta}}\left[\boldsymbol{\vartheta}_i^{(l)}\right]\right) - \mathsf{E}_{\hat{\boldsymbol{\vartheta}}}\left[\hat{\boldsymbol{\vartheta}}_i^{(l)} - \mathsf{E}_{\hat{\boldsymbol{\vartheta}}}\left[\hat{\boldsymbol{\vartheta}}_i^{(l)}\right]\right]\right|\right] \\
&= \mathsf{E}_{\boldsymbol{\vartheta}}\left[\sup_{\mathbf{x}\in\mathcal{K}}\left|\sum_{i=1}^{m}\sum_{l=1}^{L}\mathsf{E}_{\hat{\boldsymbol{\vartheta}}}\left[\boldsymbol{\vartheta}_i^{(l)} - \hat{\boldsymbol{\vartheta}}_i^{(l)}\right]\right|\right] \\
&\overset{\text{(i)}}{\leq} \mathsf{E}_{\boldsymbol{\vartheta}}\left[\mathsf{E}_{\hat{\boldsymbol{\vartheta}}}\left[\sup_{\mathbf{x}\in\mathcal{K}}\left|\sum_{i=1}^{m}\sum_{l=1}^{L}\boldsymbol{\vartheta}_i^{(l)} - \hat{\boldsymbol{\vartheta}}_i^{(l)}\right|\right]\right] \\
&= \mathsf{E}_{\boldsymbol{\vartheta},\varepsilon}\left[\mathsf{E}_{\hat{\boldsymbol{\vartheta}}}\left[\sup_{\mathbf{x}\in\mathcal{K}}\left|\sum_{i=1}^{m}\sum_{l=1}^{L}\varepsilon_{i,l}\left(\boldsymbol{\vartheta}_i^{(l)} - \hat{\boldsymbol{\vartheta}}_i^{(l)}\right)\right|\right]\right] \\
&\overset{\text{(ii)}}{\leq} 2\,\mathsf{E}_{\boldsymbol{\vartheta},\varepsilon}\left[\sup_{\mathbf{x}\in\mathcal{K}}\left|\sum_{i=1}^{m}\sum_{l=1}^{L}\varepsilon_{i,l}\boldsymbol{\vartheta}_i^{(l)}\right|\right] \qquad (4.11) \\
&= 2\sqrt{\frac{\pi}{2Lm^2}}\,\mathsf{E}\left[\sup_{\mathbf{x}\in\mathcal{K}}\left|\sum_{i=1}^{m}\sum_{l=1}^{L}\varepsilon_{i,l}\left|\left\langle\mathbf{a}_i^{(l)},\mathbf{x}_l\right\rangle\right|\right|\right] \\
&\overset{\text{(iii)}}{\leq} 4\sqrt{\frac{\pi}{2Lm^2}}\,\mathsf{E}\left[\sup_{\mathbf{x}\in\mathcal{K}}\left|\sum_{i=1}^{m}\sum_{l=1}^{L}\varepsilon_{i,l}\left\langle\mathbf{a}_i^{(l)},\mathbf{x}_l\right\rangle\right|\right] \\
&= 4\sqrt{\frac{\pi}{2Lm^2}}\,\mathsf{E}\left[\sup_{\mathbf{x}\in\mathcal{K}}\left|\sum_{l=1}^{L}\left\langle\sum_{i=1}^{m}\varepsilon_{i,l}\mathbf{a}_i^{(l)},\mathbf{x}_l\right\rangle\right|\right] \\
&= 4\sqrt{\frac{\pi}{2Lm^2}}\,\mathsf{E}\left[\sup_{\mathbf{x}\in\mathcal{K}}\left|\left\langle\sum_{i=1}^{m}\left(\varepsilon_{i,1}(\mathbf{a}_i^{(1)})^T,\ldots,\varepsilon_{i,L}(\mathbf{a}_i^{(L)})^T\right)^T,\mathbf{x}\right\rangle\right|\right] \\
&\overset{\text{(iv)}}{=} 4\sqrt{\frac{\pi}{2Lm^2}}\,\mathsf{E}\left[\sup_{\mathbf{x}\in\mathcal{K}}\left|\left\langle\sqrt{m}\mathbf{g},\mathbf{x}\right\rangle\right|\right] = \sqrt{8\pi}\frac{\gamma(\mathcal{K})}{\sqrt{mL}},
\end{aligned}$$

where (i) follows from Jensen's inequality, (ii) from the triangle inequality, (iii) is a consequence of [114, Thm. 4.12], and in (iv) we let $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{Id}_{NL})$.To prove the deviation inequality (4.10) we will first show that Z, as defined in (4.9), is Lipschitz continuous in $\mathbf{A}$. Consider two block diagonal matrices $\mathbf{A}, \mathbf{B}$ as in (4.4) and define the operator

$$Z(\mathbf{A}) := \sup_{\mathbf{x}\in\mathcal{S}}\left|\sum_{i=1}^{m}\sum_{l=1}^{L}\sqrt{\frac{\pi}{2Lm^2}}\left|\left\langle\mathbf{a}_i^{(l)},\mathbf{x}_l\right\rangle\right| - \frac{\|\mathbf{x}\|_{2,1}}{\sqrt{L}}\right|.$$

Then, we have

$$\begin{aligned}
|Z(\mathbf{A}) - Z(\mathbf{B})| &= \sup_{\mathbf{x}\in\mathcal{K}} \left| \sum_{i=1}^{m}\sum_{l=1}^{L} \sqrt{\frac{\pi}{2Lm^2}} \left| \left\langle \mathbf{a}_i^{(l)}, \mathbf{x}_l \right\rangle \right| - \frac{\|\mathbf{x}\|_{2,1}}{\sqrt{L}} \right| \\
&\quad - \sup_{\mathbf{x}\in\mathcal{S}} \left| \sum_{i=1}^{m}\sum_{l=1}^{L} \sqrt{\frac{\pi}{2Lm^2}} \left| \left\langle \mathbf{b}_i^{(l)}, \mathbf{x}_l \right\rangle \right| - \frac{\|\mathbf{x}\|_{2,1}}{\sqrt{L}} \right| \\
&\leq \sup_{\mathbf{x}\in\mathcal{K}} \left\{ \left| \sum_{i=1}^{m}\sum_{l=1}^{L} \sqrt{\frac{\pi}{2Lm^2}} \left| \left\langle \mathbf{a}_i^{(l)}, \mathbf{x}_l \right\rangle \right| - \frac{\|\mathbf{x}\|_{2,1}}{\sqrt{L}} \right| \right. \\
&\quad \left. - \left| \sum_{i=1}^{m}\sum_{l=1}^{L} \sqrt{\frac{\pi}{2Lm^2}} \left| \left\langle \mathbf{b}_i^{(l)}, \mathbf{x}_l \right\rangle \right| - \frac{\|\mathbf{x}\|_{2,1}}{\sqrt{L}} \right| \right\} \\
&\leq \sup_{\mathbf{x}\in\mathcal{K}} \left| \sum_{i=1}^{m}\sum_{l=1}^{L} \sqrt{\frac{\pi}{2Lm^2}} \left| \left\langle \mathbf{a}_i^{(l)} - \mathbf{b}_i^{(l)}, \mathbf{x}_l \right\rangle \right| \right| \\
&\leq \sup_{\mathbf{x}\in\mathcal{K}} \sqrt{\frac{\pi}{2Lm^2}} \sum_{i=1}^{m}\sum_{l=1}^{L} \left\| \mathbf{a}_i^{(l)} - \mathbf{b}_i^{(l)} \right\|_2 \|\mathbf{x}_l\|_2 \\
&\leq \sup_{\mathbf{x}\in\mathcal{K}} \sqrt{\frac{\pi}{2Lm^2}} \|\mathbf{A} - \mathbf{B}\|_F \left(\sum_{i=1}^{m}\sum_{l=1}^{L} \|\mathbf{x}_l\|_2^2 \right)^{\frac{1}{2}} \\
&\leq \sqrt{\frac{\pi}{2Lm^2}} \sqrt{m} \|\mathbf{A} - \mathbf{B}\|_F \, d(\mathcal{K}) \\
&= \frac{d(\mathcal{K})}{\sqrt{mL}} \sqrt{\frac{\pi}{2}} \|\mathbf{A} - \mathbf{B}\|_F .
\end{aligned}$$

Hence, $Z(\cdot)$ is Lipschitz continuous with constant $\frac{d(\mathcal{K})}{\sqrt{mL}}\sqrt{\frac{\pi}{2}}$. Using the measure concentration of Lipschitz functions evaluated on Gaussian random vectors [114, Eq. (1.6)], we see that

$$\mathsf{Pr}[|Z - \mathsf{E}[Z]\,| > u] \leq 2\exp\left(-\frac{2u^2 mL}{2\pi d(\mathcal{K})^2} \right).$$

Using (4.11), we have

$$\begin{aligned}
\mathsf{Pr}\left[Z - \sqrt{8\pi}\frac{\gamma(\mathcal{K})}{\sqrt{mL}} > u \right] &\leq \mathsf{Pr}[Z - \mathsf{E}[Z] > u] \leq \mathsf{Pr}[|Z - \mathsf{E}[Z]\,| > u] \\
&\leq 2\exp\left(-\frac{mLu^2}{\pi d(\mathcal{K})^2} \right),
\end{aligned}$$

which yields the claim. ■

Proof of Lemma 4.3.2 : As (4.8) is invariant under scaling of the ℓ_2-norm, it suffices to show (4.8) for all $\mathbf{z} \in \mathcal{K}_{s,L} \cap \mathbb{S}^{NL-1}$. By origin symmetry, we have that $\gamma(\mathcal{K}_{s,L} \cap \mathbb{S}^{NL-1}) = w(\mathcal{K}_{s,L} \cap \mathbb{S}^{NL-1})$.

Lemma 4.3.2 is a direct consequence of Lemmas 4.3.4 and 4.3.5. Just choose $u = \delta/2$ and $mL \geq 8\pi(\delta/2)^{-2}w(\mathcal{K}_{s,L} \cap \mathcal{B}_2(\mathbf{0},1))^2$ and note that $w(\mathcal{K}_{s,L} \cap \mathcal{B}_2(\mathbf{0},1)) \geq w(\mathcal{K}_{s,L} \cap \mathbb{S}^{NL-1})$. Then, with probability at least $1 - 2\exp\left(-mL\delta^2/(4\pi)\right)$, we have for all $\mathbf{z} \in \mathcal{K}_{s,L} \cap \mathbb{S}^{NL-1}$

$$\left| \sqrt{\frac{\pi}{2Lm^2}} \|\mathbf{A}\mathbf{z}\|_1 - \frac{\|\mathbf{z}\|_{2,1}}{\sqrt{L}} \right| \leq \sqrt{8\pi} \frac{w\left(\mathcal{K}_{s,L} \cap \mathbb{S}^{NL-1}\right)}{\sqrt{mL}} + \frac{\delta}{2} \leq \delta.$$

The statement follows for $\mathbf{z} \in \mathcal{K}_{s,L}$ by applying the above statement to $\frac{\mathbf{z}}{\|\mathbf{z}\|_2} \in \mathcal{K}_{s,L} \cap \mathbb{S}^{NL-1}$. ■

4.4 Approximation of Signal Ensembles

We are ready to state the main result of this chapter. It guarantees uniform recovery of all signal ensembles $\mathbf{x} \in \mathcal{S}_{s,L}$ by a simple hard-thresholding step and can be regarded as a generalization of Theorem 3.3.2 to joint recovery of signals sharing a common support.

Theorem 4.4.1. *Let $\mathbf{A}$ be a random $mL \times NL$ matrix as defined in* (4.4)*. Set*

$$mL \gtrsim \delta^{-2} s(\log(eN/s) + L) \tag{4.12}$$

and $\theta = \sqrt{\pi/(2Lm^2)}$.Then with probability at least $1 - 2\exp\left(-\delta^2 mL/(4\pi)\right)$ (over the entries of $\mathbf{A}$), we have for all $\mathbf{x} \in \mathcal{S}_{s,L}$ with $\|\mathbf{x}\|_2 = 1$ that

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \lesssim \sqrt{\delta}, \tag{4.13}$$

where $\hat{\mathbf{x}}$ is defined in (4.5) *and δ is the $\ell_1/\ell_{2,1}$-RIP constant of $\theta\mathbf{A}$.*

If $L \geq \log(N/s)$, the required number of measurements per signal does not depend on $s\log(N/s)$ but only on s (cf. discussion in Section 4.3). Consequently, when recovering several signals that share a common support from sign-measurements collected independently for each single signal, one can significantly reduce the necessary number of measurements by using the support structure.
For unit norm signals $\|\mathbf{x}_l\| = 1$ the error per single signal $\mathbf{x}_l$ is on average bounded by $\sqrt{\delta}$ as (4.13) becomes

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \lesssim \sqrt{L\delta}.$$

In the worst case the error is concentrated on one signal. However, if the signals all are dense on shared support set $\mathcal{T} \subset [n]$, the support will be recovered even in this case as a worse error on one signal implies less error on the remaining signals. Obviously, one needs a dense support of all signals to profit from joint recovery. If just one signal is dense on its support while the rest contains mostly zeros on $\mathcal{T}$, most of the signals do not carry valuable support information and joint recovery cannot be expected to improve performance (see also the discussion of distributed compressed sensing in Section 4.1).

The proof of Theorem 4.4.1 follows the argument of Theorem 3.3.2 (developed in [65]) but relies on the assumption that all signals $\mathbf{x}_l$ share a common ℓ_2-norm. By assumption, we have that $\operatorname{supp}(\mathbf{X}) = \operatorname{supp}(\mathbf{x}) \subset \mathcal{T}$ for some $\mathcal{T} \subset [N]$ with $|\mathcal{T}| \le s$. For $\mathbf{z} = \operatorname{vec}(Z) \in \mathbb{R}^{NL}$ let $\mathbf{z}_{\mathcal{T}} = \operatorname{vec}(\mathbf{Z}_{\mathcal{T}})$ with $\mathbf{Z}_{\mathcal{T}}$ being the matrix in which all rows not in $\mathcal{T}$ are set to zero.

Lemma 4.4.2. *If the operator $\theta\mathbf{A}$ satisfies the $\ell_1/\ell_{2,1}$-RIP on $\mathcal{K}_{s,L}$, then all $\mathbf{x} \in \mathcal{S}_{s,L}$ with $\|\mathbf{x}\|_2 = 1$ satisfy*

$$\left\| \left((\theta\mathbf{A})^T \operatorname{sign}(\mathbf{A}\mathbf{x}) \right)_{\mathcal{T}} - \mathbf{x} \right\|_2^2 \le 5\delta.$$

Proof: Define $\theta\mathbf{b} = \theta\mathbf{A}^T \operatorname{sign}(\mathbf{A}\mathbf{x}) \in \mathbb{R}^{NL}$ to be the backprojected quantized measurements. We then have

$$\left\| \left((\theta\mathbf{A})^T \operatorname{sign}(\mathbf{A}\mathbf{x}) \right)_{\mathcal{T}} - \mathbf{x} \right\|_2^2 = \|(\theta\mathbf{b})_{\mathcal{T}}\|_2^2 - 2\langle(\theta\mathbf{b})_{\mathcal{T}}, \mathbf{x}\rangle + \|\mathbf{x}\|_2^2$$

and

$$\begin{aligned} \|(\theta\mathbf{b})_{\mathcal{T}}\|_2^2 &= \langle(\theta\mathbf{b})_{\mathcal{T}}, (\theta\mathbf{b})_{\mathcal{T}}\rangle = \langle(\theta\mathbf{A})^T \operatorname{sign}(\mathbf{A}\mathbf{x}), (\theta\mathbf{b})_{\mathcal{T}}\rangle \\ &= \langle \operatorname{sign}(\mathbf{A}\mathbf{x}), (\theta\mathbf{A})(\theta\mathbf{b})_{\mathcal{T}}\rangle \le \|(\theta\mathbf{A})(\theta\mathbf{b})_{\mathcal{T}}\|_1 \\ &\le \frac{\|(\theta\mathbf{b})_{\mathcal{T}}\|_{2,1}}{\sqrt{L}} + \delta \|(\theta\mathbf{b})_{\mathcal{T}}\|_2 \le (1+\delta) \|(\theta\mathbf{b})_{\mathcal{T}}\|_2 . \end{aligned}$$

Hence, we have $\|(\theta\mathbf{b})_{\mathcal{T}}\|_2 \le 1 + \delta$ and

$$\langle(\theta\mathbf{b})_{\mathcal{T}}, \mathbf{x}\rangle = \langle \operatorname{sign}(\mathbf{A}\mathbf{x}), (\theta\mathbf{A})\mathbf{x}\rangle = \|(\theta\mathbf{A})\mathbf{x}\|_1 \ge \frac{\|\mathbf{x}\|_{2,1}}{\sqrt{L}} - \delta \|\mathbf{x}\|_2 = (1-\delta),$$

where we used that $\|\mathbf{x}\|_{2,1} = \sqrt{L}\,\|\mathbf{x}\|_2 = \sqrt{L}$ by assumption. We can conclude that

$$\left\| \left((\theta\mathbf{A})^T \operatorname{sign}(\mathbf{A}\mathbf{x}) \right)_{\mathcal{T}} - \mathbf{x} \right\|_2^2 \le (1+\delta)^2 - 2(1-\delta) + 1 \le 5\delta.$$

■

Proof of Theorem 4.4.1 : Choose $mL \gtrsim \delta^{-2} 2s(\log(en/(2s)) + L)$ such that by Lemma 4.3.2, $\theta\mathbf{A}$ satisfies the $\ell_1/\ell_{2,1}$-RIP on $\mathcal{K}_{2s,L}$ with high probability. Let $\mathcal{T} = \operatorname{supp}(\mathbf{x})$ and $\hat{\mathcal{T}} = \operatorname{supp}(\hat{\mathbf{x}})$ where $\hat{\mathbf{x}} = \tilde{\mathbb{H}}_s((\theta\mathbf{A})^T\mathbf{y})$. Note that $\hat{\mathbf{x}}$ is also the best s-row approximation of $((\theta\mathbf{A})^T\mathbf{y})_{\mathcal{T}\cup\hat{\mathcal{T}}}$. Hence,

$$\begin{aligned} \|\mathbf{x} - \hat{\mathbf{x}}\|_2 &\le \left\| ((\theta\mathbf{A})^T\mathbf{y})_{\mathcal{T}\cup\hat{\mathcal{T}}} - \hat{\mathbf{x}} \right\|_2 + \left\| ((\theta\mathbf{A})^T\mathbf{y})_{\mathcal{T}\cup\hat{\mathcal{T}}} - \mathbf{x} \right\|_2 \\ &\le 2 \left\| ((\theta\mathbf{A})^T\mathbf{y})_{\mathcal{T}\cup\hat{\mathcal{T}}} - \mathbf{x} \right\|_2 \le 2\sqrt{5\delta}. \end{aligned}$$

where we applied Lemma 4.4.2 for $\mathcal{K}_{2s,L}$ in the last inequality (note that $|\mathcal{T} \cup \hat{\mathcal{T}}| \le 2s$). ■

One might argue that the proof of Theorem 4.4.1 hardly differs from the one of Theorem 3.3.2. First, an ℓ_1/ℓ_2-RIP for $\theta\mathbf{A}$ is proven and then a simple computation shows the claim. However, the model selection $\mathcal{S}_{s,L}$ is crucial and one has to treat the matrix $\theta\mathbf{A}$ as a whole to reach the sample complexity in (4.12). Comparison with the following naive approach clarifies this point: If we have $m \gtrsim \delta^{-2}s\log(eN/s)$ for some $\delta > 0$, we know that for each $l \in [L]$ and Gaussian $\mathbf{A}^{(l)} \in \mathbb{R}^{m\times N}$ with probability exceeding $1 - C\exp(-c\delta^2 m)$

$$(1-\delta)\left\|\mathbf{z}\right\|_2 \leq \frac{\sqrt{2}}{m\sqrt{\pi}}\left\|\mathbf{A}^{(l)}\mathbf{z}\right\|_1 \leq (1+\delta)\left\|\mathbf{z}\right\|_2, \tag{4.14}$$

for all s-sparse $\mathbf{z} \in \mathbb{R}^N$ (see [148]). Applying a union bound and summing over (4.14) for $l \in [L]$ one has with probability at least $1 - C\exp(-c\delta^2 m + \log(L))$ that

$$(1-\delta)\left\|\mathbf{x}\right\|_2 \leq \left\|(\theta\mathbf{A})\mathbf{x}\right\|_1 \leq (1+\delta)\left\|\mathbf{x}\right\|_2,$$

for all $\mathbf{x} \in \mathcal{S}_{s,L}$ and $\theta = \sqrt{2}/(m\sqrt{\pi L})$. The specific choice $\delta' = \sqrt{L}\delta$ leads to comparable probabilities of success and shows that this straight-forward approach causes a worse sample complexity than (4.12).

When proving Theorem 4.4.1 we rely on the assumption that $\mathbf{x} \in \mathcal{S}_{s,L}$ which corresponds to the equivalence of $\ell_1/\ell_{2,1}$-RIP and ℓ_1/ℓ_2-RIP on $\mathcal{S}_{s,L}$ (see Section 4.3). One may relax the restriction a little by defining, for $\varepsilon \in (0,1)$, the set

$$\mathcal{S}_\varepsilon = \left\{\mathbf{z} = \operatorname{vec}(\mathbf{Z}) : \mathbf{Z} \in \mathbb{R}^{N\times L},\ \operatorname{supp}(\mathbf{Z}) \leq s,\ \|\mathbf{z}_l\|_2 \in \left[\frac{1-\varepsilon}{\sqrt{L}}\left\|\mathbf{z}\right\|_2, \frac{1+\varepsilon}{\sqrt{L}}\left\|\mathbf{z}\right\|_2\right]\right\}$$

of signal ensembles which differ in norm by a bounded perturbation. Let $\mathbf{B}$ be a matrix which satisfies the $\ell_1/\ell_{2,1}$-RIP on $\mathcal{K}_{s,L}$ with RIP-constant $\delta > 0$. As $\left\|\mathbf{x}\right\|_{2,1} \in [1-\varepsilon, 1+\varepsilon]\sqrt{L}\left\|\mathbf{x}\right\|_2$ if $\mathbf{x} \in \mathcal{S}_\varepsilon$, this implies

$$(1-\delta)(1-\varepsilon)\left\|\mathbf{x}\right\|_2 \leq \left\|\mathbf{B}\mathbf{x}\right\|_1 \leq (1+\delta)(1+\varepsilon)\left\|\mathbf{x}\right\|_2. \tag{4.15}$$

If we wish to express (4.15) as an ℓ_1/ℓ_2-RIP on $\mathcal{S}_\varepsilon$ for some $\delta' \in (0,1)$, i.e.,

$$(1-\delta')\left\|\mathbf{x}\right\|_2 \leq \left\|\mathbf{B}\mathbf{x}\right\|_1 \leq (1+\delta')\left\|\mathbf{x}\right\|_2, \tag{4.16}$$

for all $\mathbf{x} \in \mathcal{S}_\varepsilon$, it would suffice to choose δ' such that

$$(1-\delta') \leq (1-\delta)(1-\varepsilon)$$

which can be reformulated as

$$\varepsilon \leq \frac{\delta'-\delta}{1-\delta}.$$

Since the right-hand side is positive only for $\delta < \delta'$ and a decreasing function in δ for $0 \leq \delta < \delta'$ it can be upper bounded by δ'. The more general $\ell_1/\ell_{2,1}$-RIP thus becomes an ℓ_1/ℓ_2-RIP on $\mathcal{S}_\varepsilon$ for $\varepsilon \leq \delta'$ meaning that perturbations ε are only tolerated if they are sufficiently small as compared to the aspired approximation error in (4.13). Anyway, the assumption of all signals $\mathbf{x}_l$ sharing the same norm is a mild condition in our setting as the one-bit model (3.8) is blind to scaling and norm variations in signal ensembles.

Algorithm 4 : sHT$(\mathbf{y}, \mathbf{A}, s)$

```
Require: $\mathbf{Y} \in \{-1,1\}^{m\times L}$, $\mathbf{A} \in \mathbb{R}^{mL\times NL}$
$\hat{\mathbf{x}} \leftarrow \tilde{\mathbb{H}}_s(\mathbf{A}^T \mathrm{vec}(\mathbf{Y}))$        ▷ $\tilde{\mathbb{H}}_s$ is defined in (4.5)
$\hat{\mathbf{X}} \leftarrow$ reshape$(\mathbf{x}, n, L)$        ▷ reshape$(\cdot)$ reverses vec$(\cdot)$
return $\hat{\mathbf{X}}$
```

4.5 Numerical Simulation

Let us conclude the chapter by numerically illustrating the theoretical results of Section 4.4. We recover an unknown signal ensemble $\mathbf{X} \in \mathbb{R}^{N\times L}$ from its one-bit measurements $\mathbf{Y} \in \mathbb{R}^{m\times L}$ by a single hard-thresholding step which needs the measurements $\mathbf{Y}$, the block diagonal measurement matrix $\mathbf{A}$ and the sparsity level $s = |\operatorname{supp}(\mathbf{X})|$. Algorithm 4 presents this simple approximation procedure. We show two experiments which substantiate the asymptotically linear dependence of $m = \mathcal{O}(s)$ measurements per signal. As required in Lemma 4.3.2, the block diagonal measurement matrix $\mathbf{A}$ has iid Gaussian entries and is scaled by $\theta = \sqrt{\pi/(2Lm^2)}$. To create signal ensembles $\mathbf{X} \in \mathbb{R}^{N\times L}$ with $|\operatorname{supp}(\mathbf{X})| = s$, we draw a support set $\mathcal{T} \subset [N]$ uniformly at random, determine the single entries as iid Gaussians of mean 0 and variance 1, and finally normalize all single signals $\mathbf{x}_l$, $l \in [L]$.

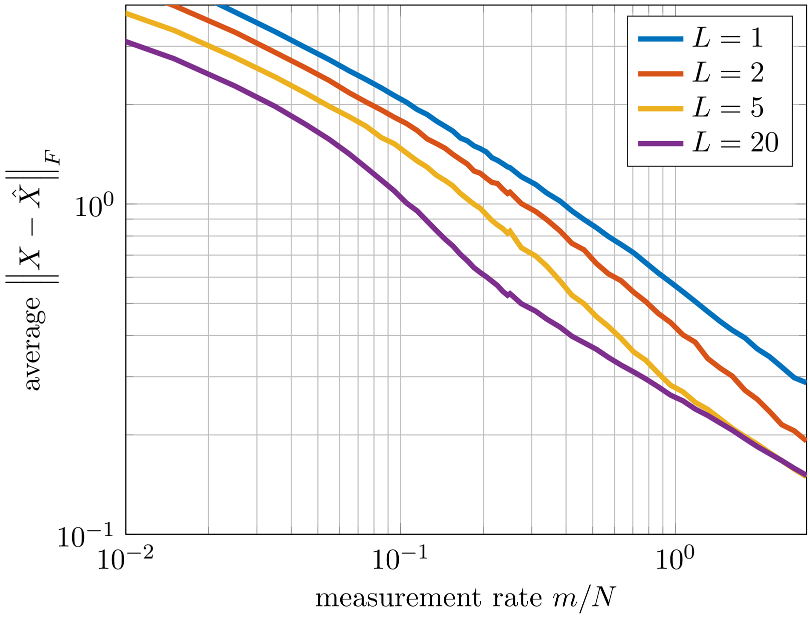


Figure 4.1: Simulated error $\|\mathbf{X} - \hat{\mathbf{X}}\|_F$ averaged over 500 experiments for $s = 5$ and $N = 100$.

In the first experiment we recover signal ensembles $\mathbf{X} \in \mathbb{R}^{N\times L}$ of signal dimension $N = 100$, ensemble size $L = 1, 2, 5, 20$, and support size $s = 5$ from their one-bit measurements $\mathbf{Y} \in \mathbb{R}^{m \times L}$. Figure 4.1 depicts the obtained approximation error $\|\mathbf{X} - \hat{\mathbf{X}}\|_F$ in Frobenius norm over the measurement rate $r = m/N$ (averaged over 500 random realizations of $\mathbf{X}$). One clearly observes an improvement for larger ensembles and a sharper transition from no recovery to practical approximation.

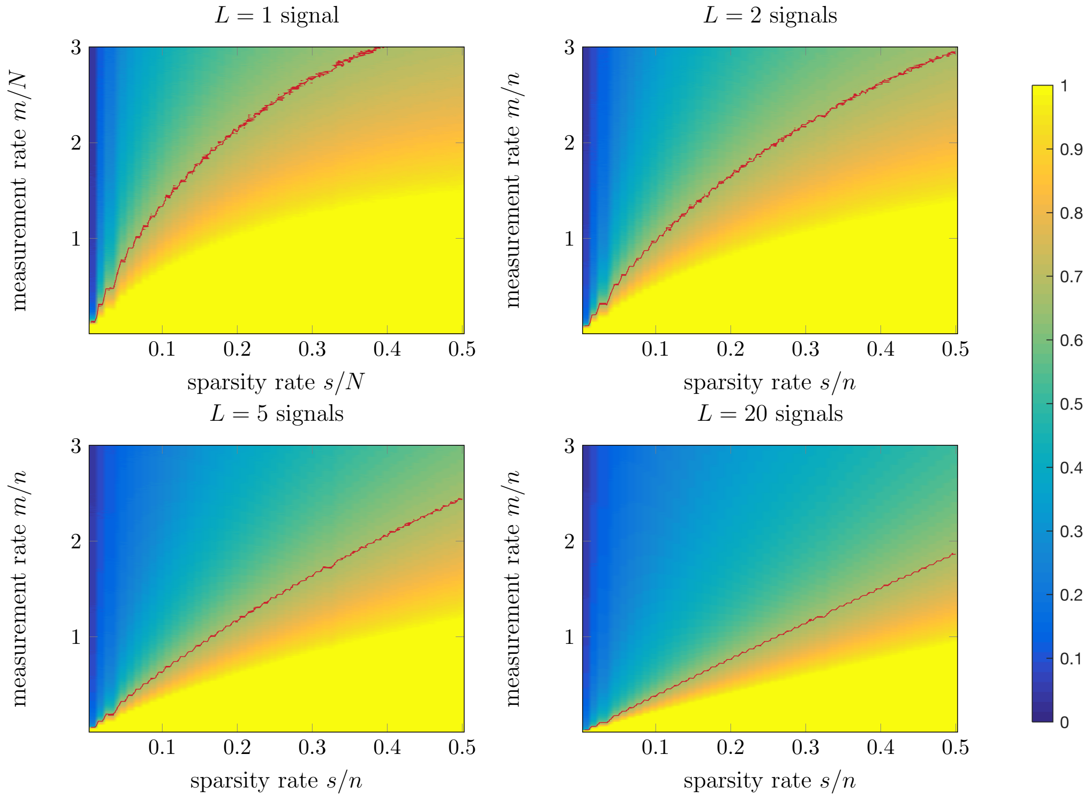


Figure 4.2: Simulated error $\|\mathbf{X} - \hat{\mathbf{X}}\|_F$ averaged over 500 experiments with $N = 100$. The red contour lines correspond to $\|\mathbf{X} - \hat{\mathbf{X}}\|_F = 2/3$.

The second experiment (see Figure 4.2) illustrates the dependence of m and s. We again approximate signal ensembles $\mathbf{X} \in \mathbb{R}^{N \times L}$ of signal dimension $N = 100$ and ensemble size $L = 1, 2, 5, 20$ from their one-bit measurements $\mathbf{Y} \in \mathbb{R}^{m \times L}$. The support size of $\mathbf{X}$ is varied from $s = 1$ to $s = 50$ while the measurement rate $r = m/N$ ranges from $r = 0.01$ up to $r = 3$. For any parameter pair (s, r) we recovered 500 random realizations of $\mathbf{X}$. The average approximation error $\|\mathbf{X} - \hat{\mathbf{X}}\|_F$ is plotted in color while a selected error level is highlighted. Comparing the different choices of L, the linear dependence of m on s for $L = 20$ and fixed error levels is clearly visible and different from the $s \ln(eN/s)$ behavior for $L = 1$.

Note that the measurement rate does not behave linearly in the plots $L = 2$ and $L = 5$ for $s/N \geq e^{1-L}$ which corresponds to the case $L \geq \log(eN/s)$. Though we claimed an $\mathcal{O}(s)$ behaviour in Section 4.3 in this case, the observation is no contradiction as in the theoretical argument it suffices to bound $(\log(eN/s) + L)/L \leq 2$. The numerical experiments with fixed L, however, show the transition from $(\log(eN/s) + L)/L \approx 2$ for small values of s/N (corresponding to large values of $\log(eN/s)$) to $(\log(eN/s)+L)/L \approx 1$ for large values of s/N (corresponding to small values of $\log(eN/s)$) causing a non-linear shape as long as L is not clearly dominating (cf. $L = 20$).

4.6 Discussion

After having examined in this chapter how the two concepts of heavily quantized measurements and distributed compressed sensing can be combined to obtain performance bounds for one-bit measurements which match corresponding unquantized distributed compressed sensing results, we see several possible directions for further research.

First, sophisticated alternatives to a single hard-thresholding step have been proposed (see [103]) which numerically outperform Algorithm 4. Explaining their improved performance in theory seems promising.

Second, the proof of Theorem 4.4.1 relies on noiseless measurements to use the equivalence of $\langle \text{sign}(\mathbf{Ax}), \mathbf{Ax} \rangle$ and $\|\mathbf{Ax}\|_1$. We suppose that it is not possible to easily modify the above proof to tolerate noise on $\mathbf{y}$. However, the noisy one-bit algorithm (3.23) could be applied in the distributed setting as well. Adapting Theorem 3.3.10 to block-diagonal measurement matrices would then lead to corresponding theoretical approximation guarantees. Moreover, the dithered one-bit model (3.10) might allow to generalize the results to subgaussian measurements.

Finally, relaxing the quantization level to multi-bit quantizers is desirable as it should decrease the approximation error (cf. [96, 94]) and thus bridges the wide performance gap between unquantized measurements and one-bit measurements.

Chapter 5

One-bit Recovery on General Manifolds

In this chapter we propose an algorithm to recover signals lying on a low-dimensional manifold $\mathcal{M} \subset \mathbb{S}^{N-1}$ from their one-bit measurements. The recovery strategies presented in Section 3.3.3 work for general signal sets $K \subset \mathbb{R}^N$. However, they become intractable if K is non-convex. One could relax K to its convex hull but there are two drawbacks. First, it is not clear if $\mathrm{conv}(K)$ can be computed in a simple way and, second, one might throw away a lot of structural information. Our approach combines geometric multi-resolution analysis, a locally linear manifold approximation, with the noisy one-bit results in Section 3.3.3 to guarantee tractable recovery even for non-convex signal sets K. After introducing geometric multi-resolution analysis, we present our algorithm and analyze its performance. Finally, we support the theoretical results by numerical simulations. The content of this chapter is joint work with Mark Iwen, Felix Krahmer, and Sara Krause-Solberg and has been published in [91] and [108].

5.1 The Geometric Multi-Resolution Analysis

Learning manifolds from samples and representing them in an efficient way is an important problem in many fields, ranging from image processing [86, 35], to analysis of electroencephalography signals [142] and computer vision [156]. It has strong connections to estimation of intrinsic dimensionality of point clouds [121, 37] and constructing dictionaries adapted to data [2]. In [5] the authors approached this problem by introducing the *geometric multi-resolution analysis (GMRA)*, a locally linear approximation of manifolds. The GMRA of a d-dimensional manifold $\mathcal{M}$ has a number of refinement levels which approximate $\mathcal{M}$ on different scales by anchor points and corresponding affine d-dimensional spaces. Figure 5.1 illustrates two different refinement levels for a simple 1-dimensional manifold. To make this precise, we first present an axiomatic definition of GMRA as given in [89]. This definition proves useful in deducing theoretical results but lacks connection to concrete applications where the structure of $\mathcal{M}$ is not known a priori. Hence, in the following we describe the original definition of a probabilistic GMRA which is well approximated by empirical data (see [5, 32, 123]) and connected to the axiomatic

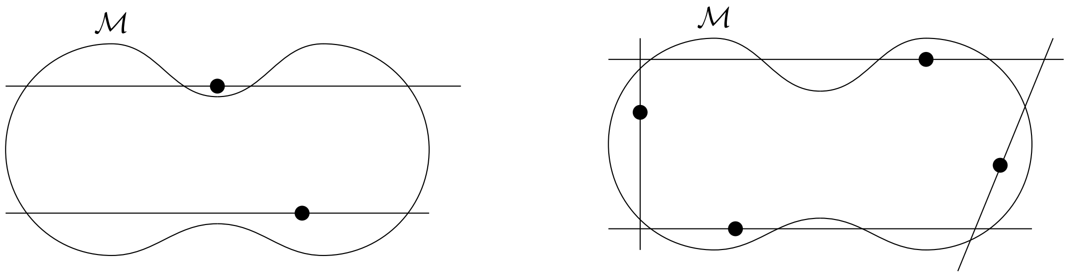


Figure 5.1: Two possible refinement levels of GMRA.

definition by applying results from [123].

5.1.1 Axiomatic GMRA

Let us begin with the axiomatic definition of GMRA as given in [89]. We denote the *tube of radius* r around a given subset $\mathcal{M} \subset \mathbb{R}^N$ by

$$\operatorname{tube}_r(\mathcal{M}) := \left\{ \mathbf{x} \in \mathbb{R}^N : \inf_{\mathbf{y} \in \mathcal{M}} \|\mathbf{x} - \mathbf{y}\|_2 \leq r \right\} \tag{5.1}$$

and let $\mathcal{M} \subset \mathbb{R}^N$ from now on be a d-dimensional manifold.

Definition 5.1.1 (GMRA Approximation to $\mathcal{M}$, [89]). *Let $J \in \mathbb{N}$ and $K_0, K_1, ..., K_J \in \mathbb{N}$. Then a* geometric multi resolution analysis (GMRA) *of $\mathcal{M}$ is a collection $\{(\mathcal{C}_j, \mathcal{P}_j)\}$, $j \in [J]$, of sets $\mathcal{C}_j = \{\mathbf{c}_{j,k}\}_{k=1}^{K_j} \subset \mathbb{R}^N$ of centers and*

$$\mathcal{P}_j = \left\{ \mathbb{P}_{j,k} : \mathbb{R}^N \to \mathbb{R}^N \mid k \in [K_j] \right\}$$

of affine projectors which approximate $\mathcal{M}$ at scale j, such that the following assumptions (1)-(3) *hold.*

(1) ***Affine Projections:*** *Every $\mathbb{P}_{j,k} \in \mathcal{P}_j$ has both an associated center $\mathbf{c}_{j,k} \in \mathcal{C}_j$ and an orthogonal matrix $\Phi_{j,k} \in \mathbb{R}^{d \times N}$, such that*

$$\mathbb{P}_{j,k}(\mathbf{z}) = \Phi_{j,k}^T \Phi_{j,k}(\mathbf{z} - \mathbf{c}_{j,k}) + \mathbf{c}_{j,k},$$

i.e., $\mathbb{P}_{j,k}$ is the projector onto some affine d-dimensional linear subspace $P_{j,k}$ containing $\mathbf{c}_{j,k}$.

(2) ***Dyadic Structure:*** *The number of centers at each level is bounded by $|\mathcal{C}_j| = K_j \leq C_{\mathcal{C}} 2^{dj}$ for an absolute constant $C_{\mathcal{C}} \geq 1$. There exist $C_1 > 0$ and $C_2 \in (0, 1]$, such that following conditions are satisfied:*

(a) $K_j \leq K_{j+1}$, for all $j \in [J-1]$.

(b) $\|\mathbf{c}_{j,k_1} - \mathbf{c}_{j,k_2}\|_2 > C_1 \cdot 2^{-j}$, for all $j \in [J]$ and $k_1 \neq k_2 \in [K_j]$.

(c) For each $j \in [J]\backslash\{0\}$ there exists a parent function $p_j : [K_j] \to [K_{j-1}]$ with

$$\|\mathbf{c}_{j,k} - \mathbf{c}_{j-1,p_j(k)}\|_2 \leq C_2 \cdot \min_{k' \in [K_{j-1}]\backslash\{p_j(k)\}} \|\mathbf{c}_{j,k} - \mathbf{c}_{j-1,k'}\|_2.$$

(3) ***Multiscale Approximation:*** *The projectors in $\mathcal{P}_j$ approximate $\mathcal{M}$ at scale j, i.e., when $\mathcal{M}$ is sufficiently smooth the affine spaces $P_{j,k}$ locally approximate $\mathcal{M}$ pointwise with error $\mathcal{O}\left(2^{-2j}\right)$. More precisely:*

(a) *There exists $j_0 \in [J-1]$, such that $\mathbf{c}_{j,k} \in \mathrm{tube}_{C_1 \cdot 2^{-j-2}}(\mathcal{M})$, for all $j > j_0 \geq 1$ and $k \in [K_j]$.*

(b) *For each $j \in [J]$ and $\mathbf{z} \in \mathbb{R}^N$ let $\mathbf{c}_{j,k_j(\mathbf{z})}$ be one of the centers closest to $\mathbf{z}$, i.e.,*

$$k_j(\mathbf{z}) \in \operatorname*{arg\,min}_{k \in [K_j]} \|\mathbf{z} - \mathbf{c}_{j,k}\|_2. \tag{5.2}$$

Then, for each $\mathbf{z} \in \mathcal{M}$ there exists a constant $C_\mathbf{z} > 0$ such that

$$\|\mathbf{z} - \mathbb{P}_{j,k_j(\mathbf{z})}(\mathbf{z})\|_2 \leq C_\mathbf{z} \cdot 2^{-2j},$$

for all $j \in [J]$. Moreover, for each $\mathbf{z} \in \mathcal{M}$ there exists $\tilde{C}_\mathbf{z} > 0$ such that

$$\|\mathbf{z} - \mathbb{P}_{j,k'}(\mathbf{z})\|_2 \leq \tilde{C}_\mathbf{z} \cdot 2^{-j},$$

for all $j \in [J]$ and $k' \in [K_j]$ satisfying

$$\|\mathbf{z} - \mathbf{c}_{j,k'}\|_2 \leq 16 \cdot \max\left\{\|\mathbf{z} - \mathbf{c}_{j,k_j(\mathbf{z})}\|_2, C_1 \cdot 2^{-j-1}\right\}.$$

By property (1), GMRA is a combination of several anchor points (the centers $\mathbf{c}_{j,k}$) and corresponding low dimensional affine spaces $P_{j,k}$. The levels j control the accuracy of the approximation. All centers are organized in a tree-like structure, cf. property (2). Property (3) then characterizes approximation criteria to be fulfilled on different refinement levels. Note that centers do not have to lie on $\mathcal{M}$ (compare Figure 5.1) but their distance to $\mathcal{M}$ is controlled by property (3a).
As the above axioms do not provide means of constructing a GMRA we now define it using probability measures on $\mathcal{M}$.

5.1.2 Probabilistic GMRA

A probabilistic GMRA of $\mathcal{M}$ with respect to a Borel probability measure Π, as introduced in [123], is a family of (piecewise linear) operators $\{\mathbb{P}_j \colon \mathbb{R}^N \to \mathbb{R}^N\}_{j\geq 0}$ of the form

$$\mathbb{P}_j(\mathbf{x}) = \sum_{k=1}^{K_j} \mathbb{1}_{C_{j,k}}(\mathbf{x}) \mathbb{P}_{j,k}(\mathbf{x}).$$

Here, $\mathbb{1}_M$ denotes the indicator function of a set M and, for each refinement level $j \geq 0$, the collection of pairs of measurable subsets and affine projections $\{(C_{j,k}, \mathbb{P}_{j,k})\}_{k=1}^{K_j}$ has the following structure.

The subsets $C_{j,k} \subset \mathbb{R}^N$ for $k = 1, \ldots, K_j$ form a partition of $\mathbb{R}^N$, i.e., they are pairwise disjoint and their union is $\mathbb{R}^N$. The affine projectors are defined by

$$\mathbb{P}_{j,k}(\mathbf{z}) = \mathbf{c}'_{j,k} + \mathbb{P}_{V_{j,k}}(\mathbf{z} - \mathbf{c}'_{j,k}),$$

where, for $\mathbf{z}_\Pi \sim \Pi$, $\mathbf{c}'_{j,k} = \mathsf{E}[\mathbf{z}_\Pi | \mathbf{z}_\Pi \in C_{j,k}] =: \mathsf{E}_{j,k}[\mathbf{z}_\Pi] \in \mathbb{R}^D$ and

$$V_{j,k} := \underset{\dim(V)=d}{\arg\min} \ \mathsf{E}_{j,k}\left[\|\mathbf{z}_\Pi - (\mathbf{c}'_{j,k} + \mathrm{Proj}_V(\mathbf{z}_\Pi - \mathbf{c}'_{j,k}))\|_2^2\right],$$

where the minimum is taken over all linear spaces V of dimension d. From now on we assume uniqueness of the subspaces $V_{j,k}$. If one thinks of Π being supported on the tube of a d-dimensional manifold, parallels to the axiomatic GMRA definition become clear. The axiomatic centers $\mathbf{c}_{j,k}$ are in this case considered to be approximately equal to the conditional means $\mathbf{c}'_{j,k}$ of some cells $C_{j,k}$ partitioning the space, and the corresponding affine projection spaces $P_{j,k}$ are spanned by eigenvectors of the d leading eigenvalues of the conditional covariance matrix

$$\mathbf{\Sigma}_{j,k} = \mathsf{E}_{j,k}\left[(\mathbf{z}_\Pi - \mathbf{c}'_{j,k})(\mathbf{z}_\Pi - \mathbf{c}'_{j,k})^T\right].$$

Defined in this way, the $\mathbb{P}_j$ correspond to projectors onto the GMRA approximations $\mathcal{M}_j$ introduced above if $\mathbf{c}_{j,k} = \mathbf{c}'_{j,k}$. From [123] we adopt the following assumptions on the entities defined above, and hence, on the distribution Π. When working with the probabilistic GMRA, we assume for all integers $j_{min} \leq j \leq j_{max}$ that **(A1)-(A4)** (see Table 5.1) hold true.

Assumption **(A1)** ensures that each partition element contains a reasonable amount of Π-mass. Assumption **(A2)** guarantees that all samples from $\Pi_{j,k}$ will lie close to its expection/center. As a result, each $\mathbf{c}'_{j,k}$ must be geometrically central within $C_{j,k}$. Together, **(A1)** and **(A2)** have the combined effect of ensuring that the probability mass of Π is equally distributed onto the different sets $C_{j,k}$, i.e., the number of points in each set $C_{j,k}$ is approximately the same, at each scale j. The third and fourth assumptions **(A3)** and **(A4)** constrain the geometry of the support of Π to being effectively d-dimensional and regular (e.g., close to a smooth d-dimensional submanifold of $\mathbb{R}^N$). We refer the reader to [123] for more detailed information regarding these assumptions.

An important class of probability measures Π fulfilling **(A1)-(A4)** is presented in [123]. For illustration of applicability we repeat it here and also discuss a method of constructing the partitions $\{C_{jk}\}_{k=1}^{K_j}$ from such probabilities measures. Let $\mathcal{M}$ be a smooth d-dimensional submanifold of $\mathbb{R}^N$ for the rest of this section.

Definition 5.1.2 ([123, Definition 3]). *Assume that $0 \leq \sigma < \tau$. The distribution Π is said to satisfy the (τ, σ)-*model assumption *if (i) there exists a smooth, compact submanifold $\mathcal{M} \hookrightarrow \mathbb{R}^N$ with reach τ such that* $\mathrm{supp}(\Pi) = \mathrm{tube}_\sigma(\mathcal{M})$, *$(ii)$ the distributions Π and $\mathcal{U}_{\mathrm{tube}_\sigma(\mathcal{M})}$ are absolutely continuous with respect to each other so the Radon-Nikodym derivative $\frac{\mathrm{d}\Pi}{\mathrm{d}\mathcal{U}_{\mathrm{tube}_\sigma(\mathcal{M})}}$ exists and satisfies*

$$0 < \phi_1 \leq \frac{\mathrm{d}\Pi}{\mathrm{d}\mathcal{U}_{\mathrm{tube}_\sigma(\mathcal{M})}} \leq \phi_2 < \infty \qquad \mathcal{U}_{\mathrm{tube}_\sigma(\mathcal{M})} - \textit{almost surely.}$$

The constants ϕ_1 and ϕ_2 are implicitly assumed to only depend on a slowly growing function of N, compare [123, Remark 4].

(A1) There exists an integer $1 \leq d \leq N$ and a positive constant $\theta_1 = \theta_1(\Pi)$ such that for all $k = 1, \dots, K_j$,

$$\Pi(C_{j,k}) \geq \theta_1 2^{-jd}.$$

(A2) Define the restricted measure $\Pi_{j,k}$ by $\Pi_{j,k}(S) := \Pi(S \cap C_{j,k})/\Pi(C_{j,k})$ for measurable $S \subset \mathbb{R}^N$. There is a positive constant $\theta_2 = \theta_2(\Pi)$ such that for all $k = 1, \dots, K_j$, if $\mathbf{z}_\Pi$ is drawn from $\Pi_{j,k}$ then, $\Pi_{j,k}$-almost surely,

$$\|\mathbf{z}_\Pi - \mathbf{c}'_{j,k}\|_2 \leq \theta_2 2^{-j}.$$

(A3) Denote the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}_{j,k}$ by $\lambda_1^{j,k} \geq \dots \geq \lambda_N^{j,k} \geq 0$. Then there exists $\sigma = \sigma(\Pi) \geq 0$, $\theta_3 = \theta_3(\Pi)$, $\theta_4 = \theta_4(\Pi) > 0$, and some $\alpha > 0$ such that for all $k = 1, \dots, K_j$,

$$\lambda_d^{j,k} \geq \theta_3 \frac{2^{-2j}}{d} \quad \text{and} \quad \sum_{l=d+1}^{N} \lambda_l^{j,k} \leq \theta_4(\sigma^2 + 2^{-2(1+\alpha)j}) \leq \frac{1}{2}\lambda_d^{j,k}.$$

(A4) There exists $\theta_5 = \theta_5(\Pi)$ such that

$$\|\mathbf{Id} - \mathbb{P}_j\|_{\infty,\Pi} \leq \theta_5(\sigma + 2^{-(1+\alpha)j}),$$

where $\|T\|_{\infty,\Pi} = \sup_{x \in \operatorname{supp}(\Pi)} \|T(x)\|_2$, for $T \colon \mathbb{R}^N \to \mathbb{R}^N$.

Table 5.1: The assumption set on Π.

Obviously, the uniform distribution on a smooth compact submanifold of $\mathbb{R}^N$ or its σ-tube satisfies the (τ, σ)-model assumption. Let us now discuss the construction of suitable partitions $\{C_{j,k}\}$ by making use of cover trees. A *cover tree* T on a finite set of samples $S \subset \mathcal{M}$ is a hierarchy of levels with the starting level containing the root point and the last level containing every point in S. To every level a set of nodes is assigned which is associated with a subset of points in S. To be precise, given a set S of n distinct points in some metric space $(\mathbb{X}, d_{\mathbb{X}})$, a cover tree T on S is a sequence of subsets $T_i \subset S, i = 0, 1, \dots$ that satisfies the following, see [20]:

(i) **Nesting:** $T_i \subset T_{i+1}$, i.e., once a point appears in T_i it is in every T_j for $j \geq i$.

(ii) **Covering:** For every $\mathbf{x} \in T_{i+1}$ there exists exactly one $\mathbf{y} \in T_i$ such that $d_{\mathbb{X}}(\mathbf{x}, \mathbf{y}) \leq 2^{-i}$. Here $\mathbf{y}$ is called the *parent* of $\mathbf{x}$.

(iii) **Separation:** For all distinct points $\mathbf{x}, \mathbf{y} \in T_i$, $d_{\mathbb{X}}(\mathbf{x}, \mathbf{y}) > 2^{-i}$.

The set T_i denotes the set of points in S associated with nodes at level i. Note that there exists $n \in \mathbb{N}$ such that $T_i = S$ for all $i \geq n$. We assume that S is sufficiently large to contain an ε-cover of $\mathcal{M}$ for $\varepsilon > 0$ sufficiently small.
Note that the axioms characterizing cover trees are strongly connected to the dyadic structure of GMRA. For a given cover tree with respect to the Euclidean distance (for construction see [20]) on a set $\mathcal{X}_n = \{X_1, \dots X_n\}$ of i.i.d. samples from the distribution Π, let $\mathbf{a}_{j,k}$ for $k = 1, \dots, K_j$ be the elements of the jth level of the cover tree, i.e. $T_j = \{\mathbf{a}_{j,k}\}_{k=1}^{K_j}$ and define

$$\kappa_j(x) = \underset{1 \leq k \leq K_j}{\arg\min} \|\mathbf{x} - \mathbf{a}_{j,k}\|_2.$$

With this a partition of $\mathbb{R}^N$ into *Voronoi regions*

$$C_{j,k} = \{\mathbf{x} \in \mathbb{R}^N : \kappa_j(\mathbf{x}) = k\} \tag{5.3}$$

can be defined. Maggioni et. al. showed in [123, Theorem 7] that by this construction all assumptions **(A1)-(A4)** can be fulfilled.

The question arises if the properties of the axiomatic definition of GMRA in Definition 5.1.1 are equally met. As only parts of the axioms are relevant for our analysis, we refrain from giving rigorous justification for all properties.

1. GMRA property (1) holds by construction if the matrices $\Phi_{j,k}$ are defined, s.t. $\Phi_{j,k}^T \Phi_{j,k} = \mathbb{P}_{V_{j,k}}$ along with any reasonable choice of centers $\mathbf{c}_{j,k}$.
2. The dyadic structure axioms (2a) – (2c) also hold as a trivial consequence of the cover tree properties (i) – (iii) above if the axiomatic centers $\mathbf{c}_{j,k}$ are chosen to be the elements of the cover tree set T_j (i.e., the $\mathbf{a}_{j,k}$ elements). By the (ρ, σ)-model assumption samples drawn from Π will have a quite uniform distribution all over supp(Π). Hence, the probabilistic centers $\mathbf{c}'_{j,k}$ of each $C_{j,k}$-set will also tend to be close to the axiomatic centers $\mathbf{c}_{j,k} = \mathbf{a}_{j,k}$ proposed here for small σ (see, e.g., assumption **(A2)** above).

3. One can deduce GMRA property (3a) from the fact that our chosen centers $\mathbf{a}_{j,k}$ belong to $\mathcal{M}$ if supp(Π) = $\mathcal{M}$ (or to a small tube around $\mathcal{M}$ if σ is small).

4. The first part of (3b) is implied by **(A4)** with the uniform constant θ_5 for all $\mathbf{x} \in \mathcal{M}$ if $\mathbf{a}_{j,k}$ is sufficiently close to $\mathbf{c}'_{j,k}$. To show the second part of (3b) note that

$$\begin{aligned}\|\mathbf{x} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2 &\leq \|\mathbf{x} - \mathbf{c}_{j,k'}\|_2 + \|\mathbf{c}_{j,k'} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2 = \|\mathbf{x} - \mathbf{c}_{j,k'}\|_2 + \|\mathbb{P}_{V_{j,k'}}(\mathbf{x} - \mathbf{c}_{j,k'})\|_2 \\ &\leq 2\|\mathbf{x} - \mathbf{c}_{j,k'}\|_2 \leq 32 \max\{\|\mathbf{x} - \mathbf{c}_{j,k_j(\mathbf{x})}\|_2, C_1 2^{-j-1}\} \\ &\leq 32 \max\{C_\epsilon 2^{-j}, C_1 2^{-j-1}\} \leq C \cdot 2^{-j},\end{aligned}$$

where in the second last step we used our cover tree properties (recall that $\mathbf{c}_{j,k} = \mathbf{a}_{j,k}$). Again, the constants $C, C_\epsilon > 0$ do not depend on the chosen $\mathbf{x} \in \mathcal{M}$ as long as S is well chosen (e.g., contains a sufficiently fine cover of $\mathcal{M}$).

Considering the GMRA axioms above we can now see that only the first part of (3b) may not hold in a satisfactory manner if we choose to set $\Phi_{j,k}^T \Phi_{j,k} = \mathbb{P}_{V_{j,k}}$ and $\mathbf{c}_{j,k} = \mathbf{a}_{j,k}$. And, even when it doesn't hold with C_z being independent of j it will still at least still hold with a worse j dependence due to assumption **(A2)**.

5.1.3 Empirical GMRA

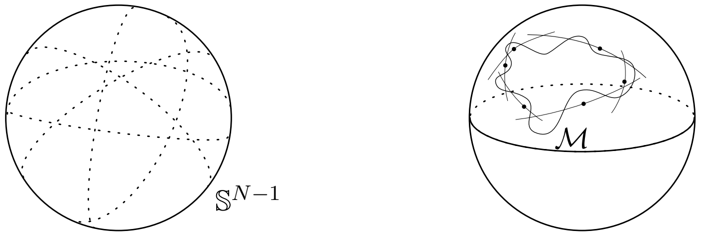
The axiomatic properties only hold as argued above if the probabilistic GMRA is constructed under knowledge of the true $\mathbb{P}_{V_{j,k}}$-subspaces. In reality this won't be the case and we are rather given training data consisting of n samples from $\mathcal{M}$, $\mathcal{X}_n = \{X_1, ..., X_n\}$, which we assume to be iid with distribution Π. These samples are used to approximate the real GMRA subspaces based on Π such that the operators $\mathbb{P}_j$ can be replaced by their estimators

$$\widehat{\mathbb{P}}_j(\mathbf{z}) = \sum_{k=1}^{K_j} \mathbb{1}_{\{\mathbf{z} \in C_{j,k}\}} \widehat{\mathbb{P}}_{j,k}(\mathbf{z}),$$

where $\{C_{j,k}\}_{k=1}^{K_j}$ is a suitable partition of $\mathbb{R}^N$ obtained from the data,

$$\begin{aligned}\widehat{\mathbb{P}}_{j,k}(\mathbf{z}) &= \widehat{\mathbf{c}}_{j,k} + \mathbb{P}_{\widehat{V}_{j,k}}(\mathbf{z} - \widehat{\mathbf{c}}_{j,k}), \\ \widehat{\mathbf{c}}_{j,k} &= \frac{1}{|\mathcal{X}_{k,j}|} \sum_{\mathbf{z} \in \mathcal{X}_{j,k}} \mathbf{z}, \\ \widehat{V}_{j,k} &= \operatorname*{arg\,min}_{\dim(V)=d} \frac{1}{|\mathcal{X}_{j,k}|} \sum_{\mathbf{z} \in \mathcal{X}_{j,k}} \|\mathbf{z} - \widehat{\mathbf{c}}_{j,k} - \mathbb{P}_V(\mathbf{z} - \widehat{\mathbf{c}}_{j,k})\|_2^2,\end{aligned}$$

and $\mathcal{X}_{j,k} = C_{j,k} \cap \mathcal{X}_n$. In other words, working with above model we have one perfect GMRA that cannot be computed (unless Π is known) but fulfills all important axiomatic properties, and an estimated GMRA that is at hand but that is only an approximation to the perfect one. Thankfully, the main results of [123] give error bounds on the difference between perfect and estimated GMRA with $\mathbf{c}_{j,k} = \widehat{\mathbf{c}}_{j,k} \approx \mathbf{c}'_{j,k} \approx \mathbf{a}_{j,k}$ that only depend on the number of samples from Π one can acquire.



(a) Tessellation of the sphere by random hyperplanes.

(b) Submanifold $\mathcal{M}$ of $\mathbb{S}^{N-1}$ and one level of GMRA.

Figure 5.2: One-bit measurements and GMRA.

Theorem 5.1.3 ([123, Theorem 2]). *Suppose that assumptions **(A1)-(A3)** are satisfied (see Table 5.1). Let $X, X_1, \dots X_n$ be an i.i.d. sample from Π and set $\bar{d} = 4d^2 \frac{\theta_2^4}{\theta_3^2}$. Then, for any $t \geq 1$ such that $t + \log(\max\{\bar{d}, 8\}) \leq \frac{1}{2}\theta_1 n 2^{-jd}$,*

$$\mathsf{E}\left[\|X - \widehat{\mathbb{P}}_j(X)\|_2^2\right] \leq 2\theta_4 \left(\sigma^2 + 2^{-2j(1+\alpha)}\right) + c_1 2^{-2j} \frac{(t + \log(\max\{\bar{d}, 8\}))d^2}{n2^{-jd}}$$

*and, if in addition **(A4)** is satisfied,*

$$\left\|\mathbf{Id} - \widehat{\mathbb{P}}_j\right\|_{\infty,\Pi} \leq \theta_5 \left(\sigma + 2^{-(1+\alpha)j}\right) + \sqrt{\frac{c_1}{2} 2^{-2j} \frac{(t + \log(\max\{\bar{d}, 8\}))d^2}{n2^{-jd}}}$$

with probability $\geq 1 - \frac{2^{jd+1}}{\theta_1}\left(e^{-t} + e^{-\frac{\theta_1}{16} n 2^{-jd}}\right)$, where $c_1 = 2\left(12\sqrt{2}\frac{\theta_2^3}{\theta_3\sqrt{\theta_1}} + 4\sqrt{2}\frac{\theta_2}{d\sqrt{\theta_1}}\right)^2$.

5.2 The One-Bit Manifold Sensing Algorithm

The problem we address is the following. We consider, for $d \ll N$, a given union of d-dimensional manifolds $\mathcal{M}$ that is a subset of the unit sphere $\mathbb{S}^{N-1}$ of a higher dimensional space $\mathbb{R}^N$ (by scaling blindness of (3.8) this restriction is reasonable). Furthermore, we imagine that we do not know $\mathcal{M}$ perfectly, and so instead we only have approximate information about $\mathcal{M}$ represented in terms of a GMRA of the manifold. Our goal is to recover an unknown signal $\mathbf{x} \in \mathcal{M}$ from m one-bit measurements

$$\mathbf{y} = \operatorname{sign}(\mathbf{A}\mathbf{x}), \tag{5.4}$$

where $\mathbf{A} \in \mathbb{R}^{m \times N}$ has Gaussian iid entries of variance $1/\sqrt{m}$, using as few measurements, m, as possible. In order to succeed using only $m \ll N$ such one-bit measurements we will use the fact that the GMRA provides structural constraints for the signal $\mathbf{x}$ to be recovered. Thus the setup connects to recent generalizations of the quantized compressed sensing problem [139] which we will exploit in our proof. Figure 5.2 illustrates the setting.

As one might expect, the complexity and structure of the GMRA for $\mathcal{M}$ will depend on the complexity of $\mathcal{M}$ itself. We work with two different measures of complexity. First, the Gaussian width and, second, the *reach* of $\mathcal{M}$ [60]. The Gaussian width and its geometrical understanding have comprehensively been discussed in Section 3.2. The notion of reach is, in contrast, more obviously linked to the geometry of $\mathcal{M}$. First, recall the definition of tube in (5.1). The domain of the nearest neighbor projection onto the closure of $\mathcal{M}$ is denoted by

$$D(\mathcal{M}) := \left\{ \mathbf{x} \in \mathbb{R}^N \ : \ \exists! \mathbf{y} \in \overline{\mathcal{M}} \text{ such that } \|\mathbf{x} - \mathbf{y}\|_2 = \inf_{\mathbf{z} \in \mathcal{M}} \|\mathbf{z} - \mathbf{y}\|_2 \right\}$$

and allows to define the reach of $\mathcal{M} \subset \mathbb{R}^N$ as

$$\operatorname{reach}(\mathcal{M}) := \sup\{ r \geq 0 \ : \ \operatorname{tube}_r(\mathcal{M}) \subset D(\mathcal{M}) \}.$$

The reach is thus the largest distance r around $\mathcal{M}$ for which the nearest neighbor projection onto the closure of $\mathcal{M}$ is well defined. Given this definition one sees, e.g., that the reach of any $d < N$ dimensional sphere of radius r in $\mathbb{R}^N$ is r, and that the reach of any $d \leq N$ dimensional convex subset of $\mathbb{R}^N$ is ∞. Note that the concept of reach is connected to the concept of manifold condition numbers in [12] where it was used to characterize embeddings of smooth manifolds by random linear projections.

To simplify notation we denote the *scale-j GMRA approximation to* $\mathcal{M}$ by

$$\mathcal{M}_j := \{ \mathbb{P}_{j,k_j(\mathbf{z})}(\mathbf{z}) \colon \mathbf{z} \in \mathcal{B}_2(\mathbf{0}, 2) \} \cap \mathcal{B}_2(\mathbf{0}, 2).$$

Note that we restrict the GMRA, for each fixed j, to the portions of the affine subspaces introduced in Definition 5.1.1 which are potentially relevant as approximations to some portion of $\mathcal{M} \subset \mathbb{S}^{N-1}$. To prevent the $\mathcal{M}_j$ above from being empty we will further assume in our results that we only use scales $j > j_0$ large enough to guarantee that $\operatorname{tube}_{C_1 2^{-j-2}}(\mathcal{M}) \subset \mathcal{B}_2(\mathbf{0}, 2)$. Hence we will have $\mathbf{c}_{j,k} \in \mathcal{B}_2(\mathbf{0}, 2)$ for all $k \in K_j$, and so $\mathcal{C}_j \subset \mathcal{M}_j$. This further guarantees that no sets $P_{j,k} \cap \mathcal{B}_2(\mathbf{0}, 2)$ are empty, and that $P_{j,k} \cap \mathcal{B}_2(\mathbf{0}, 2) \subset \mathcal{M}_j$ for all $k \in K_j$. The choice of $\mathcal{B}_2(\mathbf{0}, 2)$ is motivated by simplifying later calculations. One may instead take $\mathcal{B}_2(\mathbf{0}, r)$, for any $r > 1$. This only changes the required quality of the GMRA in Theorem 5.3.1 along with the constants E and E'.

Considering GMRA-based compressed sensing results in [89] and the noisy one-bit result Theorem 3.3.9, we suggest the following strategy for recovering an unknown $\mathbf{x} \in \mathcal{M}$ from the measurements given in (5.4): First, choose a center $\mathbf{c}_{j,k'}$ whose one-bit measurements agree with as many one-bit measurements of $\mathbf{x}$ as possible. As illustrated in Figure 5.3, this is not an optimal choice in general. Nevertheless, one can hope $P_{j,k'}$ to be a good approximation to $\mathcal{M}$ near $\mathbf{x}$. Thus, apply in the second step the noisy one-bit recovery method (3.21) on $P_{j,k'}$ to obtain an approximation of $\mathbb{P}_{j,k'}(\mathbf{x})$ which is close to $\mathbf{x}$. Note that in this second step the given measurements $\mathbf{y}$ of $\mathbf{x}$ are interpreted as noisy measurements of $\mathbb{P}_{j,k'}(\mathbf{x})$. The algorithm One-bit-Manifold-Sensing-Simple is stated explicitly in Algorithm 5.

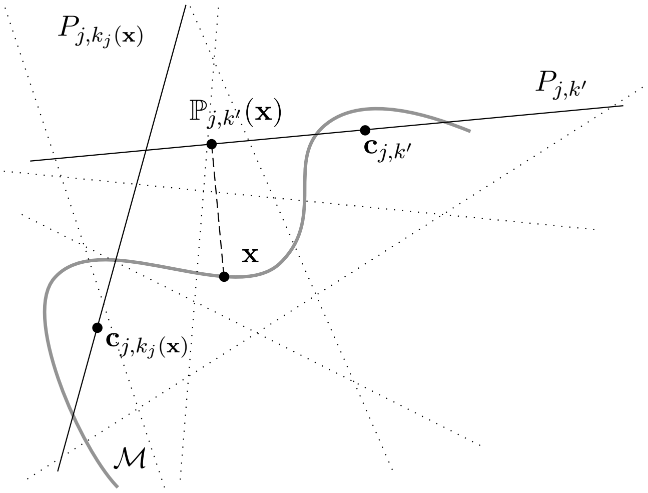


Figure 5.3: The closest center $\mathbf{c}_{j,k_j(\mathbf{x})}$ is not identified by measurements. Dotted lines represent one-bit hyperplanes.

Algorithm 5 : OMS-Simple

I. Identify a center $\mathbf{c}_{j,k'}$ close to $\mathbf{x}$ via

$$\mathbf{c}_{j,k'} \in \operatorname*{arg\,min}_{\mathbf{c}_{j,k} \in \mathcal{C}_j} \; d_H(\operatorname{sign}(\mathbf{A}\mathbf{c}_{j,k}), \mathbf{y}), \tag{5.5}$$

where d_H is the Hamming distance, i.e., $d_H(\mathbf{z}, \mathbf{z}') := |\{l : z_l \neq z'_l\}|$. If $d_H(\operatorname{sign}(\mathbf{A}\mathbf{c}_{j,k'}), \mathbf{y}) = 0$, directly choose $\mathbf{x}^* = \mathbf{c}_{j,k'}$ and omit II.

II. If there is no center in the same cell as $\mathbf{x}$ (as in Figure 5.3), solve the noisy one-bit recovery problem (3.21), i.e.,

$$\mathbf{x}^* = \operatorname*{arg\,min}_{\mathbf{z} \in \mathbb{R}^N} \sum_{l=1}^{m} (-y_l)\langle \mathbf{a}_l, \mathbf{z} \rangle, \quad \text{subject to } \mathbf{z} = \mathbb{P}_{j,k'}(\mathbf{z}) \text{ and } \|\mathbf{z}\|_2 \leq R, \tag{5.6}$$

where R is a suitably chosen parameter.

Remark 5.2.1. *The minimization in (5.5) can be efficiently calculated by exploiting tree structures in $\mathcal{C}_j$. Numerical experiments (see Section 5.5) suggest this strategy to yield adequate approximation for the center $\mathbf{c}_{j,k_j(\mathbf{x})}$ in (5.2), while being considerably faster (we observed differences in runtime up to a factor of 10).*

Though simple to understand, the constraints in (5.6) cause an issue we have to address: any optimal choice of R in (5.6) depends on $\mathbf{x}$ such that OMS-Simple requires parameter tuning, making it less practical than one might wish for.
We hence modify the constraints in (5.6) and instead minimize over the convex hull of the projection of $P_{j,k'} \cap \mathcal{B}_2(\mathbf{0}, 2)$ onto $\mathbb{S}^{N-1}$ (recall the shorthand notation $\mathbb{P}_{\mathbb{S}} = \mathbb{P}_{\mathbb{S}^{N-1}}$),

$$\operatorname{conv}\left(\mathbb{P}_{\mathbb{S}}(P_{j,k'} \cap \mathcal{B}_2(\mathbf{0}, 2))\right).$$

If $\mathbf{0} \in P_{j,k'}$ one has $\operatorname{conv}\left(\mathbb{P}_{\mathbb{S}}(P_{j,k'} \cap \mathcal{B}_2(\mathbf{0},2))\right) = P_{j,k'} \cap \mathcal{B}_2(\mathbf{0},1)$. If $\mathbf{0} \notin P_{j,k'}$ the set $\operatorname{conv}\left(\mathbb{P}_{\mathbb{S}}(P_{j,k'} \cap \mathcal{B}_2(\mathbf{0},2))\right)$ is described by the following set of convex constraints which are straightforward to implement in practice. Denote by $\mathbb{P}_{\mathbf{c}}$ the projection onto the linear space spanned by $\mathbf{c} = \mathbb{P}_{j,k'}(\mathbf{0})$. Then,

$$\mathbf{z} \in \operatorname{conv}\left(\mathbb{P}_{\mathbb{S}}(P_{j,k'} \cap \mathcal{B}_2(\mathbf{0},2))\right) \Leftrightarrow \begin{cases} \|\mathbf{z}\|_2 \leq 1, \\ \Phi_{j,k'}^T \Phi_{j,k'} \mathbf{z} + \mathbb{P}_{\mathbf{c}}(\mathbf{z}) = \mathbf{z}, \\ \langle \mathbf{z}, \mathbf{c} \rangle \geq \frac{1}{2}\|\mathbf{c}\|_2^2. \end{cases} \tag{5.7}$$

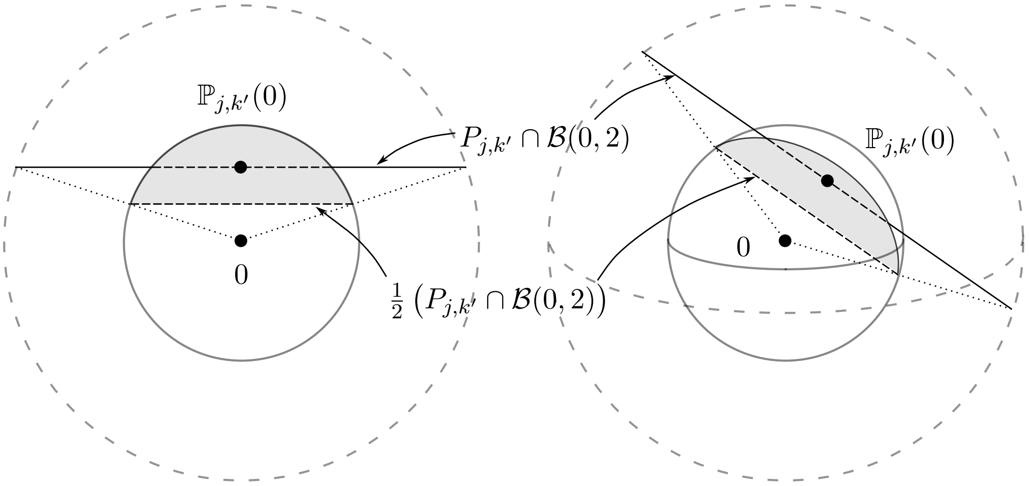

$\mathbb{P}_{j,k'}(0)$
$P_{j,k'} \cap \mathcal{B}(0,2)$
$\mathbb{P}_{j,k'}(0)$
0
0
$\frac{1}{2}\left(P_{j,k'} \cap \mathcal{B}(0,2)\right)$

Figure 5.4: Two views of an admissible set $\operatorname{conv}(\mathbb{P}_{\mathbb{S}}(P_{j,k'} \cap \mathcal{B}(0,2)))$ from (5.7) for a case with $\|\mathbf{c}\|_2 = \|\mathbb{P}_{j,k'}(\mathbf{0})\|_2 < 1$.

The first two conditions above restrict $\mathbf{z}$ to $\mathcal{B}_2(\mathbf{0},1)$ and $\operatorname{span}(P_{j,k'})$, respectively. The third condition then removes all points that are too close to the origin (see Figure 5.4). This is made explicit in the following lemma.

Lemma 5.2.2. *Let $P_{j,k'}$ be the affine subspace chosen in step I. of OMS-Simple. Define $\mathbf{c} = \mathbb{P}_{j,k'}(\mathbf{0})$. If $\mathbf{0} \notin P_{j,k'}$, the equivalence in* (5.7) *holds.*

Proof: First, assume $\mathbf{z} \in \operatorname{conv}\left(\mathbb{P}_{\mathbb{S}}(P_{j,k'} \cap \mathcal{B}_2(\mathbf{0},2))\right)$. Obviously, $\|\mathbf{z}\|_2 \leq 1$. As projecting onto the sphere is a simple rescaling, $\operatorname{conv}\left(\mathbb{P}_{\mathbb{S}}(P_{j,k'} \cap \mathcal{B}_2(\mathbf{0},2))\right) \subset \operatorname{span}(P_{j,k'})$ implying that $\Phi_{j,k'}^T \Phi_{j,k'} \mathbf{z} + \mathbb{P}_{\mathbf{c}}(\mathbf{z}) = \mathbf{z}$. For showing the third constraint note that any $\mathbf{z}' \in P_{j,k'}$ can be written as $\mathbf{z}' = \mathbf{c} + (\mathbf{z}' - \mathbf{c})$ where $\mathbf{z}' - \mathbf{c}$ is perpendicular to $\mathbf{c}$. If in addition $\|\mathbf{z}'\|_2 \leq 2$, we get

$$\langle \mathbb{P}_{\mathbb{S}}(\mathbf{z}'), \mathbf{c} \rangle = \left\langle \frac{\mathbf{z}'}{\|\mathbf{z}'\|_2}, \mathbf{c} \right\rangle = \frac{\langle \mathbf{c}, \mathbf{c} \rangle}{\|\mathbf{z}'\|_2} \geq \frac{1}{2}\|\mathbf{c}\|_2^2.$$

As $\mathbf{z}$ is a convex combination of different $\mathbb{P}_{\mathbb{S}}(\mathbf{z}')$ the constraint also holds for $\mathbf{z}$. Let $\mathbf{z}$ fulfill the three constraints. Then $\mathbf{z}' = (\|\mathbf{c}\|_2^2 / \langle \mathbf{z}, \mathbf{c} \rangle) \cdot \mathbf{z}$ satisfies $\mathbf{z}' \in P_{j,k'}$ because of the second constraint and $\langle \mathbf{z}', \mathbf{c} \rangle = \|\mathbf{c}\|_2^2$. Furthermore, by the first and third constraint $\|\mathbf{z}'\|_2 \leq (\|\mathbf{c}\|_2^2 / \langle \mathbf{z}, \mathbf{c} \rangle) \leq 2$ and hence $\mathbf{z}' \in P_{j,k'} \cap \mathcal{B}_2(\mathbf{0}, \|\mathbf{c}\|_2^2 / \langle \mathbf{z}, \mathbf{c} \rangle)$- $\subset P_{j,k'} \cap \mathcal{B}_2(\mathbf{0},2)$. As $P_{j,k'} \cap \mathcal{B}_2(\mathbf{0}, \|\mathbf{c}\|_2^2 / \langle \mathbf{z}, \mathbf{c} \rangle)$ is the convex hull of $P_{j,k'} \cap$

$(\|\mathbf{c}\|_2^2/\langle \mathbf{z},\mathbf{c}\rangle)\cdot\mathbb{S}^{N-1}$, there are $\mathbf{z}_1,...,\mathbf{z}_n \in P_{j,k'}$ and $\lambda_1,...,\lambda_n \geq 0$ with $\|\mathbf{z}_k\|_2 = \|\mathbf{c}\|_2^2/\langle \mathbf{z},\mathbf{c}\rangle$ and $\sum\lambda_k = 1$ such that $(\|\mathbf{c}\|_2^2/\langle \mathbf{z},\mathbf{c}\rangle)\cdot\mathbf{z} = \sum\lambda_k\mathbf{z}_k$. Hence, $\mathbf{z} = \sum\lambda_k(\langle \mathbf{z},\mathbf{c}\rangle/\|\mathbf{c}\|_2^2)\cdot\mathbf{z}_k$. As $(\langle \mathbf{z},\mathbf{c}\rangle/\|\mathbf{c}\|_2^2)\cdot\mathbf{z}_k \in \mathbb{P}_{\mathbb{S}}(P_{j,k'}\cap\mathcal{B}_2(\mathbf{0},2))$ we get $\mathbf{z} \in \operatorname{conv}\left(\mathbb{P}_{\mathbb{S}}(P_{j,k'}\cap\mathcal{B}_2(\mathbf{0},2))\right)$. ■

Our analysis uses that the noisy one-bit recovery results of Plan and Vershynin apply to arbitrary subsets of the unit ball $\mathcal{B}_2(\mathbf{0},1)\subset\mathbb{R}^N$ which will allow us to adapt our recovery approach. Replacing the constraints in (5.6) with those in (5.7) we obtain the following modified recovery approach, One-bit-Manifold-Sensing (see Algorithm 6).

Algorithm 6 : OMS

I. Identify a center $\mathbf{c}_{j,k'}$ close to $\mathbf{x}$ via

$$\mathbf{c}_{j,k'} \in \underset{\mathbf{c}_{j,k}\in\mathcal{C}_j}{\arg\min}\ d_H(\operatorname{sign}(A\mathbf{c}_{j,k}),\mathbf{y}). \tag{5.8}$$

where d_H is the Hamming distance, i.e., $d_H(\mathbf{z},\mathbf{z}') := |\{l : z_l \neq z'_l\}|$. If $d_H(\operatorname{sign}(\mathbf{A}\mathbf{c}_{j,k'}),\mathbf{y}) = 0$, directly choose $\mathbf{x}^* = \mathbf{c}_{j,k'}$ and omit II.

II. If there is no center lying in the same cell as $\mathbf{x}$ (see Figure 5.3), recover the projection of $\mathbf{x}$ onto $P_{j,k'}$, i.e., $\mathbb{P}_{j,k'}(\mathbf{x})$. To do so solve the convex optimization

$$\mathbf{x}^* = \underset{\mathbf{z}\in\mathbb{R}^N}{\arg\min}\sum_{l=1}^m(-y_l)\langle\mathbf{a}_l,\mathbf{z}\rangle, \quad \text{subject to } \mathbf{z}\in\operatorname{conv}\left(\mathbb{P}_{\mathbb{S}}(P_{j,k'}\cap\mathcal{B}_2(\mathbf{0},2))\right). \tag{5.9}$$

5.3 Recovery Guarantees for OMS

We are ready to state the core result of this chapter which provides for OMS recovery guarantees similar to the ones in Theorem 3.3.9 if one has access to a GMRA of $\mathcal{M}$ fulfilling the axioms in Definition 5.1.1. This section is devoted to the proof of Theorem 5.3.1 and concludes with a short discussion on the Gaussian width of Riemannian manifolds.

Theorem 5.3.1. *There exist absolute constants $E, E', c > 0$ such that the following holds. Let $\varepsilon\in(0,1)$ and assume a GMRA of $\mathcal{M}$ according to Definition 5.1.1 is given up to maximum refinement level $J \geq j := \lceil\log(1/\varepsilon)\rceil$. Further suppose that one has* $\operatorname{dist}(\mathbf{0},\mathcal{M}_j)\geq 1/2$*,* $0 < C_1 < 2^j$*, and* $\sup_{\mathbf{x}\in\mathcal{M}}\tilde{C}_{\mathbf{x}} < 2^{j-1}$*. If*

$$m \geq E C_1^{-6}\varepsilon^{-6}\max\left\{w(\mathcal{M}), \sqrt{d\log(e/\varepsilon)}\right\}^2, \tag{5.10}$$

then with probability at least $1 - 12\exp(-cC_1^2\varepsilon^2 m)$ for all $\mathbf{x}\in\mathcal{M}\subset\mathbb{S}^{N-1}$ the approximations $\mathbf{x}^*$ obtained by OMS satisfy

$$\|\mathbf{x}-\mathbf{x}^*\|_2^2 \leq E'\left(1+\tilde{C}_{\mathbf{x}}+C_1\max\left\{1,\log(C_1^{-1})\right\}\right)^2\varepsilon\log\left(\frac{e}{2\varepsilon}\right). \tag{5.11}$$

Remark 5.3.2. *The restrictions on C_1 and $\tilde{C}_{\mathbf{x}}$ are easily satisfied, e.g., if the centers form a maximal 2^{-j} packing of $\mathcal{M}$ at each scale j or if the GMRA is constructed from manifold samples as discussed in Section 5.1. In both these cases C_1 and $\tilde{C}_{\mathbf{x}}$ are in fact bounded by absolute constants. We discuss Theorem 5.3.1 for empirical GMRAs in Section 5.4.*
Note that Theorem 5.3.1 is up to the additional ε-log-factors in (5.10) *and* (5.11) *similar to Theorem 3.3.9. We thus extended Theorem 3.3.9 by a tractable approach for non-convex signal sets under mildly worse conditions.*
Numerical simulations (see Section 5.5) suggest that a slightly modified version of OMS performs better in some scenarios even though we cannot provide a rigorous theoretical justification for the modification's improved performance at present.

For proving Theorem 5.3.1 we need two technical lemmas. The first controls the Gaussian width of the union of manifold and its scale-j GMRA approximation $\mathcal{M} \cup \mathcal{M}_j$.

Lemma 5.3.3. *For $\mathcal{M}_j$, the subspace approximation in the GMRA of level $j > j_0$ for $\mathcal{M}$ of dimension $d \geq 1$, the Gaussian width of $\mathcal{M} \cup \mathcal{M}_j$ can be bounded from above and below by*

$$\max\{w(\mathcal{M}), w(\mathcal{M}_j)\} \leq w(\mathcal{M} \cup \mathcal{M}_j) \leq 2w(\mathcal{M}) + 2w(\mathcal{M}_j) + 3 \leq 2w(\mathcal{M}) + C\sqrt{dj}.$$

Remark 5.3.4. *Note that the first inequality holds for general sets, not only $\mathcal{M}$ and $\mathcal{M}_j$. Moreover, one only uses $\mathcal{M}_j \subset \mathcal{B}_2(\mathbf{0}, 2)$ to prove the second inequality. It thus holds for $\mathcal{M}_j$ replaced with arbitrary subsets of $\mathcal{B}_2(\mathbf{0}, 2)$. We might use both variations referring to Lemma 5.3.3.*

Proof: The first inequality follows by noting that

$$\begin{aligned}\max\{w(\mathcal{M}), w(\mathcal{M}_j)\} &= \max\left\{ \mathsf{E}\left[\sup_{\mathbf{z}\in\mathcal{M}} \langle \mathbf{z}, \mathbf{g}\rangle\right], \mathsf{E}\left[\sup_{\mathbf{z}\in\mathcal{M}_j} \langle \mathbf{z}, \mathbf{g}\rangle\right]\right\} \\ &\leq \mathsf{E}\left[\sup_{\mathbf{z}\in\mathcal{M}\cup\mathcal{M}_j} \langle \mathbf{z}, \mathbf{g}\rangle\right] = w(\mathcal{M} \cup \mathcal{M}_j).\end{aligned}$$

To obtain the second inequality observe that

$$\begin{aligned}w(\mathcal{M} \cup \mathcal{M}_j) \leq \gamma(\mathcal{M} \cup \mathcal{M}_j) &\leq \mathsf{E}\left[\sup_{\mathbf{z}\in\mathcal{M}} |\langle \mathbf{z}, \mathbf{g}\rangle| + \sup_{\mathbf{z}\in\mathcal{M}_j} |\langle \mathbf{z}, \mathbf{g}\rangle|\right] = \gamma(\mathcal{M}) + \gamma(\mathcal{M}_j) \\ &\leq 2w(\mathcal{M}) + 2w(\mathcal{M}_j) + \sqrt{\frac{2}{\pi}}\mathrm{dist}(\mathbf{0}, \mathcal{M}) + \sqrt{\frac{2}{\pi}}\mathrm{dist}(\mathbf{0}, \mathcal{M}_j) \\ &\leq 2\left(w(\mathcal{M}) + w(\mathcal{M}_j) + 1.5\sqrt{\frac{2}{\pi}}\right),\end{aligned} \tag{5.12}$$

where we used (3.6), the fact that $\mathcal{M} \subset \mathbb{S}^{N-1}$, and that $\mathcal{M}_j \subset \mathcal{B}_2(\mathbf{0}, 2)$.

For the last inequality we bound $w(\mathcal{M}_j)$. First, note that

$$\begin{aligned} w(\mathcal{M}_j) = \mathsf{E}\left[\sup_{\mathbf{z}\in\mathcal{M}_j} \langle \mathbf{z}, \mathbf{g}\rangle\right] &= \mathsf{E}\left[\sup_{\mathbf{z}\in\{\mathbb{P}_{j,k_j(\mathbf{z}')}(\mathbf{z}')\colon \mathbf{z}'\in\mathcal{B}_2(\mathbf{0},2)\}\cap\mathcal{B}_2(\mathbf{0},2)} \langle \mathbf{z}, \mathbf{g}\rangle\right] \\ &\leq \mathsf{E}\left[\sup_{\mathbf{z}'\in\bigcup_{k\in[K_j]} P_{j,k}\cap\mathcal{B}_2(\mathbf{0},2)} \langle \mathbf{z}', \mathbf{g}\rangle\right]. \end{aligned}$$

For all $k \in [K_j]$ there exist d-dimensional Euclidean balls $L_{j,k} \subset P_{j,k}$ of radius 2 such that $P_{j,k} \cap \mathcal{B}_2(\mathbf{0}, 2) \subset L_{j,k}$. Hence, $\bigcup_{k\in[K_j]}(P_{j,k} \cap \mathcal{B}_2(\mathbf{0}, 2)) \subset L_j := \bigcup_{k\in[K_j]} L_{j,k}$. By definition the ε-covering number of L_j (a union of K_j d-dimensional balls) can be bounded by $N(L_j, \varepsilon) \leq K_j(6/\varepsilon)^d$ which implies $\log N(L_j, \varepsilon) \leq dj\log(12C_{\mathcal{C}}/\varepsilon)$ by GMRA property (2). By Dudley's inequality in Theorem 3.2.8 and Jensen's inequality we conclude that

$$\begin{aligned} w(\mathcal{M}_j) \leq w(L_j) &\leq C_{\text{Dudley}} \int_0^2 \sqrt{\log N(L_j, \varepsilon)}\, \mathrm{d}\varepsilon \\ &\leq C_{\text{Dudley}}\sqrt{dj} \int_0^2 \sqrt{\log(12C_{\mathcal{C}}) - \log(\varepsilon)}\, \mathrm{d}\varepsilon \\ &\leq C_{\text{Dudley}}\sqrt{dj}\sqrt{2\log(12C_{\mathcal{C}}) - \int_0^2 \log(\varepsilon)\, \mathrm{d}\varepsilon} \\ &\leq C'\sqrt{dj}, \end{aligned}$$

where C' is a constant depending on C_{Dudley} and $C_{\mathcal{C}}$. Choosing $C = 2C' + 3$ yields the claim as $3\sqrt{2/\pi} \leq 3\sqrt{dj}$. ■

The second lemma quantifies the equivalence of ℓ_2-norm and normalized geodesic distance on the unit sphere.

Lemma 5.3.5. *For $\mathbf{z}, \mathbf{z}' \in \mathbb{S}^{N-1}$ one has*

$$d_G(\mathbf{z}, \mathbf{z}') \leq \|\mathbf{z} - \mathbf{z}'\|_2 \leq \pi d_G(\mathbf{z}, \mathbf{z}').$$

Proof: First observe that $\langle \mathbf{z}, \mathbf{z}'\rangle = \cos\measuredangle(\mathbf{z}, \mathbf{z}') = \cos(\pi d_G(\mathbf{z}, \mathbf{z}'))$. This yields

$$\|\mathbf{z} - \mathbf{z}'\|_2 - d_G(\mathbf{z}, \mathbf{z}') = \sqrt{2 - 2\cos(\pi d_G(\mathbf{z}, \mathbf{z}'))} - d_G(\mathbf{z}, \mathbf{z}') \geq 0$$

as the function $f(x) = \sqrt{2 - 2\cos(\pi x)} - x$ is non-negative on $[0, 1]$.
For the upper bound note the relation between the geodesic distance $\tilde{d}_G$ and the normalized geodesic distance d_G

$$\tilde{d}_G(\mathbf{z}, \mathbf{z}') = \pi d_G(\mathbf{z}, \mathbf{z}')$$

which yields

$$\|\mathbf{z} - \mathbf{z}'\|_2 \leq \tilde{d}_G(\mathbf{z}, \mathbf{z}') = \pi d_G(\mathbf{z}, \mathbf{z}').$$

■

We can now turn to the proof of Theorem 5.3.1. We show the following more detailed result which directly implies Theorem 5.3.1.

Theorem 5.3.6 (Uniform Recovery - Axiomatic Case)**.** *Let $\mathcal{M} \subset \mathbb{S}^{N-1}$ be given by its GMRA for some levels $j_0 < j \leq J$, such that $C_1 < 2^{j_0+1}$ where C_1 is the constant from GMRA properties* (2b) *and* (3a)*. Fix j and assume that* $\operatorname{dist}(\mathbf{0}, \mathcal{M}_j) \geq 1/2$*. Further, let $d \geq 1$ and*

$$m \geq 16 \max\{C', \bar{C}\} C_1^{-6} 2^{6(j+1)} (w(\mathcal{M}) + C\sqrt{dj})^2, \tag{5.13}$$

where C' is the constant from Theorem 3.3.9, $\bar{C}$ from Theorem 3.3.5, and $C > 3$ from Lemma 5.3.3. Then, with probability at least $1 - 12\exp(-c(C_1 2^{-j-1})^2 m)$ the following holds for all $\mathbf{x} \in \mathcal{M}$ with one-bit measurements $\mathbf{y} = \operatorname{sign}(A\mathbf{x})$ and GMRA constants $\tilde{C}_\mathbf{x}$ from property (3b) *satisfying $\tilde{C}_\mathbf{x} < 2^{j-1}$: The approximations $\mathbf{x}^*$ obtained by OMS fulfill*

$$\|\mathbf{x} - \mathbf{x}^*\|_2 \leq$$
$$\left(2\tilde{C}_\mathbf{x} 2^{-\frac{j}{2}} + \sqrt{\frac{C_1}{2}} \sqrt[4]{\log\left(\frac{4e}{\min\{C_1, 1\}}\right)} + \sqrt{11 C'_\mathbf{x}} \sqrt[4]{\log\left(\frac{2e}{\min\{C'_\mathbf{x}, 1\}}\right)}\right) \sqrt[4]{j} 2^{-\frac{j}{2}}.$$

Here $C'_\mathbf{x} := 2\tilde{C}_\mathbf{x} + C_1$.

Proof of Theorem 5.3.1: As $j = \lceil \log(1/\varepsilon) \rceil$, we know that $2^{-j} \leq \varepsilon \leq 2^{-j+1}$. This implies

$$\begin{aligned} m &\geq E C_1^{-6} \varepsilon^{-6} \max\left\{ w(\mathcal{M}), \sqrt{d\left(\log\left(\frac{1}{\varepsilon}\right) + 1\right)} \right\}^2 \\ &\geq 16 \max\{C', \bar{C}\} C_1^{-6} 2^{6(j+1)} (w(\mathcal{M}) + C\sqrt{dj})^2, \end{aligned}$$

for suitably chosen $E > 0$. The result follows by applying Theorem 5.3.6. ■

To prove Theorem 5.3.6 we show that the center $\mathbf{c}_{j,k'}$ identified in step I. of OMS satisfies $\|\mathbf{x} - \mathbf{c}_{j,k'}\|_2 \leq 16 \max\{\|\mathbf{x} - \mathbf{c}_{j,k_j(\mathbf{x})}\|_2, C_1 2^{-j-1}\}$. Therefore, the GMRA property (3b) provides an upper bound on $\|\mathbf{x} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2$. What remains is to then bound the gap between $\mathbb{P}_{j,k'}(\mathbf{x})$ and the approximation $\mathbf{x}^*$. This happens in two steps. First, Theorem 3.3.9 is applied to (5.9) bounding the distance between $\mathbb{P}_{j,k'}(\mathbf{x}) / \left\|\mathbb{P}_{j,k'}(\mathbf{x})\right\|_2$ and $\mathbf{x}^*$ (the true measurements $\mathbf{y}$ are interpreted as noisy version of the non-accessible one-bit measurements of $\mathbb{P}_{j,k'}(\mathbf{x}) / \left\|\mathbb{P}_{j,k'}(\mathbf{x})\right\|_2$). Second, the distance between $\mathbb{P}_{j,k'}(\mathbf{x})$ and $\mathbb{P}_{j,k'}(\mathbf{x}) / \left\|\mathbb{P}_{j,k'}(\mathbf{x})\right\|_2$ is bounded and a union bound concludes the proof.

Lemma 5.3.7. *If $m \geq \bar{C} C_1^{-6} 2^{6(j+1)} \max\{w(\mathcal{M} \cup \mathbb{P}_\mathbb{S}(\mathcal{C}_j))^2, 2/\pi\}$ the center $\mathbf{c}_{j,k'}$ chosen in step I. of OMS fulfills*

$$\|\mathbf{x} - \mathbf{c}_{j,k'}\|_2 \leq 16 \max\{\|\mathbf{x} - \mathbf{c}_{j,k_j(\mathbf{x})}\|_2, C_1 2^{-j-1}\}.$$

for all $\mathbf{x} \in \mathcal{M} \subset \mathbb{S}^{N-1}$ with probability at least $1 - 2\exp(-c(C_1 2^{-j-1})^2 m)$.

Proof: Recall that $d_A(\mathbf{z},\mathbf{z}') = \frac{1}{m} d_H(\mathrm{sign}(\mathbf{A}\mathbf{z}), \mathrm{sign}(\mathbf{A}\mathbf{z}'))$. By definition of $\mathbf{c}_{j,k'}$ in (5.8) we have that

$$d_H(\mathrm{sign}(\mathbf{A}\mathbf{c}_{j,k'}), \mathbf{y}) \le d_H(\mathrm{sign}(\mathbf{A}\mathbf{c}_{j,k_j(\mathbf{x})}), \mathbf{y}).$$

As, for all $\mathbf{z},\mathbf{z}' \in \mathbb{R}^N$, $d_H(\mathrm{sign}(\mathbf{A}\mathbf{z}), \mathrm{sign}(\mathbf{A}\mathbf{z}')) = m \cdot d_A(\mathbf{z},\mathbf{z}') = m \cdot d_A(\mathbb{P}_{\mathbb{S}}(\mathbf{z}), \mathbb{P}_{\mathbb{S}}(\mathbf{z}'))$, this is equivalent to

$$d_A(\mathbb{P}_{\mathbb{S}}(\mathbf{c}_{j,k'}), \mathbf{x}) \le d_A(\mathbb{P}_{\mathbb{S}}(\mathbf{c}_{j,k_j(\mathbf{x})}), \mathbf{x}).$$

Noting that Gaussian random vectors and Haar random vectors yield identically distributed hyperplanes, Theorem 3.3.5 transfers this bound to the normalized geodesic distance, namely

$$d_G(\mathbb{P}_{\mathbb{S}}(\mathbf{c}_{j,k'}), \mathbf{x}) \le d_G(\mathbb{P}_{\mathbb{S}}(\mathbf{c}_{j,k_j(\mathbf{x})}), \mathbf{x}) + 2\delta$$

with probability at least $1 - 2\exp(-c\delta^2 m)$ where $\delta = C_1 2^{-j-1}$. Recall that, by Lemma 5.3.5, $d_G(\mathbf{z},\mathbf{z}') \le \|\mathbf{z}-\mathbf{z}'\|_2 \le \pi d_G(\mathbf{z},\mathbf{z}')$, for all $\mathbf{z},\mathbf{z}' \in \mathbb{S}^{N-1}$, which leads to

$$\begin{aligned}\|\mathbb{P}_{\mathbb{S}}(\mathbf{c}_{j,k'}) - \mathbf{x}\|_2 &\le \pi d_G(\mathbb{P}_{\mathbb{S}}(\mathbf{c}_{j,k_j(\mathbf{x})}), \mathbf{x}) + 2\pi\delta \\ &\le \pi \|\mathbb{P}_{\mathbb{S}}(\mathbf{c}_{j,k_j(\mathbf{x})}) - \mathbf{x}\|_2 + 2\pi\delta.\end{aligned}$$

As by property (3a) the centers are close to the manifold, they are also close to the sphere and we have $\|\mathbb{P}_{\mathbb{S}}(\mathbf{c}_{j,k}) - \mathbf{c}_{j,k}\|_2 < C_1 2^{-j-2}$, for all $\mathbf{c}_{j,k} \in \mathcal{C}_j$. Hence, we conclude

$$\begin{aligned}\|\mathbf{c}_{j,k'} - \mathbf{x}\|_2 &\le \|\mathbf{c}_{j,k'} - \mathbb{P}_{\mathbb{S}}(\mathbf{c}_{j,k'})\|_2 + \|\mathbb{P}_{\mathbb{S}}(\mathbf{c}_{j,k'}) - x\|_2 \\ &\le C_1 2^{-j-2} + \pi(\|\mathbf{c}_{j,k_j(\mathbf{x})} - \mathbf{x}\|_2 + C_1 2^{-j-2}) + 2\pi\delta \\ &\le \left(\pi + \frac{\pi}{2} + 2\pi + \frac{1}{2}\right) \max\{\|\mathbf{c}_{j,k_j(\mathbf{x})} - \mathbf{x}\|_2, C_1 2^{-j-1}\} \\ &\le 16 \max\{\|\mathbf{c}_{j,k_j(\mathbf{x})} - \mathbf{x}\|_2, C_1 2^{-j-1}\}.\end{aligned}$$

■

Proof of Theorem 5.3.6: Recall that k' is the index chosen by OMS in (5.8). The proof consists of three steps. First, we apply Lemma 5.3.7 in **(I)**. By the GMRA axioms this supplies an estimate for $\|\mathbf{x} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2$ with high probability. In **(II)** we use Theorem 3.3.9 to bound the distance between $\mathbb{P}_{j,k'}(\mathbf{x})/\|\mathbb{P}_{j,k'}(\mathbf{x})\|_2$ and the minimizer $\mathbf{x}^*$ given by

$$\mathbf{x}^* = \arg\min_{\mathbf{z}} \sum_{l=1}^{m} (-y_l)\langle \mathbf{a}_l, \mathbf{z}\rangle, \quad \text{subject to } \mathbf{z} \in K := \mathrm{conv}(\mathbb{P}_{\mathbb{S}}(P_{j,k'} \cap \mathcal{B}_2(\mathbf{0},2))), \tag{5.14}$$

with high probability. By a union bound over all events, part **(III)** concludes with an estimate of the distance $\|\mathbf{x} - \mathbf{x}^*\|_2$ combining **(I)** and **(II)**.

(I) Set $\delta := C_1 2^{-j-1}$. Observing that $C_1 2^{-j-2} < 1/2$ by assumption, GMRA property (3a) yields that all centers in $\mathcal{C}_j$ are closer to $\mathbb{S}^{N-1}$ than 1/2, i.e., $1/2 \leq \|\mathbf{c}_{j,k}\|_2 \leq 3/2$. Hence, by (3.6)

$$\begin{aligned} 0 &\leq w(\mathbb{P}_{\mathbb{S}}(\mathcal{C}_j)) \leq \gamma(\mathbb{P}_{\mathbb{S}}(\mathcal{C}_j)) \leq 2\gamma(\mathcal{C}_j) \\ &\leq 4w(\mathcal{C}_j) + 2\sqrt{\frac{2}{\pi}}\operatorname{dist}(\mathbf{0}, \mathcal{C}_j) \leq 4w(\mathcal{C}_j) + 4. \end{aligned} \tag{5.15}$$

As $\mathcal{C}_j \subset \mathcal{M}_j$ we know by Lemma 5.3.3, (5.15), and Remark 5.3.4 that

$$\begin{aligned} m &\geq 4\bar{C}\delta^{-6}(2w(\mathcal{M}) + 2C\sqrt{dj})^2 \geq 4\bar{C}\delta^{-6}(2w(\mathcal{M}) + 4w(\mathcal{C}_j) + 6)^2 \\ &\geq 4\bar{C}\delta^{-6}(2w(\mathcal{M}) + w(\mathbb{P}_{\mathbb{S}}(\mathcal{C}_j)) + 2)^2 = \bar{C}\delta^{-6}(4w(\mathcal{M}) + 2w(\mathbb{P}_{\mathbb{S}}(\mathcal{C}_j)) + 4)^2 \\ &\geq \bar{C}\delta^{-6}(w(\mathcal{M} \cup \mathbb{P}_{\mathbb{S}}(\mathcal{C}_j)) + 1)^2 \geq \bar{C}\delta^{-6}\max\{w(\mathcal{M} \cup \mathbb{P}_{\mathbb{S}}(\mathcal{C}_j))^2, 2/\pi\}. \end{aligned} \tag{5.16}$$

Hence, Lemma 5.3.7 implies that

$$\|\mathbf{x} - \mathbf{c}_{j,k'}\|_2 \leq 16\max\{\|\mathbf{x} - \mathbf{c}_{j,k_j(\mathbf{x})}\|_2, C_1 2^{-j-1}\}.$$

with probability at least $1 - 2\exp(-c\delta^2 m)$. By GMRA property (3b) we now get that

$$\|\mathbf{x} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2 \leq \tilde{C}_{\mathbf{x}} 2^{-j}, \tag{5.17}$$

for some constant $\tilde{C}_{\mathbf{x}}$.

(II) Define $\alpha := \|\mathbb{P}_{j,k'}(\mathbf{x})\|_2$ and note that one has $1/2 \leq \alpha \leq 3/2$ as $\mathbf{x} \in \mathbb{S}^{N-1}$ and $\|\mathbf{x} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2 \leq \tilde{C}_{\mathbf{x}} 2^{-j} \leq 1/2$ by (5.17) and assumption. We now create the setting of Theorem 3.3.9. Define $\tilde{\mathbf{x}} := \mathbb{P}_{j,k'}(\mathbf{x})/\alpha \in \mathbb{S}^{N-1}$, $\tilde{\mathbf{y}} := \operatorname{sign}(\mathbf{A}\tilde{\mathbf{x}}) = \operatorname{sign}(\mathbf{A}\mathbb{P}_{j,k'}(\mathbf{x}))$, $K = \operatorname{conv}(\mathbb{P}_{\mathbb{S}}(P_{j,k'} \cap \mathcal{B}_2(\mathbf{0}, 2)))$, and $\tau := (2\tilde{C}_{\mathbf{x}} + C_1)2^{-j}$. If successfully applied with these quantities Theorem 3.3.9 will bound $\|\tilde{\mathbf{x}} - \mathbf{x}^*\|_2$ by

$$\begin{aligned} \|\tilde{\mathbf{x}} - \mathbf{x}^*\|_2 &\leq \sqrt{\delta\sqrt{\log\left(\frac{e}{\delta}\right)} + 11\tau\sqrt{\log\left(\frac{e}{\tau}\right)}} \\ &\leq \left(\sqrt{\frac{C_1}{2}}\sqrt[4]{\log\left(\frac{4e}{\min\{C_1, 1\}}\right)}\right. \\ &\quad \left. + \sqrt{11(2\tilde{C}_x + C_1)}\sqrt[4]{\log\left(\frac{2e}{\min\{(2\tilde{C}_x + C_1), 1\}}\right)}\right)\sqrt[4]{j}2^{-\frac{j}{2}}. \end{aligned} \tag{5.18}$$

All that remains is to verify that the conditions of Theorem 3.3.9 are met so that (5.18) is guaranteed with high probability. (Note that in our case the true measurements $\mathbf{y}$ are interpreted as corrupted version of $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}$ is interpreted as the signal which shall be approximated by (5.14).)

We first have to check $d_H(\tilde{\mathbf{y}},\mathbf{y}) \leq \tau m$. Recall that $\frac{1}{\alpha} \leq 2$ and for $\alpha > 0$ one has $\alpha w(K) = w(\alpha K)$. Applying Lemma 5.3.3 and (3.6) we have, in analogy to (5.16), that

$$\begin{aligned} m &\geq \bar{C}\delta^{-6}(4w(\mathcal{M}) + 4w(\mathcal{M}_j) + 12)^2 \geq \bar{C}\delta^{-6}\left(2w(\mathcal{M}) + 2w\left(\frac{\mathcal{M}_j}{\alpha}\right) + 12\right)^2 \\ &\geq \bar{C}\delta^{-6}\left(w\left(\mathcal{M} \cup \frac{\mathcal{M}_j}{\alpha}\right) + 7\right)^2 \geq \bar{C}\delta^{-6}\left(w\left(\left(\mathcal{M} \cup \frac{\mathcal{M}_j}{\alpha}\right) \cap \mathcal{B}_2(\mathbf{0},1)\right) + 7\right)^2. \end{aligned}$$

Note that in the third inequality a slight modification of the second inequality in Lemma 5.3.3 is used. As $\mathcal{M}_j/\alpha \subset \mathcal{B}_2(0,4)$ one has $w(\mathcal{M} \cup \mathcal{M}_j/\alpha) \leq 2w(\mathcal{M}) + 2w(\mathcal{M}_j/\alpha) + 5$ by adapting (5.12). We can now use Theorem 3.3.5, Lemma 5.3.5, and the fact that $|1-\alpha| = |\|\mathbf{x}\|_2 - \|\mathbb{P}_{j,k'}(\mathbf{x})\|_2| \leq \|\mathbf{x} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2$ to obtain

$$\begin{aligned} \frac{d_H(\tilde{\mathbf{y}},\mathbf{y})}{m} &= d_A(\tilde{\mathbf{x}},\mathbf{x}) \leq d_G(\tilde{\mathbf{x}},\mathbf{x}) + \delta \leq \|\tilde{\mathbf{x}} - \mathbf{x}\|_2 + \delta \\ &\leq \|\tilde{\mathbf{x}} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2 + \|\mathbb{P}_{j,k'}(\mathbf{x}) - \mathbf{x}\|_2 + \delta \\ &= |1-\alpha| + \|\mathbb{P}_{j,k'}(\mathbf{x}) - \mathbf{x}\|_2 + \delta \leq 2\|\mathbb{P}_{j,k'}(\mathbf{x}) - \mathbf{x}\|_2 + \delta \\ &\leq (2\tilde{C}_{\mathbf{x}} + C_1)2^{-j} = \tau \end{aligned}$$

with probability at least $1 - 2\exp(-c\delta^2 m)$. Furthermore, by a similar argumentation as in (5.15) one gets

$$w(K) = w(\mathbb{P}_{\mathbb{S}}(P_{j,k'} \cap \mathcal{B}_2(\mathbf{0},2))) \leq 4w(\mathcal{M}_j) + 4, \tag{5.19}$$

where one uses invariance of the Gaussian width under taking the convex hull (see Theorem 3.2.7), the fact that $P_{j,k'} \cap \mathcal{B}_2(\mathbf{0},2) \subset \mathcal{M}_j$, and the assumption that $1/2 \leq \operatorname{dist}(\mathcal{M}_j,\mathbf{0}) \leq 2$. In combination with Lemma 5.3.3 we have, in analogy to (5.16), that

$$m \geq 4C'\delta^{-6}(2w(\mathcal{M}) + 4w(\mathcal{M}_j) + 6)^2 \geq 4C'\delta^{-6}(w(K)+2)^2 \geq C'\delta^{-6}w(K)^2.$$

Hence, we can apply Theorem 3.3.9 to obtain with probability at least $1 - 8\exp(-c\delta^2 m)$ that

$$\|\bar{\mathbf{x}} - \tilde{\mathbf{x}}\|_2^2 \leq \delta\sqrt{\log\left(\frac{e}{\delta}\right)} + 11\tau\sqrt{\log\left(\frac{e}{\tau}\right)}.$$

The estimate (5.18) now follows.

(III) To conclude the proof we apply a union bound and obtain with probability at least $1 - 12\exp(-c\delta^2 m)$ that

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}^*\|_2 &\leq \|\mathbf{x} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2 + \|\mathbb{P}_{j,k'}(\mathbf{x}) - \tilde{\mathbf{x}}\|_2 + \|\tilde{\mathbf{x}} - \mathbf{x}^*\|_2 \\ &= \|\mathbf{x} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2 + |1-\alpha| + \|\tilde{\mathbf{x}} - \mathbf{x}^*\|_2 \\ &\leq 2\|\mathbf{x} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2 + \|\tilde{\mathbf{x}} - \mathbf{x}^*\|_2. \end{aligned}$$

GMRA property (3b) combined with (5.18) yields the final desired error bound. ■

Theorem 5.3.1 depends on the Gaussian width of $\mathcal{M}$. For general sets this quantity provides a useful measure of the set's complexity. If $\mathcal{M}$ is a compact Riemannian submanifold of $\mathbb{R}^N$ it might be more convenient to have a dependence on the geometric properties of $\mathcal{M}$ instead (e.g., its volume and reach). It is straight-forward then to deduce from [55] that $w(\mathcal{M})$ can be upper bounded in terms of the manifold's intrinsic dimension d, its d-dimensional volume $\mathrm{Vol}_d(\mathcal{M})$, and the inverse of its reach. These dependencies are intuitively to be expected as a manifold with fixed intrinsic dimension d can become more complex as either its volume or curvature (which can be bounded by the inverse of its reach) grows. The following theorem is a combination of different results in [55] and formalizes the intuition by bounding the Gaussian width of a manifold in terms of its geometric properties.

Theorem 5.3.8. *Assume $\mathcal{M} \subset \mathbb{R}^N$ is a compact d-dimensional Riemannian manifold with d-dimensional volume $\mathrm{Vol}_d(\mathcal{M})$ where $d \geq 1$. Then one can replace $w(\mathcal{M})$ in Theorem 5.3.1 by*

$$w(\mathcal{M}) \leq C{\cdot}\mathrm{diam}(\mathcal{M}){\cdot}\sqrt{d \cdot \max\left\{\log\left(c\frac{\sqrt{d}}{\min\{1, \mathrm{reach}(\mathcal{M})\}}\right), 1\right\} + \log(\max\{1, \mathrm{Vol}_d(\mathcal{M})\})}.$$

where $C, c > 0$ are absolute constants.

Proof: Denote by τ the reach and by ρ the diameter of $\mathcal{M}$. From [55, Lemma 14] we know that the covering number $N(\mathcal{M}, \varepsilon)$ of a d-dimensional Riemannian manifold $\mathcal{M}$ can be bounded by

$$N(\mathcal{M}, \varepsilon) \leq \left(\frac{2}{\varepsilon\sqrt{1 - \left(\frac{\varepsilon}{4\tau}\right)^2}}\right)^d \frac{\mathrm{Vol}_d(\mathcal{M})}{\mathrm{Vol}_d(\mathcal{B}_d)}.$$

for $\varepsilon \leq \frac{\tau}{2}$. After noting that $\mathrm{Vol}_d(\mathcal{B}_d) \geq \beta^{-1}\left(\frac{2\pi}{d}\right)^{\frac{d}{2}}$ for all $d \geq 1$ for an absolute constant $\beta > 1$, this expression may be simplified to

$$N(\mathcal{M}, \varepsilon) \leq \beta\left(\frac{\sqrt{2d}}{\sqrt{\pi}\varepsilon\sqrt{1 - \left(\frac{\varepsilon}{4\tau}\right)^2}}\right)^d \mathrm{Vol}_d(\mathcal{M}) \leq \beta\left(\frac{\sqrt{d}}{\varepsilon\sqrt{1 - \left(\frac{\varepsilon}{4\tau}\right)^2}}\right)^d \mathrm{Vol}_d(\mathcal{M}).$$

We can combine these facts with Dudley's inequality in Theorem 3.2.8 to obtain

$$\begin{aligned} w(\mathcal{M}) &\leq C' \int_0^\rho \sqrt{\log(N(\mathcal{M}, \varepsilon))}\ \mathrm{d}\varepsilon \\ &\leq C' \left(\rho \int_0^\rho \log(N(\mathcal{M}, \varepsilon))\ \mathrm{d}\varepsilon\right)^{\frac{1}{2}} \\ &= C'\sqrt{\rho}\left(\int_0^{\frac{\tau}{2}} \log(N(\mathcal{M}, \varepsilon))\ \mathrm{d}\varepsilon + \int_{\frac{\tau}{2}}^\rho \log(N(\mathcal{M}, \varepsilon))\ \mathrm{d}\varepsilon\right)^{\frac{1}{2}}, \end{aligned}$$

using Cauchy-Schwarz for the second inequality. We now bound the first integral by

$$\begin{aligned}
\int_0^{\frac{\tau}{2}} \log(N(\mathcal{M},\varepsilon))\ \mathrm{d}\varepsilon &\leq \int_0^{\frac{\tau}{2}} -d\log\left(\frac{\varepsilon}{\beta\sqrt{d}}\underbrace{\sqrt{1-\left(\frac{\varepsilon}{4\tau}\right)^2}}_{\geq\frac{1}{2},\text{ as }\varepsilon\leq\frac{\tau}{2}}\right) + \log(\mathrm{Vol}_d(\mathcal{M}))\ \mathrm{d}\varepsilon \\
&\leq \int_0^{\frac{\tau}{2}} -d\log\left(\frac{\varepsilon}{2\beta\sqrt{d}}\right)\ \mathrm{d}\varepsilon + \frac{\tau}{2}\log(\mathrm{Vol}_d(\mathcal{M})) \\
&= -d\left[\varepsilon\log\left(\frac{\varepsilon}{2\beta\sqrt{d}}\right) - \varepsilon\right]_0^{\frac{\tau}{2}} + \frac{\tau}{2}\log(\mathrm{Vol}_d(\mathcal{M})) \\
&= \frac{d\tau}{2}\left(\log\left(4\frac{\beta\sqrt{d}}{\tau}\right)+1\right) + \frac{\tau}{2}\log(\mathrm{Vol}_d(\mathcal{M})).
\end{aligned}$$

As the covering number is decreasing with increasing ε, the second integral can be bounded as follows.

$$\begin{aligned}
\int_{\frac{\tau}{2}}^{\rho} \log(N(\mathcal{M},\varepsilon))\,\mathrm{d}\varepsilon &\leq \int_{\frac{\tau}{2}}^{\rho} \log\left(N\left(\mathcal{M},\frac{\tau}{2}\right)\right)\ \mathrm{d}\varepsilon \\
&= \left(\rho-\frac{\tau}{2}\right)\left[-d\log\left(\frac{\tau}{4\beta\sqrt{d}}\right) + \log(\mathrm{Vol}_d(\mathcal{M}))\right] \\
&= d\left(\rho-\frac{\tau}{2}\right)\log\left(4\frac{\beta\sqrt{d}}{\tau}\right) + \left(\rho-\frac{\tau}{2}\right)\log(\mathrm{Vol}_d(\mathcal{M})).
\end{aligned}$$

Both together yield

$$\begin{aligned}
w(\mathcal{M}) &\leq C\sqrt{\rho}\left(\frac{d\tau}{2}\left(\log\left(4\beta\frac{\sqrt{d}}{\tau}\right)+1\right) + d\left(\rho-\frac{\tau}{2}\right)\log\left(4\beta\frac{\sqrt{d}}{\tau}\right) + \rho\log(\mathrm{Vol}_d(\mathcal{M}))\right)^{\frac{1}{2}} \\
&\leq C\sqrt{\rho}\left(d\tau\cdot\max\left\{\log\left(4\beta\frac{\sqrt{d}}{\tau}\right),1\right\}\right. \\
&\left. +d\left(2\rho-\tau\right)\cdot\max\left\{\log\left(4\beta\frac{\sqrt{d}}{\tau}\right),1\right\} + 2\rho\log(\mathrm{Vol}_d(\mathcal{M}))\right)^{\frac{1}{2}} \\
&= C\sqrt{\rho}\left(2d\rho\cdot\max\left\{\log\left(4\beta\frac{\sqrt{d}}{\tau}\right),1\right\} + 2\rho\log(\mathrm{Vol}_d(\mathcal{M}))\right)^{\frac{1}{2}} \\
&\leq C\rho\sqrt{2d\cdot\max\left\{\log\left(4\beta\frac{\sqrt{d}}{\tau}\right),1\right\} + 2\log(\mathrm{Vol}_d(\mathcal{M}))}.
\end{aligned}$$

■

5.4 OMS and the Empirical GMRA

We have noted in the beginning of this chapter that GMRAs are normally computed from samples of $\mathcal{M}$ and do not necessarily fulfill all axioms of Definition 5.1.1 perfectly, an assumption we make in Theorem 5.3.1. In this section we provide an alternative version of Theorem 5.3.6 which holds for empirical GMRAs and extends Theorem 5.3.1 as long as the sampling distribution on $\mathcal{M}$ behaves well and a sufficient number of samples is taken.

Recall from Section 5.1.2 and Section 5.1.3 the assumptions of the empirical setting. There is a probability distribution Π supported on $\mathcal{M}$ which behaves well in the sense that the assumptions **(A1)-(A4)** are fulfilled. Moreover, we are given a set of n points $\mathcal{X} = \{X_1, ..., X_n\}$ which are sampled iid from Π. The difference between probabilistic GMRA fulfilling the axioms in Definition 5.1.1 and empirical GMRA can thus be controlled by Theorem 5.1.3. OMS has to be slightly modified to conform with the empirical GMRA notation of Section 5.1.3. To this end, replace (5.8) and (5.9) by

$$\widehat{\mathbf{c}}_{j,k'} \in \operatorname*{arg\,min}_{\widehat{\mathbf{c}}_{j,k} \in \widehat{\mathcal{C}}_j} \; d_H(\operatorname{sign}(\mathbf{A}\widehat{\mathbf{c}}_{j,k}), \mathbf{y}). \tag{5.20}$$

$$\begin{cases} \mathbf{x}^* = \arg\min_{\mathbf{z} \in \mathbb{R}^N} \sum_{l=1}^{m} (-y_l) \langle \mathbf{a}_l, \mathbf{z} \rangle, \\ \text{subject to } \mathbf{z} \in \operatorname{conv}\left(\mathbb{P}_{\mathbb{S}}(\widehat{P}_{j,k'} \cap \mathcal{B}_2(\mathbf{0}, 2)) \right). \end{cases} \tag{5.21}$$

To stay consistent with the axiomatic notation we denote the sets containing the centers $\mathbf{c}'_{j,k}$ and $\widehat{\mathbf{c}}_{j,k}$ by $\mathcal{C}'_j$ and $\widehat{\mathcal{C}}_j$ respectively. We denote the empirical GMRA approximation at level j, i.e., the set $\widehat{\mathbb{P}}_j$ projects onto, by

$$\widehat{\mathcal{M}}_j = \{\widehat{\mathbb{P}}_j(\mathbf{z}) \colon \mathbf{z} \in \mathcal{B}_2(\mathbf{0}, 2)\} \cap \mathcal{B}_2(\mathbf{0}, 2)$$

and the affine subspaces by $\widehat{P}_{j,k} = \{\widehat{\mathbb{P}}_{j,k}(\mathbf{z}) \colon \mathbf{z} \in \mathbb{R}^N\}$. We again restrict the approximation to $\mathcal{B}_2(\mathbf{0}, 2)$. The single affine spaces will be non-empty as all $\widehat{\mathbf{c}}_{j,k}$ lie by definition close to $\mathcal{B}_2(\mathbf{0}, 1)$ if $\operatorname{supp}(\Pi)$ is close to $\mathcal{M}$, which we assume.

For the proof of our adapted theorem we do not need Theorem 5.1.3 in its full extent but only the following two bounds which can be deduced from (20) and (21) in [123] by setting $t = 2^{jd}$. As both appear in the proof of Theorem 5.1.3, we state them as a corollary. The interested reader may note that $n_{j,k}$ appearing in the original statements can be lower bounded by $\theta_1 n 2^{-jd}$.

Corollary 5.4.1. *Under the assumptions of Theorem 5.1.3 the following holds for any $C_1 > 0$ as long as j and α are sufficiently large and σ is sufficiently small:*

$$\mathsf{Pr}\left[\max_{k \in K_j} \left\| \mathbb{P}_{V_{j,k}} - \mathbb{P}_{\widehat{V}_{j,k}} \right\| \geq \frac{C_1}{12} 2^{-j-2}\right] \leq \frac{2}{\theta_2} 2^{jd} e^{-2^{jd} \min\left\{1, \frac{32\theta_2^2 d^2}{C_1^2}\right\}}$$

$$\mathsf{Pr}\left[\max_{k \in K_j} \left\| \mathbf{c}'_{j,k} - \widehat{\mathbf{c}}_{j,k} \right\|_2 \geq \frac{C_1}{12} 2^{-j-2}\right] \leq \frac{2}{\theta_2} 2^{jd} e^{-2^{jd} \min\left\{1, \frac{32\theta_2^2 d^2}{C_1^2}\right\}}$$

if $n \geq n_{min} = \left(2^{jd} + \log(\max\{\bar{d}, 8\})\right) \min\left\{144 \frac{\theta_2^2 d}{C_1 \theta_1 \theta_3} 2^{(d+1)j+3}, 96 \frac{\theta_2}{C_1 \theta_1} 2^{dj+1}\right\}^2.$

By Corollary 5.4.1 with probability at least $1-\mathcal{O}(2^{jd}\exp(-2^{jd}))$ the empirical centers $\widehat{\mathbf{c}}_{j,k}$ of one level j have a worst case distance to the perfect centers $\mathbf{c}'_{j,k}$ of at most $\mathcal{O}(2^{-j-2})$ if $n \gtrsim \mathcal{O}(2^{3jd})$. As a result, the empirical centers $\widehat{\mathbf{c}}_{j,k}$ will also be at most $\mathcal{O}(2^{-j-2})$ distance from their associated cover tree centers $\mathbf{a}_{j,k}$ if $n \gtrsim \mathcal{O}(2^{3jd})$ by assumption **(A2)**. The same holds true for the projectors $\mathbb{P}_{V_{j,k}}$ and $\mathbb{P}_{\widehat{V}_{j,k}}$ in operator norm. In addition to Corollary 5.4.1 we need a modified version of Lemma 5.3.3.

Lemma 5.4.2. *The Gaussian width of $\mathcal{M}\cup\mathcal{M}_j\cup\widehat{\mathcal{M}}_j$ can be bounded from above by*

$$\begin{aligned}\max\{w(\mathcal{M}),w(\mathcal{M}_j),w(\widehat{\mathcal{M}}_j)\} \le w(\mathcal{M}\cup\mathcal{M}_j\cup\widehat{\mathcal{M}}_j) &\le 2w(\mathcal{M})+2w(\mathcal{M}_j)+2w(\widehat{\mathcal{M}}_j)+5\\ &\le 2w(\mathcal{M})+C\sqrt{dj}.\end{aligned}$$

Proof: The proof follows directly the lines of the proof of Lemma 5.3.3. The additional term $w(\widehat{\mathcal{M}}_j)$ can be bounded in the same way as $w(\mathcal{M}_j)$. ■

Remark 5.4.3. *By structure of the proof one can easily obtain several subversions of the inequalities, e.g., $w(\mathcal{M}\cup\widehat{\mathcal{M}}_j)\le 2w(\mathcal{M})+2w(\widehat{\mathcal{M}}_j)+5$. We will use them while only referring to Lemma 5.4.2. Moreover, similar generalizations as in Lemma 5.3.3 apply (cf. Remark 5.3.4).*

We are ready to state an empirical version of Theorem 5.3.6 which automatically extends Theorem 5.3.1.

Theorem 5.4.4. *Let $\mathcal{M}\subset\mathbb{S}^{N-1}$ be given by its empirical GMRA for some levels $j_0\le j\le J$ from samples $X_1,...,X_n$ for $n\ge n_{min}$ (defined in Corollary 5.4.1), such that $0<C_1<2^{j_0+1}$ where C_1 is the constant from GMRA properties* (2b) *and* (3a) *for a GMRA structure constructed with centers $\mathbf{c}'_{j,k}$ and projectors $\Phi_{j,k}^T\Phi_{j,k}=\mathbb{P}_{V_{j,k}}$. Fix j and assume that* $\operatorname{dist}(\mathbf{0},\widehat{\mathcal{M}}_j)\ge 1/2$. *Further let*

$$m\ge 64\max\{C',\bar{C}\}C_1^{-6}2^{6(j+1)}(w(\mathcal{M})+C\sqrt{dj})^2.$$

where C' is the constant from Theorem 3.3.9, $\bar{C}$ from Theorem 3.3.5 and C from Lemma 5.4.2. Then, with probability at least $1-\mathcal{O}\left(2^{jd}\exp(-2^{jd})+\exp(2^{-2j}m)\right)$ the following holds for all $\mathbf{x}\in\mathcal{M}$ with one-bit measurements $y=\operatorname{sign}(\mathbf{Ax})$ and GMRA constants $\tilde{C}_{\mathbf{x}}$ from property (3b) *satisfying $\tilde{C}_{\mathbf{x}}<2^{j-1}$: The approximations $\mathbf{x}^*$ obtained by OMS with* (5.20) *and* (5.21) *fulfill*

$$\begin{aligned}&\|\mathbf{x}-\mathbf{x}^*\|_2\\ &\le\left(2\left(\tilde{C}_{\mathbf{x}}+\frac{C_1}{8}\right)2^{-\frac{j}{2}}+\sqrt{C_1}\sqrt[4]{\log\left(\frac{4e}{C_1}\right)}+\sqrt{22\tilde{C}_{\mathbf{x}}+\frac{55}{4}C_1}\sqrt[4]{\log\left(\frac{2e}{(2\tilde{C}_{\mathbf{x}}+\frac{5}{4}C_1)}\right)}\right)\sqrt[4]{j}2^{-\frac{j}{2}}.\end{aligned}$$

The proof of Theorem 5.4.4 follows the same steps as before. First, we give an empirical version of Lemma 5.3.7. Then we link $\mathbf{x}$ and $\mathbf{x}^*$ as in the proof of Theorem 5.3.6 while controlling the difference between empirical and axiomatic but unknown GMRA by Corollary 5.4.1.

Lemma 5.4.5. *Fix j sufficiently large. Under the assumptions of Theorem 5.1.3 and $n \geq n_{min}$ (defined in Corollary 5.4.1) if $m \geq \bar{C} C_1^{-6} 2^{6(j+1)} \max\{w(\mathcal{M} \cup \mathbb{P}_{\mathbb{S}}(\widehat{\mathcal{C}}_j))^2, 2/\pi\}$ the index k' of the center $\widehat{\mathbf{c}}_{j,k'}$ chosen in step I of the algorithm fulfills*

$$\|\mathbf{x} - \mathbf{c}'_{j,k'}\|_2 \leq 16 \max\{\|\mathbf{x} - \mathbf{c}'_{j,k_j(\mathbf{x})}\|_2, C_1 2^{-j-1}\}.$$

for all $\mathbf{x} \in \mathcal{M}$ with probability at least $1 - \mathcal{O}\left(2^{jd}\exp(-2^{jd}) + \exp(\delta^2 m)\right)$.

Proof: The proof will be similar to the one of Lemma 5.3.7. By definition we have

$$d_H(\operatorname{sign}(\mathbf{A}\widehat{\mathbf{c}}_{j,k'}), \mathbf{y}) \leq d_H(\operatorname{sign}(\mathbf{A}\widehat{\mathbf{c}}_{j,k_j(\mathbf{x})}), \mathbf{y}).$$

As, for all $\mathbf{z}, \mathbf{z}' \in \mathbb{S}^{N-1}$, $d_H(\operatorname{sign}(\mathbf{A}\mathbf{z}), \operatorname{sign}(\mathbf{A}\mathbf{z}')) = m \cdot d_A(\mathbf{z}, \mathbf{z}')$, this is equivalent to

$$d_A(\mathbb{P}_{\mathbb{S}}(\widehat{\mathbf{c}}_{j,k'}), \mathbf{x}) \leq d_A(\mathbb{P}_{\mathbb{S}}(\widehat{\mathbf{c}}_{j,k_j(\mathbf{x})}), \mathbf{x}).$$

Theorem 3.3.5 transfers the bound to normalized geodesic distance, namely

$$d_G(\mathbb{P}_{\mathbb{S}}(\widehat{\mathbf{c}}_{j,k'}), \mathbf{x}) \leq d_G(\mathbb{P}_{\mathbb{S}}(\widehat{\mathbf{c}}_{j,k_j(\mathbf{x})}), \mathbf{x}) + 2\delta$$

with probability at least $1 - 2\exp(-c\delta^2 m)$ where $\delta = C_1 2^{-j-1}$. By Lemma 5.3.5, $d_G(\mathbf{z}, \mathbf{z}') \leq \|\mathbf{z} - \mathbf{z}'\|_2 \leq \pi d_G(\mathbf{z}, \mathbf{z}')$, for all $\mathbf{z}, \mathbf{z}' \in \mathbb{S}^{N-1}$, which leads to

$$\begin{aligned}\|\mathbb{P}_{\mathbb{S}}(\widehat{\mathbf{c}}_{j,k'}) - \mathbf{x}\|_2 &\leq \pi d_G(\mathbb{P}_{\mathbb{S}}(\widehat{\mathbf{c}}_{j,k_j(\mathbf{x})}), \mathbf{x}) + 2\pi\delta \\ &\leq \pi \|\mathbb{P}_{\mathbb{S}}(\widehat{\mathbf{c}}_{j,k_j(\mathbf{x})}) - \mathbf{x}\|_2 + 2\pi\delta.\end{aligned}$$

We will now use the fact that by Corollary 5.4.1

$$\|\widehat{\mathbf{c}}_{j,k} - \mathbf{c}'_{j,k}\|_2 \leq \frac{C_1}{12} 2^{-j-2},$$

for all $k \in K_j$ with probability at least $1 - \mathcal{O}(2^{jd}\exp(-2^{jd}))$. From this we first deduce by GMRA property (3a) that $\|\widehat{\mathbf{c}}_{j,k} - \mathbb{P}_{\mathbb{S}}(\widehat{\mathbf{c}}_{j,k})\|_2 \leq \|\widehat{\mathbf{c}}_{j,k} - \mathbb{P}_{\mathbb{S}}(\mathbf{c}'_{j,k})\|_2 \leq \|\widehat{\mathbf{c}}_{j,k} - \mathbf{c}'_{j,k}\|_2 + \|\mathbf{c}'_{j,k} - \mathbb{P}_{\mathbb{S}}(\mathbf{c}'_{j,k})\|_2 < (C_1 + C_1/2)2^{-j-2}$ for all $\widehat{\mathbf{c}}_{j,k} \in \widehat{\mathcal{C}}_j$. Combining above estimates and using triangle inequality we obtain

$$\begin{aligned}\|\mathbf{c}'_{j,k'} - \mathbf{x}\|_2 &\leq \|\mathbf{c}'_{j,k'} - \widehat{\mathbf{c}}_{j,k'}\|_2 + \|\widehat{\mathbf{c}}_{j,k'} - \mathbb{P}_{\mathbb{S}}(\widehat{\mathbf{c}}_{j,k'})\|_2 + \|\mathbb{P}_{\mathbb{S}}(\widehat{\mathbf{c}}_{j,k'}) - \mathbf{x}\|_2 \\ &< 2C_1 2^{-j-2} + \pi\|\mathbb{P}_{\mathbb{S}}(\widehat{\mathbf{c}}_{j,k_j(x)}) - \mathbf{x}\|_2 + 2\pi\delta \\ &\leq C_1 2^{-j-1} + \pi(\|\widehat{\mathbf{c}}_{j,k_j(\mathbf{x})} - \mathbb{P}_{\mathbb{S}}(\widehat{\mathbf{c}}_{j,k_j(\mathbf{x})})\|_2 + \|\widehat{\mathbf{c}}_{j,k_j(\mathbf{x})} - \mathbf{c}'_{j,k_j(\mathbf{x})}\|_2 \\ &+ \|\mathbf{c}'_{j,k_j(\mathbf{x})} - \mathbf{x}\|_2) + 2\pi\delta \\ &< C_1 2^{-j-1} + \pi C_1 2^{-j-1} + \pi\|\mathbf{c}_{j,k_j(\mathbf{x})} - \mathbf{x}\|_2 + 2\pi\delta \\ &\leq (4\pi + 1)\max\{\|\mathbf{c}'_{j,k_j(\mathbf{x})} - \mathbf{x}\|_2, C_1 2^{-j-1}\} \\ &\leq 16\max\{\|\mathbf{c}'_{j,k_j(\mathbf{x})} - \mathbf{x}\|_2, C_1 2^{-j-1}\}.\end{aligned}$$

A union bound over both probabilities yields the result. ■

Proof of Theorem 5.4.4 : The proof consists of the same three steps as the one of Theorem 5.3.6. First, we apply Lemma 5.4.5 in **(I)**. By the GMRA axioms this supplies an estimate for $\|\mathbf{x} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2$ with high probability (recall that $\mathbb{P}_{j,k'}(\mathbf{x})$ will be $\mathbb{P}_{V_{j,k'}}(\mathbf{x} - \mathbf{c}'_{j,k'}) + \mathbf{c}'_{j,k'}$ in this case). In **(II)** we use **(I)** to deduce a bound on $\|\mathbf{x} - \widehat{\mathbb{P}}_{j,k'}(\mathbf{x})\|_2$, and then use Theorem 3.3.9 to bound the distance between $\widehat{\mathbb{P}}_{j,k'}(\mathbf{x})/\|\widehat{\mathbb{P}}_{j,k'}(\mathbf{x})\|_2$ and the minimizer $\mathbf{x}^*$ of

$$\begin{aligned} \mathbf{x}^* = \arg\min_{\mathbf{z}} \sum_{l=1}^{m} (-y_l)\langle \mathbf{a}_l, \mathbf{z}\rangle, \\ \text{subject to } \mathbf{z} \in K := \operatorname{conv}\left(\mathbb{P}_{\mathbb{S}}(\widehat{P}_{j,k'} \cap \mathcal{B}_2(\mathbf{0}, 2))\right), \end{aligned} \tag{5.22}$$

with high probability. Taking the union bound over all events, part **(III)** then concludes with an estimate of the distance $\|\mathbf{x} - \mathbf{x}^*\|_2$ by combining **(I)** and **(II)**.

(I) Set $\delta = C_1 2^{-j-1}$ and recall that $C_1 2^{-j-2} < 1/2$ by assumption which implies by GMRA property (3a) that all centers in $\mathcal{C}'_j$ are closer to $\mathbb{S}^{N-1}$ than $1/2$, i.e. $1/2 \le \|\mathbf{c}'_{j,k}\|_2 \le 3/2$. Moreover, Corollary 5.4.1 holds with probability at least $1 - \mathcal{O}(2^{jd}\exp(-2^{jd}))$ and implies $\|\widehat{\mathbf{c}}_{j,k} - \mathbf{c}'_{j,k}\|_2 \le (C_1/12)2^{-j-2} \le 1/4$. Hence, by triangle inequality $1/4 \le \|\widehat{\mathbf{c}}_{j,k}\|_2 \le 7/4$. From this and (3.6) we deduce

$$\begin{aligned} w(\mathbb{P}_{\mathbb{S}}(\widehat{\mathcal{C}}_j)) \le \gamma(\mathbb{P}_{\mathbb{S}}(\widehat{\mathcal{C}}_j)) \le 4\gamma(\widehat{\mathcal{C}}_j) \le 8w(\widehat{\mathcal{C}}_j) + 4\sqrt{\frac{2}{\pi}}\operatorname{dist}(\mathbf{0}, \widehat{\mathcal{C}}_j) \\ \le 8w(\widehat{\mathcal{C}}_j) + 8. \end{aligned} \tag{5.23}$$

As $\widehat{\mathcal{C}}_j \subset \widehat{\mathcal{M}}_j$ we know by Lemma 5.4.2 and (5.23) that

$$\begin{aligned} m &\ge 16\bar{C}\delta^{-6}(2w(\mathcal{M}) + 2C\sqrt{dj})^2 \ge 16\bar{C}\delta^{-6}(2w(\mathcal{M}) + 4w(\widehat{\mathcal{C}}_j) + 10)^2 \\ &\ge 4\bar{C}\delta^{-6}(4w(\mathcal{M}) + 8w(\widehat{\mathcal{C}}_j) + 20)^2 \ge 4\bar{C}\delta^{-6}(4w(\mathcal{M}) + w(\mathbb{P}_{\mathbb{S}}(\widehat{\mathcal{C}}_j)) + 12)^2 \\ &\ge \bar{C}\delta^{-6}(8w(\mathcal{M}) + 2w(\mathbb{P}_{\mathbb{S}}(\widehat{\mathcal{C}}_j)) + 24)^2 \ge \bar{C}\delta^{-6}(w(\mathcal{M} \cup \mathbb{P}_{\mathbb{S}}(\widehat{\mathcal{C}}_j)) + 19)^2 \\ &\ge \bar{C}\delta^{-6}\max\left\{w(\mathcal{M} \cup \mathbb{P}_{\mathbb{S}}(\widehat{\mathcal{C}}_j))^2, \frac{2}{\pi}\right\}. \end{aligned} \tag{5.24}$$

Hence, Lemma 5.4.5 implies

$$\|\mathbf{x} - \mathbf{c}'_{j,k'}\|_2 \le 16\max\left\{\left\|\mathbf{x} - \mathbf{c}'_{j,k_j(\mathbf{x})}\right\|_2, C_1 2^{-j-1}\right\}$$

with probability at least $1 - \mathcal{O}\left(2^{jd}\exp(-2^{jd}) + \exp(\delta^2 m)\right)$. By GMRA property (3b) we get

$$\|\mathbf{x} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2 \le \tilde{C}_{\mathbf{x}} 2^{-j}, \tag{5.25}$$

for some constant $\tilde{C}_{\mathbf{x}}$.

(II) Define $\widehat{\alpha} := \left\| \widehat{\mathbb{P}}_{j,k'}(\mathbf{x}) \right\|_2$. Note that $\|\mathbf{x} - \mathbf{c}'_{j,k'}\|_2 \leq 4$ as $\mathbf{x} \in \mathbb{S}^{N-1}$ and all $\mathbf{c}'_{j,k}$ are close to the sphere by assumption. Hence,

$$\begin{aligned}\|\mathbb{P}_{j,k'}(\mathbf{x}) - \widehat{\mathbb{P}}_{j,k'}(\mathbf{x})\|_2 &= \|\mathbf{c}'_{j,k'} + \mathbb{P}_{V_{j,k'}}(\mathbf{x} - \mathbf{c}'_{j,k'}) - \widehat{\mathbf{c}}_{j,k'} - \mathbb{P}_{\widehat{V}_{j,k'}}(\mathbf{x} - \widehat{\mathbf{c}}_{j,k'})\|_2 \\ &\leq \|\mathbf{c}'_{j,k'} - \widehat{\mathbf{c}}_{j,k'}\|_2 + \|\mathbb{P}_{V_{j,k'}} - \mathbb{P}_{\widehat{V}_{j,k'}}\|\|\mathbf{x} - \mathbf{c}'_{j,k'}\|_2 + \|\mathbf{c}'_{j,k'} - \widehat{\mathbf{c}}_{j,k'}\|_2 \\ &\leq \frac{2}{12}C_1 2^{-j-2} + \frac{1}{12}C_1 2^{-j-2}\|\mathbf{x} - \mathbf{c}'_{j,k'}\|_2 \leq \frac{1}{2}C_1 2^{-j-2}\end{aligned}$$

by application of Corollary 5.4.1. This implies $1/4 \leq \widehat{\alpha} \leq 7/4$ as $\mathbf{x} \in \mathbb{S}^{N-1}$ and

$$\begin{aligned}\|\mathbf{x} - \widehat{\mathbb{P}}_{j,k'}(\mathbf{x})\|_2 &\leq \|\mathbf{x} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2 + \|\mathbb{P}_{j,k'}(\mathbf{x}) - \widehat{\mathbb{P}}_{j,k'}(\mathbf{x})\|_2 \\ &\leq \tilde{C}_{\mathbf{x}} 2^{-j} + \frac{1}{2}C_1 2^{-j-2} \leq \frac{3}{4}\end{aligned} \tag{5.26}$$

by (5.25) and the assumption that $\max\{\tilde{C}_{\mathbf{x}}, C_1/4\} \cdot 2^{-j} \leq 1/2$. As before we create the setting of Theorem 3.3.9.

Define $\tilde{\mathbf{x}} := \widehat{\mathbb{P}}_{j,k'}(\mathbf{x})/\widehat{\alpha} \in \mathbb{S}^{N-1}$, $\tilde{\mathbf{y}} := \text{sign}(A\tilde{\mathbf{x}}) = \text{sign}(A\widehat{\mathbb{P}}_{j,k'}(\mathbf{x}))$, $K = \text{conv}(\mathbb{P}_{\mathbb{S}}(\widehat{P}_{j,k'} \cap \mathcal{B}_2(\mathbf{0}, 2)))$ and $\tau := (2\tilde{C}_{\mathbf{x}} + \frac{5}{4}C_1)2^{-j}$. If applied to this, Theorem 3.3.9 would give the desired bound on $\|\tilde{\mathbf{x}} - \mathbf{x}^*\|_2$. We first have to check $d_H(\tilde{\mathbf{y}}, \mathbf{y}) \leq \tau m$. Recall that $\frac{1}{\widehat{\alpha}} \leq 4$ and as $\widehat{\alpha} > 0$ one has $\widehat{\alpha} w(K) = w(\widehat{\alpha} K)$. By applying Lemma 5.4.2 again we have that

$$\begin{aligned}m &\geq 64\bar{C}\delta^{-6}(w(\mathcal{M}) + C\sqrt{dj})^2 \geq 16\bar{C}\delta^{-6}(2w(\mathcal{M}) + 2w(\widehat{\mathcal{M}}_j) + 5)^2 \\ &= \bar{C}\delta^{-6}(8w(\mathcal{M}) + 8w(\widehat{\mathcal{M}}_j) + 20)^2 \geq \bar{C}\delta^{-6}\left(8w(\mathcal{M}) + 2w\left(\frac{\widehat{\mathcal{M}}_j}{\widehat{\alpha}}\right) + 20\right)^2 \\ &\geq \bar{C}\delta^{-6}\max\left\{w\left(\left(\mathcal{M} \cup \frac{\widehat{\mathcal{M}}_j}{\widehat{\alpha}}\right) \cap B(\mathbf{0}, 1)\right)^2, \frac{2}{\pi}\right\}.\end{aligned}$$

We may now use Theorem 3.3.5, Lemma 5.3.5 and $\|\tilde{\mathbf{x}} - \widehat{\mathbb{P}}_{j,k'}(\mathbf{x})\|_2 = |1 - \widehat{\alpha}| \leq \|\mathbf{x} - \widehat{\mathbb{P}}_{j,k'}(\mathbf{x})\|_2$ to obtain

$$\begin{aligned}\frac{d_H(\tilde{\mathbf{y}}, \mathbf{y})}{m} &= d_A(\tilde{\mathbf{x}}, \mathbf{x}) \leq d_G(\tilde{\mathbf{x}}, \mathbf{x}) + 2\delta \leq \|\tilde{\mathbf{x}} - \mathbf{x}\|_2 + 2\delta \\ &\leq \|\tilde{\mathbf{x}} - \widehat{\mathbb{P}}_{j,k'}(\mathbf{x})\|_2 + \|\widehat{\mathbb{P}}_{j,k'}(\mathbf{x}) - \mathbf{x}\|_2 + 2\delta \\ &\leq 2\|\widehat{\mathbb{P}}_{j,k'}(\mathbf{x}) - \mathbf{x}\|_2 + 2\delta \leq 2\tilde{C}_{\mathbf{x}} 2^{-j} + C_1 2^{-j-2} + 2\delta \\ &\leq (2\tilde{C}_{\mathbf{x}} + \frac{5}{4}C_1)2^{-j} = \tau\end{aligned}$$

with probability at least $1 - 2\exp(-c\delta^2 m)$. Assuming the above events hold true we can apply Theorem 3.3.9 as by Lemma 5.4.2, in analogy to (5.24) and (5.19), that

$$\begin{aligned}m &\geq 4C'\delta^{-6}(2w(\mathcal{M}) + 4w(\mathcal{M}_j) + 4w(\widehat{\mathcal{M}}_j) + 10)^2 \\ &\geq C'\delta^{-6}w(\mathbb{P}_{\mathbb{S}}(\widehat{P}_{j,k'} \cap \mathcal{B}_2(\mathbf{0}, 2))) \\ &\geq C'\delta^{-6}w(K)^2\end{aligned}$$

and obtain with probability at least $1 - 8\exp(-c\delta^2 m)$

$$\|\tilde{\mathbf{x}} - \mathbf{x}^*\|_2^2 \leq \delta\sqrt{\log\left(\frac{e}{\delta}\right)} + 11\tau\sqrt{\log\left(\frac{e}{\tau}\right)}. \tag{5.27}$$

(III) We conclude as in Theorem 5.3.6. Recall that $\|\tilde{\mathbf{x}} - \widehat{\mathbb{P}}_{j,k'}(\mathbf{x})\|_2 = |1-\alpha| \leq \|\mathbf{x} - \mathbb{P}_{j,k'}(\mathbf{x})\|_2 \leq (\tilde{C}_{\mathbf{x}} + \frac{C_1}{8})2^{-j}$. By union bound we obtain with probability at least $1 - \mathcal{O}\left(2^{jd}\exp(-2^{jd}) + \exp(\delta^2 m)\right)$

$$\begin{aligned}
\|\mathbf{x} - \mathbf{x}^*\|_2 &\leq \|\mathbf{x} - \widehat{\mathbb{P}}_{j,k'}(\mathbf{x})\|_2 + \|\widehat{\mathbb{P}}_{j,k'}(\mathbf{x}) - \tilde{\mathbf{x}}\|_2 + \|\tilde{\mathbf{x}} - \mathbf{x}^*\|_2 \\
&\leq 2\|\mathbf{x} - \widehat{\mathbb{P}}_{j,k'}(\mathbf{x})\|_2 + \sqrt{\delta\sqrt{\log\left(\frac{e}{\delta}\right)} + 11\tau\sqrt{\log\left(\frac{e}{\tau}\right)}} \\
&\leq 2\left(\tilde{C}_{\mathbf{x}} + \frac{C_1}{8}\right)2^{-j} + \sqrt{C_1}2^{\frac{-j-1}{2}}\sqrt[4]{\log\left(\frac{e}{C_1 2^{-j-1}}\right)} \\
&\quad + \sqrt{22\tilde{C}_{\mathbf{x}} + \frac{55}{4}C_1}2^{-\frac{j}{2}}\sqrt[4]{\log\left(\frac{e}{(2\tilde{C}_{\mathbf{x}} + \frac{5}{4}C_1)2^{-j}}\right)} \\
&\leq \left(2\left(\tilde{C}_{\mathbf{x}} + \frac{C_1}{8}\right)2^{-\frac{j}{2}} + \sqrt{C_1}\sqrt[4]{\log\left(\frac{4e}{C_1}\right)}\right. \\
&\quad \left. + \sqrt{22\tilde{C}_{\mathbf{x}} + \frac{55}{4}C_1}\sqrt[4]{\log\left(\frac{2e}{(2\tilde{C}_{\mathbf{x}} + \frac{5}{4}C_1)}\right)}\right)\sqrt[4]{j}2^{-\frac{j}{2}}.
\end{aligned}$$

■

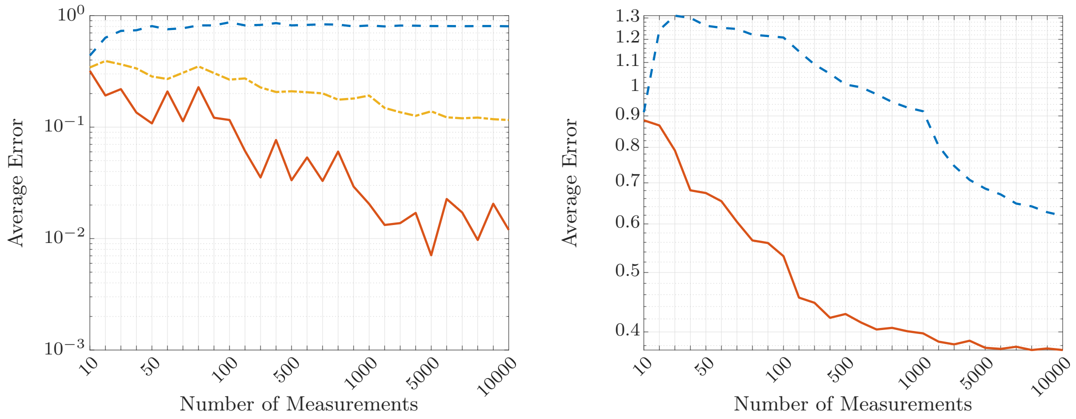
5.5 Numerical Simulation

In this section we present various numerical experiments to benchmark OMS. The GMRAs we work with are constructed using code provided by Maggioni[1]. We compared the performance of OMS for two exemplary choices of $\mathcal{M}$, namely, a simple 2-dim sphere embedded in $\mathbb{R}^{20}$ (20000 data points sampled from the 2-dimensional sphere $\mathcal{M}$ embedded in $\mathbb{S}^{20-1}$) and the MNIST data set [113] of handwritten digits "1" (3000 data points in $\mathbb{R}^{784}$) where the restriction to the single digit 1 was done to keep the underlying manifold as simple as possible. In each of the experiments in Sections 5.5.1-5.5.4 we first computed a GMRA up to refinement level $j_{\max} = 10$ and then recovered 100 randomly chosen $\mathbf{x} \in \mathcal{M}$ from their one-bit measurements by applying OMS. Depicted is the averaged relative error between $\mathbf{x}$ and its approximation $\mathbf{x}^*$, i.e., $\|\mathbf{x} - \mathbf{x}^*\|_2/\|\mathbf{x}\|_2$ which is equal to the absolute error $\|\mathbf{x} - \mathbf{x}^*\|_2$ for $\mathcal{M} \subset \mathbb{S}^{N-1}$. Note the different approximation error ranges of the sphere and the MNIST experiments when comparing both settings.

[1] The code is available at `http://www.math.jhu.edu/~mauro/#tab_code`.

5.5.1 OMS-simple vs. OMS

The first test illustrates recovery performance of the two algorithms presented above, namely OMS-Simple for $R \in \{0.5, 1, 1.5\}$ and OMS. The results are depicted in Figure 5.5. Note that only $R = 1.5$ and, in the case of the 2-sphere, $R = 1$ are depicted as in the respective other cases for each number of measurements most of the trials did not yield a feasible solution in (5.6) so the average was not well-defined. One can observe that for both data sets OMS outperforms OMS-Simple which is not surprising as OMS does not rely on a suitable parameter choice. This observation is also the reason for us to restrict the theoretical analysis to OMS. The more detailed approximation of the toy example (2-dimensional sphere) is due to its simpler structure and lower dimensional setting. This difference in approximation quality can also be observed in Sections 5.5.2-5.5.4.



(a) 2-Sphere

(b) MNIST

Figure 5.5: Comparison of OMS-Simple for $R = 1$ (dotted-dashed, yellow), $R = 1.5$ (dashed, blue) and OMS (solid, red).

5.5.2 Modifying OMS

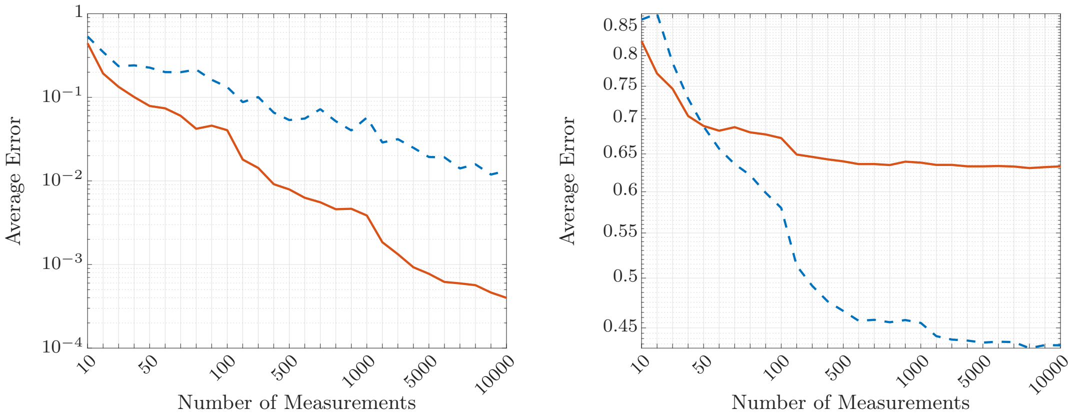
In the second experiment we compared OMS to a slightly different version in which (5.9) is replaced by

$$\mathbf{x}^* = \arg\min_{\mathbf{z}\in\mathbb{R}^N} \sum_{l=1}^{m} [(-y_l)\langle \mathbf{a}_l, \mathbf{z}\rangle]_+ , \quad \text{subject to } \mathbf{z} \in \operatorname{conv}\left(\mathbb{P}_{\mathbb{S}}(P_{j,k'} \cap \mathcal{B}_2(0,2))\right),$$

where $[t]_+ = \max\{0, t\}$ denotes the positive part of $t \in \mathbb{R}$. This is motivated by the observation that (3.21) can be re-stated equivalently as (3.22). Hence, (3.21) punishes incorrect measurements of a feasible point $\mathbf{z} \in K$ by its distance to the 'measurements border' $H_{\mathbf{a}_l}$ while rewarding correct ones. The second part which rewards might cause problems as it pushes minimizers away from the hyperplanes $H_{\mathbf{a}_l}$ of correct measurements. If the true $\mathbf{x}$, however, lies close to one of them, this may be suboptimal. Hence, we dropped the rewarding term in (3.22) leading to

$$\operatorname*{arg\,min}_{\mathbf{z}\in\mathbb{R}^D} \sum_{l=1}^{m} \left[(-y_l)\langle \mathbf{a}_l, \mathbf{z}\rangle\right]_+ , \quad \text{subject to } \mathbf{z} \in K, \tag{5.28}$$

which is still convex but performs better numerically in some cases. As depicted in Figure 5.6, the version with $[\cdot]_+$ clearly outperforms the one without if $\mathcal{M}$ is the 2-dimensional sphere. In contrast, if $\mathcal{M}$ is more complex (MNIST data), the $[\cdot]_+$ formulation clearly fails. We have no satisfactory explanation for this difference in behavior so far.



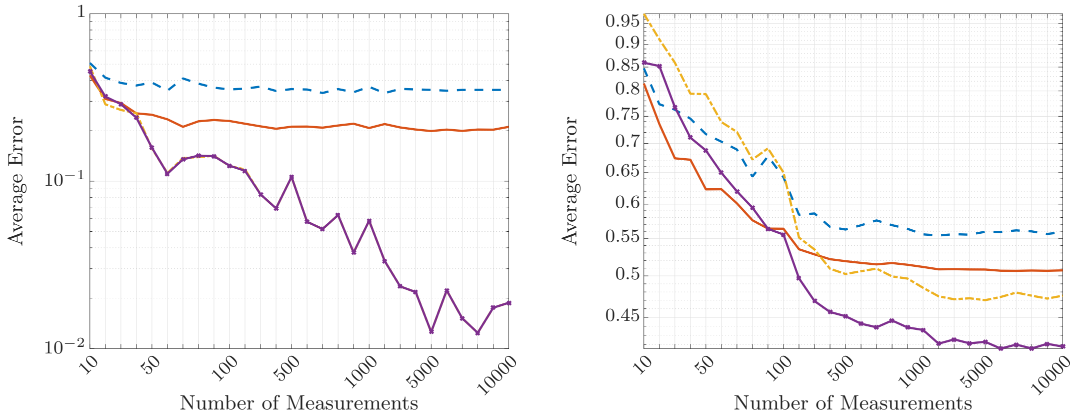
(a) 2-Sphere

(b) MNIST

Figure 5.6: Comparison of OMS (dashed, blue) and a modified version (solid, red) as described in (5.28).

5.5.3 Are Two Steps Necessary?

One might wonder if the two steps in OMS-Simple and OMS are necessary at all. Wouldn't it be sufficient to use the center $\mathbf{c}_{j,k'}$ determined in step I. as an approximation of $\mathbf{x}$? If the GMRA is fine enough, this indeed is the case. If one only has access to a rather rough GMRA, the simulations in Figure 5.7 show that the second step makes a notable difference in approximation quality. This behavior suits Lemma 5.3.7. The lemma guarantees a good approximation of $\mathbf{x}$ by $\mathbf{c}_{j,k'}$ as long as $\mathbf{x}$ is well approximated by an optimal center. In the MNIST case, one can observe that the second step only improves performance if the number of one-bit measurements is sufficiently high. For a small set of measurements the centers might yield better approximation as they lie close to $\mathcal{M}$ by GMRA property (3a). On the other hand, only parts of the affine spaces (the ones inside $\mathcal{B}_2(\mathbf{0}, 2)$) are practical for approximation and a certain number of measurements is necessary to restrict II. to the relevant parts.



(a) 2-Sphere

(b) MNIST

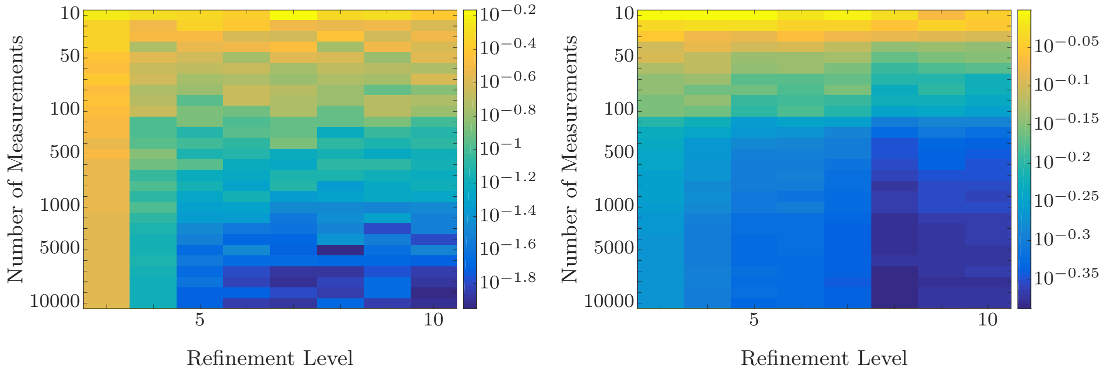
Figure 5.7: Comparison of the following: Approximation by step I. of OMS when using tree structure (dashed, blue) and when comparing all centers (solid, red); approximation by step I.+II. of OMS when using tree structure (dashed with points, yellow) and when comparing all centers (solid with points, purple).

5.5.4 Tree vs. No Tree

In the fourth test we checked if approximation still works when not all possible centers are compared in step I. of OMS but their tree structure is used. This means to find an optimal center one compares on the first refinement level all centers, and then continues in each subsequent level solely with the children of the k best centers (in the presented experiments we chose $k = 10$). Of course, the chosen center will not be optimal as not all centers are compared (see Figure 5.7). In the simple 2-dimensional sphere setting, step II. can compensate the worse approximation quality of I. with tree search. Figure 5.7 hardly shows a difference in final approximation quality in both cases. In the MNIST setting one can observe a considerable difference even when performing two steps.

5.5.5 A Change of Refinement Level

The last experiment (see Figure 5.8) examines the influence of the refinement level j on the approximation error. For small j (corresponding to a rough GMRA) a high number of measurements can hardly improve the approximation quality while for large j (corresponding to a fine GMRA) the approximation error decreases with increasing measurement rates. This behavior is as expected. A rough GMRA cannot profit much from many measurements as the GMRA approximation itself yields a lower bound on obtainable approximation error. For fine GMRAs the behavior along the measurement axis is similar to above experiments. Note that further increase of j for the same range of measurements does not improve accuracy.



(a) 2-Sphere (b) MNIST

Figure 5.8: Approximation error of OMS for different refinement levels j and numbers of measurements.

5.6 Alternative Width Bounds

For obtaining the lower bounds on m in (5.13) and (5.10) we made use of Lemma 5.3.3 leading to the influence of j which is suboptimal for fine scales (i.e., j large). To improve on this for large j one can exploit two alternative versions of the lemma which apply to a more restrictive definition of $\mathcal{M}_j$, namely,

$$\mathcal{M}_j^{rel} := \{\mathbb{P}_{j,k_j(\mathbf{z})}(\mathbf{z}) \colon \mathbf{z} \in \mathcal{M}\} \cap \mathcal{B}_2(\mathbf{0}, 2).$$

In order to derive the bounds three technical tools are needed. Recall the definition of $C_\mathbf{z}$ in GMRA property (3b).

Lemma 5.6.1. *Set $C_\mathcal{M} := \sup_{\mathbf{z}\in\mathcal{M}} C_\mathbf{z}$. Then, $\mathcal{M}_j^{rel} \subset \text{tube}_{C_\mathcal{M} 2^{-2j}}(\mathcal{M})$.*

Proof: If $\mathbf{x} \in \mathcal{M}_j^{rel}$ there exists $\mathbf{z}_\mathbf{x} \in \mathcal{M}$ such that $\mathbf{x} = \mathbb{P}_{j,k_j(\mathbf{z}_\mathbf{x})}(\mathbf{z}_\mathbf{x})$. The Euclidean distance $d(\mathbf{x},\mathcal{M})$ therefore satisfies

$$d(\mathbf{x},\mathcal{M}) = \inf_{\mathbf{z}\in\mathcal{M}} \|\mathbf{x}-\mathbf{z}\|_2 \leq \|\mathbb{P}_{j,k_j(\mathbf{z}_\mathbf{x})}(\mathbf{z}_\mathbf{x}) - \mathbf{z}_\mathbf{x}\|_2 \leq C_\mathcal{M} 2^{-2j}$$

by property (3b). ■

Given a subset $S \subset \mathbb{R}^N$, we let $P(S,\varepsilon)$ denote the maximal packing number of S (i.e., the maximum cardinality of a subset of S whose elements are all separated from one another with Euclidean distance $\varepsilon > 0$).

Lemma 5.6.2. *Set $C_\mathcal{M} := \sup_{\mathbf{z}\in\mathcal{M}} C_\mathbf{z}$. Then $N(\mathcal{M}_j^{rel},\varepsilon) \leq N(\mathcal{M},\varepsilon/2)$ for all $\varepsilon \geq 2C_\mathcal{M}2^{-2j}$.*

Proof: First note that for all $\eta \geq \rho := C_\mathcal{M}2^{-2j}$ Lemma 5.6.1 implies that

$$\mathcal{M}_j^{rel} \subset \text{tube}_\rho(\mathcal{M}) \subset \bigcup_{\mathbf{z}\in M_\eta} \mathcal{B}_2(\mathbf{z}, 2\eta),$$

where M_η is an η-cover of $\mathcal{M}$. Thus, for all $\varepsilon \geq 2\eta \geq 2\rho$

$$N(\mathcal{M}_j^{rel}, \varepsilon) \leq N\left(\bigcup_{\mathbf{z}\in M_\eta} \mathcal{B}_2(\mathbf{z}, 2\eta), \varepsilon\right) \leq N(\mathcal{M}, \eta) = N\left(\mathcal{M}, \frac{\varepsilon}{2}\right). \qquad \blacksquare$$

Lemma 5.6.3. *$N(\mathcal{M}_j^{rel}, \varepsilon) \leq (6/\varepsilon)^d N(\mathcal{M}, \varepsilon)$ for all $\varepsilon \leq \frac{1}{4}C_1 2^{-j}$ as long as $j > j_0$ (see properties* (3a) *and* (2b)*).*

Proof: By properties (3a) and (2b) every center $\mathbf{c}_{j,k}$ has an associated $\mathbf{p}_{j,k} \in \mathcal{M}$ such that both $\mathcal{B}_2\left(\mathbf{p}_{j,k}, C_1 2^{-j-2}\right) \subset \mathcal{B}_2\left(\mathbf{c}_{j,k}, C_1 2^{-j-1}\right)$ and $\mathcal{B}_2\left(\mathbf{p}_{j,k}, C_1 2^{-j-2}\right) \cap \mathcal{B}_2\left(\mathbf{c}_{j,k'}, C_1 2^{-j-1}\right) = \emptyset$ for all $k \neq k'$. Let $\tilde{P}_j := \{\mathbf{p}_{j,k} \mid k \in [K_j]\}$. Consequently, we have that $K_j = |\tilde{P}_j|$ and $\|\mathbf{p}_{j,k} - \mathbf{p}_{j,k'}\|_2 \geq C_1 2^{-j-1}$ for all $k \neq k'$. Since $\tilde{P}_j$ is a $C_1 2^{-j-1}$-packing of $\mathcal{M}$ we can further see that

$$K_j \leq P\left(\mathcal{M}, C_1 2^{-j-1}\right) \leq N\left(\mathcal{M}, C_1 2^{-j-2}\right) \leq N(\mathcal{M}, \varepsilon),$$

for all $\varepsilon \leq C_1 2^{-j-2}$. Now, $\mathcal{M}_j^{rel} \subset L_j$, where L_j is defined as in the proof of Lemma 5.3.3 (this proof also discusses its covering numbers). As a result we have that

$$N(\mathcal{M}_j^{rel}, \varepsilon) \leq N(L_j, \varepsilon) \leq K_j (6/\varepsilon)^d \leq N(\mathcal{M}, \varepsilon) \cdot (6/\varepsilon)^d$$

holds for all $\varepsilon \leq C_1 2^{-j-2}$. $\blacksquare$

We can now show the first alternative version of Lemma 5.3.3 which is independent of the refinement level j.

Lemma 5.6.4 (A Bound of the Gaussian Width for Fine Scales)**.** *If $j \geq \log_2(N)$ and $\max\{1, \sup_{\mathbf{z}\in\mathcal{M}} C_\mathbf{z}\} =: C_\mathcal{M} < \infty$, we obtain*

$$\begin{aligned}\max\{w(\mathcal{M}), w(\mathcal{M}_j^{rel})\} \leq w(\mathcal{M} \cup \mathcal{M}_j^{rel}) &\leq 2w(\mathcal{M}) + 2w(\mathcal{M}_j^{rel}) + 3 \\ &\leq C(w(\mathcal{M}) + 1)\log(N).\end{aligned}$$

Proof: We aim to bound $w(\mathcal{M}_j^{rel})$ in terms of $w(\mathcal{M})$. By the two-sided Sudakov inequality [166] and Lemma 5.6.1 we get that

$$\begin{aligned}w(\mathcal{M}_j^{rel}) &\leq C\log(N) \sup_{\varepsilon\geq 0} \varepsilon\sqrt{\log N(\mathcal{M}_j^{rel}, \varepsilon))} \\ &\leq C\log(N)\left(\sup_{0\leq\varepsilon\leq 2C_\mathcal{M}2^{-2j}} \varepsilon\sqrt{\log N(\mathrm{tube}_{C_\mathcal{M}2^{-2j}}(\mathcal{M}), \varepsilon)}\right. \\ &\quad \left. + \sup_{\varepsilon\geq 2C_\mathcal{M}2^{-2j}} \varepsilon\sqrt{\log N(\mathcal{M}_j^{rel}, \varepsilon)}\right) \\ &\leq C\log(N)\left(\sup_{0\leq\varepsilon\leq 2C_\mathcal{M}2^{-2j}} \varepsilon\sqrt{\log N(\mathcal{B}_2(\mathbf{0}, 1 + C_\mathcal{M}), \varepsilon))}\right. \\ &\quad \left. + \sup_{\varepsilon\geq 2C_\mathcal{M}2^{-2j}} \varepsilon\sqrt{\log N(\mathcal{M}, \varepsilon/2)}\right),\end{aligned}$$

where the last inequality follows from $\text{tube}_{C_{\mathcal{M}}2^{-2j}}(\mathcal{M}) \subset \mathcal{B}_2(\mathbf{0}, 1+C_{\mathcal{M}})$ and Lemma 5.6.2. Appealing to the Sudakov inequality once more to bound the second term above we learn that

$$\begin{aligned} w(\mathcal{M}_j^{rel}) \leq C\log(N) & \left(\sup_{0\leq\varepsilon\leq 2C_{\mathcal{M}}2^{-2j}} \varepsilon\sqrt{\log N(\mathcal{B}_2(\mathbf{0}, 1+C_{\mathcal{M}}), \varepsilon))} \right. \\ & \left. +2\sup_{\varepsilon\geq 0} \frac{\varepsilon}{2}\sqrt{\log N(\mathcal{M}, \varepsilon/2)} \right) \\ & \leq C\log(N) \left(\sup_{0\leq\varepsilon\leq 2C_{\mathcal{M}}2^{-2j}} \varepsilon\sqrt{\log N(\mathcal{B}_2(\mathbf{0}, 1+C_{\mathcal{M}}), \varepsilon))} + 2c\, w(\mathcal{M}) \right). \end{aligned}$$

To bound the first term above we note that the covering number of $\mathcal{B}_2(\mathbf{0}, 1+C_{\mathcal{M}})$ can be bounded as

$$N\left(\mathcal{B}_2(\mathbf{0}, 1+C_{\mathcal{M}}), \varepsilon\right) = N\left(\mathcal{B}_2(\mathbf{0}, 1), \frac{\varepsilon}{1+C_{\mathcal{M}}}\right) \leq \left(1+\frac{2+2C_{\mathcal{M}}}{\varepsilon}\right)^N \leq \left(\frac{4C_{\mathcal{M}}+4}{\varepsilon}\right)^N.$$

As $\varepsilon \mapsto \varepsilon\sqrt{\log(\frac{4C_{\mathcal{M}}+4}{\varepsilon})}$ is non decreasing for $\varepsilon \in (0, 2C_{\mathcal{M}}2^{-2j})$, we obtain by assuming that $\log_2(N) \leq j$

$$\begin{aligned} \sup_{0\leq\varepsilon\leq 2C_{\mathcal{M}}2^{-2j}} \varepsilon\sqrt{\log N(\mathcal{B}_2(\mathbf{0}, 1+C_{\mathcal{M}}), \varepsilon))} & \leq \sup_{0\leq\varepsilon\leq 2C_{\mathcal{M}}2^{-2j}} \varepsilon\sqrt{N\log\left(\frac{4C_{\mathcal{M}}+4}{\varepsilon}\right)} \\ & \leq 2C_{\mathcal{M}}2^{-2j}\sqrt{N\log\left(\left(4+\frac{4}{C_{\mathcal{M}}}\right)\cdot 2^{2j-1}\right)} \\ & \leq C_A\left(\frac{\sqrt{2j-1}}{2^j}\right)\left(\frac{\sqrt{N}}{2^j}\right) \leq C', \end{aligned}$$

where C' is an absolute constant. The computation in (5.12) finishes the proof. ■

It is no surprise that for general $\mathcal{M} \in \mathbb{S}^{N-1}$ the width bound for $w(\mathcal{M}_j)$ (resp. $w(\mathcal{M}_j^{rel})$) depends on either j or $\log(N)$. The proximity of $\mathcal{M}_j^{rel}$ to $\mathcal{M}$ in Lemma 5.6.4 only implies $\mathcal{M}_j^{rel} \subset \text{tube}_{C_{\mathcal{M}}2^{-2j}}$ and a large ambient dimension N will lead to a higher complexity of the tube. In Lemma 5.3.3 the proximity argument is omitted using the maximal number of affine d-dimensional spaces in $\mathcal{M}_j$. This argument does consequently not depend on N but on the refinement level j.
The next lemma requires more geometric structure by assuming that $\mathcal{M}$ is a Riemannian Manifold. It improves on both Lemma 5.3.3 and 5.6.4 for such $\mathcal{M}$ by yielding a width bound which is independent of both j and N for all j sufficiently large.

Lemma 5.6.5 (A Bound of the Gaussian Width for Approximations to Riemannian Manifolds)**.** *Let $\mathcal{M} \subset \mathbb{S}^{N-1}$ be a compact d-dimensional Riemannian manifold with d-dimensional volume* $\text{Vol}_d(\mathcal{M})$ *where $d \geq 1$. Suppose that $j > \max\{j_0, \log_2(8C_{\mathcal{M}}/C_1)\}$ for*

$C_{\mathcal{M}} := \max\{1, \sup_{\mathbf{z}\in\mathcal{M}} C_{\mathbf{z}}\}$. *Then, there exist absolute constants* $C, c > 0$ *such that*

$$\begin{aligned}\max\{w(\mathcal{M}), w(\mathcal{M}_j^{rel})\} &\leq w(\mathcal{M} \cup \mathcal{M}_j^{rel})\\ &\leq C\sqrt{d\left(1+\log\left(c\frac{\sqrt{d}}{\operatorname{reach}(\mathcal{M})}\right)\right)+\log(\max\{1, \operatorname{Vol}_d(\mathcal{M})\})}.\end{aligned}$$

Here the constants C_z *and* C_1 *come from GMRA properties* (3b) *and* (3a).

Proof: Let $2C_{\mathcal{M}}2^{-2j} \leq \tilde{\varepsilon} \leq \frac{1}{4}C_1 2^{-j}$. We aim to bound $w(\mathcal{M}_j^{rel})$ in terms of covering numbers for $\mathcal{M}$. To do this we will use Dudley's inequality in combination with the knowledge that $\mathcal{M}_j^{rel} \subset \mathcal{B}_2(\mathbf{0}, 2)$ (by definition). By Dudley's inequality (Theorem 3.2.8)

$$\begin{aligned}w(\mathcal{M}_j^{rel}) &\leq C' \int_0^4 \sqrt{\log(N(\mathcal{M}_j^{rel}, \varepsilon))}\ \mathrm{d}\varepsilon\\ &\leq C'\left(\int_0^{\tilde{\varepsilon}} \sqrt{\log(N(\mathcal{M}_j^{rel}, \varepsilon))}\ \mathrm{d}\varepsilon + \int_{\tilde{\varepsilon}}^4 \sqrt{\log(N(\mathcal{M}_j^{rel}, \varepsilon))}\ \mathrm{d}\varepsilon\right),\end{aligned}$$

where C' is an absolute constant. Appealing to Lemmas 5.6.3 and 5.6.2 for the first and second terms above, we can see that

$$\begin{aligned}w(\mathcal{M}_j^{rel}) &\leq C'\left(\int_0^{\tilde{\varepsilon}} \sqrt{\log((6/\varepsilon)^d N(\mathcal{M}, \varepsilon))}\ \mathrm{d}\varepsilon + \int_{\tilde{\varepsilon}}^4 \sqrt{\log(N(\mathcal{M}, \varepsilon/2))}\ \mathrm{d}\varepsilon\right)\\ &\leq C' \int_0^4 \sqrt{\log((6/\varepsilon)^d N(\mathcal{M}, \varepsilon/2))}\ \mathrm{d}\varepsilon\\ &= 2C' \int_0^2 \sqrt{d\log(3/\eta) + \log(N(\mathcal{M}, \eta))}\ \mathrm{d}\eta\\ &\leq 2C'\sqrt{\int_0^2 d\log(3/\eta)\ \mathrm{d}\eta + \int_0^2 \log(N(\mathcal{M}, \eta))\ \mathrm{d}\eta},\end{aligned}$$

where the last step follows from the Cauchy-Schwarz inequality. We may bound the second term as in the proof of Theorem 5.3.8. Doing so we obtain

$$\begin{aligned}w(\mathcal{M}_j^{rel}) &\leq C''\sqrt{\int_0^2 d\log(3/\eta)\ \mathrm{d}\eta + d\left(1+\log\left(c'\frac{\sqrt{d}}{\tau}\right)\right) + \log(\operatorname{Vol}_d(\mathcal{M}))}\\ &\leq C'''\sqrt{d\left(1+\log\left(c'\frac{\sqrt{d}}{\tau}\right)\right) + \log(\operatorname{Vol}_d(\mathcal{M}))},\end{aligned}$$

where τ is the reach of $\mathcal{M}$, and C''', c' are an absolute constants. The computation in (5.12) together with Theorem 5.3.8 concludes the proof. ■

As the above lemmas only apply to $\mathcal{M}_j^{rel}$, using these alternatives makes some modifications in the proof of Theorem 5.3.6 necessary:

In **(I)**, e.g., one has to guarantee that $\mathcal{C}_j \subset \mathcal{M}_j^{rel}$, i.e., that each center $\mathbf{c}_{j,k}$ is a best approximation for some part of the manifold. This assumption is reasonable if the centers are constructed as means of small manifold patches, a common approach in empirical applications (cf. Section 5.1.3).
When working with $\mathcal{M}_j^{rel}$ one has to guarantee in **(II)** that k' obtained in **(I)** fulfills $k' \approx k_j(\mathbf{x})$ as $\mathcal{M}_j^{rel}$ does not include many near-optimal centers for each point on $\mathcal{M}$. In addition, the optimization which has to be performed in this case in step II. of OMS becomes more involved. We will leave such variants to future work.

5.7 Discussion

In this chapter we proposed a tractable algorithm to approximate data lying on low-dimensional manifolds from compressive one-bit measurements to complement the theoretical results of Plan and Vershynin on one-bit sensing for general sets in [139]. We linked the theoretical understanding of one-bit measurements as tessellations of the sphere [140] to the GMRA techniques introduced in [5] and analyzed the interplay between a given manifold and its GMRA approximation's complexity measured in terms of the Gaussian mean width. To illustrate applicability of our results we showed that they even hold for manifolds learned from samples. Several interesting questions remain for future research:

First, it is straight-forward to generalize and improve the theory by replacing Theorem 3.3.5 and Theorem 3.3.9 with the quite recent results Theorem 3.3.8 and Theorem 3.3.10. In Theorem 5.3.1 this would allow $\mathcal{M}$ to be not on the sphere and reduce the sufficient number of measurements to obtain a squared error of $\varepsilon > 0$ from $\mathcal{O}(\varepsilon^{-6})$ to $\mathcal{O}(\varepsilon^{-2})$. As the dithered one-bit model (which has to be used then) causes additional notational complexity, we defer those considerations to the near future.

Second, the experiments in Section 5.5.4 suggest the use of tree structure within $\mathcal{C}_j$ to reduce computational complexity. Indeed OMS does still yield comparable results if I. is replaced by a tree based search which has the advantage of being computable much faster than the minimization over all possible centers. Obtaining theoretical error bounds in this case would be desirable, as well as considering the use of other related fast nearest neighbor methods from computer science [87].

Third, the reader might have noticed in the empirical setting of GMRA and in Section 5.4, that **(A2)** in combination with Lemma 5.4.5 seemingly renders II. of OMS useless. As Section 5.5.3 though shows, the second step of OMS yields a notable improvement even with an empirically constructed GMRA. Seemingly, even with **(A2)** not strictly fulfilled the empirical GMRA techniques remain valid, and II. of OMS of value. Understanding this phenomenon should lead to more relaxed assumptions than **(A1)**-**(A4)**.

Fourth, one might consider versions of OMS for additional empirical GMRA variants including those which rely on adaptive constructions [117], GMRA constructions in which subspaces that minimize different criteria are used to approximate the data in each partition element (see, e.g., [88]), and distributed GMRA constructions which are built up across networks using distributed clustering [11] and SVD [90] algorithms. Such variants could reduce the overall computational storage and/or runtime requirements of OMS in different practical situations.

Finally, as already pointed out in Section 5.5.2 we do not yet understand the influence of an inserted positive part $[\cdot]_+$ in II.. There seem to be cases in which a massive improvement can be observed and others in which the performance completely deteriorates. The explanation is probably decoupled from this work and OMS.

Chapter 6

Interlude: From Multi-Bit Recovery to Matrix Sensing

In this short chapter we discuss numerical experiments motivating our study of multi-bit recovery. We then point out that quantized compressed sensing and classification in machine learning are two sides of the same coin. Building upon this relation we describe how our work on matrix sensing (presented in Chapter 7) originated from the multi-bit recovery problem described in Section 3.4. This chapter contains unpublished joint work with Lars Palzer.

6.1 The Influence of Bit-Depth

Figure 6.1 and Figure 6.2 show the outcomes of different numerical experiments which consist in recovering randomly drawn s-sparse unit-norm signals $\mathbf{x} \in \mathbb{R}^N$, for $N = 100$, from uniformly quantized Gaussian measurements. We use for recovery multi-bit basis pursuit (cf. Section 3.4), which we describe in detail in the next section.

Comparing approximation quality in mean-squared and worst case error shows an exponential decrease of both errors in bit-depth, i.e. number of bits per measurement. This observation matches the theoretical performance of multi-bit compressed sensing predicted in Theorem 3.4.2. Figure 6.1 depicts for different quantization levels the mean squared error $\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2$ in dB when recovering $\mathbf{x}$ from m measurements (n dB $= 10^{\frac{-n}{10}}$, i.e., 0 dB $= 1$, -10 dB $= 0.1$, -20 dB $= 0.01$, etc.). The sparsity s of $\mathbf{x}$ varies from 1 to 100 and m from 0 to 400. Note that the measurement rate is defined as m/N. The experiment illustrates two important facts. First, one-bit quantization only works well under very strong sparsity assumptions. Second, the different multi-bit quantizer improve approximation quality step-by-step and lead to an almost smooth connection between one-bit compressed sensing and classical un-quantized compressed sensing.

In the second experiment, we are interested in the trade-off between number of measurements and bit-depth per measurement. We let the sparsity s vary from 0 to 100 and the number of available bits R (measured in Bits/N) from 0 to 400. The eight diagrams in Figure 6.2 depict empirical probabilities of obtaining a mean squared error $\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 < 0.1$ for different choices of B bits per measurement (note that $m = R/B$). The results suggest

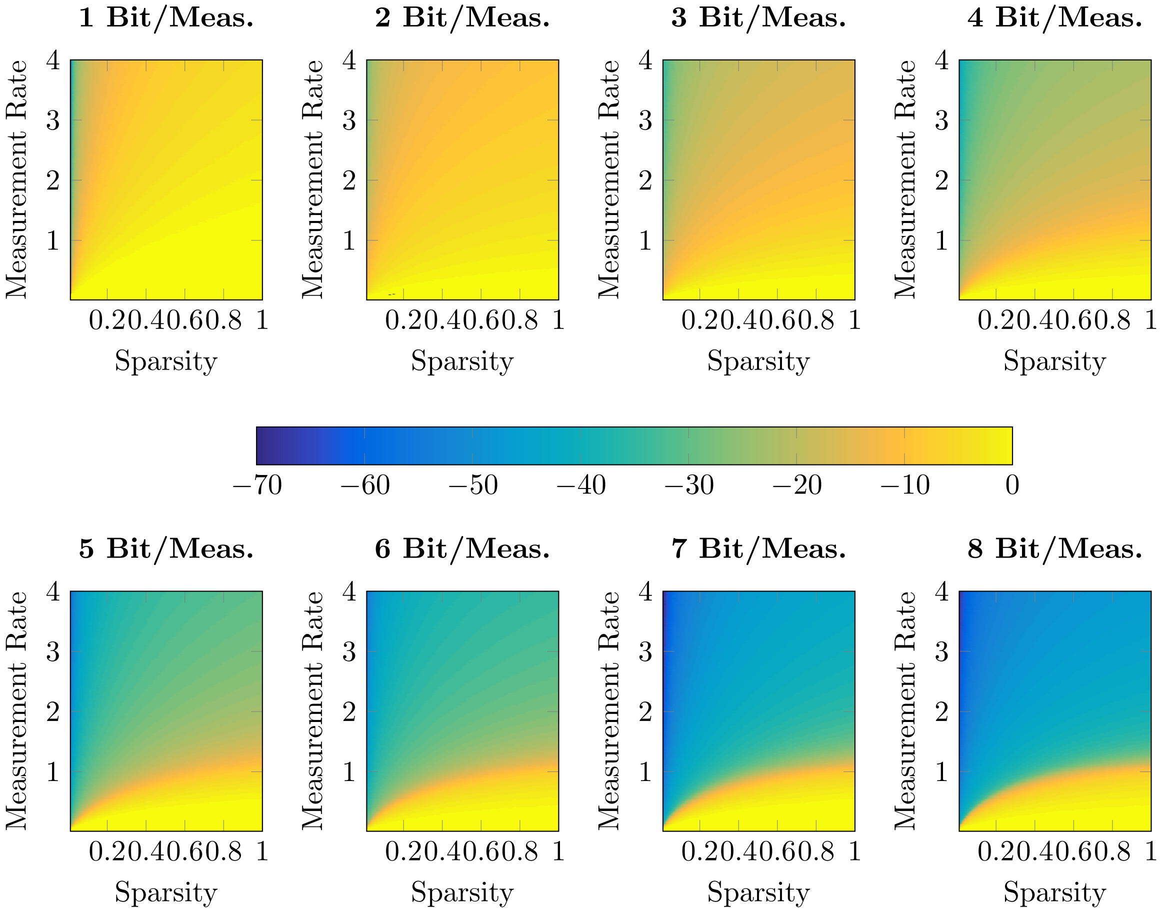


Figure 6.1: Median MSE over 500 experiments.

to choose a certain number of bits per measurement to maximize the region of successful recovery if some given approximation error shall be attained. Both experiments showed the qualitatively same behavior in worst case error.

In [112] the authors examine the trade-off between number of measurements and bit-depth per measurement when considering a fixed bit budget and draw the conclusion that an optimal choice depends on the concrete application. If high noise is expected, one should reduce the number of bits per measurement to have more measurements and hence more robust recovery. If the expected noise level is small it pays off to increase the number of bits per measurement to improve approximation quality. They observe as well that for obtaining a certain approximation error one should neither choose too few nor too many bits per measurement, the former only yielding rough approximation while the latter reducing the number of measurement below the lower bound necessary to apply compressed sensing techniques (cf. Figure 6.2).
In order to profit from bit-depth in the low-noise regime, it is crucial to have reliable reconstruction methods which exploit the improved accuracy best possible. When compared to

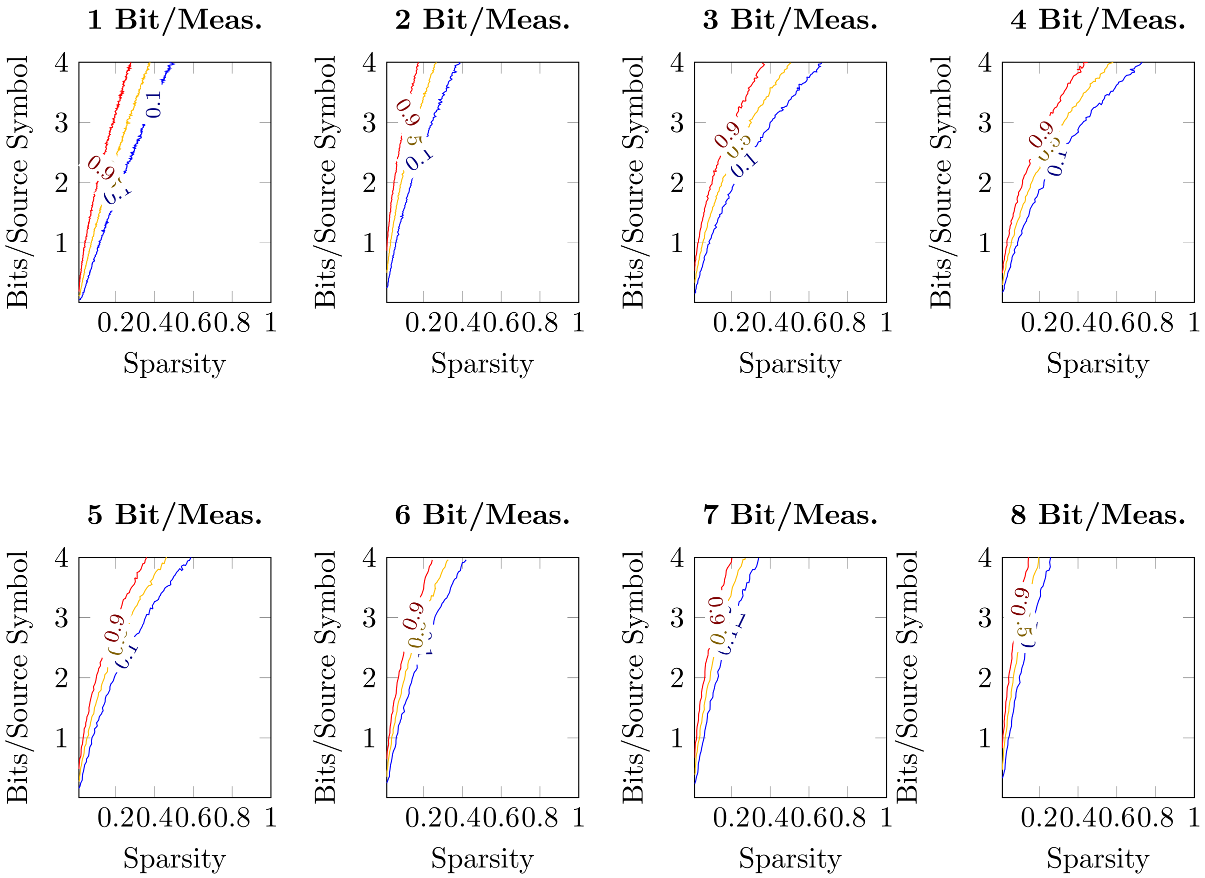
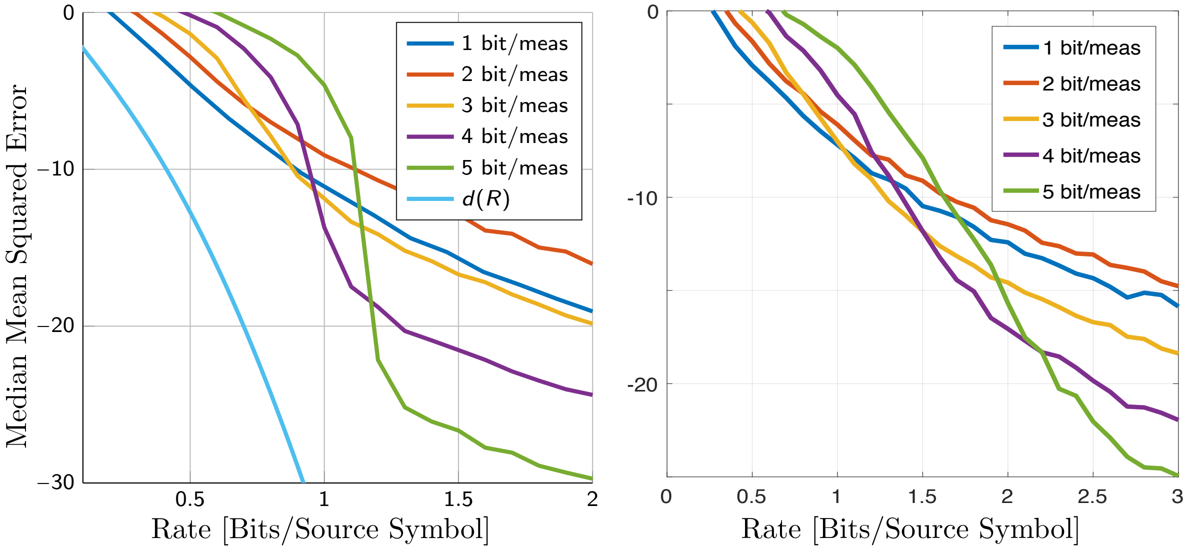


Figure 6.2: Empirical probability of MSE< −10 dB over 500 experiments.



(a) AMP (b) Multi-Bit BP

Figure 6.3: Comparison of AMP and Multi-Bit BP

approximate message passing algorithms (a class of more sophisticated algorithms we do not explain in detail here), the multi-bit basis pursuit performs poorly, see Figure 6.3. Is it possible to explain this performance gap or to find a simple alternative to basis pursuit coming with better worst-case approximation guarantees?

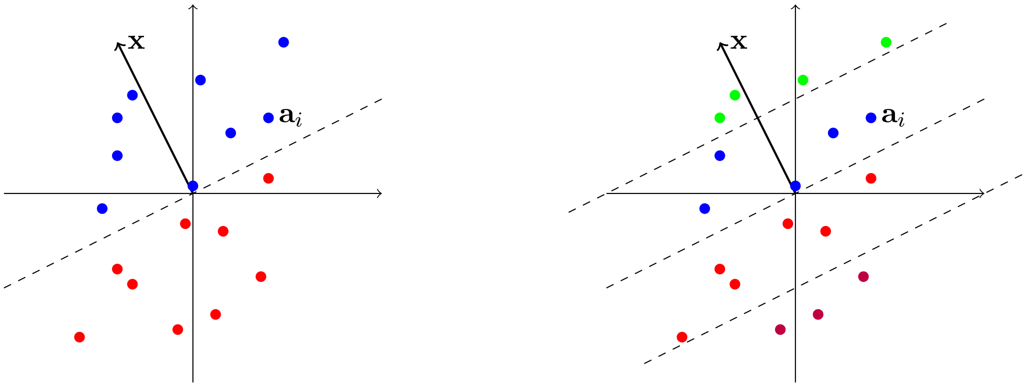
6.2 The Relation Between Quantized Compressed Sensing and Machine Learning

Recall from Sections 3.3 and 3.4 how we imagined one-bit and multi-bit quantization. Each quantized measurement characterizes one or several hyperplanes such that the collection of all measurements provides a tessellation of the space into quantization cells. Let us change our perspective. For simplicity, we concentrate on the one-bit setting first. Instead of identifying the measurement vectors $\mathbf{a}_i$ (rows of $\mathbf{A}$) with hyperplanes, we might see them as points in space which are labeled by $y_i = \operatorname{sign}(\langle \mathbf{a}_i, \mathbf{x} \rangle) \in \{-1, +1\}$ as depicted in Figure 6.4 (a).

We are now in the setting of binary classification problems where the unknown classifier is the hyperplane $H_{\mathbf{x}} = \{\mathbf{z} \in \mathbb{R}^N : \langle \mathbf{z}, \mathbf{x} \rangle = 0\}$ defined by $\mathbf{x}$. It is well-known that linear binary classification problems can be solved by *support vector machines* (SVM) [154] which search for a hyperplane separating the points according to their labels and maximize the margin between hyperplane and point clouds by computing

$$\mathbf{w} = \operatorname*{arg\,min}_{\hat{\mathbf{w}} \in \mathbb{R}^N, \tau \in \mathbb{R}} \|\hat{\mathbf{w}}\|_2^2, \quad \text{subject to } y_i \left(\langle \hat{\mathbf{w}}, \mathbf{a}_i \rangle - \tau \right) \geq 1, \text{ for all } i \in [m]. \tag{6.1}$$

Note that τ controls offset of the hyperplane $H_{\mathbf{w}}$ while the right-hand side of the constraints



(a) One-bit measurements

(b) Multi-bit measurements

Figure 6.4: Quantized compressed sensing as classification problem.

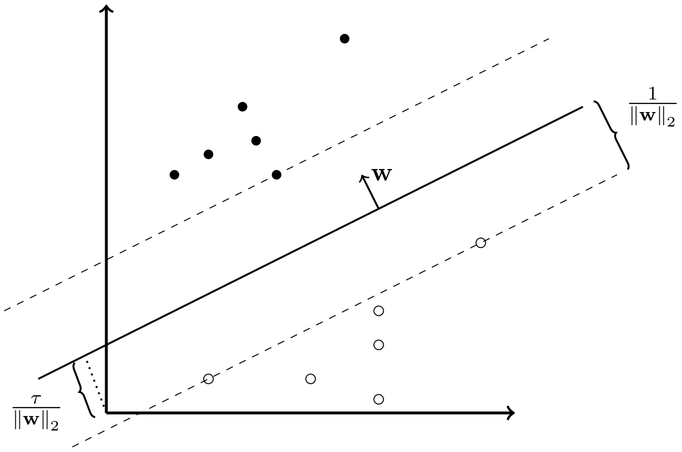


Figure 6.5: Separation of two point clouds with maximal margin.

causes the margin (cf. Figure 6.5). If in addition sparsity of the wanted support vector is assumed, one can replace the squared ℓ_2-norm in (6.1) by $\|\cdot\|_1$, a setting which has been studied in [105]. Comparing Figure 6.4 (a) to Figure 6.5 it is clear that τ and the right-hand side of the constraints in (6.1) are zero when one considers quantized compressed sensing as classification problem. This leads from (6.1) to

$$\mathbf{w} = \arg\min_{\hat{\mathbf{w}} \in \mathbb{R}^N} \|\hat{\mathbf{w}}\|_1, \quad \text{subject to } y_i\langle \hat{\mathbf{w}}, \mathbf{a}_i\rangle \geq 0, \text{ for all } i \in [m] \tag{6.2}$$

which is equivalent to (3.11) (up to the additional constraint $\|\mathbf{Az}\|_1 = m$).

Let us consider the multi-bit quantization depicted in Figure 6.4 (b). Instead of having two classes of points, we get depending on the number of bits per measurement $B \in \mathbb{N}$, 2^B classes we can distinguish. Moreover, we know that there exists a bundle of parallel hyperplanes which separate the points and share $\mathbf{x}$ as normal vector. If we write I_b for the set of $\mathbf{a}_i$ which belong to class $b \in [2^B]$ and introduce the auxiliary variables

$$z_i^b := \begin{cases} -1 & \text{if } \mathbf{a}_i \in I_{\tilde{b}}, \text{ for } \tilde{b} \leq b, \\ 1 & \text{else,} \end{cases}$$

(6.2) can be generalized to

$$\mathbf{w} = \arg\min_{\hat{\mathbf{w}} \in \mathbb{R}^N} \|\hat{\mathbf{w}}\|_1, \quad \text{subject to } \begin{cases} z_i^b \left(\langle \hat{\mathbf{w}}, \mathbf{a}_i \rangle - \tau(\Delta, b, \|\mathbf{x}\|_2) \right) \geq 0, \\ \text{for all } i \in [m], b \in [2^B] \end{cases} \tag{6.3}$$

where the shifts $\tau(\Delta, b, \|\mathbf{x}\|_2)$ depend on the quantizers refinement level Δ and the norm of $\mathbf{x}$ (if those parameters are unknown, one may also optimize over τ as in (6.1)). To put it simply, (6.3) minimizes the ℓ_1-norm with respect to the quantized measurements and is thus equivalent to multi-bit basis pursuit [128].

Remark 6.2.1. *The just observed relation can be used to transfer results on quantized compressed sensing in Section 3.3 and 3.4 to the framework of classification problems in machine learning. Theorem 3.3.1, for example, characterizes how many Gaussian samples are necessary to solve a binary classification problem up-to expected misclassification error $\mathcal{O}(\varepsilon)$ in $\mathbb{R}^N$ when the classifier is linear and orthogonal to some sparse vector while Theorem 3.3.9 additionally respects misclassifications on the training data. These are valuable results if one assumes that affiliation with a class only depends on few relevant parameters (corresponding to sparsity of the classifier's normal vector).*
Interpreting the dithered results Theorem 3.3.10 and 3.4.2 is more involved. Here the dither adds a specific random perturbation to each of the $\mathbf{a}_i$ which might be hard to justify if the $\mathbf{a}_i$ resemble empirical data.

Interpreting quantized compressed sensing as classification problem clearly shows that only a small fraction of the measurements is used in obtaining the estimate of $\mathbf{x}$. As support vector machines mainly rely on the points closest to the classifying hyperplane (called support vectors), only those $\mathbf{a}_i$ which lie close to $H_\mathbf{x}$ determine the approximation quality of (6.2) and (6.3).
But Figure 6.4 (b) illustrates an interesting geometric property of $\mathbf{x}$. If one centers the point clouds I_b around the origin and writes the resulting points into a matrix, one of the singular vectors will be close to $\mathbf{x}$, a sparse vector. Consequently, it might be possible to improve on the support vector machine ansatz by efficiently calculating singular value decompositions of matrices which have sparse singular vectors.
Unfortunately, numerical simulations show that this strategy does not yield practical results. The problem is that sparsity of $\mathbf{x}$ implies only sparsity on one singular vector and that this specific singular vector belongs to the smallest singular value which is most sensitive to noise. However, the algorithm we have examined performs well in recovery of low-rank matrices with sparse singular vectors from unquantized compressed sensing measurements. Therefore, let us leave in the following chapter the framework of quantized compressed sensing and turn to matrix sensing.

Chapter 7

ATLAS: Matrix Sensing with Combined Structures

In the last chapter we discuss recovery of matrices with multiple structures, in particular low-rankness and sparsity, from unquantized compressed sensing measurements. We present recent work on this topic including sparse power factorization, the so far stand-alone algorithm for recovery of matrices which are low-rank and sparse. After proposing a new algorithm, A-T-LAS$_{2,1}$, we analyze its behavior, compare it numerically to sparse power factorization, and discuss its potential generalizations. The content of the chapter is joint work with Massimo Fornasier and Valeriya Naumova, and has been published in [63] and [64].

7.1 Matrix Sensing and Multiple Structures

Starting with [29, 144] the theory of compressed sensing has been generalized to matrix signals. Motivated by applications as low-dimensional embedding of data into Euclidean space [120] and low-order realizations of linear systems [59], one is interested in finding a matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ of minimal rank fulfilling

$$\mathbf{y} = \mathcal{A}(\mathbf{X}), \tag{7.1}$$

where the linear measurement operator $\mathcal{A}\colon \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ and the measurements $\mathbf{y} \in \mathbb{R}^m$ are given. Just like ℓ_0-minimization (2.2), solving

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}} \operatorname{rank}(\mathbf{Z}), \quad \text{subject to } \mathcal{A}(\mathbf{Z}) = \mathbf{y}, \tag{7.2}$$

is NP-hard in general. In order to solve (7.1), one has to replace (7.2) with a tractable program. By singular value decomposition we can write $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ where $\mathbf{U} \in \mathbb{R}^{n_1 \times n_1}, \mathbf{V} \in \mathbb{R}^{n_2 \times n_2}$ are orthogonal matrices and $\mathbf{\Sigma} \in \mathbb{R}^{n_1 \times n_2}$ is diagonal. Let us denote by $\boldsymbol{\sigma} \in \mathbb{R}^{\min\{n_1,n_2\}}$ the vector of singular values $\sigma_1 \geq ... \geq \sigma_{\min\{n_1,n_2\}} \geq 0$. Based on the observation that $\operatorname{rank}(\mathbf{Z}) = \|\boldsymbol{\sigma}\|_0$, the idea of [144] is to relax (7.2) to

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}} \|\mathbf{Z}\|_*, \quad \text{subject to } \mathcal{A}(\mathbf{Z}) = \mathbf{y}, \tag{7.3}$$

where $\|\mathbf{Z}\|_* = \|\boldsymbol{\sigma}\|_1$ denotes the so-called *nuclear norm.* When transferring the concepts presented in Chapter 2 to this setting, one can show that certain random operators $\mathcal{A}$ satisfy with high probability a restricted isometry property on the set of $n_1 \times n_2$ rank-R matrices for

$$m \gtrsim R \max\{n_1, n_2\}. \tag{7.4}$$

Consequently, any low-rank matrix $\mathbf{X}$ can be uniquely recovered from (7.1) by the convex program (7.3). Moreover, the bound in (7.4) corresponds up-to a constant to the number of parameters which are necessary to describe an $n_1 \times n_2$ rank-R matrix and thus scales optimally in the intrinsic dimension of the set of low-rank matrices. As in Chapter 2, the RIP allows to obtain robust and stable recovery results.

Things become interesting as soon as sparsity is considered in addition. By viewing $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ as vector in $\mathbb{R}^{n_1 n_2}$, the classical compressed sensing machinery can be applied as well. If $\mathbf{X}$, for example, has only $s_1 \ll n_1$ non-zero rows and $s_2 \ll n_2$ non-zero columns, the results in Chapter 2 guarantee that $\mathbf{X}$ can be recovered from

$$m \gtrsim (s_1 s_2) \log \left(\frac{e(n_1 n_2)}{(s_1 s_2)} \right) \tag{7.5}$$

measurements of type (7.1) in a stable and robust way. But is it possible to profit of both structures, low-rankness and sparsity, at once if $\mathbf{X}$ is (s_1, s_2)-column/row-sparse and has rank $R \ll \min\{s_1, s_2\}$? To describe a matrix of this type, less than $R(s_1 + s_2)$ parameters are sufficient which can be considerably less than (7.4) and (7.5).

One might wonder if the assumption of combined structures is useful. A practical motivation is given by *blind deconvolution* [84]. In blind deconvolution one is interested in recovering two unknown vectors $\mathbf{w}$ and $\mathbf{p}$ solely from their convolutional product

$$\mathbf{y} = \mathbf{w} * \mathbf{p} = \left(\sum_{i=1}^{m} w_i p_{(k-i) \bmod m} \right)_{k=1}^{m} \tag{7.6}$$

In imaging applications $\mathbf{p}$ represents the picture and $\mathbf{w}$ an unknown blurring kernel [155]. In signal transmission $\mathbf{p}$ is a coded message and $\mathbf{w}$ models the properties of the transmission channel [71]. Independent of the concrete application (7.6) is highly underdetermined and contains ambiguities which might be reduced and handled by sparsity and low-rankness assumptions.

Let us focus here on the transmission of signals to elaborate on the last statement. In [3] the authors used that, by bilinearity of the convolution, (7.6) can be represented as a linear map acting on the outer product $\mathbf{w}\mathbf{p}^T$, a technique commonly known as *lifting.* They assumed in addition that the channel properties $\mathbf{w}$ and the message $\mathbf{p}$ lie in lower dimensional subspaces and are of the form $\mathbf{w} = \mathbf{B}\mathbf{h}$ and $\mathbf{p} = \mathbf{C}\mathbf{x}$ with $\mathbf{h} \in \mathbb{C}^{n_1}$ and $\mathbf{x} \in \mathbb{C}^{n_2}$ being coefficient vectors encoding channel and message ($\mathbf{B}$ and $\mathbf{C}$ are suitable transformation matrices). They could hence re-write (7.6) as

$$\mathbf{y} = \mathcal{A}(\mathbf{h}\mathbf{x}^*)$$

Algorithm 7 : SPF$(\mathbf{y}, \mathcal{A}, \mathbf{V}_0, s_1, s_2, R, L)$

```
Require: $\mathcal{A} \colon \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{V}_0 \in \mathbb{R}^{n_2 \times R}$, $s_1, s_2, R \in \mathbb{N}$, number of iterations $L$
$l = 0$                                                        ▷ initialize

while $l < L$ do
    $\mathbf{V}_l \leftarrow \mathrm{orth}([\mathbf{V}_l, \mathbf{V}_0])$                          ▷ orthonormalize
    if $s_1 < n_1$ then
        $\tilde{\mathbf{U}} = \mathbf{HTP}(\mathcal{A}_{\mathbf{V}_l}, \mathbf{y}, s_1)$                  ▷ fix $\mathbf{V}_l$ and apply HTP
    else
        $\tilde{\mathbf{U}} = \arg\min_{\mathbf{U} \in \mathbb{R}^{n_1 \times R}} \|\mathbf{y} - \mathcal{A}_{\mathbf{V}_l}(\mathbf{U})\|_2^2$     ▷ fix $\mathbf{V}_l$ and solve least squares
    end if
    Let $\mathbf{U}_{l+1}$ be the best rank-$R$ approximation of $\tilde{\mathbf{U}}$

    $\mathbf{U}_{l+1} \leftarrow \mathrm{orth}([\mathbf{U}_{l+1}, \mathbf{U}_0])$                      ▷ orthonormalize
    if $s_2 < n_2$ then
        $\tilde{\mathbf{V}} = \mathbf{HTP}(\mathcal{A}_{\mathbf{U}_{l+1}}, \mathbf{y}, s_2)$               ▷ fix $\mathbf{U}_{l+1}$ and apply HTP
    else
        $\tilde{\mathbf{V}} = \arg\min_{\mathbf{V} \in \mathbb{R}^{n_2 \times R}} \left\|\mathbf{y} - \mathcal{A}_{\mathbf{U}_{l+1}}(\mathbf{V})\right\|_2^2$  ▷ fix $\mathbf{U}_{l+1}$ and solve least squares
    end if
    Let $\mathbf{V}_{l+1}$ be the best rank-$R$ approximation of $\tilde{\mathbf{V}}$

    $l = l + 1$
end while

return $\mathbf{X}_{\mathrm{SPF}} = \mathbf{U}_l \mathbf{V}_l^T$
```

where a rank-1 matrix $\mathbf{hx}^* \in \mathbb{C}^{n_1 \times n_2}$ has to be recovered from m linear measurements. As explained above, this can be solved by nuclear norm minimization under suitable assumptions on $\mathbf{A}$.

In *blind demixing* [118, 119, 102] or MIMO channel identification [45] a receiver gets the overlay of R different convolutions which translates in the lifted formulation into the recovery of rank-R matrices from linear measurements of type

$$\mathbf{y} = \mathcal{A}\left(\sum_{r=1}^{R} \mathbf{h}_r \mathbf{x}_r^*\right).$$

As already mentioned in [102], one can typically impose extra structure like sparsity on the channel impulse responses h to further reduce the number of measurements m. In this case one hopes to profit from low-rankness and sparsity at the same time.

A first throwback is an observation of Oymak et. al. in [135]: the mere convex combination of regularizers for different sparsity structures does not allow in general outperforming the recovery guarantees of the "best" one of them alone. Consequently, in order to improve recovery further, one has to go beyond linear combinations of already known convex

Algorithm 8 : HTP$(\mathbf{y}, \Phi, s, L)$

```
Require: $\Phi\colon \mathbb{R}^{n\times R} \to \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^m$, $s \in \mathbb{N}$, number of iterations $L$
$l = 0$, $\mathbf{Z}_0 = \mathbf{0} \in \mathbb{R}^{n\times R}$                                ▷ initialize

while $l < L$ do
    $\tilde{\mathbf{Z}} = \mathbf{Z}_l + \Phi^*(\mathbf{y} - \Phi(\mathbf{Z}_l))$                        ▷ gradient step
    Let $J \subset [n]$ be the index set of the $s$ rows of $\tilde{\mathbf{Z}}$
with largest $\ell_2$-norm and $\Pi_J$ the projection onto the row set.
    $\mathbf{Z}_{l+1} = \arg\min_{\mathbf{Z}\colon \Pi_J\mathbf{Z}=\mathbf{Z}} \|\mathbf{y} - \Phi(\mathbf{Z})\|_2^2$          ▷ least square fit
    $l = l + 1$
end while

return $\mathbf{Z}_{\text{HTP}} = \mathbf{Z}_l$
```

regularizers.

In [9] the authors overcome the aforementioned limitations of purely convex approaches by assuming a nested structure of the measurement operator $\mathcal{A}$ and applying basic solvers for low-rank resp. row-sparse recovery in two consecutive steps which is an elegant approach but clearly restricts possible choices for $\mathcal{A}$.

Lee et. al. in contrast proposed and analyzed in [116] the so-called *sparse power factorization* (SPF) to recover $\mathbf{X}$ from noisy measurements

$$\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\eta} = \begin{pmatrix} \langle \mathbf{A}_1, \mathbf{X}\rangle_F \\ \vdots \\ \langle \mathbf{A}_m, \mathbf{X}\rangle_F \end{pmatrix} + \boldsymbol{\eta}$$

where $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius scalar product, $\boldsymbol{\eta} \in \mathbb{R}^m$ models additive noise, and the matrices $\mathbf{A}_i \in \mathbb{R}^{n_1\times n_2}$, for $i \in [m]$, characterize the operator $\mathcal{A}$. SPF is a modified version of Power Factorization (see [99]) which recovers low-rank matrices by representing them as product of two matrices $\mathbf{X} = \mathbf{U}\mathbf{V}^T$ and then applying alternating minimization over the (de)composing matrix $\mathbf{U}, \mathbf{V}$. In addition to power factorization, SPF introduces *hard-thresholding pursuit* (HTP) to each of the alternating steps to enforce additional sparsity of the columns of $\mathbf{U}$ and/or $\mathbf{V}$. HTP is closely related to iterative hard-thresholding and presented in Algorithm 8. Be aware in Algorithm 7 that orth$(\mathbf{Z})$ computes an orthonormal basis of span$(\mathbf{Z})$ and that

$$\mathcal{A}_{\mathbf{V}}(\mathbf{U}) := \begin{pmatrix} \langle \mathbf{A}_1\mathbf{V}, \mathbf{U}\rangle_F \\ \vdots \\ \langle \mathbf{A}_m\mathbf{V}, \mathbf{U}\rangle_F \end{pmatrix} \quad \text{and} \quad \mathcal{A}_{\mathbf{U}}(\mathbf{V}) := \begin{pmatrix} \langle \mathbf{A}_1^T\mathbf{U}, \mathbf{V}\rangle_F \\ \vdots \\ \langle \mathbf{A}_m^T\mathbf{U}, \mathbf{V}\rangle_F \end{pmatrix}$$

fulfill $\mathcal{A}(\mathbf{U}\mathbf{V}^T) = \mathcal{A}_{\mathbf{V}}(\mathbf{U}) = \mathcal{A}_{\mathbf{U}}(\mathbf{V})$. To analyze SPF, Lee et. al. introduce the following RIP.

Definition 7.1.1 (Rank-R and (s_1, s_2)-jointly-sparse RIP)**.** *A linear operator $\mathcal{A}\colon \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ satisfies the rank-R and (s_1, s_2)-jointly-sparse RIP with isometry constant $\delta \in (0,1)$ if*

$$(1-\delta)\|\mathbf{Z}\|_F^2 \le \|\mathcal{A}(\mathbf{Z})\|_2^2 \le (1+\delta)\|\mathbf{Z}\|_F^2, \tag{7.7}$$

for all $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$ *such that* $\operatorname{rank}(\mathbf{Z}) \leq R$, $\|\mathbf{Z}\|_{0,2} \leq s_1$, *and* $\|\mathbf{Z}^T\|_{0,2} \leq s_2$ *where* $\|\cdot\|_{0,2}$ *denotes, by abuse of notation, the number of nonzero rows of a matrix.*

They show that Gaussian measurement operators fulfill above RIP with high probability if the number of measurements satisfies

$$m \gtrsim R(s_1 + s_2) \log\left(\max\left\{\frac{en_1}{s_1}, \frac{en_2}{s_2}\right\}\right) \tag{7.8}$$

and that up to the log-factor this is at the information theoretical limit. Based on the RIP they give a recovery guarantee for SPF.

Theorem 7.1.2 ([116, Theorem 7])**.** *Suppose the following:*

(i) $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ *denotes the singular value decomposition of a rank-R matrix* $X \in \mathbb{R}^{n_1 \times n_2}$, *where* $\mathbf{U} \in \mathbb{R}^{n_1 \times R}$ *and* $\mathbf{V} \in \mathbb{R}^{n_2 \times R}$ *are row-s_1-sparse and row-s_2-sparse, respectively.*

(ii) The condition number of $\mathbf{X}$ *(restricted to the non-zero singular values) is not greater than* τ.

(iii) $\mathcal{A}$ *satisfies the rank-$2R$ and $(3s_1, 3s_2)$-jointly-sparse RIP with isometry constant* $\delta = 0.04/\tau$.

(iv) $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\eta}$ *where* $\boldsymbol{\eta}$ *and* $\mathcal{A}(\mathbf{X})$ *satisfy*

$$\frac{\|\mathbf{X}\|_F}{\|\mathbf{X}\|} \cdot \frac{\|\boldsymbol{\eta}\|_2}{\|\mathcal{A}(\mathbf{X})\|_2} \leq \nu$$

with $\nu = 0.04/\tau$.

(v) The initialization $(\mathbf{U}_0, \mathbf{V}_0)$ *satisfies*

$$\max\left\{\|\mathbb{P}_{\mathcal{R}(\mathbf{U})^\perp}\mathbb{P}_{\mathbb{R}(\mathbf{U}_0)}\|, \|\mathbb{P}_{\mathcal{R}(\mathbf{V})^\perp}\mathbb{P}_{\mathcal{R}(\mathbf{V}_0)}\|\right\} < 0.95$$

where $\mathbb{P}$ *denotes the orthogonal projection and* $\mathcal{R}(\mathbf{U})$ *the range of the columns of* $\mathbf{U}$.

Then the output $(\mathbf{X}_l)_{l\in\mathbb{N}}$ *of SPF satisfies*

$$\limsup_{l\to\infty} \frac{\|\mathbf{X}_l - \mathbf{X}\|_F}{\|\mathbf{X}\|_F} \leq (55\tau^2 + 3\tau + 3)\frac{\|\boldsymbol{\eta}\|_2}{\|\mathcal{A}(\mathbf{X})\|_2}.$$

Moreover, the convergence is linear, i.e. for any $\varepsilon > 0$, *there exists* $l_0 = \mathcal{O}(\log(1/\varepsilon))$ *that satisfies*

$$\frac{\|\mathbf{X}_{l_0} - \mathbf{X}\|_F}{\|\mathbf{X}\|_F} \leq (55\tau^2 + 3\tau + 3)\frac{\|\boldsymbol{\eta}\|_2}{\|\mathcal{A}(\mathbf{X})\|_2} + \varepsilon.$$

Remark 7.1.3. *Sparse power factorization has been introduced and analyzed in the more general complex setting. To conform with the rest of the thesis, we restrict it here to real valued matrices.*

Theorem 7.1.2 states that using suitable initialization and assuming the noise level to be sufficiently small, SPF approximates low-rank and row- and/or column-sparse matrices $\mathbf{X}$ from a nearly optimal number of measurements. If $\mathbf{X}$ is rank-R, has s_1-sparse columns and s_2-sparse rows, $m \gtrsim R(s_1 + s_2)\log(\max\{en_1/s_1, en_2/s_2\})$ measurements suffice which is up to the log-factor at the information theoretical bound. Note that all columns (resp. rows) need to share a common support in this setting. As it has been shown in [116] that SPF outperforms methods based on convex relaxation, we will use it as a benchmark for recovery. However, SPF and its analysis are heavily based on the assumption that the operator $\mathcal{A}$ possesses a suitable restricted isometry property and cannot be applied to arbitrary inverse problems of type (7.1). As we will see in the next sections, this shortcoming can be dealt with by considering soft-thresholding instead.

7.2 Alternating Tikhonov Regularization and LASSO

Let us first clarify our problem setting. We recall that the reduced singular value decomposition of a matrix $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$ is given by

$$\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{r=1}^{\operatorname{rank}(\mathbf{Z})} \sigma_r \mathbf{u}^r (\mathbf{v}^r)^T, \tag{7.9}$$

where $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values $\sigma_1 \geq ... \geq \sigma_{\operatorname{rank}(\mathbf{Z})} > 0$ while $\mathbf{U} \in \mathbb{R}^{n_1 \times \operatorname{rank}(\mathbf{Z})}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times \operatorname{rank}(\mathbf{Z})}$ have orthonormal columns which are called left and right singular vectors. Hence, each non-zero singular value σ_r has one left singular vector $\mathbf{u}^r$ and one right singular vector $\mathbf{v}^r$.

In contrast to [116] where the singular vectors of the signal have to be sparse and share a common support, we just assume the unknown signal $\hat{\mathbf{X}}$ of rank $R > 0$ to possess a decomposition of the form

$$\hat{\mathbf{X}} = \sum_{r=1}^{R} \hat{\mathbf{u}}^r (\hat{\mathbf{v}}^r)^T, \tag{7.10}$$

where the vectors $\hat{\mathbf{u}}^r$ and $\hat{\mathbf{v}}^r$ are effectively s_1/s_2-sparse (recall Definition 3.2.1). Note that we require $\hat{\mathbf{u}}^r$ and $\hat{\mathbf{v}}^r$ neither to be exactly sparse nor to share a common support. We call the vectors $\hat{\mathbf{u}}^r$ (resp. $\hat{\mathbf{v}}^r$) the left (resp. right) component vectors of $\hat{\mathbf{X}}$. From the context it will be clear to which decomposition they are referred. As we do not require orthogonality of the components, (7.10) does not need to be the SVD of $\hat{\mathbf{X}}$, although this case is also covered by our analysis. If $\hat{\mathbf{X}}$ of rank $R > 0$ possesses an SVD as in (7.10) for $\|\hat{\mathbf{v}}^r\|_2 = \sigma_r$, $r \in [R]$, then for any $0 < p < \infty$

$$\|\hat{\mathbf{X}}\|_p^p = \sum_{r=1}^{R} (\|\hat{\mathbf{u}}^r\|_2 \|\hat{\mathbf{v}}^r\|_2)^p. \tag{7.11}$$

If the decomposition (7.10) does not coincide with the SVD of $\hat{\mathbf{X}}$, then $\hat{\mathbf{u}}^1, \dots, \hat{\mathbf{u}}^R$ are anyhow linearly independent by $\operatorname{rank}(\hat{\mathbf{X}}) = R$ and assuming that the eigenvalues of the

Gramian of the vectors $\hat{\mathbf{u}}^1/\|\hat{\mathbf{u}}^1\|_2, \dots, \hat{\mathbf{u}}^R/\|\hat{\mathbf{u}}^R\|_2$ are in some positive bounded interval, one has

$$\begin{aligned}\|\hat{\mathbf{X}}\|_F^2 &= \sum_{j=1}^{n_2}\sum_{i=1}^{n_1}\left|\sum_{r=1}^{R}\hat{\mathbf{u}}_i^r\hat{\mathbf{v}}_j^r\right|^2 = \sum_{j=1}^{n_2}\left\|\sum_{r=1}^{R}\frac{\hat{\mathbf{u}}^r}{\|\hat{\mathbf{u}}^r\|_2}\|\hat{\mathbf{u}}^r\|_2\hat{\mathbf{v}}_j^r\right\|_2^2 \\ &\simeq \sum_{j=1}^{n_2}\sum_{r=1}^{R}\|\hat{\mathbf{u}}^r\|_2^2|\hat{\mathbf{v}}_j^r|^2 = \sum_{r=1}^{R}(\|\hat{\mathbf{u}}^r\|_2\|\hat{\mathbf{v}}^r\|_2)^2.\end{aligned}$$

From this equivalence and the equivalence of ℓ_p-quasi-norms and Schatten-p-quasi-norms for $0 < p \le 2$, one further obtains as a relaxation of (7.11)

$$c_{\hat{\mathbf{U}}}^{-1}R^{p/2-1}\sum_{r=1}^{R}(\|\hat{\mathbf{u}}^r\|_2\|\hat{\mathbf{v}}^r\|_2)^p \le \|\mathbf{X}\|_p^p \le C_{\hat{\mathbf{U}}}R^{1-p/2}\sum_{r=1}^{R}(\|\hat{\mathbf{u}}^r\|_2\|\hat{\mathbf{v}}^r\|_2)^p, \tag{7.12}$$

for positive constants $c_{\hat{\mathbf{U}}}, C_{\hat{\mathbf{U}}} > 0$, which depend on the largest and smallest eigenvalues of the Gramian of the vectors $\hat{\mathbf{u}}^1/\|\hat{\mathbf{u}}^1\|_2, \dots, \hat{\mathbf{u}}^R/\|\hat{\mathbf{u}}^R\|_2$. We use (7.12) mainly for $p = 2/3$ below.

Remark 7.2.1. *For simplicity, we focus in the following on decompositions* (7.10) *with effectively sparse right component vectors and arbitrary left component vectors. Conceptually straight-forward, but perhaps tedious modifications of the arguments lead to similar results in the left-sided and both-sided sparse case. We will comment on this whenever appropriate.*

As in the last section, we are given some linear measurement operator $\mathcal{A}\colon \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ and the vector of measurements $\mathbf{y} \in \mathbb{R}^m$, which is obtained from $\hat{\mathbf{X}}$ by

$$\mathbf{y} = \mathcal{A}(\hat{\mathbf{X}}) + \boldsymbol{\eta} = \frac{1}{\sqrt{m}}\begin{pmatrix}\langle \mathbf{A}_1, \hat{\mathbf{X}}\rangle_F \\ \vdots \\ \langle \mathbf{A}_m, \hat{\mathbf{X}}\rangle_F\end{pmatrix} + \boldsymbol{\eta}. \tag{7.13}$$

The operator $\mathcal{A}$ is completely characterized by the m matrices $\mathbf{A}_i \in \mathbb{R}^{n_1\times n_2}$ and individual measurements correspond to Frobenius products $\langle \mathbf{A}_i, \hat{\mathbf{X}}\rangle_F = \operatorname{trace}(\mathbf{A}_i\hat{\mathbf{X}}^T)$. Additive noise is modeled by $\boldsymbol{\eta} \in \mathbb{R}^m$ of which only the ℓ_2-norm is assumed to be known.

We propose to recover and decompose $\hat{X}$ by a variational approach and to minimize the following multi-penalty functional $J_{\alpha,\beta}^R\colon \mathbb{R}^{n_1} \times ... \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times ... \times \mathbb{R}^{n_2} \to \mathbb{R}$ which is defined, for $\alpha, \beta > 0$, by

$$J_{\alpha,\beta}^R(\mathbf{u}^1, \dots, \mathbf{u}^R, \mathbf{v}^1, \dots, \mathbf{v}^R) := \left\|\mathbf{y} - \mathcal{A}\left(\sum_{r=1}^{R}\mathbf{u}^r(\mathbf{v}^r)^T\right)\right\|_2^2 + \alpha\sum_{r=1}^{R}\|\mathbf{u}^r\|_2^2 + \beta\sum_{r=1}^{R}\|\mathbf{v}^r\|_1, \tag{7.14}$$

where α, β are regularization parameters. The functional in (7.14) is motivated by the success of multi-penalty regularization in recent works [131, 75, 40]. We denote a global minimizer of (7.14) by

$$(\mathbf{u}_{\alpha,\beta}^1, \dots, \mathbf{u}_{\alpha,\beta}^R, \mathbf{v}_{\alpha,\beta}^1, \dots, \mathbf{v}_{\alpha,\beta}^R).$$

Note that $J^R_{\alpha,\beta}$ also applies to matrices by viewing each $2R$-tuple $(\mathbf{u}^1, \dots, \mathbf{u}^R, \mathbf{v}^1, \dots, \mathbf{v}^R)$ as the matrix $\mathbf{X} = \sum_{r=1}^R \mathbf{u}^r(\mathbf{v}^r)^T$. Let us denote by

$$\mathbf{X}_{\alpha,\beta} = \sum_{r=1}^{R} \mathbf{u}^r_{\alpha,\beta}(\mathbf{v}^r_{\alpha,\beta})^T$$

the matrix corresponding to $(\mathbf{u}^1_{\alpha,\beta}, \dots, \mathbf{u}^R_{\alpha,\beta}, \mathbf{v}^1_{\alpha,\beta}, \dots, \mathbf{v}^R_{\alpha,\beta})$. The functional $J^R_{\alpha,\beta}$ has a restricted domain (the decomposition can only consist of R vector pairs) to enforce low-rankness of $\mathbf{X}_{\alpha,\beta}$ and uses a non-smooth term $\|\cdot\|_1$ to promote sparsity in right singular vectors of $\mathbf{X}_{\alpha,\beta}$. Despite the convex multi-penalty regularization term $\alpha \sum_{r=1}^R \|\mathbf{u}^r\|_2^2 + \beta \sum_{r=1}^R \|\mathbf{v}^r\|_1$, the functional (7.14) is highly non-convex, hence, it is not affected by the above mentioned negative results of Oymak et. al. [135].

We approach its minimization by using the following alternating algorithm based on simple iterative soft-thresholding, to which we refer as **A**lternating **T**ikhonov regularization and **Las**so (A-T-LAS$_{2,1}$). However, for simplicity and ease of notation we write ATLAS throughout the chapter.

$$(\text{A-T-LAS}_{2,1}) \begin{cases} \mathbf{u}^1_{k+1} = \arg\min_{\mathbf{u}} & \left\| \left(\mathbf{y} - \mathcal{A}\left(\sum_{r=2}^R \mathbf{u}^r_k \mathbf{v}^{r\,T}_k \right) \right) - \mathcal{A}(\mathbf{u}\mathbf{v}^{1\,T}_k) \right\|_2^2 \\ & +\alpha\|\mathbf{u}\|_2^2 + \frac{1}{2\lambda^1_k}\|\mathbf{u} - \mathbf{u}^1_k\|_2^2, \\ \mathbf{v}^1_{k+1} = \arg\min_{\mathbf{v}} & \left\| \left(\mathbf{y} - \mathcal{A}\left(\sum_{r=2}^R \mathbf{u}^r_k \mathbf{v}^{r\,T}_k \right) \right) - \mathcal{A}(\mathbf{u}^1_{k+1}\mathbf{v}^T) \right\|^2 \\ & +\beta\|\mathbf{v}\|_1 + \frac{1}{2\mu^1_k}\|\mathbf{v} - \mathbf{v}^1_k\|_2^2, \\ & \vdots \\ \mathbf{u}^R_{k+1} = \arg\min_{\mathbf{u}} & \left\| \left(\mathbf{y} - \mathcal{A}\left(\sum_{r=1}^{R-1} \mathbf{u}^r_{k+1} \mathbf{v}^r_{k+1}{}^T \right) \right) - \mathcal{A}(\mathbf{u}\mathbf{v}^{R\,T}_k) \right\|_2^2 \\ & +\alpha\|\mathbf{u}\|_2^2 + \frac{1}{2\lambda^R_k}\|\mathbf{u} - \mathbf{u}^R_k\|_2^2, \\ \mathbf{v}^R_{k+1} = \arg\min_{\mathbf{v}} & \left\| \left(\mathbf{y} - \mathcal{A}\left(\sum_{r=1}^{R-1} \mathbf{u}^r_{k+1} \mathbf{v}^r_{k+1}{}^T \right) \right) - \mathcal{A}(\mathbf{u}^R_{k+1}\mathbf{v}^T) \right\|^2 \\ & +\beta\|\mathbf{v}\|_1 + \frac{1}{2\mu^R_k}\|\mathbf{v} - \mathbf{v}^R_k\|_2^2, \end{cases}$$

In each iteration above, the terms $\|\mathbf{u} - \mathbf{u}^r_k\|_2^2$ and $\|\mathbf{v} - \mathbf{v}^r_k\|_2^2$ are added to provide theoretical convergence guarantees for the sequence $(\mathbf{u}^1_k, ..., \mathbf{v}^R_k)$ under suitable choice of the $2R$ positive sequences of parameters $(\lambda^1_k)_{k\in\mathbb{N}}, \dots, (\lambda^R_k)_{k\in\mathbb{N}}$, $(\mu^R_k)_{k\in\mathbb{N}}, \dots, (\mu^R_k)_{k\in\mathbb{N}} > 0$. In practice, ATLAS converges without those terms.

As most of the non-convex minimization algorithms, empirical performances of ATLAS likely depends on a proper initialization $(\mathbf{u}^1_0, \dots, \mathbf{u}^R_0, \mathbf{v}^1_0, \dots, \mathbf{v}^R_0)$. Initialization by the leading right singular vectors of $\mathcal{A}^*(\mathbf{y})$, where $\mathcal{A}^*$ denotes the adjoint of $\mathcal{A}$, ensures empirically stable recovery in the experiments (Section 7.7). However, we do not provide any theoretical guarantees for this observation.

One of the virtues of the algorithm is the explicit formulas for computation of the successive iterations, resulting in low computational complexity. Although relying on alternating minimization as SPF, at first glance, ATLAS may seem to be quite similar to SPF, it exhibits important positive differences: by using convex relaxation (ℓ_1-norm minimization) at each iteration, instead of solving a non-convex problem (ℓ_0-minimization) as in SPF, we can extend the approximation guarantees to the case of a high level of noise and effective sparsity of decomposition vectors, which are cases not covered by the theoretical guarantees of SPF. By virtue of the Lipschitz-continuity of soft-thresholding, we obtain approximation guarantees, also for the situation where neither restricted isometry property of the measurement operator $\mathcal{A}$ nor conditions on the support distribution of $\hat{\mathbf{X}}$ are assumed. In particular, while SPF can be considered as an alternating minimization over matrices, ATLAS alternates on R pairs of vectors. This enable us to drop the assumption of a common support for all columns (resp. rows) as in SPF.

7.3 Properties of Minimizers

Let us begin with some basic properties which minimizers of $J^R_{\alpha,\beta}$ have under very general assumptions. The first result bounds the error which is caused in measurements by $\mathbf{X}_{\alpha,\beta}$ in comparison to $\hat{\mathbf{X}}$.

Proposition 7.3.1. *Assume* $(\mathbf{u}^1_{\alpha,\beta}, ..., \mathbf{v}^R_{\alpha,\beta})$ *is a global minimizer of* $J^R_{\alpha,\beta}$ *and* $\hat{\mathbf{X}}$ *is fulfilling the noisy measurements* $\mathbf{y} = \mathcal{A}(\hat{\mathbf{X}}) + \boldsymbol{\eta}$. *Then,*

$$\|\mathbf{y} - \mathcal{A}(\mathbf{X}_{\alpha,\beta})\|_2^2 \leq \|\boldsymbol{\eta}\|_2^2 + C_{2,1}\sqrt[3]{\alpha\beta^2}\sum_{r=1}^{R}(\|\hat{\mathbf{u}}^r\|_2\|\hat{\mathbf{v}}^r\|_1)^{\frac{2}{3}}, \tag{7.15}$$

where $C_{2,1}$ *is the constant from Lemma 7.3.2.*

To prove Proposition 7.3.1, we need the following technical lemma.

Lemma 7.3.2. *Let* $\alpha, \beta, a, b, p, q > 0$. *Then*

$$f : \mathbb{R}^+ \to \mathbb{R}, \quad f(\lambda) := \lambda^p \alpha a + \frac{1}{\lambda^q}\beta b,$$

attains its minimum at $\tilde{\lambda} = \left(\frac{q}{p}\frac{\beta b}{\alpha a}\right)^{\frac{1}{p+q}}$ *and has the minimal value*

$$\min f = f(\tilde{\lambda}) = C_{p,q}(\alpha a)^{\frac{q}{p+q}}(\beta b)^{\frac{p}{p+q}},$$

where $C_{p,q} = \left(\frac{q}{p}\right)^{\frac{p}{p+q}} + \left(\frac{p}{q}\right)^{\frac{q}{p+q}}$.

Proof: The result is obtained by differentiation of f and by searching for its derivative's zeros. ■

Proof of Proposition 7.3.1: By applying Lemma 7.3.2 R times using $p = 2, q = 1, a = \|\hat{\mathbf{u}}^r\|_2^2, b = \|\hat{\mathbf{v}}^r\|_1$ we get $\tilde{\lambda}_1, ..., \tilde{\lambda}_R$, such that

$$\begin{aligned} J^R_{\alpha,\beta}(\tilde{\lambda}_1\hat{\mathbf{u}}^1, ..., \tilde{\lambda}_R\hat{\mathbf{u}}^R, \frac{1}{\tilde{\lambda}_1}\hat{\mathbf{v}}^1, ..., \frac{1}{\tilde{\lambda}_R}\hat{\mathbf{v}}^R) &= \|\mathbf{y} - \mathcal{A}(\hat{X})\|_2^2 + \sum_{r=1}^{R} C_{2,1}\sqrt[3]{\alpha\beta^2}\sqrt[3]{\|\hat{\mathbf{u}}^r\|_2^2\|\hat{\mathbf{v}}^r\|_1^2} \\ &= \|\boldsymbol{\eta}\|_2^2 + C_{2,1}\sqrt[3]{\alpha\beta^2}\sum_{r=1}^{R}(\|\hat{\mathbf{u}}^r\|_2\|\hat{\mathbf{v}}^r\|_1)^{\frac{2}{3}}. \end{aligned} \tag{7.16}$$

Note that, although not explicitly labeled, each $\tilde{\lambda}_r$ depends on the choice of α and β as well as on a, b, p, and q. The minimality of $(\mathbf{u}^1_{\alpha,\beta}, ..., \mathbf{v}^R_{\alpha,\beta})$ implies

$$\begin{aligned}\|\mathbf{y} - \mathcal{A}(X_{\alpha,\beta})\|_2^2 &\leq J^R_{\alpha,\beta}(\mathbf{u}^1_{\alpha,\beta}, ..., \mathbf{v}^R_{\alpha,\beta}) \leq J^R_{\alpha,\beta}(\tilde{\lambda}_1 \hat{\mathbf{u}}^1, ..., \tilde{\lambda}_R \hat{\mathbf{u}}^R, \frac{1}{\tilde{\lambda}_1}\hat{\mathbf{v}}^1, ..., \frac{1}{\tilde{\lambda}_R}\hat{\mathbf{v}}^R) \\ &= \|\boldsymbol{\eta}\|_2^2 + C_{2,1}\sqrt[3]{\alpha\beta^2}\sum_{r=1}^R (\|\hat{\mathbf{u}}^r\|_2 \|\hat{\mathbf{v}}^r\|_1)^{\frac{2}{3}}\end{aligned}$$

which is the claim. ■

By similar proof techniques, one can also control the norms of the components $\mathbf{u}^r_{\alpha,\beta}$ and $\mathbf{v}^r_{\alpha,\beta}$, for $r \in [R]$.

Lemma 7.3.3. *Assume $(\mathbf{u}^1_{\alpha,\beta}, ..., \mathbf{v}^R_{\alpha,\beta})$ is a global minimizer of $J^R_{\alpha,\beta}$ and $\hat{\mathbf{X}}$ is fulfilling the noisy measurements $\mathbf{y} = \mathcal{A}(\hat{\mathbf{X}}) + \boldsymbol{\eta}$. If $\|\mathbf{y} - \mathcal{A}(\mathbf{X}_{\alpha,\beta})\|_2 \geq \|\boldsymbol{\eta}\|_2$, we have*

$$\begin{aligned}\sum_{r=1}^R \|\mathbf{u}^r_{\alpha,\beta}\|_2^2 &\leq C_{2,1}\sqrt[3]{\frac{\beta^2}{\alpha^2}}\sum_{r=1}^R (\|\hat{\mathbf{u}}^r\|_2\|\hat{\mathbf{v}}^r\|_1)^{\frac{2}{3}}, \\ \sum_{r=1}^R \|\mathbf{v}^r_{\alpha,\beta}\|_1 &\leq C_{2,1}\sqrt[3]{\frac{\alpha}{\beta}}\sum_{r=1}^R (\|\hat{\mathbf{u}}^r\|_2\|\hat{\mathbf{v}}^r\|_1)^{\frac{2}{3}},\end{aligned} \tag{7.17}$$

and

$$\sum_{r=1}^R \left(\|\mathbf{u}^r_{\alpha,\beta}\|_2\|\mathbf{v}^r_{\alpha,\beta}\|_1\right)^{\frac{2}{3}} \leq \sum_{r=1}^R (\|\hat{\mathbf{u}}^r\|_2\|\hat{\mathbf{v}}^r\|_1)^{\frac{2}{3}}$$

where $C_{2,1}$ is the constant from Lemma 7.3.2.

Proof: From (7.16) in the proof of Proposition 7.3.1 we obtain

$$\begin{aligned}&\|\mathbf{y} - \mathcal{A}(\mathbf{X}_{\alpha,\beta})\|_2^2 + \sum_{r=1}^R \left(\alpha\|\mathbf{u}^r_{\alpha,\beta}\|_2^2 + \beta\|\mathbf{v}^r_{\alpha,\beta}\|_1\right) \\ &\qquad = J^R_{\alpha,\beta}(\mathbf{u}^1_{\alpha,\beta}, ..., \mathbf{v}^R_{\alpha,\beta}) \leq J^R_{\alpha,\beta}(\tilde{\lambda}_1 \hat{\mathbf{u}}^1, ..., \tilde{\lambda}_R \hat{\mathbf{u}}^R, \frac{1}{\tilde{\lambda}_1}\hat{\mathbf{v}}^1, ..., \frac{1}{\tilde{\lambda}_R}\hat{\mathbf{v}}^R) \\ &\qquad = \|\boldsymbol{\eta}\|_2^2 + C_{2,1}\sqrt[3]{\alpha\beta^2}\sum_{r=1}^R (\|\hat{\mathbf{u}}^r\|_2\|\hat{\mathbf{v}}^r\|_1)^{\frac{2}{3}}\end{aligned}$$

The first part of the claim follows by subtracting $\|\mathbf{y} - \mathcal{A}(\mathbf{X}_{\alpha,\beta})\|_2^2$ on both sides, leaving out half of the terms on the left-hand side, and dividing by α (resp. β). To show the second part, note that by minimality of $(\mathbf{u}^1_{\alpha,\beta}, ..., \mathbf{v}^R_{\alpha,\beta})$ and Lemma 7.3.2

$$\sum_{r=1}^R \left(\alpha\|\mathbf{u}^r_{\alpha,\beta}\|_2^2 + \beta\|\mathbf{v}^r_{\alpha,\beta}\|_1\right) = C_{2,1}\sqrt[3]{\alpha\beta^2}\sum_{r=1}^R \left(\|\mathbf{u}^r_{\alpha,\beta}\|_2\|\mathbf{v}^r_{\alpha,\beta}\|_1\right)^{\frac{2}{3}}$$

and hence

$$\begin{aligned}\|\mathbf{y}-\mathcal{A}(\mathbf{X}_{\alpha,\beta})\|_2^2 + C_{2,1}\sqrt[3]{\alpha\beta^2}\sum_{r=1}^R \left(\|\mathbf{u}_{\alpha,\beta}^r\|_2\|\mathbf{v}_{\alpha,\beta}^r\|_1\right)^{\frac{2}{3}} \\ = J_{\alpha,\beta}^R(\mathbf{u}_{\alpha,\beta}^1,...,\mathbf{v}_{\alpha,\beta}^R) \leq J_{\alpha,\beta}^R(\tilde{\lambda}_1\hat{\mathbf{u}}^1,...,\tilde{\lambda}_R\hat{\mathbf{u}}^R,\frac{1}{\tilde{\lambda}_1}\hat{\mathbf{v}}^1,...,\frac{1}{\tilde{\lambda}_R}\hat{\mathbf{v}}^R) \\ = \|\boldsymbol{\eta}\|_2^2 + C_{2,1}\sqrt[3]{\alpha\beta^2}\sum_{r=1}^R (\|\hat{\mathbf{u}}^r\|_2\|\hat{\mathbf{v}}^r\|_1)^{\frac{2}{3}}.\end{aligned}$$

Subtracting $\|\mathbf{y}-\mathcal{A}(\mathbf{X}_{\alpha,\beta})\|_2^2$ on both sides and dividing by $C_{2,1}\sqrt[3]{\alpha\beta^2}$ concludes the proof. ■

The two estimates in (7.17) point out an interesting property of $J_{\alpha,\beta}^R$. If one chooses the parameters α and β of different magnitude, either the left or the right components of a minimizer $(\mathbf{u}_{\alpha,\beta}^1,...,\mathbf{v}_{\alpha,\beta}^R)$ can be forced to become smaller in norm, while the grip on the others is lost. If α and β are chosen to be equal the norm bounds are balanced and one obtains

$$\sum_{r=1}^R \left(\|\mathbf{u}_{\alpha,\beta}^r\|_2^2 + \|\mathbf{v}_{\alpha,\beta}^r\|_1\right) \leq C_{2,1}\sum_{r=1}^R (\|\hat{\mathbf{u}}^r\|_2\|\hat{\mathbf{v}}^r\|_1)^{\frac{2}{3}}.$$

The assumption $\|\mathbf{y}-\mathcal{A}(\mathbf{X}_{\alpha,\beta})\|_2 \geq \|\boldsymbol{\eta}\|_2$ is not restrictive. As soon as $\|\mathbf{y}-\mathcal{A}(\mathbf{X}_{\alpha,\beta})\|_2 = \|\boldsymbol{\eta}\|_2$ one doesn't have to diminish α and β any further. There is no hope in obtaining an accuracy below noise level and any smaller parameter choice will lead to overfitting phenomena.

The ℓ_1-regularization in $J_{\alpha,\beta}^R$ provides means to bound $\|\mathbf{v}_{\alpha,\beta}^r\|_1$, for $r \in [R]$. By this, one can control effective sparsity of the minimizer's right components.

Lemma 7.3.4. *Assume $\mathcal{A}\colon \mathbb{R}^{n_1\times n_2}\to\mathbb{R}^m$ is a linear operator and $\mathbf{y}\in\mathbb{R}^m$. Let $(\mathbf{u}_{\alpha,\beta}^1,...,\mathbf{v}_{\alpha,\beta}^R)$ be a minimizer of $J_{\alpha,\beta}^R$. For all $r\in[R]$ we have that if $\|\mathbf{v}_{\alpha,\beta}^r\|_2 \geq \|\mathbf{y}\|_2^2/\gamma$ for some $\gamma > 0$, then*

$$\frac{\|\mathbf{v}_{\alpha,\beta}^r\|_1}{\|\mathbf{v}_{\alpha,\beta}^r\|_2} < \frac{\gamma}{\beta}.$$

Proof: By comparing $J_{\alpha,\beta}^R(\mathbf{u}_{\alpha,\beta}^1,...,\mathbf{v}_{\alpha,\beta}^r)$ to $J_{\alpha,\beta}^R(0,...,0)$, we get

$$\sum_{r=1}^R \left(\alpha\|\mathbf{u}_{\alpha,\beta}^r\|_2^2 + \beta\|\mathbf{v}_{\alpha,\beta}^r\|_1\right) \leq J_{\alpha,\beta}^R(\mathbf{u}_{\alpha,\beta}^1,...,\mathbf{v}_{\alpha,\beta}^R) \leq J_{\alpha,\beta}^R(0,...,0) = \|\mathbf{y}\|_2^2.$$

This implies $\|\mathbf{v}_{\alpha,\beta}^r\|_1 < \|\mathbf{y}\|_2^2/\beta$. As by assumption $\|\mathbf{v}_{\alpha,\beta}^r\|_2 \geq \|\mathbf{y}\|_2^2/\gamma$, we conclude

$$\frac{\|\mathbf{v}_{\alpha,\beta}^r\|_1}{\|\mathbf{v}_{\alpha,\beta}^r\|_2} < \frac{\|\mathbf{y}\|_2^2}{\beta}\frac{\gamma}{\|\mathbf{y}\|_2^2} = \frac{\gamma}{\beta}.$$

■

Lemma 7.3.4 states that those vectors $\mathbf{v}^r_{\alpha,\beta}$ which lie not too close to zero are effectively sparse. Numerical experiments suggest that if $\hat{\mathbf{X}}$ has s-sparse right components $\hat{\mathbf{v}}^r$, ATLAS leads to solutions with sparse right components $\mathbf{v}^r_{\alpha,\beta}$. The theoretical necessity of considering not only sparsity but effective sparsity in this case is caused by a missing bound on the support size of the vectors $\mathbf{v}^r_{\alpha,\beta}$.

Considering (7.10) and (7.13), Proposition 7.3.1 and Lemma 7.3.4 state that $\mathbf{X}_{\alpha,\beta}$ is even without any requirements on $\mathcal{A}$ a reasonable approximation of $\hat{\mathbf{X}}$, i.e., it is of rank R, yields similar measurements, and has effectively sparse right components. However, the parameters α and β have to be chosen with care, neither too small nor too large. Moreover, Lemma 7.3.3 shows that α and β have to be chosen of similar magnitude. Otherwise either left or right components of $\mathbf{X}_{\alpha,\beta}$ cannot be controlled.

7.4 Signal Sets and Stable Embeddings

Though applicable in general, the results of Section 7.3 do not provide approximation guarantees comparable to Theorem 7.1.2. To obtain a statement of similar flavor, we have to make the definition of our signal set more precise and to find operators which embed the signal sets in a stable way like the RIP in Definition 7.1.1. Let us first characterize a set of matrices allowing decompositions of the form (7.10) with sparse left and right components. We define, for $\Gamma \geq 1$,

$$\begin{aligned} S^{R,\Gamma}_{s_1,s_2} = \{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2} : \exists\, & \mathbf{u}^1, ..., \mathbf{u}^R \in \mathbb{R}^{n_1},\ \mathbf{v}^1, ..., \mathbf{v}^R \in \mathbb{R}^{n_2}, \\ & \text{and } \boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_R)^T \in \mathbb{R}^R, \text{ s.t.} \\ & \quad \mathbf{Z} = \sum_{r=1}^R \sigma_r \mathbf{u}^r (\mathbf{v}^r)^T, \\ & \text{where } |\operatorname{supp}(\mathbf{u}^r)| \leq s_1,\ |\operatorname{supp}(\mathbf{v}^r)| \leq s_2, \\ & \|\mathbf{u}^r\|_2 = \|\mathbf{v}^r\|_2 = 1, \text{ for all } r \in [R], \\ & \text{and } \|\boldsymbol{\sigma}\|_2 \leq \Gamma\}. \end{aligned} \tag{7.18}$$

It contains all matrices $\mathbf{Z}$ which can be decomposed into three matrices $\mathbf{U\Sigma V}^T$ such that $\mathbf{U} \in \mathbb{R}^{n_1 \times R}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times R}$ have s_1-sparse (resp. s_2-sparse) unit norm columns and $\mathbf{\Sigma} \in \mathbb{R}^{R \times R}$ is the diagonal matrix defined by $\boldsymbol{\sigma}$. The set is restricted to decompositions with $\|\mathbf{\Sigma}\|_F \leq \Gamma$. The important difference w.r.t. [116] is that the columns do not need to share a common support. Moreover, we do not require $\mathbf{U}$ and $\mathbf{V}$ to be orthogonal matrices. In particular, all matrices $\mathbf{X}$ with rank less or equal R, s_1-sparse (resp. s_2-sparse) left and right singular vectors, and $\|\mathbf{X}\|_F \leq \Gamma$ are in $S^{R,\Gamma}_{s_1,s_2}$. In this case $\|\mathbf{\Sigma}\|_F = \|\mathbf{X}\|_F$. We call such an admissible decomposition $\mathbf{U\Sigma V}^T$ in (7.18) a Sparse Decomposition (SD) of $\mathbf{Z}$. Note that the SD is not unique and that the SVD of $\mathbf{Z}$ is not necessarily a SD of $\mathbf{Z}$.
The second set is a further generalization of $S^{R,\Gamma}_{s_1,s_2}$. We drop the sparsity assumption and

replace it by effective sparsity. Define, for $\Gamma \geq 1$,

$$\begin{aligned} K_{s_1,s_2}^{R,\Gamma} = \{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2} \colon & \exists\, \mathbf{u}^1, ..., \mathbf{u}^R \in K_{n_1,s_1},\ \mathbf{v}^1, ..., \mathbf{v}^R \in K_{n_2,s_2}, \\ & \text{and } \boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_R)^T \in \mathbb{R}^R, \text{ s.t.} \\ & \mathbf{Z} = \sum_{r=1}^{R} \sigma_r \mathbf{u}^r (\mathbf{v}^r)^T, \\ & \text{where } \|\mathbf{u}^r\|_2 = \|\mathbf{v}^r\|_2 = 1, \text{ for all } r \in [R], \\ & \text{and } \|\boldsymbol{\sigma}\|_2 \leq \Gamma\} \end{aligned} \tag{7.19}$$

which is a relaxed version of $S_{s_1,s_2}^{R,\Gamma}$ as $S_{s_1,s_2}^{R,\Gamma} \subset K_{s_1,s_2}^{R,\Gamma}$. The class of matrices $K_{s_1,s_2}^{R,\Gamma}$ is to a certain extent closed under summation: in fact if $\mathbf{Z} \in K_{s_1,s_2}^{R,\Gamma}$ and $\hat{\mathbf{Z}} \in K_{\hat{s}_1,\hat{s}_2}^{R,\hat{\Gamma}}$ then

$$\mathbf{Z} - \hat{\mathbf{Z}} \in K_{\max\{s_1,\hat{s}_1\},\max\{s_2,\hat{s}_2\}}^{2R,\sqrt{\Gamma^2+\hat{\Gamma}^2}}. \tag{7.20}$$

We call such an admissible decomposition $\mathbf{Z} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ in (7.19) an effectively Sparse Decomposition of $\mathbf{Z}$ and use the same shorthand notation, i.e., SD. The context makes clear which decomposition is meant. Any $\hat{\mathbf{X}}$ decomposed as in (7.10) belongs to $K_{n_1,s}^{R,\Gamma}$ if $\sum_{r=1}^{R} \|\hat{\mathbf{u}}^r\|_2^2 \|\hat{\mathbf{v}}^r\|_2^2 \leq \Gamma^2$. Having the sets $S_{s_1,s_2}^{R,\Gamma}$ and $K_{s_1,s_2}^{R,\Gamma}$ at hand we now define corresponding RIPs.

Definition 7.4.1 (Additive Rank-R and (effectively) (s_1, s_2)-sparse RIP$_\Gamma$)**.** *A linear operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ satisfies the additive rank-R and (s_1, s_2)-sparse RIP$_\Gamma$ with isometry constant $\delta > 0$ if*

$$\left| \|\mathcal{A}(\mathbf{Z})\|_2^2 - \|\mathbf{Z}\|_F^2 \right| \leq \delta, \tag{7.21}$$

for all $\mathbf{Z} \in S_{s_1,s_2}^{R,\Gamma}$. If (7.21) holds for all $\mathbf{Z} \in K_{s_1,s_2}^{R,\Gamma}$, we say $\mathcal{A}$ has the additive rank-R and effectively (s_1, s_2)-sparse RIP$_\Gamma$. Note that the rank-R and effectively (s_1, s_2)-sparse RIP$_\Gamma$ implies the rank-R and (s_1, s_2)-sparse RIP$_\Gamma$ as $S_{s_1,s_2}^{R,\Gamma} \subset K_{s_1,s_2}^{R,\Gamma}$.

We comment on the additive form of (7.21) in Remark 7.4.3 below.
A linear operator $\mathbf{A}$ of the form (7.13) which is drawn from a subgaussian distribution fulfills the above introduced RIPs with high probability. This is stated in the following Lemma. Recall the definition of subgaussian random variables in Definition 3.3.6.

Lemma 7.4.2 (RIP for Subgaussian Operators)**.** *Let $\Gamma \geq 1$ and let $\mathcal{A}\colon \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ be the linear measurement operator of form (7.13). Assume, all $\mathbf{A}_i$, for $1 \leq i \leq m$, have iid $\mathcal{K}$-subgaussian entries $a_{i,j,k}$ with mean 0 and variance 1. If*

$$m \gtrsim \left(\frac{\delta}{\Gamma^2 R} \right)^{-2} R(s_1 + s_2) \log\left(\max\{n_1, n_2\}\right) \tag{7.22}$$

then $\mathcal{A}$ has the additive rank-R and (s_1, s_2)-sparse RIP$_\Gamma$ with isometry constant $\delta \in (0, \Gamma^2 R)$ with probability at least $1 - 2\exp(-C(\delta/\Gamma^2 R)^2 m)$ where $C > 0$ is a constant depending on $\mathcal{K}$. If

$$m \gtrsim \left(\frac{\delta}{\Gamma^2 R} \right)^{-2} R(s_1 + s_2) \log^3\left(\max\{n_1, n_2\}\right) \tag{7.23}$$

then $\mathcal{A}$ has the additive rank-R and effectively (s_1, s_2)-sparse RIP$_\Gamma$ with isometry constant $\delta \in (0, \Gamma^2 R)$ with probability at least $1 - 2\exp(-C'(\delta/\Gamma^2 R)^2 m)$ where $C' > 0$ is a constant depending on $\mathcal{K}$.

Lemma 7.4.2 states, for $\delta = d(\Gamma^2 R)$, $d \in (0,1)$, that $m \approx \mathcal{O}\left(d^{-2} R(s_1 + s_2)\right)$ subgaussian measurements are sufficient to have δ-stable embeddings of $S^{R,\Gamma}_{s_1,s_2}$ and $K^{R,\Gamma}_{s_1,s_2}$ (cf. [140, Def. 1.1 & Thm. 1.5]). Note that $\Gamma^2 R$ is the squared Frobenius diameter of $S^{R,\Gamma}_{s_1,s_2}$ and $K^{R,\Gamma}_{s_1,s_2}$. As we restrict ourselves below to s-effective sparse right component vectors of $\hat{\mathbf{X}}$, we only use the rank-R and (effectively) (n_1, s)-sparse RIP$_\Gamma$. For the presented results to have some meaning, a typical dimensional setting is $R \ll s \approx n_1 \ll n_2$. In fact, if n_1 were close to n_2 in magnitude, the sparsity s of the right component vectors would not be useful to reduce the order of the measurements, as they would already be of order $n_1 \approx n_2$. Moreover, if R were close to n_1, the matrix would not be low-rank as n_1 would be the maximal possible rank.

Definition 7.4.1 and Lemma 7.4.2 allow more general settings. In [116] the authors give information theoretical lower bounds on the necessary number of measurements for reconstructing low-rank matrices with sparse singular vectors (sharing a common support), namely $m \gtrsim R(s_1 + s_2)$. As we do not require orthogonality of SDs in $S^{R,\Gamma}_{s_1,s_2}$ resp. $K^{R,\Gamma}_{s_1,s_2}$ (excluding a scaling invariant RIP which is independent of the set diameter, see Remark 7.4.3), the bounds in (7.22) and (7.23) are close to information theoretic limits. We are not aware of any information theoretical lower bounds for our more general setting.

Remark 7.4.3. *The additive RIP in* (7.21) *differs from the commonly used multiplicative RIPs of the form*

$$(1-\delta)\|\mathbf{Z}\|_F^2 \leq \|\mathcal{A}(\mathbf{Z})\|_2^2 \leq (1+\delta)\|\mathbf{Z}\|_F^2 \tag{7.24}$$

as it is not scaling invariant and $\mathcal{A}(\mathbf{Z}) = \mathcal{A}(\mathbf{Z}')$ does not imply $\mathbf{Z} = \mathbf{Z}'$ but only $\|\mathbf{Z}-\mathbf{Z}'\|_2^2 \leq \delta$. In fact it is not possible to derive a classical scaling invariant RIP like (7.24) *on $K^{R,\Gamma}_{s_1,s_2}$ under similar conditions as* (7.23). *The main problem is non-orthogonality of the SD. A simple example illustrates this point: Assume $m \simeq (n_1 + s)\log^3(\max\{n_1, n_2\})$ and the linear operator $\mathcal{A}$ fulfills* (7.24) *for all $\mathbf{Z} \in K^{2,1}_{n_1,s}$. Choose some $\mathbf{u} \in \mathbb{R}^{n_1}, \mathbf{v}_1 \in \mathbb{R}^{n_2}$ of unit norm and $\|\mathbf{v}_1\|_1 \leq \sqrt{s}/2$. Define $\mathbf{v}_2 := -\mathbf{v}_1 + \varepsilon\mathbf{w}$ for any $\mathbf{w} \in \mathbb{R}^{n_2}$ and choose $\varepsilon > 0$ sufficiently small to ensure $\|\mathbf{v}_2\|_1 \leq \sqrt{s}$ and $\|\mathbf{v}_2\|_2 \approx 1$. Then $\mathbf{Z} := (1/2)\mathbf{u}\mathbf{v}_1^T + (1/2)\mathbf{u}\mathbf{v}_2^T \in K^{2,1}_{n_1,s}$ and* (7.24) *holds. But this implies by definition of $\mathbf{Z}$ and scaling invariance of* (7.24) *that*

$$(1-\delta)\|\mathbf{u}\mathbf{w}^T\|_F^2 \leq \|\mathbf{A}(\mathbf{u}\mathbf{w}^T)\|_2^2 \leq (1+\delta)\|\mathbf{u}\mathbf{w}^T\|_F^2$$

which means the RIP directly extends to all rank-1 matrices (not only those with sparse right component). If $n_1, s \ll n_2$, this is a clear contradiction to information theoretical lower bounds, as corresponding RIPs would require at least $m \simeq \max\{n_1, n_2\}$ (see [28, Section 2.1]).

For proving Lemma 7.4.2 we need bounds on the covering numbers of $S^{R,\Gamma}_{s_1,s_2}$ and $K^{R,\Gamma}_{s_1,s_2}$. The bound for $N(S^{R,\Gamma}_{s_1,s_2}, \|\cdot\|_F, \varepsilon)$ below is an adaption of [28, Lemma 3.1].

Lemma 7.4.4. *Let $S^{R,\Gamma}_{s_1,s_2}$ be the set defined in* (7.18). *Then, for all* $0 < \varepsilon < 1$, *one has*

$$\begin{aligned} &\log(N(S^{R,\Gamma}_{s_1,s_2}, \|\cdot\|_F, \varepsilon)) \\ &\le R(s_1+s_2+1)\log\left(\frac{18\Gamma R}{\varepsilon}\right) + Rs_1\log\left(\frac{en_1}{s_1}\right) + Rs_2\log\left(\frac{en_2}{s_2}\right). \end{aligned} \tag{7.25}$$

Proof: Recall, each $\mathbf{Z} \in S^{R,\Gamma}_{s_1,s_2}$ can be represented as $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ with $\mathbf{U} = (\mathbf{u}^1, ..., \mathbf{u}^R)$, $\mathbf{V} = (\mathbf{v}^1, ..., \mathbf{v}^R)$ where all unit norm columns $\mathbf{u}^r \in \mathbb{R}^{n_1}$ are s_1-sparse, all unit norm columns $\mathbf{v}^r \in \mathbb{R}^{n_2}$ are s_2-sparse, and $\|\mathbf{\Sigma}\|_F \le \Gamma$. Let us first consider the larger set $S = \{\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \colon \mathbf{U} \in Q^R_{n_1,s_1}, \mathbf{\Sigma} \in D_\Gamma,$ and $\mathbf{V} \in Q^R_{n_2,s_2}\}$ where D_Γ is the set of $R \times R$ diagonal matrices with Frobenius norm less or equal Γ and $Q^R_{n,s} = \{\mathbf{W} \in \mathbb{R}^{n\times R} \colon \|\mathbf{W}\|_F \le \sqrt{R}$ and all columns $\mathbf{w}^r$ are s-sparse$\}$. Then, we know that $S^{R,\Gamma}_{s_1,s_2} \subset S$. We construct an $(\varepsilon/2)$-net $\tilde{S}$ of S by covering the sets of permissible $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}$ and conclude the proof by applying the well-known relation $N(K, \|\cdot\|, \varepsilon) \le N(K', \|\cdot\|, \varepsilon/2)$ which holds whenever $K \subset K'$.

Recall that if B is a unit ball in D dimensions (with respect to some norm $\|\cdot\|_B$) there exists an ε-net $\tilde{B}$ (i.e., for all $\mathbf{b} \in B$ there is some $\tilde{\mathbf{b}} \in \tilde{B}$ with $\|\mathbf{b} - \tilde{\mathbf{b}}\|_B \le \varepsilon$) with $\tilde{B} \subset B$ and $|\tilde{B}| \le (3/\varepsilon)^D$. See for example [28, Section 3]. Moreover, note that $N(K, \|\cdot\|, \varepsilon) = N(cK, \|\cdot\|, c\varepsilon)$ for any set K and $c > 0$. Hence, for any scaled unit ball cB there exists an ε-net $\tilde{B} \subset cB$ and $|\tilde{B}| \le (3c/\varepsilon)^D$.

Let $\tilde{D}_\Gamma$ be an $(\varepsilon/(6R))$-net of D_Γ which is of size $|\tilde{D}_\Gamma| \le (18\Gamma R/\varepsilon)^R$. For $\mathbf{W} \in \mathbb{R}^{n\times R}$ denote by $\text{supp}(\mathbf{W}) = \{\text{supp}(\mathbf{w}^1), ..., \text{supp}(\mathbf{w}^R)\}$ and by $\text{supp}(\mathbf{W}) \Subset \text{supp}(\mathbf{W}')$ that $\text{supp}(\mathbf{w}^r) \subset \text{supp}((\mathbf{w}')^r)$, for all $r \in [R]$. Define the set of all possible supports of maximal size

$$T^R_{n,s} = \{\text{supp}(\mathbf{W}) \colon \mathbf{W} \in \mathbb{R}^{n\times R} \text{ and all columns } \mathbf{w}^r \text{ have exactly } s \text{ non-zero entries}\}.$$

For any fixed $\theta \in T^R_{n,s}$ the set $\{\mathbf{W} \in Q^R_{n,s} \colon \text{supp}(\mathbf{W}) \Subset \theta\}$ is an $\mathbb{R}^{s\times R}$ Frobenius ball of radius $\sqrt{R}$ embedded into $\mathbb{R}^{n\times R}$ and $Q^R_{n,s} = \bigcup_{\theta\in T^R_{n,s}} \{\mathbf{W} \in Q^R_{n,s} \colon \text{supp}(\mathbf{W}) \Subset \theta\}$. Hence, there is an $(\varepsilon/(6\Gamma\sqrt{R}))$-net $\tilde{Q}^R_{n,s}$ of $Q^R_{n,s}$ with

$$|\tilde{Q}^R_{n,s}| \le |T^R_{n,s}| \left(\frac{18\Gamma R}{\varepsilon}\right)^{Rs} \le \binom{n}{s}^R \left(\frac{18\Gamma R}{\varepsilon}\right)^{Rs} \le \left(\frac{en}{s}\right)^{Rs} \left(\frac{18\Gamma R}{\varepsilon}\right)^{Rs}$$

We define now $\tilde{S} = \{\tilde{\mathbf{Z}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T \colon \tilde{\mathbf{U}} \in \tilde{Q}^R_{n_1,s_1}, \tilde{\mathbf{\Sigma}} \in \tilde{D}_\Gamma,$ and $\tilde{\mathbf{V}} \in \tilde{Q}^R_{n_2,s_2}\}$. It is clear that

$$|\tilde{S}| \le |\tilde{Q}^R_{n_1,s_1}| \cdot |\tilde{D}_\Gamma| \cdot |\tilde{Q}^R_{n_2,s_2}| \le \left(\frac{18\Gamma R}{\varepsilon}\right)^{R(s_1+s_2+1)} \left(\frac{en_1}{s_1}\right)^{Rs_1} \left(\frac{en_2}{s_2}\right)^{Rs_2}.$$

Let us conclude by showing $\tilde{S}$ is indeed an $(\varepsilon/2)$-net for S. Given any $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \in S$, there exists $\tilde{\mathbf{Z}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T \in \tilde{S}$ with $\|\mathbf{U} - \tilde{\mathbf{U}}\|_F \le \varepsilon/(6\Gamma\sqrt{R})$, $\|\mathbf{\Sigma} - \tilde{\mathbf{\Sigma}}\|_F \le \varepsilon/(6R)$, and $\|\mathbf{V}-\tilde{\mathbf{V}}\|_F \leq \varepsilon/(6\Gamma\sqrt{R})$. We can estimate

$$\begin{aligned}\|\mathbf{Z}-\tilde{\mathbf{Z}}\|_F &\leq \|(\mathbf{U}-\tilde{\mathbf{U}})\boldsymbol{\Sigma}\mathbf{V}^T\|_F + \|\tilde{\mathbf{U}}(\boldsymbol{\Sigma}-\tilde{\boldsymbol{\Sigma}})\mathbf{V}^T\|_F + \|\tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}(\mathbf{V}-\tilde{\mathbf{V}})^T\|_F \\ &\leq \frac{\varepsilon}{6\Gamma\sqrt{R}}\Gamma\sqrt{R} + \sqrt{R}\frac{\varepsilon}{6R}\sqrt{R} + \sqrt{R}\Gamma\frac{\varepsilon}{6\Gamma\sqrt{R}} \\ &\leq \frac{\varepsilon}{2}\end{aligned}$$

where we used triangle inequality in the first line and $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F\|\mathbf{B}\|_F$ in the second. ■

To derive a similar bound on $N(K_{s_1,s_2}^{R,\Gamma}, \|\cdot\|_F, \varepsilon)$ recall Lemma 3.2.2 which characterizes the covering number of $K_{n,s} \subset \mathbb{R}^n$.

Lemma 7.4.5. *Let $K_{s_1,s_2}^{R,\Gamma}$ be the set defined in* (7.19). *Assume w.l.o.g. that $s_1/n_1 \leq s_2/n_2$. Then, for all $0 < \varepsilon < 6\Gamma\sqrt{R}$, one has*

$$\begin{aligned}&\log(N(K_{s_1,s_2}^{R,\Gamma}, \|\cdot\|_F, \varepsilon)) \\ &\leq \begin{cases} R(n_1+n_2+1)\log\left(\frac{36\Gamma R}{\varepsilon}\right) & 0<\varepsilon<12\Gamma\sqrt{\frac{Rs_1}{n_1}}, \\ \frac{144\Gamma^2R^2s_1}{\varepsilon^2}\log\left(\frac{9\varepsilon n_1}{6\Gamma\sqrt{R}s_1}\right) + R(n_2+1)\log\left(\frac{36\Gamma R}{\varepsilon}\right) & 12\Gamma\sqrt{\frac{Rs_1}{n_1}}\leq\varepsilon<12\Gamma\sqrt{\frac{Rs_2}{n_2}}, \\ \frac{144\Gamma^2R^2(s_1+s_2)}{\varepsilon^2}\log\left(\frac{9\varepsilon n_1}{6\Gamma\sqrt{R}s_1}\right) + R\log\left(\frac{18\Gamma R}{\varepsilon}\right) & 12\Gamma\sqrt{\frac{Rs_2}{n_2}}\leq\varepsilon<6\Gamma\sqrt{R}. \end{cases}\end{aligned} \tag{7.26}$$

Proof: Let $\tilde{K}_{n,s}$ be a minimal $\varepsilon/(6\Gamma\sqrt{R})$-net for $K_{n,s}$ in Euclidean norm. Let D_Γ be the set of $R\times R$ diagonal matrices with Frobenius-norm less or equal Γ. As discussed in the proof of Lemma 7.4.4, one has that $N(D_\Gamma, \|\cdot\|_F, \varepsilon) \leq (3\Gamma/\varepsilon)^R$. Denote by $\tilde{D}_\Gamma$ a minimal $(\varepsilon/(6R))$-net of D_Γ and define the sets

$$\begin{aligned}K = \{&\mathbf{Z}\in\mathbb{R}^{n_1\times n_2}\colon \mathbf{Z}=\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \\ &\text{with } \mathbf{u}^r\in K_{n_1,s_1},\ \mathbf{v}^r\in K_{n_2,s_2} \text{ for all } r\in[R], \text{ and } \|\boldsymbol{\Sigma}\|_F\leq\Gamma\} \\ \tilde{K} = \{&\tilde{\mathbf{Z}}\in\mathbb{R}^{n_1\times n_2}\colon \tilde{\mathbf{Z}}=\tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{V}}^T \\ &\text{with } \tilde{\mathbf{u}}^r\in \tilde{K}_{n_1,s_1},\ \tilde{\mathbf{v}}^r\in \tilde{K}_{n_2,s_2} \text{ for all } r\in[R], \text{ and } \tilde{\boldsymbol{\Sigma}}\in\tilde{D}_\Gamma\}.\end{aligned}$$

We first show that $\tilde{K}$ is an $(\varepsilon/2)$-net of K. Let $\mathbf{Z}=\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\in K$ be given. There exists $\tilde{\mathbf{Z}}=\tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{V}}^T\in\tilde{K}$ with $\|\mathbf{u}^r-\tilde{\mathbf{u}}^r\|_2\leq\varepsilon/(6\Gamma\sqrt{R})$, $\|\mathbf{v}^r-\tilde{\mathbf{v}}^r\|_2\leq\varepsilon/(6\Gamma\sqrt{R})$, for all $r\in[R]$, and $\|\boldsymbol{\Sigma}-\tilde{\boldsymbol{\Sigma}}\|_F\leq\varepsilon/(6R)$. Therefore, $\|\mathbf{U}-\tilde{\mathbf{U}}\|_F^2 = \sum_{r=1}^R\|\mathbf{u}^r-\tilde{\mathbf{u}}^r\|_2^2 \leq (\varepsilon/(6\Gamma))^2$ and $\|\mathbf{V}-\tilde{\mathbf{V}}\|_F^2\leq(\varepsilon/(6\Gamma))^2$. Moreover, $\|\mathbf{U}\|_F^2=\sum_{r=1}^R\|\mathbf{u}^r\|_2^2\leq R$ (the same holds for $\mathbf{V},\tilde{\mathbf{U}},\tilde{\mathbf{V}}$) and $\|\mathbf{U}\boldsymbol{\Sigma}\|_F\leq\|\boldsymbol{\Sigma}\|_F$ (the same holds for $\boldsymbol{\Sigma}\mathbf{V}^T,\tilde{\mathbf{U}}\boldsymbol{\Sigma},\boldsymbol{\Sigma}\tilde{\mathbf{V}}^T$). We now obtain by triangle inequality and the fact that $\|\mathbf{AB}\|_F\leq\|\mathbf{A}\|_F\|\mathbf{B}\|_F$

$$\begin{aligned}\|\mathbf{Z}-\tilde{\mathbf{Z}}\|_F &\leq \|(\mathbf{U}-\tilde{\mathbf{U}})\boldsymbol{\Sigma}\mathbf{V}^T\|_F + \|\tilde{\mathbf{U}}(\boldsymbol{\Sigma}-\tilde{\boldsymbol{\Sigma}})\mathbf{V}^T\|_F + \|\tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}(\mathbf{V}-\tilde{\mathbf{V}})^T\|_F \\ &\leq \frac{\varepsilon}{6\Gamma}\Gamma + \sqrt{R}\frac{\varepsilon}{6R}\sqrt{R} + \Gamma\frac{\varepsilon}{6\Gamma} \leq \frac{\varepsilon}{2}.\end{aligned}$$

Since $K_{s_1,s_2}^{R,\Gamma}\subset K$ one has $N(K_{s_1,s_2}^{R,\Gamma},\|\cdot\|_F,\varepsilon)\leq N(K,\|\cdot\|_F,\varepsilon/2)$. Hence,

$$N(K_{s_1,s_2}^{R,\Gamma},\|\cdot\|_F,\varepsilon)\leq|\tilde{K}|\leq|\tilde{K}_{n_1,s_1}|^R|\tilde{D}_\Gamma||\tilde{K}_{n_2,s_2}|^R$$

which yields the claim by applying Lemma 3.2.2. ■

Lemma 7.4.2 can now be proven by applying the following bound on suprema of chaos processes [107, Theorems 1.4 & 3.1] in combination with the bounds on the covering numbers $N(S_{s_1,s_2}^{R,\Gamma}, \|\cdot\|_F, \varepsilon)$ and $N(K_{s_1,s_2}^{R,\Gamma}, \|\cdot\|_F, \varepsilon)$ in Lemma 7.4.4 and 7.4.5. We recall below the relevant result in the form presented in [102]. The appearing γ_2-functional is defined in [107] and can be bounded by

$$\gamma_2\left(\mathcal{H}, \|\cdot\|_{2\to 2}\right) \lesssim \int_0^{d_{2\to 2}(\mathcal{H})} \sqrt{\log N\left(\mathcal{H}, \|\cdot\|_{2\to 2}, \varepsilon\right)}\, \mathrm{d}\varepsilon, \tag{7.27}$$

in the case of a set of matrices $\mathcal{H}$ equipped with the operator norm. Here and below $d_{\boxdot}(\mathcal{H}) = \sup_{\mathbf{H}\in\mathcal{H}} \|\mathbf{H}\|_{\boxdot}$, where $\boxdot$ is a generic norm.

Theorem 7.4.6. *Let $\mathcal{H}$ be a symmetric set of matrices, i.e., $\mathcal{H} = -\mathcal{H}$, and let $\boldsymbol{\xi}$ be a random vector whose entries ξ_i are independent $\mathcal{K}$-subgaussian random variables with mean 0 and variance 1. Set*

$$\begin{aligned} E &= \gamma_2\left(\mathcal{H}, \|\cdot\|_{2\to 2}\right)\left(\gamma_2\left(\mathcal{H}, \|\cdot\|_{2\to 2}\right) + d_F(\mathcal{H})\right) \\ V &= d_{2\to 2}\left(\mathcal{H}\right)\left(\gamma_2\left(\mathcal{H}, \|\cdot\|_{2\to 2}\right) + d_F(\mathcal{H})\right) \\ U &= d_{2\to 2}^2\left(\mathcal{H}\right) \end{aligned}$$

Then, for $t > 0$,

$$\mathsf{Pr}\left[\sup_{\mathbf{H}\in\mathcal{H}} \left| \|\mathbf{H}\boldsymbol{\xi}\|_{\ell_2}^2 - \mathsf{E}\left[\|\mathbf{H}\boldsymbol{\xi}\|_2^2\right] \right| \geq c_1 E + t\right] \leq 2\exp\left(-c_2 \min\left(\frac{t^2}{V^2}, \frac{t}{U}\right)\right).$$

The constants c_1 and c_2 are universal and only depend on $\mathcal{K}$.

We refer the reader to [107] and [102] for further details.

Proof of Lemma 7.4.2 : The proof consists of three main parts. We start in **(I)** by fitting our setting into the one of Theorem 7.4.6. In **(IIa)** resp. **(IIb)** the γ_2-functional gets bounded for $S_{s_1,s_2}^{R,\Gamma}$ and $K_{s_1,s_2}^{R,\Gamma}$, and in **(III)** we conclude by applying Theorem 7.4.6. Note that the computations of **(IIa)** and **(IIb)** can be found in the Appendix.

(I) We first switch the roles of our random measurement operator $\mathcal{A}$ applied to the fixed matrices $\mathbf{Z}$ to have fixed operators $\mathbf{H}_{\mathbf{Z}}$ applied to a random vector $\boldsymbol{\xi}$. Observe, for all $\mathbf{Z} \in \mathbb{R}^{n_1\times n_2}$, that

$$\begin{aligned} \mathcal{A}(\mathbf{Z}) &= \frac{1}{\sqrt{m}} \begin{pmatrix} \langle \mathrm{vec}(\mathbf{A}_1), \mathrm{vec}(\mathbf{Z})\rangle \\ \vdots \\ \langle \mathrm{vec}(\mathbf{A}_m), \mathrm{vec}(\mathbf{Z})\rangle \end{pmatrix} \\ &= \frac{1}{\sqrt{m}} \begin{pmatrix} \mathrm{vec}(\mathbf{Z})^T & 0 & \cdots \\ & \ddots & \\ \cdots & 0 & \mathrm{vec}(\mathbf{Z})^T \end{pmatrix} \cdot \begin{pmatrix} \mathrm{vec}(\mathbf{A}_1) \\ \vdots \\ \mathrm{vec}(\mathbf{A}_m) \end{pmatrix} = \mathbf{H}_{\mathbf{Z}} \cdot \boldsymbol{\xi}, \end{aligned}$$

where $\mathbf{H_Z} \in \mathbb{R}^{m \times mn_1n_2}$ is a matrix depending on $\mathbf{Z}$ and $\boldsymbol{\xi} \in \mathbb{R}^{mn_1n_2}$ has i.i.d. $\mathcal{K}$-subgaussian entries ξ_l of mean 0 and variance 1. We define $\mathcal{H}_S = \{\mathbf{H_Z} \colon \mathbf{Z} \in S_{s_1,s_2}^{R,\Gamma}\}$. Note that the mapping $\mathbf{Z} \mapsto \mathbf{H_Z}$ is an isometric linear bijection. In particular, we have $\|\mathbf{H_Z}\|_F = \|\mathbf{Z}\|_F$ and $\|\mathbf{H_Z}\|_{2\to 2} = \|\mathbf{Z}\|_F/\sqrt{m}$. For $\mathbf{Z} \in S_{s_1,s_2}^{R,\Gamma}$ it holds that $\|\mathbf{Z}\|_F \le \|\mathbf{U}\|_F \|\mathbf{\Sigma}\mathbf{V}^T\|_F \le \Gamma\sqrt{R}$. Hence, $d_F(\mathcal{H}_S) \le \Gamma\sqrt{R}$ and $d_{2\to 2}(\mathcal{H}_S) \le \Gamma\sqrt{R}/\sqrt{m}$.

(IIa) Since $\|\mathbf{H_Z}\|_{2\to 2} = \|\mathbf{Z}\|_F/\sqrt{m}$ and $\mathbf{Z} \mapsto \mathbf{H_Z}$ is a linear bijection, it follows that $N(\mathcal{H}_S, \|\cdot\|_{2\to 2}, \varepsilon) = N(S, \|\cdot\|_F, \sqrt{m}\varepsilon)$. We can estimate by (7.27) and Lemma B.1.1

$$\begin{aligned}
\gamma_2\left(\mathcal{H}_S, \|\cdot\|_{2\to 2}\right) &\lesssim \int_0^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \sqrt{\log N\left(\mathcal{H}_S, \|\cdot\|_{2\to 2}, \varepsilon\right)} dx\varepsilon \\
&= \int_0^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \sqrt{\log N\left(S_{s_1,s_2}^{R,\Gamma}, \|\cdot\|_F, \sqrt{m}\varepsilon\right)} d\varepsilon \\
&\le \sqrt{\frac{C_S \Gamma^2 R^2 (s_1+s_2) \log\left(\max\left\{n_1,n_2\right\}\right)}{m}} =: \mathcal{L}_S,
\end{aligned}$$

for some constant $C_S > 0$.

(IIb) In the same manner we obtain a bound on $\gamma_2(\mathcal{H}_K, \|\cdot\|_{2\to 2})$ where $\mathcal{H}_K = \{\mathbf{H_Z} \colon \mathbf{Z} \in K_{s_1,s_2}^{R,\Gamma}\}$. Recall that $\|\mathbf{H_Z}\|_F = \|\mathbf{Z}\|_F$, $\|\mathbf{H_Z}\|_{2\to 2} = \|\mathbf{Z}\|_F/\sqrt{m}$ and $\mathbf{Z} \mapsto \mathbf{H_Z}$ is an linear bijection. This implies $N(\mathcal{H}_K, \|\cdot\|_{2\to 2}, \varepsilon) = N(K_{s_1,s_2}^{R,\Gamma}, \|\cdot\|_F, \sqrt{m}\varepsilon)$. Note that $d_F(\mathcal{H}_K) \le \Gamma\sqrt{R}$ and $d_{2\to 2}(\mathcal{H}_K) \le \Gamma\sqrt{R}/\sqrt{m}$. We obtain by (7.27) and Lemma B.1.1

$$\begin{aligned}
\gamma_2(\mathcal{H}_K, \|\cdot\|_{2\to 2}) &\lesssim \int_0^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \sqrt{\log N(\mathcal{H}_K, \|\cdot\|_{2\to 2}, \varepsilon)}\, d\varepsilon \\
&= \int_0^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \sqrt{\log N(K_{s_1,s_2}^{R,\Gamma}, \|\cdot\|_F, \sqrt{m}\varepsilon)}\, dx\varepsilon \\
&\le \sqrt{\frac{C_K \Gamma^2 R^2 (s_1+s_2) \log^3(\max\{n_1,n_2\})}{m}} =: \mathcal{L}_K,
\end{aligned}$$

for some constant $C_K > 0$.

(III) The final part of the proof is now equal for both sets $S_{s_1,s_2}^{R,\Gamma}$ and $K_{s_1,s_2}^{R,\Gamma}$. We write $\mathcal{L}$ for $\mathcal{L}_S$ resp. $\mathcal{L}_K$ and assume $m \gtrsim C_S d^{-2} R(s_1+s_2) \log\left(\max\left\{n_1,n_2\right\}\right)$ resp. $m \gtrsim C_K d^{-2} R(s_1+s_2) \log^3(\max\{n_1,n_2\})$, for some $0 < d < 1$. Then, $\mathcal{L} \le \Gamma\sqrt{R}d$ and

$$\mathcal{L}^2 + \Gamma\sqrt{R}\mathcal{L} \le \Gamma^2 R(d^2+d) \le 2\Gamma^2 Rd. \tag{7.28}$$

We obtain the following bounds on the quantities (cf. Theorem 7.4.6):

$$E \le \mathcal{L}^2 + \Gamma\sqrt{R}\mathcal{L}, \qquad V \le \frac{\Gamma\sqrt{R}\mathcal{L} + \Gamma^2 R}{\sqrt{m}}, \qquad U \le \frac{\Gamma^2 R}{m}. \tag{7.29}$$

Observing now that $\mathsf{E}\left[\|\mathbf{H}_{\mathbf{Z}}\boldsymbol{\xi}\|_2^2\right] = \|\mathbf{H}_{\mathbf{Z}}\|_F^2 = \|\mathbf{Z}\|_F^2$ and recalling $\Gamma \geq 1$ we finally get, for $\delta \geq 3c_1\Gamma^2 Rd$ (which implies by (7.28) that $\delta \geq c_1 E + c_1\Gamma^2 Rd$),

$$\begin{aligned}
&\mathsf{Pr}\left[\sup_{\mathbf{Z}\in S} \left|\|\mathcal{A}(\mathbf{Z})\|_2^2 - \|\mathbf{Z}\|_F^2\right| \geq \delta\right] \\
&\leq \ \mathsf{Pr}\left[\sup_{\mathbf{H}_{\mathbf{Z}}\in\mathcal{H}} \left|\|\mathbf{H}_{\mathbf{Z}}\boldsymbol{\xi}\|_2^2 - \ \mathsf{E}\left[\|\mathbf{H}_{\mathbf{Z}}\boldsymbol{\xi}\|_2^2\right]\right| \geq c_1 E + c_1\Gamma^2 Rd\right] \\
&\leq 2\exp\left(-c_2 \min\left\{m\frac{c_1^2\Gamma^4R^2d^2}{\Gamma^2R(\mathcal{L}+\Gamma\sqrt{R})^2}, m\frac{c_1\Gamma^2Rd}{\Gamma^2R}\right\}\right) \\
&\leq 2\exp\left(-Cd^2m\right),
\end{aligned}$$

where $C > 0$ is a positive constant which depends on $\mathcal{K}$. In the last step we used that $\mathcal{L} + \Gamma\sqrt{R} \in [\Gamma\sqrt{R}, 2\Gamma\sqrt{R}]$ (because $0 < \mathcal{L} < \Gamma\sqrt{R}$). ■

7.5 Approximation Under RIP Assumptions

We are ready state an approximation result which we introduce as follows: If one assumes RIP, any appropriate global minimizer of $J_{\alpha,\beta}^R$ provides a good approximation to $\hat{\mathbf{X}}$ depending on the magnitude of α and β, the sparsity s, the RIP constant δ and the magnitude of $\hat{\mathbf{X}}$ measured in an appropriate Schatten quasi-norm. The approximation is worsened in an additive way by noise level. For a given minimizer $(\mathbf{u}_{\alpha,\beta}^1, ..., \mathbf{u}_{\alpha,\beta}^R, \mathbf{v}_{\alpha,\beta}^1, ..., \mathbf{v}_{\alpha,\beta}^R)$ of $J_{\alpha,\beta}^R$ we denote

$$\mathbf{X}_{\alpha,\beta} = \mathbf{U}_{\alpha,\beta}\mathbf{\Sigma}_{\alpha,\beta}\mathbf{V}_{\alpha,\beta}^T = \sum_{r=1}^{R} (\sigma_{\alpha,\beta})_r \frac{\mathbf{u}_{\alpha,\beta}^r}{\|\mathbf{u}_{\alpha,\beta}^r\|_2}\left(\frac{\mathbf{v}_{\alpha,\beta}^r}{\|\mathbf{v}_{\alpha,\beta}^r\|_2}\right)^T, \tag{7.30}$$

where $(\sigma_{\alpha,\beta})_r = \|\mathbf{u}_{\alpha,\beta}^r\|_2\|\mathbf{v}_{\alpha,\beta}^r\|_2$, for all $r \in [R]$, and $\mathbf{\Sigma}_{\alpha,\beta}$ is the diagonal matrix defined by the vector $\sigma_{\alpha,\beta}$.

Theorem 7.5.1. *Fix the positive constants $\alpha, \beta > 0$, $\Gamma \geq 1$, and the effective sparsity indicator level $1 \leq s \leq n_2$. Let $\mathcal{A}$ have the additive rank-$2R$ effectively $(n_1, \max\{s, (\gamma/\beta)^2\})$-sparse RIP$_{(c+1)\Gamma}$ with RIP-constant $0 < \delta < 1$, for a fixed choice of $\gamma > 0$ and $c \geq 1$. If $\hat{\mathbf{X}} \in K_{n_1,s}^{R,\Gamma}$ of rank R and $\mathbf{y} = \mathcal{A}(\hat{\mathbf{X}}) + \boldsymbol{\eta} \in \mathbb{R}^m$, then*

$$\|\hat{\mathbf{X}} - \mathbf{X}_{\alpha,\beta}\|_F \leq \sqrt{s^{\frac{1}{3}}R^{\frac{2}{3}}C_{2,1}c_{\hat{U}}\sqrt[6]{\alpha\beta^2}\|\hat{\mathbf{X}}\|_{\frac{2}{3}}^{\frac{1}{3}}} + 2\|\boldsymbol{\eta}\|_2 + \sqrt{\delta}, \tag{7.31}$$

for any global minimizer $(\mathbf{u}_{\alpha,\beta}^1, ..., \mathbf{v}_{\alpha,\beta}^R)$ of $J_{\alpha,\beta}^R$ that fulfills $\|\mathbf{v}_{\alpha,\beta}^r\|_2 \geq (\|\hat{\mathbf{X}}\|_F + \|\boldsymbol{\eta}\|_2 + \sqrt{\delta})^2/\gamma$ for all $r \in [R]$ and $\|\boldsymbol{\sigma}_{\alpha,\beta}\|_F \leq c\Gamma$ in (7.30). In this case, in particular, $\mathbf{X}_{\alpha,\beta} \in K_{n_1,(\gamma/\beta)^2}^{R,c\Gamma}$ with the SD in (7.30).

Proof: As $\|\mathbf{y}\|_2 \leq \|\mathcal{A}(\hat{\mathbf{X}})\|_2 + \|\boldsymbol{\eta}\|_2 \leq (\|\mathbf{X}\|_F + \sqrt{\delta}) + \|\boldsymbol{\eta}\|_2$, Lemma 7.3.4 applies and yields that $\mathbf{X}_{\alpha,\beta}$ is in $K_{n_1,(\gamma/\beta)^2}^{R,c\Gamma}$. Combined with $\hat{\mathbf{X}} \in K_{n_1,s}^{R,\Gamma}$, we know from

(7.20) that the difference $\hat{\mathbf{X}} - \mathbf{X}_{\alpha,\beta} \in K^{2R,(c+1)\Gamma}_{n_1,\max\{s,(\gamma/\beta)^2\}}$. Hence, we apply the rank-$2R$ and effectively $(n_1, \max\{s, (\gamma/\beta)^2\})$-sparse $\mathrm{RIP}_{(c+1)\Gamma}$ of $\mathcal{A}$ to obtain (note that $|a^2 - b^2| \leq \delta$ implies $|a - b| \leq \sqrt{\delta}$, for $a, b > 0$)

$$
\begin{aligned}
\|\hat{\mathbf{X}} - \mathbf{X}_{\alpha,\beta}\|_F &\leq \|\mathcal{A}(\hat{\mathbf{X}}) - \mathcal{A}(\mathbf{X}_{\alpha,\beta})\|_2 + \sqrt{\delta} \leq (\|\mathbf{y} - \mathcal{A}(\mathbf{X}_{\alpha,\beta})\|_2 + \|\boldsymbol{\eta}\|_2) + \sqrt{\delta} \\
&\leq \sqrt{s^{\frac{1}{3}} R^{\frac{2}{3}} C_{2,1} c_{\hat{\mathbf{U}}} \sqrt[3]{\alpha\beta^2} \|\hat{\mathbf{X}}\|_{\frac{2}{3}}^{\frac{2}{3}} + \|\boldsymbol{\eta}\|_2^2} + \|\boldsymbol{\eta}\|_2 + \sqrt{\delta} \\
&\leq \sqrt{s^{\frac{1}{3}} R^{\frac{2}{3}} C_{2,1} c_{\hat{\mathbf{U}}}} \sqrt[6]{\alpha\beta^2} \|\hat{\mathbf{X}}\|_{\frac{2}{3}}^{\frac{1}{3}} + 2\|\boldsymbol{\eta}\|_2 + \sqrt{\delta}.
\end{aligned}
$$

In the third inequality we used Proposition 7.3.1 in combination with $\|\hat{\mathbf{v}}^r\|_1 \leq \sqrt{s}\|\hat{\mathbf{v}}^r\|_2$ and

$$
\sum_{r=1}^{R} (\|\hat{\mathbf{u}}^r\|_2 \|\hat{\mathbf{v}}^r\|_1)^{\frac{2}{3}} \leq s^{\frac{1}{3}} \sum_{r=1}^{R} (\|\hat{\mathbf{u}}^r\|_2 \|\hat{\mathbf{v}}^r\|_2)^{\frac{2}{3}} \leq c_{\hat{\mathbf{U}}} R^{\frac{2}{3}} s^{\frac{1}{3}} \|\hat{\mathbf{X}}\|_{\frac{2}{3}}^{\frac{2}{3}},
$$

where we used again (7.12) for $p = 2/3$. ■

There are some aspects of this result we would like to stress:

If we could take the limits $\alpha \to 0$ and $\beta \to 0$, the error in (7.31) would vanish up to noise-level and RIP-constant. However, this limit cannot be performed as there are important restrictions dictated by the need of fulfilling simultaneously the RIP and the assumptions on $\mathbf{X}_{\alpha,\beta}$. If β is getting small the conditions for having RIP degenerate, i.e., reconstruction for a fixed number of measurements only works up to a minimal β. Letting α to zero while keeping β fixed leads to minimizers which violate the lower bound on $\|\mathbf{v}^r_{\alpha,\beta}\|_2$ or the upper bound on $\|\boldsymbol{\sigma}_{\alpha,\beta}\|_2$. To see this, note that by Lemma 7.3.3 small α leads to strict bounds on $\|\mathbf{v}^r_{\alpha,\beta}\|_2$ and weak bounds on $\|\mathbf{u}^r_{\alpha,\beta}\|_2$.

If $\hat{\mathbf{X}} \in K^{R,\Gamma}_{n_1,s}$ and the SD of $\hat{\mathbf{X}}$ coincides with its SVD, then in view of the identity (7.11) the factor $c_{\hat{\mathbf{U}}} R^{2/3}$ in the error estimates (7.31) and (7.32) can be substituted by 1, hence there would be no dependence on the rank R.

In order to clarify how $(\gamma/\beta)^2$ and s are related in the RIP in Theorem 7.5.1 and Corollary 7.5.2, let us assume for simplicity that the SD of $\hat{\mathbf{X}}$ coincides with its SVD and $\alpha = \beta$. Consequently, to get a meaningful approximation result, in (7.31) α and β have to be chosen of order $\mathcal{O}(s^{-\frac{1}{3}})$, i.e., $(\gamma/\beta)^2$ is of order $\mathcal{O}(s^{\frac{2}{3}})$ which means that an $(n_1, \gamma^2 s)$-sparse $\mathrm{RIP}_{(c+1)\Gamma}$ is sufficient for recovery.

The result only applies to minimizers whose scaling matrix $\mathbf{\Sigma}_{\alpha,\beta}$ is bounded in Frobenius norm and whose right components $\mathbf{v}^r_{\alpha,\beta}$ are not too close to zero. The first requirement is necessary as the RIP is restricted to SDs with scaling matrices within a ball around zero. The second one is needed to show some level of effective sparsity of the minimizers $\mathbf{X}_{\alpha,\beta}$. While effective sparsity of (right) component vectors of $\mathbf{X}_{\alpha,\beta}$ is naturally wished and expected if $\hat{\mathbf{X}} \in K^{R,\Gamma}_{n_1,s}$, we were not able in all cases to show *exact* sparsity of (right) component vectors of $\mathbf{X}_{\alpha,\beta}$ if $\hat{\mathbf{X}} \in S^{R,\Gamma}_{n_1,s}$, but again only their effective sparsity. Hence, we are bound to using as an artifact of the proof the stronger effectively (s_1, s_2)-sparse RIP_Γ for theoretical analysis also in this case. In numerical experiments, however, for $\hat{\mathbf{X}} \in S^{R,\Gamma}_{n_1,s}$ the obtained minimizers $\mathbf{X}_{\alpha,\beta}$ are empirically *exactly* sparse (not just effectively sparse)

and, hence, the weaker rank-$2R$ (s_1, s_2)-sparse RIP$_\Gamma$ might suffice in practice. The latter can already be guaranteed for a smaller number of measurements.
As the reader may notice, all technical results of Section 7.3 can be adapted to effective sparsity on the left components $(\mathbf{u}^1_{\alpha,\beta}, ..., \mathbf{u}^R_{\alpha,\beta})$. This can be done by replacing ℓ_2-norms by corresponding ℓ_1-norms in $J^R_{\alpha,\beta}$. The proof of Lemma 7.3.4, which guarantees effective sparsity of the right components, is independent of the minimization of the left components. Therefore, Lemma 7.3.4 applies also to the left components if ℓ_2-norms are replaced by ℓ_1-norms in $J^R_{\alpha,\beta}$. Theorem 7.5.1 then can be adapted to this setting in a straightforward way.
It is important to require $\operatorname{rank}(\hat{\mathbf{X}}) = R$ as otherwise the equivalence of Schatten-norm and normed SD cannot be guaranteed as (7.12). If the SD of $\hat{\mathbf{X}}$ coincides with its SVD though, the rank condition may be dropped.

By choosing α and β in relation to the noise-to-signal ratio $\|\boldsymbol{\eta}\|_2^2/\|\hat{\mathbf{X}}\|_{\frac{2}{3}}^{\frac{2}{3}}$ we obtain the following version of Theorem 7.5.1, which has the form of a typical compressed sensing recovery bound. Assuming the RIP, the approximation error is linear in noise level while the slope of the linear function depends on sparsity level and possibly the rank. However, for fixed number of measurements the RIP fails for exceedingly small noise. Hence, the result is valid only for sufficiently small signal-to-noise ratio. As we will show in Section 7.7, this apparently counter-intuitive result is factual and not an artifact of the proof technique. A possible intuitive explanation is that $J^R_{\alpha,\beta}$ becomes a mere least-squares without sparsifying effect for α and β close to zero, which is caused by vanishing noise.

Corollary 7.5.2. *Let* $\hat{\mathbf{X}} \in K^{R,\Gamma}_{n_1,s}$ *with* $\operatorname{rank}(\hat{\mathbf{X}}) = R$ *fulfill the noisy measurements* $\mathbf{y} = \mathcal{A}(\hat{\mathbf{X}}) + \boldsymbol{\eta}$ *and let* $\alpha = \beta = \|\boldsymbol{\eta}\|_2^2/\|\hat{\mathbf{X}}\|_{\frac{2}{3}}^{\frac{2}{3}} < 1$. *Assume* $\mathcal{A}$ *has for some* $\gamma > 0$ *and* $c \geq 1$ *the additive rank-*$2R$ *effectively* $\left(n_1, \max\{s, \gamma^2(\|\hat{\mathbf{X}}\|_{\frac{2}{3}}^{\frac{2}{3}}/\|\boldsymbol{\eta}\|_2^2)^2\}\right)$*-sparse RIP*$_{(c+1)\Gamma}$ *with RIP-constant* $0 < \delta < 1$. *Then, for* $\mathbf{X}_{\alpha,\beta}$ *with* $\|\boldsymbol{\Sigma}_{\alpha,\beta}\|_F \leq c\Gamma$ *and* $\|\mathbf{v}^r_{\alpha,\beta}\|_2 \geq (\|\hat{\mathbf{X}}\|_F + \|\boldsymbol{\eta}\|_2 + \sqrt{\delta})^2/\gamma$, $r \in [R]$, *we have*

$$\|\hat{\mathbf{X}} - \mathbf{X}_{\alpha,\beta}\|_F \leq \left(2\sqrt{c_{\hat{\mathbf{U}}} R^{2/3} s^{1/3}} + 2\right) \|\boldsymbol{\eta}\|_2 + \sqrt{\delta}. \tag{7.32}$$

Remark 7.5.3. *One could object that the simple zero solution* $\bar{\mathbf{X}} = 0$ *is already a competitor in case of large noise* $\|\boldsymbol{\eta}\|_2 \geq \Xi(m)\|\hat{\mathbf{X}}\|_F$, *i.e.,*

$$\|\hat{\mathbf{X}} - \bar{\mathbf{X}}\|_F \leq \Xi(m)^{-1}\|\boldsymbol{\eta}\|_2. \tag{7.33}$$

However, for a larger number m *of measurements we can consider lower level of noise, i.e.,* $\Xi(m) \to 0$ *and the bound* (7.33) *would explode, while* (7.32) *would remain effective.*

7.6 Local Convergence of ATLAS

So far an important question has not been posed. The above results only apply to global minimizers of $J^R_{\alpha,\beta}$ which is a highly non-convex functional. One wonders if the alternating minimization defined in (A-T-LAS$_{2,1}$) is able to provide minimizers. We give a partial answer to this issue. By adapting results of Attouch et. al. in [7] we show convergence of ATLAS and that there is a neighborhood $\mathcal{U}_{(\mathbf{u}^1_{\alpha,\beta},...,\mathbf{v}^R_{\alpha,\beta})}$ of a global minimizer $(\mathbf{u}^1_{\alpha,\beta},...,\mathbf{v}^R_{\alpha,\beta})$ such that the sequence $(\mathbf{u}^1_k,...,\mathbf{v}^R_k)$ defined by (A-T-LAS$_{2,1}$) converges to $(\mathbf{u}^1_{\alpha,\beta},...,\mathbf{v}^R_{\alpha,\beta})$ of $J^R_{\alpha,\beta}$ if the initialization lies within $\mathcal{U}_{(u^1_{\alpha,\beta},...,v^R_{\alpha,\beta})}$. However, we do not give proof for any initialization to fulfill the requirement. This is an open issue for future research, but recent promising results [68, 67] may shed light on how to attack the problem also for ATLAS. The techniques in [7] provide tools to analyze the rate of convergence of ATLAS as well. However, additional work is necessary to estimate the appearing parameters for $J^R_{\alpha,\beta}$.

We begin by a generalization of the basic conditions of [7]. Let L be a functional of the following form:

$$(H)\begin{cases} L(\mathbf{u}^1,\dots,\mathbf{u}^R,\mathbf{v}^1,\dots,\mathbf{v}^R)=\sum_{r=1}^R f_r(\mathbf{u}^r)+Q(\mathbf{u}^1,\dots,\mathbf{v}^R)+\sum_{r=1}^R g_r(\mathbf{v}^r),\\ f_r:\mathbb{R}^{n_1}\to\mathbb{R}\cup\{\infty\},\ g_r:\mathbb{R}^{n_2}\to\mathbb{R}\cup\{\infty\} \text{ are proper lower semicontinuous, for } 1\le r\le R,\\ Q:\mathbb{R}^{n_1}\times\dots\times\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\times\dots\times\mathbb{R}^{n_2}\to\mathbb{R} \text{ is a } C^1 \text{ function,}\\ \nabla Q \text{ is Lipschitz continuous on bounded subsets of } \mathbb{R}^{n_1}\times\dots\times\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\times\dots\times\mathbb{R}^{n_2}. \end{cases}$$

For given $(\mathbf{u}^1_0,\dots,\mathbf{v}^R_0)\in(\mathbb{R}^{n_1})^R\times(\mathbb{R}^{n_2})^R$ and fixed sequences $(\lambda^1_k)_{k\in\mathbb{N}},\dots,(\lambda^R_k)_{k\in\mathbb{N}}$, $(\mu^R_k)_{k\in\mathbb{N}},\dots,(\mu^R_k)_{k\in\mathbb{N}}$ assume that

$$(H_1)\begin{cases} \inf L>-\infty,\\ L(\cdot,\mathbf{u}^2_0,\dots,\mathbf{v}^R_0) \text{ is proper,}\\ \text{for some positive } r_-<r_+ \text{ the sequences } \lambda^1_k,\dots,\mu^R_k \text{ belong to } (r_-,r_+). \end{cases}$$

The adapted main result of [7] guarantees convergence of the so-called Proximal Alternating Minimization

$$\text{(PAM)}\begin{cases} \mathbf{u}^1_{k+1}=\arg\min_{\mathbf{u}\in\mathbb{R}^{n_1}} L(\mathbf{u},\mathbf{u}^2_k,\dots,\mathbf{u}^R_k,\mathbf{v}^1_k,\dots,\mathbf{v}^R_k)+\frac{1}{2\lambda^1_k}\|\mathbf{u}-\mathbf{u}^1_k\|^2_2,\\ \mathbf{v}^1_{k+1}=\arg\min_{\mathbf{v}\in\mathbb{R}^{n_2}} L(\mathbf{u}^1_{k+1},\mathbf{u}^2_k,\dots,\mathbf{u}^R_k,\mathbf{v},\mathbf{v}^2_k\dots,\mathbf{v}^R_k)+\frac{1}{2\mu_k}\|\mathbf{v}-\mathbf{v}^1_k\|^2_2,\\ \vdots\\ \mathbf{u}^R_{k+1}=\arg\min_{\mathbf{u}\in\mathbb{R}^{n_1}} L(\mathbf{u}^1_{k+1},\dots,\mathbf{u}^{R-1}_{k+1},\mathbf{u},\mathbf{v}^1_{k+1},\dots,\mathbf{v}^{R-1}_{k+1},\mathbf{v}^R_k)+\frac{1}{2\lambda_k}\|\mathbf{u}-\mathbf{u}^R_k\|^2_2,\\ \mathbf{v}^R_{k+1}=\arg\min_{\mathbf{v}\in\mathbb{R}^{n_2}} L(\mathbf{u}^1_{k+1},\dots,\mathbf{u}^R_{k+1},\mathbf{v}^1_{k+1},\dots,\mathbf{v}^{R-1}_{k+1},\mathbf{v})+\frac{1}{2\mu_k}\|\mathbf{v}-\mathbf{v}^R_k\|^2_2, \end{cases} \tag{7.34}$$

to a stationary point of L if L fulfills (H), (H_1), and the so called Kurdyka-Lojasiewicz property, which requires L to behave well around stationary points. If the initialization $(\mathbf{u}^1_0,\dots,\mathbf{v}^R_0)$ of (PAM) lies, in addition, sufficiently close to a global minimizer $(\mathbf{u}^1_*,\dots,\mathbf{v}^R_*)$ of L, (PAM) converges to a global minimizer of L.

Definition 7.6.1 (Kurdyka-Lojasiewicz Property)**.** *A proper lower semicontinuous function $f:\mathbb{R}^n\to\mathbb{R}\cup\{\infty\}$ is said to have the Kurdyka-Lojasiewicz property (KL-property) at $\bar{\mathbf{x}}\in\operatorname{dom}\partial f$ (here ∂f denotes the subdifferential of f and $\operatorname{dom}\partial f$ the domain on which ∂f takes finite values) if there exist $\eta\in(0,\infty]$, a neighborhood U of $\bar{\mathbf{x}}$ and a continuous concave function $\varphi:[0,\infty)\to\mathbb{R}_+$ such that*

- $\varphi(0) = 0$,
- *φ is C^1 on $(0, \eta)$,*
- *$\varphi'(t) > 0$, for all $t \in (0, \eta)$,*
- *and, for all $\mathbf{x} \in U \cap \{\mathbf{x} \in \mathbb{R}^n : f(\bar{\mathbf{x}}) < f(\mathbf{x}) < f(\bar{\mathbf{x}}) + \eta\}$, the KL-inequality holds:*

$$\varphi'(f(\mathbf{x}) - f(\bar{\mathbf{x}}))\ \operatorname{dist}(\mathbf{0}, \partial f(\mathbf{x})) \geq 1.$$

Theorem 7.6.2. *Assume that L satisfies (H) and (H_1). If L has the Kurdyka-Lojasiewicz property at its global minimizer $(\mathbf{u}_*^1, \dots, \mathbf{v}_*^R)$, then there exist $\varepsilon, \eta > 0$, such that the initial conditions*

$$\|(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) - (\mathbf{u}_*^1, \dots, \mathbf{v}_*^R)\|_2 < \varepsilon, \quad \min L < L(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) < \min L + \eta,$$

imply that the iterations $(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$ generated by (PAM) converge to a point in $\arg\min L$. *If L has the Kurdyka-Lojasiewicz property at each point of its domain, then, independent of initialization, either $\|(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)\|_2 \to \infty$ or $(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$ converges to a stationary point of L.*

By applying Theorem 7.6.2 to $L = J_{\alpha,\beta}^R$ and ATLAS we obtain convergence to stationary points and local convergence to global minimizers as the sequence $(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$ is bounded by coercivity of $J_{\alpha,\beta}^R$. One can check that conditions (H), (H_1) are fulfilled by $J_{\alpha,\beta}^R$ and ATLAS for a suitable choice of the sequences $(\lambda_k^1)_{k\in\mathbb{N}}, \dots, (\lambda_k^R)_{k\in\mathbb{N}}$, $(\mu_k^R)_{k\in\mathbb{N}}, \dots, (\mu_k^R)_{k\in\mathbb{N}}$. It remains to validate the KL-property. As mentioned in [7, Section 4.3], all semialgebraic functions satisfy the KL-property at each point with $\varphi(t) = ct^{1-\theta}$ for some $\theta \in [0,1) \cap \mathbb{Q}$ and $c > 0$. Hence, by showing that $J_{\alpha,\beta}^R$ is semialgebraic, we get the KL-property for free. But we pay the price of having no better knowledge on the parameters ε and η in Theorem 7.6.2, which characterize the convergence radius. Let us conclude by showing that $J_{\alpha,\beta}^R$ is semialgebraic, i.e., $\operatorname{graph}(J_{\alpha,\beta}^R) \subset \mathbb{R}^{Rn_1+Rn_2} \times \mathbb{R}$ is a semialgebraic set.

A set in $\mathbb{R}^n$ is called semialgebraic if it can be written as a finite union of sets of the form

$$\{\mathbf{x} \in \mathbb{R}^n \ : \ p_i(\mathbf{x}) = 0,\ q_i(\mathbf{x}) > 0,\ i = 1, \dots, k\},$$

where p_i, q_i are real polynomials and $k \in \mathbb{N}$. First, the absolute value of one component of a vector $h(\mathbf{x}) := |x_l|$ is a semialgebraic function as

$$\begin{aligned}\operatorname{graph}(h) = \{(\mathbf{x}, r) \in \mathbb{R}^n \times \mathbb{R} : x_i + r = 0,\ x_i < 0\} \cup \{(\mathbf{x}, r) \in \mathbb{R}^n \times \mathbb{R} : x_i = 0,\ r = 0\} \\ \cup \{(\mathbf{x}, r) \in \mathbb{R}^n \times \mathbb{R} : x_i - r = 0,\ -x_i < 0\}.\end{aligned}$$

Second, it is clear that polynomials p are semialgebraic as $\operatorname{graph}(p) = \{(\mathbf{x}, r) \in \mathbb{R}^n \times \mathbb{R} : p(\mathbf{x}) - r = 0\}$ and, third, composition, finite sums and finite products of semialgebraic functions are semialgebraic. The semialgebraicity of $J_{\alpha,\beta}^R$ follows as

$$J_{\alpha,\beta}^R(\mathbf{u}^1, \dots, \mathbf{v}^R) = \sum_{l=1}^{m} |y_l - \sum_{r=1}^{R} \langle \mathbf{A}_l, \mathbf{u}^r \mathbf{v}^{rT} \rangle_F|^2 + \alpha \sum_{r=1}^{R} \sum_{l=1}^{n_1} |u_l^r|^2 + \beta \sum_{r=1}^{R} \sum_{l=1}^{n_2} |v_l^r|$$

is just a finite composition of semialgebraic basic units.

If one can estimate the Kurdyka-Lojasiewicz parameters, the above considerations characterize convergence rates of ATLAS as the following theorem states.

Theorem 7.6.3. *Assume that L satisfies (H) and (H_1). Assume further that $(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$ converges to $(\mathbf{u}_\infty^1, \dots, \mathbf{v}_\infty^R)$ and L has the Kurdyka-Lojasiewicz property at $(\mathbf{u}_\infty^1, \dots, \mathbf{v}_\infty^R)$ with $\varphi(t) = ct^{1-\theta}$, for $\theta \in [0,1)$ and $c > 0$. Then the following hold:*

(i) If $\theta = 0$, the sequence $(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$ converges in a finite number of steps.

(ii) If $\theta \in (0, \frac{1}{2}]$, there exist $c' > 0$ and $\tau \in [0,1)$ such that

$$\left\|(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R) - (\mathbf{u}_\infty^1, \dots, \mathbf{v}_\infty^R\right\|_2 \leq c'\tau^k.$$

(iii) If $\theta \in (\frac{1}{2}, 1)$, there exists $c' > 0$ such that

$$\left\|(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R) - (\mathbf{u}_\infty^1, \dots, \mathbf{v}_\infty^R\right\|_2 \leq c'k^{-\frac{1-\theta}{2\theta-1}}.$$

We refrain from presenting the proofs of Theorem 7.6.2 and 7.6.3 here as they are straight-forward modifications of the arguments in [7]. The interested reader finds them in the Appendix.

7.7 Numerical Simulation

After having obtained some theoretical insight on the proposed optimization problem, we provide an implementation of ATLAS and discuss its predicted behavior in numerical experiments. Therefore, we begin by presenting the implementation that has been used in all experiments. As in practice ATLAS converges even without the auxiliary terms introduced in (A-T-LAS$_{2,1}$), for sake of simplicity we drop those terms. By the alternating form of ATLAS one must solve several Tikhonov regularization resp. ℓ_1-LASSO problems. Note that for the Tikhonov regularization

$$\mathbf{u} = \arg\min_{\mathbf{z} \in \mathbb{R}^n} \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2 + \alpha \|\mathbf{z}\|_2^2,$$

with $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{y} \in \mathbb{R}^m$, and $\alpha > 0$, the solution is explicitly given by $\mathbf{u} = (\alpha \mathbf{Id} + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$. Solutions to ℓ_1-LASSO

$$\mathbf{v} = \arg\min_{\mathbf{z} \in \mathbb{R}^n} \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2 + \beta \|\mathbf{z}\|_1,$$

for some $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{y} \in \mathbb{R}^m$ and $\beta > 0$ can be well approximated by Iterative Soft-Thresholding Algorithm (ISTA) as seen in Section 2.3.3. Hence, a suitable implementation of ATLAS is given in Algorithm 9. Necessary modifications in case of sparse left component vectors of $\hat{\mathbf{X}}$ are straightforward.

Let us turn toward numerical simulations. First, we check if the approximation results stated in Theorem 7.5.1 and Corollary 7.5.2 describe the qualitative and quantitative behavior of the approximation error well. Then, we compare ATLAS to Sparse Power Factorization (SPF), see Section 7.1. We used the leading singular vectors of $\mathcal{A}^*(\mathbf{y})$ to initialize both algorithms, which is likely not an optimal choice and certainly may cause loss of performance for both algorithms, but it is nevertheless sufficient to illustrate certain comparisons numerically.

Algorithm 9 : ATLAS$(\mathbf{y}, \mathcal{A}, R, \mathbf{v}_0^1, ..., \mathbf{v}_0^R, \alpha, \beta, L)$

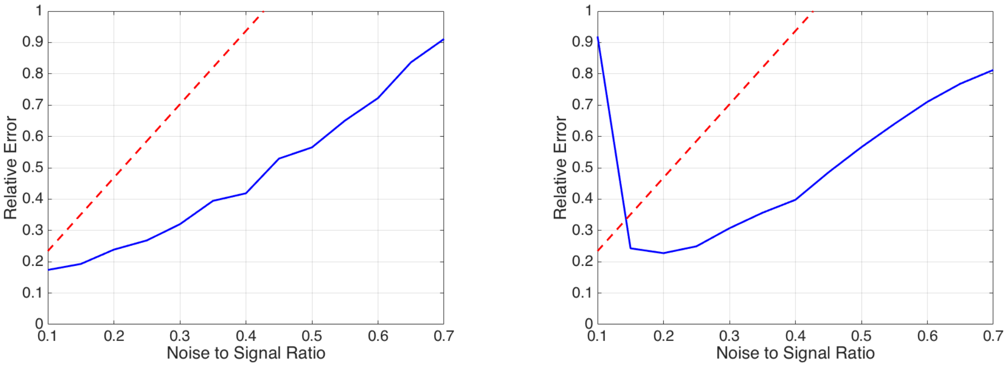
```
Require: $\mathbf{y} \in \mathbb{R}^m$, $\mathcal{A}\colon \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$, rank $R$, $\mathbf{v}_0^1, ..., \mathbf{v}_0^R \in \mathbb{R}^{n_2}$, $\alpha, \beta > 0$, and number of
    iterations $L$
$l = 0$                                                              ▷ initialize

while $l < L$ do
    for $r = 1, ..., R$ do
        $\tilde{\mathbf{y}} = \mathbf{y} - \mathcal{A}\left(\sum_{\tilde{r}<r} \mathbf{u}_l^{\tilde{r}}(\mathbf{v}_l^{\tilde{r}})^T + \sum_{\tilde{r}>r} \mathbf{u}_{l-1}^{\tilde{r}}(\mathbf{v}_{l-1}^{\tilde{r}})^T\right)$      ▷ Fix $\mathbf{u}^{\tilde{r}}$, $\mathbf{v}^{\tilde{r}}$ with $\tilde{r} \neq r$
        $\mathbf{u}_l^r = \left(\alpha \mathbf{Id} + \mathbf{A}_\mathbf{v}(\mathbf{v}_{l-1}^r)^T \mathbf{A}_\mathbf{v}(\mathbf{v}_{l-1}^r)\right)^{-1} \mathbf{A}_\mathbf{v}(\mathbf{v}_{l-1}^r)^T \tilde{\mathbf{y}}$      ▷ $\mathcal{A}(\mathbf{u}\mathbf{v}^T) = \mathbf{A}_\mathbf{v}(\mathbf{v}) \cdot \mathbf{u}$
        $\mathbf{v}_l^r = \mathbf{ISTA}(\tilde{\mathbf{y}}, \mathbf{A}_\mathbf{u}(\mathbf{u}_l^r), \mathbf{v}_{l-1}^r, \beta)$      ▷ $\mathcal{A}(\mathbf{u}\mathbf{v}^T) = \mathbf{A}_\mathbf{u}(\mathbf{u}) \cdot \mathbf{v}$
    end for
end while
return $(\mathbf{u}_{\mathrm{ATLAS}}^1, ..., \mathbf{v}_{\mathrm{ATLAS}}^R) = (\mathbf{u}_L^1, ..., \mathbf{v}_L^R)$
```

7.7.1 Validation of Corollary 7.5.2

Figure 7.1 shows the average approximation error of 100 randomly drawn $\hat{\mathbf{X}} \in \mathbb{R}^{16\times 100}$, $\|\hat{\mathbf{X}}\|_F = 10$, with $\operatorname{rank}(\hat{\mathbf{X}}) = 1$ (resp. $\operatorname{rank}(\hat{\mathbf{X}}) = 5$) and 10-sparse right singular vector(s) from $m = 90$ (resp. $m = 400$) noisy measurements $\mathbf{y} = \mathcal{A}(\hat{\mathbf{X}}) + \boldsymbol{\eta}$. The parameters have been chosen exemplarily for purpose of illustration. The operator $\mathcal{A}$ is drawn once at random. The error bound from Corollary 7.5.2 is plotted as dashed red line, whereas the average approximation errors are in blue. Though not tight, the theoretical bound seems to describe the linear dependence of the approximation error on noise level appropriately. In addition, Figure 7.1 (b) shows a breakdown of approximation for noise to signal ratios below ≈ 0.25. This occurrence is not surprising as the assumptions of Corollary 7.5.2 include a lower-bound on the noise-to-signal ratio for a fixed number of measurements. Below a certain value the RIP requirements will be too strong for $\mathcal{A}$ to fulfill it, the RIP breaks down, and the recovery guarantees fail.

7.7.2 Validation of Theorem 7.5.1

In the second experiment, we study the influence of parameters α and β on the reconstruction accuracy. In particular, we vary the parameters α and β when reconstructing one randomly drawn $\hat{\mathbf{X}} \in \mathbb{R}^{16\times 100}$, $\|\hat{\mathbf{X}}\|_F = 10$, with $\operatorname{rank}(\hat{\mathbf{X}}) = 1$ and 10-sparse right singular vector from 90 measurements without noise. Again parameter choice is exemplary. We compare the three settings: (a) $\alpha = \beta$, (b) $\alpha = 0.01\beta$ and (c) $\alpha = 100\beta$ in Figure 7.2.



(a) $R = 1$ (b) $R = 5$

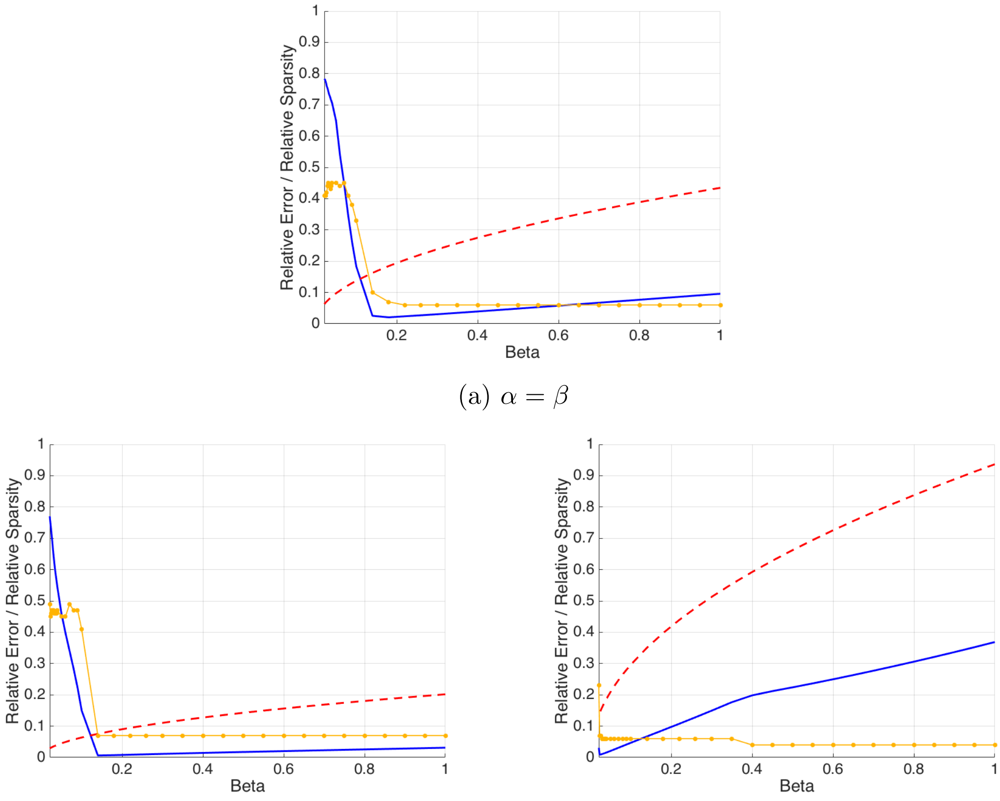
Figure 7.1: Approximation quality depending on noise level. The x-axis shows noise to signal ratio $\|\boldsymbol{\eta}\|_2/\|\hat{\mathbf{X}}\|_F$ while the y-axis presents approximation error relative to $\|\hat{\mathbf{X}}\|_F$. One can see the comparison of approximation results (solid blue) and theoretical bound (dashed red)

One can observe a decrease of approximation error for $\alpha, \beta \to 0$ up to a certain threshold, under which the approximation seemingly fails. While this threshold lies at $\beta \approx 0.15$ in (a) and (b) it is hardly recognizable in (c). At the same time (a) and (b) show a much smaller approximation error. These observations suggest that the choice of α strongly influences the approximation quality of ATLAS. This is consistent with Theorem 7.5.1, as a smaller α leads to a smaller theoretical approximation error bound.
Even though (a) and (b) show a linear decrease in approximation error which is in contrast to the square-root behavior of the theoretical bound, (c) suggests that the error, indeed, behaves similar to the theoretical bound.
Figure 7.2 shows that the sparsity level remains stable for sufficiently large β and breaks down precisely at the same threshold as the approximation error, coinciding with the violation of the RIP conditions.

For a better understanding of ATLAS we made a third experiment reconstructing one randomly drawn $\hat{\mathbf{X}} \in \mathbb{R}^{16\times 100}$ with $\operatorname{rank}(\hat{\mathbf{X}}) = 1$ and 10-sparse right singular vector for different values of $\|\hat{\mathbf{X}}\|_F$ from 90 measurements. The noise level was set to 0 and the parameters to $\alpha = \beta = 0.5$. The outcome is depicted in Figure 7.3. One can see the relative approximation error decreasing with the magnitude of $\hat{\mathbf{X}}$ as expected from the bound of Theorem 7.5.1. This seemingly confirms the theoretical dependence of reconstruction error on $\|\hat{\mathbf{X}}\|_{\frac{2}{3}}^{\frac{1}{3}}$.

7.7.3 ATLAS vs SPF

After confirming the theoretical results numerically, we now turn to the comparison of ATLAS with its state-of-the-art counterpart SPF. To our knowledge, SPF is the only algorithm available so far in matrix sensing, which exploits low-rankness and sparsity constraints together and comes with near-optimal recovery guarantees (not relying on a



(a) $\alpha = \beta$

(b) $\alpha = 0.01\beta$

(c) $\alpha = 100\beta$

Figure 7.2: Approximation quality and sparsity depending on parameter size. The approximation error (solid blue) and the theoretical bound (dashed red) are measured relative to $\|\hat{\mathbf{X}}\|_F$ while sparsity of the right singular vector (dotted yellow) is relative to n_2.

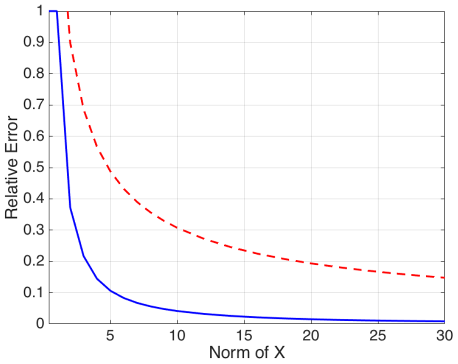


Figure 7.3: Approximation error depending on the magnitude of $\hat{\mathbf{X}}$ in Frobenius norm. Approximation error (solid blue) and theoretical bound (dashed red) are relative to $\|\hat{\mathbf{X}}\|_F$.

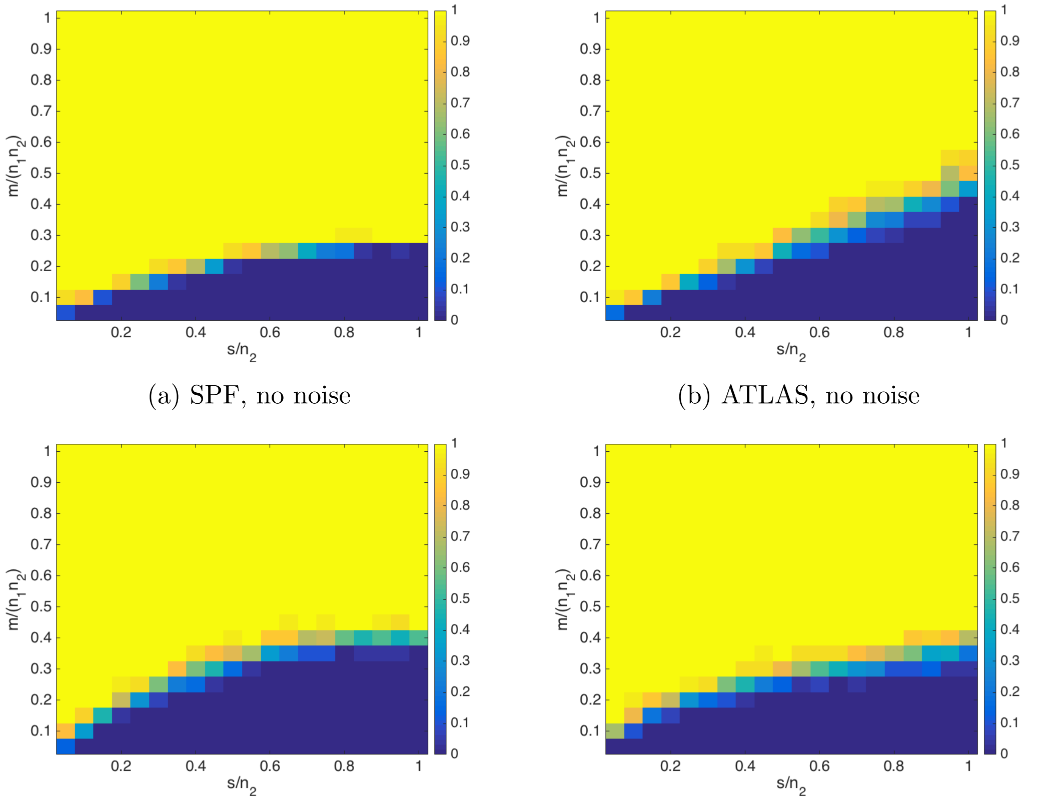
special structure of $\mathcal{A}$ as in [9]). As [116] contains exhaustive numerical comparisons of SPF and low-rank (resp. sparse) recovery strategies based on convex relaxation, SPF suffices for numerical benchmark tests. From the structure of the algorithms and their respective theoretical analysis one would expect SPF to yield more accurate reconstruction in the noiseless-to-low-noise setting, while ATLAS should prove to be more reliable if noise becomes large. This theoretical expectation is confirmed by the following experiments.

In Figure 7.4 we compare for $s/n_2 \in [0,1]$ and $m/(n_1 n_2)$ the number of successful recoveries of 30 randomly drawn $\hat{\mathbf{X}} \in \mathbb{R}^{4\times 128}$, $\|\hat{\mathbf{X}}\|_F = 10$, with $\operatorname{rank}(\hat{\mathbf{X}}) = 1$ and s-sparse right singular vectors from m measurements. The dimensions of $\hat{\mathbf{X}}$ are chosen accordingly to similar experiments in [116]. We set the noise level to 0 (resp. $0.3\|\hat{\mathbf{X}}\|_F$) and count the recovery successful if $\|\hat{\mathbf{X}} - \mathbf{X}_{\text{appr}}\|_F / \|\hat{\mathbf{X}}\|_F \leq 0.2$ (resp. 0.4). In order to compare the noisy and noiseless cases, we fix $\alpha = \beta = 0.5$ for both, which is a reasonable choice for high noise level, but perhaps sub-optimal if the noise level is low. Selected quantiles are directly compared in Figure 7.5 for convenience.
As expected, SPF outperforms ATLAS if there is no noise. In case of strong noise on the measurements, the situation changes. In particular, we observe the improved performance of ATLAS, whereas the SPF performance remarkably deteriorates.

To further quantify this effect, we perform the experiments reflected in Figure 7.6. For varying number of measurements we compare average approximation error and recovery probability of SPF and ATLAS for 30 randomly chosen $\hat{\mathbf{X}} \in \mathbb{R}^{16\times 100}$, $\|\hat{\mathbf{X}}\|_F = 10$, with $\operatorname{rank}(\hat{\mathbf{X}}) = 5$ and 10-sparse right singular vectors which either share a common support or may have various support sets. The parameters are chosen as $\alpha = \beta = 0.5$. One can clearly see that SPF outperforms ATLAS even in the noisy case for common support sets of the singular vectors. This is not surprising as ATLAS makes no use of the additional information provided by shared support sets. If the singular vectors, however, do not share a common support set, ATLAS shows its strength in the noisy setting. SPF which needs pre-information on the row-/column-sparsity $\tilde{s}$ of $\hat{\mathbf{X}}$ has to be initialized with $\tilde{s} = Rs$ as in the general case all support sets may differ.

7.7.4 Initialization

We close the section by a simple test on the influence of initialization. The plots in Figure 7.7 compared for $s/n_2 \in [0, 0.5]$ and $m/(n_1 n_2) \in [0,1]$ the number of successful recoveries of 20 randomly drawn $\hat{\mathbf{X}} \in \mathbb{R}^{8\times 128}$, $\|\hat{\mathbf{X}}\|_F = 10$, with $\operatorname{rank}(\hat{\mathbf{X}}) \in \{1,3\}$ and s-sparse right singular vectors from m measurements. The noise level was set to $0.3\|\hat{\mathbf{X}}\|_F$ and recovery was counted successful if $\|\hat{\mathbf{X}} - \mathbf{X}_{\text{appr}}\|_F / \|\hat{\mathbf{X}}\|_F \leq 0.4$. We compare initialization by the leading singular vectors of $\mathcal{A}^*(\mathbf{y})$ and by the leading singular vectors of $\mathbf{X} + \mathbf{Z}$ where $\mathbf{Z}$ is drawn at random, and scaled to $\|\mathbf{Z}\|_F = 100$ (strong perturbation) resp. $\|\mathbf{Z}\|_F = 0.2$ (mild perturbation).
For $\operatorname{rank}(\hat{\mathbf{X}}) = 1$ we note remarkably that the convergence radius of ATLAS is seemingly very large (yet not global), as the phase transition diagrams in Figure 7.7 do not show significant variations from choosing as initialization the leading singular vectors of $\mathcal{A}^*(\mathbf{y})$ and those of small random perturbation. Instead for $\operatorname{rank}(\hat{\mathbf{X}}) = 3$, initialization plays a



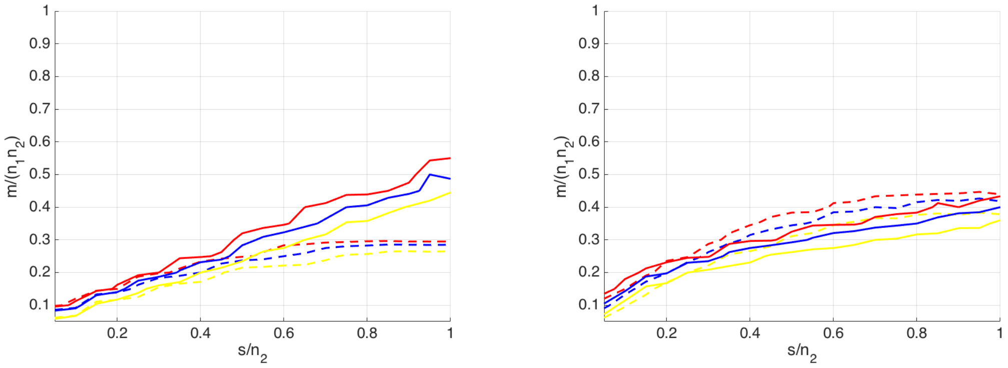
(a) SPF, no noise

(b) ATLAS, no noise

(c) SPF, with relatively strong noise

(d) ATLAS, with relatively strong noise

Figure 7.4: Phase transition diagrams comparing SPF and ATLAS with and without noise on the measurements. Empirical recovery probability is depicted by color from zero (blue) to one (yellow)



(a) No noise

(b) Noise

Figure 7.5: Recovery probability comparison of SPF (dashed) and ATLAS (solid). Plotted are the thresholds for 90% (red), 70% (blue) and 30% (yellow) successful recoveries. A recovery was counted successful if $\|\hat{\mathbf{X}} - \mathbf{X}_{\text{appr}}\|_F / \|\hat{\mathbf{X}}\|_F \leq 0.2$ (resp. 0.4)

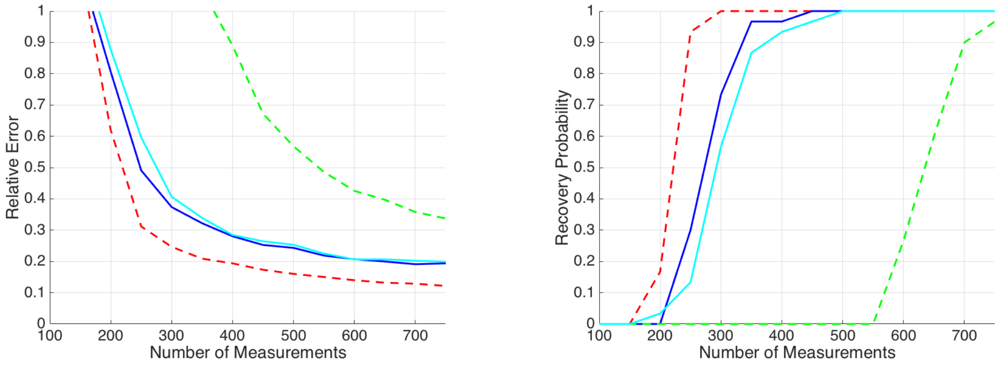


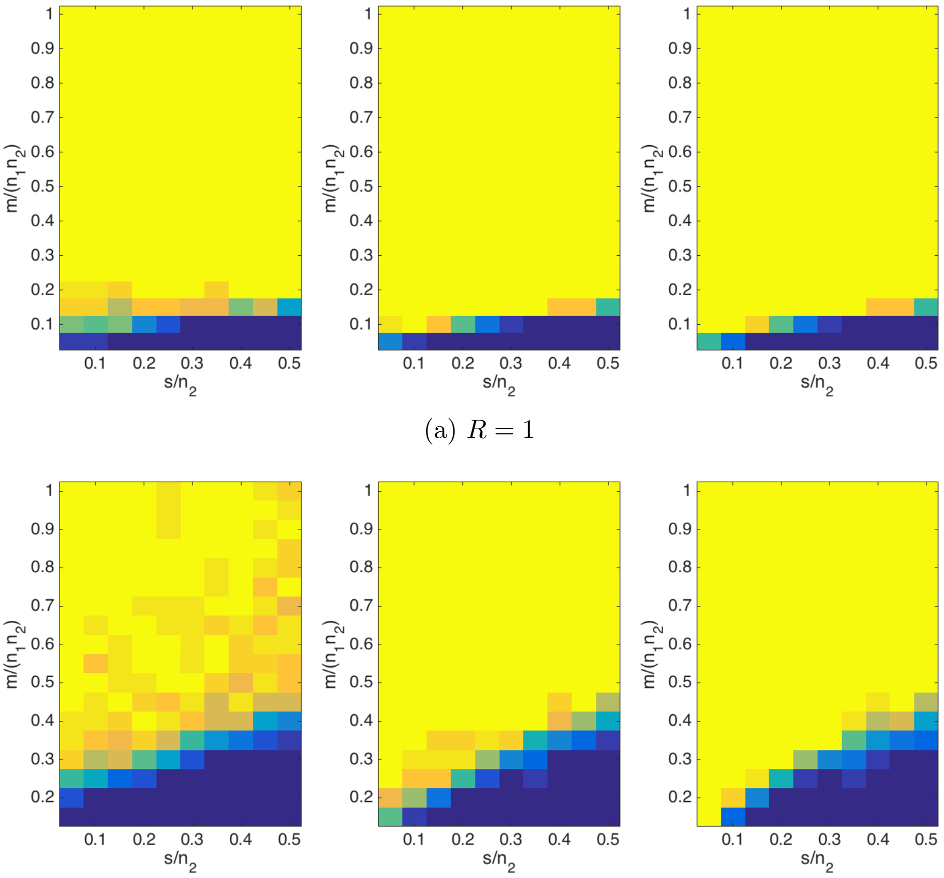
Figure 7.6: Comparison of SPF and ATLAS with and without common support for $R = 5$. Depicted are average approximation error relative to $\|\mathbf{X}\|_F$ and empirical recovery probabilities of SPF (dashed) and ATLAS (solid). Common Support: SPF (red) vs ATLAS (blue). Arbitrary Support: SPF (green) vs ATLAS (cyan).

more important role in performance and the initialization by leading singular vectors of $\mathcal{A}^*(\mathbf{y})$ does not yield optimal performance.

7.8 Discussion

In this chapter we deduced general bounds on the performance of the proposed algorithm, ATLAS, and a necessary number of measurements for subgaussian measurements to approximate effectively sparse and low-rank matrices. The theoretical results were confirmed in numerical experiments. ATLAS is especially effective in the most realistic setting of ineliminable noise and, hence, it complements the state-of-the-art algorithm SPF of Lee et. al. in [116], which works well for low level of noise or exact measurements. Moreover, ATLAS tackles the recovery of a significantly larger class of matrices than SPF, matrices with non-orthogonal rank-1 decompositions and effectively sparse components.
We wish to conclude by emphasizing the last point. In Section 7.1 we motivated the recovery of sparse and low-rank matrices by blind demixing, a specific signal processing application. The more general setting we consider for ATLAS, however, notably enlarges the scope of possible applications.

Principal Component Analysis (PCA) [101] is a classical tool for processing large amounts of data and performing data analysis such as dimensionality reduction and factor extraction. It has been widely used in various areas ranging from engineering and technology to social sciences and biology. We illustrate PCA by considering a simple example of a grocery store, which has n_1 regular customers and n_2 products. Let $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ be such that $X_{i,j}$ is the probability of customer i buying product j. It is reasonable to assume that there are only $R \ll \min\{n_1, n_2\}$ underlying basic factors like age, income, family size, etc. which govern the customer's purchase behavior. For each basic factor $r \in [R] := \{1, ..., R\}$ one defines two vectors: a vector $\mathbf{u}^r \in \mathbb{R}^{n_1}$ of components u_i^r encoding for each user $i \in [n_1]$



(a) $R = 1$

(b) $R = 3$

Figure 7.7: Comparison of different initializations for ATLAS for (7.13) with noise $\boldsymbol{\eta} \neq 0$ on the measurements, namely, initialization with a strongly perturbed approximation $\mathbf{X}_0 \approx \hat{\mathbf{X}}$ (left), initialization by the leading singular vectors of $\mathcal{A}^*(\mathbf{y})$ (middle), and initialization with a mildly perturbed approximation $\mathbf{X}_0 \approx \hat{\mathbf{X}}$ (right). Empirical recovery probability is depicted by color from zero (blue) to one (yellow).

how much they are affected by the factor r, and a vector $\mathbf{v}^r \in \mathbb{R}^{n_2}$ encoding the probability of buying product j if having factor r. Then, one can decompose

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^T = \sum_{r=1}^{R} \mathbf{u}^r (\mathbf{v}^r)^T \tag{7.35}$$

as the product of two matrices $\mathbf{U} \in \mathbb{R}^{n_1 \times R}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times R}$ with columns $\mathbf{u}^r$ and $\mathbf{v}^r$. Even if the product $\mathbf{U}\mathbf{V}^T$ is only approximately X, the decomposition into orthogonal principal components $\mathbf{U}$ and loadings $\mathbf{V}$ is appealing for more interpretability and having less data to store ($\mathcal{O}(n_1 R + n_2 R)$ instead of $\mathcal{O}(n_1 n_2)$). Both, $\mathbf{U}$ and $\mathbf{V}$ can be simply obtained by calculating a rank-reduced SVD of $\mathbf{X}$.

However, if we want to understand which factors mostly affect customer's behavior, PCA might not be the best option, since principal components are usually a linear combination of all original variables. To further improve interpretability and reduce the number of explicitly used variables, sparse PCA [173, 38], which promotes sparsity of the loadings $\mathbf{v}^r$ in (7.35), has been proposed. Sparse PCA trades orthogonality of the principal components for sparse solutions. In the aforementioned example of the grocery store, it is quite reasonable to assume sparsity of the probability distributions $\mathbf{v}^r$, as certain factors normally are more correlated with the probability of purchase of few specific items.

For some applications one may not have access to the complete matrix $\mathbf{X}$ but only to a partial indirect information, i.e., one has only $m \ll n_1 n_2$ pieces of information describing $\mathbf{X}$. In the example of the grocery store this may model the situation where customers do not possess all a fidelity card, which allows to identify them individually. Each day $d \in [D]$ the store caches in a certain amount of money y_l^d corresponding to purchases of a random subset $T_d \subset [n_1]$ of its customers ($l \in \mathbb{N}$ is a fixed index, whose role will soon become clear). If $\mathbf{p}^l \in \mathbb{R}^{n_2}$ is a vector encoding the prices p_j^l of each product j and $\mathcal{P}_{i,d} \subset [n_2]$ is the set of products purchased by customer i on a day d, we can express the takings as

$$y_l^d = \sum_{i \in T_d} \sum_{j \in \mathcal{P}_{i,d}} p_j^l,$$

If we assume that each customer i visits the grocery store with probability q_i, we can compute the expected takings as

$$\mathsf{E}_{T_d, \mathcal{P}_{\cdot,d}} \left[\sum_{i \in T_d} \sum_{j \in \mathcal{P}_{i,d}} p_j^l \right] = \sum_{i=1}^{n_1} q_i \sum_{j=1}^{n_2} X_{i,j} p_j^l.$$

Choosing D sufficiently large, the law of large numbers guarantees that

$$\lim_{D \to \infty} \frac{1}{D} \sum_{d=1}^{D} y_l^d = \mathsf{E}_{T_d, \mathcal{P}_{\cdot,d}} \left[\sum_{i \in T_d} \sum_{j \in \mathcal{P}_{i,d}} p_j^l \right],$$

in probability and almost surely. Moreover, by Central Limit Theorem, we may model the average takings of D days as

$$\frac{1}{D} \sum_{d=1}^{D} y_l^d = \sum_{i=1}^{n_1} q_i \sum_{j=1}^{n_2} X_{i,j} p_j^l + \eta_l^D,$$

for a suitable Gaussian noise η_l^D. By defining $y_l = \frac{1}{D}\sum_{d=1}^{D} y_l^d$, we can rewrite the above equation as

$$y_l = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(q_i p_j^l)X_{i,j} + \eta_l^D = \langle \mathbf{A}_l, \mathbf{X}\rangle_F + \eta_l^D$$

where the matrix $\mathbf{A}_l \in \mathbb{R}^{n_1\times n_2}$ has entries $(q_i p_j^l)_{i,j}$, and $\langle\cdot,\cdot\rangle_F$ is the Frobenius scalar product.

Tracking the daily sales over a time period of $m\cdot D$ days and perturbing the prizes in each subperiod $l \in [m]$ randomly would result in m inaccurate linear measurements, where each single measurement is a random average over the entries of $\mathbf{X}$ with an ineliminable additive noise η_l^D. (The random fluctuation of prizes is applied by groceries also for rotating promotions on products. Periodic price reductions, or sales, constitute a widely observed phenomenon in retailing. Sales occur on a regular basis, which suggests that they are not entirely due to random variations such as shocks to inventory holdings or demand.) The whole measurement process can be written as

$$\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\eta}$$

where $\mathcal{A}\colon \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ is a linear operator defined by the matrices $\mathbf{A}_1, ..., \mathbf{A}_m$ and $\boldsymbol{\eta} = (\eta_1^D, \dots, \eta_m^D)^T \in \mathbb{R}^m$ models the noise.

Since the probability distributions $\mathbf{v}^r$ cannot be expected to share a common support, the sparse principal components are not necessarily orthogonal, and there is a considerable amount of noise induced by the model, ATLAS should proof more useful than SPF in performing sparse PCA from inaccurate and incomplete measurements. Moreover, we have several promising extensions of ATLAS in mind which can be tackled by our tools.

First, replacing the ℓ_2- and ℓ_1-norms by ℓ_p- and ℓ_q-(quasi)-norms, for $p \geq 2$ and $0 < q < 2$, yields the functional

$$J_{\alpha,\beta}^{R,p,q}(\mathbf{u}^1,\dots,\mathbf{u}^R,\mathbf{v}^1,\dots,\mathbf{v}^R) := \left\| \mathbf{y} - \mathcal{A}\left(\sum_{r=1}^{R} \mathbf{u}^r(\mathbf{v}^r)^T\right)\right\|_2^2 + \alpha\sum_{r=1}^{R}\|\mathbf{u}^r\|_p^p + \beta\sum_{r=1}^{R}\|\mathbf{v}^r\|_q^q.$$

and, in turn, the algorithm

$$(\text{A-T-LAS}_{p,q})\begin{cases}
\mathbf{u}_{k+1}^1 = \arg\min_{\mathbf{u}} & \left\|\left(\mathbf{y} - \mathcal{A}\left(\sum_{r=2}^{R}\mathbf{u}_k^r\mathbf{v}_k^{r\,T}\right)\right) - \mathcal{A}(\mathbf{u}\mathbf{v}_k^{1\,T})\right\|_2^2 \\
 & +\alpha\|\mathbf{u}\|_p^p + \frac{1}{2\lambda_k^1}\|\mathbf{u}-\mathbf{u}_k^1\|_2^2, \\
\mathbf{v}_{k+1}^1 = \arg\min_{\mathbf{v}} & \left\|\left(\mathbf{y} - \mathcal{A}\left(\sum_{r=2}^{R}\mathbf{u}_k^r\mathbf{v}_k^{r\,T}\right)\right) - \mathcal{A}(\mathbf{u}_{k+1}^1\mathbf{v}^T)\right\|^2 \\
 & +\beta\|\mathbf{v}\|_q^q + \frac{1}{2\mu_k^1}\|\mathbf{v}-\mathbf{v}_k^1\|_2^2, \\
 & \vdots \\
\mathbf{u}_{k+1}^R = \arg\min_{\mathbf{u}} & \left\|\left(\mathbf{y} - \mathcal{A}\left(\sum_{r=1}^{R-1}\mathbf{u}_{k+1}^r{\mathbf{v}_{k+1}^r}^T\right)\right) - \mathcal{A}(\mathbf{u}\mathbf{v}_k^{R\,T})\right\|_2^2 \\
 & +\alpha\|\mathbf{u}\|_p^p + \frac{1}{2\lambda_k^R}\|\mathbf{u}-\mathbf{u}_k^R\|_2^2, \\
\mathbf{v}_{k+1}^R = \arg\min_{\mathbf{v}} & \left\|\left(\mathbf{y} - \mathcal{A}\left(\sum_{r=1}^{R-1}\mathbf{u}_{k+1}^r{\mathbf{v}_{k+1}^r}^T\right)\right) - \mathcal{A}(\mathbf{u}_{k+1}^R\mathbf{v}^T)\right\|^2 \\
 & +\beta\|\mathbf{v}\|_q^q + \frac{1}{2\mu_k^R}\|\mathbf{v}-\mathbf{v}_k^R\|_2^2,
\end{cases}$$

As for $q < 1$ even the single component minimizations become non-convex, this setting needs special care. One would need non-standard iterative thresholding methods, which have been developed and studied, e.g., in [131]. As q-quasi-norms, for $q < 1$, have proved particularly effective in enforcing sparsity, this additional technical difficulties are worth to overcome.

Second, in recommendation systems, one usually imposes additionally non negativity constraints on the obtained matrices. We could easily implement them in ATLAS by asymmetric ℓ_1-regularization. Define for $\mathbf{z} \in \mathbb{R}^n$ and $\theta > 0$

$$\|\mathbf{z}\|_{1,\theta}^{+} := \sum_{i=1}^{n} |z_i|_{\theta}^{+}, \quad |x|_{\theta}^{+} := \begin{cases} x & x \geq 0 \\ \theta|x| & \text{else.} \end{cases}$$

For θ becoming large, the regularization by $\|\cdot\|_{1,\theta}^{+}$ enforces sparsity and non-negativity. Replacing the ℓ_1-norm in ATLAS by $\|\cdot\|_{1,\theta}^{+}$ would result in the simple modification of ISTA (Algorithm 3) where the soft-thresholding operator $\mathbb{S}_\beta$ is substituted with

$$\mathbb{S}_{\beta,\theta}(\mathbf{z}) = \begin{pmatrix} S_{\beta,\theta}(z_1) \\ \vdots \\ S_{\beta,\theta}(z_{n_2}) \end{pmatrix}, \quad \text{where } S_{\beta,\theta}(z_i) = \begin{cases} z_i - \frac{\beta}{2} & z_i > \frac{\beta}{2} \\ 0 & -\theta\frac{\beta}{2} \leq z_i \leq \frac{\beta}{2} \\ z_i + \theta\frac{\beta}{2} & z_i < -\theta\frac{\beta}{2} \end{cases}.$$

Note that in the limit case $\theta \to \infty$ the operator $\mathbb{S}_{\beta,\theta}$ is a shifted ReLU function. Choosing θ sufficiently large or considering the limit $\theta \to \infty$ would lead to non-negative sparse PCA [172] from incomplete and inaccurate measurements with further applications in economics [98], biology [8], and computer vision [115].

Third, as a byproduct of our generalization of sparse PCA, we introduce, in our view, the right class of matrices and corresponding RIPs which might allow to study SPF in the more general setting of matrices having non-orthogonal, effectively sparse decompositions.

Apart from those extensions of algorithm and theory, there are two issues one has to deal with in future work.

The current results demand a careful choice of parameters at noise level. This drawback of multi-penalty regularization is well-known and could be attacked by implementing LASSO-path. LASSO-path has been recently extended to the multi-penalty setting in case of superposition of the signals [74], where the authors provided an efficient procedure for the construction of regions containing structurally similar solutions. In addition, $J_{\alpha,\beta}^{R}$ depends by construction heavenly on pre-knowledge of the rank R. One might ask how to get good estimates for R in case the rank is unknown.

As mentioned above, initialization is crucial for good performances of the algorithm. It is currently unclear how a good initialization can be obtained to guarantee convergence of the whole procedure to global minimizers. This question is closely connected to the fundamental problem in non-convex optimization how to initialize gradient-descent methods. In fact, alternating minimization is somewhat related to gradient-descent. While in gradient-descent one determines an optimal descent direction and then approximates the optimal step size, alternating minimization strongly restricts the directions in space in order to calculate optimal step sizes. Lee et. al. proposed an initialization, which worked in their setting if one assumes a strong decay of the singular values. Possibly one could prove this initialization to be sufficiently good in our setting as well, also in the light of recently improved analysis [67].

Appendix A

Parameter Choice for LASSO via OMP

In this chapter we examine a parameter choice strategy for LASSO, introduced in (2.12), based on a first approximation by OMP, see Algorithm 1. The idea is to find a parameter for LASSO minimizing the least squared distance between the LASSO solution and the OMP approximation. We provide theoretical guarantees that the additional application of LASSO with automatically tuned parameter cannot worsen the first approximation of OMP. Moreover, numerical experiments suggest LASSO to improve approximation results as soon as signals are not perfectly sparse and there is noise on the measurements. The results of this section are joint work with Judith Wewerka and were presented in similar form in her Master's thesis [169].

A.1 Problem Setup

Recall from Section 2.3.3 that LASSO is closely related to Basis Pursuit and provides sparse solutions to (2.4) by minimizing the weighted sum of a squared data fidelity and an ℓ_1-regularization term. Moreover, its minimizers can be approximated by ISTA, see Algorithm 3. Let us denote minimizers of LASSO by

$$\mathbf{x}_\alpha = \operatorname*{arg\,min}_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{y}-\mathbf{A}\mathbf{z}\|_2^2 + \alpha\|\mathbf{z}\|_1, \tag{A.1}$$

where $\alpha > 0$ is a regularization parameter that trades off between accuracy in fitting $\mathbf{y}$ and sparsity of the solution. A proper choice of α is crucial in order to neither overfitting the noise nor producing too sparse solutions [109]. If one chooses a large α, the solution $\mathbf{x}$ will be shrinking to zero; conversely, a small α leads to an ordinary least squares fit (cf. Section 2.3.3).

There are various approaches to select the regularization parameter of LASSO: cross-validation [54] and generalized cross-validation [72], Stein's unbiased estimate of risk [160], L-curve [83] and U-curve [109] criteria, Akaike Information Criterion [4] and Bayesian Information Criterion [149]. The correct parameter choice is still a current problem. A fixed choice of α is effective as long as the dimension N is small. However, with large dimension we get a large selection bias [152].

Inspired by parameter choice strategies for Tikhonov regularization in [167], we examine how to choose α from the given noisy data $\mathbf{y}$ without relying on any knowledge of the noise level. To do so, we start with the observation that under mild conditions on $\mathbf{A}$ the regularized solution $\mathbf{x}_\alpha$ is unique [161, Lemma 4] for all $\alpha > 0$ and the optimal choice of α to recover $\mathbf{x}$ can be described by

$$\alpha^* = \arg\min_{\alpha>0} \|\mathbf{x}_\alpha - \mathbf{x}\|_2. \tag{A.2}$$

Of course, the original signal $\mathbf{x}$ is unknown. Hence, we compute a first approximation $\mathbf{x}_{\text{OMP}}$ by using the greedy Orthogonal Matching Pursuit (OMP) and approximate α^* by

$$\hat{\alpha} = \arg\min_{\alpha>0} \|\mathbf{x}_\alpha - \mathbf{x}_{\text{OMP}}\|_2. \tag{A.3}$$

By this approach we aim at exploiting the advantages of OMP, overcome its downsides and additionally profit from the positive aspects of LASSO. OMP is known for its speed [132], its ease of implementation, and its efficiency when the signal is highly sparse [162]. The major advantage of LASSO is that it is quite robust to noise and not restricted to exactly sparse solutions (in contrast to OMP) but may be computationally intensive [153].

A.2 Theoretical Considerations

We provide now a simple theoretical justification for sequentially applying OMP and LASSO using the basic result Theorem 2.3.5. To be precise, we show that given an RIP of $\mathbf{A}$ and a first approximation $\mathbf{x}_{\text{OMP}}$ computed by OMP with approximation error $\|\mathbf{x}_{\text{OMP}} - \mathbf{x}\|_2 \leq \varepsilon$, for some $\varepsilon > 0$, the LASSO solution $\mathbf{x}_{\hat{\alpha}}$ with $\hat{\alpha}$ defined in (A.3) satisfies $\|\mathbf{x}_{\hat{\alpha}} - \mathbf{x}\|_2 \leq 2\varepsilon + C\sigma_k(\mathbf{x})_2 + D\eta$ where C and D are the constants from Theorem 2.3.5. This means choosing the LASSO parameter by OMP will produce a solution at least as good as the first OMP guess (keeping in mind that one in general cannot hope for a smaller error than best k-term approximation $\sigma_k(\mathbf{x})$ and noise level η). Moreover, when having strong noise and/or recovering effectively sparse $\mathbf{x}$ one might expect LASSO to improve on OMP as LASSO is known to be robust and does not enforce sparsity as strictly as OMP (the numerical experiments in Section A.3 fortify this intuition).
First recall the definitions of $\mathbf{x}_\alpha$, α_* and $\hat{\alpha}$ in (A.1), (A.2) and (A.3). We additionally define, for $\alpha > 0$,

$$\eta_\alpha = \|\mathbf{y} - \mathbf{A}\mathbf{x}_\alpha\|_2,$$

as in Theorem 2.3.5. Note that, under the reasonable assumption $\eta \leq \|\mathbf{y}\|_2$, Theorem 2.3.5 applies to all sufficiently large α as $\eta_\alpha \to \|\mathbf{y}\|_2$ for $\alpha \to \infty$ ($\mathbf{x}_\alpha \to \mathbf{0}$ for $\alpha \to \infty$) and $\eta_\alpha \to 0$ for $\alpha \to 0$. One plain observation is that α lower bounds η_α.

Lemma A.2.1. *Under the assumptions of Theorem 2.3.5, we have that*

$$\eta_\alpha \geq \frac{1}{1+\delta}\alpha,$$

for all $\alpha > 0$ with $\mathbf{x}_\alpha \neq \mathbf{0}$.

Proof: Using the Karush-Kuhn-Tucker conditions on (A.1) we get

$$\mathbf{A}^T(\mathbf{A}\mathbf{x}_\alpha - \mathbf{y}) = -\alpha \operatorname{sign}(\mathbf{x}_\alpha). \tag{A.4}$$

As $\mathbf{x}_\alpha \neq \mathbf{0}$ there exists $i \in \{1, ..., N\}$ s.t. $(x_\alpha)_i \neq 0$. Hence, (A.4) yields

$$\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}_\alpha - \mathbf{y} \rangle = -\alpha \operatorname{sign}((x_\alpha)_i)$$

where $\mathbf{a}_i \in \mathbb{R}^m$ denotes the i-th column of $\mathbf{A}$. Taking absolute value on both sides and applying Cauchy-Schwarz, we obtain

$$\begin{aligned}\alpha = |\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}_\alpha - \mathbf{y} \rangle| &\leq \|\mathbf{a}_i\|_2 \|\mathbf{A}\mathbf{x}_\alpha - \mathbf{y}\|_2 \\ &\leq (1+\delta)\eta_\alpha.\end{aligned}$$

In the last line, we used that by the RIP of $\mathbf{A}$ and $\mathbf{a}_i = \mathbf{A}\mathbf{e}_i$

$$(1-\delta) \leq \|\mathbf{a}_i\|_2 \leq (1+\delta)$$

■

Lemma A.2.1 implies that Theorem 2.3.5 applies to all $\alpha \geq (1+\delta_1)\eta$ for which $\mathbf{x}_\alpha \neq \mathbf{0}$. This, however, does not lead to the desired approximation, as the magnitude of η_α (appearing in the error bound of Theorem 2.3.5) might be considerably larger than η.
A second, and more helpful, observation is that under mild assumptions there exists some $\bar{\alpha} > 0$ for which $\eta_{\bar{\alpha}} = \eta$.

Lemma A.2.2. *If $\eta \leq \|\mathbf{y}\|_2$, there exists $\bar{\alpha} > 0$ with $\eta_{\bar{\alpha}} = \eta$.*

Proof: Observe that for any convergent parameter sequence $\alpha_n \to \alpha$, the LASSO functionals

$$J_{\alpha_n}(\mathbf{z}) = \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2 + \alpha\|\mathbf{z}\|_1$$

Γ-converge to J_α. Hence, the mapping $\alpha \mapsto \eta_\alpha$ is continuous (this still holds true if minimizers of the LASSO problem are not unique as by [161, Lemma 1] the value of η_α is invariant under the choice of minimizer). By the considerations above and the intermediate value theorem, the claim follows. ■

We are ready to state and prove the main result of this section as already outlined above.

Theorem A.2.3. *Suppose that the $2s$-th restricted isometry constant of $\mathbf{A} \in \mathbb{R}^{m\times N}$ satisfies $\delta < 4/\sqrt{41} \approx 0.6246$. Then, for any $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^m$ with $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \eta \leq \|\mathbf{y}\|_2$ the following holds.*
If the first OMP approximation $\mathbf{x}_{OMP}$ satisfies $\|\mathbf{x}_{OMP} - \mathbf{x}\|_2 \leq \varepsilon$, for some $\varepsilon > 0$, the solution $\mathbf{x}_{\hat{\alpha}}$ of problem (A.1) with $\hat{\alpha}$ defined in (A.3) *approximates $\mathbf{x}$ with the error*

$$\|\mathbf{x}_{\hat{\alpha}} - \mathbf{x}\|_2 \leq 2\varepsilon + C\sigma_k(\mathbf{x})_2 + D\eta$$

where C and D are the constants from Theorem 2.3.5.

Proof: By definition $\mathbf{x}_{\alpha^*}$ satisfies

$$\|\mathbf{x}_{\alpha^*} - \mathbf{x}\|_2 \leq \|\mathbf{x}_{\bar{\alpha}} - \mathbf{x}\|_2 \leq C\sigma_k(\mathbf{x})_2 + D\eta$$

where we applied Theorem 2.3.5 to $\bar{\alpha}$ from Lemma A.2.2 in the second inequality. By triangle inequality we obtain

$$\begin{aligned}\|\mathbf{x}_{\hat{\alpha}} - \mathbf{x}\|_2 \leq \|\mathbf{x}_{\hat{\alpha}} - \mathbf{x}_{\mathrm{OMP}}\|_2 + \|\mathbf{x}_{\mathrm{OMP}} - \mathbf{x}\|_2 &\leq \|\mathbf{x}_{\alpha^*} - \mathbf{x}_{OMP}\|_2 + \varepsilon \\ &\leq \|\mathbf{x}_{\alpha^*} - \mathbf{x}\|_2 + \|\mathbf{x} - \mathbf{x}_{\mathrm{OMP}}\|_2 + \varepsilon \\ &\leq 2\varepsilon + C\sigma_k(\mathbf{x})_2 + D\eta.\end{aligned}$$

■

A.3 Numerical Simulation

The last section showed that LASSO whose parameter is adapted to a first OMP approximation does approximate the true signal at least as good as OMP. In this section we want to provide numerical evidence that the additional application of LASSO can improve approximation quality. To this end, we both ran a toy example with signals drawn randomly from an ℓ_q-ball and tested recovery of MRI images from subsampled Fourier measurements. For solving LASSO we always use ISTA and for determining $\hat{\alpha}$ a simple grid search over different $\alpha \in (0, 1)$.

A.3.1 Sampling on ℓ_q-balls

In an artificial setting, we test how reliable OMP is in finding a good parameter α for LASSO and if LASSO can improve on the OMP approximation when the original signal is not exactly sparse but only effectively sparse. To this end, we draw signals $\mathbf{x} \in \mathbb{R}^N$ at random from ℓ_q-balls, for $0 < q \leq 1$ (cf. Section 3.2 and refer to [110] for further details on sampling uniform distributions on ℓ_q-balls).

In the first experiment we compare the optimal LASSO parameter α^* to $\hat{\alpha}$ obtained from (A.3) for 100 random realizations of $\mathbf{x} \in \mathcal{B}_q(\mathbf{0}, 1)$, $N = 200$, $q = 0.7$, and $m = 50$ Gaussian measurements, see Figure A.1. The noise level η is here set to zero. One can see that $\hat{\alpha}$ approximates α^* quite well in all realizations.

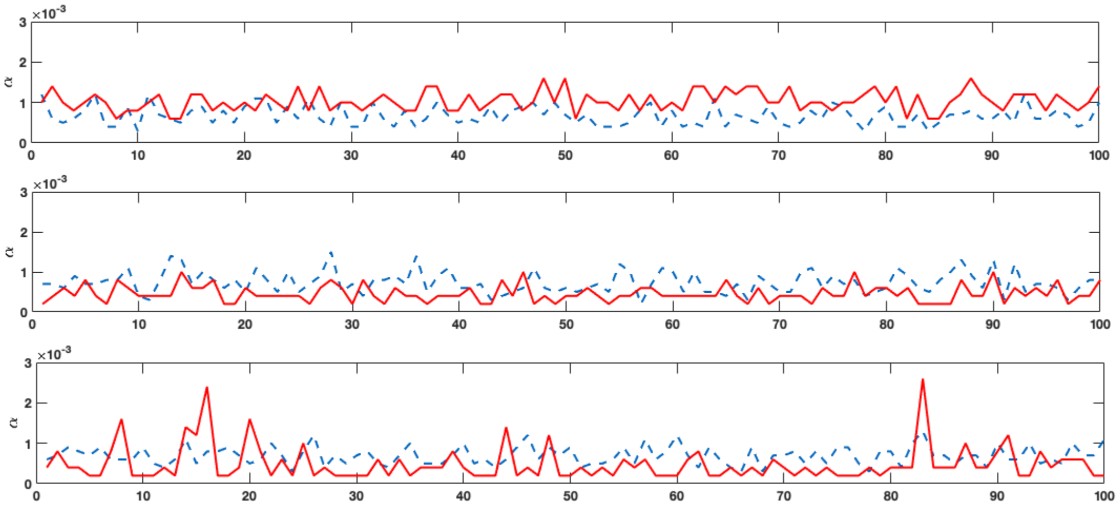


Figure A.1: Optimal parameter α^* (dashed blue) and corresponding approximation $\hat{\alpha}$ (solid red) for 100 ℓ_q-samples.

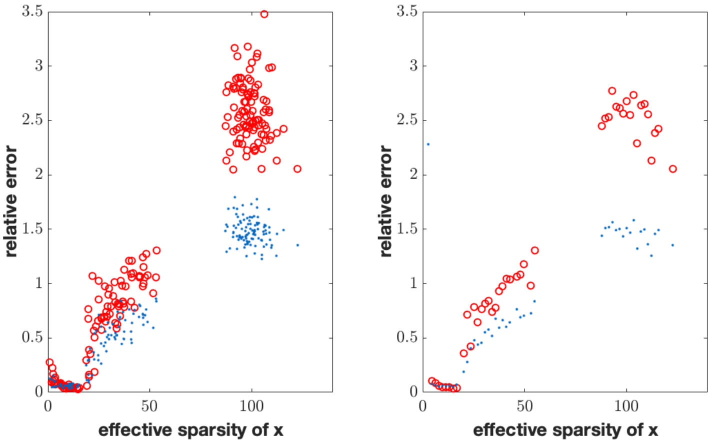


Figure A.2: Effective sparsity of $\mathbf{x}$ versus relative error of reconstruction by OMP (red circles) and LASSO (blue dots) for all samples (left) and in average (right).

In a second experiment we compared how well OMP and LASSO with $\hat{\alpha}$ reconstruct the signals $\mathbf{x}$ from their Gaussian measurements. Here, $N = 200$, $m = 80$, $q = 1$, and $\eta = 0$. We grouped signals of similar effective sparsity. As the ℓ_1-ball signals were mainly of effective sparsity $80-120$, we artificially added sparse random signals with sparsity $1-50$ to cover regimes for which the measurements satisfy the RIP. Figure A.2 depicts the outcome. Observe that LASSO with $\hat{\alpha}$ always performs at least as good as OMP. Moreover, as soon as OMP fails (sparsity ≥ 20) LASSO notably improves the approximation.

A.3.2 Compressed Sensing of MRI

After promising results in the toy scenario, we move to more realistic data. We now aim at recovering MRI images from randomly subsampled Fourier measurements by OMP and LASSO. To reduce the computational effort, a small 32×32 batch from an MRI brain image is picked as $\mathbf{x}$ (see Figure A.3), i.e. the ambient dimension is $N = 1024$. The measurement vector $\mathbf{y}$ consists of m Fourier coefficients of the discrete Fourier transform of our batch which are drawn uniformly at random from the N possible ones.

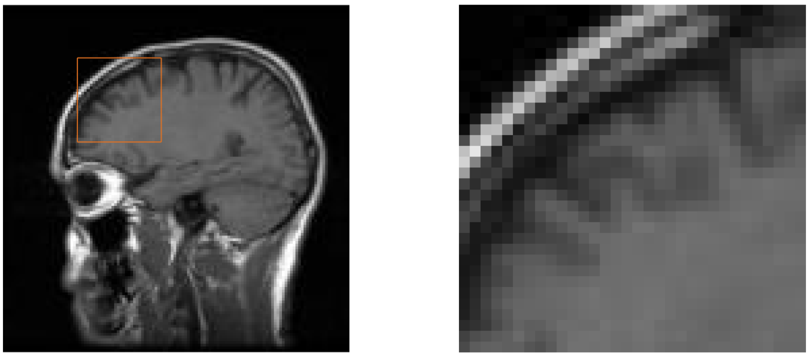


Figure A.3: Original MRI. The sample we reconstruct is inside the orange square (left) and the enlarged sample in the right image.

When transforming $\mathbf{x}$ to $\mathcal{W}(\mathbf{x})$ in a suitable Wavelet domain, we may assume it to become close to sparse. We used Haar-wavelets in our experiments and computed the signals effective sparsity s by $(\|\mathcal{W}(\mathbf{x})\|_1/\|\mathcal{W}(\mathbf{x})\|_2)^2$ yielding $s = 145$. On the one hand, this rather high value shows that Haar-wavelets are a suboptimal choice for representing the images. On the other hand, in real applications one normally has not a perfect representation domain at hand. So let us check how OMP and LASSO with $\hat{\alpha}$ perform here.

Figure A.4 depicts the reconstruction of the patch from $m = N/2 = 512$ measurements where 10% noise is added. OMP has been stopped after $s = 145$ iterations. LASSO with $\hat{\alpha}$ clearly improves on the OMP reconstruction. In Figure A.5 similar results are shown for $m = N/3 \approx 340$ measurements with 5% noise added.
Knowing s in advance is, in general, not possible. One might only have a very rough estimate of the expected sparsity. The last experiment (see Figure A.6) suggests LASSO with $\hat{\alpha}$ to be quite robust with respect to different stopping criteria k of OMP. This confirms the observations of the previous section and makes the automatically tuned LASSO a valuable postprocessing step for OMP. Here, $m = N/3 \approx 340$ measurements have been distorted by 10% noise.

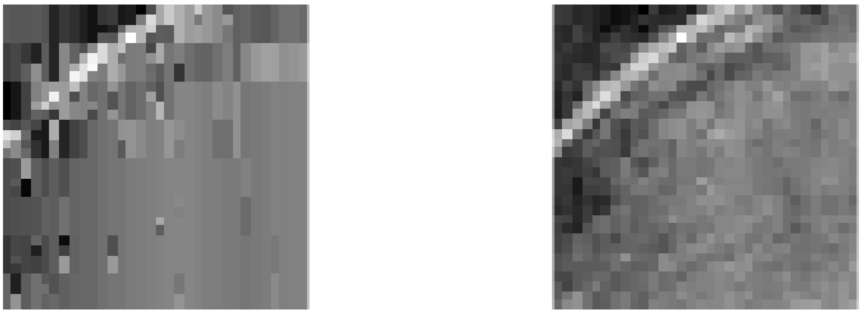


Figure A.4: Reconstruction of the original image via OMP with s iterations (left) and additional application of LASSO (right).

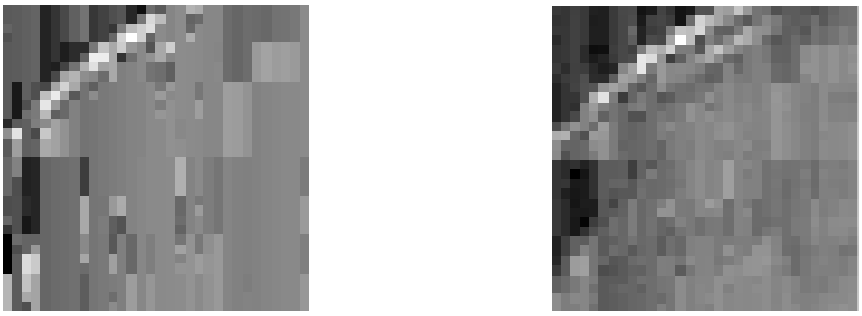


Figure A.5: Reconstruction of the original image via OMP with s iterations (left) and additional application of LASSO (right).

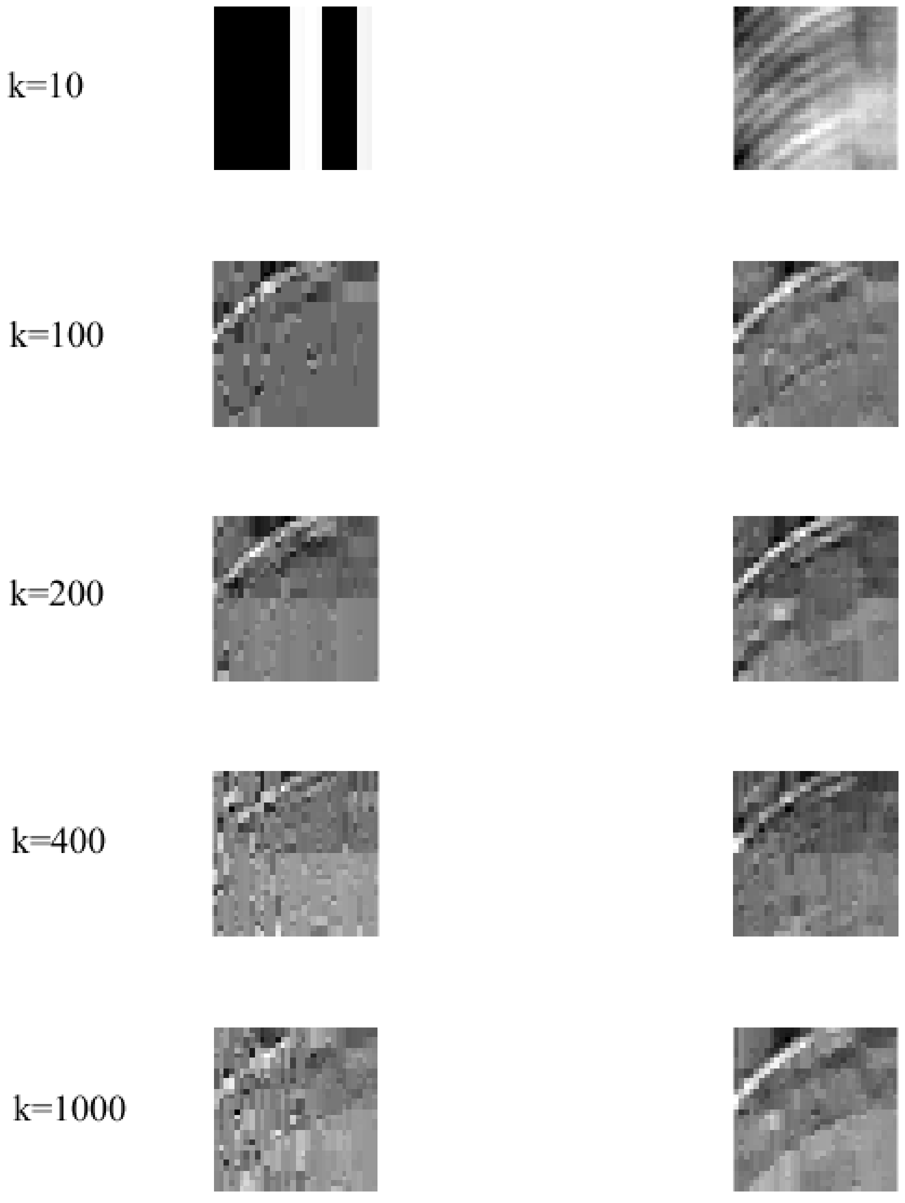


Figure A.6: Comparison of OMP (left) and LASSO (right) reconstruction for different stopping points of OMP.

Appendix B

Auxiliary Results

In this chapter we present the proofs of two results used in Chapter 7. They are joint work with Massimo Fornasier and Valeriya Naumova and have been published in [64].

B.1 Bounds for Gamma Functional

We calculate here the two integral estimates which are needed in the proof of Lemma 7.4.2.

Lemma B.1.1. *If* $\Gamma \geq 1$, *we have for the sets* $S_{s_1,s_2}^{R,\Gamma}$ *and* $K_{s_1,s_2}^{R,\Gamma}$ *defined in* (7.18) *and* (7.19) *that*

$$\int_0^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \sqrt{\log N\left(S_{s_1,s_2}^{R,\Gamma}, \|\cdot\|_F, \sqrt{m}\varepsilon\right)}\, \mathrm{d}\varepsilon \leq \sqrt{\frac{C_S\Gamma^2R^2(s_1+s_2)\log\left(\max\left\{n_1,n_2\right\}\right)}{m}}$$
$$\int_0^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \sqrt{\log N(K_{s_1,s_2}^{R,\Gamma}, \|\cdot\|_F, \sqrt{m}\varepsilon)}\, d\varepsilon \leq \sqrt{\frac{C_K\Gamma^2R^2(s_1+s_2)\log^3(\max\{n_1,n_2\})}{m}}$$

where $C_S, C_K > 0$ *are constants.*

Proof: For the first estimate apply Lemma 7.4.4 to obtain

$$\begin{aligned}
&\int_0^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \sqrt{\log N\left(S_{s_1,s_2}^{R,\Gamma}, \|\cdot\|_F, \sqrt{m}\varepsilon\right)}\, \mathrm{d}\varepsilon \\
&\leq \sqrt{\int_0^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} 1\, d\varepsilon \int_0^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \log N\left(S_{s_1,s_2}^{R,\Gamma}, \|\cdot\|_F, \sqrt{m}\varepsilon\right)\, \mathrm{d}\varepsilon} \\
&\leq \sqrt{\frac{\Gamma^2R^2(s_1+s_2+1)\Big(1+\log\left(18\sqrt{R}\right)\Big) + \Gamma^2R^2s_1\log\left(\frac{en_1}{s_1}\right) + \Gamma^2R^2s_2\log\left(\frac{en_2}{s_2}\right)}{m}} \\
&\leq \sqrt{\frac{C_S\Gamma^2R^2(s_1+s_2)\log\left(\max\left\{n_1,n_2\right\}\right)}{m}},
\end{aligned}$$

where we used Cauchy-Schwarz inequality in the first step and the fact that $\sqrt{R} \leq \max\{n_1, n_2\}$ in the last inequality. $C_S > 0$ is an appropriate constant.

To obtain the second estimate let us first assume $s_1/n_1 \leq s_2/n_2$. We apply Lemma 7.4.5 and find

$$\begin{aligned}
&\int_0^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \sqrt{\log N(K_{s_1,s_2}^{R,\Gamma}, \|\cdot\|_F, \sqrt{m}\varepsilon)}\, d\varepsilon \\
&\leq \int_0^{12\Gamma\sqrt{\frac{Rs_1}{mn_1}}} \sqrt{R(n_1+n_2+1)\log\left(\frac{36\Gamma R}{\sqrt{m}\varepsilon}\right)}\, \mathrm{d}\varepsilon \\
&+ \int_{12\Gamma\sqrt{\frac{Rs_1}{mn_1}}}^{12\Gamma\sqrt{\frac{Rs_2}{mn_2}}} \sqrt{\frac{144\Gamma^2R^2s_1}{m\varepsilon^2}\log\left(\frac{9\sqrt{m}\varepsilon n_1}{6\Gamma\sqrt{R}s_1}\right)}\, \mathrm{d}\varepsilon \\
&+ \int_{12\Gamma\sqrt{\frac{Rs_1}{mn_1}}}^{12\Gamma\sqrt{\frac{Rs_2}{mn_2}}} \sqrt{R(n_2+1)\log\left(\frac{36\Gamma R}{\sqrt{m}\varepsilon}\right)}\, \mathrm{d}\varepsilon \\
&+ \int_{12\Gamma\sqrt{\frac{Rs_2}{mn_2}}}^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \sqrt{\frac{144\Gamma^2R^2(s_1+s_2)}{m\varepsilon^2}\log\left(\frac{9\sqrt{m}\varepsilon n_1}{6\Gamma\sqrt{R}s_1}\right)}\, \mathrm{d}\varepsilon \\
&+ \int_{12\Gamma\sqrt{\frac{Rs_2}{mn_2}}}^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \sqrt{R\log\left(\frac{18\Gamma R}{\sqrt{m}\varepsilon}\right)} \\
&= I_1 + I_2 + I_3 + I_4 + I_5.
\end{aligned}$$

We now estimate the five integrals. We use the short notation $a_i = 12\Gamma\sqrt{\frac{Rs_i}{mn_i}}$ for $i = 1, 2$ and $b = \frac{\Gamma\sqrt{R}}{\sqrt{m}}$. The first integral can be bounded by

$$\begin{aligned}
I_1 &\leq \left(\int_0^{a_1} 1\, \mathrm{d}\varepsilon \int_0^{a_1} R(n_1+n_2+1)\log\left(\frac{36\Gamma R}{\sqrt{m}\varepsilon}\right)\, \mathrm{d}\varepsilon\right)^{\frac{1}{2}} \\
&\leq \left(a_1R(n_1+n_2+1)\left[\varepsilon\left(1+\log\left(\frac{36\Gamma R}{\sqrt{m}\varepsilon}\right)\right)\right]_{\varepsilon=0}^{a_1}\right)^{\frac{1}{2}} \\
&= \left(\frac{144\Gamma^2R^2s_1(n_1+n_2+1)}{mn_1}\left(1+\log\left(3\sqrt{\frac{Rn_1}{s_1}}\right)\right)\right)^{\frac{1}{2}} \\
&\leq \begin{cases} \left(\frac{432\Gamma^2R^2s_1}{m}\left(1+\log\left(3\sqrt{\frac{Rn_1}{s_1}}\right)\right)\right)^{\frac{1}{2}} & n_1 \geq n_2 \\ \left(\frac{432\Gamma^2R^2s_2}{m}\left(1+\log\left(3\sqrt{\frac{Rn_1}{s_1}}\right)\right)\right)^{\frac{1}{2}} & \text{else} \end{cases}
\end{aligned}$$

where we used in the last step the assumption $s_1/n_1 \leq s_2/n_2$. As can be seen later, the case distinction is irrelevant in the final estimate. Let us now turn to the second integral.

$$\begin{aligned}
I_2 &= \sqrt{\frac{144\Gamma^2 R^2 s_1}{m}} \int_{a_1}^{a_2} \frac{1}{\varepsilon} \sqrt{\log\left(\frac{9\sqrt{m}\varepsilon n_1}{6\Gamma\sqrt{R}s_1}\right)}\, \mathrm{d}\varepsilon \\
&= \sqrt{\frac{144\Gamma^2 R^2 s_1}{m}} \left[\frac{2}{3}\log^{\frac{3}{2}}\left(\frac{9\sqrt{m}\varepsilon n_1}{6\Gamma\sqrt{R}s_1}\right)\right]_{\varepsilon=a_1}^{a_2} \\
&= \left(\frac{64\Gamma^2 R^2 s_1}{m}\right)^{\frac{1}{2}} \left(\log^{\frac{3}{2}}\left(\frac{18\sqrt{s_2}n_1}{\sqrt{n_2}s_1}\right) - \log^{\frac{3}{2}}\left(\frac{18\sqrt{n_1}}{\sqrt{s_1}}\right)\right) \\
&\leq \left(\frac{64\Gamma^2 R^2 s_1}{m}\log^3(18 n_1)\right)^{\frac{1}{2}}
\end{aligned}$$

The third integral is similar to the first. Again the case distinction does not play a major role in the end.

$$\begin{aligned}
I_3 &\leq \left((a_2-a_1)R(n_2+1)\left[\varepsilon\left(1+\log\left(\frac{36\Gamma R}{\sqrt{m\varepsilon}}\right)\right)\right]_{\varepsilon=a_1}^{a_2}\right)^{\frac{1}{2}} \\
&= \left((a_2-a_1)R(n_2+1)\left[a_2\left(1+\log\left(3\sqrt{\frac{Rn_2}{s_2}}\right)\right) - a_1\left(1+\log\left(3\sqrt{\frac{Rn_1}{s_1}}\right)\right)\right]\right)^{\frac{1}{2}} \\
&\leq \left((a_2-a_1)^2 R(n_2+1)\left(1+\log\left(3\sqrt{\frac{Rn_1}{s_1}}\right)\right)\right)^{\frac{1}{2}} \\
&\leq \left((a_2^2+a_1^2)R(n_2+1)\left(1+\log\left(3\sqrt{\frac{Rn_1}{s_1}}\right)\right)\right)^{\frac{1}{2}} \\
&= \left(\frac{144\Gamma^2 R^2}{m}\left(\frac{s_2(n_2+1)}{n_2}+\frac{s_1(n_2+1)}{n_1}\right)\left(1+\log\left(3\sqrt{\frac{Rn_1}{s_1}}\right)\right)\right)^{\frac{1}{2}} \\
&\leq \begin{cases} \left(\frac{432\Gamma^2 R^2(s_1+s_2)}{m}\left(1+\log\left(3\sqrt{\frac{Rn_1}{s_1}}\right)\right)\right)^{\frac{1}{2}} & n_1 \geq n_2, \\ \left(\frac{432\Gamma^2 R^2 s_2}{m}\left(1+\log\left(3\sqrt{\frac{Rn_1}{s_1}}\right)\right)\right)^{\frac{1}{2}} & \text{else.} \end{cases}
\end{aligned}$$

In the third and the last line we again used $s_1/n_1 \leq s_2/n_2$. The fourth integral is similar to the second.

$$
\begin{aligned}
I_4 &= \sqrt{\frac{144\Gamma^2 R^2(s_1+s_2)}{m}} \int_{a_2}^{b} \frac{1}{\varepsilon}\sqrt{\log\left(\frac{9\sqrt{m}\varepsilon n_1}{6\Gamma\sqrt{R}s_1}\right)}\, \mathrm{d}\varepsilon \\
&= \sqrt{\frac{144\Gamma^2 R^2(s_1+s_2)}{m}} \left[\frac{2}{3}\log^{\frac{3}{2}}\left(\frac{9\sqrt{m}\varepsilon n_1}{6\Gamma\sqrt{R}s_1}\right)\right]_{\varepsilon=a_2}^{b} \\
&= \left(\frac{64\Gamma^2 R^2(s_1+s_2)}{m}\right)^{\frac{1}{2}} \left(\log^{\frac{3}{2}}\left(\frac{3n_1}{2s_1}\right) - \log^{\frac{3}{2}}\left(\frac{18\sqrt{s_2}n_1}{\sqrt{n_2}s_1}\right)\right) \\
&\leq \left(\frac{64\Gamma^2 R^2(s_1+s_2)}{m}\log^3(18n_1)\right)^{\frac{1}{2}}
\end{aligned}
$$

The last integral is similar to the third.

$$
\begin{aligned}
I_5 &\leq \left((b-a_2)R\left[\varepsilon\left(1+\log\left(\frac{18\Gamma R}{\sqrt{m}\varepsilon}\right)\right)\right]_{\varepsilon=a_2}^{b}\right)^{\frac{1}{2}} \\
&\leq \left((b-a_2)^2 R\left(1+\log\left(18\sqrt{\frac{Rn_2}{s_2}}\right)\right)\right)^{\frac{1}{2}} \\
&\leq \left((b^2+a_2^2)R\left(1+\log\left(18\sqrt{\frac{Rn_2}{s_2}}\right)\right)\right)^{\frac{1}{2}} \\
&= \left(\left(\frac{\Gamma^2R^2}{m}+\frac{144\Gamma^2R^2s_2}{mn_2}\right)\left(1+\log\left(18\sqrt{\frac{Rn_2}{s_2}}\right)\right)\right)^{\frac{1}{2}} \\
&\leq \left(\frac{145\Gamma^2R^2}{m}\left(1+\log\left(18\sqrt{\frac{Rn_2}{s_2}}\right)\right)\right)^{\frac{1}{2}}
\end{aligned}
$$

Let us put all estimates together. If $s_1/n_1 \geq s_2/n_2$, the involved entities would just switch their roles. Hence, we obtain

$$
\int_0^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \sqrt{\log N(K, \|\cdot\|_F, \sqrt{m}\varepsilon)}\, d\varepsilon \leq \sqrt{\frac{C_K\Gamma^2R^2(s_1+s_2)\log^3(\max\{n_1,n_2\})}{m}}
$$

for some constant $C_K > 0$. ■

B.2 ATLAS: Proof of Convergence

In this section we show the convergence of ATLAS to global minimizers as presented in Theorem 7.6.2 and 7.6.3 by adapting results from [7]. In particular, we first present two technical lemmas (Lemma B.2.1 & Lemma B.2.2), which are essentially generalizations of results in [7]. These lemmas are used to prove the central theorem (here Theorem B.2.3) of Attouch et. al. in our slightly more general setting. Finally, the theorem on local convergence, Theorem 7.6.2, and the theorem on convergence rates, Theorem 7.6.3, can be derived from Theorem B.2.3. We refer the interested reader to [7] for further details and provide references to the original work in brackets. Recall the assumption sets (H) and (H_1) from Section 7.6.

Lemma B.2.1 ([7, Lemma 5]). *Under assumptions (H) and (H_1) the sequences $\mathbf{u}_k^1, \ldots, \mathbf{v}_k^R$ are well-posed in the sense that all minimizations in (7.34) have unique and finite solutions. Moreover,*

(i)

$$L(\mathbf{u}_k^1, \ldots, \mathbf{v}_k^R) + \sum_{r=1}^{R} \frac{1}{2\lambda_{k-1}^r} \|\mathbf{u}_k^r - \mathbf{u}_{k-1}^r\|_2^2 + \sum_{r=1}^{R} \frac{1}{2\mu_{k-1}^r} \|\mathbf{v}_k^r - \mathbf{v}_{k-1}^r\|_2^2 \le L(\mathbf{u}_{k-1}^1, \ldots, \mathbf{v}_{k-1}^R),$$

for all $k \ge 1$, hence $L(\mathbf{u}_k^1, \ldots, \mathbf{v}_k^R)$ is non-increasing.

(ii)

$$\sum_{k=1}^{\infty} \left(\|\mathbf{u}_k^1 - \mathbf{u}_{k-1}^1\|_2^2 + \cdots + \|\mathbf{v}_k^R - \mathbf{v}_{k-1}^R\|_2^2\right) < \infty,$$

hence $\lim_{k\to\infty} \left(\|\mathbf{u}_k^1 - \mathbf{u}_{k-1}^1\|_2 + \cdots + \|\mathbf{v}_k^R - \mathbf{v}_{k-1}^R\|_2\right) = 0$.

(iii) For $k \ge 1$, define

$$(\tilde{\mathbf{u}}_k^1, \ldots, \tilde{\mathbf{v}}_k^R) := \begin{pmatrix} \nabla_{\mathbf{u}^1} Q(\mathbf{u}_k^1, \ldots, \mathbf{v}_k^R) - \nabla_{\mathbf{u}^1} Q(\mathbf{u}_k^1, \mathbf{u}_{k-1}^2, \ldots, \mathbf{v}_{k-1}^R) \\ \nabla_{\mathbf{v}^1} Q(\mathbf{u}_k^1, \ldots, \mathbf{v}_k^R) - \nabla_{\mathbf{v}^1} Q(\mathbf{u}_k^1, \mathbf{u}_{k-1}^2, \ldots, \mathbf{v}_k^1, \mathbf{v}_{k-1}^2, \ldots, \mathbf{v}_{k-1}^R) \\ \vdots \\ 0 \end{pmatrix}$$
$$- \begin{pmatrix} \frac{1}{\lambda_{k-1}^1}(\mathbf{u}_k^1 - \mathbf{u}_{k-1}^1) \\ \frac{1}{\mu_{k-1}^1}(\mathbf{v}_k^1 - \mathbf{v}_{k-1}^1) \\ \vdots \\ \frac{1}{\mu_{k-1}^R}(\mathbf{v}_k^R - \mathbf{v}_{k-1}^R) \end{pmatrix}.$$

Then $(\tilde{\mathbf{u}}_k^1, \ldots, \tilde{\mathbf{v}}_k^R) \in \partial L(\mathbf{u}_k^1, \ldots, \mathbf{v}_k^R)$ and for all bounded subsequences $(\mathbf{u}_{k'}^1, \ldots, \mathbf{v}_{k'}^R)$ we have $(\tilde{\mathbf{u}}_{k'}^1, \ldots, \tilde{\mathbf{v}}_{k'}^R) \to 0$, hence $\operatorname{dist}(0, \partial L(\mathbf{u}_{k'}^1, \ldots, \mathbf{v}_{k'}^R)) \to 0$*, for $k' \to \infty$.*

Proof: From $\inf L > -\infty$ and (H) it follows that the functions to be minimized in (7.34) are bounded below, coercive and lower semicontinuous and, therefore, the sequence $(\mathbf{u}_k^1, \ldots, \mathbf{v}_k^R)$ is well-posed.

(i) Using the minimizing properties of $\mathbf{u}_k^1, \ldots, \mathbf{v}_k^R$ from (7.34), we obtain

$$\begin{aligned}
L(\mathbf{u}_{k-1}^1, \ldots, \mathbf{v}_{k-1}^R) &\geq L(\mathbf{u}_k^1, \mathbf{u}_{k-1}^2, \ldots, \mathbf{v}_{k-1}^1, \ldots, \mathbf{v}_{k-1}^R) + \frac{1}{2\lambda_{k-1}^1}\|\mathbf{u}_k^1 - \mathbf{u}_{k-1}^1\|_2^2 \\
&\geq \left(L(\mathbf{u}_k^1, \mathbf{u}_{k-1}^2, \ldots, \mathbf{u}_{k-1}^R, \mathbf{v}_k^1, \mathbf{v}_{k-1}^2, \ldots, \mathbf{v}_{k-1}^R) + \frac{1}{2\mu_{k-1}^1}\|\mathbf{v}_k^1 - \mathbf{v}_{k-1}^1\|_2^2 \right) \\
&\quad + \frac{1}{2\lambda_{k-1}^1}\|\mathbf{u}_k^1 - \mathbf{u}_{k-1}^1\|_2^2 \\
&\quad\quad \vdots \\
&\geq L(\mathbf{u}_k^1, \ldots, \mathbf{v}_k^R) + \sum_{r=1}^R \frac{1}{2\lambda_{k-1}^r}\|\mathbf{u}_k^r - \mathbf{u}_{k-1}^r\|_2^2 + \sum_{r=1}^R \frac{1}{2\mu_{k-1}^r}\|\mathbf{v}_k^r - \mathbf{v}_{k-1}^r\|_2^2.
\end{aligned}$$

(ii) From *(i)* and (H_1) one has, for every $K \in \mathbb{N}$,

$$\begin{aligned}
\frac{1}{2r_+}\sum_{k=1}^K \left(\|\mathbf{u}_k^1 - \mathbf{u}_{k-1}^1\|_2^2 + \cdots + \|\mathbf{v}_k^R - \mathbf{v}_{k-1}^R\|_2^2\right) &\leq \sum_{k=1}^K \left(L(\mathbf{u}_{k-1}^1, \ldots, \mathbf{v}_{k-1}^R) - L(\mathbf{u}_k^1, \ldots, \mathbf{v}_k^R)\right) \\
&= L(\mathbf{u}_0^1, \ldots, \mathbf{v}_0^R) - L(\mathbf{u}_K^1, \ldots, \mathbf{v}_K^R) \\
&< L(\mathbf{u}_0^1, \ldots, \mathbf{v}_0^R) - \inf L < \infty.
\end{aligned}$$

By letting $K \to \infty$ we get the claim.

(iii) By definition of $\mathbf{u}_k^1$, $\mathbf{0}$ must lie in the subdifferential of $\boldsymbol{\xi} \mapsto L(\boldsymbol{\xi}, \mathbf{u}_{k-1}^2, \ldots, \mathbf{v}_{k-1}^R) + \frac{1}{2\lambda_{k-1}^1}\|\boldsymbol{\xi} - \mathbf{u}_{k-1}^1\|_2^2$ at $\mathbf{u}_k^1$. As a similar fact holds true for the other sequences, one gets, for all $1 \leq r \leq R$

$$\begin{aligned}
0 &\in \frac{1}{\lambda_{k-1}^r}(\mathbf{u}_k^r - \mathbf{u}_{k-1}^r) + \partial_{\mathbf{u}^r} L(\mathbf{u}_k^1, \ldots, \mathbf{u}_k^r, \mathbf{u}_{k-1}^{r+1}, \ldots, \mathbf{u}_{k-1}^R, \mathbf{v}_k^1, \ldots, \mathbf{v}_k^{r-1}, \mathbf{v}_{k-1}^r, \ldots, \mathbf{v}_{k-1}^R), \\
0 &\in \frac{1}{\mu_{k-1}^r}(\mathbf{v}_k^r - \mathbf{v}_{k-1}^r) + \partial_{\mathbf{v}^r} L(\mathbf{u}_k^1, \ldots, \mathbf{u}_k^r, \mathbf{u}_{k-1}^{r+1}, \ldots, \mathbf{u}_{k-1}^R, \mathbf{v}_k^1, \ldots, \mathbf{v}_k^r, \mathbf{v}_{k-1}^{r+1}, \ldots, \mathbf{v}_{k-1}^R).
\end{aligned}$$

The structure of L implies $\partial_{\mathbf{u}^r} L(\mathbf{u}_k^1, \ldots, \mathbf{u}_k^r, \mathbf{u}_{k-1}^{r+1} \ldots, \mathbf{u}_{k-1}^R, \mathbf{v}_k^1, \ldots, \mathbf{v}_k^{r-1}, \mathbf{v}_{k-1}^r, \ldots, \mathbf{v}_{k-1}^R) = \partial f_r(\mathbf{u}_k^r) + \nabla_{\mathbf{u}^r} Q(\mathbf{u}_k^1, \ldots, \mathbf{u}_k^r, \mathbf{u}_{k-1}^{r+1} \ldots, \mathbf{u}_{k-1}^R, \mathbf{v}_k^1, \ldots, \mathbf{v}_k^{r-1}, \mathbf{v}_{k-1}^r, \ldots, \mathbf{v}_{k-1}^R)$ and a similar equation for the v-components. Hence, one may rewrite the inclusions above:

$$\begin{aligned}
-\frac{1}{\lambda_{k-1}^1}(\mathbf{u}_k^1 - \mathbf{u}_{k-1}^1) - (\nabla_{\mathbf{u}^1} Q(\mathbf{u}_k^1, \mathbf{u}_{k-1}^2, \ldots, \mathbf{v}_{k-1}^R) - \nabla_{\mathbf{u}^1} Q(\mathbf{u}_k^1, \ldots, \mathbf{v}_k^R)) \\
\in \partial f_1(\mathbf{u}_k^1) + \nabla_{\mathbf{u}^1} Q(\mathbf{u}_k^1, \ldots, \mathbf{v}_k^R), \\
\vdots \\
-\frac{1}{\mu_{k-1}^R}(\mathbf{v}_k^R - \mathbf{v}_{k-1}^R) \in \partial g_R(\mathbf{v}_k^R) + \nabla_{\mathbf{v}^R} Q(\mathbf{u}_k^1, \ldots, \mathbf{v}_k^R).
\end{aligned}$$

Together with [7, Proposition 3] this yields the claim. ■

Lemma B.2.2 ([7, Proposition 6])**.** *Assume (H) and (H_1) hold. Let $(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$ be a sequence defined by (7.34) and $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$ be a (possibly empty) set of limit points. Then,*

(i) if $(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$ is bounded, then $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$ is nonempty, compact and connected and $\operatorname{dist}((\mathbf{u}_k^1, \dots, \mathbf{v}_k^R), \omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)) \to 0$ as $k \to \infty$,

(ii) $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) \subset \operatorname{crit} L$, where $\operatorname{crit} L$ denotes a set of critical points of L,

(iii) L is finite and constant on $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$, and equal to $\inf_{k\in\mathbb{N}} L(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R) = \lim_{k\to\infty} L(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$.

Proof: (i) If $(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$ is bounded, there exists a convergent subsequence, which implies $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$ is nonempty. It also follows $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$ is bounded. Let now $(\hat{\mathbf{u}}^1, \dots, \hat{\mathbf{v}}^R) \notin \omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$ be given. There must exist some $\varepsilon > 0$ with $(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R) \notin B((\hat{\mathbf{u}}^1, \dots, \hat{\mathbf{v}}^R), \varepsilon)$, for all $k \in \mathbb{N}$. But then $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) \cap B((\hat{\mathbf{u}}^1, \dots, \hat{\mathbf{v}}^R), \varepsilon) = \emptyset$. This proves $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$ is closed and, hence, compact. Let us assume $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$ is not connected and let $\omega_c(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) \subset \omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$ be a connected component. Then, $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) \setminus \omega_c(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) \neq \emptyset$ and there exists some $\varepsilon > 0$ such that

$$\omega_c^\varepsilon(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) \cap \omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) \setminus \omega_c(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) = \emptyset,$$

where $\omega_c^\varepsilon(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$ is an ε-neighborhood of $\omega_c(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$. We know from Lemma B.2.1 (ii) that

$$\lim_{k\to\infty} \left(\|\mathbf{u}_k^1 - \mathbf{u}_{k-1}^1\|_2 + \dots + \|\mathbf{v}_k^R - \mathbf{v}_{k-1}^R\|_2\right) = 0.$$

Combined with $\omega_c(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$ and $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) \setminus \omega_c(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$ being sets of limit points of $(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$, it implies the existence of a subsequence $(\mathbf{u}_{k'}^1, \dots, \mathbf{v}_{k'}^R) \subset \omega_c^\varepsilon(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) \setminus \omega_c^{\frac{\varepsilon}{2}}(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$. As this subsequence is bounded, it must have a limit point and $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) \cap \omega_c^\varepsilon(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) \setminus \omega_c^{\frac{\varepsilon}{2}}(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R) \neq \emptyset$. Contradiction. The last part of (i) can be proven in a similar way. If $\operatorname{dist}((\mathbf{u}_k^1, \dots, \mathbf{v}_k^R), \omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)) \nrightarrow 0$, there must exist a subsequence that keeps distance to $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$. But this subsequence again must have a limit point which obviously lies in $\omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$. Contradiction.

(ii) We have, for all $k \geq 1$, $\boldsymbol{\xi}^r \in \mathbb{R}^{n_1}$, $\boldsymbol{\zeta}^r \in \mathbb{R}^{n_2}$, that

$$L(\mathbf{u}_k^1, \mathbf{u}_{k-1}^2, \dots, \mathbf{v}_{k-1}^R) + \frac{1}{2\lambda_{k-1}^1}\|\mathbf{u}_k^1 - \mathbf{u}_{k-1}^1\|_2^2 \leq L(\boldsymbol{\xi}^1, \mathbf{u}_{k-1}^2, \dots, \mathbf{v}_{k-1}^R) + \frac{1}{2\lambda_{k-1}^1}\|\boldsymbol{\xi}^1 - \mathbf{u}_{k-1}^1\|_2^2$$

$$\vdots$$

$$L(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R) + \frac{1}{2\mu_{k-1}^R}\|\mathbf{v}_k^R - \mathbf{v}_{k-1}^R\|_2^2 \leq L(\mathbf{u}_k^1, \dots, \mathbf{v}_k^{R-1}, \boldsymbol{\zeta}^R) + \frac{1}{2\mu_{k-1}^R}\|\boldsymbol{\zeta}^R - \mathbf{v}_{k-1}^R\|_2^2$$

Using the bounds on λ_k^r and μ_k^r and the special form of L one gets

$$\begin{aligned}
&f_1(\mathbf{u}_k^1) + Q(\mathbf{u}_k^1, \mathbf{u}_{k-1}^2, \dots, \mathbf{v}_{k-1}^R) + \frac{1}{2r_+}\|\mathbf{u}_k^1 - \mathbf{u}_{k-1}^1\|_2^2 \\
&\le f_1(\boldsymbol{\xi}^1) + Q(\boldsymbol{\xi}^1, \mathbf{u}_{k-1}^2, \dots, \mathbf{v}_{k-1}^R) + \frac{1}{2r_-}\|\boldsymbol{\xi}^1 - \mathbf{u}_{k-1}^1\|_2^2 \\
&\vdots \\
&g_R(\mathbf{v}_k^R) + Q(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R) + \frac{1}{2r_+}\|\mathbf{v}_k^R - \mathbf{v}_{k-1}^R\|_2^2 \\
&\le g_R(\boldsymbol{\zeta}^R) + Q(\mathbf{u}_k^1, \dots, \mathbf{v}_k^{R-1}, \boldsymbol{\zeta}^R) + \frac{1}{2r_-}\|\boldsymbol{\zeta}^R - \mathbf{v}_{k-1}^R\|_2^2
\end{aligned}$$

Let $(\overline{\mathbf{u}}^1, \dots, \overline{\mathbf{v}}^R) \in \omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$. There exists a subsequence $(\mathbf{u}_{k'}^1, \dots, \mathbf{v}_{k'}^R)$ of $(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$ with $(\mathbf{u}_{k'}^1 \dots, \mathbf{v}_{k'}^R) \to (\overline{\mathbf{u}}^1, \dots, \overline{\mathbf{v}}^R)$. Together with Lemma B.2.1.(ii) this gives

$$\liminf_{k'\to\infty} f_r(\mathbf{u}_{k'}^r) + Q(\overline{\mathbf{u}}^1, \dots, \overline{\mathbf{v}}^R) \le f_r(\boldsymbol{\xi}^r) + Q(\overline{\mathbf{u}}^1, \dots, \boldsymbol{\xi}^r, \dots, \overline{\mathbf{v}}^R) + \frac{1}{2r_-}\|\boldsymbol{\xi}^r - \overline{\mathbf{u}}^r\|_2^2,$$

for all $1 \le r \le R$. We can now set $\boldsymbol{\xi}^r = \overline{\mathbf{u}}^r$ to obtain

$$\liminf_{k'\to\infty} f_r(\mathbf{u}_{k'}^r) \le f_r(\overline{\mathbf{u}}^r).$$

This and f_r being lower semicontinuous yields

$$\lim_{k'\to\infty} f_r(\mathbf{u}_{k'}^r) = f_r(\overline{\mathbf{u}}^r).$$

Repeating the argument for g_r, $1 \le r \le R$, and recalling the continuity of Q we obtain $L(\mathbf{u}_{k'}^1, \dots, \mathbf{v}_{k'}^R) \to L(\overline{\mathbf{u}}^1, \dots, \overline{\mathbf{v}}^R)$. Combined with Lemma B.2.1 (iii) and the closedness properties of ∂L(see Remark 1(b) in [7]) proves $\mathbf{0} \in \partial L(\overline{\mathbf{u}}^1, \dots, \overline{\mathbf{v}}^R)$.

(iii) As we just seen, for any point $(\overline{\mathbf{u}}^1, \dots, \overline{\mathbf{v}}^R) \in \omega(\mathbf{u}_0^1, \dots, \mathbf{v}_0^R)$, there exists a subsequence $(\mathbf{u}_{k'}^1, \dots, \mathbf{v}_{k'}^R)$ of $(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$ with $L(\mathbf{u}_{k'}^1 \dots, \mathbf{v}_{k'}^R) \to L(\overline{\mathbf{u}}^1, \dots, \overline{\mathbf{v}}^R)$. Then $L(\overline{\mathbf{u}}^1, \dots, \overline{\mathbf{v}}^R) = \inf L(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$ as $L(\mathbf{u}_k^1, \dots, \mathbf{v}_k^R)$ is non-increasing. This holds for every limit point. Hence, L is finite and constant on the set of limit points. ■

Following the notation in [7] we write

$$\begin{aligned}
\mathbf{z}_k &:= (\mathbf{u}_k^1, \dots, \mathbf{v}_k^R), & l_k &:= L(\mathbf{z}_k), \\
\overline{\mathbf{z}} &:= (\overline{\mathbf{u}}^1, \dots, \overline{\mathbf{v}}^R), & \bar{l} &:= L(\overline{\mathbf{z}}).
\end{aligned}$$

The next theorem essentially asserts that a sequence $\mathbf{z}_k$ starting in the neighborhood of a point $\overline{\mathbf{z}}$, as described in (B.2), and not improving $L(\overline{\mathbf{z}})$, as given in (B.1), converges to a critical point near $\overline{\mathbf{z}}$.

Theorem B.2.3 ([7, Theorem 8])**.** *Let L satisfy (H), (H_1) and have the KL-property at some $\overline{\mathbf{z}}$. Denote by U, η and $\varphi : [0, \eta) \to \mathbb{R}$ the objects connected to the KL-property of L at $\overline{\mathbf{z}}$. Let $\rho > 0$ be chosen such that $B(\overline{\mathbf{z}}, \rho) \subset U$. Let $\mathbf{z}_k$ be generated by (7.34) with $\mathbf{z}_0$ as initial point. Let us assume that*

$$\bar{l} < l_k < \bar{l} + \eta, \tag{B.1}$$

for all $k \geq 0$, and

$$M\varphi(l_0 - \bar{l}) + 2\sqrt{2r_+}\sqrt{l_0 - \bar{l}} + \|\mathbf{z}_0 - \overline{\mathbf{z}}\|_2 < \rho \tag{B.2}$$

where $M = 2r_+(C\sqrt{2R} + \frac{1}{r_-})$ and C is a Lipschitz-constant for ∇Q on $B(\overline{\mathbf{z}}, \sqrt{2R}\rho)$. Then, the sequence $\mathbf{z}_k$ converges to a critical point of L and the following holds, for all $k \geq 0$:

(i) $\mathbf{z}_k \in B(\overline{\mathbf{z}}, \rho)$

(ii) $\sum_{i=k+1}^{\infty} \|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2 \leq M\varphi(l_k - \bar{l}) + \sqrt{2r_+}\sqrt{l_k - \bar{l}}$.

Proof: We may without loss of generality assume $L(\overline{\mathbf{z}}) = 0$ (replace L by $L - L(\overline{\mathbf{z}})$).

With Lemma B.2.1.(i) we have

$$l_i - l_{i+1} \geq \frac{1}{2r_+}\|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2^2, \tag{B.3}$$

for all $i \geq 0$. Moreover, $\varphi'(l_i)$ makes sense in view of (B.1) and $\varphi'(l_i) > 0$. Hence,

$$\varphi'(l_i)(l_i - l_{i+1}) \geq \frac{\varphi'(l_i)}{2r_+}\|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2^2.$$

Owing to φ being concave, we obtain

$$\varphi(l_i) - \varphi(l_{i+1}) \geq \frac{\varphi'(l_i)}{2r_+}\|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2^2, \tag{B.4}$$

for all $i \geq 0$. Let us first check (i) for $k = 0$ and $k = 1$. We know from (B.2) that $\mathbf{z}_0$ lies in $B(\overline{\mathbf{z}}, \rho)$. Furthermore, (B.3) yields

$$\frac{1}{2r_+}\|\mathbf{z}_1 - \mathbf{z}_0\|_2^2 \leq l_0 - l_1 \leq l_0$$

which gives

$$\|\mathbf{z}_1 - \overline{\mathbf{z}}\|_2 \leq \|\mathbf{z}_1 - \mathbf{z}_0\|_2 + \|\mathbf{z}_0 - \overline{\mathbf{z}}\|_2 \leq \sqrt{2r_+}\sqrt{l_0} + \|\mathbf{z}_0 - \overline{\mathbf{z}}\|_2 < \rho.$$

Let us now prove by induction that $\mathbf{z}_k \in B(\overline{\mathbf{z}}, \rho)$, for all $k \geq 0$. We assume this holds true up to some $k \geq 0$. Hence, for $0 \leq i \leq k$, using $\mathbf{z}_i \in B(\overline{\mathbf{z}}, \rho)$ and $0 < l_i < \eta$ we can write the KL-inequality

$$\varphi'(l_i)\,\mathrm{dist}(\mathbf{0}, \partial L(\mathbf{z}_i)) \geq 1.$$

Lemma B.2.1.(iii) states that

$$\mathbf{z}_i^* := \begin{pmatrix} \nabla_{\mathbf{u}^1} Q(\mathbf{u}_i^1, \ldots, \mathbf{v}_i^R) - \nabla_{\mathbf{u}^1} Q(\mathbf{u}_i^1, \mathbf{u}_{i-1}^2, \ldots, \mathbf{v}_{i-1}^R) \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{1}{\lambda_{i-1}^1}(\mathbf{u}_i^1 - \mathbf{u}_{i-1}^1) \\ \vdots \\ \frac{1}{\mu_{i-1}^R}(\mathbf{v}_i^R - \mathbf{v}_{i-1}^R) \end{pmatrix}.$$

is an element of $\partial L(\mathbf{z}_i)$. So, we have

$$\varphi'(l_i) \|\mathbf{z}_i^*\|_2 \geq 1, \tag{B.5}$$

for all $1 \leq i \leq k$. Let us now examine $\|\mathbf{z}_i^*\|_2$, for $1 \leq i \leq k$. On the one hand,

$$\left\| \left(\frac{1}{\lambda_{i-1}^1}(\mathbf{u}_i^1 - \mathbf{u}_{i-1}^1), \ldots, \frac{1}{\mu_{i-1}^R}(\mathbf{v}_i^R - \mathbf{v}_{i-1}^R) \right) \right\|_2 \leq \frac{1}{r_-} \|\mathbf{z}_i - \mathbf{z}_{i-1}\|_2.$$

On the other hand, for arbitrary $s_t \in \{i-1, i\}$, $t \in \{1, \ldots, 2R\}$,

$$\begin{aligned} \|(\mathbf{u}_{s_1}^1, \ldots, \mathbf{v}_{s_{2R}}^R) - (\overline{\mathbf{u}}^1, \ldots, \overline{\mathbf{v}}^R)\|_2^2 &= \|\mathbf{u}_{s_1}^1 - \overline{\mathbf{u}}^1\|_2^2 + \cdots + \|\mathbf{v}_{s_{2R}}^R - \overline{\mathbf{v}}^R\|_2^2 \\ &\leq \|\mathbf{z}_{s_1} - \overline{\mathbf{z}}\|_2^2 + \cdots + \|\mathbf{z}_{s_{2R}} - \overline{\mathbf{z}}\|_2^2 \leq 2R\rho^2. \end{aligned}$$

Hence, $(\mathbf{u}_{s_1}^1, \ldots, \mathbf{v}_{s_{2R}}^R)$ and $\mathbf{z}_i$ lie in $B(\overline{\mathbf{z}}, \sqrt{2R}\rho)$. We can use Lipschitz-continuity of ∇Q to obtain

$$\|\nabla_{\boldsymbol{\xi}} Q(\mathbf{u}_{s_1}^1, \ldots, \mathbf{v}_{s_{2R}}^R) - \nabla_{\boldsymbol{\xi}} Q(\mathbf{u}_i^1, \ldots, \mathbf{v}_i^R)\|_2 \leq C \|\mathbf{z}_i - \mathbf{z}_{i-1}\|_2,$$

for any $\boldsymbol{\xi} \in \{\mathbf{u}^1, \ldots, \mathbf{v}^R\}$, which implies

$$\left\| \begin{pmatrix} \nabla_{\mathbf{u}^1} Q(\mathbf{u}_i^1, \ldots, \mathbf{v}_i^R) - \nabla_{\mathbf{u}^1} Q(\mathbf{u}_i^1, \mathbf{u}_{i-1}^2, \ldots, \mathbf{v}_{i-1}^R) \\ \vdots \\ 0 \end{pmatrix} \right\|_2 \leq C\sqrt{2R} \|\mathbf{z}_i - \mathbf{z}_{i-1}\|_2.$$

We get

$$\|\mathbf{z}_i^*\|_2 \leq (C\sqrt{2R} + \frac{1}{r_-}) \|\mathbf{z}_i - \mathbf{z}_{i-1}\|_2, \tag{B.6}$$

for all $1 \leq i \leq k$. Now (B.5) yields

$$\varphi'(l_i) \geq \frac{1}{C\sqrt{2R} + \frac{1}{r_-}} \|\mathbf{z}_i - \mathbf{z}_{i-1}\|_2^{-1}, \quad 1 \leq i \leq k,$$

and combined with (B.4)

$$\varphi(l_i) - \varphi(l_{i+1}) \geq \frac{1}{M} \frac{\|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2^2}{\|\mathbf{z}_i - \mathbf{z}_{i-1}\|_2}, \quad 1 \leq i \leq k.$$

This is equivalent to

$$\|\mathbf{z}_i - \mathbf{z}_{i-1}\|_2^{\frac{1}{2}} (M(\varphi(l_i) - \varphi(l_{i+1})))^{\frac{1}{2}} \geq \|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2$$

and, using $ab \leq (a^2 + b^2)/2$, gives

$$\|\mathbf{z}_i - \mathbf{z}_{i-1}\|_2 + M(\varphi(l_i) - \varphi(l_{i+1})) \geq 2\|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2, \quad 1 \leq i \leq k. \tag{B.7}$$

Summation over i leads to

$$\|\mathbf{z}_1 - \mathbf{z}_0\|_2 + M(\varphi(l_1) - \varphi(l_{k+1})) \geq \sum_{i=1}^{k} \|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2 + \|\mathbf{z}_{k+1} - \mathbf{z}_k\|_2.$$

Therefore, by using the monotonicity properties of φ and l_k

$$\|\mathbf{z}_1 - \mathbf{z}_0\|_2 + M\varphi(l_0) \geq \sum_{i=1}^{k} \|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2.$$

Finally,

$$\|\mathbf{z}_{k+1} - \overline{\mathbf{z}}\|_2 \leq \sum_{i=1}^{k} \|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2 + \|\mathbf{z}_1 - \overline{\mathbf{z}}\|_2 \leq M\varphi(l_0) + 2\sqrt{2r_+}\sqrt{l_0} + \|\mathbf{z}_0 - \overline{\mathbf{z}}\|_2 < \rho$$

which closes the induction and proves (i). Moreover, (B.7) holds for all $i \geq 1$. We can sum from k to K and get

$$\|\mathbf{z}_k - \mathbf{z}_{k-1}\|_2 + M(\varphi(l_k) - \varphi(l_{K+1})) \geq \sum_{i=k}^{K} \|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2 + \|\mathbf{z}_{K+1} - \mathbf{z}_K\|_2.$$

For $K \to \infty$, this becomes

$$\sum_{i=k}^{\infty} \|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2 \leq \|\mathbf{z}_k - \mathbf{z}_{k-1}\|_2 + M\varphi(l_k). \tag{B.8}$$

We conclude with (B.3) proving (ii)

$$\sum_{i=k}^{\infty} \|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2 \leq M\varphi(l_k) + \sqrt{2r_+}\sqrt{l_{k-1}} \leq M\varphi(l_{k-1}) + \sqrt{2r_+}\sqrt{l_{k-1}}.$$

This implies $\mathbf{z}_k$ is convergent and, therefore, its limit is a critical point as guaranteed by Lemma B.2.2. ■

We can now prove Theorem 7.6.2 and Theorem 7.6.3 which are straight-forward adaptions of the original results Theorem 10 and Theorem 11 in [7]. We state them for the sake of completeness.

Proof of Theorem 7.6.2 : To show the first part of the statement, note that by Theorem B.2.3 we get convergence of $(\mathbf{u}_k^1, ..., \mathbf{v}_k^R)$ to some $(\bar{\mathbf{u}}^1, ..., \bar{\mathbf{v}}^R) \in \operatorname{crit} L$ with $L(\bar{\mathbf{u}}^1, ..., \bar{\mathbf{v}}^R) \in [\min L, \min L + \eta)$. If $L(\mathbf{u}_*^1, ..., \mathbf{v}_*^R) \neq L(\bar{\mathbf{u}}^1, ..., \bar{\mathbf{v}}^R)$, we get by applying the KL-inequality that

$$\varphi'\left(L(\mathbf{u}_*^1, ..., \mathbf{v}_*^R) - L(\bar{\mathbf{u}}^1, ..., \bar{\mathbf{v}}^R)\right) \operatorname{dist}(\mathbf{0}, \partial(\mathbf{u}_*^1, ..., \mathbf{v}_*^R)) \geq 1$$

which contradicts $\mathbf{0} \in \partial(\mathbf{u}_*^1, ..., \mathbf{v}_*^R)$. The second part follows directly from Theorem B.2.3. ■

Proof of Theorem 7.6.3 : We use the notations of Theorem B.2.3 and assume for simplicity that $l_k \to 0$. Hence, by Lemma B.2.2 we have $L(\mathbf{u}_\infty^1, ..., \mathbf{v}_\infty^R) = 0$.

Assume first that $\theta = 0$. If (l_k) is stationary, then the same holds for $(\mathbf{u}_k^1, ..., \mathbf{v}_k^R)$ by Lemma B.2.1 (*i*). If (l_k) is not stationary, then the KL-inequality yields for any sufficiently large k that

$$c \operatorname{dist}(\mathbf{0}, \partial L(\mathbf{u}_k^1, ..., \mathbf{v}_k^R)) \geq 1$$

which contradicts Lemma B.2.1 (*iii*).

Assume $\theta > 0$. For any $k \in \mathbb{N}$, set $\Delta_k = \sum_{i=k}^{\infty} \|\mathbf{z}_{i+1} - \mathbf{z}_i\|_2$ which is finite by Theorem B.2.3. Since $\|\mathbf{z}_k - \mathbf{z}_\infty\|_2 \leq \Delta_k$ it suffices to bound Δ_k. By (B.8) we know that

$$\Delta_k \leq M\varphi(l_k) + (\Delta_{k-1} - \Delta_k). \tag{B.9}$$

The KL inequality yields

$$\varphi'(l_k) \operatorname{dist}(\mathbf{0}, \partial L(\mathbf{u}_k^1, ..., \mathbf{v}_k^R)) = c(1-\theta) l_k^{-\theta} \operatorname{dist}(\mathbf{0}, \partial L(\mathbf{u}_k^1, ..., \mathbf{v}_k^R)) \geq 1$$

and thus

$$l_k^\theta \leq c(1-\theta) \operatorname{dist}(L(\mathbf{u}_k^1, ..., \mathbf{v}_k^R)).$$

Using (B.6) and the definition of $\mathbf{z}_k^*$, we get

$$\operatorname{dist}(\mathbf{0}, \partial L(\mathbf{u}_k^1, ..., \mathbf{v}_k^R)) \leq \left(C + \frac{1}{r_-}\right)(\Delta_{k-1} - \Delta_k).$$

By combining the above estimates we obtain for some $K > 0$

$$\varphi(l_k) = c l_k^{1-\theta} \leq K(\Delta_{k-1} - \Delta_k)^{\frac{1-\theta}{\theta}}.$$

which shows with (B.9) that

$$\Delta_k \leq MK(\Delta_{k-1} - \Delta_k)^{\frac{1-\theta}{\theta}} + (\Delta_{k-1} - \Delta_k).$$

The statements (*ii*) and (*iii*) now follow from [6, Theorem 2]. ■

Bibliography

[1] D. Achlioptas, "Database-friendly random projections," *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 274–281, 2001.

[2] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, p. 4311, 2006.

[3] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *IEEE Transactions on Information Theory*, vol. 60, no. 3, pp. 1711–1732, 2014.

[4] H. Akaike, "Information theory and the maximum likelihood principle," in *International Symposium on Information Theory*, 1973.

[5] W. Allard, G. Chen, and M. Maggioni, "Multi-scale geometric methods for data sets II: Geometric multi-resolution analysis," *Applied and Computational Harmonic Analysis*, vol. 32, no. 3, pp. 435–462, 2012.

[6] H. Attouch and J. Bolte, "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features," *Mathematical Programming*, vol. 116, no. 1-2, pp. 5–16, 2009.

[7] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 438–457, 2010.

[8] L. Badea and D. Tilivea, "Sparse factorizations of gene expression data guided by binding data," in *Biocomputing 2005*, 2005, pp. 447–458.

[9] S. Bahmani and J. Romberg, "Near-optimal estimation of simultaneously sparse and low-rank matrices from nested linear measurements," *Information and Inference: A Journal of the IMA*, vol. 5, no. 3, pp. 331–351, 2016.

[10] K. Ball, "An elementary introduction to modern convex geometry," *Flavors of geometry*, vol. 31, pp. 1–58, 1997.

[11] S. Bandyopadhyay, C. Giannella, U. Maulik, H. Kargupta, K. Liu, and S. Datta, "Clustering distributed data streams in peer-to-peer environments," *Inform. Sciences*, vol. 176, no. 14, pp. 1952 – 1985, 2006.

[12] R. G. Baraniuk and M. B. Wakin, "Random projections of smooth manifolds," in *Foundations of Computational Mathematics*, 2006, pp. 941–944.

[13] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[14] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[15] R. G. Baraniuk, S. Foucart, D. Needell, Y. Plan, and M. Wootters, "Exponential decay of reconstruction error from binary measurements of sparse signals," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3368–3385, 2017.

[16] D. Baron, M. F. Duarte, M. B. Wakin, S. Sarvotham, and R. G. Baraniuk, "Distributed compressive sensing," *arXiv:0901.3403*, 2009.

[17] J. J. Benedetto, A. M. Powell, and Ö. Yılmaz, "Second-order Sigma-Delta (ΣΔ) quantization of finite frame expansions," *Appl. Comput. Harmon. Anal.*, vol. 20, no. 1, pp. 126 – 148, 2006.

[18] ——, "Sigma-Delta (ΣΔ) quantization and finite frames," *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 1990–2005, 2006.

[19] J. Bennett and S. Lanning, "The netflix prize," in *Proceedings of KDD cup and workshop*, vol. 2007, 2007, p. 35.

[20] A. Beygelzimer, S. Kakade, and J. Langford, "Cover trees for nearest neighbor," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 97–104.

[21] D. Bilyk and M. T. Lacey, "Random tessellations, restricted isometric embeddings, and one bit sensing," *arXiv:1512.06697*, 2015.

[22] P. Boufounos, L. Jacques, F. Krahmer, and R. Saab, "Quantization and compressive sensing," in *Compressed Sensing and its Applications.* Springer, 2015, pp. 193–237.

[23] P. T. Boufounos, "Greedy sparse signal reconstruction from sign measurements," in *Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, 2009, pp. 1305–1309.

[24] ——, "Universal rate-efficient scalar quantization," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1861 – 1872, 2012.

[25] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *42nd Annual Conference on Information Sciences and Systems (CISS)*, 2008, pp. 16–21.

[26] T. T. Cai and A. Zhang, "Sparse representation of a polytope and recovery of sparse signals and low-rank matrices." *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 122–132, 2014.

[27] A. R. Calderbank and I. Daubechies, "The pros and cons of democracy," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1721–1725, 2002.

[28] E. J. Candès and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[29] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, p. 717, 2009.

[30] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[31] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[32] G. Chen, M. Iwen, S. Chin, and M. Maggioni, "A fast multiscale framework for data in high-dimensions: Measure estimation, anomaly detection, and compressive measurements," in *2012 Visual Communications and Image Processing*, 2012, pp. 1–6.

[33] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[34] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," *Journal of the American mathematical society*, vol. 22, no. 1, pp. 211–231, 2009.

[35] R. R. Coifman, S. Lafon, M. Maggioni, Y. Keller, A. D. Szlam, F. J. Warner, and S. W. Zucker, "Geometries of sensor outputs, inference, and information processing," in *Intelligent Integrated Microsystems*, vol. 6232, 2006, p. 623209.

[36] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[37] J. A. Costa and A. O. Hero, "Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets," in *12th European Signal Processing Conference*, 2004, pp. 369–372.

[38] A. d'Aspremont, L. E. Ghaoui, M. I. Jordan, and G. R. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," in *Advances in neural information processing systems*, 2005, pp. 41–48.

[39] I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Communications on pure and applied mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[40] ——, “Sparsity-enforcing regularisation and ISTA revisited,” *Inverse Problems*, vol. 32, no. 10, p. 104001, 2016.

[41] M. A. Davenport and M. B. Wakin, “Analysis of orthogonal matching pursuit using the restricted isometry property,” *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4395–4401, 2010.

[42] M. A. Davenport, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, “A simple proof that random matrices are democratic,” *arXiv:0911.0736*, 2009.

[43] M. E. Davies and Y. C. Eldar, “Rank awareness in joint sparse recovery,” *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 1135–1146, 2012.

[44] G. Davis, S. Mallat, and M. Avellaneda, “Adaptive greedy approximations,” *Constructive approximation*, vol. 13, no. 1, pp. 57–98, 1997.

[45] Z. Ding and Y. Li, *Blind equalization and identification.* CRC press, 2001.

[46] S. Dirksen, H. C. Jung, and H. Rauhut, “One-bit compressed sensing with partial gaussian circulant matrices,” *arXiv:1710.03287*, 2017.

[47] S. Dirksen and S. Mendelson, “Non-gaussian hyperplane tessellations and robust one-bit compressed sensing,” *arXiv:1805.09409*, 2018.

[48] D. L. Donoho, “Compressed sensing,” *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[49] ——, “For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution,” *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 6, pp. 797–829, 2006.

[50] D. L. Donoho and X. Huo, “Uncertainty principles and ideal atomic decomposition,” *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[51] D. L. Donoho and B. F. Logan, “Signal recovery and the large sieve,” *SIAM Journal on Applied Mathematics*, vol. 52, no. 2, pp. 577–591, 1992.

[52] A. Dvoretzky, “Some results on convex bodies and Banach spaces,” *International Symposium on Linear Spaces*, 1961.

[53] ——, “A theorem on convex bodies and applications to Banach spaces,” *Proceedings of the National Academy of Sciences*, vol. 45, no. 2, pp. 223–226, 1959.

[54] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap.* CRC press, 1994.

[55] A. Eftekhari and M. B. Wakin, “New analysis of manifold embeddings and signal recovery from compressive measurements,” *Applied Computational Harmonic Analysis*, vol. 39, no. 1, pp. 67–109, 2015.

[56] A. Eftekhari, H. L. Yap, C. J. Rozell, and M. B. Wakin, “The restricted isometry property for random block diagonal matrices,” *Applied and Computational Harmonic Analysis*, vol. 38, no. 1, pp. 1–31, 2015.

[57] Y. Eldar and M. Mishali, “Robust recovery of signals from a structured union of subspaces,” *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[58] Y. C. Eldar and H. Rauhut, “Average case analysis of multichannel sparse recovery using convex relaxation,” *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 505–519, 2010.

[59] M. Fazel, H. Hindi, and S. P. Boyd, “A rank minimization heuristic with application to minimum order system approximation,” in *Proceedings of the 2001 American Control Conference*, vol. 6, 2001, pp. 4734–4739.

[60] H. Federer, “Curvature measures,” *Transactions of the AMS*, vol. 93, no. 3, pp. 418–491, 1959.

[61] J. Feng and F. Krahmer, “An RIP-based approach to Sigma-Delta quantization for compressed sensing,” *IEEE Signal Processing Letters*, vol. 21, no. 11, pp. 1351–1355, 2014.

[62] J. Feng, F. Krahmer, and R. Saab, “Quantized compressed sensing for partial random circulant matrices,” 2017, arXiv:1702.04711.

[63] M. Fornasier, J. Maly, and V. Naumova, “Robust recovery of low-rank matrices using multi-penalty regularization,” *NIPS workshop Optimization for Machine Learning*, 2017.

[64] ——, “Sparse PCA from inaccurate and incomplete measurements,” *arXiv:1801.06240*, 2018.

[65] S. Foucart, “Flavors of compressive sensing,” in *Approximation Theory XV: San Antonio 2016*. Springer, 2016.

[66] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Birkhäuser Basel, 2013.

[67] J. Geppert, F. Krahmer, and D. Stöger, “Sparse power factorization: Balancing peakiness and sample complexity,” *arXiv:1804.09097*, 2018.

[68] J. A. Geppert, F. Krahmer, and D. Stöger, “Refined performance guarantees for sparse power factorization,” in *International Conference on Sampling Theory and Applications (SampTA)*, 2017, pp. 509–513.

[69] A. Giannopoulos and V. Milman, “Asymptotic convex geometry short overview,” in *Different faces of geometry.* Springer, 2004, pp. 87–162.

[70] A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss, “Approximation of functions over redundant dictionaries using coherence,” in *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms.* Society for Industrial and Applied Mathematics, 2003, pp. 243–252.

[71] D. Godard, “Self-recovering equalization and carrier tracking in two-dimensional data communication systems,” *IEEE Transactions on Communications*, vol. 28, no. 11, pp. 1867–1875, 1980.

[72] G. H. Golub, M. Heath, and G. Wahba, “Generalized cross-validation as a method for choosing a good ridge parameter,” *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

[73] V. K. Goyal, M. Vetterli, and N. T. Thao, “Quantized overcomplete expansions in $\mathbf{R}^N$: analysis, synthesis, and algorithms,” *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 16–31, 1998.

[74] M. Grasmair, T. Klock, and V. Naumova, “Adaptive multi-penalty regularization based on a generalized LASSO path,” *arXiv*, 2017.

[75] M. Grasmair and V. Naumova, “Conditions on optimal support recovery in unmixing problems by means of multi-penalty regularization,” *Inverse Problems*, vol. 32, no. 10, p. 104007, 2016.

[76] R. Gray, “Oversampled Sigma-Delta modulation,” *IEEE Transactions on Communications*, vol. 35, no. 5, pp. 481–489, 1987.

[77] R. M. Gray and D. L. Neuhoff, “Quantization,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[78] O. Guédon and E. Milman, “Interpolating thin-shell and sharp large-deviation estimates for isotropic log-concave measures,” *Geometric and Functional Analysis*, vol. 21, no. 5, p. 1043, 2011.

[79] C. S. Güntürk, “Harmonic analysis of two problems in signal quantization and compression,” Ph.D. dissertation, Princeton University, 2000.

[80] S. Güntürk, M. Lammers, A. Powell, R. Saab, and Ö. Yılmaz, “Sobolev duals for random frames and $\Sigma\Delta$ quantization of compressed sensing measurements,” *Foundations of Computational Mathematics*, vol. 13, no. 1, pp. 1–36, 2013.

[81] V. Gupta, B. Kailkhura, T. Wimalajeewa, S. Liu, and P. K. Varshney, “Joint sparsity pattern recovery with 1-bit compressive sensing in sensor networks,” in *49th Asilomar Conference on Signals, Systems and Computers*, 2015, pp. 1472–1476.

[82] J. P. Haldar, D. Hernando, and Z.-P. Liang, “Compressed-sensing mri with random encoding,” *IEEE Transactions on Medical Imaging*, vol. 30, no. 4, pp. 893–903, 2011.

[83] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the L-curve," *SIAM review*, vol. 34, no. 4, pp. 561–580, 1992.

[84] S. Haykin, "The blind deconvolution problem," *Blind Deconvolution*, p. 1, 1994.

[85] M. A. Herman and T. Strohmer, "High-resolution radar via compressed sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 6, pp. 2275–2284, 2009.

[86] W. Hong, J. Wright, K. Huang, and Y. Ma, "Multiscale hybrid linear models for lossy image representation," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3655–3671, 2006.

[87] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, 1998, pp. 604–613.

[88] M. A. Iwen and F. Krahmer, "Fast subspace approximation via greedy least-squares," *Constructive Approximation*, vol. 42, no. 2, pp. 281–301, 2015.

[89] M. A. Iwen and M. Maggioni, "Approximation of points on low-dimensional manifolds via random linear projections," *Information and Inference: A Journal of the IMA*, vol. 2, no. 1, p. 1, 2013.

[90] M. A. Iwen and B. W. Ong, "A distributed and incremental SVD algorithm for agglomerative data analysis on large networks," *SIAM Journal on Matrix Analysis and Applications*, vol. 37, no. 4, pp. 1699–1718, 2016.

[91] M. A. Iwen, F. Krahmer, S. Krause-Solberg, and J. Maly, "On recovery guarantees for one-bit compressed sensing on manifolds," *arXiv:1807.06490*, 2018.

[92] L. Jacques, J. Laskas, P. Boufounos, and R. Baraniuk, "Robust 1-bit compressed sensing via binary stable embeddings of sparse vectors," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2082–2102, 2013.

[93] L. Jacques, "A quantized Johnson–Lindenstrauss lemma: The finding of Buffon's needle," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 5012–5027, 2015.

[94] ——, "Error decay of (almost) consistent signal estimations from quantized gaussian random projections," *IEEE Transactions on Information Theory*, vol. 62, no. 8, pp. 4696–4709, 2016.

[95] ——, "Small width, low distortions: quantized random embeddings of low-complexity sets," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5477–5495, 2017.

[96] L. Jacques, K. Degraux, and C. D. Vleeschouwer, "Quantized iterative hard thresholding: Bridging 1-bit and high-resolution quantized compressed sensing," in *International Conference on Sampling Theory and Applications*, 2013.

[97] L. Jacques, D. K. Hammond, and J. M. Fadili, "Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 559–571, 2011.

[98] R. Jagannathan and T. Ma, "Risk reduction in large portfolios: Why imposing the wrong constraints helps," *The Journal of Finance*, vol. 58, no. 4, pp. 1651–1683, 2003.

[99] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674, 2013.

[100] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemporary mathematics*, vol. 26, no. 189-206, p. 1, 1984.

[101] I. Jolliffe, "Principal component analysis," in *International encyclopedia of statistical science*. Springer, 2011, pp. 1094–1096.

[102] P. Jung, F. Krahmer, and D. Stöger, "Blind demixing and deconvolution at near-optimal rate," *IEEE Transactions on Information Theory*, vol. 64, pp. 704–727, 2018.

[103] S. Kafle, B. Kailkhura, T. Wimalajeewa, and P. K. Varshney, "Decentralized joint sparsity pattern recovery using 1-bit compressive sensing," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2016, pp. 1354–1358.

[104] K. Knudson, R. Saab, and R. Ward, "One-bit compressive sensing with norm estimation," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2748–2758, 2016.

[105] A. Kolleck and J. Vybíral, "Non-asymptotic analysis of l1-norm support vector machines," *IEEE Transactions on Information Theory*, 2017.

[106] F. Krahmer, R. Saab, and Ö. Yilmaz, "Sigma-delta quantization of sub-gaussian frame expansions and its application to compressed sensing," *Information and Inference: A Journal of the IMA*, vol. 3, no. 1, pp. 40–58, 2014.

[107] F. Krahmer, S. Mendelson, and H. Rauhut, "Suprema of chaos processes and the restricted isometry property," *Communications on Pure and Applied Mathematics*, vol. 67, no. 11, pp. 1877–1904, 2014.

[108] S. Krause-Solberg and J. Maly, "A tractable approach for one-bit compressed sensing on manifolds," in *International Conference on Sampling Theory and Applications (SampTA)*, 2017, pp. 667–671.

[109] D. Krawczyk-StańDo and M. Rudnicki, "Regularization parameter selection in discrete ill-posed problems - the use of the U-curve," *International Journal of Applied Mathematics and Computer Science*, vol. 17, no. 2, pp. 157–164, 2007.

[110] V. Lacko and R. Harman, "A conditional distribution approach to uniform sampling on spheres and balls in l_p spaces," *Metrika*, vol. 75, no. 7, pp. 939–951, 2012.

[111] J. N. Laska, P. T. Boufounos, M. A. Davenport, and R. G. Baraniuk, "Democracy in action: Quantization, saturation, and compressive sensing," *Applied and Computational Harmonic Analysis*, vol. 31, no. 3, pp. 429–443, 2011.

[112] J. N. Laska, "Regime change: Sampling rate vs. bit-depth in compressive sensing," Ph.D. dissertation, Rice University, 2011.

[113] Y. LeCun, C. Cortes, and C. J. Burges, "The MNIST database of handwritten digits," online database, 19.12.2017.

[114] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer Verlag, 2002.

[115] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

[116] K. Lee, Y. Wu, and Y. Bresler, "Near-optimal compressed sensing of a class of sparse low-rank matrices via sparse power factorization," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1666–1698, 2018.

[117] W. Liao and M. Maggioni, "Adaptive geometric multiscale approximations for intrinsically low-dimensional data," *arXiv:1611.01179v2*, 2017.

[118] S. Ling and T. Strohmer, "Blind deconvolution meets blind demixing: Algorithms and performance bounds," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4497–4520, 2017.

[119] ——, "Regularized gradient descent: a non-convex recipe for fast joint blind deconvolution and demixing," *Information and Inference: A Journal of the IMA*, 2017.

[120] N. Linial, E. London, and Y. Rabinovich, "The geometry of graphs and some of its algorithmic applications," *Combinatorica*, vol. 15, no. 2, pp. 215–245, 1995.

[121] A. V. Little, Y.-M. Jung, and M. Maggioni, "Multiscale estimation of intrinsic dimensionality of data sets." in *AAAI fall symposium: manifold learning and its applications*, vol. 9, 2009, p. 04.

[122] B. F. Logan, "Properties of high-pass signals," Ph.D. dissertation, Columbia University, 1965.

[123] M. Maggioni, S. Minsker, and N. Strawn, "Multiscale dictionary learning: Non-asymptotic bounds and robustness," *Journal of Machine Learning Research*, vol. 7, no. 1, pp. 43–93, 2016.

[124] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," Courant Institute of Mathematical Sciences New York United States, Tech. Rep., 1993.

[125] J. Maly and L. Palzer, "Analysis of hard-thresholding for distributed compressed sensing with one-bit measurements," *to appear in Information and Inference: A journal of the IMA; arXiv:1805.03486*, 2018.

[126] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Reconstruction and subgaussian operators in asymptotic geometric analysis," *Geometric and Functional Analysis*, vol. 17, no. 4, pp. 1248–1282, 2007.

[127] V. Milman, "Surprising geometric phenomena in high-dimensional convexity theory," in *European Congress of Mathematics*. Springer, 1998, pp. 73–91.

[128] A. Moshtaghpour, L. Jacques, V. Cambareri, K. Degraux, and C. De Vleeschouwer, "Consistent basis pursuit for signal and matrix estimates in quantized compressed sensing," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 25–29, 2016.

[129] M. Murphy, M. Alley, J. Demmel, K. Keutzer, S. Vasanawala, and M. Lustig, "Fast ℓ_1-spirit compressed sensing parallel imaging mri: Scalable parallel implementation and clinically feasible runtime," *IEEE Transactions on Medical Imaging*, vol. 31, no. 6, pp. 1250–1262, 2012.

[130] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM journal on computing*, vol. 24, no. 2, pp. 227–234, 1995.

[131] V. Naumova and S. Peter, "Minimization of multi-penalty functionals by alternating iterative thresholding and optimal parameter choices," *Inverse Problems*, vol. 30, no. 12, p. 125003, 2014.

[132] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Foundations of Computational Mathematics*, vol. 9, no. 3, pp. 317–334, 2009.

[133] S. R. Norsworthy, R. Schreier, and G. C. Temes, Eds., *Delta-Sigma-Converters: Theory, Design and Simulation.* Wiley-IEEE, 1996.

[134] B. Oliver, J. Pierce, and C. E. Shannon, "The philosophy of PCM," *Proceedings of the IRE*, vol. 36, no. 11, pp. 1324–1331, 1948.

[135] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2886–2908, 2015.

[136] N. Parikh, S. Boyd *et al.*, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[137] G. Pisier, *The volume of convex bodies and Banach space geometry.* Cambridge University Press, 1999, vol. 94.

[138] Y. Plan and R. Vershynin, "One-bit compressed sensing by linear programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1275–1297, 2013.

[139] ——, “Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach,” *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 482–494, 2013.

[140] ——, “Dimension reduction by random hyperplane tessellations,” *Discrete & Computational Geometry*, vol. 51, no. 2, pp. 438–461, 2014.

[141] A. M. Powell and J. T. Whitehouse, “Error bounds for consistent reconstruction: random polytopes and coverage processes,” *Foundations of Computational Mathematics*, vol. 16, no. 2, pp. 395–423, 2016.

[142] L. S. Prichep, E. Causevic, R. R. Coifman, R. Isenhart, A. Jacquin, E. R. John, M. Maggioni, and F. J. Warner, “Qeeg-based classification with wavelet packet and microstate features for triage applications in the ER,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, 2006.

[143] X. Rao and V. K. N. Lau, “Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems,” *IEEE Transaction on Signal Processing*, vol. 62, no. 12, pp. 3261 – 3271, 2014.

[144] B. Recht, M. Fazel, and P. A. Parrilo, “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[145] M. Rudelson and R. Vershynin, “On sparse reconstruction from Fourier and Gaussian measurements,” *Communications on Pure and Applied Mathematics*, vol. 61, no. 8, pp. 1025–1045, 2008.

[146] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D: nonlinear phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[147] R. Saab, R. Wang, and Ö. Yilmaz, “Quantization of compressive samples with stable and robust recovery,” *Applied Computational Harmonic Analysis*, vol. 44, no. 1, pp. 123 – 143, 2018.

[148] G. Schechtman, “Two observations regarding embedding subsets of euclidean spaces in normed spaces,” *Advances in Mathematics*, vol. 200, no. 1, pp. 125–135, 2006.

[149] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[150] C. E. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[151] ——, “Coding theorems for a discrete source with a fidelity criterion,” *IRE International Convention Record*, vol. 7, pp. 142–163, 1959.

[152] X. Shen and J. Ye, “Adaptive Model Selection,” *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 210–221, 2002.

[153] C. S. Signorino and A. Kirchner, “Using LASSO to Model Interactions and Nonlinearities in Survey Data,” *Survey Practice*, vol. 11, no. 1, p. 2716, 2018.

[154] I. Steinwart and A. Christmann, *Support vector machines.* Springer Science & Business Media, 2008.

[155] T. G. Stockham, T. M. Cannon, and R. B. Ingebretsen, “Blind deconvolution through digital signal processing,” *Proceedings of the IEEE*, vol. 63, no. 4, pp. 678–692, 1975.

[156] Y. Sugaya and K. Kanatani, “Multi-stage unsupervised learning for multi-body motion segmentation,” *IEICE Transactions on Information and Systems*, vol. 87, no. 7, pp. 1935–1942, 2004.

[157] M. Talagrand, *The generic chaining: upper and lower bounds of stochastic processes.* Springer Science & Business Media, 2006.

[158] ——, *Upper and lower bounds for stochastic processes: modern methods and classical problems.* Springer Science & Business Media, 2014, vol. 60.

[159] Y. Tian, W. Xu, Y. Wang, and H. Yang, “A distributed compressed sensing scheme based on one-bit quantization,” in *IEEE 79th Vehicular Technology Conference (VTC Spring)*, 2014, pp. 1–6.

[160] R. Tibshirani, “Regression shrinkage and selection via the LASSO,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[161] R. J. Tibshirani, “The LASSO problem and uniqueness,” *Electronic Journal of Statistics*, vol. 7, pp. 1456–1490, 2013.

[162] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[163] J. A. Tropp, “Greed is good: Algorithmic results for sparse approximation,” *IEEE Transactions on Information theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[164] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” in *Compressed Sensing: Theory and Applications.* Cambridge University Press, 2012, pp. 210–268.

[165] ——, “Estimation in high dimensions: a geometric perspective,” in *Sampling theory, a renaissance.* Springer, 2015, pp. 3–66.

[166] ——, *High-dimensional probability: An introduction with applications in data science.* Cambridge University Press, 2018, vol. 47.

[167] E. D. Vito, M. Fornasier, and V. Naumova, “A machine learning approach to optimal tikhonov regularisation I: Affine manifolds,” *arXiv:1610.01952*, 2016.

[168] M. B. Wakin, "The geometry of low-dimensional signal models," Ph.D. dissertation, Rice University, 2007.

[169] J. Wewerka, "Near-optimal data-driven l1-regularization," *Master's Thesis*, 2018.

[170] Y. Wu, Y.-J. Zhu, Q.-Y. Tang, C. Zhu, W. Liu, R.-B. Dai, X. Liu, E. X. Wu, L. Ying, and D. Liang, "Accelerated MR diffusion tensor imaging using distributed compressed sensing," *Magnetic Resonance in Medicine*, vol. 71, no. 2, pp. 764 – 772, 2014.

[171] C. Xu and L. Jacques, "Quantized compressive sensing with RIP matrices: The benefit of dithering," *arXiv:1801.05870*, 2018.

[172] R. Zass and A. Shashua, "Nonnegative sparse PCA," in *Advances in neural information processing systems*, 2007, pp. 1561–1568.

[173] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.