Technische Universität München

Department of Mathematics

Master's Thesis

Non-Convex Approaches to Compressed Sensing and Robust Recovery of Simultaneously Structured Signals from Inaccurate and Incomplete Information

Konstantin Riedl

Supervisor: Prof. Dr. Massimo Fornasier
Advisor: Dr. Johannes Maly
Submission Date: November 19, 2019

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Konstantin Riedl

Munich, November 19, 2019

Abstract

Signals having multiple structures simultaneously appear in several applications in signal processing and machine learning throughout science and engineering. One is often confronted with the problem of their robust and resource efficient recovery from inaccurate and incomplete information.

This thesis is concerned with the reconstruction of low-rank matrices possessing a non-orthogonal decomposition with partially sparse and non-negative component vectors. Our employed numerical algorithm is based on alternating minimization of a highly non-convex multi-penalty functional associated with the matrix decomposition. It comprises both data fidelity and the individual low-dimensional structures. A particular focus of our proposed method is on structure identification.

In order to analyze this approach theoretically as well as numerically we provide concise overviews of the fields of compressed sensing and low-rank matrix recovery before outlining recent developments regarding simultaneously structured models. By introducing a novel class of effectively sparse and low-rank matrices together with a suitable restricted isometry property, we are able to show successful recovery up to noise level from a number of random measurements scaling, up to a polylogarithmic factor, linearly in the intrinsic dimension of the signal. With this we improve upon previous results.

Zusammenfassung

Signale, welche mehrere Strukturen gleichzeitig besitzen, treten in etlichen Anwendungen aus Wissenschaft und Technik in Bereichen der Signalverarbeitung und dem maschinellen Lernen auf. Oft steht man dem Problem ihrer stabilen und ressourcenschonenden Wiederherstellung aus fehlerhafter und unvollständiger Information gegenüber.

Diese Arbeit beschäftigt sich mit der Rekonstruktion von Matrizen mit niedrigerem Rang, welche eine nichtorthogonale Zerlegung mit zum Teil dünn besetzten und nichtnegativen Komponenten besitzen. Unser verwendeter numerischer Algorithmus basiert auf alternierender Minimierung eines stark nichtkonvexen Funktionals, welches aus zahlreichen mit der Matrixzerlegung assoziierten Straftermen besteht. Es vereint Datentreue mit den individuellen niedrigdimensionalen Strukturen. Ein spezieller Fokus unserer vorgeschlagen Methode liegt dabei auf der Identifikation der involvierten Strukturen.

Um jenen Ansatz theoretisch sowie numerisch zu analysieren, geben wir prägnante Überblicke über die Felder des Compressed Sensing und der Wiederherstellung von Matrizen mit niedrigem Rang, bevor wir aktuelle Entwicklungen bezüglich gleichzeitig strukturierter Modelle skizzieren. Indem wir eine neuartige Klasse von effektiv dünn besetzten Matrizen niedrigen Ranges zusammen mit einer passenden eingeschränkten Isometrieeigenschaft einführen, können wir deren erfolgreiche Wiederherstellung bis auf Fehlerniveau zeigen, falls sich die Anzahl an zufälligen Messungen bis auf einen Faktor polylogarithmischer Ordnung proportional zur intrinsischen Dimension des Signals verhält. Wir verbessern damit frühere Resultate.

Acknowledgements

During my time as a Master's student at the chair for Applied and Numerical Analysis and Optimization and Data Analysis, I enjoyed the pleasure to be introduced to the rapidly evolving mathematical field of compressed sensing and machine learning by attending the lecture "Mathematical Foundations of Artificial Neural Networks" held for its first time by Prof. Massimo Fornasier. He provided us students an insight into very active mathematical research by presenting parts of the work [FVD19], he and his colleagues were working on at that time.

Attracted by the intriguing topic, Prof. Massimo Fornasier's honest commitment to my development already during the lecture and the chance to work on an interesting research topic, I was glad to have the opportunity to conduct my Master's thesis with him as my supervisor. For this I would like to express my deep gratitude towards him. I benefited from his inspiring, motivating and encouraging nature as well as his enthusiasm for mathematics in many ways, personally and professionally.

Moreover, I wish to sincerely thank my advisor Dr. Johannes Maly for all the fruitful discussions we had, all the mails we wrote and for being available whenever I had a question, an idea or a problem. His expertise and his constant guidance were invaluable for me finding my way in this broad and rather new field of mathematics. I am greatly appreciative of his innumerable helpful suggestions, improvements and advices.

Furthermore, I gratefully acknowledge the compute and data resources provided by the Leibniz Supercomputing Centre (LRZ).

And last, but undeniably most importantly, I am very much obliged to my friends and family for their never-ending support and encouragement.

Contents

Introduction

Mankind is about to reach an unprecedented and unimaginable scale of annually generated data. Within the research project DATA AGE 2025 the International Data Corporation (IDC) published the White Paper "The Digitalization of the World: From Edge to Core" [RGR18]. They forecast that the overall generated and captured data per year, which they call the Global Datasphere, will grow from 41 zettabytes in 2019 to 175 zettabytes by the year 2025. One zettabyte is equal to one thousand exabytes or one trillion gigabytes. To put these numbers into perspective, let us state a quote from 2010 by former Google CEO ERIC SCHMIDT [Van13].

> *There were five exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days.*
>
> —ERIC SCHMIDT

Even though the numbers in the first part of the statement are certainly exaggerated [Moo11], the underlying message stands for its own. And at least the data created, captured or replicated by humans every two days in near future will exceed the data generated between the origin of humanity and the early 2000s.

Due to this data deluge, nowadays available processing power as well as data storage capacities and transmission possibilities are brought to their limits [Eco10].

This raises the question if all the data acquired by modern high-resolution sensors—and therefore stored and processed, even if this takes only milliseconds—is actually necessary. To illustrate that this is far from being the case, let us give an example from everybody's daily life—a photo from a customary smartphone camera. In the case of an ordinary 10 megapixel camera, a total number of ten million individual measurements is required to be taken, before all the raw data for the later image is available. This digital negative needs roughly 10MB of storage (assuming an 8-bit color depth). However, that's typically not what is saved to our photo library. Image compression techniques, such as, e.g., JPEG, drastically reduce the size by about 80% without admitting any noticeable loss of quality visible to the human eye. Similar methods exist for digital audio, like, e.g., MP3, and for video footage, like, e.g., MPEG.

Any of those compression techniques relies on the empirically observed fact that real-word data is compressible—at least with respect to an appropriate basis or, more generally, a suitable dictionary or frame. That means, we are able to approximate the coefficients of the signal in this basis by a sparse vector without allowing a significant error.

Thinking through this process of first acquiring the full data by taking a high number of measurements and subsequently discarding a large proportion thereof, makes this seem to be a waste of resources. This may be not decisive when taking a snapshot with a commercial camera, but if sensors are expensive, which, for instance, is the case for infrared light, it is desirable to reduce the number of required sensors to take pictures. A proof of concept that in principle one single sensor suffices to generate enough data to be able to recover a whole image, was given with the single-pixel camera [Mac09, DDT$^+$08]. A different instance, where it is also very reasonable to rethink if every measurement and the related time to generate a high-resolution image is necessary, is magnetic resonance imaging (MRI) [LDSP08].

The desire of acquiring a compressed version of a compressible signal with significantly less measurements, which were originally believed to be incomplete and therefore useless information, gave rise to the theory of compressed sensing. This field of very active development is also known as compressive sampling or sparse recovery. The seminal works [CRT06a, CT06, CRT06b] by Candès, Romberg and Tao and [Don06] by Donoho initiated the research. Similar, but mostly application-oriented approaches can be found already much earlier, see, e.g., [FR15, Section 6.2] for an overview of such early findings.

Having this in mind, one may wonder about the value of incomplete data. A very prominent example outlining how this could constitute is the Netflix Prize problem [BL07], where one aims at recommending movies to a user by analyzing the preferences of other but similar customers. More precisely, in collaborative filtering [GNOT92] in general, one is interested in predicting a particular user's taste based on collected available interests of other users in an automated way. The underlying idea relates again to some sort of parsimony, namely low-rankness of the data matrix. This is reasonable as in practice only few factors contribute to a user's preferences.

In order to overcome such issues, sparse recovery and compressed sensing were generalized to matrices. The framework for the matrix completion problem as described above emerged with [CR09, CT10]. A more general problem formulation, the so-called matrix sensing problem, was investigated in [RFP10].

Problems involving low-rank matrices and the need to recover them from measurements arise in various applications in science and engineering. To name just a few examples, in quantum state tomography approximately pure quantum states are modeled as positive semidefinite matrices of low rank with unit trace [GLF$^+$10, Gro11]. In control and system theory the state of a low-order time-invariant system can be described by a low-rank Hankel matrix, which one may want to detect in system identification problems [LV09]. Also in machine learning and data mining sparsity and low-rankness are ubiquitous, underlying several data models in various applications of the fields, like, e.g., face or voice recognition, intelligent searching, natural language processing and medical diagnosis.

Some of these models exhibit multiple structures simultaneously, making it desirable to take advantage of the signal lying in several different unions of low-dimensional manifolds at the same time. A very common instance in this setting are low-rank matrices, which are additionally sparse in the sense that they admit some kind of sparse decomposition. In order to illustrate how these different structures arise in a real-world problem, let us take up the example of a recommendation system for a grocery store described in [FMN19]. Therefore, imagine a data matrix with rows corresponding to the customers of the store and columns representing the available products. Each entry of this matrix describes the probability that a certain customer purchases a certain product. Since there are just a few basic factors, which have an impact on the purchase behavior, such as gender, age, financial status, family or lifestyle, it is reasonable to suspect that this matrix is of low rank. Any customer is influenced by these basic factors in an individual manner. In

turn, any basic factor determines a very specific and pronounced buying pattern, which motivates to assume that besides low-rankness also sparsity is involved in our model. The first observation can be rephrased to how much a certain basis factor influences each customer. This gives rise to two component vectors for each of the basic factors, which eventually form a non-orthogonal rank-1 decomposition with partially sparse component vectors of our low-rank data matrix. In fact, even more structure can be found in this model. After a proper rescaling one component vector can be regarded as a discrete probability distribution, i.e., it has only positive entries, which add up to one. This model is very similar to the one from the Netflix prize problem mentioned previously as a paradigm for matrix completion. However, due to customers tending to maintain their purchase behavior even in case of small random prize fluctuations, the problem of recovering the low-rank data matrix can be regarded as an instance of matrix sensing by acquiring information from aggregated revenues.
The study of simultaneously structured models is a relatively new but upcoming branch of the field of matrix recovery. In turn, the idea of reducing complexity from a set of data by extracting the most relevant directions and revealing the essential underlying information is comparably old [Pea01]. Principal component analysis [Jol02] established itself in dimension reduction and unsupervised learning. However, despite being widely spread throughout statistics, data analysis and machine learning, it lacks interpretability, which is a severe disadvantage. To impose additional structure enhancing the desired interpretability, sparse principal component analysis was introduced in [ZHT06]. Instead of requiring orthogonality of the principal components, sparsity of the same is promoted. This leads to a non-orthogonal low-rank decomposition or approximation involving partially sparse components. The recovery of such matrices from a few linear measurements was investigated from a theoretical point of view in [OJF$^+$15] showing that tractable convex approaches are not capable of taking advantage of multiple structures. Non-convex formulations, however, are able to recover low-rank matrices possessing a sparse decomposition from few measurements of the order of the information theoretic limit.

Organization. The thesis is organized as follows. Chapter 1 gives a concise overview of the quickly evolving mathematical field of compressed sensing. In Chapter 2 we extend this framework to the recovery of low-rank matrices from inaccurate and incomplete information. Based thereon we investigate simultaneously structured signals in Chapter 3. In particular, low-rank matrices which admit a non-orthogonal sparse decomposition are considered. The main focus of Chapter 4 is also on this signal class. We propose a novel, highly non-convex approach based on alternating minimization to recover such matrices from noisy and only partially available linear measurements. This generalizes the work of [FMN19]. Eventually, Chapter 5 supports the preceding chapter with numerical experiments. Beyond that, we analyze numerically if the number of necessary measurements can be reduced if a further structure, namely positivity in the right components of the low-rank decomposition, is added. We conclude with a further discussion.

Notation. Let us provide an overview of the notation being used throughout the thesis. We denote the set of natural numbers by $\mathbb{N} = \{1, 2, \dots\}$ and abbreviate $[n] = \{1, \dots, n\}$ for $n \in \mathbb{N}$. Scalars are denoted by math italic uppercase and lowercase letters. For matrices and vectors we use uppercase and lowercase bold letters, respectively, i.e., $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$ is a matrix and $\mathbf{z} \in \mathbb{R}^N$ a vector. Consequently, the ith row of the matrix $\mathbf{Z}$ is denoted by $\mathbf{z}^i$, the jth column by $\mathbf{z}_j$ and the entry they have in common by z_{ij}. For the latter, we may also make use of the notation $(\mathbf{Z})_{ij}$. Analogously, both z_i and $(\mathbf{z})_i$ denote the ith entry of the vector $\mathbf{z}$.

We furthermore introduce the support of $\mathbf{z}$ as $\operatorname{supp}(\mathbf{z}) = \{i \in [N] : z_i \neq 0\}$ and refer to its size as the ℓ_0-norm $\|\mathbf{z}\|_0$ of $\mathbf{z}$. However, we want to emphasize that this notion is misleading, as $\|\cdot\|_0$ is not even a quasi-norm, since it is not absolutely homogeneous. Moreover, we will heavily use the ℓ_q-(quasi)-norms, which are defined for $0 < q \leq \infty$ as

$$\|\mathbf{z}\|_q = \left(\sum_{i=1}^{N} |z_i|^q\right)^{1/q} \qquad \text{and} \qquad \|\mathbf{z}\|_\infty = \sup_{i\in[N]} |z_i|,$$

respectively. They are norms for $1 \leq q \leq \infty$ and (quasi)-norms for $0 < q < 1$, since they only obey the quasi-triangle inequality in this case. The so-called ℓ_0-norm can be interpreted as the limit of the ℓ_q-(quasi)-norms for $q \to 0$, since $\|\mathbf{z}\|_q^q \to \|\mathbf{z}\|_0$, cf. Section 1.2. Based on the former definition we introduce the N-dimensional ℓ_q-(quasi)-norm-balls of radius r and centered in $\mathbf{z}$ as $\mathcal{B}_q^N(\mathbf{z}, r)$ and abbreviate $\mathcal{B}_q^N(\mathbf{0}, 1)$ with $\mathcal{B}_q^N$. They are convex, when a norm is underlying. The $(N-1)$-dimensional Euclidean unit sphere, i.e., $\partial\mathcal{B}_2^N$, is denoted by $\mathbb{S}^{N-1} \subset \mathbb{R}^N$.

As for vectors, we will employ a variety of matrix norms. Therefore let us first recall some notions from basic linear algebra. Namely, by $\operatorname{rank}(\mathbf{Z})$ we denote the rank of the matrix $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$. Furthermore, we write $\mathcal{R}(\mathbf{Z})$ for the range or image of the matrix $\mathbf{Z}$, which is of dimension $\operatorname{rank}(\mathbf{Z})$. The nullspace or kernel of $\mathbf{Z}$ is denoted by $\mathcal{N}(\mathbf{Z})$ or $\ker(\mathbf{Z})$ and is $(n_2 - \operatorname{rank}(\mathbf{Z}))$-dimensional according to the rank-nullity theorem.

Now, let us introduce the singular value decomposition of a rank-R matrix $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$ as the product

$$\mathbf{Z} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_{r=1}^{R} \sigma_r \mathbf{u}_r \mathbf{v}_r^T,$$

where $\sigma_1(\mathbf{Z}) \geq \cdots \geq \sigma_R(\mathbf{Z}) > 0$ denote the singular values of $\mathbf{Z}$. They may also appear arranged in a vector $\boldsymbol{\sigma}(\mathbf{Z}) \in \mathbb{R}^R$. For ease of notation and as already done above, we sometimes abbreviate $\boldsymbol{\sigma} = \boldsymbol{\sigma}(\mathbf{Z})$ and analogously for the singular vectors and the individual singular values. The left and right singular vectors, $\mathbf{u}_1, \ldots, \mathbf{u}_R \in \mathbb{R}^{n_1}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_R \in \mathbb{R}^{n_2}$, are collected in the matrices $\mathbf{U} \in \mathbb{R}^{n_1 \times R}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times R}$ and form an orthonormal basis of $\mathcal{R}(\mathbf{Z})$ and $\mathcal{N}(\mathbf{Z})^\perp$, respectively. That means, $\mathbf{U} \in \mathbb{R}^{n_1 \times R}$ as well as $\mathbf{V} \in \mathbb{R}^{n_2 \times R}$ are orthogonal rectangular matrices, i.e., they satisfy $\mathbf{U}^T\mathbf{U} = \mathrm{Id} = \mathbf{V}^T\mathbf{V}$. Here, the identity matrix, mapping $\mathbb{R}^R$ into itself, is denoted by Id, which, more generally, labels the identity operator on arbitrary spaces. $\boldsymbol{\Sigma} \in \mathbb{R}^{R \times R}$ is a positive definite diagonal matrix, such that $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\sigma})$. The operator diag maps a vector $\boldsymbol{\sigma}$ onto a diagonal matrix $\boldsymbol{\Sigma}$, such that $\sigma_{ii} = \sigma_i$ holds for all i. Note, however, that the same notation may be also used to extract the diagonal of a maybe non-diagonal matrix and rearrange it into a vector. Positive definiteness of $\boldsymbol{\Sigma}$ can be denoted in symbols by $\boldsymbol{\Sigma} \succ 0$. Accordingly, we will use the character $\succeq$ to indicate that a matrix is positive semidefinite. The singular value decomposition introduced before is also known as the rank-reduced singular value decomposition. In contrast, a full singular value decomposition of the form $\mathbf{Z} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ comes with orthogonal squared matrices $\mathbf{U} \in \mathbb{R}^{n_1 \times n_1}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times n_2}$ satisfying $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathrm{Id} \in \mathbb{R}^{n_1 \times n_1}$ and analogously for $\mathbf{V}$. The diagonal matrix $\boldsymbol{\Sigma}$ is then of the same size as the original matrix and has the R singular values on its diagonal along with zeros filling up the diagonal.

We can now naturally generalize the concept of the ℓ_q-(quasi)-norms to matrices by defining the Schatten-q-(quasi)-norm $\|\mathbf{Z}\|_q$ of the matrix $\mathbf{Z}$ as the ℓ_q-(quasi)-norm of its associated vector $\boldsymbol{\sigma}(\mathbf{Z})$ of singular values, i.e.,

$$\|\mathbf{Z}\|_q = \|\boldsymbol{\sigma}(\mathbf{Z})\|_q = \left(\sum_{r=1}^{R} \sigma_r(\mathbf{Z})^q\right)^{1/q} \qquad \text{and} \qquad \|\mathbf{Z}\|_\infty = \sigma_1(\mathbf{Z}),$$

respectively. A few candidates deserve some special attention and come along with their own labeling. First, for $q = \infty$, we obtain the operator norm $\|\mathbf{Z}\| = \|\mathbf{Z}\|_\infty$. Second, for $q = 2$, the definition yields the Frobenius norm $\|\mathbf{Z}\|_F = \|\mathbf{Z}\|_2 = \operatorname{tr}(\mathbf{Z}^T\mathbf{Z})^{1/2}$, which is induced by the Frobenius scalar product $\langle \mathbf{Z}_1, \mathbf{Z}_2 \rangle_F = \operatorname{tr}(\mathbf{Z}_1^T\mathbf{Z}_2) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\mathbf{Z}_1)_{ij}(\mathbf{Z}_2)_{ij}$, where $\mathbf{Z}_\ell \in \mathbb{R}^{n_1 \times n_2}$ for $\ell = 1, 2$. Last, for $q = 1$, we get the nuclear norm $\|\mathbf{Z}\|_* = \|\mathbf{Z}\|_1$, whose $*$-notation originates from being the dual norm of the operator norm $\|\mathbf{Z}\|$. Moreover, we can identify the rank of $\mathbf{Z}$ with $\|\mathbf{Z}\|_0$, which gives the number of singular values, i.e., $\|\mathbf{Z}\|_0 = \operatorname{rank} \mathbf{Z}$. However, this is of course no quasi-norm.

Returning to vectors, for $\mathbf{z} \in \mathbb{R}^N$ and an index set $T \subset [N]$ we introduce the restriction of $\mathbf{z}$ to this set as $\mathbf{z}|_T$ by setting all entries outside of T, i.e., on T^c, to zero. This generalizes to matrices in the form of submatrices. We denote a submatrix of $\mathbf{Z}$ consisting of the columns with indices $j \in \Lambda \subset [n_2]$ by $\mathbf{Z}|_\Lambda$. This works analogously for $\mathbf{Z}|^\Lambda$. Also associated with matrices is the vectorization $\operatorname{vec} \mathbf{Z} \in \mathbb{R}^{n_1 n_2}$ of the matrix $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$, which is the vector resulting from stacking the columns of $\mathbf{Z}$ on top of each other.

To denote an operator mapping matrices to vectors we use calligraphic capital letters, such as $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$, $\mathbf{Z} \mapsto \mathbf{y} = \mathcal{A}(\mathbf{Z})$. If this is a linear map it can be described by an $(m \times n_1 n_2)$-dimensional matrix by employing the vectorization of $\mathbf{Z}$. This can also be reshaped properly into m individual Frobenius scalar products involving $(n_1 \times n_2)$-dimensional matrices $\mathbf{A}_i$ for $i \in [m]$, cf. Section 2.2. We moreover use the calligraphic letter $\mathcal{O}$ as part of the Landau notation.

Additionally to matrices we use bold capital letters also for subspaces and insinuate that it is in general clear from the context whether we refer to a matrix or the subspace spanned by its columns.

We use the notation $\mathbf{x}^k$ also for the kth iterate in an iterative algorithm. This causes a certain ambiguity, however, it will be clear from the context to what we refer.

The indicator function of an arbitrary set $K \subset \mathbb{R}^N$ is denoted by $\mathbb{1}_K(\mathbf{z})$. Furthermore, we call $K^\#$ an ϵ-net or ϵ-cover of K with respect to a metric d, if for any $\mathbf{z} \in K$ there exists a $\mathbf{z}^\# \in K^\#$ such that $d(\mathbf{z}, \mathbf{z}^\#) \leq \epsilon$. The ϵ-covering number $N(K, d, \epsilon)$ of K is the smallest cardinality of any ϵ-net, which we call minimal ϵ-net. In this thesis we only consider internal covers, i.e., we require $K^\# \subset K$. If d is induced by a norm $\|\cdot\|$, we use the notation $N(K, \|\cdot\|, \epsilon)$.

Let $(\Omega, \mathcal{F}, \mathsf{P})$ denote a probability space. The probability of an event $F \in \mathcal{F}$ is denoted by $\mathsf{P}(F)$, the expectation of a random variable $X : \Omega \to E$, where $(E, \mathcal{E})$ is a measurable space, is denoted by $\mathsf{E}X = \int_\Omega X(\omega)\, d\mathsf{P}(\omega)$. Moreover, we denote the normal distribution with expectation $\boldsymbol{\mu} \in \mathbb{R}^N$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}$ by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

For the sake of simplicity we write $a \lesssim b$ to express that there exists an absolute constant $C > 0$ such that $a \leq Cb$. The same applies to $\gtrsim$ and with $a \simeq b$ we mean that $a \lesssim b$ and $a \gtrsim b$ hold simultaneously.

Further notation is in general introduced when it first appears.

Chapter 1

Compressed Sensing

In this first chapter we give an outline of the vast field of compressed sensing. Starting from basic notions such as sparsity and compressibility, we investigate the question of how to identify sparse signals using a number of measurements which is related to their intrinsic information content rather than the ambient dimension. We particularly focus on lower bounds on the number of necessary measurements, on how to design a measurement process and on how to realize the recovery efficiently via specialized algorithms.

For a more comprehensive presentation providing a much deeper insight we refer to the monograph [FR13] by Foucart and Rauhut as well as the compendium [FR15] by Fornasier and Rauhut and the paper [CDD09] by Cohen, Dahmen and DeVore. For Section 1.4 in particular we recommend [For10] by Fornasier.

1.1 Sparsity and Compressibility

We motivated in the introduction that real-word signals can usually be well approximated by sparse expansions with respect to an appropriate basis. This, however, is based solely on an empirical observation, which seems to underly nature in many situations. In order to make the notions of sparsity and compressibility more rigorous, let us start by defining the set of s-sparse vectors Σ_s^N. A signal $\mathbf{z} \in \mathbb{R}^N$ is called s-sparse, if at most s of its entries are non-zero, i.e.,

$$\|\mathbf{z}\|_0 := |\operatorname{supp}(\mathbf{z})| \leq s. \tag{1.1}$$

Despite being not even a quasi-norm, $\|\cdot\|_0$ is colloquially termed ℓ_0-norm. For convenience, let us establish the notion of the relative sparsity of an s-sparse signal $\mathbf{z}$ as the quotient s/N. Moreover, let us briefly comment on the geometry of the set of s-sparse vectors. Given a fixed support set of size s, all s-sparse vectors supported thereon form an s-dimensional subspace. As there are $\binom{N}{s}$ different support sets, Σ_s^N is the union of just as many subspaces of dimension s. Since the sum of two s-sparse vectors from different subspaces may have up to $2s$ non-zero components, this set is non-convex.

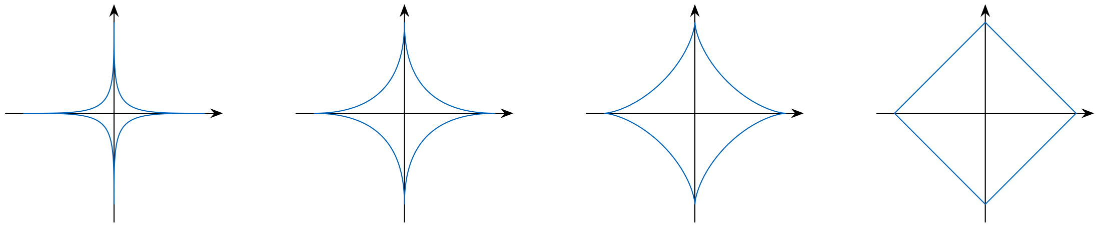
As already mentioned, assuming merely compressibility instead of sparsity is more realistic in applications. Compressible vectors are characterized by the property that they can be well approximated by sparse ones. In order to make this more precise, for $p > 0$ let us define the best s-term approximation of $\mathbf{z}$ as

$$\mathbf{z}_{[s]} := \operatorname*{arg\,inf}_{\tilde{\mathbf{z}} \in \Sigma_s^N} \|\mathbf{z} - \tilde{\mathbf{z}}\|_p. \tag{1.2}$$

Independently of p, $\mathbf{z}_{[s]}$ can be obtained by retaining only the s in absolute value largest components of $\mathbf{z}$ and setting all the remaining ones to zero. The quality of this non-linear approximation is quantified by the best s-term approximation error of a vector $\mathbf{z} \in \mathbb{R}^N$ with respect to the ℓ_p-(quasi)-norm, given by

$$\sigma_s(\mathbf{z})_p := \left\|\mathbf{z} - \mathbf{z}_{[s]}\right\|_p = \inf_{\tilde{\mathbf{z}} \in \Sigma_s^N} \left\|\mathbf{z} - \tilde{\mathbf{z}}\right\|_p. \tag{1.3}$$

Now, a signal $\mathbf{z} \in \mathbb{R}^N$ is called compressible if $\sigma_s(\mathbf{z})_p$ decays quickly in s. A prominent example for compressible vectors are the ℓ_q-balls $\mathcal{B}_q^N$ for small $q \leq 1$, which are sketched in Figure 1.1 for different values of q.



(a) $q = 1/3$ (b) $q = 1/2$ (c) $q = 2/3$ (d) $q = 1$

Figure 1.1. Geometry of the ℓ_q-(quasi)-norm-balls $\mathcal{B}_q^2$ for different values of q.

The following lemma formalizes this observation by establishing a bound on the best s-term approximation error in terms of suitable ℓ_q-(quasi)-norms.

Lemma 1.1 (Stechkin's Inequality)**.** *For $p \geq q > 0$ and $\mathbf{z} \in \mathbb{R}^N$ it holds*

$$\sigma_s(\mathbf{z})_p \leq s^{-(1/q - 1/p)} \left\|\mathbf{z}\right\|_q. \tag{1.4}$$

Proof. Without loss of generality let us assume that $\mathbf{z}$ is given in its nonincreasing rearrangement, i.e., $|z_i| \geq |z_j|$ for all $i < j$. Then, $s|z_s|^q \leq \sum_{i=1}^s |z_i|^q \leq \|\mathbf{z}\|_q^q$. Therefore,

$$\sigma_s(\mathbf{z})_p^p = \sum_{i=s+1}^N |z_i|^p \leq |z_s|^{p-q} \sum_{i=s+1}^N |z_i|^q \leq s^{-(p-q)/q} \|\mathbf{z}\|_q^{p-q} \|\mathbf{z}\|_q^q \leq s^{-(p-q)/q} \|\mathbf{z}\|_q^p, \tag{1.5}$$

which implies the claim by taking the pth root on both sides. □

A variation of this lemma and therefore a different definition of compressibility resembling the one of sparsity can be found in [FR13, Proposition 2.11]. There, a vector $\mathbf{z} \in \mathbb{R}^N$ is called compressible if, for some threshold $t > 0$, the number of its significant, i.e., in absolute value larger than t, components is small. The ℓ_q-(quasi)-norm in equation (1.4) can then essentially be replaced by the weak ℓ_q-(quasi)-norm, cf. [DeV98].

1.2 Sparse Solutions of Underdetermined Systems

As already mentioned in the introduction, sensors measuring all kinds of data are everywhere. In natural sciences, such as physics, chemistry or biology, as well as applied sciences, such as engineering and technology or medicine, the quantities of interest are frequently only available indirectly in terms of their measurements. This often necessitates to recover a high-dimensional signal $\mathbf{x} \in \mathbb{R}^N$ from the observed data $\mathbf{y} \in \mathbb{R}^m$. Assuming that the measurement procedure is linear, it can be described by

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{1.6}$$

where the matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ models the information acquisition process and is referred to as the encoder or encoding matrix. In the following we require $\mathbf{A}$ to have full rank, i.e., $\operatorname{rank}(\mathbf{A}) = \min\{m, N\}$. From standard linear algebra we know that, in general, $m = N$ is needed for unique solvability of equation (1.6). We, in turn, are interested in the undersampled case $m < N$. Without further assumptions on the signal $\mathbf{x}$, an identification of the desired one is impossible as there are infinitely many solutions to the, in this case, underdetermined linear system.

However, as motivated in the introduction, typical real-world signals admit a certain structural feature, namely compressibility or in an idealized setting, sparsity. That means, even though being embedded into the high-dimensional space $\mathbb{R}^N$, their intrinsic complexity is much smaller.

In this case, the natural question to ask is whether we are able to reduce the number of necessary measurements such that it depends on the intrinsic information content rather than the ambient dimension. That this is indeed theoretically possible is the subject of the following statement.

Theorem 1.2. *Suppose that the measurement matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ is such that every set of $2s$ columns of $\mathbf{A}$ is linearly independent. Then every s-sparse vector $\mathbf{x} \in \mathbb{R}^N$ can be reconstructed uniquely from its measurements $\mathbf{y} = \mathbf{A}\mathbf{x}$.*

Proof. Let us assume that unique reconstruction cannot be performed, i.e., there are two different s-sparse vectors $\mathbf{x}, \mathbf{x}'$ such that $\mathbf{A}\mathbf{x} = \mathbf{y} = \mathbf{A}\mathbf{x}'$. Then, by linearity, $\mathbf{A}(\mathbf{x}-\mathbf{x}') = \mathbf{0}$ and since $\mathbf{x} - \mathbf{x}'$ is $2s$-sparse, $2s$ columns of $\mathbf{A}$ are linearly dependent. □

Moreover, one can actually show that the condition on the matrix $\mathbf{A}$ from Theorem 1.2 is also necessary, cf. [FR13, Theorem 2.13]. Therefrom, for the number of required measurements we deduce $m \geq 2s$, as any submatrix containing $2s$ columns of $\mathbf{A}$ needs to be of rank $2s$. Vandermonde-type matrices, for instance, can provide suitably designed measurement matrices, which ensure this property, cf. [FR13, Theorem 2.14 and 2.15].

Let us put this differently. Assuming that we are capable of designing our measurement process appropriately, i.e., such that it can be described by the matrix $\mathbf{A}$, we are able to guarantee that the non-convex optimization program

$$\min_{\mathbf{z} \in \mathbb{R}^N} \|\mathbf{z}\|_0 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y} \tag{1.7}$$

yields the correct solution $\mathbf{x}$. We call this the ℓ_0-minimization problem and denote its minimizer by $\Delta_0(\mathbf{y})$. In general, typically non-linear maps of the type $\Delta : \mathbb{R}^m \to \mathbb{R}^N$, implementing a reconstruction method trying to recover the signal $\mathbf{x}$ from the information held by the measurements $\mathbf{y}$ are referred to as decoders[1].

Prohibitively, however, the optimization problem (1.7), related to the decoder Δ_0, turns out to be NP-hard for arbitrary matrices $\mathbf{A}$ and right-hand sides $\mathbf{y}$ [Nat95]. This entails

[1] We want to emphasize at this point that compressed sensing brings together the—in classical theory strictly separated—worlds of sampling, which is totally linear, and compression, which is highly non-linear.

that any algorithmic approach becomes computationally intractable for a large surrounding dimension N.

Furthermore, for the case that N is large, suitable measurement matrices $\mathbf{A}$ are poorly conditioned, which leads to numerical instabilities when recovering a signal $\mathbf{x}$ from measurements of the form (1.6).

Though being foremost of theoretical interest, the non-convex and discontinuous ℓ_0-minimization problem (1.7) serves as starting point for essentially two classes of tractable approaches. Firstly, greedy methods established themselves. A well-known representative of this class is orthogonal matching pursuit, which will be addressed briefly in Subsection 1.4.1. Secondly, methods based on convex relaxation were proposed, including convex optimization based techniques and iterative thresholding methods, such as iterative soft thresholding. Convex relaxation will be from a theoretical perspective the main focus of Section 1.3 and from an algorithmic one the main part of Section 1.4.

Before that, however, we want to give a quick overview of the path from non-convex and discontinuous ℓ_0-minimization over non-convex but continuous relaxation thereof towards the in so many ways successful convex and continuous ℓ_1-minimization problem. This will then finally be introduced at the beginning of Section 1.3.

The key observation for relaxation is the convergence of the qth powers of the ℓ_q-(quasi)-norms to the ℓ_0-norm monotonously from below, i.e.,

$$\|\mathbf{z}\|_q^q = \sum_{i=1}^{N} |z_i|^q \longrightarrow \sum_{i=1}^{N} \mathbb{1}_{\{z_i \neq 0\}} = \|\mathbf{z}\|_0 \tag{1.8}$$

monotonously from below for $q \to 0$ monotonously. As a consequence, the continuous ℓ_q-(quasi)-norms can be seen as approximations to the discontinuous and non-convex ℓ_0-norm. For $q < 1$, however, the ℓ_q-quasi-norms are still non-convex, which results in the relaxed ℓ_q-minimization problem

$$\min_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{z}\|_q^q \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}, \tag{1.9}$$

which is still NP-hard [GJY11]. Note that, without imposing certain conditions on $\mathbf{A}$ and the sparsity of $\mathbf{x}$, (1.9) is just an approximation to (1.7). Analogously to before, we denote the minimizer of (1.9) by $\Delta_q(\mathbf{y})$.

1.3 Recovery of Sparse Signals

Candès, Romberg and Tao [CRT06a] and Donoho [Don06] proposed the decoder

$$\min_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{z}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}, \tag{1.10}$$

which is called ℓ_1-minimization or basis pursuit. It can be regarded as the convex relaxation of ℓ_0-minimization (1.7) and admits several promising properties. First, it is a convex optimization problem and can even be recast as a linear program [CDS98]. Therefore, standard techniques from convex optimization can be employed, such as the simplex algorithm or the interior-point method. Second, and this was one of the remarkable observations of the papers cited at the beginning of this section, conditions on the matrix $\mathbf{A}$ and the sparsity of $\mathbf{x}$ were established which assure that ℓ_1-minimization recovers the correct solution $\mathbf{x}$. Furthermore, it was shown that such measurement matrices can be constructed at random with high probability and, moreover, that the necessary number of measurements behaves like

$$m \gtrsim s \log\left(\frac{eN}{s}\right). \tag{1.11}$$

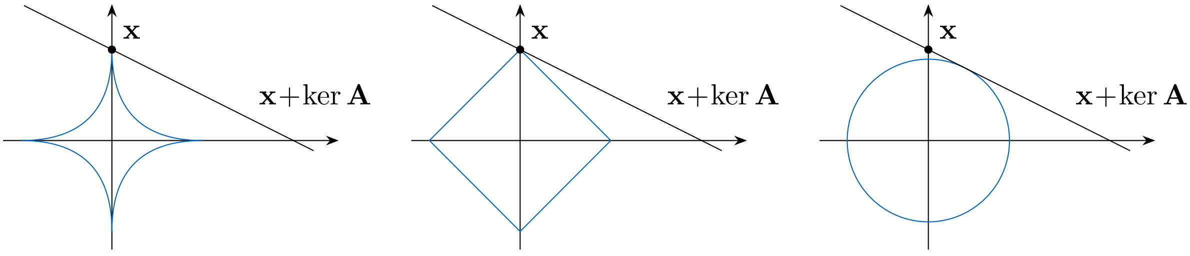
Compared to the lower bound following from Theorem 1.2, $m \geq 2s$, this is worse, as the ambient dimension N is involved. However, it appears only logarithmically. In fact, this number is necessary to guarantee stable recovery. Third, the decoder Δ_1, given by (1.10), is stable. The notion of stability refers to possible sparsity defects mirroring typical disturbances to signals from real-world applications, cf. the concept of compressibility from Section 1.1. And fourth, the decoder is also robust with respect to noise in the measurements [CT05, CRT06b]. That means, if we are facing the situation, where we wish to recover the signal $\mathbf{x}$ from inaccurate measurements

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\eta}, \tag{1.12}$$

the quality of the recovery is only affected by an additional error of order $\mathcal{O}(\|\boldsymbol{\eta}\|_2)$. Here, $\boldsymbol{\eta} \in \mathbb{R}^m$ denotes unknown noise corrupting the sensing process. We will discuss this more closely in Subsection 1.3.5, but also refer to [FR13, Section 4.3] for a detailed presentation. At first, however, let us consider the noiseless case.

1.3.1 On how ℓ_1-Minimization Promotes Sparsity

An illustrative and intuitive explanation why ℓ_1-minimization works so well is given in Figure 1.2. We compare it with the on the right-hand side depicted ℓ_2-minimization, which is strongly related to Tikhonov regularization[2]. Furthermore, for comparison and in order to ease the imagination of ℓ_0-minimization, non-convex ℓ_q-minimization[3] is sketched for $q = 1/2$ on the left-hand side.



(a) $q = 1/2$ (b) $q = 1$ (c) $q = 2$

Figure 1.2. Comparison of ℓ_1-minimization (middle) with ℓ_2-minimization (right) and ℓ_q-minimization for $q = 1/2$ (left). Depicted are, for each of the three situations, the solution set $\{\mathbf{z} : \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}\}$ and the suitably inflated ℓ_q-(quasi)-norm ball.

The figure suggests that, contrarily to ℓ_2-minimization, ℓ_1-minimization is capable of recovering the sparsest solution to the linear system (1.6), i.e., it holds $\Delta_1(\mathbf{y}) = \Delta_0(\mathbf{y})$. However, there are certain situations where this cannot be guaranteed, namely when the

[2] In statistics, the term ridge regression is the more common one.

[3] Originated from a play on words, the regularized version of (1.9) is also called bridge regression as it literally builds a bridge between ℓ_0-penalization and ridge regression [Tib96].

$(N-m)$-dimensional subspace $\ker \mathbf{A}$ is aligned with one of the faces of the ℓ_1-norm ball. Therefore, a necessary condition on the kernel of $\mathbf{A}$, the so-called null space property, was introduced in [CDD09]. Actually, without being mentioned explicitly though, it appeared already much earlier. It turns out that this property is equivalent to sparse recovery via ℓ_1-minimization, as we will show in Theorem 1.4.

1.3.2 The Null Space Property

We argued intuitively that the kernel of the encoder $\mathbf{A}$ plays the central role for the question whether the problems (1.7) and (1.10) admit the same solution. The following property of the measurement matrix turns out to be fundamental.

Definition 1.3 (Null Space Property (NSP)). *A matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$ satisfies the null space property of order s with constant $0 < \gamma < 1$, if*

$$\|\mathbf{z}|_S\|_1 \leq \gamma \|\mathbf{z}|_{S^c}\|_1 \tag{1.13}$$

for all sets $S \subset [N]$ such that $|S| \leq s$ and for all $\mathbf{z} \in \ker \mathbf{A}\backslash\{\mathbf{0}\}$. Moreover, if we refer to it without mentioning the constant γ, we require $\|\mathbf{z}|_S\|_1 < \|\mathbf{z}|_{S^c}\|_1$ in place of (1.13).

The null space property essentially states that the kernel of the measurement matrix $\mathbf{A}$ must not contain vectors, where some entries are significantly larger in magnitude than the remaining ones. This needs to hold in particular for sparse and compressible vectors. Namely, if this were not the case, they could not be distinguished from zero and thus a stable recovery would not be possible.
In the following theorem we now formalize what was already mentioned previously. The null space property is a necessary and sufficient condition for recovering sparse vectors exactly via ℓ_1-minimization.

Theorem 1.4 (Equivalence between the Null Space Property and ℓ_1-Recovery). *A given matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$ satisfies the null space property of order s, if and only if every s-sparse vector $\mathbf{x} \in \mathbb{R}^N$ is the unique solution to the ℓ_1-minimization problem (1.10) with $\mathbf{y} = \mathbf{A}\mathbf{x}$.*

Proof. Let us assume that $\mathbf{A}$ has the null space property of order s. Furthermore, let $\mathbf{x} \in \Sigma_s^N$ and denote its support by S. Then, for a vector $\mathbf{z} \in \mathbf{x} + \ker \mathbf{A}$, $\mathbf{z} \neq \mathbf{x}$, we observe that $\mathbf{v} := \mathbf{x} - \mathbf{z} \in \ker \mathbf{A}\backslash\{\mathbf{0}\}$. Thus, by the null space property,

$$\begin{aligned} \|\mathbf{x}\|_1 \leq \|\mathbf{x}-\mathbf{z}|_S\|_1 + \|\mathbf{z}|_S\|_1 &= \|\mathbf{v}|_S\|_1 + \|\mathbf{z}|_S\|_1 \\ &< \|\mathbf{v}|_{S^c}\|_1 + \|\mathbf{z}|_S\|_1 = \|-\mathbf{z}|_{S^c}\|_1 + \|\mathbf{z}|_S\|_1 = \|\mathbf{z}\|_1, \end{aligned} \tag{1.14}$$

which shows the optimality of $\mathbf{x}$.
Conversely, let us assume that every s-sparse vector $\mathbf{x} \in \mathbb{R}^N$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{Az} = \mathbf{Ax}$. Then, for any index set S and any vector $\mathbf{v} \in \ker \mathbf{A}\backslash\{\mathbf{0}\}$, $\mathbf{v}|_S$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{Az} = \mathbf{Av}|_S$. However, as $\mathbf{v} \in \ker \mathbf{A}$, we have $\mathbf{A}(-\mathbf{v}|_{S^c}) = \mathbf{Av}|_S$ and $-\mathbf{v}|_{S^c} \neq \mathbf{v}|_S$ as $\mathbf{v} \neq \mathbf{0}$. Consequently, $\|\mathbf{v}|_S\|_1 < \|-\mathbf{v}|_{S^c}\|_1$. □

Remark 1.5 (Null Space Property for the Decoder Δ_q). Replacing condition (1.13) in Definition 1.3 by $\|\mathbf{z}|_S\|_q^q \leq \gamma\|\mathbf{z}|_{S^c}\|_q^q$ provides a suitable null space property for ℓ_q-minimization, see, e.g., [GPYZ15, Pet16]. Then, Theorem 1.4 can be reformulated analogously for

ℓ_q-minimization. Moreover, it can be shown that ℓ_q-minimization provides an approximation to (1.7), which improves for $q \to 0$. Stated more precisely, for $0 < q < p \leq 1$, sparse recovery via ℓ_p-minimization implies sparse recovery via ℓ_q-minimization [FR13, Theorem 4.9].

For practical applications it is indispensable for our reconstruction method to be stable in the sense that a compressible signal $\mathbf{x}$ can be recovered by only admitting an error, which is comparable to the best s-term approximation error. That this is the case for ℓ_1-minimization is made more rigorous in the next theorem.

Theorem 1.6 ([FR13, Theorem 4.12]). *Let us assume that $\mathbf{A} \in \mathbb{R}^{m \times N}$ has the null space property of order s with constant $0 < \gamma < 1$. Then, for any $\mathbf{x} \in \mathbb{R}^N$, a solution $\hat{\mathbf{x}}$ to the ℓ_1-minimization problem* (1.10) *with $\mathbf{y} = \mathbf{A}\mathbf{x}$ fulfills*

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_1 \leq \frac{2(1+\gamma)}{1-\gamma}\sigma_s(\mathbf{x})_1. \tag{1.15}$$

Proof. First observe that $\mathbf{v} := \mathbf{x} - \hat{\mathbf{x}} \in \ker \mathbf{A}$. Moreover, as $\hat{\mathbf{x}}$ is a solution to (1.10), it holds $\|\hat{\mathbf{x}}\|_1 \leq \|\mathbf{x}\|_1$. Let us denote the set of the s, in absolute value largest entries of $\mathbf{x}$, by S. Then,

$$\|\hat{\mathbf{x}}|_S\|_1 + \|\hat{\mathbf{x}}|_{S^c}\|_1 \leq \|\mathbf{x}|_S\|_1 + \|\mathbf{x}|_{S^c}\|_1 \tag{1.16}$$

and by the reverse triangle inequality

$$\|\mathbf{x}|_S\|_1 - \|\mathbf{v}|_S\|_1 + \|\mathbf{v}|_{S^c}\|_1 - \|\mathbf{x}|_{S^c}\|_1 \leq \|\mathbf{x}|_S\|_1 + \|\mathbf{x}|_{S^c}\|_1. \tag{1.17}$$

From this we deduce

$$\|\mathbf{v}|_{S^c}\|_1 \leq 2\|\mathbf{x}|_{S^c}\|_1 + \|\mathbf{v}|_S\|_1 \leq 2\sigma_s(\mathbf{x})_1 + \gamma\|\mathbf{v}|_{S^c}\|_1, \tag{1.18}$$

where, besides the definition of best s-term approximation, the null space property of order s with constant γ was utilized in the last inequality. We can reformulate this equivalently and obtain

$$\|\mathbf{v}|_{S^c}\|_1 \leq \frac{2}{1-\gamma}\sigma_s(\mathbf{x})_1. \tag{1.19}$$

Eventually, employing the null space property once more in the next-to-last inequality, we get with inequality (1.19) that

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_1 = \|\mathbf{v}\|_1 = \|\mathbf{v}|_S\|_1 + \|\mathbf{v}|_{S^c}\|_1 \leq (1+\gamma)\|\mathbf{v}|_{S^c}\|_1 \leq \frac{2(1+\gamma)}{1-\gamma}\sigma_s(\mathbf{x})_1, \tag{1.20}$$

yielding the claim. □

Because of Theorem 1.6, the null space property of order s with constant γ is often referred to as the stable null space property of order s with constant γ.
A common term, related to (1.15), appearing frequently in the literature is instance optimality [CDD09]. An encoder-decoder pair $(\mathbf{A}, \Delta)$ is referred to as instance optimal of order s with constant C with respect to the (quasi)-norm $\|\cdot\|_X$, if

$$\|\mathbf{z} - \Delta(\mathbf{A}\mathbf{z})\|_X \leq C\sigma_s(\mathbf{z})_X \tag{1.21}$$

for all $\mathbf{z} \in \mathbb{R}^N$. An immediate consequence of instance optimal of order s is that s-sparse signals $\mathbf{z}$ can be recovered exactly.

Remark 1.7 (Instance Optimality of the Decoder Δ_q)**.** Assuming the modified null space property introduced in Remark 1.5 for the measurement matrix $\mathbf{A}$ one can show that the pair $(\mathbf{A}, \Delta_q)$ is actually instance optimal with respect to the (quasi)-norm $\|\cdot\|_q$, see, e.g., [GPYZ15, Pet16].

Despite having these favorable results it is inconvenient to work with the null space property in practice, as it is hard to verify for a given matrix $\mathbf{A}$. In its stead, a stronger condition was introduced in [CT06].

1.3.3 The Restricted Isometry Property

From standard linear algebra it is well-known that a linear map, described by a matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$, cannot preserve distances between arbitrary points, when $m < N$, i.e., it cannot be an isometry. However, as we are not interested in sensing arbitrary vectors $\mathbf{z} \in \mathbb{R}^N$, there is no need for $\mathbf{A}$ to retain the whole geometry of $\mathbb{R}^N$. It suffices if $\mathbf{A}$ behaves like an isometry when restricted to Σ_s^N. This was firstly investigated in [CT06] under the name uniform uncertainty principle and is now known as the restricted isometry property [CT05].

Definition 1.8 (Restricted Isometry Property (RIP))**.** *A matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$ satisfies the restricted isometry property of order s with isometry constant $0 < \delta < 1$, if*

$$(1-\delta)\|\mathbf{z}\|_2^2 \leq \|\mathbf{A}\mathbf{z}\|_2^2 \leq (1+\delta)\|\mathbf{z}\|_2^2 \tag{1.22}$$

for all $\mathbf{z} \in \Sigma_s^N$.

Clearly, if $\mathbf{A}$ has the restricted isometry property of order $2s$, any two different s-sparse vectors $\mathbf{z}$ and $\mathbf{z}'$ are distinguishable by their measurements. Consequently, $\mathbf{x}$ is the unique s-sparse solution amenable to the measurements $\mathbf{y} = \mathbf{A}\mathbf{x}$, cf. [CT05, Lemma 1.2].

Remark 1.9. In some situations it is relevant to be able to consider RIP constants δ associated with different orders s. To distinguish them, we add the order as a subscript, i.e., we write δ_s.

The restricted isometry property is stronger than the null space property, as specified in the following result. Its proof employs a common technique in the compressed sensing literature.

Theorem 1.10 (The Restricted Isometry Property implies the Null Space Property, [FR15, Lemma 2])**.** *Let us assume that $\mathbf{A} \in \mathbb{R}^{m\times N}$ has the restricted isometry property of order $s+t$ with constant $0 < \delta < 1$. Then, $\mathbf{A}$ has the null space property of order s with constant $\gamma = \sqrt{\frac{s(1+\delta)}{t(1-\delta)}}$.*

Proof. Let $\mathbf{v} \in \ker \mathbf{A}$ and define a partition

$$\mathcal{T} = \{T_\ell : |T_0| = s \text{ and } |T_\ell| = t \text{ for all } 0 < \ell < \lceil (n-s)/t \rceil\}_{\ell=0}^{\lceil (n-s)/t \rceil} \tag{1.23}$$

of $[n]$ associated with a nonincreasing rearrangement of $\mathbf{v}$, i.e., for all $\ell \geq 1$ it holds

$$|v_i| \leq |v_j| \quad \text{for all } i \in T_\ell \text{ and } j \in T_{\ell-1}. \tag{1.24}$$

Then, after employing Cauchy-Schwarz inequality, the restricted isometry property, using that $\mathbf{v} \in \ker \mathbf{A}$, applying the triangle inequality and eventually the restricted isometry property once more, we obtain

$$\begin{aligned} \|\mathbf{v}|_{T_0}\|_1 &\leq \sqrt{s}\|\mathbf{v}|_{T_0}\|_2 \leq \sqrt{s}\|\mathbf{v}|_{T_0 \cup T_1}\|_2 \leq \sqrt{s}\sqrt{\frac{1}{1-\delta}}\|\mathbf{A}\mathbf{v}|_{T_0 \cup T_1}\|_2 \\ &= \sqrt{s}\sqrt{\frac{1}{1-\delta}}\left\|\mathbf{A}\mathbf{v}|_{T_2 \cup T_3 \cup \cdots \cup T_{\lceil (n-s)/t \rceil}}\right\|_2 \leq \sqrt{s}\sqrt{\frac{1}{1-\delta}} \sum_{\ell=2}^{\lceil (n-s)/t \rceil} \|\mathbf{A}\mathbf{v}|_{T_\ell}\|_2 \\ &\leq \sqrt{s}\sqrt{\frac{1+\delta}{1-\delta}} \sum_{\ell=2}^{\lceil (n-s)/t \rceil} \|\mathbf{v}|_{T_\ell}\|_2. \end{aligned} \tag{1.25}$$

In order to upper bound the last term, we note that by definition of the partition $\mathcal{T}$, firstly, for $\ell \geq 2$, summation over j yields

$$|v_i| \leq t^{-1}\left\|\mathbf{v}|_{T_{\ell-1}}\right\|_1 \tag{1.26}$$

for all $i \in T_\ell$. Secondly, taking the ℓ^2-norm over $i \in T_\ell$ shows

$$\|\mathbf{v}|_{T_\ell}\|_2 \leq t^{-1/2}\left\|\mathbf{v}|_{T_{\ell-1}}\right\|_1. \tag{1.27}$$

Using this estimate in (1.25) yields

$$\|\mathbf{v}|_{T_0}\|_1 \leq \sqrt{\frac{s}{t}}\sqrt{\frac{1+\delta}{1-\delta}} \sum_{\ell=2}^{\lceil (n-s)/t \rceil} \left\|\mathbf{v}|_{T_{\ell-1}}\right\|_1 \leq \sqrt{\frac{s}{t}}\sqrt{\frac{1+\delta}{1-\delta}}\left\|\mathbf{v}|_{T_0^c}\right\|_1, \tag{1.28}$$

establishing the null space property of order s with the claimed constant γ. □

It remains to discuss two relevant connected questions for RIP matrices. Namely, how they can be constructed and which size of m is necessary therefore. It is first worth mentioning that currently no deterministic measurement matrices are available, which are provably optimal in the sense that they match the bounds on the number of measurements we will derive in Subsection 1.3.4. To date, this is an open problem. However, randomization helps out, as recognized in [CRT06a, CT06]. Random matrices, such as Gaussian or Bernoulli matrices, turn out to satisfy the restricted isometry property with high probability provided m fulfills (1.11). This is made more precise below in Theorem 1.11 for measurement matrices with Gaussian entries. The proof we present, exploits a link between the Johnson-Lindenstrauss Lemma [JL84] and the restricted isometry property and was given in [BDDW08].

Theorem 1.11 (Gaussian Matrices are RIP Matrices, cf. [BDDW08, Theorem 5.2])**.** *Let $\mathbf{A} \in \mathbb{R}^{m \times N}$ be a random matrix with i.i.d. mean-zero Gaussian entries of unit variance, and assume that*

$$m \geq Cs \log\left(\frac{eN}{s}\right) \tag{1.29}$$

holds for a constant $C > 0$, which only depends on $0 < \delta < 1$. Then, with probability at least $1 - \exp(-cm)$, where $c > 0$ denotes a constant, which only depends on δ as well, the matrix $\frac{1}{\sqrt{m}}\mathbf{A}$ satisfies the restricted isometry property of order s with isometry constant δ.

Theorem 1.11 was established for several other distributions, especially for sub-Gaussian distributions, such as the Bernoulli distribution. The decisive point is whether a concentration inequality of the type (1.30) can be derived for the particular underlying distribution, cf. [Ach03].

Lemma 1.12 (Concentration Inequality for Gaussian Matrices, [Ver12, Theorem 5.16])**.** *Let $\mathbf{A} \in \mathbb{R}^{m \times N}$ be a random matrix with entries sampled i.i.d. according to $\mathcal{N}(0,1)$. Then, for all $\mathbf{z} \in \mathbb{R}^N$ and $0 < \epsilon < 1$, it holds*

$$\mathsf{P}\left(\left| \left\| \frac{1}{\sqrt{m}} \mathbf{A}\mathbf{z} \right\|_2^2 - \|\mathbf{z}\|_2^2 \right| \geq \epsilon \|\mathbf{z}\|_2^2 \right) \leq 2 \exp\left(-cm\epsilon^2\right), \tag{1.30}$$

where $c > 0$ is an absolute constant.

Sketch of Proof. The claim follows from Bernstein inequality for sub-exponential random variables after observing that $Z_i = |\langle \mathbf{a}^i, \mathbf{z} \rangle|^2 - \|\mathbf{z}\|_2^2$ is a mean-zero sub-exponential random variable for all $i \in [m]$. □

Besides a suitable concentration inequality, such as Lemma 1.12 in the case of Gaussian random matrices, a covering argument plays an important role when verifying a restricted isometry property.

Proof of Theorem 1.11. Without loss of generality we can restrict the proof to $\|\mathbf{z}\|_2 = 1$, as $\mathbf{A}$ is linear. For the moment, let us fix a support set $S \subset [N]$ of size s and define the set of s-sparse vectors supported thereon by Σ_S. Note that there are in total $\binom{N}{s}$ such sets $\Sigma_S \subset \mathbb{S}^{N-1}$. Now, choose a minimal $\delta/4$-net $(\Sigma_S)^\# \subset \Sigma_S$. It is well-known (see, e.g., [Pis89, Lemma 4.16]) that such a net can be found of cardinality $\left|(\Sigma_S)^\#\right| \leq (12/\delta)^s$, as the support set S is fixed. A union bound to the set $(\Sigma_S)^\#$ yields by applying Lemma 1.12 with $\epsilon = \delta/8$ that

$$\left(1 - \frac{\delta}{8}\right) \|\mathbf{z}^\#\|_2^2 \leq \left\| \frac{1}{\sqrt{m}} \mathbf{A}\mathbf{z}^\# \right\|_2^2 \leq \left(1 + \frac{\delta}{8}\right) \|\mathbf{z}^\#\|_2^2 \tag{1.31}$$

holds for all $\mathbf{z}^\# \in (\Sigma_S)^\#$ with probability at least $1 - 2(12/\delta)^s \exp\left(-cm\delta^2/64\right)$. Let us now define the constant $A > 0$ as the smallest number such that

$$\left\| \frac{1}{\sqrt{m}} \mathbf{A}\mathbf{z} \right\|_2^2 \leq (1 + A) \|\mathbf{z}\|_2^2 \tag{1.32}$$

for all $\mathbf{z} \in \Sigma_S$. We want to show $A \leq \delta$. For any $\mathbf{z} \in \Sigma_S$ we can choose a $\mathbf{z}^\# \in (\Sigma_S)^\#$ with $\|\mathbf{z} - \mathbf{z}^\#\|_2 \leq \delta/4$. As $\mathbf{z} - \mathbf{z}^\# \in \Sigma_S$, we observe

$$\left\| \frac{1}{\sqrt{m}} \mathbf{A}\mathbf{z} \right\|_2^2 \leq \left(\left\| \frac{1}{\sqrt{m}} \mathbf{A}\mathbf{z}^\# \right\|_2 + \left\| \frac{1}{\sqrt{m}} \mathbf{A}\left(\mathbf{z} - \mathbf{z}^\#\right) \right\|_2 \right)^2 \leq \left(\sqrt{1 + \frac{\delta}{8}} + \sqrt{1 + A}\, \frac{\delta}{4} \right)^2. \tag{1.33}$$

As A is minimal by assumption, this implies

$$1 + A \leq \left(\sqrt{1 + \frac{\delta}{8}} + \sqrt{1 + A}\, \frac{\delta}{4} \right)^2, \tag{1.34}$$

which in turn shows $A \leq \frac{(1+\delta/8)}{(1-\delta/4)^2} - 1 \leq \delta$. The lower bound follows therefrom, as by the reverse triangle inequality it holds

$$\left\| \frac{1}{\sqrt{m}} \mathbf{A}\mathbf{z} \right\|_2^2 \geq \left(\left\| \frac{1}{\sqrt{m}} \mathbf{A}\mathbf{z}^{\#} \right\|_2 - \left\| \frac{1}{\sqrt{m}} \mathbf{A}\left(\mathbf{z} - \mathbf{z}^{\#}\right) \right\|_2 \right)^2 \geq \left(\sqrt{1 - \frac{\delta}{8}} - \sqrt{1+\delta}\,\frac{\delta}{4} \right)^2 \geq 1-\delta. \tag{1.35}$$

This establishes the restricted isometry property relative to the fixed support set S. It remains to extend this result to the whole set Σ_s^N by a union bound. As there are in total $\binom{N}{s} \leq (eN/s)^s$ such s-dimensional subspaces Σ_S, the restricted isometry property will fail with probability of at most $2(12eN/(\delta s))^s \exp\left(-cm\delta^2/64\right)$. The assertion follows by using assumption (1.29). This is made more verbose in [BDDW08, Theorem 5.2]. □

Despite being theoretically of high interest, Gaussian measurement matrices have limited practical application for two main reasons. First, it is difficult to design real-world sensing devices such that they can be represented by such random matrices. Second, they are almost surely dense matrices and therefore difficult to store and admit no fast matrix-vector multiplication. However, several research was done on structured random matrices, like, e.g., random partial Fourier matrices or partial random circulant and Toeplitz matrices [Rau10].

For the sake of completeness, let us state a stability result analogously to the one in Theorem 1.6 under the assumption of the restricted isometry property. In fact, it is a direct consequence thereof when applying Theorem 1.10.

Corollary 1.13. *Let us assume that* $\mathbf{A} \in \mathbb{R}^{m \times N}$ *has the restricted isometry property of order* $3s$ *with constant* $0 < \delta < 1/3$*. Then, for any* $\mathbf{x} \in \mathbb{R}^N$*, a solution* $\hat{\mathbf{x}}$ *to the* ℓ_1*-minimization problem* (1.10) *with* $\mathbf{y} = \mathbf{A}\mathbf{x}$ *fulfills*

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_1 \leq C\sigma_s(\mathbf{x})_1, \tag{1.36}$$

where $C = \frac{2(1+\gamma)}{1-\gamma}$ *with* $\gamma = \sqrt{\frac{(1+\delta)}{2(1-\delta)}}$.

Furthermore, under the same condition on the measurement matrix $\mathbf{A}$, a bound on the reconstruction error with respect to the ℓ_2-norm can be established.

Theorem 1.14 ([FR15, Theorem 2])**.** *Let us assume that* $\mathbf{A} \in \mathbb{R}^{m \times N}$ *has the restricted isometry property of order* $3s$ *with constant* $0 < \delta < 1/3$*. Then, for any* $\mathbf{x} \in \mathbb{R}^N$*, a solution* $\hat{\mathbf{x}}$ *to the* ℓ_1*-minimization problem* (1.10) *with* $\mathbf{y} = \mathbf{A}\mathbf{x}$ *fulfills*

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq C\frac{\sigma_s(\mathbf{x})_1}{\sqrt{s}}, \tag{1.37}$$

where $C = \frac{2}{1-\gamma}\left(\frac{\gamma+1}{\sqrt{2}} + \gamma\right)$ *with* $\gamma = \sqrt{\frac{(1+\delta)}{2(1-\delta)}}$.

In what follows we want to provide a proof of this theorem, as its concept will reappear in the matrix setting at a later point, cf. Theorem 2.5. However, we make a slightly stronger assumption, namely $\delta_{3s} + 3\delta_{4s} < 2$ and arrive at a modified constant. Therefore, we adapt the proof of Theorem 1.2 from [CRT06b].

Proof of Theorem 1.14. Since $\mathbf{x}$ is feasible and $\hat{\mathbf{x}}$ a solution to (1.10) with $\mathbf{y} = \mathbf{A}\mathbf{x}$, we have $\|\hat{\mathbf{x}}\|_1 \leq \|\mathbf{x}\|_1$. Let us define $\mathbf{h} = \mathbf{x} - \hat{\mathbf{x}} \in \ker \mathbf{A}$ and $T_0 = \operatorname{supp}(\mathbf{x}_{[s]})$. Then, by utilizing the reverse triangle inequality we observe

$$\begin{aligned} \|\mathbf{x}|_{T_0}\|_1 - \|\mathbf{h}|_{T_0}\|_1 - \left\|\mathbf{x}|_{T_0^c}\right\|_1 + \left\|\mathbf{h}|_{T_0^c}\right\|_1 &\leq \|(\mathbf{x}-\mathbf{h})|_{T_0}\|_1 + \left\|(\mathbf{x}-\mathbf{h})|_{T_0^c}\right\|_1 \\ &= \|\mathbf{x}-\mathbf{h}\|_1 = \|\hat{\mathbf{x}}\|_1 \leq \|\mathbf{x}\|_1, \end{aligned} \tag{1.38}$$

which simplifies to

$$\left\|\mathbf{h}|_{T_0^c}\right\|_1 \leq \|\mathbf{h}|_{T_0}\|_1 + 2\left\|\mathbf{x}|_{T_0^c}\right\|_1. \tag{1.39}$$

Let us now divide T_0^c into subsets of size $3s$ associated with a non-increasing rearrangement of $\mathbf{h}|_{T_0^c}$, i.e., we introduce the partition $\mathcal{T}$ from Theorem 1.10 with $t = 3s$ for $\mathbf{h}$, however, after having fixed T_0 in advance. That means, property (1.24) holds only for all $\ell \geq 2$. With this decomposition, we can show that the ℓ_2-norm of $\mathbf{h}$ is concentrated on $T_0 \cup T_1$. In fact, as the kth largest entry of $\mathbf{h}|_{T_0^c}$ can be upper bounded by $\left\|\mathbf{h}|_{T_0^c}\right\|_1/k$ we obtain

$$\left\|\mathbf{h}|_{(T_0\cup T_1)^c}\right\|_2^2 \leq \left\|\mathbf{h}|_{T_0^c}\right\|_1^2 \sum_{k=3s+1}^{N} \frac{1}{k^2} \leq \left\|\mathbf{h}|_{T_0^c}\right\|_1^2 \sum_{k=3s+1}^{N} \left(\frac{1}{k-1} - \frac{1}{k}\right) \leq \frac{\left\|\mathbf{h}|_{T_0^c}\right\|_1^2}{3s}, \tag{1.40}$$

having exploited that the last sum is a telescopic sum. Combining the last two inequalities and using Cauchy-Schwarz inequality thereafter yields

$$\left\|\mathbf{h}|_{(T_0\cup T_1)^c}\right\|_2 \leq \frac{\left\|\mathbf{h}|_{T_0^c}\right\|_1}{\sqrt{3s}} \leq \frac{\|\mathbf{h}|_{T_0}\|_1 + 2\left\|\mathbf{x}|_{T_0^c}\right\|_1}{\sqrt{3s}} \leq \frac{1}{\sqrt{3}}\|\mathbf{h}|_{T_0}\|_2 + \frac{2}{\sqrt{3s}}\left\|\mathbf{x}|_{T_0^c}\right\|_1. \tag{1.41}$$

We can use this bound to establish a bound on $\|\mathbf{h}\|_2$, namely

$$\begin{aligned} \|\mathbf{h}\|_2 &\leq \|\mathbf{h}|_{T_0\cup T_1}\|_2 + \left\|\mathbf{h}|_{(T_0\cup T_1)^c}\right\|_2 \leq \|\mathbf{h}|_{T_0\cup T_1}\|_2 + \frac{1}{\sqrt{3}}\|\mathbf{h}|_{T_0}\|_2 + \frac{2}{\sqrt{3s}}\left\|\mathbf{x}|_{T_0^c}\right\|_1 \\ &\leq \left(1 + \frac{1}{\sqrt{3}}\right)\|\mathbf{h}|_{T_0\cup T_1}\|_2 + \frac{2}{\sqrt{3s}}\left\|\mathbf{x}|_{T_0^c}\right\|_1. \end{aligned} \tag{1.42}$$

After using both triangle inequalities, an application of the restricted isometry property of order $4s$ and $3s$ with the respective isometry constants δ_{4s} and δ_{3s} yields

$$\begin{aligned} \|\mathbf{A}\mathbf{h}\|_2 &= \left\|\mathbf{A}\mathbf{h}|_{T_0\cup T_1} + \sum_{\ell\geq 2}\mathbf{A}\mathbf{h}|_{T_\ell}\right\|_2 \geq \|\mathbf{A}\mathbf{h}|_{T_0\cup T_1}\|_2 - \sum_{\ell\geq 2}\|\mathbf{A}\mathbf{h}|_{T_\ell}\|_2 \\ &\geq \sqrt{1-\delta_{4s}}\|\mathbf{h}|_{T_0\cup T_1}\|_2 - \sqrt{1+\delta_{3s}}\sum_{\ell\geq 2}\|\mathbf{h}|_{T_\ell}\|_2 \\ &\geq \sqrt{1-\delta_{4s}}\|\mathbf{h}|_{T_0\cup T_1}\|_2 - \sqrt{1+\delta_{3s}}\left(\frac{1}{\sqrt{3}}\|\mathbf{h}|_{T_0}\|_2 + \frac{2}{\sqrt{3s}}\left\|\mathbf{x}|_{T_0^c}\right\|_1\right) \\ &\geq \left(\sqrt{1-\delta_{4s}} - \frac{1}{\sqrt{3}}\sqrt{1+\delta_{3s}}\right)\|\mathbf{h}|_{T_0\cup T_1}\|_2 - \frac{2}{\sqrt{3s}}\sqrt{1+\delta_{3s}}\left\|\mathbf{x}|_{T_0^c}\right\|_1. \end{aligned} \tag{1.43}$$

The next-to-last inequality avails itself of the common technique in compressed sensing used in the proof of Theorem 1.10. More precisely, repeating these computations we obtain $\sum_{\ell\geq 2}\|\mathbf{h}|_{T_\ell}\|_2 \leq \frac{1}{\sqrt{3s}}\sum_{\ell\geq 2}\left\|\mathbf{h}|_{T_{\ell-1}}\right\|_1 \leq \frac{1}{\sqrt{3s}}\left\|\mathbf{h}|_{T_0^c}\right\|_1$. Now, using (1.39) and Cauchy-Schwarz inequality we arrive at $\sum_{\ell\geq 2}\|\mathbf{h}|_{T_\ell}\|_2 \leq \frac{1}{\sqrt{3s}}\|\mathbf{h}|_{T_0}\|_1 + \frac{2}{\sqrt{3s}}\left\|\mathbf{x}|_{T_0^c}\right\|_1 \leq \frac{1}{\sqrt{3}}\|\mathbf{h}|_{T_0}\|_2 + \frac{2}{\sqrt{3s}}\left\|\mathbf{x}|_{T_0^c}\right\|_1$.

Finally, since $\mathbf{h} \in \ker \mathbf{A}$, $\|\mathbf{A}\mathbf{h}\|_2 = 0$ and thus (1.43) provides a bound on $\|\mathbf{h}|_{T_0 \cup T_1}\|_2$. This in turn can be used to complete the bound (1.42) on $\|\mathbf{h}\|_2$ as follows,

$$\|\mathbf{h}\|_2 \leq \left(\left(1 + \frac{1}{\sqrt{3}}\right) \frac{1}{C_\delta} \sqrt{1 + \delta_{3s}} + 1\right) \frac{2}{\sqrt{3s}} \left\|\mathbf{x}|_{T_0^c}\right\|_1 \leq C \frac{\sigma_s(\mathbf{x})_1}{\sqrt{s}}, \tag{1.44}$$

abbreviating $C_\delta = \sqrt{1 - \delta_{4s}} - \frac{1}{\sqrt{3}} \sqrt{1 + \delta_{3s}}$. Simple algebra reveals that the denominator C_δ is greater than 0 if $\delta_{3s} + 3\delta_{4s} < 2$, which completes the proof. □

Remark 1.15. There is nothing particularly special about having chosen the subsets, T_0^c is divided into, to be of size $3s$. The proof shows immediately that any $b > 1$ would have been possible yielding the condition $\delta_{bs} + b\delta_{(b+1)s} < b - 1$. This reveals a trade-off between the order of the restricted isometry property and the condition on the RIP constants. Improving both the required order and a sufficient bound on δ in statements like Corollary 1.13 and Theorem 1.14 was the focus of a vast amount of research. For instance, in [Can08], the results of the previous two statements were shown, for a different but well-behaved constant C, under the assumption of a restricted isometry property of order $2s$ with $0 < \delta < \sqrt{2} - 1$.

Having these reconstruction and stability results at hand, an intriguing question arises. Namely, how well the combination of a measurement matrix satisfying the restricted isometry property and ℓ_1-minimization performs compared to theoretically optimal encoder-decoder pairs $(\mathbf{A}, \Delta)$. This is the content of the upcoming subsection. Before that, however, we want to comment on ℓ_q-minimization.

Remark 1.16 (Restricted Isometry Property for ℓ_q-Minimization)**.** Building upon Theorem 1.10 one may ask whether the restricted isometry property also implies the modified null space property for ℓ_q-minimization introduced in Remark 1.5. And indeed, we formulate Theorem 1.17 below to show that this is the case. Ideas of its proof, which resembles the one of Theorem 1.10, will reappear later on in Lemma 4.9.

Theorem 1.17 (The Restricted Isometry Property implies the Null Space Property for ℓ_q-minimization)**.** *Let $0 < q \leq 1$ and let us assume that $\mathbf{A} \in \mathbb{R}^{m \times N}$ has the restricted isometry property of order $s + t$ with constant $0 < \delta < 1$. Then, $\mathbf{A}$ has the null space property for ℓ_q-minimization of order s with constant $\gamma = \left(\frac{s}{t}\right)^{1-q/2} \left(\frac{1+\delta}{1-\delta}\right)^{q/2}$.*

Proof. Analogously to the proof of Theorem 1.10, let $\mathbf{v} \in \ker \mathbf{A}$ and define a partition $\mathcal{T}$ as in (1.23). Then, we utilize Lemma A.2(ii), which is a consequence of Hölder's inequality, the restricted isometry property and the fact that $\mathbf{v} \in \ker \mathbf{A}$ to obtain

$$\begin{aligned}
\|\mathbf{v}|_{T_0}\|_q^q &\leq s^{1-q/2} \|\mathbf{v}|_{T_0}\|_2^q \leq s^{1-q/2} \|\mathbf{v}|_{T_0 \cup T_1}\|_2^q \leq s^{1-q/2} \left(\frac{1}{1-\delta}\right)^{q/2} \|\mathbf{A}\mathbf{v}|_{T_0 \cup T_1}\|_2^q \\
&= s^{1-q/2} \left(\frac{1}{1-\delta}\right)^{q/2} \left\|\mathbf{A}\mathbf{v}|_{T_2 \cup T_3 \cup \cdots \cup T_{\lceil (n-s)/t \rceil}}\right\|_2^q \\
&= s^{1-q/2} \left(\frac{1}{1-\delta}\right)^{q/2} \left\|\mathbf{A}\mathbf{v}|_{T_2} + \mathbf{A}\mathbf{v}|_{T_3} + \cdots + \mathbf{A}\mathbf{v}|_{T_{\lceil (n-s)/t \rceil}}\right\|_2^q.
\end{aligned} \tag{1.45}$$

In order to bound the norm in the last expression, we observe that the vector-valued function $f(\,\cdot\,) := \|\cdot\|_2^q$ is subadditive for $0 < q \leq 1$. To see this, note that $f(\mathbf{w}) = g(\|\mathbf{w}\|_2)$

with the concave and increasing function $g : [0, \infty) \to [0, \infty),\ w \mapsto w^q$. As $g(0) = 0$, by concavity, the real-valued function g itself is subadditive. Thus, using that g is an increasing function in the first and exploiting that g is subadditive in the second inequality, we get $f(\mathbf{w}_1+\mathbf{w}_2) = g(\|\mathbf{w}_1+\mathbf{w}_2\|_2) \leq g(\|\mathbf{w}_1\|_2+\|\mathbf{w}_2\|_2) \leq g(\|\mathbf{w}_1\|_2)+g(\|\mathbf{w}_2\|_2) = f(\mathbf{w}_1)+f(\mathbf{w}_2)$. With this we can further bound $\|\mathbf{v}|_{T_0}\|_q^q$ by employing the restricted isometry property once more, which yields

$$\|\mathbf{v}|_{T_0}\|_q^q \leq s^{1-q/2} \left(\frac{1}{1-\delta}\right)^{q/2} \sum_{\ell=2}^{\lceil (n-s)/t \rceil} \|\mathbf{A}\mathbf{v}|_{T_\ell}\|_2^q \leq s^{1-q/2} \left(\frac{1+\delta}{1-\delta}\right)^{q/2} \sum_{\ell=2}^{\lceil (n-s)/t \rceil} \|\mathbf{v}|_{T_\ell}\|_2^q. \tag{1.46}$$

It remains to upper bound the last term. We note that by definition of the partition $\mathcal{T}$ for all $\ell \geq 1$ it also holds

$$|v_i|^q \leq |v_j|^q \quad \text{for all } i \in T_\ell \text{ and } j \in T_{\ell-1}. \tag{1.47}$$

For $\ell \geq 2$, summation over j yields $|v_i|^q \leq t^{-1}\big\|\mathbf{v}|_{T_{\ell-1}}\big\|_q^q$ for all $i \in T_\ell$, or equivalently

$$|v_i| \leq t^{-1/q}\big\|\mathbf{v}|_{T_{\ell-1}}\big\|_q. \tag{1.48}$$

Taking the ℓ_2-norm over $i \in T_\ell$ subsequently shows

$$\|\mathbf{v}|_{T_\ell}\|_2 \leq t^{1/2-1/q}\big\|\mathbf{v}|_{T_{\ell-1}}\big\|_q. \tag{1.49}$$

Finally, using this estimate in (1.46) results in

$$\|\mathbf{v}|_{T_0}\|_q^q \leq \left(\frac{s}{t}\right)^{1-q/2} \left(\frac{1+\delta}{1-\delta}\right)^{q/2} \sum_{\ell=2}^{\lceil (n-s)/t \rceil} \|\mathbf{v}|_{T_{\ell-1}}\|_q^q \leq \left(\frac{s}{t}\right)^{1-q/2} \left(\frac{1+\delta}{1-\delta}\right)^{q/2} \|\mathbf{v}|_{T_0^c}\|_q^q, \tag{1.50}$$

which establishes the null space property for ℓ_q-minimization of order s with the claimed constant γ. □

Remark 1.18 (A modified Restricted Isometry Property for ℓ_q-Minimization)**.** In [CS08], the authors propose a different notion of the restricted isometry property. In order to adapt Definition 1.8 to the ℓ_q-case, condition (1.22) is modified and replaced by

$$(1-\delta)\|\mathbf{z}\|_2^q \leq \|\mathbf{A}\mathbf{z}\|_q^q \leq (1+\delta)\|\mathbf{z}\|_2^q. \tag{1.51}$$

Instead of quantifying how well Σ_s^N can be isometrically embedded into $\mathbb{R}^m$ by the matrix $\mathbf{A}$ with respect to the ℓ_2-norm, the ℓ_q-(quasi)-norm is taken as measure. Eventually, in the case of a random Gaussian measurement matrix, they are able to derive that the number of measurement necessary to ensure, with high probability, unique recoverability of s-sparse signals $\mathbf{x}$ from measurements $\mathbf{y}$ via ℓ_q-minimization (1.9) behaves like

$$m \geq C_1(q)s + qC_2(q)s\log\left(\frac{eN}{s}\right). \tag{1.52}$$

Here, C_1 and C_2 are constants depending on q such that $qC_2(q)$ vanishes for $q \to 0$. Thus, remarkably, also the dependency on the ambient dimension N vanishes. This can be seen as an interpolation between the ℓ_1- and the ℓ_0-case in the following sense. Theorem 1.2

provides an ultimate lower bound on the necessary number of measurements, which is $2s$ and thus linear in the sparsity. Theoretically, any s-sparse vector $\mathbf{x}$ can be reconstructed uniquely from its corresponding measurements $\mathbf{y}$ via ℓ_0-minimization (1.7), though, as discussed, not stably. For the other case, namely $q = 1$, from, e.g., Corollary 1.13 and Theorem 1.11 we deduce that, with high probability, s-sparse vectors $\mathbf{x}$ can be recovered from the measurements $\mathbf{y}$ via ℓ_1-minimization (1.10) if the measurement matrix is a Gaussian matrix and m behaves like $m \geq Cs\log(eN/s)$.

1.3.4 Performance of ℓ_1-Minimization

In the following we consider the performance of encoder-decoder pairs $(\mathbf{A}, \Delta)$ from a more general and theoretical perspective. Instead of asking the typical compressed sensing question related to good pairs $(\mathbf{A}, \Delta)$, we focus on the performance and properties of optimal encoder-decoder pairs, in particular with regard to the number of necessary measurements [CDD09]. To this end we compare the worst reconstruction error relative to a subset $K \subset \mathbb{R}^N$ for the best encoder-decoder pair $(\mathbf{A}, \Delta)$,

$$E_m(K)_X = \inf_{(\mathbf{A},\Delta)\in\mathcal{A}_{m,N}} \sup_{\mathbf{z}\in K} \|\mathbf{z} - \Delta(\mathbf{A}\mathbf{z})\|_X, \tag{1.53}$$

with the best s-term approximation error of this set. Here, $\mathcal{A}_{m,N}$ denotes the set of all possible pairs $(\mathbf{A}, \Delta)$, where $\mathbf{A} \in \mathbb{R}^{m\times N}$ describes the encoder and $\Delta : \mathbb{R}^m \to \mathbb{R}^N$ is any function modeling the decoder. Issues of this type are common in the field of information based complexity.

We are now interested in finding the smallest m such that

$$E_m(K)_X \lesssim \sigma_s(K)_X, \tag{1.54}$$

where $\sigma_s(K)_X = \sup_{\mathbf{z}\in K} \sigma_s(\mathbf{z})_X$ denotes the best s-term approximation error of K with respect to the norm $\|\cdot\|_X$. It turns out that the quantity $E_m(K)_X$ is fundamentally linked to the concept of Gelfand widths, as we will see in Theorem 1.21.

Definition 1.19 (Gelfand Width)**.** *For a compact set K in a Banach space X the Gelfand width of order m is given by*

$$d^m(K)_X = \inf_{\operatorname{codim}(Y)\leq m} \sup_{\mathbf{z}\in K\cap Y} \|\mathbf{z}\|_X. \tag{1.55}$$

The following lemma and many variants involving different sets K, such as the ℓ_q-balls $\mathcal{B}_q^N$, and various normed spaces X were established for Kolmogorov widths [GG84, Glu84], which are dual to the Gelfand widths. However, classical theory merely addressed the case $q \geq 1$. In [FPRU10] an entirely new approach relying on compressed sensing techniques was taken to investigate the Gelfand widths of the non-convex ℓ_q-(quasi)-norm-balls $\mathcal{B}_q^N$.

Lemma 1.20 (Bounds on the Gelfand Width for the ℓ_q-(Quasi)-Norm-Ball, [FPRU10, Theorem 1.1])**.** *Let $0 < q \leq 1$ and $q < p \leq 2$. Then, for $K = \mathcal{B}_q^N$ and $X = \ell_p^N$ we have*

$$c_{q,p} \min\left\{1, \frac{\log(eN/m)}{m}\right\}^{1/q-1/p} \leq d^m(K)_X \leq C_{q,p} \min\left\{1, \frac{\log(eN/m)}{m}\right\}^{1/q-1/p}, \tag{1.56}$$

where the constants $c_{q,p}$ and $C_{q,p}$ depend solely on p and q. For the bound $c_{q,p}$ involved in the lower bound we explicitly have $c_{q,p} = (1/2)^{1/q}(cq)^{1/q-1/p}$, where $c \approx 0.073$ is an absolute constant.

In the next theorem we now outline the relation between the optimal worst reconstruction error $E_m(K)_X$ and the Gelfand width $d^m(K)_X$.

Theorem 1.21 ([CDD09, Lemma 2.1])**.** *Let K be a closed compact set such that $K = -K$ and $K + K \subset C_0 K$ for a constant $C_0 > 0$. Then, for any norm $\|\cdot\|_X$ on $\mathbb{R}^N$, it holds*

$$d^m(K)_X \leq E_m(K)_X \leq C_0 d^m(K)_X. \tag{1.57}$$

Sketch of Proof. By identifying a subspace Y of codimension less or equal than m with the kernel of a suitable matrix $\mathbf{A}$, we observe

$$d^m(K)_X = \inf_{\mathbf{A}\in\mathbb{R}^{m\times N}} \sup_{\mathbf{v}\in K\cap\ker\mathbf{A}} \|\mathbf{v}\|_X. \tag{1.58}$$

To show the lower bound, let $(\mathbf{A}, \Delta)$ denote an encoder-decoder pair and set $\mathbf{z} = \Delta(\mathbf{0})$. Now, for $\mathbf{v} \in \ker\mathbf{A}$, we obtain $\|\mathbf{v} - \mathbf{z}\|_X \geq \|\mathbf{v}\|_X$ or $\|-\mathbf{v} - \mathbf{z}\|_X \geq \|\mathbf{v}\|_X$. Exploiting the assumed symmetry of K we derive

$$d^m(K)_X \leq \sup_{\mathbf{v}\in K\cap\ker\mathbf{A}} \|\mathbf{v} - \mathbf{z}\|_X = \sup_{\mathbf{v}\in K\cap\ker\mathbf{A}} \|\mathbf{v} - \Delta(\mathbf{A}\mathbf{v})\|_X \leq \sup_{\mathbf{v}\in K} \|\mathbf{v} - \Delta(\mathbf{A}\mathbf{v})\|_X. \tag{1.59}$$

This yields the lower bound after taking the infimum over $\mathcal{A}_{m,N}$.
For proving the upper bound let us first choose a subspace Y as specified at the beginning and associate the encoder $\mathbf{A}$ with $Y^\perp$. In turn, define the decoder Δ as follows. Given $\mathbf{y}$ in the image of K under the linear map $\mathbf{A}$, we choose $\Delta(\mathbf{y}) \in K \cap \{\mathbf{z} : \mathbf{A}\mathbf{z} = \mathbf{y}\}$. Utilizing the assumed symmetry and inclusion of the Minkowski sum of K, we observe

$$E_m(K)_X \leq \sup_{\mathbf{v}\in K} \left(\sup_{\mathbf{v}'\in K\cap\{\mathbf{z}:\mathbf{A}\mathbf{z}=\mathbf{A}\mathbf{v}\}} \|\mathbf{v} - \mathbf{x}'\|_X \right) \leq \sup_{\mathbf{v}\in C_0 K\cap\ker\mathbf{A}} \|\mathbf{v}\|_X \leq C_0 \sup_{\mathbf{v}\in K\cap\ker\mathbf{A}} \|\mathbf{v}\|_X. \tag{1.60}$$

Taking the infimum over the subspaces Y provides the upper bound. □

Combining Theorem 1.21 and Lemma 1.20 and applying them in the setting $K = \mathcal{B}_1^N$ and $X = \ell_2^N$ establishes, for m and N large enough, the bounds

$$\sqrt{\frac{\log(eN/m)}{m}} \lesssim E_m(\mathcal{B}_1^N)_{\ell_2^N} \lesssim \sqrt{\frac{\log(eN/m)}{m}} \tag{1.61}$$

on the optimal performance for the recovery of vectors with ℓ_1-norm bounded by one, when measuring the approximation error in the ℓ_2-norm.
Particularly the lower bound is of interest for compressed sensing, as we can deduce by requiring equation (1.54) and utilizing Stechkin's Inequality, Lemma 1.1, that for the number of necessary measurements it has to hold $m \gtrsim s\log(eN/m)$. Therefrom, by modifying the constants, one can deduce

$$m \gtrsim s \log\left(\frac{eN}{s}\right), \tag{1.62}$$

see, e.g., [FR13, Lemma C.6(c)]. This shows that the minimal number of measurements necessary for stable recovery is given by (1.62), which moreover matches the bound from Theorem 1.11, i.e., Gaussian measurements achieve the optimal measurement size.
We want to end this subsection with a few notes on ℓ_q-minimization.

Remark 1.22 (Performance of ℓ_q-Minimization)**.** First, by combining Theorem 1.21 in the exactly same manner with Lemma 1.20, bounds on the quantity $E_m(\mathcal{B}_q^N)_{\ell_2^N}$ can be derived. They can be interpreted as bounds on the optimal performance for the recovery of compressible vectors.
Second, based thereon, bounds on the measurements similar to (1.62) can be obtained for stable recovery with ℓ_q-minimization. More precisely, in order to assure (1.54) in the setting $K = \mathcal{B}_q^N$ and $X = \ell_2^N$, for a lower bound on the measurements we derive

$$m \gtrsim q(1/2)^{(2/(2-q))} s \log\left(eN/s\right). \tag{1.63}$$

This improves the required number of measurements as q tends to 0. For a similar result we refer to [FPRU10, Theorem 2.7], where it is shown that $m \gtrsim qs \log\left(eN/s\right)$ is a necessary requirement for stability in the sense of instance optimality of order s with respect to the quasi-norm $\|\cdot\|_q$. Even though the right-hand side in (1.63) converges to 0 for $q \to 0$, we want to emphasize that a term of order $\mathcal{O}(s)$ needs to be added according to the note after Theorem 1.2. Due to the volumetric argument used in the proof of Lemma 1.20, terms of such low-dimensional order vanish. However, the important observation here is that the dependency on the dimension N disappears, which is consistent with the findings in [CS08], cf. Remark 1.18.
And third, it is furthermore possible to extended Theorem 1.21 in a straightforward manner to quasi-norms $\|\cdot\|_X$ on $\mathbb{R}^N$ by replacing the lower bound with $c^{-1}d^m(K)_X$, where $c > 1$ denotes the respective quasi-norm constant, i.e., the smallest number c such that $\|\mathbf{w}_1+\mathbf{w}_2\|_X \le c(\|\mathbf{w}_1\|_X+\|\mathbf{w}_2\|_X)$ holds for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^N$, cf. [FPRU10, Proposition 1.2].

1.3.5 Robustness with respect to Measurement Noise

In most practical applications the measurement process is affected by noise, which corrupts the measured data. This can be due to a disturbing environment or caused by the sensor itself. In any case, the noiseless acquisition model (1.6) needs to be adapted as already described in (1.12) in order to incorporate such perturbations. For an unknown noise vector $\boldsymbol{\eta} \in \mathbb{R}^m$ we model the corrupted measurement process by

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\eta} \tag{1.64}$$

and are interested in recovering $\mathbf{x}$ from $\mathbf{y}$ up to an error of order $\mathcal{O}(\|\boldsymbol{\eta}\|_2)$. If a reconstruction scheme achieves this, it is called robust with respect to measurement noise. In literature, many different noise models were considered. The two most prominent ones include a random dense noise vector $\boldsymbol{\eta}$, which can be interpreted as a permanent present corruption due to the environment [CRT06b], and a random sparse vector $\boldsymbol{\eta}$, which can model an unreliable measurement device or transmission channel [CT05, CRTV05]. In the following we restrict ourselves to the first case and refer to the literature for the second.
As in the noiseless case, we are interested in finding the sparsest solution $\mathbf{x}$ to (1.64). This means we consider the noise-aware ℓ_0-minimization problem

$$\min_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{z}\|_0 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \le \eta, \tag{1.65}$$

where $\eta \geq 0$ is chosen suitably such that $\eta \geq \|\mathbf{Ax} - \mathbf{y}\|_2 = \|\boldsymbol{\eta}\|_2$. Following the idea of ℓ_1-minimization in the setting without noise, it is natural to analogously consider the convex relaxation

$$\min_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{Az} - \mathbf{y}\|_2 \leq \eta. \tag{1.66}$$

This is a convex optimization problem and commonly known as quadratically constrained or noise-aware ℓ_1-minimization, or basis pursuit denoising.
A different approach is to consider the so-called least absolute shrinkage and selection operator (LASSO)

$$\min_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{Az} - \mathbf{y}\|_2^2 + \beta\|\mathbf{z}\|_1 \tag{1.67}$$

for a regularization parameter $\beta \geq 0$. This was introduced by Tibshirani in [Tib96]. Both optimization problems are indeed equivalent for a specific, but on the minimizer dependent, choice of the parameters η and β. More rigorously the following holds, cf. [FR13, Proposition 3.2 and Theorem B.28].

Lemma 1.23. *Let $\mathbf{A} \in \mathbb{R}^{m\times N}$ and $\mathbf{y} \in \mathbb{R}^m$.*

(i) Let $\eta > 0$. If $\hat{\mathbf{x}}$ is a minimizer of (1.66), there exists a parameter $\beta \geq 0$ such that $\hat{\mathbf{x}}$ is a minimizer of (1.67).

(ii) Conversely, let $\beta \geq 0$. If $\hat{\mathbf{x}}$ is a minimizer of (1.67), then there exists a parameter $\eta \geq 0$ such that $\hat{\mathbf{x}}$ is a minimizer of (1.66).

Proof. Firstly, for (i), we observe that (1.66) is equivalent to

$$\min_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{Az} - \mathbf{y}\|_2^2 \leq \eta^2, \tag{1.68}$$

which in turn has the Lagrange function

$$\mathcal{L}(\mathbf{z}, \lambda) = \|\mathbf{z}\|_1 + \lambda\left(\|\mathbf{Az} - \mathbf{y}\|_2^2 - \eta^2\right). \tag{1.69}$$

As $\eta > 0$ and $\mathbf{A}$ has full rank there exists a strictly feasible $\mathbf{z}$ of (1.66). Thus Slater's condition is fulfilled and strong duality holds. Consequently, there exists the primal-dual optimal pair $(\hat{\lambda}, \hat{\mathbf{x}})$ for a $\hat{\lambda} \geq 0$. From the saddle-point property it then follows that $\mathcal{L}(\hat{\mathbf{x}}, \hat{\lambda}) \leq \mathcal{L}(\mathbf{z}, \hat{\lambda})$ for all $\mathbf{z} \in \mathbb{R}^N$, i.e., $\hat{\mathbf{x}}$ minimizes the function $\mathcal{L}(\,\cdot\,, \hat{\lambda})$. In the case $\hat{\lambda} > 0$ the claim follows by setting $\beta = 1/\hat{\lambda} \geq 0$. However, for $\hat{\lambda} = 0$ we note that the saddle-point property implies $\hat{\mathbf{x}} = \mathbf{0}$, since $\|\hat{\mathbf{x}}\|_1 = \mathcal{L}(\hat{\mathbf{x}}, 0) \leq \mathcal{L}(\mathbf{z}, 0)$ for all $\mathbf{z} \in \mathbb{R}^N$, i.e., in particular for $\mathbf{z} = \mathbf{0}$. Now, let us choose $\beta = 2\|\mathbf{A}\|\|\mathbf{y}\|_2$. Then,

$$\begin{aligned}
&\|\mathbf{Az} - \mathbf{y}\|_2^2 + \beta\|\mathbf{z}\|_1 = \|\mathbf{Az}\|_2^2 - 2\langle \mathbf{Az}, \mathbf{y}\rangle + \|\mathbf{y}\|_2^2 + \beta\|\mathbf{z}\|_1 \\
&\quad \geq \|\mathbf{Az}\|_2^2 - 2\|\mathbf{Az}\|_2\|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2 + \beta\|\mathbf{z}\|_1 = \|\mathbf{Az}\|_2\left(\|\mathbf{Az}\|_2 - 2\|\mathbf{y}\|_2\right) + \|\mathbf{y}\|_2^2 + \beta\|\mathbf{z}\|_1 \\
&\quad \geq -2\|\mathbf{Az}\|_2\|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2 + \beta\|\mathbf{z}\|_1 = -2\|\mathbf{Az}\|_2\|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2 + 2\|\mathbf{A}\|\|\mathbf{y}\|_2\|\mathbf{z}\|_1 \\
&\quad \geq -2\|\mathbf{A}\|\|\mathbf{z}\|_2\|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2 + 2\|\mathbf{A}\|\|\mathbf{y}\|_2\|\mathbf{z}\|_2 = \|\mathbf{y}\|_2^2 = \|\mathbf{A0} - \mathbf{y}\|_2^2 + \beta\|\mathbf{0}\|_1,
\end{aligned} \tag{1.70}$$

where in the last inequality, besides the definition of the operator norm, it is used that $\|\mathbf{z}\|_1 \geq \|\mathbf{z}\|_2$. Thus, $\mathbf{0}$ is a minimizer of (1.67) as claimed.

Secondly, for (ii), we set $\eta = \|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_2$ and note that for $\mathbf{z}$ with $\|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta$ by optimality of $\hat{\mathbf{x}}$ we observe

$$\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_2^2 + \beta\|\hat{\mathbf{x}}\|_1 \leq \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2^2 + \beta\|\mathbf{z}\|_1 \leq \|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_2^2 + \beta\|\mathbf{z}\|_1, \tag{1.71}$$

which implies the claim. □

In order to ensure robustness of the noise-aware ℓ_1-minimization (1.66) the null space property from Definition 1.3 is not sufficient and needs an adjustment.

Definition 1.24 (Robust Null Space Property (NSP))**.** *A matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$ satisfies the robust null space property of order s with constants $0 < \gamma < 1$ and $\tau > 0$, if*

$$\|\mathbf{z}|_S\|_1 \leq \gamma\|\mathbf{z}|_{S^c}\|_1 + \tau\|\mathbf{A}\mathbf{z}\|_2, \tag{1.72}$$

for all sets $S \subset [N]$ such that $|S| \leq s$ and for all $\mathbf{z} \in \mathbb{R}^N$.

We want to point out that the robust null space property requires that condition (1.72) holds for all $\mathbf{z} \in \mathbb{R}^N$ and not just for all $\mathbf{z} \in \ker \mathbf{A}\backslash\{\mathbf{0}\}$ as in the null space property. Thus, it is a strengthened version, i.e., it implies the null space property. Moreover, analogously to the noiseless case, the parameter γ controls stability and novelly, robustness is controlled by both parameters γ and τ. This is made precise in the next theorem, which is the noisy version of Theorem 1.6.

Theorem 1.25 ([FR13, Theorem 4.19])**.** *Let us assume that $\mathbf{A} \in \mathbb{R}^{m\times N}$ has the robust null space property of order s with constants $0 < \gamma < 1$ and $\tau > 0$. Then, for any $\mathbf{x} \in \mathbb{R}^N$, a solution $\hat{\mathbf{x}}$ to (1.66) with $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\eta}$ and $\eta \geq \|\boldsymbol{\eta}\|_2$ fulfills*

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_1 \leq \frac{2(1+\gamma)}{1-\gamma}\sigma_s(\mathbf{x})_1 + \frac{4\tau}{1-\gamma}\eta. \tag{1.73}$$

As the null space property also the robust null space property is implied by a restricted isometry property in the following way.

Theorem 1.26 ([FR13, Theorem 6.13])**.** *Let us assume that $\mathbf{A} \in \mathbb{R}^{m\times N}$ has the restricted isometry property of order $2s$ with constant $0 < \delta < 4/\sqrt{41}$. Then, $\mathbf{A}$ has the robust null space property of order s with constants $0 < \gamma < 1$ and $\tau > 0$ depending only on δ.*

For the sake of completeness we want to conclude with a recovery result for the ℓ_2-norm. It is the noisy version of the result referred to in Remark 1.15 and portrays the robustness of (1.66) to noise nicely.

Theorem 1.27 ([Can08, Theorem 1.2])**.** *Let us assume that $\mathbf{A} \in \mathbb{R}^{m\times N}$ has the restricted isometry property of order $2s$ with constant $0 < \delta < \sqrt{2} - 1$. Then, for any $\mathbf{x} \in \mathbb{R}^N$, a solution $\hat{\mathbf{x}}$ to (1.66) with $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\eta}$ and $\eta \geq \|\boldsymbol{\eta}\|_2$ fulfills*

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq C_0\frac{\sigma_s(\mathbf{x})_1}{\sqrt{s}} + C_1\eta, \tag{1.74}$$

with well-behaved constants C_0 and C_1.

As in the previous parts we also want to address ℓ_q-minimization briefly.

Remark 1.28 (ℓ_q-Minimization in the Case of Noise)**.** In [GPYZ15] two different modifications of the robust null space property were introduced. The first and more natural one replaces (1.72) by $\|\mathbf{z}|_S\|_q^q \leq \gamma \|\mathbf{z}|_{S^c}\|_q^q + \tau \|\mathbf{A}\mathbf{z}\|_2$ and assures a corresponding recovery result, where the error is measured in the metric $d(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_q^q$ and $\sigma_s(\mathbf{x})_q^q$ replaces the corresponding stability error term on the right-hand side. The second one, though, generalizes this and reads for $0 < p \leq q \leq 1$ as $\|\mathbf{z}|_S\|_q^q \leq \frac{\gamma}{s^{q/p-1}} \|\mathbf{z}|_{S^c}\|_p^p + \frac{\tau}{s^{q/p-1}} \|\mathbf{A}\mathbf{z}\|_2$. For theoretical results and proofs thereon we refer to the literature.

Regarding stability and robustness in case of noise, in [SY10] a generalization of Theorem 1.2 from [CRT06b] was formulated for ℓ_q-minimization. The latter essentially states the same as Theorem 1.27, however, under a different assumption on the RIP constants, namely $\delta_{3s} + 3\delta_{4s} < 2$. We presented the noiseless version of this proof to show Theorem 1.14. The only adaption, which needs to be made, is to replace $\|\mathbf{A}\mathbf{h}\|_2 = 0$ by $\|\mathbf{A}\mathbf{h}\|_2 \leq 2\eta$ in the chain of inequalities (1.43). The former paper, in turn, showed that the condition on the RIP constants for ℓ_q-minimization is weaker the smaller q gets, while the recovery result remains comparable.

Moreover, in [CS08] it is claimed that robustness with respect to measurement noise enhances for smaller q's, whereas stability with respect to defects in the sparsity initially improves before it worsens as q decreases. An analysis thereof, including conditions for a restricted isometry property, is provided in [SCOY08]. However, they also observed that the numerical results only partially confirm their theoretically funded expectations.

1.4 Numerical Algorithms for Compressed Sensing

After having provided an insight into the theory of compressed sensing in the previous sections, for the remainder of this chapter we want to turn our attention to efficient algorithms tackling the posed optimization problems.

To begin with, we want to very briefly outline orthogonal matching pursuit, which is a greedy method and successively builds a sparse solution to (1.64). After that we turn towards ℓ_1-minimization and start with a very simple idea, namely reformulating the maybe noise-aware optimization problem into a linear or second-order cone program, respectively. Eventually, we investigate a more problem-specific class of algorithms, namely iterative thresholding based techniques, such as the iterative soft thresholding algorithm.

However, many more methods were proposed and analyzed in recent years, see, e.g., [For10, FR15] for an overview. Without going into further detail we want to mention a few of them. The iteratively reweighted least squares (IRLS) method is also an iterative approach, which transforms the ℓ_1-minimization problem into a weighted ℓ_2-minimization problem [ODBP15]. The homotopy method is a direct approach attempting to trace the sparse solution to an ℓ_1-regularized least squares functional with respect to the regularization parameter. Least angle regression (LARS) is a slightly modified version thereof [EHJT04].

1.4.1 Orthogonal Matching Pursuit

As already pointed out, orthogonal matching pursuit (OMP), see, e.g., [Tro04, TG07], is a greedy algorithm. In general, a greedy method follows the strategy of finding the globally optimal solution by iteratively choosing the instance which promises the best solution for the moment. In most cases, though, this does not result in global but only local optimizers. However, the method has the advantages of being easy to understand and to implement.

Basically, orthogonal matching pursuit replaces ℓ_0-minimization by the optimization problem

$$\min_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{A}\mathbf{z}-\mathbf{y}\|_2 \quad \text{subject to } \|\mathbf{z}\|_0 \leq s, \tag{1.75}$$

which—assuming uniqueness of the respective minimizers—is equivalent to noise-aware ℓ_0-minimization (1.65), see, Lemma A.4.

Let us now describe how the algorithm proceeds. Therefore, we assume that the sparsity s of the signal $\mathbf{x}$, which we want to recover from the measurements (1.64), is known. The algorithm is initialized with $\hat{\mathbf{x}}^0_{\mathrm{OMP}} = \mathbf{0}$ and the corresponding residuum $\mathbf{r}^0 = \mathbf{y}$ as well as support set $\Lambda^0 = \emptyset$ are defined. The latter will be increased greedily until it reaches the desired size s, which is also the stopping criterion. At each iteration $k = 1, \ldots, s$ the index $j \in [N]$ maximizing the scalar product $\langle \mathbf{a}_j, \mathbf{r}^{k-1}\rangle$ is added to Λ^k. This ensures that the ℓ_2-norm of the residual is reduced as much as possible in this iteration, when setting

$$\hat{\mathbf{x}}^k_{\mathrm{OMP}} = \operatorname*{arg\,min}_{\hat{\mathbf{z}}:\mathrm{supp}\,(\hat{\mathbf{z}})\subset\Lambda^k} \|\mathbf{A}\hat{\mathbf{z}}-\mathbf{y}\|_2. \tag{1.76}$$

As (1.76) is essentially a k-dimensional least squares problem, it can be solved with standard tools from numerical linear algebra in a fast and stable manner. Finally, it remains to update the residual via $\mathbf{r}^k = \mathbf{y} - \mathbf{A}\hat{\mathbf{x}}^k_{\mathrm{OMP}}$. After s steps the algorithm finishes with $\hat{\mathbf{x}}_{\mathrm{OMP}} = \hat{\mathbf{x}}^s_{\mathrm{OMP}}$.

Of course, orthogonal matching pursuit as presented here is amenable to various adaptions. For instance, available prior information can be included or different stopping criteria can be used to promote sparser or less sparse solutions, or to guarantee a specific bound on the norm of the final residual.

From an analytic point of view, under a restricted isometry property data fidelity of the solution can be assured. Moreover, in the noiseless case exact recovery can be guaranteed, see, e.g., [DW10].

A more enhanced greedy method, which is ultimately based on orthogonal matching pursuit, is compressive sampling matching pursuit (CoSaMP), see, e.g., [NT09].

1.4.2 Linear and Second-Order Cone Programming

For this subsection, let us focus on the convex relaxation (1.66) of (1.65). It turns out that the noise-aware ℓ_1-minimization problem can be recast as the second order cone program

$$\min_{\hat{\boldsymbol{\zeta}}\in\mathbb{R}^{2N}} \sum_{i=1}^{2N} \hat{\zeta}_i \quad \text{subject to } \hat{\boldsymbol{\zeta}} \geq 0, \ \left\|(\mathbf{A}|-\mathbf{A})\,\hat{\boldsymbol{\zeta}} - \mathbf{y}\right\|_2 \leq \eta. \tag{1.77}$$

Here, the solution $\hat{\mathbf{x}}$ to (1.66) can be simply recovered via $\hat{\mathbf{x}} = (\mathrm{Id}\,|-\mathrm{Id})\,\hat{\boldsymbol{\xi}}$, where we denote the minimizer of (1.77) by $\hat{\boldsymbol{\xi}}$. In the noiseless case this even reduces to a linear program.

For both problems, standard algorithms from convex optimization can be employed, such as interior point methods, or in the case of a linear program the simplex method [BV04]. Furthermore, there are also several software package available.

However, due to the general applicability of such methods we cannot expect optimal performance and anticipate that specialized algorithms are capable of outperforming them.

1.4.3 Iterative Thresholding Algorithms

Motivated by the equivalence between noise-aware ℓ_1-minimization (1.66) and the least absolute shrinkage and selection operator (1.67), let us for the moment consider the optimization problem of the very general form

$$\min_{\mathbf{z}\in\mathbb{R}^N} f(\mathbf{z}) + g(\mathbf{z}). \tag{1.78}$$

Here, $f : \mathbb{R}^N \to \mathbb{R}$ is a convex and differentiable function, which—in our setting—will be a measure for data fidelity, whereas $g : \mathbb{R}^N \to [0, \infty]$ is a lower semi-continuous function, which is assumed to be convex. Furthermore it is typically non-differentiable and will describe a penalty term in our context. Therefore, we cannot utilize a simple gradient descent method for minimization, since this would require a gradient. However, so-called forward-backward splitting methods (FBS), also known as proximal gradient methods, can be used to approach such a problem. In order to reason this terminology, let us firstly introduce the proximal mapping of a function g as

$$\operatorname{prox}_g(\mathbf{z}) = \operatorname*{arg\,min}_{\mathbf{v}\in\mathbb{R}^N} g(\mathbf{v}) + \frac{1}{2}\|\mathbf{v}-\mathbf{z}\|_2^2. \tag{1.79}$$

It is well-known that strong convexity of the objective function guarantees uniqueness of the minimizer making the proximal operator well-defined. We want to point out that this could not be assured in general if g was non-convex.

Since the proximal mapping returns a point, which is on the one hand close to the minimizer of g but also not far from $\mathbf{z}$ on the other hand, it can be used iteratively for minimization. Let us now explain why this is usually called backward gradient descent step. Therefore, by definition of the subdifferential, we note that the proximal mapping satisfies $\operatorname{prox}_g(\mathbf{z}) \in \mathbf{z} - \partial g(\operatorname{prox}_g(\mathbf{z}))$, which resembles a common gradient descent step. Though, with the difference that the subgradient is evaluated at the endpoint rather than the starting point $\mathbf{z}$.

Now, by performing iteratively and alternatively a (forward) gradient descent step for the convex and smooth component f and subsequently a backward gradient descent step for the convex but non-smooth component g, we obtain the forward-backward splitting method as described in Algorithm 1 for the minimization of their sum.

Algorithm 1 Forward-Backward Splitting Method (FBS)

```
Input: Functions $f, g : \mathbb{R}^N \to \mathbb{R}$, step sizes $(t^k)_{k=1}^K$ and number of iterations $K$.
Output: Minimizer $\hat{\mathbf{x}}_{\mathrm{FBS}}$.
1: Set $\hat{\mathbf{x}}_{\mathrm{FBS}}^0 = \mathbf{0} \in \mathbb{R}^N$ and $k = 0$.
2: while $k \le K$ and stopping criterion not fulfilled
3:     Set $k = k + 1$.
4:     $\hat{\mathbf{x}}_{\mathrm{FBS}}^k = \operatorname{prox}_{t^k g}\left(\hat{\mathbf{x}}_{\mathrm{FBS}}^{k-1} - t^k \nabla f(\hat{\mathbf{x}}_{\mathrm{FBS}}^{k-1})\right)$
5: end while
6: Set $\hat{\mathbf{x}}_{\mathrm{FBS}} = \hat{\mathbf{x}}_{\mathrm{FBS}}^k$.
```

For a convergence analysis we refer to [GSB14] and just make the remark that a sufficient criterion in the case of a constant step size is $t^k = t < 2/L$ for all k, where L denotes the Lipschitz constant of ∇f[4].

[4] For $f(\mathbf{z}) = \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2^2$ we immediately obtain $(\nabla f)(\mathbf{z}) = 2\mathbf{A}^T(\mathbf{A}\mathbf{z} - \mathbf{y})$ and thus $L = 2\|\mathbf{A}^T\mathbf{A}\|$.

Before coming to the first iterative thresholding algorithm, which can be derived immediately from what we observed, we want to spend a few words on the situation, where this is not the case, namely when g is non-convex. In general, as the proximal mapping may become a point-to-set mapping, also known as the proximal correspondence, in such a setting neither convergence to a global minimizer nor independence of the initial iterate can be guaranteed. Nevertheless, having this in mind and taking suitable precautions, allows to use the forward-backward splitting method also for non-convex problems [GSB14, Section 3.3].

Iterative Soft Thresholding. Applying the forward-backward splitting method in the setting $f(\mathbf{z}) = \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2^2$ and $g(\mathbf{z}) = \beta\|\mathbf{z}\|_1$ results in the iterative soft thresholding algorithm (ISTA), which solves the LASSO problem (1.67), cf. [DDDM04]. Its popularity is favored by its easy applicability and evaluation possibility as an explicit expression of the proximal mapping of $\|\cdot\|_1$ can be given in terms of the soft thresholding operator. A derivation thereof can be found in, e.g., [For10, Lemma 4.1]. Since the ℓ_1-norm is fully separable, the proximal mapping can be evaluated component-wise, i.e.,

$$\operatorname{prox}_{\beta\|\cdot\|_1}(\mathbf{z}) = \mathbb{S}_{2\beta}(\mathbf{z}) = \left(S_{2\beta}(z_i)\right)_{i=1}^N, \tag{1.80}$$

where $S_\beta : \mathbb{R} \to \mathbb{R}$ denotes the scalar soft thresholding operator with threshold β and is given by

$$S_\beta(z) = \begin{cases} z - \frac{\beta}{2} & \text{if } z > \frac{\beta}{2}, \\ 0 & \text{if } |z| \le \frac{\beta}{2}, \\ z + \frac{\beta}{2} & \text{if } z < -\frac{\beta}{2}. \end{cases} \tag{1.81}$$

A visualization of this function is given in Figure 1.3 on the very right.
When using the constant step size $t^k = 1/2$ for all k, line 4 in Algorithm 1 becomes the update rule of the iterative soft thresholding algorithm

$$\hat{\mathbf{x}}^k_{\text{ISTA}} = \mathbb{S}_\beta\left(\hat{\mathbf{x}}^{k-1}_{\text{ISTA}} - \mathbf{A}^T(\mathbf{A}\hat{\mathbf{x}}^{k-1}_{\text{ISTA}} - \mathbf{y})\right). \tag{1.82}$$

For a detailed convergence analysis thereof we refer to [For10, Section 4.1], where besides a proof of strong convergence, which can be guaranteed under the assumption $\|\mathbf{A}\| < \sqrt{2}$, also acceleration techniques such as the decreasing iterative soft thresholding algorithm (D-ISTA) are presented. We want to emphasize at this point that the proof of convergence relies solely on tools of convex analysis. Consequently, iterative soft thresholding yields a meaningful solution with good data fidelity, i.e., a small residual, and with a small ℓ_1-norm even if there is no restricted isometry property fulfilled. More precisely, see, e.g., [Mal19], for a minimizer $\hat{\mathbf{x}}$ of (1.67) it holds for these two quantities

$$\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_2^2 \le \|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_2^2 + \beta\|\hat{\mathbf{x}}\|_1 \le \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \beta\|\mathbf{x}\|_1 = \|\boldsymbol{\eta}\|_2^2 + \beta\|\mathbf{x}\|_1 \tag{1.83}$$

and

$$\|\hat{\mathbf{x}}\|_1 \le \frac{1}{\beta}\left(\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_2^2 + \beta\|\hat{\mathbf{x}}\|_1\right) \le \frac{1}{\beta}\left(\|\mathbf{A}\mathbf{0} - \mathbf{y}\|_2^2 + \beta\|\mathbf{0}\|_1\right) = \frac{1}{\beta}\|\mathbf{y}\|_2^2. \tag{1.84}$$

Thus, there is a trade-off between data fidelity and sparsity, which can be controlled by β. Note furthermore that the assumption on the spectral norm of $\mathbf{A}$ is in practice not restrictive, as it can be assured by an appropriate rescaling of $\mathbf{A}$, $\mathbf{y}$ and β in any case. Let us end this paragraph with the following corollary, which is essentially a consequence of Theorem 1.27 and applies to a limit point $\hat{\mathbf{x}}_{\mathrm{ISTA}}$ of the iterative soft thresholding algorithm.

Corollary 1.29 (cf. [Mal19, Theorem 2.3.5]). *Let us assume that* $\mathbf{A} \in \mathbb{R}^{m \times N}$ *has the restricted isometry property of order* $2s$ *with constant* $0 < \delta < \sqrt{2} - 1$. *Then, for a minimizer* $\hat{\mathbf{x}}_\beta$ *of* (1.67) *it holds*

$$\|\mathbf{x} - \hat{\mathbf{x}}_\beta\|_2 \leq C_0 \frac{\sigma_s(\mathbf{x})_1}{\sqrt{s}} + C_1 \eta_\beta, \tag{1.85}$$

if β *was chosen such that* $\eta_\beta := \|\mathbf{A}\hat{\mathbf{x}}_\beta - \mathbf{y}\| \geq \eta$. *The constants* C_0 *and* C_1 *are the ones from Theorem 1.27.*

Proof. According to Lemma 1.23(ii) $\hat{\mathbf{x}}_\beta$ is a minimizer of (1.66) with η_β instead of η. Thus, Theorem 1.27 can be used with η_β instead of η, which yields the claim. □

Iterative Bridge Thresholding. Due to the non-convex penalty term $g(\mathbf{z}) = \beta\|\mathbf{z}\|_q^q$, the theory underlying the forward-backward splitting methods cannot be applied directly if $0 < q < 1$, as it relies heavily on tools from convex analysis. However, in [BLR15] an alternative approach, the generalized gradient projection method (GGPM), was proposed, which aims at the minimization of functionals that are the sum of a smooth part f and a non-smooth and non-convex part g. More precisely[5], let us assume that the function $f : \mathbb{R}^N \to [0, \infty)$ is differentiable and has a Lipschitz continuous derivative with Lipschitz constant L. In turn, the function $g : \mathbb{R}^N \to [0, \infty]$ is assumed to be proper, lower semicontinuous and coercive. The generalized gradient projection method now follows a similar idea as the forward-backward splitting method, namely alternating between forward and backward gradient steps on the two distinct parts, respectively. Its update rule is given by

$$\begin{aligned} \hat{\mathbf{x}}^k_{\mathrm{GGPM}} &\in \operatorname{prox}_{t^k g}\left(\hat{\mathbf{x}}^{k-1}_{\mathrm{GGPM}} - t^k \nabla f(\hat{\mathbf{x}}^{k-1}_{\mathrm{GGPM}})\right) \\ &= \underset{\mathbf{v} \in \mathbb{R}^N}{\arg\min}\; t^k g(\mathbf{v}) + \frac{1}{2}\left\|\mathbf{v} - \left(\hat{\mathbf{x}}^{k-1}_{\mathrm{GGPM}} - t^k \nabla f(\hat{\mathbf{x}}^{k-1}_{\mathrm{GGPM}})\right)\right\|_2^2 \end{aligned} \tag{1.86}$$

for suitable step sizes $(t^k)_{k=1}^K$. The difference and difficulty, though, is that the proximal mapping of the non-convex g may be set-valued, i.e., there can exist several global and moreover various local minimizers for the optimization problem in (1.79). Despite this fact, under the additional assumption that the step size is constant and fulfills $t^k = t < 1/L$, it was shown in [BLR15] that a sequence generated in the described manner can firstly be defined in the sense that the right-hand side in (1.86) is not empty. Secondly, at least it decreases the objective function, but is neither guaranteed to reach a global minimizer in general nor to converge at all. However, any global minimizer is a fixed point of the algorithm. For a more detailed presentation we refer to the cited literature.

A similar approach was taken in [ABRS10, ABS13]. Under a slightly stronger assumption, namely that $f + g$ additionally fulfills the so-called Kurdyka-Łojasiewicz inequality as well as that g is continuous, convergence of the sequence of iterates to a critical point of the objective function can be established if the generated sequence is bounded.

[5] For an application in infinite dimensional spaces, we refer to the cited literature.

For a proof that our objective function $\|\mathbf{Az}-\mathbf{y}\|_2^2+\beta\|\mathbf{z}\|_q^q$ indeed satisfies the Kurdyka-Łojasiewicz inequality we refer to the discussion after Example 5.4(b) from [ABS13] and the literature referenced therein. Therefore, these convergence results can be applied.
It remains to investigate the proximal mappings of the ℓ_q-(quasi)-norms. Exploiting the fact that $g(\mathbf{z})=\beta\|\mathbf{z}\|_q^q$ is fully separable, for its multivalued proximal mapping one can derive

$$\operatorname{prox}_{\beta\|\,\cdot\,\|_q^q}(\mathbf{z})=\mathbb{B}_{2\beta}(\mathbf{z})=\left(B_{2\beta}(z_i)\right)_{i=1}^N\,, \tag{1.87}$$

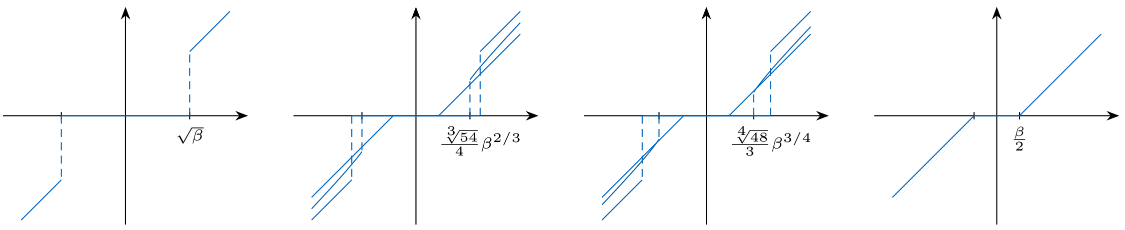
where $B_\beta^q:\mathbb{R}\rightrightarrows\mathbb{R}$ denotes the multivalued scalar bridge-q thresholding operator with threshold β. Admitting a slight abuse of notation—it is given by

$$B_\beta^q(z)=\begin{cases}0 & \text{if } |z|\le\tau_\beta^q,\\ \left(\,\cdot\,+\frac{\beta}{2}q\operatorname{sgn}(\,\cdot\,)|\cdot|^{q-1}\right)^{-1}(z) & \text{if } |z|\ge\tau_\beta^q,\end{cases} \tag{1.88}$$

with threshold $\tau_\beta^q=\frac{2-q}{2-2q}\left(\beta(1-q)\right)^{1/(2-q)}$. Note that the non-convexity induces the discontinuity at the threshold τ_β^q, where both 0 and $(\beta(1-q))^{1/(2-q)}$ are suitable minimizers. However, these are the only two points, where the minimizer is not unique.
For a complete derivation that $B_{2\beta}^q(z)$ is the solution to the one-dimensional optimization problem

$$\min_{v\in\mathbb{R}}\beta|v|^q+\frac{1}{2}(v-z)^2 \tag{1.89}$$

we refer to [BLR15, Lemma 5.1]. Nevertheless, let us for the moment assume that the regularization parameter q is rational, i.e., $q=a/b$ for $a,b\in\mathbb{N}$. Then it turns out that the global minimizer is, respectively, that the global minimizers are among the roots of a certain polynomial of degree $2b-a$ and the candidate 0. According to the Abel-Ruffini theorem there is in general only a closed-form algebraic expression of these roots if the degree is less than or equal to four. For these cases, besides the standard quadratic formula, Cardano's and Ferrari's formula provide such an expression for a third and fourth order polynomial, respectively. This applies to $q=1/2$ and $q=2/3$. The final choice among these candidates can then be made by comparing the objective values [KF09]. In fact, for these two cases, a complete analytic expression of the thresholding operator (1.88) was derived in [XCXZ12] for $B_\beta^{1/2}$ and in [CSX13] for $B_\beta^{2/3}$. These two operators are depicted in the two middle subfigures of Figure 1.3.



(a) H_β, $q=0$ (b) $B_\beta^{1/2}$, $q=1/2$ (c) $B_\beta^{2/3}$, $q=2/3$ (d) S_β, $q=1$

Figure 1.3. Scalar thresholding operators associated with different values of q. For reference, in the cases where $0<q<1$ we include the hard and soft thresholding operator.

We note that B_β^q evolves from the former investigated soft thresholding operator to the hard thresholding operator as the parameter q tends from 1 to 0, cf. [Lor04]. The latter

will be introduced in the subsequent paragraph as the solution to the limiting case $q = 0$ in (1.89), i.e., by using that $|v|^q \to \mathbb{1}_{v\neq 0}$ as $q \to 0$.
Conclusively let us state the iterative bridge-q thresholding algorithm (IBTA) for the constant step size $t^k = 1/2$. It has the update rule

$$\hat{\mathbf{x}}^k_{\mathrm{IBTA}} \in \mathbb{B}^q_\beta \left(\hat{\mathbf{x}}^{k-1}_{\mathrm{IBTA}} - \mathbf{A}^T(\mathbf{A}\hat{\mathbf{x}}^{k-1}_{\mathrm{IBTA}} - \mathbf{y})\right). \tag{1.90}$$

The results from above apply if $\|\mathbf{A}\| < 1$. In the case that iterative bridge-q thresholding finds a global minimizer, analogously to (1.83) and (1.84) we obtain for data fidelity and sparsity

$$\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_2^2 \leq \|\boldsymbol{\eta}\|_2^2 + \beta\|\mathbf{x}\|_q^q \quad \text{and} \quad \|\hat{\mathbf{x}}\|_q^q \leq \frac{1}{\beta}\|\mathbf{y}\|_2^2, \tag{1.91}$$

respectively.

Iterative Hard Thresholding and Iterative Best s-Term Approximation. After having examined bridge-q thresholding let us lastly also discuss case, where the thresholding operator can be associated with the ℓ_0-norm. There exist two different versions of the so-called iterative hard thresholding algorithm. The first one builds, analogously to the previous considerations, on the regularized version of (1.75), which can be obtained from (1.78) with $f(\mathbf{z}) = \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2^2$ and $g(\mathbf{z}) = \beta\|\mathbf{z}\|_0$. For the second one, in turn, the ℓ_0-constrained optimization problem (1.75) serves as starting point, which was also underlying orthogonal matching pursuit and was shown to be an equivalent formulation of (1.65).
Let us start with the first one. In order to derive the hard thresholding algorithm for the ℓ_0-regularized problem we define the set-valued hard thresholding operator as

$$\mathbb{H}_\beta(\mathbf{z}) = \operatorname*{arg\,min}_{\mathbf{v}\in\mathbb{R}^N} \beta\|\mathbf{v}\|_0 + \|\mathbf{v} - \mathbf{z}\|_2^2. \tag{1.92}$$

It is easy to verify that $\mathbb{H}_\beta(\mathbf{z}) = (H_\beta(z_i))_{i=1}^N$ where H_β denotes—again with a slight abuse of notation—the scalar hard thresholding operator

$$H_\beta(z) = \begin{cases} 0 & \text{if } |z| \leq \sqrt{\beta}, \\ z & \text{if } |z| \geq \sqrt{\beta}, \end{cases} \tag{1.93}$$

which is plotted in Figure 1.3 on the very left. This can be done by firstly observing that the objective function in (1.92) is fully separable, yielding a minimization problem over the function $\beta\mathbb{1}_{v\neq 0} + (v - z)^2$. The form of the operator then follows by comparing the values of the reasonable candidates for minimization, $v = 0$ and $v = z$, and choosing the better one. It shall be noted that there is no unique choice for $z = \pm\sqrt{\beta}$.
This motivates to define the update rule of the iterative hard thresholding algorithm by

$$\hat{\mathbf{x}}^k_{\mathrm{IHT}} \in \mathbb{H}_\beta \left(\hat{\mathbf{x}}^{k-1}_{\mathrm{IHT}} - \mathbf{A}^T(\mathbf{A}\hat{\mathbf{x}}^{k-1}_{\mathrm{IHT}} - \mathbf{y})\right) \tag{1.94}$$

in order to approach a minimizer of the ℓ_0-regularized minimization problem. However, as g in non-convex we cannot take advantage of the analysis of the forward-backward splitting methods. In [BD08] a full proof of convergence is provided, which essentially consists of two parts. At first, one shows that after a finite number of iterations the

sets of zero and non-zero entries are fixed. Thus, the algorithm reduces to a Landweber iteration, which is guaranteed to converge linearly assuming that $\|\mathbf{A}\| < 1$.
Even though the iterative hard thresholding algorithm guarantees to find a local minimizer of the ℓ_0-regularized minimization problem [BD08, Theorem 3], a solution is not guaranteed to be sparse, which is undesirable. This behaves differently for the second version, i.e., when considering the ℓ_0-constrained optimization problem, as only sparse solutions are feasible. Of course, this requires the knowledge of the sparsity in advance. A suitable algorithm, called the s-sparse algorithm or the iterative best s-term approximation, was introduced in [BD08] as

$$\hat{\mathbf{x}}^k_{\mathrm{IBA}} = \left(\hat{\mathbf{x}}^{k-1}_{\mathrm{IBA}} - \mathbf{A}^T(\mathbf{A}\hat{\mathbf{x}}^{k-1}_{\mathrm{IBA}} - \mathbf{y})\right)_{[s]}, \tag{1.95}$$

where the hard thresholding operator is replaced by the non-linear best s-term approximation. Using similar techniques as before, convergence to a local minimizer of (1.75) can be established under the assumption $\|\mathbf{A}\| < 1$, see, e.g., [BD08, Section 3].
At the end of this paragraph let us provide the following convergence result from [BD09] in the case of $\mathbf{A}$ having a restricted isometry property.

Theorem 1.30 ([BD09, Theorem 4])**.** *Let us assume that $\mathbf{A} \in \mathbb{R}^{m\times N}$ has the restricted isometry property of order $3s$ with constant $0 < \delta < 1/32$. Then, at iteration k, it holds*

$$\|\mathbf{x} - \hat{\mathbf{x}}^k_{\mathrm{IBA}}\|_2 \le 2^{-k}\|\mathbf{x}_{[s]}\|_2 + 6\epsilon_s, \tag{1.96}$$

where $\epsilon_s := \sigma_s(\mathbf{x})_2 + \frac{1}{\sqrt{s}}\sigma_s(\mathbf{x})_1 + \|\boldsymbol{\eta}\|_2$. Furthermore, after at most $K = \left\lceil \log_2\left(\|\mathbf{x}_{[s]}\|_2/\epsilon_s\right)\right\rceil$ iterations it holds

$$\|\mathbf{x} - \hat{\mathbf{x}}^K_{\mathrm{IBA}}\|_2 \le 7\epsilon_s. \tag{1.97}$$

Further Iterative Thresholding Methods. To end this section on iterative thresholding based methods, we want to point out that in recent years several different ideas to modify and improve the existing algorithms were proposed. Some of them may have originated in a different context and were then transferred to the setting of sparse recovery and compressed sensing. We want to name just a few of them. Firm thresholding, which interpolates between soft and hard thresholding, was introduced in [GB97] and analyzed in [FR08]. A smoother version thereof is garrote thresholding, proposed in [Bre95]. A further thresholding operator lying between soft and hard thresholding was put forward in [FW10].

Chapter 2

Low-Rank Matrix Recovery

We extend the sparse recovery and compressed sensing framework in this chapter in a straightforward way to matrices. At first we briefly address the illustrative example of matrix completion, where we would also like to take the opportunity to reveal a subtle difference between sparse recovery and compressed sensing. Afterwards we turn towards matrix sensing, where we investigate under which conditions on the measurement process low-rank matrices can be recovered from a minimal number of measurements. This is followed by an outline of a selection of the in practice most used algorithms to tackle this sort of recovery problem.

For an overview of low-rank matrix recovery from incomplete observations we refer to the paper [DR16] of the same name by Davenport and Romberg, which provides a general but concise outline of the broad field. For matrix completion in specific we recommend [CR09] and for the more general case of matrix sensing we suggest [RFP10].

2.1 Matrix Completion

As sketched in the introduction, recommender systems are concerned with the recovery of a partially unknown data matrix from some of its known entries. To elaborate on this connection, let us use the example of the Netflix Prize problem. Imagine that each of the n_1 customers of the platform has the opportunity to rate some of the n_2 available movies according to his or her preferences with a number from, say, 1 to 10. This results in an $n_1 \times n_2$ data matrix $\mathbf{X}$, in which the ratings of each customer for the respective movies are stored. Since most users usually only rate few movies, most parts of the matrix are empty. However, a company like Netflix is interested in these very entries in order to be able to recommend a specific users unseen movies, which he or she might like. In mathematical terms, one is interested in recovering the complete data matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ from some of its sampled entries x_{ij} for $(i, j) \in \Omega$, where Ω denotes a subset of $[n_1] \times [n_2]$. This subset corresponds to the available ratings and we assume that there are m such tuples in Ω, i.e., the total number of ratings submitted by all users is m.

Evidently, without having additional information about the matrix $\mathbf{X}$ this problem is highly ill-posed. However, in many practical applications it turns out to be reasonable to assume that the data matrix $\mathbf{X}$ is of low rank R or close to low rank. In our example this kind of sparsity is legitimate to be expected as in general only very few factors contribute to a person's individual preferences.

These considerations result in the rank-minimization problem

$$\min_{\mathbf{Z}\in\mathbb{R}^{n_1\times n_2}} \operatorname{rank}\mathbf{Z} \quad \text{subject to } z_{ij} = x_{ij} \text{ for all } (i,j)\in\Omega, \tag{2.1}$$

which is, similarly to the ℓ_0-minimization problem (1.7), unfortunately NP-hard. Before highlighting a significant difference to the compressed sensing framework from Chapter 1, let us draw a connection to this optimization problem. To this end let us recall the singular value decomposition (SVD) of a rank-R matrix $\mathbf{Z}\in\mathbb{R}^{n_1\times n_2}$ as the decomposition $\mathbf{Z} = \mathbf{U\Sigma V}^T$. Furthermore, denote the vector containing the singular values of $\mathbf{Z}$ by $\boldsymbol{\sigma} = \operatorname{diag}\boldsymbol{\Sigma}$. For the Schatten-$q$-(quasi)-norm $\|\mathbf{Z}\|_q$ of $\mathbf{Z}$ it holds $\|\mathbf{Z}\|_q = \|\boldsymbol{\sigma}\|_q$, as we saw in the notations paragraph in the introduction. Consequently, we can associate the rank of $\mathbf{Z}$ with the Schatten-0-norm $\|\mathbf{Z}\|_0$, which counts the singular values and is actually not even a quasi-norm. Moreover, the ℓ_1-norm of the vector containing the singular values matches the nuclear norm $\|\mathbf{Z}\|_*$ of $\mathbf{Z}$.

Inspired by ℓ_1-minimization in the sparse recovery and compressed sensing framework, the nuclear norm minimization problem

$$\min_{\mathbf{Z}\in\mathbb{R}^{n_1\times n_2}} \|\mathbf{Z}\|_* \quad \text{subject to } z_{ij} = x_{ij} \text{ for all } (i,j)\in\Omega \tag{2.2}$$

was proposed in [CR09], which is convex and thus computationally tractable. Further details about this optimization problem in a more general form are addressed in the following Section 2.2 on matrix sensing.

Let us now discuss a delicate difference compared to the compressed sensing framework. In contrast thereto, in this section we do not have the freedom to design a measurement operator according to our wishes and needs. Instead, our measurements are determined by the sampling operator. Similarly it behaves in the area of sparse signal recovery or sparse approximation, where one has no control on the decoding process. Even though there is this slight difference in terminology and philosophy, the terms are sometimes used interchangeably.

In the setting of matrix completion, we are therefore obliged to pose assumptions on the data matrix $\mathbf{X}$ in order to make it recoverable. To clarify this, consider the example, where $\mathbf{X}$ is a rank-2 matrix such that $\mathbf{u}_1 = 1/\sqrt{2}(\mathbf{e}_i + \mathbf{e}_{i+1})$, $\mathbf{u}_2 = 1/\sqrt{2}(\mathbf{e}_i - \mathbf{e}_{i+1})$ and $\mathbf{v}_1 = 1/\sqrt{2}(\mathbf{e}_j + \mathbf{e}_{j+1})$, $\mathbf{v}_2 = 1/\sqrt{2}(\mathbf{e}_j - \mathbf{e}_{j+1})$, denoting the ith unit vector of the appropriate dimension by $\mathbf{e}_i$. Then this matrix has only at most 4 non-zero entries located within a 2×2-square with upper left corner at position (i,j). To recover this matrix a substantial proportion of the entries would need to be sampled or phrased differently, $\mathbf{X}$ is likely to lie in the null space of sampling operators for moderate m. Consequently, it is reasonable to require that the singular vectors are sufficiently uncorrelated with the canonical basis. They need to be spread out. A suitable type of measure, the coherence, was introduced in [CR09].

Definition 2.1 (Coherence, [CR09, Definition 1.2]). *Let $\mathbf{U}$ be a subspace of $\mathbb{R}^N$ of dimension R and $P_\mathbf{U}$ be the orthogonal projection onto $\mathbf{U}$. Then the coherence of $\mathbf{U}$ (vis-à-vis the standard basis $\{\mathbf{e}_i\}_{i=1}^N \subset \mathbb{R}^N$) is defined to be*

$$\mu(\mathbf{U}) = \frac{N}{R}\max_{i\in[N]}\|P_\mathbf{U}\mathbf{e}_i\|_2^2. \tag{2.3}$$

Note that $\mu(\mathbf{U}) \in [1, N/R]$, where minimal coherence is achieved when all entries in the with the subspace associated matrix $\mathbf{U}$ have the same magnitude $1/\sqrt{N}$. In contrast, as soon as $\mathbf{U}$ contains an element of the canonical basis, the coherence is maximal. Now, if a matrix $\mathbf{X}$ has a column and row space of low coherence, it is firstly unlikely for $\mathbf{X}$ to lie in the null space of the sampling operator and secondly each entry of $\mathbf{X}$ provides about the same amount of information about the matrix, cf. [Rec11].
Under a suitable incoherence assumption on the low-rank matrix $\mathbf{X}$, recovery via nuclear norm minimization can be assured with high probability. This is made more rigorous in the following theorem, for whose proof we refer to the literature.

Theorem 2.2 (Matrix Completion, [Rec11, Theorem 2])**.** *Let $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ be a matrix of rank R with singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Without loss of generality, let $n_1 \leq n_2$. Assume that $\mathbf{X}$ obeys the following two properties.*

A0 *The row and column spaces have coherences bounded from above by some $\mu_0 > 0$.*

A1 *The matrix $\mathbf{U}\mathbf{V}^T$ has a maximum entry bounded by $\mu_1 \sqrt{R/(n_1 n_2)}$ in absolute value for some $\mu_1 > 0$.*

Suppose that m entries of $\mathbf{X}$ are observed with locations sampled uniformly at random. Then, if

$$m \geq 32 \max\{\mu_0, \mu_1^2\} R(n_1 + n_2)\beta \log^2(2n_2) \tag{2.4}$$

for some $\beta > 1$, the minimizer to nuclear norm minimization (2.9) is unique and equal to $\mathbf{X}$ with probability at least $1 - 6\log(n_2)(n_1+n_2)^{2-2\beta} - n_2^{2-2\sqrt{\beta}}$.

To close this section let us discuss the required lower bound (2.4) on the number of observed entries by comparing it to the information theoretic limit. Any low-rank-R matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ has

$$R + \sum_{i=1}^{R}(n_1 - i) + \sum_{j=1}^{R}(n_2 - j) = R(n_1 + n_2 - R) \tag{2.5}$$

degrees of freedom. Thus, if $m < R(n_1+n_2-R)$, no matter which method is used, matrix completion is impossible from a purely information theoretic point of view. Notably, Theorem 2.2 reaches up to a (poly)logarithmic factor the information theoretic limit. Due to the coupon collector's effect, which occurs since the entries are assumed to be sampled uniformly at random, there is moreover no hope to improve (2.4) beyond a term of the order $n_2 \log(n_2)$, as this is the number of required samples to ensure that every column is taken into account at least once with high probability, cf. [CT10, Section 1.7].

2.2 Matrix Sensing

After having introduced the low-rank matrix recovery problem in the illustrative setting of matrix completion, in the following, we want to generalize the problem in the sense that we consider arbitrary rank-R matrices $\mathbf{X}$. Let us firstly investigate the geometry of this set. Therefore, let

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{2.6}$$

be a singular value decomposition of $\mathbf{X}$. Similarly to the set of s-sparse vectors, the set of rank-R matrices $\mathcal{S}^R$ forms a union of R-dimensional subspaces, where each subspace is associated with two fixed orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$, corresponding to the left and right singular vectors. In contrast, however, the union here contains uncountably many subspaces as $\mathbf{U}$ and $\mathbf{V}$ can vary continuously. It is furthermore straightforward to see that the set of rank-R matrices is non-convex.

We now aim at recovering a low-rank-R matrix $\mathbf{X}$ from few linear functionals about the matrix. Therefore, let us introduce the linear measurement operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ together with the corresponding measurement vector

$$\mathbf{y} = \mathcal{A}(\mathbf{X}). \tag{2.7}$$

This sensing process of acquiring m measurements can be parametrized by just as many matrices $\mathbf{A}_i \in \mathbb{R}^{n_1 \times n_2}$, where the individual measurements are obtained according to $y_i = \langle \mathbf{A}_i, \mathbf{X} \rangle_F = \operatorname{tr}(\mathbf{A}_i^T \mathbf{X})$ for $i \in [m]$.

For reconstructing $\mathbf{X}$ from incomplete measurements $\mathbf{y}$ we propose, following the ideas of (2.1), the rank-minimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}} \operatorname{rank} \mathbf{Z} \quad \text{subject to } \mathcal{A}(\mathbf{Z}) = \mathbf{y}. \tag{2.8}$$

Needlessly to say, problem (2.8) is NP-hard since it contains the ℓ_0-minimization problem as a special case. To be a little more specific, imagine that $\mathbf{Z}$ was constrained to be a diagonal matrix. It is then easy to see that the rank-minimization problem (2.8) reduces to the ℓ_0-minimization problem (1.7). For a detailed comparison of these two problem types, rank minimization on the one hand and support or cardinality minimization on the other, we refer to [RFP10, Section 2].

Let us now turn towards the convex relaxation of the non-convex rank-minimization problem (2.8). For this purpose, we note that the nuclear norm is the convex envelope of the rank on the set $\{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2} : \|\mathbf{Z}\| \leq 1\}$. It is straightforward to check that for any $\mathbf{Z}$ it holds $\operatorname{rank} \mathbf{Z} \geq \|\mathbf{Z}\|_* / \|\mathbf{Z}\|$, showing that on the given set the nuclear norm bounds the rank from below and is moreover convex. For a verification that this is in fact the tightest lower bound, which is also convex, see, e.g., [Faz02, Theorem 1]. This proves that the nuclear norm can be regarded as the convex envelope of the rank and motivates to consider the nuclear norm minimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}} \|\mathbf{Z}\|_* \quad \text{subject to } \mathcal{A}(\mathbf{Z}) = \mathbf{y}, \tag{2.9}$$

which is a convex optimization problem. To this end it can be tackled efficiently via semidefinite programming (SDP) and is consequently no longer NP-hard. Such an idea of relaxation in the matrix setting can be found already earlier as heuristics, for instance, in [Faz02, Section 5].

As the nuclear norm is the convex envelope of the rank, the rank of the minimizer to (2.8) can be lower bounded in terms of the minimizer to (2.9) in the following sense. Let $\mathbf{X}$ denote the solution to the former rank-minimization problem, and $\widehat{\mathbf{X}}$ the respective one to the nuclear norm minimization problem. Then, utilizing the properties of the particular optimizers in the first and last inequality, respectively, we observe the chain of inequalities

$$\frac{\|\widehat{\mathbf{X}}\|_*}{\|\mathbf{X}\|} \leq \frac{\|\mathbf{X}\|_*}{\|\mathbf{X}\|} \leq \operatorname{rank} \mathbf{X} \leq \operatorname{rank} \widehat{\mathbf{X}}. \tag{2.10}$$

Besides the lower bound on the rank we obtain an immediate upper bound as well. This raises the fundamental question under which conditions on the measurement operator $\mathcal{A}$ it can be guaranteed in the first place that the solution to (2.9) coincides with the one to (2.8), i.e., $\widehat{\mathbf{X}} = \mathbf{X}$.

2.2.1 A Restricted Isometry Property for Matrix Sensing

It turns out that a natural generalization of the vector-valued version, Definition 1.8, to matrices suffices. Apart from replacing the vector norm with the corresponding matrix norm, the set of s-sparse vectors, for which we required to have a near-isometry, is replaced by the set of at most rank-R matrices. This restricted isometry property was introduced and analyzed in [RFP10].

Definition 2.3 (Rank-R Restricted Isometry Property (Matrix RIP))**.** *A linear operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ satisfies the rank-R restricted isometry property with isometry constant $0 < \delta < 1$, if*

$$(1-\delta)\|\mathbf{Z}\|_F^2 \leq \|\mathcal{A}(\mathbf{Z})\|_2^2 \leq (1+\delta)\|\mathbf{Z}\|_F^2 \tag{2.11}$$

for all matrices $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$ of rank at most R.

Completely analogous to the sparse case, if $\mathcal{A}$ has the rank-$2R$ restricted isometry property, any two different rank-R matrices $\mathbf{Z}$ and $\mathbf{Z}'$ are distinguishable by their measurements. In particular, $\mathbf{X}$ is the only rank-R matrix satisfying $\mathbf{y} = \mathcal{A}(\mathbf{X})$, cf. [RFP10, Theorem 3.2].

Remark 2.4. Note that Definition 3.1 in [RFP10] uses an alternative version of the matrix restricted isometry property compared to our Definition 2.3. However, in order to stay consistent with the vector-valued analog, Definition 1.8, we use the version stated above. In fact, up to an algebraic modification of the RIP constants, both definitions are equivalent. More precisely, for $a, b \geq 0$ and $0 < \delta < 1$, it holds in general that for

$$(1-\delta)a^2 \leq b^2 \leq (1+\delta)a^2 \quad \text{and} \quad (1-\delta')a \leq b \leq (1+\delta')a,$$

the former implies the latter with $\delta' = \delta$, whereas, vice versa, the latter with $\delta' = \delta/3$ implies the former. Both versions, the one with as well as the one without squares, also appear in the compressed sensing literature.

Let us now address the question of how well nuclear norm minimization performs under the assumption that the measurement operator $\mathcal{A}$ has a suitable matrix restricted isometry property. Therefore, let us establish a bound on the reconstruction error in the Frobenius norm. In particular, we obtain that a rank-R matrix $\mathbf{X}$ is recovered exactly from its measurements, which itself was already proven in [RFP10, Theorem 3.3]. However, as we will show in the subsequent theorem, also for not necessarily low-rank matrices $\mathbf{X}$, a solution to the nuclear norm minimization problem obeys a reasonable stability guarantee. This is based on the slightly more general situation presented in [FCRP08, Theorem 4], where the ideas from [RFP10] were extended.

Theorem 2.5 ([FCRP08, Theorem 4])**.** *Let us assume that $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ has the rank-$5R$ restricted isometry property such that $3\delta_{5R} + 2\delta_{3R} < 1$ holds for the isometry* *constants. Then, for any matrix* $\mathbf{X}$*, a solution* $\widehat{\mathbf{X}}$ *to the nuclear norm minimization problem* (2.9) *with* $\mathbf{y} = \mathcal{A}(\mathbf{X})$ *fulfills*

$$\|\mathbf{X} - \widehat{\mathbf{X}}\|_F \leq C \frac{\|\mathbf{X} - \mathbf{X}_{[R]}\|_*}{\sqrt{R}}, \tag{2.12}$$

where $\mathbf{X}_{[R]}$ *denotes the rank-*R *matrix that best approximates* $\mathbf{X}$*.*

To prove this theorem we need the following two technical lemmas, for whose proof we refer to [RFP10].

Lemma 2.6 ([RFP10, Lemma 2.3])**.** *Let* $\mathbf{X}$ *and* $\mathbf{H}$ *be matrices of the same dimensions. If their row and column spaces are orthogonal, i.e.,* $\mathbf{X}\mathbf{H}^T = \mathbf{0}$ *and* $\mathbf{X}^T\mathbf{H} = \mathbf{0}$*, then the nuclear norm is additive, i.e.,* $\|\mathbf{X} + \mathbf{H}\|_* = \|\mathbf{X}\|_* + \|\mathbf{H}\|_*$*.*

Lemma 2.7 ([RFP10, Lemma 3.4])**.** *Let* $\widetilde{\mathbf{X}}$ *and* $\mathbf{H}$ *be matrices of the same dimensions and moreover let* $\widetilde{\mathbf{X}}$ *have rank* R*. Then there exist matrices* $\mathbf{H}_0$ *and* $\mathbf{H}_c$ *such that*

(i) $\mathbf{H} = \mathbf{H}_0 + \mathbf{H}_c$*,*

(ii) $\operatorname{rank} \mathbf{H}_0 \leq 2R$*,*

(iii) $\widetilde{\mathbf{X}}\mathbf{H}_c^T = \mathbf{0}$ *and* $\widetilde{\mathbf{X}}^T\mathbf{H}_c = \mathbf{0}$*,*

(iv) $\langle \mathbf{H}_0, \mathbf{H}_c \rangle_F = 0$*.*

Sketch of Proof. Let $\widetilde{\mathbf{X}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^T$ denote a full singular value decomposition and note that $\widetilde{\mathbf{\Sigma}}$ is block diagonal with a diagonal matrix $\widetilde{\mathbf{\Sigma}}_{11} \in \mathbb{R}^{R \times R}$ and a rectangular matrix $\widetilde{\mathbf{\Sigma}}_{22} = \mathbf{0}$. Based thereon, set $\widehat{\mathbf{H}} = \widetilde{\mathbf{U}}^T\mathbf{H}\widetilde{\mathbf{V}}$ and disassemble it into blocks $\widehat{\mathbf{H}}_{11}$, $\widehat{\mathbf{H}}_{12}$, $\widehat{\mathbf{H}}_{21}$ and $\widehat{\mathbf{H}}_{22}$ aligned with the ones of $\widetilde{\mathbf{\Sigma}}$. Now, for

$$\mathbf{H}_0 = \widetilde{\mathbf{U}} \begin{pmatrix} \widehat{\mathbf{H}}_{11} & \widehat{\mathbf{H}}_{12} \\ \widehat{\mathbf{H}}_{21} & \mathbf{0} \end{pmatrix} \widetilde{\mathbf{V}}^T \quad \text{and} \quad \mathbf{H}_c = \widetilde{\mathbf{U}} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{H}}_{22} \end{pmatrix} \widetilde{\mathbf{V}}^T \tag{2.13}$$

it is straightforward to check that the desired properties hold. □

The following proof resembles the one of Theorem 1.14 and was presented in [FCRP08]. Only some details need to be adapted to fit into the matrix setting.

Proof of Theorem 2.5. Since $\mathbf{X}$ is feasible and $\widehat{\mathbf{X}}$ a solution to the nuclear norm minimization problem (2.9) with $\mathbf{y} = \mathcal{A}(\mathbf{X})$, we have by optimality $\|\widehat{\mathbf{X}}\|_* \leq \|\mathbf{X}\|_*$. Let $\mathbf{X} = \mathbf{X}_{[R]} + \mathbf{X}_c$ and let us now define $\mathbf{H} = \mathbf{X} - \widehat{\mathbf{X}} \in \ker \mathcal{A}$. According to Lemma 2.7, there exist matrices $\mathbf{H}_0$ and $\mathbf{H}_c$ such that $\mathbf{H} = \mathbf{H}_0 + \mathbf{H}_c$, $\operatorname{rank} \mathbf{H}_0 \leq 2R$, $\mathbf{X}_{[R]}\mathbf{H}_c^T = \mathbf{0}$, $\mathbf{X}_{[R]}^T\mathbf{H}_c = \mathbf{0}$ and $\langle \mathbf{H}_0, \mathbf{H}_c \rangle_F = 0$. Then, by utilizing Lemma 2.6 for the first inequality and the reverse triangle inequality twice for the second, we observe

$$\begin{aligned} \|\mathbf{X}_{[R]}\|_* + \|\mathbf{H}_c\|_* - \|\mathbf{X}_c\|_* - \|\mathbf{H}_0\|_* &\leq \|\mathbf{X}_{[R]} - \mathbf{H}_c\|_* - \|\mathbf{X}_c\|_* - \|\mathbf{H}_0\|_* \\ &\leq \|\mathbf{X}_{[R]} + \mathbf{X}_c - \mathbf{H}_0 - \mathbf{H}_c\|_* \\ &= \|\mathbf{X} - \mathbf{H}\|_* = \|\widehat{\mathbf{X}}\|_* \leq \|\mathbf{X}\|_*, \end{aligned} \tag{2.14}$$

which simplifies to

$$\|\mathbf{H}_c\|_* \leq \|\mathbf{H}_0\|_* + 2\|\mathbf{X}_c\|_*, \tag{2.15}$$

having used that $\|\mathbf{X}\|_* = \|\mathbf{X}_{[R]}\|_* + \|\mathbf{X}_c\|_*$. Let us now disassemble $\mathbf{H}_c$ into a sum of rank-$3R$ matrices $\mathbf{H}_1, \mathbf{H}_2, \ldots$ associated with a non-increasing rearrangement of the singular values of $\mathbf{H}_c$. Therefore, let $\mathbf{H}_c = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \mathbf{U}\,\mathrm{diag}\,(\boldsymbol{\sigma})\mathbf{V}^T$ denote the singular value decomposition of $\mathbf{H}_c$. Note that the singular values are already arranged in a non-decreasing order. Thus[6] we can define the index sets $T_\ell = \{i : 3R(\ell-1)+1 \leq i \leq 3R\ell\}$ for $\ell \geq 1$ and the associated matrices $\mathbf{H}_\ell = \mathbf{U}\,\mathrm{diag}(\boldsymbol{\sigma}|_{T_\ell})\mathbf{V}^T$. Due to this construction, it follows that for all $\ell \geq 1$ we have $\sigma_k \leq \frac{1}{3R}\sum_{j\in T_\ell}\sigma_j$ for all $k \in T_{\ell+1}$ and in consequence $\|\mathbf{H}_{\ell+1}\|_F^2 \leq \frac{1}{3R}\|\mathbf{H}_\ell\|_*^2$. Using this as well as inequality (2.15) and the fact that $\mathbf{H}_0$ is at most rank-$2R$, we can derive the bound

$$\begin{aligned}\sum_{\ell\geq 2}\|\mathbf{H}_\ell\|_F &\leq \frac{1}{\sqrt{3R}}\sum_{\ell\geq 1}\|\mathbf{H}_\ell\|_* = \frac{1}{\sqrt{3R}}\|\mathbf{H}_c\|_* \\ &\leq \frac{1}{\sqrt{3R}}\left(\|\mathbf{H}_0\|_* + 2\|\mathbf{X}_c\|_*\right) \leq \sqrt{\frac{2}{3}}\|\mathbf{H}_0\|_F + \frac{2}{\sqrt{3R}}\|\mathbf{X}_c\|_*.\end{aligned} \tag{2.16}$$

Now, by using both triangle inequalities and by applying the rank-$5R$ restricted isometry property with isometry constants δ_{5R} and δ_{3R}, respectively, we obtain

$$\begin{aligned}\|\mathcal{A}(\mathbf{H})\|_2 &= \left\|\mathcal{A}(\mathbf{H}_0+\mathbf{H}_1) + \sum_{\ell\geq 2}\mathcal{A}(\mathbf{H}_\ell)\right\|_2 \geq \|\mathcal{A}(\mathbf{H}_0+\mathbf{H}_1)\|_2 - \sum_{\ell\geq 2}\|\mathcal{A}(\mathbf{H}_\ell)\|_2 \\ &\geq \sqrt{1-\delta_{5R}}\|\mathbf{H}_0+\mathbf{H}_1\|_F - \sqrt{1+\delta_{3R}}\sum_{\ell\geq 2}\|\mathbf{H}_\ell\|_F \\ &\geq \sqrt{1-\delta_{5R}}\|\mathbf{H}_0+\mathbf{H}_1\|_F - \sqrt{1+\delta_{3R}}\left(\sqrt{\frac{2}{3}}\|\mathbf{H}_0\|_F + \frac{2}{\sqrt{3R}}\|\mathbf{X}_c\|_*\right) \\ &\geq \left(\sqrt{1-\delta_{5R}} - \sqrt{\frac{2}{3}}\sqrt{1+\delta_{3R}}\right)\|\mathbf{H}_0+\mathbf{H}_1\|_F - \frac{2}{\sqrt{3R}}\sqrt{1+\delta_{3R}}\|\mathbf{X}_c\|_*.\end{aligned} \tag{2.17}$$

The second inequality holds since $\mathrm{rank}\,(\mathbf{H}_0+\mathbf{H}_1) \leq 5R$, the next-to-last makes use of (2.16) and the last follows as $\mathbf{H}_0$ and $\mathbf{H}_1$ are orthogonal due to the construction of $\mathbf{H}_0$ and $\mathbf{H}_c$ in Lemma 2.7, which implies $\|\mathbf{H}_0+\mathbf{H}_1\|_F \geq \|\mathbf{H}_0\|_F$.

Finally, since $\mathbf{H} \in \ker\mathcal{A}$, we have $\|\mathcal{A}(\mathbf{H})\|_2 = 0$. With this we can rearrange (2.17) such that it provides a bound on $\|\mathbf{H}_0+\mathbf{H}_1\|_F$. This can be used to derive a bound on $\|\mathbf{H}\|_F$ as follows,

$$\begin{aligned}\|\mathbf{H}\|_F &\leq \|\mathbf{H}_0+\mathbf{H}_1\|_F + \sum_{\ell\geq 2}\|\mathbf{H}_\ell\|_F \leq \left(1+\sqrt{\frac{2}{3}}\right)\|\mathbf{H}_0+\mathbf{H}_1\|_F + \frac{2}{\sqrt{3R}}\|\mathbf{X}_c\|_* \\ &\leq \left(\left(1+\sqrt{\frac{2}{3}}\right)\frac{1}{C_\delta}\sqrt{1+\delta_{3R}}+1\right)\frac{2}{\sqrt{3R}}\|\mathbf{X}_c\|_* \leq C\frac{\|\mathbf{X}_c\|_*}{\sqrt{R}},\end{aligned} \tag{2.18}$$

abbreviating $C_\delta = \sqrt{1-\delta_{5R}} - \sqrt{\frac{2}{3}}\sqrt{1+\delta_{3R}}$. It follows again from simple algebraic computations that C_δ is greater than 0 if $3\delta_{5R} + 2\delta_{3R} < 1$, completing the proof. □

[6] Note that these steps mimic the common technique in compressed sensing, which was used in the proof of Theorem 1.10.

Furthermore, by utilizing the same proof technique and modifying only a few steps of the argument, a bound on the approximation error with respect to the nuclear norm can be established.

Theorem 2.8 ([FCRP08, Theorem 5]). *Let us assume that $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ has the rank-$5R$ restricted isometry property such that $3\delta_{5R} + 2\delta_{3R} < 1$ holds for the isometry constants. Then, for any matrix $\mathbf{X}$, a solution $\widehat{\mathbf{X}}$ to the nuclear norm minimization problem (2.9) with $\mathbf{y} = \mathcal{A}(\mathbf{X})$ fulfills*

$$\|\mathbf{X} - \widehat{\mathbf{X}}\|_* \leq C\big\|\mathbf{X} - \mathbf{X}_{[R]}\big\|_*. \tag{2.19}$$

The preceding theorem tells us that, under the assumption of a sufficient matrix restricted isometry property, nuclear norm minimization is, up to a moderate constant, quantitatively as good as approximation with the rank-R matrix $\mathbf{X}_{[R]}$ that best approximates $\mathbf{X}$. Note that $\mathbf{X}_{[R]}$ can be obtained by a truncated singular value decomposition of $\mathbf{X}$ according to the Eckart-Young theorem [HJ13, Subsection 7.4.2].
It remains to figure out which measurement operators are capable of guaranteeing a matrix restricted isometry property with high probability. Following [CP11] we will show that certain random operators are very likely to provide a near-isometry when the number of measurements is commensurate with the degrees of freedom (2.5) of an rank-R matrix $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$. The most well-known class of probability distributions obeying a suitable concentration of measure, which takes the center stage for establishing the rank-R restricted isometry property, is an ensemble with i.i.d. Gaussian entries.

Definition 2.9 (Gaussian Measurement Ensembles, cf. [CP11, Definition 2.2]). *The measurement operator $\mathcal{A}$ is called a Gaussian measurement ensemble if, for $i \in [m]$, each individual measurement matrix $\mathbf{A}_i$ contains i.i.d. mean-zero Gaussian entries of unit variance and if these individual measurement matrices are independent from each other as well.*

Remark 2.10. Contrarily to Definition 2.2 in [CP11], we require the entries of $\mathcal{A}$ to be standard normally distributed instead of having variance $1/m$. This is consistent with the considered measurement matrix in Theorem 1.11 for the vector-valued analog.

Now, if $\mathcal{A}$ is a Gaussian measurement ensemble we observe that $\mathsf{E}\big[\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{Z})\|_2^2\big] = \|\mathbf{Z}\|_F^2$ for any matrix $\mathbf{Z}$ and, as we will show in the subsequent theorem, that, under certain conditions, $\frac{1}{\sqrt{m}}\mathcal{A}$ has the rank-R restricted isometry property with high probability.

Theorem 2.11 (Gaussian Measurement Ensembles are RIP Operators, cf. [CP11, Theorem 2.3]). *Let $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ be a Gaussian measurement ensemble and assume that*

$$m \geq CR(n_1 + n_2 + 1) \tag{2.20}$$

holds for a constant $C > 0$, which only depends on $0 < \delta < 1$. Then, with probability at least $1 - 2\exp(-dm)$, where $d > 0$ denotes a constant, which only depends on δ as well, the operator $\frac{1}{\sqrt{m}}\mathcal{A}$ satisfies the rank-R restricted isometry property with isometry constant δ.

The remarkable novelty of this result, compared to previous versions such as [RFP10, Theorem 4.2] is that the number of necessary measurements is up to a constant factor at the information theoretic limit (2.5) and in particular involves no extra (poly)logarithmic

factors. This distinguishes this result also from related requirements on the measurement process in the compressed sensing framework.
For its proof we need two results of independent interest. The first is a standard concentration inequality for Gaussian random matrices in the form of the one in Lemma 1.12. In fact, Lemma 2.12 can be traced back to the former. The second provides an upper bound on the covering number for the set of low-rank matrices in order to extend the concentration inequality from single matrices to a finite cover of low-rank matrices with bounded Frobenius norm. Therefrom a uniform result can be derived. This differs slightly from the proceeding in the proof of the vector-valued analog, Theorem 1.11, where the separate subspaces were covered individually before a uniform bound over all such subspaces was applied. However, as we have uncountably many R-dimensional subspaces this is not possible.

Lemma 2.12. *Let $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ be a Gaussian measurement ensemble. Then, for all $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$ and $0 < \epsilon < 1$, it holds*

$$\mathsf{P}\left(\left| \left\| \frac{1}{\sqrt{m}} \mathcal{A}(\mathbf{Z}) \right\|_2^2 - \|\mathbf{Z}\|_F^2 \right| \geq \epsilon \|\mathbf{Z}\|_F^2 \right) \leq 2 \exp\left(-cm\epsilon^2\right), \tag{2.21}$$

where $c > 0$ is an absolute constant.

Proof. By setting $\mathbf{z} = \operatorname{vec}(\mathbf{Z}) \in \mathbb{R}^{n_1 n_2}$ and $\mathbf{A} = (\operatorname{vec}(\mathbf{A}_1), \ldots, \operatorname{vec}(\mathbf{A}_m))^T \in \mathbb{R}^{m \times (n_1 n_2)}$, the statement follows directly by applying Lemma 1.12 and noting that $\|\mathbf{z}\|_2^2 = \|\mathbf{Z}\|_F^2$ and $\mathcal{A}(\mathbf{Z}) = \mathbf{A} \operatorname{vec}(\mathbf{Z})$. □

Lemma 2.13 (Metric Entropy of the Set of Low-Rank Matrices, cf. [CP11, Lemma 3.1])**.** *Let $\mathcal{S}^{R,\Gamma} = \{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2} : \operatorname{rank} \mathbf{Z} \leq R \text{ and } \|\mathbf{Z}\|_F \leq \Gamma\}$. Then, for $0 < \epsilon \leq \Gamma$, there exists an ϵ-net $\left(\mathcal{S}^{R,\Gamma}\right)^{\#}$ with respect to the Frobenius norm obeying*

$$\left| \left(\mathcal{S}^{R,\Gamma}\right)^{\#} \right| \leq (18\Gamma/\epsilon)^{R(n_1+n_2+1)}, \tag{2.22}$$

i.e., for the metric entropy of $\mathcal{S}^{R,\Gamma}$ it holds

$$\log N(\mathcal{S}^{R,\Gamma}, \|\cdot\|_F, \epsilon) \leq R(n_1 + n_2 + 1) \log\left(18\Gamma/\epsilon\right). \tag{2.23}$$

Proof. Let $\mathbf{Z} = \mathbf{U\Sigma V}^T$ denote the singular value decomposition of any $\mathbf{Z} \in \mathcal{S}^{R,\Gamma}$ and note that $\|\mathbf{\Sigma}\|_F \leq \Gamma$. In order to construct an ϵ-net for $\mathcal{S}^{R,\Gamma}$, we cover the sets of permissible $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ individually. Let us start with $\mathbf{\Sigma}$. Therefore, let us consider the set D_Γ of diagonal $R \times R$ matrices with Frobenius norm bounded by Γ. For the covering number of this set it holds

$$N(D_\Gamma, \|\cdot\|_F, \epsilon) = N(\mathcal{B}_2^R(\mathbf{0}, \Gamma), \|\cdot\|_2, \epsilon) \leq \left(1 + \frac{2}{\epsilon/\Gamma}\right)^R \leq \left(\frac{3\Gamma}{\epsilon}\right)^R, \tag{2.24}$$

where the next-to-last inequality is a consequence of [Pis89, Lemma 4.16]. The last bound uses the assumption $\epsilon/\Gamma \leq 1$. With respect to $\|\cdot\|_F$, let $(D_\Gamma)^{\#}$ denote a minimal $\epsilon/3$-net. Thus, $\left|(D_\Gamma)^{\#}\right| \leq (9\Gamma/\epsilon)^R$. Now, let us turn to $\mathbf{U}$ and note that $\mathbf{V}$ can be considered analogously by replacing n_1 by n_2. To this end let $O_{n_1,R} = \{\mathbf{U} \in \mathbb{R}^{n_1 \times R} : \mathbf{U}^T\mathbf{U} = \mathbf{Id}\}$. In order to cover this set, introduce the $\|\cdot\|_{2,\infty}$ norm via $\|\mathbf{U}\|_{2,\infty} = \max_{i \in [R]} \|\mathbf{u}_i\|_2$ and note that $O_{n_1,R}$ is a subset of the unit ball $\mathcal{B}^{n_1}_{2,\infty}(\mathbf{0},1)$ under this norm. Hence, by using the bound on the covering number from above in this setting, there exists an $\epsilon/(3\Gamma)$-net $(O_{n_1,R})^{\#}$ for $O_{n_1,R}$ with $\left|(O_{n_1,R})^{\#}\right| \leq (18\Gamma/\epsilon)^{Rn_1}$. Note that an additional factor of 2 appears since $N(O_{n_1,R}, \|\cdot\|, \epsilon) \leq N(\mathcal{B}^{n_1}_{2,\infty}(\mathbf{0},1), \|\cdot\|, \epsilon/2)$ as $O_{n_1,R} \subset \mathcal{B}^{n_1}_{2,\infty}(\mathbf{0},1)$. Now define

$$(\mathcal{S}^{R,\Gamma})^{\#} = \left\{\mathbf{Z}^{\#} = \mathbf{U}^{\#}\mathbf{\Sigma}^{\#}(\mathbf{V}^{\#})^{T} : \mathbf{\Sigma}^{\#} \in (D_\Gamma)^{\#}, \mathbf{U}^{\#} \in (O_{n_1,R})^{\#} \text{ and } \mathbf{V}^{\#} \in (O_{n_2,R})^{\#}\right\} \tag{2.25}$$

and observe that $\left|(\mathcal{S}^{R,\Gamma})^{\#}\right| \leq \left|(O_{n_1,R})^{\#}\right|\left|(D_\Gamma)^{\#}\right|\left|(O_{n_2,R})^{\#}\right| \leq (18\Gamma/\epsilon)^{R(n_1+n_2+1)}$. It remains to show that $(\mathcal{S}^{R,\Gamma})^{\#}$ is an ϵ-net of $\mathcal{S}^{R,\Gamma}$ with respect to the Frobenius norm. Therefore, for a fixed $\mathbf{Z} = \mathbf{U\Sigma V}^T \in \mathcal{S}^{R,\Gamma}$ let $\mathbf{\Sigma}^{\#} \in (D_\Gamma)^{\#}$ with $\left\|\mathbf{\Sigma} - \mathbf{\Sigma}^{\#}\right\|_F \leq \epsilon/3$, $\mathbf{U}^{\#} \in (O_{n_1,R})^{\#}$ and $\mathbf{V}^{\#} \in (O_{n_2,R})^{\#}$ with $\left\|\mathbf{U} - \mathbf{U}^{\#}\right\|_{2,\infty} \leq \epsilon/(3\Gamma)$ and $\left\|\mathbf{V} - \mathbf{V}^{\#}\right\|_{2,\infty} \leq \epsilon/(3\Gamma)$, respectively. Then, by triangle inequality,

$$\begin{aligned}\left\|\mathbf{Z} - \mathbf{Z}^{\#}\right\|_F &= \left\|\mathbf{U\Sigma V}^T - \mathbf{U}^{\#}\mathbf{\Sigma}^{\#}(\mathbf{V}^T)^{\#}\right\|_F \\ &\leq \left\|\left(\mathbf{U} - \mathbf{U}^{\#}\right)\mathbf{\Sigma V}^T\right\|_F + \left\|\mathbf{U}^{\#}\left(\mathbf{\Sigma} - \mathbf{\Sigma}^{\#}\right)\mathbf{V}^T\right\|_F + \left\|\mathbf{U}^{\#}\mathbf{\Sigma}^{\#}\left(\mathbf{V}^T - (\mathbf{V}^T)^{\#}\right)\right\|_F \\ &= \left\|\left(\mathbf{U} - \mathbf{U}^{\#}\right)\mathbf{\Sigma}\right\|_F + \left\|\mathbf{\Sigma} - \mathbf{\Sigma}^{\#}\right\|_F + \left\|\mathbf{\Sigma}^{\#}\left(\mathbf{V}^T - (\mathbf{V}^T)^{\#}\right)\right\|_F \\ &\leq \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon,\end{aligned} \tag{2.26}$$

having exploited the orthogonality of the matrices $\mathbf{V}$ and $\mathbf{U}^{\#}$ in the third line and $\left\|\left(\mathbf{U} - \mathbf{U}^{\#}\right)\mathbf{\Sigma}\right\|_F = \sum_{i=1}^R \sigma_{ii}^2\|\mathbf{u}_i - \mathbf{u}_i^{\#}\|_2^2 \leq \|\mathbf{\Sigma}\|_F^2\|\mathbf{U} - \mathbf{U}^{\#}\|_{2,\infty}^2 \leq \Gamma^2(\epsilon/(3\Gamma))^2 = (\epsilon/3)^2$ in the inequality of the last line. This proves the claim. □

Proof of Theorem 2.11. Without loss of generality we can restrict ourselves to $\|\mathbf{Z}\|_F = 1$. Let us in a first step show that, with high probability, $\frac{1}{\sqrt{m}}\mathcal{A}$ is a near-isometry on a covering set of $\partial\mathcal{S}^{R,1} = \{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2} : \operatorname{rank}\mathbf{Z} \leq R \text{ and } \|\mathbf{Z}\|_F = 1\}$, which is known to obey

$$\left|\left(\partial\mathcal{S}^{R,1}\right)^{\#}\right| \leq \left(144\sqrt{2}/\delta\right)^{R(n_1+n_2+1)} \tag{2.27}$$

according to Lemma 2.13 with $\epsilon = \delta/(4\sqrt{2})$ and an additional factor of 2 coming up due to the monotonicity property of the covering number as $\partial\mathcal{S}^{R,1} \subset \mathcal{S}^{R,1}$. Using a union bound argument yields, for a constant $d > 0$, which may only depend on δ, by applying Lemma 2.12, that

$$\begin{aligned}\mathsf{P}\left(\max_{\mathbf{Z}^{\#} \in (\partial\mathcal{S}^{R,1})^{\#}}\left|\left\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{Z}^{\#})\right\|_2^2 - 1\right| \geq \delta/2\right) &\leq 2\left(144\sqrt{2}/\delta\right)^{R(n_1+n_2+1)}\exp\left(-cm\delta^2/4\right) \\ &\leq 2\exp\left(-dm\right),\end{aligned} \tag{2.28}$$

having used the assumption $m \geq CR(n_1 + n_2 + 1)$. Requiring $C > (\log(144\sqrt{2}/\delta))/(d\delta^2)$ guarantees that the constant $d = c\delta^2/4 - 1/C\log(144\sqrt{2}\delta)$ is positive. It remains to extend the derived result in a second step to the whole set $\partial\mathcal{S}^{R,1}$. Therefore, let us define the constant $B > 0$ as

$$B = \sup_{\mathbf{Z} \in \partial\mathcal{S}^{R,1}}\left\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{Z})\right\|_2. \tag{2.29}$$

Now, for any $\mathbf{Z} \in \partial\mathcal{S}^{R,1}$ there exists $\mathbf{Z}^{\#} \in \left(\partial\mathcal{S}^{R,1}\right)^{\#}$ with $\left\|\mathbf{Z} - \mathbf{Z}^{\#}\right\|_F \leq \delta/(4\sqrt{2})$, which yields with high probability

$$\left\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{Z})\right\|_2 \leq \left\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{Z}-\mathbf{Z}^{\#})\right\|_2 + \left\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{Z}^{\#})\right\|_2 \leq \left\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{Z}-\mathbf{Z}^{\#})\right\|_2 + (1+\delta/2). \tag{2.30}$$

By decomposing the rank-$2R$ matrix $\Delta\mathbf{Z} = \mathbf{Z} - \mathbf{Z}^{\#}$ into two orthogonal rank-R matrices $\Delta\mathbf{Z}_1$ and $\Delta\mathbf{Z}_2$, e.g., by splitting the singular value decomposition, we observe

$$\begin{aligned}\left\|\frac{1}{\sqrt{m}}\mathcal{A}(\Delta\mathbf{Z})\right\|_2 &\leq \left\|\frac{1}{\sqrt{m}}\mathcal{A}(\Delta\mathbf{Z}_1)\right\|_2 + \left\|\frac{1}{\sqrt{m}}\mathcal{A}(\Delta\mathbf{Z}_2)\right\|_2 \leq B\left(\|\Delta\mathbf{Z}_1\|_F + \|\Delta\mathbf{Z}_2\|_F\right) \\ &\leq B\sqrt{2}\|\Delta\mathbf{Z}\|_F = B\sqrt{2}\left\|\mathbf{Z}-\mathbf{Z}^{\#}\right\|_F \leq B\delta/4.\end{aligned} \tag{2.31}$$

We used that $\Delta\mathbf{Z}_i/\|\Delta\mathbf{Z}_i\|_F \in \partial\mathcal{S}^{R,1}$ for $i = 1, 2$ in the second inequality together with the definition of B and the Pythagoras theorem $\|\Delta\mathbf{Z}_1\|_F^2 + \|\Delta\mathbf{Z}_2\|_F^2 = \|\Delta\mathbf{Z}\|_F^2$ in the third inequality. With this we can derive

$$\left\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{Z})\right\|_2 \leq B\delta/4 + (1+\delta/2), \tag{2.32}$$

from which $B \leq B\delta/4 + (1+\delta/2)$ follows as the previous inequality holds for all $\mathbf{Z} \in \partial\mathcal{S}^{R,1}$. This, in turns, entails $B \leq (1+\delta/4)/(1+\delta/2) \leq 1+\delta$ establishing the upper bound of the restricted isometry property with Remark 2.4 in mind. Finally, the lower bound follows since

$$\begin{aligned}\left\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{Z})\right\|_2 &\geq \left\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{Z}^{\#})\right\|_2 - \left\|\frac{1}{\sqrt{m}}\mathcal{A}(\Delta\mathbf{Z})\right\|_2 \\ &\geq (1-\delta/2) - B\delta/4 \geq (1-\delta/2) - (1+\delta)\delta/4 \geq 1-\delta,\end{aligned} \tag{2.33}$$

which completes the proof. □

2.2.2 Matrix Sensing in the Presence of Noise

Due to the relevance of the low-rank matrix recovery problem throughout science and applied mathematics it is inevitable to investigate the situation when the measurements are affected by noise, i.e.,

$$\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\eta} \tag{2.34}$$

for a noise vector $\boldsymbol{\eta} \in \mathbb{R}^m$ with $\|\boldsymbol{\eta}\|_2 \leq \eta$. In this setting, noise-aware nuclear norm minimization

$$\min_{\mathbf{Z}\in\mathbb{R}^{n_1\times n_2}} \|\mathbf{Z}\|_* \quad \text{subject to } \|\mathcal{A}(\mathbf{Z}) - \mathbf{y}\|_2 \leq \eta \tag{2.35}$$

stably and robustly recovers a low-rank matrix $\widehat{\mathbf{X}}$ according to a modification of Theorem 2.5. Additionally to the approximation error term, which compares the solution $\widehat{\mathbf{X}}$ to the best rank-R approximation, a measurement error of the order $\mathcal{O}(\eta)$ appears. The adjustment of the proof therefore works analogously to the vector-valued case, which was addressed in Remark 1.28. By noting that $\|\mathcal{A}(\mathbf{H})\|_2 \leq \|\mathcal{A}(\mathbf{X}) - \mathbf{y}\|_2 + \|\mathbf{y} - \mathcal{A}(\widehat{\mathbf{X}})\|_2 \leq 2\eta$ and using this in (2.17), we obtain the additional measurement error term. This was carried out in [FCRP08] as an extension to the preceding paper [RFP10].

2.3 Numerical Algorithms for Low-Rank Matrix Recovery

As we did for compressed sensing in Section 1.4, we want to provide an outline of some state-of-the-art methods for low-rank matrix recovery from incomplete and inaccurate information. For a more wide-ranging overview with several references to the literature we refer to [DR16, Section 3].
The first two subsections below are generalizations of the concepts of the preceded Subsections 1.4.2 and 1.4.3, respectively. In the former we exploit the equivalence of nuclear norm minimization and a suitable semidefinite program, whereas we describe matrix conform versions of the iterative soft and hard thresholding algorithms in the latter. In the last subsection we present the power factorization method which is an alternating minimization approach relying on a decomposition of the form $\mathbf{X} = \mathbf{U}\mathbf{V}^T$. Note that the matrices $\mathbf{U}$, $\mathbf{V}$ may not be associated with a singular value decomposition here. To some extent, this idea also lays the foundation for the numerical methods used at a later point.

2.3.1 Semidefinite Programming

In the noiseless case, nuclear norm minimization (2.9) is equivalent to the semidefinite program

$$\min_{\substack{\mathbf{Z}\in\mathbb{R}^{n_1\times n_2},\\ \mathbf{W}_i\in\mathbb{R}^{n_i\times n_i} \text{ for } i=1,2}} \frac{1}{2}(\operatorname{tr}\mathbf{W}_1 + \operatorname{tr}\mathbf{W}_2) \quad \text{subject to} \begin{pmatrix} \mathbf{W}_1 & \mathbf{Z} \\ \mathbf{Z}^T & \mathbf{W}_2 \end{pmatrix} \succeq 0,\ \mathcal{A}(\mathbf{Z}) = \mathbf{y}, \tag{2.36}$$

see, e.g., [RFP10, Section 2] for a derivation thereof. A minor adjustment makes this also amenable to noise, namely by noting that $\|\mathcal{A}(\mathbf{Z}) - \mathbf{y}\|_2 \leq \eta$ can be expressed as a linear matrix inequality, since for any $\mathbf{R} \in \mathbb{R}^{\nu_1\times\nu_2}$ it holds

$$\|\mathbf{R}\| \leq \eta \iff \lambda_{\max}(\mathbf{R}^T\mathbf{R}) \leq \eta^2 \iff \eta^2\mathbf{Id}_{\nu_2} - \mathbf{R}^T\mathbf{R} \succeq 0 \iff \begin{pmatrix} \eta\mathbf{Id}_{\nu_1} & \mathbf{R} \\ \mathbf{R}^T & \eta\mathbf{Id}_{\nu_2} \end{pmatrix} \succeq 0. \tag{2.37}$$

The last equivalence follows from a Schur complement argument, which is explained in Lemma A.5. This observation can be applied to $\mathbf{R} = \mathcal{A}(\mathbf{Z}) - \mathbf{y}$, resulting in a semidefinite program in the case of noise.

2.3.2 Iterative Soft and Hard Thresholding for Matrices

Due to the analogy of the working principle of iterative thresholding based methods, we refrain from going into detail for all methods presented in Subsection 1.4.3. Instead, we derive the matrix version of iterative soft thresholding and refer at this point to the adaptability of the argument for bridge thresholding as well as the first version of hard thresholding. For the second version, however, we sketch the iterative best rank-R approximation algorithm, which is also known as the singular value projection method.

Iterative Soft Thresholding. In the spirit of forward-backward splitting methods, the matrix-valued least absolute shrinkage and selection operator (LASSO)

$$\min_{\mathbf{Z}\in\mathbb{R}^{n_1\times n_2}} \|\mathcal{A}(\mathbf{Z}) - \mathbf{y}\|_2^2 + \beta\|\mathbf{Z}\|_*, \tag{2.38}$$

which is an unconstrained convex optimization problem, serves as a starting point for defining the update rule

$$\widehat{\mathbf{X}}^k_{\mathrm{ISTA}} = \operatorname{prox}_{t^k\beta\|\cdot\|_*}\left(\widehat{\mathbf{X}}^{k-1}_{\mathrm{ISTA}} - t^k\mathcal{A}^*\big(\mathcal{A}(\widehat{\mathbf{X}}^{k-1}_{\mathrm{ISTA}}) - \mathbf{y}\big)\right) \tag{2.39}$$

for suitable step sizes $(t^k)_{k=1}^K$. This behaves completely analogous to the vector-valued case. It only remains to investigate the proximal mapping of the nuclear norm, i.e., the optimizer of

$$\min_{\mathbf{W}\in\mathbb{R}^{n_1\times n_2}} \beta\|\mathbf{W}\|_* + \frac{1}{2}\|\mathbf{W}-\mathbf{Z}\|_F^2. \tag{2.40}$$

Noting that, by definition, $\|\mathbf{W}\|_*$ only depends on the singular values of $\mathbf{W}$ and thus the choice of $\mathbf{U}(\mathbf{W})$ and $\mathbf{V}(\mathbf{W})$ can be solely made dependent upon the minimization of $\|\mathbf{W}-\mathbf{Z}\|_F$, it is straightforward to establish a connection to the proximal mapping of the ℓ_1-norm. Therefore, we assume in the following that the matrices containing the right and left singular vectors arise from a full singular value decomposition, i.e., they are square matrices. Then, as the Frobenius norm is invariant under orthogonal transformations and furthermore the sum of the squared entries of the matrix, it is most reasonable to choose $\mathbf{U}(\mathbf{W}) = \mathbf{U}(\mathbf{Z})$ and analogously for the right singular vectors. This restricts the sum to the diagonal. Having reduced (2.40) by this means to $\min_{\boldsymbol{\sigma}(\mathbf{W})\in\mathbb{R}^{\min\{n_1,n_2\}}} \beta\|\boldsymbol{\sigma}(\mathbf{W})\|_1 + \frac{1}{2}\|\boldsymbol{\sigma}(\mathbf{W})-\boldsymbol{\sigma}(\mathbf{Z})\|_2^2$ we can utilize the knowledge of the proximal mapping of the ℓ_1-norm to obtain

$$\operatorname{prox}_{\beta\|\cdot\|_*}(\mathbf{Z}) = \mathbf{U}(\mathbf{Z})\operatorname{diag}\left(\mathbb{S}_{2\beta}(\boldsymbol{\sigma}(\mathbf{Z}))\right)\mathbf{V}(\mathbf{Z})^T. \tag{2.41}$$

Consequently, the proximal mapping of the nuclear norm is essentially singular value soft-thresholding. For performing the iterative soft thresholding algorithm in the setting of matrices, in turn, this necessitates to compute a singular value decomposition of $\widehat{\mathbf{X}}^{k-1}_{\mathrm{ISTA}} - t^k\mathcal{A}^*\big(\mathcal{A}(\widehat{\mathbf{X}}^{k-1}_{\mathrm{ISTA}}) - \mathbf{y}\big)$ at each iteration k. If $\mathcal{A}$ and $\mathcal{A}^*$ admit some kind of structure, which is typically the case in applications, however, not if $\mathcal{A}$ is a Gaussian measurement ensemble, the computation of the singular value decomposition is the most expensive computational cost, which demands $\mathcal{O}\left(n_1n_2\min\{n_1,n_2\}\right)$ floating-point operations (flops). Nevertheless, techniques from randomized linear algebra can bring improvements in many different ways but at the expenses of exactness of the decomposition [HMT11]. In general, randomized algorithms have a lower computational complexity, in the case of computing a singular value decomposition, for instance, one can achieve an improvement to $\mathcal{O}\left(n_1n_2\log(\min\{n_1,n_2\})\right)$ flops. Additionally, they are more robust than standard algorithms, accessible to parallelization and more memory efficient.

Singular Value Projection. Analogously to the iterative best s-term approximation algorithm, the idea underlying the matrix-valued analog, the iterative best rank-R approximation algorithm, is to allow only rank-R matrices throughout the iterations. This results in the update rule

$$\widehat{\mathbf{X}}^k_{\mathrm{SVP}} = \left(\widehat{\mathbf{X}}^{k-1}_{\mathrm{SVP}} - t^k\mathcal{A}^*\big(\mathcal{A}(\widehat{\mathbf{X}}^{k-1}_{\mathrm{SVP}}) - \mathbf{y}\big)\right)_{[R]} \tag{2.42}$$

for suitable step sizes $(t^k)_{k=1}^K$. This method was introduced and analyzed in [JMD10] under the name singular value projection (SVP). As for its vector-valued relative, prior knowledge, in this case of the rank, is required. Regarding the computational complexity, in order to obtain the R dominant components of the singular value decomposition, only $\mathcal{O}(n_1 n_2 R)$ flops are required when applying standard methods for the decomposition. This compares to $\mathcal{O}(n_1 n_2 \log(R))$ flops for randomized algorithms as addressed above. However, even though the singular value projection method guarantees convergence with a geometric convergence rate, it typically requires many iterations for precise solutions, which motivated to include a Newton-step in order to speed up convergence, cf. [JMD10, Subsection 2.3]. The subsequent theorem establishes recovery up to noise level for rank-R matrices under a suitable restricted isometry property.

Theorem 2.14 ([JMD10, Theorem 1.2])**.** *Let us assume that $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ has the rank-$2R$ restricted isometry property with constant $0 < \delta < 1/3$ and let $\mathbf{X}$ be a rank-R matrix. Then, for $\epsilon \geq 0$, singular value projection with constant step size $t^k = 1/(1+\delta)$ outputs a matrix $\widehat{\mathbf{X}}^K_{\mathrm{SVP}}$ of rank at most R such that $\|\mathcal{A}(\widehat{\mathbf{X}}^K_{\mathrm{SVP}}) - \mathbf{y}\|_2^2 \leq C\|\boldsymbol{\eta}\|_2^2 + \epsilon^2$ and*

$$\|\mathbf{X} - \mathbf{X}^K_{\mathrm{SVP}}\|_F^2 \leq \frac{C\|\boldsymbol{\eta}\|_2^2 + \epsilon^2}{1-\delta} \tag{2.43}$$

after at most $K = \left\lceil \frac{1}{\log(1/D)} \log\left(\frac{\|\mathbf{y}\|_2^2}{2(C\|\boldsymbol{\eta}\|_2^2 + \epsilon^2)}\right)\right\rceil$ iterations for universal constants C, D.

2.3.3 Alternating Minimization

In several applications, a bilinear decomposition $\mathbf{X} = \mathbf{U}\mathbf{V}^T$ of the rank-R matrix $\mathbf{X}$ allows for more interpretability, such as in the example of the grocery store from the introduction. In what follows, $\mathbf{U} \in \mathbb{R}^{n_1 \times R}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times R}$ do, in general, not arise from a singular value decomposition of $\mathbf{X}$. Instead, we allow them to be non-orthogonal matrices. Moreover, for $i \in [R]$, we refer to the vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ as the left and right component vectors of $\mathbf{X}$, respectively. Having prescribed the desired matrix rank directly into the two factor matrices $\mathbf{U}$ and $\mathbf{V}$, turns the rank restricted version of the rank-minimization problem (2.8), i.e.,

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}} \|\mathcal{A}(\mathbf{Z}) - \mathbf{y}\|_2 \quad \text{subject to } \operatorname{rank} \mathbf{Z} \leq R, \tag{2.44}$$

into the problem of finding the two component matrices $\mathbf{U}$ and $\mathbf{V}$ solving the optimization problem

$$\min_{\substack{\widetilde{\mathbf{U}} \in \mathbb{R}^{n_1 \times R}, \\ \widetilde{\mathbf{V}} \in \mathbb{R}^{n_2 \times R}}} \|\mathcal{A}(\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T) - \mathbf{y}\|_2. \tag{2.45}$$

Note that the equivalence of the noise-aware version of (2.8) and (2.44) follows by redoing the proof of Lemma A.4 for the combination of the rank and the Frobenius norm.

Due to the bilinear nature of the matrix factorization the minimization problem (2.45) is non-convex. However, it becomes convex as soon as one factor is fixed, which motivates to employ an alternating minimization procedure, i.e., keeping $\mathbf{U}$ or $\mathbf{V}$ fixed and minimizing over the other, before switching their roles and repeating. This provides an approach to obtain an approximate solution to (2.45) and is known under the name power factorization (PF), which was proposed in the setting of low-rank matrix recovery in [PHH09]. We summarize this procedure in Algorithm 2.

Algorithm 2 Power Factorization (PF)

```
Input: Measurement operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$, measurements $\mathbf{y} \in \mathbb{R}^m$, rank $R$ and
    number of iterations $K$.
Output: Minimizer $\widehat{\mathbf{X}}_{\mathrm{PF}}$.
1: Initialize $\widehat{\mathbf{V}}^0 \in \mathbb{R}^{n_2 \times R}$ to be the $R$ leading right singular vectors of $\mathcal{A}^*(\mathbf{y})$ and set
    $k = 0$.
2: while $k \leq K$ and stopping criterion not fulfilled
3:     Set $k = k + 1$.
4:     $\widehat{\mathbf{U}}^k = \arg\min_{\widehat{\mathbf{U}} \in \mathbb{R}^{n_1 \times R}} \left\| \mathcal{A}\big(\widehat{\mathbf{U}}(\widehat{\mathbf{V}}^{k-1})^T\big) - \mathbf{y} \right\|_2$
5:     $\widehat{\mathbf{V}}^k = \arg\min_{\widehat{\mathbf{V}} \in \mathbb{R}^{n_2 \times R}} \left\| \mathcal{A}\big(\widehat{\mathbf{U}}^k \widehat{\mathbf{V}}^T\big) - \mathbf{y} \right\|_2$
6: end while
7: Set $\widehat{\mathbf{X}}_{\mathrm{PF}} = \widehat{\mathbf{U}}^k(\widehat{\mathbf{V}}^k)^T$.
```

The two convex minimization problems at each iteration step are least-squares problems of the form

$$\min_{\hat{\mathbf{u}} \in \mathbb{R}^{n_1 R}} \left\| \mathbf{A}_{\widehat{\mathbf{V}}} \hat{\mathbf{u}} - \mathbf{y} \right\|_2 \tag{2.46}$$

in case of line 4 in Algorithm 2 and analogously for the minimization in $\widehat{\mathbf{V}}$, i.e., line 5. Here, the matrix $\mathbf{A}_{\widehat{\mathbf{V}}} \in \mathbb{R}^{m \times n_1 R}$ parametrizes the action of $\mathcal{A}$ for fixed $\widehat{\mathbf{V}}$ such that for all $\widehat{\mathbf{U}}$ it holds $\mathbf{A}_{\widehat{\mathbf{V}}} \operatorname{vec}(\widehat{\mathbf{U}}) = \mathcal{A}(\widehat{\mathbf{U}}\widehat{\mathbf{V}}^T)$. The solution to (2.46) is then given in terms of the Moore-Penrose inverse $\mathbf{A}_{\widehat{\mathbf{V}}}^{\dagger}\mathbf{y}$. This can be solved easily and efficiently using standard techniques from numerical linear algebra in $\mathcal{O}\left(mn_1n_2R + ((n_1R)^2 m + (n_1R)^3)\right)$ flops, where the terms correspond to the computation of $\mathbf{A}_{\widehat{\mathbf{V}}}$ and the application of the pseudoinverse, respectively. Note that we did not assume any special structure about the measurement operator $\mathcal{A}$, which could lower the cost of emerging matrix-matrix products. This behaves analogously when keeping $\widehat{\mathbf{U}}$ fixed.

At this point two comments on the convex subproblems as well as the overall non-convex minimization problem are in order. First, due to the non-convexity, power factorization is prone to local minimizers and relies substantially on a good initialization. And second, the relatively easy least-squares subproblems are readily amenable for exploiting additional structure by regularizing them. This may be, for instance, sparsity or non-negativity in the individual component vectors. However, before taking a closer look at matrix recovery from multiple structures—what we will do for the remainder of the thesis—we want to conclude this part with a theoretical performance analysis result of alternating minimization under the assumption of the restricted isometry property. The result is taken from [JNS13], where power factorization was analyzed in the noiseless case in the settings of matrix sensing and matrix completion. The method of proof is based on the observation that our method can bee seen as a perturbed version of the power method.

Theorem 2.15 ([JNS13, Theorem 2.2]). *Let us assume that $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ has the rank-$2R$ restricted isometry property with constant $\delta < \frac{(\sigma_1(\mathbf{X}))^2}{(\sigma_R(\mathbf{X}))^2}\frac{1}{100R}$, where $\mathbf{X}$ is a rank-R matrix. Then, for $\epsilon > 0$, power factorization in form of Algorithm 2 yields a matrix $\widehat{\mathbf{X}}_{\mathrm{PF}}^K = \widehat{\mathbf{U}}^K\big(\widehat{\mathbf{V}}^K\big)^T$ satisfying*

$$\left\| \mathbf{X} - \widehat{\mathbf{X}}_{\mathrm{PF}}^K \right\|_F \leq \epsilon \tag{2.47}$$

after at most $K = \lceil 2\log(\|\mathbf{X}\|_F/\epsilon) \rceil$ iterations.

Chapter 3

Matrix Sensing from Multiple Structures

After having investigated how to recover low-rank matrices from few linear measurements in the preceding chapter, we consider the problem of recovering matrices with multiple structures in this chapter. Starting with an outline on how simultaneously structured models emerge, we raise the question of how to benefit from these multiple structures. In this context we address a fundamental limitation of convex approaches based on multi-objective optimization. As a consequence thereof we direct our attention to non-convex recovery methods.

For a detailed analysis of matrix recovery from multiple structures we recommend the paper [OJF+15] by Oymak, Jalali, Fazel, Eldar and Hassibi as well as [MHWG14] by Mu, Huang, Wright and Goldfarb.

3.1 Sparse Principal Component Analysis

We want to elucidate the relevance of simultaneously structured models and their recovery from few linear measurements by the example of the grocery store which was already sketched in the introduction. Below, however, we do this more detailed and rigorous. Therefore, following [FMN19, Section 1], let us consider a grocery store with n_1 customers and n_2 products. We denote the matrix containing the probabilities x_{ij} that customer i purchases product j by $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$. Under the reasonable assumption that a customer's purchase behavior is only influenced by $R \leq \min\{n_1, n_2\}$ basis factors for a comparably small R, we can assign two vectors $\mathbf{u}_r \in \mathbb{R}^{n_1}$ and $\mathbf{v}_r \in \mathbb{R}^{n_2}$ to each basis factor in the following way. The ith entry $(\mathbf{u}_r)_i$ of the left component vector $\mathbf{u}_r$ encodes to what extent the rth basis factor affects customer i. In turn, the jth component $(\mathbf{v}_r)_j$ of the right component vector $\mathbf{v}_r$ describes the probability that product j is bought under the assumption of being affected by the rth basis factor. This motivates the non-orthogonal low-rank factorization

$$\mathbf{X} = \mathbf{U}\mathbf{V}^T = \sum_{r=1}^{R} \mathbf{u}_r \mathbf{v}_r^T, \tag{3.1}$$

where the two matrices $\mathbf{U} \in \mathbb{R}^{n_1 \times R}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times R}$ contain the R left and right component vectors $\mathbf{u}_r$ and $\mathbf{v}_r$, respectively. We assume each set to be linearly independent, yet, we do not assume the vectors to be mutually orthogonal.

This distinguishes the decomposition (3.1) from the widely used principal component analysis (PCA) [Jol02], which was introduced over a century ago in [Pea01]. The idea thereof is to find a new basis for the data set, which represents the data better and moreover results from a linear transformation of the original basis. These new variables, the so-called principal components, shall capture the maximal variance successively in the sense that the first principal component describes the best-fitting line and is complemented to the best-fitting plane by the second one and so forth. This is based on the thought that a large variance represents the interesting dynamics, whereas small variances are associated with noise. Returning to the change of basis, we note that, in general, there are many old variables, which are additionally correlated. In contrast, the new variables are uncorrelated and it is typically possible to retain most of the variance and consequently the relevant structure of the data set by taking into account only the first few principal components. This makes principal component analysis appealing for dimension reduction techniques. It is moreover intimately related to the singular value decomposition. Let us assume that $\mathbf{X}$ has column mean zero and let $\mathbf{X} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^T$ denote a singular value decomposition of $\mathbf{X}$. Then, the columns of $\widetilde{\mathbf{U}}$ form the new orthonormal basis of the data with corresponding variances $\tilde{\sigma}_{rr}^2$ for the rth principal component. The column $\tilde{\mathbf{v}}_r$ of the matrix $\widetilde{\mathbf{V}}$ is the loading of the corresponding rth principal component.

Despite assuring minimal information loss when representing the data insinuating a lower-dimensional structure and guaranteeing that the principal components are uncorrelated, there is one severe drawback when it comes to interpretability of the principal components. In general, each principal component is a combination of all original variables and the corresponding loading is a dense vector. This makes it typically really hard to assign a real-world interpretation to these decisive directions. Consequently, in order to ease interpretability it is desirable to reduce the number of explicitly involved old variables in the principal components, i.e., we request sparse loadings of the principal components.

Sparse principal component analysis (sparse PCA) was proposed in [ZHT06] by Zou, Hastie and Tibshirani as a remedy and an extension to regular principal component analysis especially for high dimensions. It gives up orthogonality of the principal components in order to promote sparsity of the respective loadings. The underlying idea is to use the fact that principal component analysis can be formulated as a regression-type optimization problem [ZHT06, Theorem 3]. Then, in order to achieve the desired sparsity, the ridge penalty can be replaced or supplemented by an ℓ_1-penalization term, which is also known as LASSO penalty. The resulting regression model is known as the LASSO [Tib96], cf. (1.67), or as the elastic net [ZH05], respectively.

Let us now come back to the example of the grocery store from the beginning. Since a specific basic factor entails a very specific and pronounced buying pattern, it is reasonable to assume that the right component vectors $\mathbf{v}_r$ are sparse for all $r \in [R]$. As they can be moreover interpreted as discrete probability vectors, it is also meaningful to assume even further structure such as positivity, i.e., $(\mathbf{v}_r)_j \geq 0$ for all $j \in [n_2]$ and for all $r \in [R]$ as well as $\sum_{j=1}^{n_2} (\mathbf{v}_r)_j = 1$ for all $r \in [R]$. The latter can be easily ensured by a proper rescaling of the two component vectors.

In most applications we only have partial access to certain entries of $\mathbf{X}$ or even just indirect information in terms of linear measurements $\mathbf{y} = \mathcal{A}(\mathbf{X})$. In the following we assume that there is no data from personalized fidelity cards available, i.e., the store is not able to associate certain purchases with certain customers. That means we rely solely on information obtained from measurements and thus consider the matrix sensing case.

Let us sketch hereinafter, as proposed in [FMN19, Section 1], how to learn $\mathbf{X}$ from aggregated revenues utilizing small random price fluctuations. Therefore, let us consider

m periods of constant prices with D days each. That means we track the daily sales over mD days in total. For each price period $\ell \in [m]$ let us denote the vector, which encodes the prices p_j^ℓ of all products j, by $\mathbf{p}^\ell \in \mathbb{R}^{n_2}$. The total revenues $y_{\ell,d}$ of the grocery store on day d in price period ℓ can now be obtained as follows. Assuming that a random subset $T_d \subset [n_1]$ of all customers visits the store on that day d and that each customer's shopping cart can be modeled by a random subset $\mathcal{P}_{i,d} \subset [n_2]$ of all available products, we can compute

$$y_{\ell,d} = \sum_{i \in T_d} \sum_{j \in \mathcal{P}_{i,d}} p_j^\ell. \tag{3.2}$$

Recalling that the entry x_{ij} in the ith row and jth column of $\mathbf{X}$ gives the probability that customer i purchases product j, the expected revenues of the grocery store on one day d in price period ℓ are

$$\mathsf{E}\left[y_{\ell,d}\right] = \sum_{i=1}^{n_1} q_i \sum_{j=1}^{n_2} x_{ij} p_j^\ell, \tag{3.3}$$

where q_i denotes the probability that customer i visits the store. We collect these probabilities in a vector $\mathbf{q} \in \mathbb{R}^{n_1}$ and want to note that this vector could also depend on the price period ℓ, which would correspond to the case that some customers are attracted by special offers created by the price modifications, in case that they are communicated. Moreover, we notice that the quantity $\mathsf{E}\left[y_{\ell,d}\right]$ is constant over a fixed period ℓ, i.e., independent of d. Let us therefore denote the over D days averaged takings in price period ℓ by $y_\ell = \frac{1}{D}\sum_{d=1}^{D} y_{\ell,d}$. By the law of large numbers we observe that $\lim_{D\to\infty} y_\ell = \mathsf{E}\left[y_{\ell,d}\right]$ in probability and almost surely. According to the central limit theorem we obtain $y_\ell = \mathsf{E}\left[y_{\ell,d}\right] + \eta_{\ell,D}$, where $\eta_{\ell,D}$ denotes suitable Gaussian noise for D large enough. This can be now rewritten as

$$y_\ell = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(q_i p_j^\ell\right) x_{ij} + \eta_{\ell,D} = \langle \mathbf{A}_\ell, \mathbf{X} \rangle_F + \eta_{\ell,D}, \tag{3.4}$$

where the matrices $\mathbf{A}_\ell \in \mathbb{R}^{n_1 \times n_2}$ are the outer products $\mathbf{q}(\mathbf{p}^\ell)^T$ for $\ell \in [m]$, i.e., for their entries it holds $(a_\ell)_{ij} = q_i p_j^\ell$ for all $i \in [n_1]$ and $j \in [n_2]$. By using the random price perturbations in each period $\ell \in [m]$, we obtain m individual inaccurate measurements y_ℓ of the data matrix $\mathbf{X}$. We can rearrange them in the form

$$\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\eta}, \tag{3.5}$$

where the measurement operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ collects the individual matrices $\mathbf{A}_\ell$ for $\ell \in [m]$ and $\boldsymbol{\eta}$ denotes ineliminable noise with entries $\eta_{\ell,D}$.
As it is desirable that the number of measurements m is small compared to the dimension of the ambient space $\mathbb{R}^{n_1 \times n_2}$, which is $n_1 n_2$, we want to briefly address information theoretic limitations on the required size of the measurements. Therefore, recall the sparse low-rank decomposition (3.1), which can be reformulated as

$$\mathbf{X} = \sum_{r=1}^{R} \sigma_r \frac{\mathbf{u}_r}{\|\mathbf{u}_r\|_2} \left(\frac{\mathbf{v}_r}{\|\mathbf{v}_r\|_2}\right)^T \tag{3.6}$$

for $\sigma_r = \|\mathbf{u}_r\|_2 \|\mathbf{v}_r\|_2$. Let us now assume that the left component vectors $\mathbf{u}^r$ are s_1-sparse and that the right component vectors $\mathbf{v}^r$ are s_2-sparse, i.e., $\mathbf{u}^r \in \Sigma_{s_1}^{n_1}$ and $\mathbf{v}^r \in \Sigma_{s_2}^{n_2}$ for all $r \in [R]$. By counting the number of degrees of freedom in (3.6), which turns out to equal $\sum_{r=1}^{R}(1 + (s_1 - 1) + (s_2 - 1)) = R(s_1 + s_2 - 1)$, we conclude that a lower bound of $m \geq R(s_1 + s_2 - 1)$ on the number of measurements is an absolutely necessary condition to recover a rank-R matrix $\mathbf{X}$ with (s_1, s_2)-sparse non-orthogonal rank-1 decomposition. In what follows, we denote the set of such (s_1, s_2)-sparse rank-R matrices by $\mathcal{S}_{s_1,s_2}^{R}$, i.e.,

$$\begin{aligned}\mathcal{S}_{s_1,s_2}^{R} = \Big\{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2} : \exists\, \mathbf{u}_1, \dots, \mathbf{u}_R \in \Sigma_{s_1}^{n_1},\ \mathbf{v}_1, \dots, \mathbf{v}_R \in \Sigma_{s_2}^{n_2},\ \text{and } \boldsymbol{\sigma} \in \mathbb{R}^R,\\ \text{s.t. } \mathbf{Z} = \sum_{r=1}^{R} \sigma_r \mathbf{u}_r \mathbf{v}_r^T,\\ \text{where } \|\mathbf{u}_r\|_2 = \|\mathbf{v}_r\|_2 = 1\ \forall r \in [R]\Big\}.\end{aligned} \tag{3.7}$$

This matrix model was utilized and analyzed in [FMN19, Subsection 3.2]. For a matrix $\mathbf{X} \in \mathcal{S}_{s_1,s_2}^{R}$, a decomposition of the form $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ as in (3.7), where $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\sigma})$, is called sparse decomposition (SD) of $\mathbf{X}$. This factorization may not be confused with the singular value decomposition, which is in general different from a sparse decomposition. Moreover, the sparse decomposition is not necessarily unique and does not admit orthogonal left and right component vectors, cf. [ROV14, Proposition 6].

Let us now compare the previously determined number of degrees of freedom and the related bound on the number of required measurements to the situations where we would exploit only one of the two structures. We start with sparsity solely. To this end note that an (s_1, s_2)-sparse rank-R matrix has at most Rs_1s_2 non-zero entries. Thus, by regarding the matrix as a long vector with columns stacked on top of each other, we deduce that we require $m \gtrsim Rs_1s_2$ measurements for recovery according to Theorem 1.2. An additional multiplicative logarithmic term $\log\left(\frac{en_1n_2}{Rs_1s_2}\right)$ ensures stability, cf. Corollary 1.13 in combination with Theorem 1.11. Vice versa, considering low-rankness regardless of sparsity, according to Theorem 2.8 combined with Theorem 2.11, any $(n_1 \times n_2)$-dimensional matrix of rank R can be determined with $m \gtrsim R(n_1 + n_2)$ measurements. This raises the question whether we can go below these lower bounds when taking both structures into consideration simultaneously. We will deal with this question for the rest of the thesis beginning in the upcoming section.

Before that, however, let us mention a different well-studied situation of multiple structures, which distinguishes itself in several aspects from the one we will focus on. Therefore, consider for the moment a matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$, which is the sum of a low-rank-R component $\mathbf{L}$ and an s-sparse[7] component $\mathbf{S}$, i.e., we have the additive decomposition

$$\mathbf{M} = \mathbf{L} + \mathbf{S}. \tag{3.8}$$

Assuming that we can observe the matrix $\mathbf{M}$ directly, we aim at recovering both parts of the matrix individually and exactly. It was shown in [CLMW11] that this is indeed possible under suitable conditions, including an incoherence assumption on the left and right singular spaces of $\mathbf{L}$. Moreover, recovery can be done using a tractable convex optimization program, whose objective function is a combination of the nuclear norm $\|\mathbf{L}\|_*$ and the ℓ_1-norm $\|\operatorname{vec}\mathbf{S}\|_1$. This is called principal component pursuit (PCP). A compressed version of principal component pursuit was proposed in [WGMM13] requiring

[7] A matrix $\mathbf{S}$ is called s-sparse, if it contains at most s entries, i.e., its vectorization $\operatorname{vec}\mathbf{S}$ is s-sparse in the usual sense.

$m \gtrsim (R(n_1+n_2-R)+s)\log^2(\max\{n_1,n_2\})$ measurements, which matches the number of degrees of freedom of the additive decomposition up to the polylogarithmic term. The practical interest of this problem stems from the possibility to understand this approach as a robust principal component analysis, since it provides the opportunity to determine the principal components of $\mathbf{L}$ even if a proportion of the observed entries is significantly corrupted or missing at all, which is modeled by the matrix $\mathbf{S}$.

3.2 On Limitations of Convex Optimization

At the beginning of this section we want to take a more general point of view and consider arbitrarily simultaneously structured models. As usual, the signal to be recovered is denoted by $\mathbf{X}$. Despite formulating the statements in the following from the perspective of matrices, i.e., $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$, the results also apply to vectors and tensors by adapting them suitably. Following the problem formulation from [OJF+15] by Oymak et al., let us assume that $\mathbf{X}$ has τ low-dimensional structures $S_1, \dots, S_\tau$ at the same time, such as low-rankness or different types of sparsity. When aiming at recovering the signal, it is tempting to minimize the convex relaxations for the individual structures simultaneously or, as a weakened form, to consider a suitable linear combination of the same as a convex relaxation for the simultaneously structured object. Roughly speaking, however, it turns out that this composite optimization does, in general, not improve the number of required measurements significantly. To elaborate on this let, us denote the penalty norm promoting structure S_t by $\|\cdot\|_{(t)}$ for $t \in [\tau]$ and consider the scalarized multi-objective minimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}} f(\mathbf{Z}) = h\big(\|\mathbf{Z}\|_{(1)}, \dots, \|\mathbf{Z}\|_{(\tau)}\big) \quad \text{subject to } \mathcal{A}(\mathbf{Z}) = \mathbf{y} \tag{3.9}$$

for a scalar-valued non-negative convex and non-decreasing function h. The already addressed linear and moreover convex combination is obtained for $h(\mathbf{w}) = \sum_{t=1}^{\tau} \lambda_t w_t$ with positive scalars $\lambda_t > 0$ for all $t \in [\tau]$.
The crucial question is whether and under which conditions $\mathbf{X}$ is the unique solution to (3.9). We will discuss this issue in the following by presenting the geometrical argument from [MHWG14] by Mu et al. First of all, we notice that $\mathbf{X}$ in fact uniquely solves (3.9) if and only if $\mathcal{C}(f,\mathbf{X}) \cap \ker \mathcal{A} = \{\mathbf{0}\}$, where $\mathcal{C}(f,\mathbf{X}) = \operatorname{cone}\{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2} : f(\mathbf{X}+\mathbf{Z}) \le f(\mathbf{X})\}$ denotes the descent cone of f at $\mathbf{X}$. Since $\ker \mathcal{A}$ is a random $(n_1 n_2 - m)$-dimensional subspace, the smaller $\mathcal{C}(f,\mathbf{X})$ is, the more probable unique recoverability of $\mathbf{X}$ is. In order to quantify the size of the descent cone we will introduce the statistical dimension [ALMT13] in a moment in Definition 3.2. Before that, however, let us investigate the structure of $\mathcal{C}(f,\mathbf{X})$. Therefore, we consider its polar cone $\mathcal{C}(f,\mathbf{X})^\circ = \operatorname{cone} \partial f(\mathbf{X})$, see, e.g., [Roc70, Theorem 23.7]. Having in mind that for linear combinations of proper convex functions under certain rather weak regularity conditions, namely that the relative interiors of their domains have at least a point in common, it holds $\partial(f_1+f_2)(\mathbf{X}) = \partial f_1(\mathbf{X}) + \partial f_2(\mathbf{X})$, where the addition on the right-hand side is the Minkowski sum, we investigate the subdifferential $\partial\|\mathbf{X}\|_{(t)}$ of a single norm first. To this end let us define a measure for the alignment of a signal with respect to a set.

Definition 3.1 (Correlation)**.** *The correlation between a matrix $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$ and a set*

$S \subset \mathbb{R}^{n_1 \times n_2}$ is defined as

$$\rho(\mathbf{Z}, S) = \inf_{\mathbf{S} \in S \setminus \{\mathbf{0}\}} \frac{|\langle \mathbf{Z}, \mathbf{S} \rangle_F|}{\|\mathbf{Z}\|_F \|\mathbf{S}\|_F}. \tag{3.10}$$

We want to remark that in the previous definition there is nothing special about taking matrices. In fact, due to the relation of the Frobenius and the Euclidean scalar product, our definition can be obtained from the version for vectors by interpreting vectors as vectorizations of matrices. Now notice that $\rho(\mathbf{Z}, S) \in [0,1]$ and, geometrically speaking, $\cos^{-1}(\rho(\mathbf{Z}, S))$ gives the largest angle between $\mathbf{Z}$ and any point of S. For the subdifferential of some norm $\|\cdot\|_{(t)}$ at $\mathbf{X}$ in specific, we observe

$$\rho(\mathbf{X}, \partial\|\mathbf{X}\|_{(t)}) = \frac{\|\mathbf{X}\|_{(t)}}{\sup_{\mathbf{S} \in \partial\|\mathbf{X}\|_{(t)}} \|\mathbf{S}\|_F \|\mathbf{X}\|_F} \geq \frac{\|\mathbf{X}\|_{(t)}}{L_t \|\mathbf{X}\|_F}, \tag{3.11}$$

denoting the Lipschitz constant of $\|\cdot\|_{(t)}$ by L_t. Note that the equality follows since the subdifferential of a norm is the set of points, where Hölder's inequality is tight. More precisely, $\mathbf{S} \in \partial\|\mathbf{X}\|_{(t)}$ is equivalent to $\|\mathbf{X}\|_{(t)} = \langle \mathbf{S}, \mathbf{X} \rangle_F$ and $\|\mathbf{S}\|_{(t),*} \leq 1$, where $\|\cdot\|_{(t),*}$ denotes the dual norm of $\|\cdot\|_{(t)}$. The quantity $\kappa_t = \|\mathbf{X}\|_{(t)}/(L_t\|\mathbf{X}\|_F)$ can be regarded as a measure for the model complexity, which can be exemplified with sparse vectors and the ℓ_1-norm. For an s-sparse vector $\mathbf{x} \in \mathbb{R}^N$ with entries of the same magnitude it holds $\kappa = \|\mathbf{x}\|_1/(L_{\|\cdot\|_1}\|\mathbf{x}\|_2) = \sqrt{s/N}$. For further examples see, e.g., [OJF+15, Table 2]. Moreover, we can use κ_t to define a with $\partial\|\cdot\|_{(t)}$ associated angle $\theta_t = \cos^{-1}(\kappa_t)$ with which it holds

$$\partial\|\mathbf{X}\|_{(t)} \subset \operatorname{circ}(\mathbf{X}, \theta_t) = \{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2} : \langle \mathbf{Z}, \mathbf{X} \rangle \leq \cos(\theta_t)\}. \tag{3.12}$$

Here, $\operatorname{circ}(\mathbf{X}, \theta_t)$ denotes the circular cone with axis $\mathbf{X}$ and angle θ_t, which is defined as shown. The resulting situation is visualized in Figure 3.1.

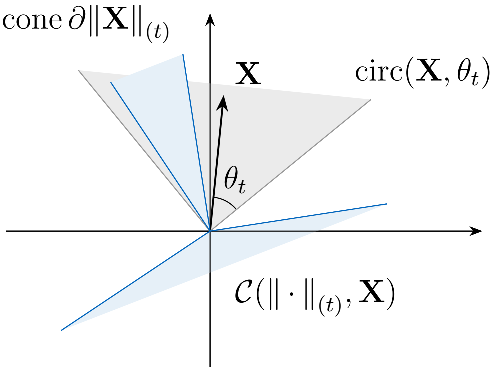

cone $\partial\|\mathbf{X}\|_{(t)}$
$\mathbf{X}$
$\operatorname{circ}(\mathbf{X}, \theta_t)$
θ_t
$\mathcal{C}(\|\cdot\|_{(t)}, \mathbf{X})$

Figure 3.1. Descent cone $\mathcal{C}(\|\cdot\|_{(t)}, \mathbf{X})$ of some structure promoting norm $\|\cdot\|_{(t)}$ at $\mathbf{X}$ together with its polar cone, cone $\partial\|\mathbf{X}\|_{(t)}$, which is enclosed by the circular cone, $\operatorname{circ}(\mathbf{X}, \theta_t)$, with axis $\mathbf{X}$ and angle $\theta_t = \cos^{-1}(\|\mathbf{X}\|_{(t)}/(L_t\|\mathbf{X}\|_F))$, cf. [MHWG14, Figure 1].

Let us now turn back to multiple regularizing norms and consider a composite convex regularizer $f(\mathbf{Z})$ of the form $\sum_{t=1}^{\tau} \lambda_t \|\mathbf{Z}\|_{(t)}$ with $\lambda_t > 0$ for all $t \in [\tau]$. To any regularizer $\|\cdot\|_{(t)}$ we associate an angle $\theta_t = \cos^{-1}(\kappa_t) = \cos^{-1}(\|\mathbf{X}\|_{(t)}/(L_t\|\mathbf{X}\|_F))$ and a corresponding circular cone $\operatorname{circ}(\mathbf{X}, \theta_t)$. However, as the axis thereof is fixed, we conclude

$$\partial f(\mathbf{X}) = \sum_{t=1}^{\tau} \lambda_t \partial\|\mathbf{X}\|_{(t)} \subset \sum_{t=1}^{\tau} \operatorname{circ}(\mathbf{X}, \theta_t) \subset \operatorname{circ}(\mathbf{X}, \max_{t \in [\tau]} \theta_t) \tag{3.13}$$

and consequently observe that the fact that several structures were involved disappears on the right-hand side, which is only affected by the largest angle $\max_{t\in[\tau]}\theta_t$. Broadly speaking, the largest angle also corresponds to the strongest or best of the structures. By reconsidering sparse vectors, this can be illustrated. We note that the more significant the sparsity structure of $\mathbf{x}$ is, the smaller is κ and thus the larger is the resulting $\theta=\cos^{-1}(\kappa)$. Moreover, and leading back to the required number of measurements, a small angle $\max_{t\in[\tau]}\theta_t$ enforces narrow polar cones $\mathcal{C}(\|\cdot\|_{(t)},\mathbf{X})^\circ = \operatorname{cone}\partial\|\mathbf{X}\|_{(t)} \subset \operatorname{circ}(\mathbf{X},\max_{t\in[\tau]}\theta_t)$ and consequently large descent cones $\mathcal{C}(\|\cdot\|_{(t)},\mathbf{X})$. These, in turn, increase the probability that a randomly orientated subspace $\ker\mathcal{A}$ intersects one of these cones. To make this more rigorous let us define the statistical dimension of a convex cone $\mathcal{C}$, which was introduced in [ALMT13] for the vector case.

Definition 3.2 (Statistical Dimension)**.** *Let $\mathbf{Z}\in\mathbb{R}^{n_1\times n_2}$ be a random matrix with i.i.d. mean-zero Gaussian entries of unit variance. The statistical dimension of a closed convex cone $\mathcal{C}\subset\mathbb{R}^{n_1\times n_2}$ is defined as*

$$\delta(\mathcal{C}) = \mathsf{E}_{\mathbf{Z}}\left[\|P_{\mathcal{C}}(\mathbf{Z})\|_F^2\right], \tag{3.14}$$

where $P_{\mathcal{C}}$ denotes the projection onto the cone, i.e., $P_{\mathcal{C}}(\mathbf{Z}) = \arg\min_{\widetilde{\mathbf{Z}}\in\mathcal{C}}\|\widetilde{\mathbf{Z}}-\mathbf{Z}\|_F$.

The former definition can be easily traced back to the vector version by considering the vectorizations of the respective quantities. The statistical dimension extends the concept of a dimension from subspaces to convex cones. In particular, it is non-negative, bounded by the ambient dimension and monotonous in the sense that $\delta(\mathcal{C}_1)\le\delta(\mathcal{C}_2)$ if $\mathcal{C}_1\subset\mathcal{C}_2$ for two closed convex cones $\mathcal{C}_1$ and $\mathcal{C}_2$. Furthermore, it fulfills a complementarity condition, namely $\delta(\mathcal{C})+\delta(\mathcal{C}^\circ)=n$, when n is the ambient dimension. For a linear subspace $\mathcal{L}$ it moreover holds $\delta(\mathcal{L})=\dim(\mathcal{L})$. To give an example let us consider the nonnegative orthant in n dimensions, i.e., $\mathbb{R}_+^n$. As it is a self-dual cone the complementarity property yields $\delta(\mathbb{R}_+^n)=\frac{1}{2}n$. A closely related quantity, see, e.g., [Ver15, Definition 3.4], is the Gaussian mean width, which aims at capturing the complexity of an arbitrary bounded set by averaging over widths induced by intersections with randomly oriented one-dimensional subspaces.

After this small excursion on the statistical dimension of a cone, let us now return to the question about the probability that the random (n_1n_2-m)-dimensional subspace $\ker\mathcal{A}$ has a non-trivial intersection with a fixed convex cone $\mathcal{C}\subset\mathbb{R}^{n_1\times n_2}$. In [ALMT13] this question was explored extensively with the result that there occurs a sharp phase transition, which is determined by the statistical dimension of the cone $\mathcal{C}$. If the codimension m of the randomly oriented subspace $\ker\mathcal{A}$ is larger than the statistical dimension $\delta(\mathcal{C})$, the probability of sharing a non-trivial intersection with $\mathcal{C}$ is small. In turn, if m is smaller than $\delta(\mathcal{C})$, it is very probable that $\mathcal{C}\cap\ker\mathcal{A}\neq\{\mathbf{0}\}$. As we are primarily interested in the latter case, we formalize this statement in the subsequent lemma.

Lemma 3.3 (Corollary of [ALMT13, Theorem 7.1])**.** *Let $\mathcal{A}:\mathbb{R}^{n_1\times n_2}\to\mathbb{R}^m$ be a Gaussian measurement ensemble and $\mathcal{C}$ a convex cone. Then, if $m\le\delta(\mathcal{C})$, it holds*

$$\mathsf{P}\left(\mathcal{C}\cap\ker\mathcal{A}=\{\mathbf{0}\}\right)\le 4\exp\left(-\frac{(\delta(\mathcal{C})-m)^2}{16\delta(\mathcal{C})}\right). \tag{3.15}$$

Sketch of Proof. The statement follows from the first part of Theorem 7.1 in [ALMT13] by considering $\lambda=\delta(\mathcal{C})-m$, cf. [MHWG14, Corollary 4]. □

In order to apply this result, let us investigate the statistical dimension of the descent cone of a composite regularizer of the form $\sum_{t=1}^{\tau} \lambda_t \|\cdot\|_{(t)}$ with positive scalars λ_t, i.e., the linear combination of the individual structure promoting norms.

Lemma 3.4 (Collorary of [MHWG14, Lemma 2]). *For the statistical dimension of the descent cone $\mathcal{C}(\sum_{t=1}^{\tau} \lambda_t \|\cdot\|_{(t)}, \mathbf{X})$ with positive scalars λ_t it holds*

$$\delta\Big(\mathcal{C}\Big(\sum_{t=1}^{\tau} \lambda_t \|\cdot\|_{(t)}, \mathbf{X}\Big)\Big) \geq n_1 n_2 \kappa^2 - 2, \tag{3.16}$$

where $\kappa = \min_{t\in[\tau]} \kappa_t$.

Proof. By utilizing the complementarity condition of the statistical dimension in the first line and the monotonicity property in combination with the inclusion in (3.13) in the second we observe

$$\begin{aligned}\delta\Big(\mathcal{C}\Big(\sum_{t=1}^{\tau} \lambda_t \|\cdot\|_{(t)}, \mathbf{X}\Big)\Big) &= n_1 n_2 - \delta\Big(\mathcal{C}\Big(\sum_{t=1}^{\tau} \lambda_t \|\cdot\|_{(t)}, \mathbf{X}\Big)^{\circ}\Big) = n_1 n_2 - \operatorname{cone} \partial\Big(\sum_{t=1}^{\tau} \lambda_t \|\mathbf{X}\|_{(t)}\Big) \\ &\geq n_1 n_2 - \operatorname{circ}(\mathbf{X}, \max_{t\in[\tau]} \theta_t) \geq n_1 n_2 \big(1 - \sin^2(\max_{t\in[\tau]} \theta_t)\big) - 2 \\ &= n_1 n_2 \cos^2(\max_{t\in[\tau]} \theta_t) - 2 = n_1 n_2 \min_{t\in[\tau]} \kappa_t^2 - 2.\end{aligned} \tag{3.17}$$

The second inequality in the second line follows from Lemma 2 in [MHWG14], which is an improvement to [ALMT13, Proposition 3.4]. □

By combining these two results and recalling that κ_t^2 can be interpreted as the relative intrinsic dimension associated with structure S_t, we are now able to formally state the negative result first discovered by Oymak et al. Namely that composite optimization using a linear combination of individual structure promoting norms can do no better in the sense of improving the required number of measurements than exploiting only the best of these structures alone.

Theorem 3.5 ([MHWG14, Theorem 5]). *Let $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ be a Gaussian measurement ensemble and let $\mathbf{X} \neq \mathbf{0}$. Suppose that for each $t \in [\tau]$ the norm $\|\cdot\|_{(t)}$ is Lipschitz continuous with constant L_t. Moreover, let us denote $\kappa = \min_{t\in[\tau]} \kappa_t$, where $\kappa_t = \|\mathbf{X}\|_{(t)}/(L_t \|\mathbf{X}\|_F)$. Then, if $m \leq n_1 n_2 \kappa^2 - 2$, it holds that*

$$\mathsf{P}\left(\mathbf{X} \text{ uniquely solves } (3.9)\right) \leq 4 \exp\left(-\frac{(n_1 n_2 \kappa^2 - 2 - m)^2}{16(n_1 n_2 \kappa^2 - 2)}\right), \tag{3.18}$$

where the function f in (3.9) is of the form $f = \sum_{t=1}^{\tau} \lambda_t \|\cdot\|_{(t)}$ with positive scalars λ_t.

Proof. First recall that

$$\mathsf{P}\left(\mathbf{X} \text{ uniquely solves } (3.9)\right) = \mathsf{P}\left(\mathcal{C}(f, \mathbf{X}) \cap \ker \mathcal{A} = \{\mathbf{0}\}\right). \tag{3.19}$$

Then, the statement follows directly from Lemma 3.3 by applying it to the setting where $\mathcal{C}$ is the descent cone of f at $\mathbf{X}$ and using the lower bound on the statistical dimension of from Lemma 3.4. □

With this statement we highlighted one side of the coin concerning the sharp phase transition in recoverability regarding the required number of measurements, which is determined by a geometric invariant, the statistical dimension. Vice versa, it can be also shown using similar tools that if m is sufficiently large compared to the statistical dimension, recovery is certain with exceeding probability. However, as this limit is dictated by a single structure only, this is highly unsatisfactory for our purposes as it does not allow to benefit from multiple structures simultaneously.
As a consequence, we are obliged to go beyond tractable convex optimization based approaches in order to enhance the required number of measurements. This is the responsibility of the following section.
Before coming to that, a short annotation regarding the recovery of low-rank tensors, which play an increasingly important role in the training of neural networks, shall be provided, as this issue was the motivation for the paper [MHWG14]. The previously described negative result, namely, applies to their recovery. In the numerical algebra of tensors it is a well-known fact that already the computation of the rank of a tensor, which is the number of rank-1 tensor products in a canonical polyadic decomposition and therefore also called CP rank, is NP-hard. Consequently, concepts like the multilinear rank of an oder-d tensor, also known as the Tucker rank, were introduced and investigated. It is a d-dimensional vector containing the ranks of all d distinct unfoldings of the tensor and can be revealed by the higher order singular value decomposition (HOSVD), a particular type of the Tucker decomposition. Tensors of low Tucker rank can now be seen as signals which have multiple structures simultaneously by being of low rank along each mode.

3.3 The Power and Perils of Non-Convex Recovery

A significantly improved performance achieving nearly order-optimal recovery guarantees in the case of simultaneously structured models requires non-convex methods, as we saw in the preceding section. In order to demonstrate the superiority of non-convex regularization we sketch results from [OJF$^+$15] on simultaneously sparse and low-rank matrices. Therefore, let us introduce a slightly modified and more special matrix model compared to the set $\mathcal{S}_{s_1,s_2}^R$ of (s_1, s_2)-sparse rank-R matrices from (3.7). $\widetilde{\mathcal{S}}_{s_1,s_2}^R$ denotes this set of $(n_1 \times n_2)$-dimensional low-rank-R matrices, which are zero outside an $(s_1 \times s_2)$-dimensional submatrix, i.e.,

$$\begin{aligned}\widetilde{\mathcal{S}}_{s_1,s_2}^R = \Big\{\mathbf{Z} = \sum_{r=1}^{R} \sigma_r \mathbf{u}_r \mathbf{v}_r^T \in \mathcal{S}_{s_1,s_2}^R :\ \operatorname{supp}(\mathbf{u}_r) = \operatorname{supp}(\mathbf{u}_1),\\ \text{and } \operatorname{supp}(\mathbf{v}_r) = \operatorname{supp}(\mathbf{v}_1)\ \forall r \in [R]\Big\}.\end{aligned} \tag{3.20}$$

This set is also called the set of (s_1, s_2)-jointly-sparse rank-R matrices. The authors of the addressed paper investigated recovery methods based on convex as well as non-convex programs and exposed a gap in their performance regarding the required number of measurements, which we will outline below. In order to do so, we need to introduce one further matrix norm as well as its non-convex counterpart. For a matrix $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$, $\|\mathbf{Z}\|_{1,2}$ denotes the ℓ_1-norm of the ℓ_2-norms of the columns of $\mathbf{Z}$, whereas $\|\mathbf{Z}\|_{0,2}$ is the number of non-zero columns of $\mathbf{Z}$.

Theorem 3.6 ([OJF⁺15, Theorem 3]). *Let $\mathcal{A}: \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ be a Gaussian measurement ensemble and consider recovering $\mathbf{X} \in \widetilde{\mathcal{S}}^R_{s_1,s_2}$ via the program*

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}} f(\mathbf{Z}) \quad \textit{subject to } \mathcal{A}(\mathbf{Z}) = \mathbf{y}. \tag{3.21}$$

Then, for some constants $c_1, c_2 > 0$, the following hold.

(i) If $f(\mathbf{Z}) = \beta_1 \|\mathbf{Z}\|_{1,2} + \beta_2 \|\mathbf{Z}^T\|_{1,2} + \|\mathbf{Z}\|_*$ for regularization parameters $\beta_1, \beta_2 > 0$, the convex optimization program (3.21) will fail to recover $\mathbf{X}$ with probability at least $1 - \exp(-c_1 m_0)$, whenever $m \leq c_2 m_0$, where $m_0 = \min\{n_1 s_2, n_2 s_1, R(n_1 + n_2)\}$.

(ii) If $f(\mathbf{Z}) = \frac{1}{s_2}\|\mathbf{Z}\|_{0,2} + \frac{1}{s_1}\|\mathbf{Z}^T\|_{0,2} + \frac{1}{R}\operatorname{rank}(\mathbf{Z})$, the non-convex optimization program (3.21) will uniquely recover $\mathbf{X}$ with probability at least $1 - \exp(-c_1 m)$, whenever $m \geq c_2 \max\{s_2 \log(en_2/s_2), s_1 \log(en_1/s_1), R(s_1 + s_2)\}$.

Sketch of Proof. Firstly, (i) follows using similar arguments as we exploited to prove Theorem 3.5. A comparable result of the latter in the framework of this theorem can be found in [OJF⁺15, Theorem 2].
Secondly, for (ii), note that f obeys the triangle inequality and it moreover holds $f(\mathbf{X}) = 3$. Therefore, if we have $f(\mathbf{H}) > 6$ for all $\mathbf{H} \in \ker \mathcal{A}$, unique recoverability follows from $f(\mathbf{Z}) \geq f(\mathbf{Z} - \mathbf{X}) - f(-\mathbf{X}) > 3$ for all feasible $\mathbf{Z}$. In order to show the former we observe that $f(\widetilde{\mathbf{H}}) \leq 6$ implies $\widetilde{\mathbf{H}} \in \widetilde{\mathcal{S}}^{6R}_{6s_1,6s_2}$ and thus that $\widetilde{\mathcal{S}}^{6R}_{6s_1,6s_2} \cap \ker \mathcal{A} = \{\mathbf{0}\}$ suffices. This, in turn, can be shown by utilizing that a random linear subspace $\ker \mathcal{A}$ of codimension m has trivial intersection with the cone $\widetilde{\mathcal{S}}^{6R}_{6s_1,6s_2}$ with high probability whenever m is sufficiently large, cf. Lemma 3.3. This requires the computation of the statistical dimension of the respective cone. An alternative argument based on covering numbers was given in [OJF⁺15, Lemma 14 and Lemma 15]. □

Having a closer look at the lower bound on the number of necessary measurements for non-convex recovery, i.e., Theorem 3.6(ii), reveals that, up to logarithmic factors, reliable recovery is possible with high probability from as many measurements as degrees of freedom, which is $R(s_1 + s_2 - R)$. In order to derive the latter number, note that it suffices to consider a singular value decomposition of the inscribed $(s_1 \times s_2)$-dimensional rank-R matrix, from which one can directly count the $R + \sum_{r=1}^{R} (s_1 - r) + (s_2 - r)$ degrees of freedom. This significantly enhances the performance of the respective convexification, which itself results in an orderwise suboptimal sample size, cf. Theorem 3.6(i) and the discussion in Section 3.2.
Unfortunately, though, non-convex optimization problems are considered to be computationally intractable in general. Due to their non-convexity they are susceptible to spurious local minima making global convergence guarantees impractical. Moreover, initialization typically plays a decisive role leading to the whole performance of the algorithm being closely related to and heavily relying on a good initialization of the method. Characteristically for non-convex approaches, finding suitable initializations is an open problem in most instances and if at all only heuristics without any provable convergence guarantees are available.
In recent years, however, substantial progress was made in the fields of non-convex and non-smooth optimization, mainly driven by the undeniable success of signal processing and machine learning [JK17]. This is due to the immense diversity of modeling possibilities

non-convex formulations provide compared to merely convex problems. Nowadays, non-convex optimization is not considered to be such a delicate endeavor any more as a couple of years ago. A crucial observation in this context was that in typical applications present structure of the problem can be used in favor of the non-convex numerical method. This helps to make such approaches more tractable and lower the aversion towards them. Despite that, though, the addressed difficulties do not disappear out of nothing and have to be kept in mind.

Understanding both, opportunities and threats, non-convex formulations offer and gaining a much deeper insight is currently a vast area of very active research with several fields of mathematics contributing.

In fact, we already addressed two very relevant and also well-understood non-convex techniques, namely projected gradient descent, which is a subcase of the forward-backward splitting methods, introduced in Subsection 1.4.3, and alternating minimization, presented in Subsection 2.3.3. The latter also lays the foundation for a state-of-the-art non-convex algorithm for compressed sensing of sparse and low-rank matrices, which is described in the first part of the subsequent section.

3.4 Numerical Algorithms for the Recovery of Simultaneously Sparse and Low-Rank Matrices

Following on from the preceding Sections 1.4 and 2.3 on numerical algorithms for the recovery of sparse vectors and low-rank matrices, respectively, in this section we describe two non-convex approaches for the compressed recovery of simultaneously sparse and low-rank matrices.

The first subsection below addresses a sparse version of the alternating minimization algorithm introduced in Subsection 2.3.3. The so-called sparse power factorization additionally imposes a sparsity structure on the matrices $\mathbf{U}$ and $\mathbf{V}$ from the decomposition $\mathbf{X} = \mathbf{U}\mathbf{V}^T$ using hard thresholding pursuit. This method can be considered as a state-of-the-art benchmark for competing algorithms. After that, in the second subsection, we present a numerical algorithm minimizing a multi-penalty functional, which encompasses data fidelity as well as low-rankness of the matrix and sparsity of its non-orthogonal decomposition. This approach was dubbed Alternating Tikhonov regularization and LASSO. It is an iterative alternating minimization approach that alternates, however, directly on vector pairs instead of matrices. Sparsity in the component vectors is promoted by employing iterative soft thresholding.

3.4.1 Sparse Power Factorization

As already brought up towards the end of Subsection 2.3.3 on the power factorization method, it is possible to impose further structure on the matrices $\mathbf{U}$ and $\mathbf{V}$ from the bilinear decomposition $\mathbf{X} = \mathbf{U}\mathbf{V}^T$. This can be done by modifying their update rules, which are simple least-squares problems in the first place. In [LWB18] the sparse power factorization (SPF) was proposed and investigated by Lee, Wu and Bresler as an alternating minimization algorithm for the compressed recovery of (s_1, s_2)-jointly-sparse rank-R matrices $\mathbf{X} \in \widetilde{\mathcal{S}}^R_{s_1,s_2}$. The row-sparsity of the factor matrices $\mathbf{U}$ and $\mathbf{V}$ is imposed by means of Algorithm 4, which is a matrix-conform version of iterative best s-term approximation, cf. Paragraph 1.4.3(3). However, several other methods for compressed sensing of sparse vectors could be employed as well, such as, e.g., iterative soft thresholding, cf. Paragraph 1.4.3(1). Sparse Power Factorization is given below in Algorithm 3 in form of the subspace-concatenated version proposed in the cited literature.

Algorithm 3 Sparse Power Factorization (SPF)

```
Input: Measurement operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$, measurements $\mathbf{y} \in \mathbb{R}^m$, rank $R$, sparsities $s_1, s_2$ and number of iterations $K$ and $L$, where $K$ relates to the power factorization iterations and $L$ to the iterations of the iterative best $s$-term approximation.
Output: Minimizer $\widehat{\mathbf{X}}_{\mathrm{SPF}}$.
 1: Initialize $\widehat{\mathbf{U}}^0 \in \mathbb{R}^{n_1 \times R}$ and $\widehat{\mathbf{V}}^0 \in \mathbb{R}^{n_2 \times R}$ sufficiently, see, e.g., [LWB18, Algorithms 6 and 7] and set $k = 0$.
 2: while $k \leq K$ and stopping criterion not fulfilled
 3:     Set $k = k + 1$.
 4:     Compute $\widehat{\mathbf{V}}^{k-1} = \operatorname{orth}\big((\widehat{\mathbf{V}}^{k-1}, \widehat{\mathbf{V}}^0)\big)$, where $\operatorname{orth}(\,\cdot\,)$ returns an orthonormal basis for the range $\mathcal{R}(\,\cdot\,)$ of the argument.
 5:     if $s_1 < n_1$ then
 6:         Construct $\mathcal{A}_{\widehat{\mathbf{V}}^{k-1}} : \mathbb{R}^{n_1 \times R} \to \mathbb{R}^m$ such that it parametrizes the action of $\mathcal{A}$ for fixed $\widehat{\mathbf{V}}^{k-1}$, i.e., such that for all $\widehat{\mathbf{U}}$ it holds $\mathcal{A}_{\widehat{\mathbf{V}}^{k-1}}(\widehat{\mathbf{U}}) = \mathcal{A}\big(\widehat{\mathbf{U}}(\widehat{\mathbf{V}}^{k-1})^T\big)$.
 7:         $\widetilde{\mathbf{U}}^k = \mathrm{MatrixIBA}(\mathcal{A}_{\widehat{\mathbf{V}}^{k-1}}, \mathbf{y}, s_1, L)$, as described in Algorithm 4
 8:     else
 9:         $\widetilde{\mathbf{U}}^k = \arg\min_{\widetilde{\mathbf{U}} \in \mathbb{R}^{n_1 \times R}} \big\|\mathcal{A}\big(\widetilde{\mathbf{U}}(\widehat{\mathbf{V}}^{k-1})^T\big) - \mathbf{y}\big\|_2$
10:     end if
11:     Let $\widehat{\mathbf{U}}^k$ be the best rank-$R$ approximation of $\widetilde{\mathbf{U}}^k$.
12:     Compute $\widehat{\mathbf{U}}^k = \operatorname{orth}\big((\widehat{\mathbf{U}}^k, \widehat{\mathbf{U}}^0)\big)$.
13:     if $s_2 < n_2$ then
14:         Construct $\mathcal{A}_{\widehat{\mathbf{U}}^k} : \mathbb{R}^{n_2 \times R} \to \mathbb{R}^m$ analogously to line 6.
15:         $\widetilde{\mathbf{V}}^k = \mathrm{MatrixIBA}(\mathcal{A}_{\widehat{\mathbf{U}}^k}, \mathbf{y}, s_2, L)$
16:     else
17:         $\widetilde{\mathbf{V}}^k = \arg\min_{\widetilde{\mathbf{V}} \in \mathbb{R}^{n_2 \times R}} \big\|\mathcal{A}\big(\widehat{\mathbf{U}}^k \widetilde{\mathbf{V}}^T\big) - \mathbf{y}\big\|_2$
18:     end if
19:     Let $\widehat{\mathbf{V}}^k$ be the best rank-$R$ approximation of $\widetilde{\mathbf{V}}^k$.
20: end while
21: Set $\widehat{\mathbf{X}}_{\mathrm{SPF}} = \widehat{\mathbf{U}}^k(\widehat{\mathbf{V}}^k)^T$.
```

The method was extensively tested numerically showing a significantly improved performance compared to methods based on convex programming. In fact, near-optimal performance was reported based on a variety of numerical tests. Advantageously as well, the algorithm comes with an own, albeit computationally expensive, initialization procedure. Yet, using this initialization yields a provably near-optimal performance guarantee.
The theoretical aspects of sparse power factorization shall also be the last concern of this part. In order to investigate the performance of the algorithm, the authors introduced a suitable restricted isometry property for the relevant matrix set $\widetilde{\mathcal{S}}_{s_1,s_2}^R$ of (s_1, s_2)-jointly-sparse rank-R matrices.

Definition 3.7 (Low-Rank and Jointly-Sparse Restricted Isometry Property)**.** *A linear operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ satisfies the rank-R and (s_1, s_2)-jointly-sparse restricted isom-*

Algorithm 4 Iterative Best s-Term Approx. for Row-Sparse Matrices (MatrixIBA)

```
Input: Measurement operator $\mathcal{A}_{\text{red}} : \mathbb{R}^{n \times R} \to \mathbb{R}^m$, measurements $\mathbf{y} \in \mathbb{R}^m$, row-sparsity $s$
    and number of iterations $L$.
Output: Final iterate $\mathbf{W}_{\text{MatrixIBA}}$.
Initialize $\mathbf{W}^0 = \mathbf{0} \in \mathbb{R}^{n \times R}$ and set $\ell = 0$.
while $\ell \leq L$ and stopping criterion not fulfilled
    Set $\ell = \ell + 1$.
    Compute $\widetilde{\mathbf{W}} = \mathbf{W}^{\ell-1} - \mathcal{A}_{\text{red}}^*(\mathcal{A}_{\text{red}}(\mathbf{W}^{\ell-1}) - \mathbf{y})$.
    Let $J \subset [n]$ denote the index set of the $s$ rows of $\widetilde{\mathbf{W}}$ with largest $\ell_2$-norm and
    $P_J$ the projection onto the row set.
    $\mathbf{W}^\ell = \arg\min_{\mathbf{W} \in \mathbb{R}^{n \times R} : P_J(\mathbf{W}) = \mathbf{W}} \|\mathcal{A}_{\text{red}}(\mathbf{W}) - \mathbf{y}\|_2$
end while
Set $\mathbf{W}_{\text{MatrixIBA}} = \mathbf{W}^\ell$.
```

etry property with isometry constant $0 < \delta < 1$ *if*

$$(1-\delta)\|\mathbf{Z}\|_F^2 \leq \|\mathcal{A}(\mathbf{Z})\|_2^2 \leq (1+\delta)\|\mathbf{Z}\|_F^2 \tag{3.22}$$

for all $\mathbf{Z} \in \widetilde{\mathcal{S}}_{s_1,s_2}^R$.

Comparably to the results from compressed sensing of sparse vectors and low-rank matrices, Theorems 1.11 and 2.11, respectively, it transpires that Gaussian measurement ensembles also provide, with exceeding probability, suitable operators fulfilling the rank-R and (s_1, s_2)-jointly-sparse restricted isometry property provided the number of measurements is sufficiently large. This threshold turns out to be at the information theoretical limit up to a log-factor in the inverse of the relative sparsities s_1/n_1 and s_2/n_2. More precisely, it holds the subsequent result.

Theorem 3.8 (Gaussian Measurement Ensembles have the Low-Rank and Jointly-Sparse RIP, cf. [LWB18, Theorem 2])**.** *Let* $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ *be a Gaussian measurement ensemble and assume that*

$$m \geq CR(s_1 + s_2 + 1)\log\left(\max\left\{\frac{en_1}{s_1}, \frac{en_2}{s_2}\right\}\right) \tag{3.23}$$

holds for a constant $C > 0$*, which only depends on* $0 < \delta < 1$*. Then, with probability at least* $1 - \exp(-dm)$*, where* $d > 0$ *denotes a constant, which only depends on* δ *as well, the operator* $\frac{1}{\sqrt{m}}\mathcal{A}$ *satisfies rank-*R *and* (s_1, s_2)*-jointly-sparse RIP with isometry constant* δ*.*

Sketch of Proof. Essentially, the proof resembles the one of Theorem 2.11. The major modification concerns the usage of Lemma 2.13. Instead of applying it to the $(n_1 \times n_2)$-dimensional matrices themselves, we use it for the $(s_1 \times s_2)$-dimensional submatrices. First, we note that the set of interest, $\partial\mathcal{S}^{R,1} = \{\mathbf{Z} \in \mathbb{R}^{s_1 \times s_2} : \operatorname{rank}\mathbf{Z} \leq R \text{ and } \|\mathbf{Z}\|_F = 1\}$, can be covered with precision $\delta/(4\sqrt{2})$ by a discrete set obeying

$$\left|\left(\partial\mathcal{S}^{R,1}\right)^{\#}\right| \leq \left(144\sqrt{2}/\delta\right)^{R(s_1+s_2+1)}. \tag{3.24}$$

Second, we observe that the matrix set $\widetilde{\mathcal{S}}^R_{s_1,s_2} \cap \{\mathbf{Z} \in \mathbb{R}^{n_1\times n_2} : \|\mathbf{Z}\|_F = 1\}$ is a union of $\binom{n_1}{s_1}\binom{n_2}{s_2} \leq (en_1/s_1)^{s_1}(en_2/s_2)^{s_2}$ sets of type $\partial\mathcal{S}^{R,1}$. Therefore, we deduce that a $\delta/(4\sqrt{2})$-cover of the former set fulfills

$$\begin{aligned}\left|\left(\widetilde{\mathcal{S}}^R_{s_1,s_2}\right)^{\#}\right| &\leq (en_1/s_1)^{s_1}(en_2/s_2)^{s_2}\left(144\sqrt{2}/\delta\right)^{R(s_1+s_2+1)} \\ &\leq \left(144\sqrt{2}\max\{en_1/s_1, en_2/s_2\}/\delta\right)^{R(s_1+s_2+1)},\end{aligned} \tag{3.25}$$

admitting a slight abuse of notation by omitting the restriction to matrices of unit Frobenius norm. The logarithm of the right-hand side in the proceeded equation dictates the necessary number of measurements, which follows by modifying the respective argument in the proof of Theorem 2.11 suitably. The remainder works alike. □

As we mentioned earlier, sparse power factorization comes with a provable recovery guarantee from a nearly optimal number of measurements. This is summarized in the following theorem, for whose proof we refer to the literature. It shall be pointed out that the made assumptions demand a carefully designed initialization, what is in fact crucial for the performance. Furthermore, the noise level is required to be sufficiently small.

Theorem 3.9 (Performance of Sparse Power Factorization, cf. [LWB18, Theorem 7]). *Suppose that the following assumptions hold.*

(i) $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ *denotes a singular value decomposition of* $\mathbf{X} \in \widetilde{\mathcal{S}}^R_{s_1,s_2}$.

(ii) The condition number of $\mathbf{X}$ *is no greater than* κ.

(iii) $\mathcal{A}$ *satisfies the rank-*$2R$ *and* $(3s_1, 3s_2)$*-jointly-sparse restricted isometry property with isometry constant* $\delta = 0.04/\kappa$.

(iv) $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\eta}$, *where* $\boldsymbol{\eta}$ *and* $\mathcal{A}(\mathbf{X})$ *satisfy*

$$\frac{\|\mathbf{X}\|_F}{\|\mathbf{X}\|}\frac{\|\boldsymbol{\eta}\|_2}{\|\mathcal{A}(\mathbf{X})\|_2} \leq \frac{0.04}{\kappa}. \tag{3.26}$$

(v) The initialization $(\widehat{\mathbf{U}}^0, \widehat{\mathbf{V}}^0)$ *satisfies* $\max\left\{\left\|P_{\mathcal{R}(\mathbf{U})^\perp}P_{\mathcal{R}(\widehat{\mathbf{U}}^0)}\right\|, \left\|P_{\mathcal{R}(\mathbf{V})^\perp}P_{\mathcal{R}(\widehat{\mathbf{V}}^0)}\right\|\right\} < 0.95$, *where* P *denotes the orthogonal projection.*

Then, the iterates $\left(\widehat{\mathbf{X}}^k = \widehat{\mathbf{U}}^k(\widehat{\mathbf{V}}^k)^T\right)_{k\geq 0}$ *of Algorithm 3 satisfy*

$$\limsup_{k\to\infty}\frac{\|\widehat{\mathbf{X}}^k - \mathbf{X}\|_F}{\|\mathbf{X}\|_F} \leq \frac{(55\kappa^2 + 3\kappa + 3)\|\boldsymbol{\eta}\|_2}{\|\mathcal{A}(\mathbf{X})\|_2}. \tag{3.27}$$

Moreover, the convergence is linear, i.e., for any $\epsilon > 0$, *there exists* $k_0 = \mathcal{O}(\log(1/\epsilon))$ *that satisfies*

$$\frac{\|\widehat{\mathbf{X}}^{k_0} - \mathbf{X}\|_F}{\|\mathbf{X}\|_F} \leq \frac{(55\kappa^2 + 3\kappa + 3)\|\boldsymbol{\eta}\|_2}{\|\mathcal{A}(\mathbf{X})\|_2} + \epsilon. \tag{3.28}$$

With this we conclude this subsection and turn towards a novel alternating minimization algorithm, which was introduced only very lately.

3.4.2 Alternating Tikhonov Regularization and LASSO

Motivated by the convincing performance of approaches based on alternating minimization and recent results in multi-penalty regularization [NP14], an algorithm relying on the alternating minimization of a certain multi-penalty functional was proposed in [FMN19] by Maly, Fornasier and Naumova. Their method is called Alternating Tikhonov regularization and LASSO (ATLAS). It assures both convergence and approximation guarantees and allows for more freedom in the recoverable class of signals compared to sparse power factorization. More specifically, we return to the setting of (s_1, s_2)-sparse low-rank-R matrices, which possess a non-orthogonal rank-1 decomposition, i.e., we aim at recovering $\mathbf{X} \in \mathcal{S}^R_{s_1,s_2}$ from inaccurate and incomplete measurements

$$\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\eta}. \tag{3.29}$$

In fact, the authors carried out the theoretical analysis for an even enlarged signal class by introducing the concept of effective sparsity, cf. Definition 4.3. Concerning this, we will go into further detail in the following Chapter 4, where we propose a more general version of the herein outlined numerical algorithm that contains ATLAS as a special case. The multi-penalty functional $\mathcal{J}^R_{\alpha,\beta} : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_2} \to \mathbb{R}$ leading to ATLAS is specifically designed for matrices, which have only sparse right component vectors, i.e., are elements of $\mathcal{S}^R_{n_1,s_2}$, and is given by

$$\mathcal{J}^R_{\alpha,\beta}(\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_R, \tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_R) = \left\| \mathcal{A}\left(\sum_{r=1}^{R} \tilde{\mathbf{u}}_r \tilde{\mathbf{v}}_r^T \right) - \mathbf{y} \right\|_2^2 + \alpha \sum_{r=1}^{R} \|\tilde{\mathbf{u}}_r\|_2^2 + \beta \sum_{r=1}^{R} \|\tilde{\mathbf{v}}_r\|_1 \tag{3.30}$$

for regularization parameters $\alpha, \beta > 0$. Note, however, that sparsity in the left component vectors can be easily promoted by replacing the Tikhonov regularization terms $\|\tilde{\mathbf{u}}_r\|_2^2$ with the sparsity promoting ℓ_1-norm-regularizers $\|\tilde{\mathbf{u}}_r\|_1$.

The operating principle of ATLAS is now to alternate on R pairs of left and right component vectors $\tilde{\mathbf{u}}_r$ and $\tilde{\mathbf{v}}_r$. The former entails a standard Tikhonov regularization problem, whose solution can be given explicitly via the Moore-Penrose inverse. The latter yields a LASSO problem of the form (1.67), which can be tackled using iterative soft thresholding, cf. Paragraph 1.4.3(1). This approach allows to enlarge the class of recoverable matrices by giving up the joint-sparsity requirement.

In [FMN19] local convergence of ATLAS to global minimizers of the functional $\mathcal{J}^R_{\alpha,\beta}$ was established. Moreover, such global optima were analyzed and shown to have meaningful properties one would expect from solutions to the inverse problem (3.29), such as having a small measurement misfit and an in some sense sparse decomposition. Eventually, the number of required measurements for recovery with high probability was proven to be at the information theoretic limit up to a (poly)logarithmic factor in the ambient dimension of the component vectors.

We refrain from going into further theoretical and numerical detail at this time, as we will investigate this closely in the more general setting of the following Chapter 4. Furthermore, we do not give a pseudocode here, as Algorithm 5 on page 78 reduces to ATLAS by setting $p = 2$ and $q = 1$.

Chapter 4

Compressed Sensing of Simultaneously Sparse and Low-Rank Matrices— A Multi-Penalty Approach relying on Non-Convex Regularizers

With the knowledge of the preceded parts in mind, in this chapter we propose to approach the inverse problem of recovering a simultaneously sparse and low-rank matrix from inaccurate and incomplete measurements using a method based on alternating minimization of a highly non-convex multi-penalty functional. This functional comprises both data fidelity and the two low-dimensional structures. Whereas our algorithm enforces low-rankness directly, multiple non-convex regularizers are employed to promote sparsity in the components of a non-orthogonal low-rank factorization individually. We start the chapter with a brief recapitulation of our problem setup before we summarize properties of global optima of the multi-penalty functional. This is followed by the introduction of our numerical method, which we call Alternating Ridge and Bridge or ℓ_0-Regression. We outline how to prove convergence of the algorithm and address the question of initialization. Afterwards, we turn towards the theoretical facets of recovering simultaneously sparse and low-rank matrices by introducing two classes of matrices along with a restricted isometry property. Finally, we investigate the number of required measurements, which turns out to be optimal up to a polylogarithmic factor in the ambient dimension of the component vectors and the rank.

The present Chapter 4 and the subsequent Chapter 5 can be regarded and read in parallel as they are separated by decoupling numerics from theory, rather than being separated chronologically and substantively.

The results presented in these two chapters are unpublished joint work with Johannes Maly and Massimo Fornasier. They generalize the findings of their work together with Valeriya Naumova, which was published in [FMN19].

4.1 Problem Formulation and Our Contribution

For the reader's convenience, let us recapitulate the formulation of the problem setup from a purely mathematical point of view. A motivating example was already given in

Section 3.1 on a recommendation system for a grocery store.
We aim at recovering a simultaneously sparse and low-rank matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ from m of its inaccurate and incomplete linear measurements, which are gathered in the measurement vector

$$\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\eta}. \tag{4.1}$$

In this sensing process, $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ describes a suitable measurement operator and $\boldsymbol{\eta} \in \mathbb{R}^m$ ineliminable noise. The action of the former can be parametrized, as described at the beginning of Section 2.2, by m separate Frobenius scalar products $\langle \mathbf{A}_i, \mathbf{X} \rangle_F$ with matrices $\mathbf{A}_i \in \mathbb{R}^{n_1 \times n_2}$. Moreover, from the latter we are only aware of an upper bound $\eta > 0$ on its Euclidean norm, i.e., $\|\boldsymbol{\eta}\|_2 \leq \eta$. The types of signals we consider and want to recover are assumed to be represented by matrices that have a certain structure by admitting a not-necessarily-orthogonal low-rank-R decomposition of the form

$$\mathbf{X} = \sum_{r=1}^{R} \mathbf{u}_r \mathbf{v}_r^T \tag{4.2}$$

with s_1-sparse left and s_2-sparse right component vectors, i.e., $\mathbf{u}_r \in \Sigma_{s_1}^{n_1}$ and $\mathbf{v}_r \in \Sigma_{s_2}^{n_2}$ for all $r \in [R]$, respectively. Such matrices are called (s_1, s_2)-sparse rank-R matrices and they form the set $\mathcal{S}_{s_1,s_2}^R$, cf. equation (3.7).
The work and methodology in this chapter are encouraged by the results of [FMN19] and multi-penalty regularization in general, combined with the theoretically better and closer-optimal performance of non-convex regularizers for the compressed recovery of sparse vectors, cf. Remarks 1.18 and 1.22. For the recovery of an (s_1, s_2)-sparse low-rank-R matrix $\mathbf{X}$ possessing a decomposition as in (4.2) from linear inaccurate and incomplete measurements $\mathbf{y}$ of the form (4.1), we propose the alternating minimization of the multi-penalty functional $\mathcal{J}_{\alpha,\beta}^{p,q,R} : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_2} \to \mathbb{R}$. It is given by

$$\mathcal{J}_{\alpha,\beta}^{p,q,R}(\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_R, \tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_R) = \left\| \mathcal{A}\left(\sum_{r=1}^{R} \tilde{\mathbf{u}}_r \tilde{\mathbf{v}}_r^T \right) - \mathbf{y} \right\|_2^2 + \alpha \sum_{r=1}^{R} \|\tilde{\mathbf{u}}_r\|_p^p + \beta \sum_{r=1}^{R} \|\tilde{\mathbf{v}}_r\|_q^q \tag{4.3}$$

for regularizing (quasi)-norm parameters $0 < p, q \leq 2$ and associated regularization parameters $\alpha, \beta > 0$. This functional is highly non-convex, for one thing due to the generalized bilinear factorization of $\widetilde{\mathbf{X}} = \sum_{r=1}^{R} \tilde{\mathbf{u}}_r (\tilde{\mathbf{v}}_r)^T$ in the least-squares term and for the other due to the multi-penalty regularization term, which is permitted to include non-convex regularizers for the individual component vectors. This favors and allows for successful recovery from highly incomplete measurements, cf. Sections 3.2 and 3.3.
A global minimizer of the functional $\mathcal{J}_{\alpha,\beta}^{p,q,R}$ is denoted by

$$\left(\left(\hat{\mathbf{u}}_{\alpha,\beta}^{p,q}\right)_1, \ldots, \left(\hat{\mathbf{u}}_{\alpha,\beta}^{p,q}\right)_R, \left(\hat{\mathbf{v}}_{\alpha,\beta}^{p,q}\right)_1, \ldots, \left(\hat{\mathbf{v}}_{\alpha,\beta}^{p,q}\right)_R \right). \tag{4.4}$$

Sometimes, admitting a minor inaccuracy, we directly refer to the resulting matrix $\widehat{\mathbf{X}}_{\alpha,\beta}^{p,q}$ as the minimizer of the functional (4.3). It is given by

$$\widehat{\mathbf{X}}_{\alpha,\beta}^{p,q} = \sum_{r=1}^{R} \left(\hat{\mathbf{u}}_{\alpha,\beta}^{p,q}\right)_r \left(\left(\hat{\mathbf{v}}_{\alpha,\beta}^{p,q}\right)_r \right)^T. \tag{4.5}$$

Let us spend a few words on the idea behind the multi-penalty functional $\mathcal{J}_{\alpha,\beta}^{p,q,R}$ proposed in (4.3). It comprises data fidelity in form of the squared Euclidean norm of the residual $\mathbf{y}-\mathcal{A}(\widetilde{\mathbf{X}})$ together with low-rankness of the matrix $\widetilde{\mathbf{X}}$ as well as sparsity in one or both vector components of its non-orthogonal decomposition. The former of the two structures is hard-coded into the formulation by allowing only R individual vector pairs $(\tilde{\mathbf{u}}_r, \tilde{\mathbf{v}}_r)$ in the low-rank decomposition. The latter, in turn, is promoted by the regularizing ℓ_p- and ℓ_q-(quasi)-norms of the left and right component vectors $\tilde{\mathbf{u}}_r$ and $\tilde{\mathbf{v}}_r$, respectively.
On the basis of this functional we propose an algorithm, which we title **A**lternating **R**idge and **B**ridge or ℓ_0-**R**egression (ARBeR[8]). It can be regarded as a generalization of the in [FMN19] introduced method ATLAS and is in particular designed for the same class of matrices, namely right-sided sparse and low-rank matrices, i.e., matrices in the set $\mathcal{S}_{n_1,s_2}^R$. However, as we introduce it, it is amenable also to sparsity in both, rows and columns. One is even able to impose different regularizing (quasi)-norms for the two component vectors as it is the case in (4.3). As its relative it is based on alternating minimization and alternates on R pairs of left and right component vectors $\tilde{\mathbf{u}}_r$ and $\tilde{\mathbf{v}}_r$. More precisely, in each iteration we successively solve the individual vector-valued minimization problems

$$\hat{\mathbf{u}}_r^k = \underset{\hat{\mathbf{u}}\in\mathbb{R}^{n_1}}{\arg\min} \left\| \mathbf{y} - \mathcal{A}\Big(\sum_{\rho<r} \hat{\mathbf{u}}_\rho^k (\hat{\mathbf{v}}_\rho^k)^T\Big) - \mathcal{A}\big(\hat{\mathbf{u}}(\hat{\mathbf{v}}_r^{k-1})^T\big) - \mathcal{A}\Big(\sum_{\rho>r} \hat{\mathbf{u}}_\rho^{k-1} (\hat{\mathbf{v}}_\rho^{k-1})^T\Big) \right\|_2^2 + \alpha \|\hat{\mathbf{u}}\|_p^p \tag{4.6}$$

and

$$\hat{\mathbf{v}}_r^k = \underset{\hat{\mathbf{v}}\in\mathbb{R}^{n_2}}{\arg\min} \left\| \mathbf{y} - \mathcal{A}\Big(\sum_{\rho<r} \hat{\mathbf{u}}_\rho^k (\hat{\mathbf{v}}_\rho^k)^T\Big) - \mathcal{A}\big(\hat{\mathbf{u}}_r^k \hat{\mathbf{v}}^T\big) - \mathcal{A}\Big(\sum_{\rho>r} \hat{\mathbf{u}}_\rho^{k-1} (\hat{\mathbf{v}}_\rho^{k-1})^T\Big) \right\|_2^2 + \beta \|\hat{\mathbf{v}}\|_q^q \tag{4.7}$$

for each $r \in [R]$. In contrast to ATLAS, even these individual vector-valued minimization problems are non-convex if $0 < p < 1$ or $0 < q < 1$, respectively. However, compared to the joint-minimization with respect to all $2R$ component vectors, these programs are tractable and the solutions can be well-approximated by iterative bridge thresholding, cf. Paragraph 1.4.3(2).
Let us now give a concise outline for the rest of this chapter. Beginning with properties of global minimizers of the multi-penalty functional $\mathcal{J}_{\alpha,\beta}^{p,q,R}$ under very weak conditions in Section 4.2, we investigate the functional more closely. We verify that global optima provide reasonable solutions to the inverse problem. In this context we introduce the novel concept of ℓ_q-effective sparsity, which weakens the notion of sparsity in a graduated sense. To illustrate the associated set of vectors we discuss astonishing phenomena from high-dimensional geometry. In Section 4.3 we formalize the algorithm ARBeR and describe how to tackle the individual minimization problems. Moreover, we sketch a proof of convergence when the method is initialized sufficiently well, yet, provide no universal initialization strategy. However, we suggest initialization with the leading left and right singular vectors of $\mathcal{A}^*(\mathbf{y})$, which empirically proved itself to show good results. Additionally, we recommend a multilevel-type strategy that is capable of enhancing recovery in severely non-convex cases. After having presented these results demanding minimal requirements, we introduce a suitably designed restricted isometry property together with a new matrix model in the subsequent Section 4.4. This model generalizes the set of sparse and low-rank matrices possessing a non-orthogonal decomposition of the type (4.2) by

[8] Observe that the **e** in ARBeR stands for the **e** in ℓ_0 when being pronounced as **e**ll-zero.

involving the formerly motivated set of effectively sparse vectors. Based thereon we prove a recovery result under the assumption of the restricted isometry property. The question of the required number of measurements for which this restricted isometry property and consequently recoverability is ensured with high probability is the content of Section 4.5. In order to determine a lower bound on this number, we make use of tools from the theory of stochastic processes, in specific a bound on suprema of chaos processes, which is based on a chaining method [KMR14]. For its application we bound the appearing Talagrand's γ_2-functional in terms of the metric entropy of the newly introduced sets of matrices having an (effectively) sparse and low-rank decomposition by employing Dudley's inequality. We eventually show that the necessary number of Gaussian measurements is optimal up to a polylogarithmic factor in the ambient dimension of the component vectors and the rank. In this final part of the theoretical analysis and also in the subsequent numerical considerations, we are mainly interested in two special instances of the multi-penalty functional (4.3) related to the kind of matrix we want to recover. First, in case that $\mathbf{X}$ possesses a decomposition with sparse left and right component vectors, it is reasonable to use the same regularizing (quasi)-norm for both of them, i.e., we set $p = q$ and let $0 < q \leq 1$. Second, if only the right component vector is known to be sparse, we use $p = 2$ and $0 < q \leq 1$. In both cases, though, we write $\mathcal{J}_{\alpha,\beta}^{q,R}$ instead of $\mathcal{J}_{\alpha,\beta}^{q,q,R}$ and $\mathcal{J}_{\alpha,\beta}^{2,q,R}$, respectively. They are distinguishable from one another by the signal set under consideration, i.e., whether we investigate $\mathcal{S}_{s_1,s_2}^{R}$ or $\mathcal{S}_{n_1,s_2}^{R}$.

Regarding notation as well, let us recall the version of the sparse decomposition (SD) with normalized component vectors, which was also used in the definition of the set of (s_1, s_2)-sparse rank-R matrices (3.7). For the matrix $\mathbf{X}$ with low-rank decomposition (4.2) let us define quasi-singular values $\sigma_r = \|\mathbf{u}_r\|_2\|\mathbf{v}_r\|_2$ for $r \in [R]$, which may not be confused with singular values. Then,

$$\mathbf{X} = \underline{\mathbf{U}}\mathbf{\Sigma}\underline{\mathbf{V}}^T = \sum_{r=1}^{R} \sigma_r \frac{\mathbf{u}_r}{\|\mathbf{u}_r\|_2}\left(\frac{\mathbf{v}_r}{\|\mathbf{v}_r\|_2}\right)^T, \tag{4.8}$$

where the matrices $\underline{\mathbf{U}}$ and $\underline{\mathbf{V}}$ contain the normalized left and right component vectors as columns. $\mathbf{\Sigma}$ is a diagonal matrix, whose rth diagonal entry is σ_r. Since the two former matrices may not be orthogonal in general, we only have an equivalence $\|\mathbf{X}\|_F \simeq \|\mathbf{\Sigma}\|_F$ instead of an equality between those two quantities. The hidden constants correspond to the square roots of the smallest and largest eigenvalue of the Gramian of $\underline{\mathbf{U}}$. More precisely, see, e.g., Lemma A.6 for a proof, we have $c_{\underline{\mathbf{U}}}\|\mathbf{\Sigma}\|_F \leq \|\mathbf{X}\|_F \leq C_{\underline{\mathbf{U}}}\|\mathbf{\Sigma}\|_F$ with constants $c_{\underline{\mathbf{U}}} = \sqrt{\lambda_{\min}(\underline{\mathbf{U}}^T\underline{\mathbf{U}})}$ and $C_{\underline{\mathbf{U}}} = \sqrt{\lambda_{\max}(\underline{\mathbf{U}}^T\underline{\mathbf{U}})}$. Based on this equivalence and the in Lemma A.2 established equivalence of the ℓ_q-(quasi)-norms, we find the following relation between the Schatten-q-(quasi)-norms and the ℓ_q-(quasi)-norms of the vector $\boldsymbol{\sigma}$ of quasi-singular values σ_r for $0 < q \leq 2$, namely

$$c_{\underline{\mathbf{U}}}R^{1/2-1/q}\left(\sum_{r=1}^{R}(\|\mathbf{u}_r\|_2\|\mathbf{v}_r\|_2)^q\right)^{1/q} \leq \|\mathbf{X}\|_q \leq C_{\underline{\mathbf{U}}}R^{1/q-1/2}\left(\sum_{r=1}^{R}(\|\mathbf{u}_r\|_2\|\mathbf{v}_r\|_2)^q\right)^{1/q}. \tag{4.9}$$

Exemplarily, let us reason the upper bound as follows,

$$\|\mathbf{X}\|_q \leq R^{1/q-1/2}\|\mathbf{X}\|_F \leq C_{\underline{\mathbf{U}}}R^{1/q-1/2}\|\mathbf{\Sigma}\|_F = C_{\underline{\mathbf{U}}}R^{1/q-1/2}\|\boldsymbol{\sigma}\|_2 \leq C_{\underline{\mathbf{U}}}R^{1/q-1/2}\|\boldsymbol{\sigma}\|_q, \tag{4.10}$$

where $\boldsymbol{\sigma} = \mathrm{diag}(\mathbf{\Sigma})$. The lower bound follows analogously.

4.2 On Global Minimizers of the Non-Convex Multi-Penalty Functional

As already announced in the outline at the end of the preceded section, we start our analysis under very mild assumptions on the inverse problem with properties of global minimizers $\big((\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_1,\dots,(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_R,(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_1,\dots,(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_R\big)$ of the functional $\mathcal{J}^{p,q,R}_{\alpha,\beta}$. The features that we expect to be promoted by the proposed multi-penalty functional are data fidelity together with the two low-dimensional structures, low-rankness and sparsity in the left and right component vectors of the decomposition.

4.2.1 Data Fidelity of Global Optima

Let us begin by establishing an upper bound on the measurement misfit of the global minimizer, i.e., the residual $\mathbf{y}-\mathcal{A}(\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta})$. The following result is a natural generalization of Proposition 3.1 from [FMN19].

Proposition 4.1 (Measurement Misfit for Global Minimizers)**.** *Let us assume that* $\mathbf{X}$ *admits a decomposition as in* (4.2) *and fulfills the noisy measurements* $\mathbf{y}=\mathcal{A}(\mathbf{X})+\boldsymbol{\eta}$. *Moreover, let* $\big((\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_1,\dots,(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_R,(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_1,\dots,(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_R\big)$ *denote a global minimizer of* $\mathcal{J}^{p,q,R}_{\alpha,\beta}$. *Then*

$$\big\|\mathbf{y}-\mathcal{A}(\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta})\big\|_2^2 \le \|\boldsymbol{\eta}\|_2^2 + C_{pq}\left(\alpha^q\beta^p\right)^{\frac{1}{p+q}}\sum_{r=1}^{R}\left(\|\mathbf{u}_r\|_p\|\mathbf{v}_r\|_q\right)^{\frac{pq}{p+q}}, \tag{4.11}$$

where $C_{pq}=\left(\frac{p}{q}\right)^{\frac{q}{p+q}}+\left(\frac{q}{p}\right)^{\frac{p}{p+q}}$.

Proof. By using the global optimality of $\big((\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_1,\dots,(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_R,(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_1,\dots,(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_R\big)$ together with the definition of the multi-penalty functional $\mathcal{J}^{p,q,R}_{\alpha,\beta}$ we obtain

$$\begin{aligned}\big\|\mathbf{y}-\mathcal{A}(\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta})\big\|_2^2 &\le \big\|\mathbf{y}-\mathcal{A}(\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta})\big\|_2^2+\alpha\sum_{r=1}^{R}\big\|\big(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta}\big)_r\big\|_p^p+\beta\sum_{r=1}^{R}\big\|\big(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta}\big)_r\big\|_q^q\\ &=\mathcal{J}^{p,q,R}_{\alpha,\beta}\left(\big(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta}\big)_1,\dots,\big(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta}\big)_R,\big(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta}\big)_1,\dots,\big(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta}\big)_R\right)\\ &\le \mathcal{J}^{p,q,R}_{\alpha,\beta}\left(\lambda_1\mathbf{u}_1,\dots,\lambda_R\mathbf{u}_R,\lambda_1^{-1}\mathbf{v}_1,\dots,\lambda_R^{-1}\mathbf{v}_R\right)\\ &=\big\|\mathbf{y}-\mathcal{A}(\mathbf{X})\big\|_2^2+\alpha\sum_{r=1}^{R}\lambda_r^p\|\mathbf{u}_r\|_p^p+\beta\sum_{r=1}^{R}\lambda_r^{-q}\|\mathbf{v}_r\|_q^q\\ &=\|\boldsymbol{\eta}\|_2^2+\sum_{r=1}^{R}\left(\alpha\lambda_r^p\|\mathbf{u}_r\|_p^p+\beta\lambda_r^{-q}\|\mathbf{v}_r\|_q^q\right)\end{aligned} \tag{4.12}$$

for any set of positive parameters $(\lambda_r)_{r=1}^R$. The right-hand side in this expression can be optimized with respect to these parameters. An application of Lemma A.8 to each of the R terms $\alpha\lambda_r^p\|\mathbf{u}_r\|_p^p+\beta\lambda_r^{-q}\|\mathbf{v}_r\|_q^q$ individually yields

$$\begin{aligned}\big\|\mathbf{y}-\mathcal{A}(\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta})\big\|_2^2 &\le \|\boldsymbol{\eta}\|_2^2+\sum_{r=1}^{R}\left(C_{pq}\big(\alpha\|\mathbf{u}_r\|_p^p\big)^{\frac{q}{p+q}}\big(\beta\|\mathbf{v}_r\|_q^q\big)^{\frac{p}{p+q}}\right)\\ &\le \|\boldsymbol{\eta}\|_2^2+C_{pq}(\alpha^q\beta^p)^{\frac{1}{p+q}}\sum_{r=1}^{R}\left(\|\mathbf{u}_r\|_p\|\mathbf{v}_r\|_q\right)^{\frac{pq}{p+q}},\end{aligned} \tag{4.13}$$

where C_{pq} denotes the constant from Lemma A.8. □

This proposition shows that the norm of the residual can be bounded up to the noise level and an additional term in the sparsity promoting norms of the respective component vectors. This summand can be controlled by the regularization parameters α and β. In fact, it seems that this term can be made arbitrarily small, if the parameters are. However, this causes certain difficulties. First of all, extending Lemma 3.2 from [FMN19], we show that one needs to choose parameters of similar magnitude in order to maintain control of both groups of component vectors. Furthermore, if the parameters become too small, it is natural to suspect that data fidelity is achieved at the expense of the sparsity structure, what we will see in Proposition 4.4.

Lemma 4.2 (Boundedness of Global Minimizers in Sparsity Promoting Norm)**.** *Let us assume that* $\mathbf{X}$ *admits a decomposition as in* (4.2) *and fulfills the noisy measurements* $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\eta}$*. Moreover, let* $\big((\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_1, \dots, (\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_R, (\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_1, \dots, (\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_R\big)$ *denote a global minimizer of* $\mathcal{J}^{p,q,R}_{\alpha,\beta}$*. Then, if* $\big\|\mathbf{y} - \mathcal{A}(\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta})\big\|_2 \geq \|\boldsymbol{\eta}\|_2$*, it hold*

$$\sum_{r=1}^{R} \big\|(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_r\big\|_p^p \leq C_{pq} \left(\frac{\beta}{\alpha}\right)^{\frac{p}{p+q}} \sum_{r=1}^{R} \left(\|\mathbf{u}_r\|_p \|\mathbf{v}_r\|_q\right)^{\frac{pq}{p+q}} \tag{4.14}$$

and

$$\sum_{r=1}^{R} \big\|(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_r\big\|_q^q \leq C_{pq} \left(\frac{\alpha}{\beta}\right)^{\frac{q}{p+q}} \sum_{r=1}^{R} \left(\|\mathbf{u}_r\|_p \|\mathbf{v}_r\|_q\right)^{\frac{pq}{p+q}}, \tag{4.15}$$

where $C_{pq} = \left(\frac{p}{q}\right)^{\frac{q}{p+q}} + \left(\frac{q}{p}\right)^{\frac{p}{p+q}}$*. Under the same assumptions it furthermore holds*

$$\sum_{r=1}^{R} \left(\big\|(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_r\big\|_p \big\|(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_r\big\|_q\right)^{\frac{pq}{p+q}} \leq \sum_{r=1}^{R} \left(\|\mathbf{u}_r\|_p \|\mathbf{v}_r\|_q\right)^{\frac{pq}{p+q}}. \tag{4.16}$$

Proof. By revisiting the proof of Proposition 4.1 we also deduce

$$\begin{aligned}\big\|\mathbf{y} - \mathcal{A}(\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta})\big\|_2^2 + \alpha \sum_{r=1}^{R} \big\|\big(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta}\big)_r\big\|_p^p + \beta \sum_{r=1}^{R} \big\|\big(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta}\big)_r\big\|_q^q \\ \leq \|\boldsymbol{\eta}\|_2^2 + C_{pq}(\alpha^q \beta^p)^{\frac{1}{p+q}} \sum_{r=1}^{R} \left(\|\mathbf{u}_r\|_p \|\mathbf{v}_r\|_q\right)^{\frac{pq}{p+q}}.\end{aligned} \tag{4.17}$$

Utilizing the assumption we conclude

$$\alpha \sum_{r=1}^{R} \big\|\big(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta}\big)_r\big\|_p^p + \beta \sum_{r=1}^{R} \big\|\big(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta}\big)_r\big\|_q^q \leq C_{pq}(\alpha^q \beta^p)^{\frac{1}{p+q}} \sum_{r=1}^{R} \left(\|\mathbf{u}_r\|_p \|\mathbf{v}_r\|_q\right)^{\frac{pq}{p+q}}, \tag{4.18}$$

showing the first two statements when considering only one of the two terms on the left-hand side at a time.
In order to show the last inequality, note that we have

$$\mathcal{J}^{p,q,R}_{\alpha,\beta}\left(\big(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta}\big)_1, \dots, \big(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta}\big)_R\right) = \inf_{\lambda_1,\dots,\lambda_R > 0} \mathcal{J}^{p,q,R}_{\alpha,\beta}\left(\lambda_1 \big(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta}\big)_1, \dots, \lambda_R^{-1} \big(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta}\big)_R\right) \tag{4.19}$$

The inequality direction "$\leq$" follows directly from optimality of $\big((\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_1, \dots, (\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_R\big)$. Additionally, as $\lambda_r = 1$ for all $r \in [R]$ is a valid choice of parameters, the right-hand side is bounded by $\mathcal{J}^{p,q,R}_{\alpha,\beta}\big((\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_1, \dots, (\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_R\big)$ and thus we also have the other direction "$\geq$". From this equality we follow

$$\alpha \sum_{r=1}^{R} \big\| \big(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta}\big)_r \big\|_p^p + \beta \sum_{r=1}^{R} \big\| \big(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta}\big)_r \big\|_q^q = C_{pq}(\alpha^q \beta^p)^{\frac{1}{p+q}} \sum_{r=1}^{R} \Big(\big\| \big(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta}\big)_r \big\|_p \big\| \big(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta}\big)_r \big\|_q \Big)^{\frac{pq}{p+q}} \quad (4.20)$$

by using analogous arguments as at the beginning of this proof. With this we obtain the first equality in the following. Moreover, we denote optimal parameters in the sense of Lemma A.8 by $\tilde{\lambda}_r$. Thus, using the argument from the proof of Proposition 4.1, we have

$$\begin{aligned} &\big\| \mathbf{y} - \mathcal{A}(\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta}) \big\|_2^2 + C_{pq}(\alpha^q \beta^p)^{\frac{1}{p+q}} \sum_{r=1}^{R} \Big(\big\| \big(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta}\big)_r \big\|_p \big\| \big(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta}\big)_r \big\|_q \Big)^{\frac{pq}{p+q}} \\ &\qquad = \mathcal{J}^{p,q,R}_{\alpha,\beta} \Big(\big(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta}\big)_1, \dots, \big(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta}\big)_R \Big) \leq \mathcal{J}^{p,q,R}_{\alpha,\beta} \big(\tilde{\lambda}_1 \mathbf{u}_1, \dots, \tilde{\lambda}_R^{-1} \mathbf{v}_R \big) \qquad (4.21) \\ &\qquad \leq \|\boldsymbol{\eta}\|_2^2 + C_{pq}(\alpha^q \beta^p)^{\frac{1}{p+q}} \sum_{r=1}^{R} \big(\|\mathbf{u}_r\|_p \|\mathbf{v}_r\|_q \big)^{\frac{pq}{p+q}}, \end{aligned}$$

which assures the claim after using the assumption. □

Before addressing properties associated with the two low-dimensional structures, we introduce the concept of ℓ_q-effective sparsity in the next subsection, as this will be used to investigate the sparsity in the left and right component vectors.

4.2.2 Effective Sparsity

It is a well-known consequence of Hölder's inequality that the ℓ_q-(quasi)-norm of an s-sparse vector $\mathbf{z} \in \mathbb{R}^N$ can be bounded by

$$\|\mathbf{z}\|_q \leq s^{1/q - 1/2} \|\mathbf{z}\|_2. \qquad (4.22)$$

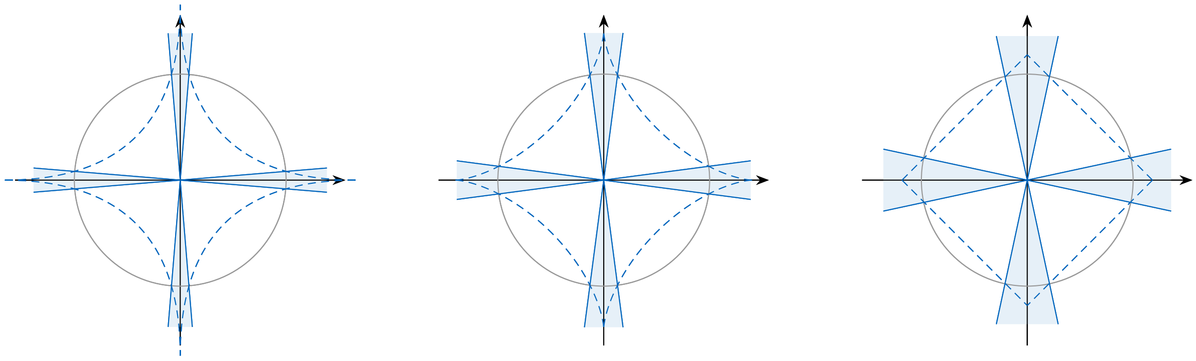
For a proof thereof we refer to Remark A.3 in the appendix. This property motivates to generalize the notion of sparsity by introducing a larger set of vectors obeying inequality (4.22). This so-called set of ℓ_q-effectively s-sparse vectors is given in Definition 4.3 below. A related, yet, slightly different version of this concept was introduced in [PV13, Section 3] for the case $q = 1$.

Definition 4.3 (ℓ_q-Effectively Sparse Vectors). *For $0 < q \leq 2$, $N \in \mathbb{N}$ and $1 \leq s \leq N$,*

$$K_s^{q,N} = \big\{ \mathbf{z} \in \mathbb{R}^N : \|\mathbf{z}\|_q \leq s^{1/q - 1/2} \|\mathbf{z}\|_2 \big\} \qquad (4.23)$$

defines the set of ℓ_q-effectively s-sparse vectors in $\mathbb{R}^N$.

Let us reflect upon this definition. First, we note that despite not having excluded the situations $q = 2$ and $s = N$, the naming is inappropriate in these instances, as we have $K_s^{q,N} = \mathbb{R}^N$. In all other cases the set of ℓ_q-effectively sparse vectors is strictly smaller. Second, as already mentioned, we have the inclusion $\Sigma_s^N \subset K_s^{q,N}$. Beyond that, the set is monotonous in the sparsity as well as in the regularizing (quasi)-norm parameter q, i.e., we have both $K_{s_1}^{q,N} \subset K_{s_2}^{q,N}$ for $s_1 \leq s_2$ and $K_s^{q_1,N} \subset K_s^{q_2,N}$ for $q_1 \leq q_2$. The former follows immediately and the latter uses Hölder's inequality as shown in Lemma A.9 in the appendix. As q approaches zero, the set $K_s^{q,N}$ of ℓ_q-effectively s-sparse vectors approaches the set Σ_s^N of exactly s-sparse vectors. These facts are also suggested by Figure 4.1 below, which we want to describe more closely in the following.



(a) $q \ll 1$, N small or $q < 1$, N large

(b) $q < 1$, N small or $q \leq 1$, N large

(c) $q = 1$, $N = 2$

Figure 4.1. Geometry of the set of ℓ_q-effectively sparse vectors $K_s^{q,N}$ for different values of q and in different dimensions N.

Let us for the moment focus on the two-dimensional situation which our geometrical intuition is accustomed to and which we can illustrate. Each individual subfigure depicts, for a different value of q, a properly scaled ℓ_q-(quasi)-norm-ball of radius $s^{1/q-1/2}$ together with the sphere of a unit ℓ_2-norm ball, i.e., $\mathbb{S}^1$. The shaded non-convex cone then pictures the set $K_s^{q,2}$, exploiting the general observation that $K_s^{q,N} = \operatorname{cone}\left(\mathbb{S}^{N-1} \cap \mathcal{B}_q^N(\mathbf{0}, s^{1/q-1/2})\right)$. In order to understand the behavior in high dimensions we pause for a brief excursion to high-dimensional geometry. We discuss two surprising and counterintuitive phenomena regarding the geometry of convex bodies and ℓ_q-(quasi)-norm balls in high-dimensional spaces. They even merge into one another.

First of all let us compute the volume of an arbitrary unit ℓ_q-(quasi)-norm ball. By using the Cavalieri principle, we obtain the following recursive formula

$$\begin{aligned}
\lambda^N(\mathcal{B}_q^N) &= \int_{-1}^{1} \int_{\mathbf{z} \in \mathbb{R}^{N-1} : |z_1|^q + \dots + |z_{N-1}|^q \leq 1 - |z_N|^q} 1 \, \mathrm{d}\lambda^{N-1}(\mathbf{z}) \, \mathrm{d}\lambda(z_N) \\
&= \int_{-1}^{1} \int_{\mathcal{B}_q^{N-1}(\mathbf{0}, (1-|z_N|^q)^{1/q})} 1 \, \mathrm{d}\lambda^{N-1} \, \mathrm{d}\lambda(z_N) \\
&= \int_{-1}^{1} \lambda^{N-1}\left(\mathcal{B}_q^{N-1}(\mathbf{0}, (1-|z_N|^q)^{1/q})\right) \mathrm{d}\lambda(z_N) \qquad (4.24) \\
&= \lambda^{N-1}\left(\mathcal{B}_q^{N-1}\right) \int_{-1}^{1} (1-|z|^q)^{(N-1)/q} \, \mathrm{d}\lambda(z) \\
&= \lambda^{N-1}\left(\mathcal{B}_q^{N-1}\right) 2 \frac{\Gamma\left(\frac{1}{q}+1\right)\Gamma\left(\frac{N-1}{q}+1\right)}{\Gamma\left(\frac{N}{q}+1\right)},
\end{aligned}$$

where λ^N denotes the N-dimensional Lebesgue measure. The last step involves the Gamma function Γ and requires some computation, which is omitted here. As $\lambda^1(\mathcal{B}_q^1) = 2$

we conclude

$$\lambda^N(\mathcal{B}_q^N) = \frac{\left(2\Gamma\left(\frac{1}{q}+1\right)\right)^N}{\Gamma\left(\frac{N}{q}+1\right)}. \tag{4.25}$$

This recovers the familiar volume expression of the Euclidean unit ball, namely $\frac{\pi^{N/2}}{\Gamma(N/2+1)}$, when setting $q = 2$. Furthermore, we also obtain easy volume formulas for other (quasi)-norm balls, for instance, $\frac{2^N}{N!}$ and $\frac{4^N}{(2N)!}$ for the ℓ_1-norm ball and the $\ell_{1/2}$-quasi-norm ball, respectively. It is straightforward to see that the volumes approach zero as N tends to infinity, meaning that high-dimensional balls contain almost no volume. To make this more rigorous we utilize Sterling's approximation $\Gamma(z+1) \simeq \sqrt{2\pi z}\left(\frac{z}{e}\right)^z$ in equation (4.25) in order to highlight the asymptotic behavior

$$\lambda^N(\mathcal{B}_q^N) \sim \frac{2}{\sqrt{N}}\left(\frac{8\pi}{q}\right)^{(N-1)/2}\frac{1}{N^{N/q}}. \tag{4.26}$$

In fact, this also shows that the decay is superexponential and stronger the smaller q is. The former observation becomes even more counterintuitive when realizing that despite having incredibly small volume, each individual ℓ_q-(quasi)-norm ball contains all unit vectors of the canonical basis, from which there are 2^N in N dimensions. Seeming somewhat contradictory this raises the question how ℓ_q-(quasi)-norm balls in particular and convex bodies in general look like. Heuristically, as summarized in [Ver15], a convex set consist of a bulk and lots of outliers. The former contains almost all the volume despite being small in diameter. The latter, in turn, reach far into space, yet, contribute almost nothing to the volume. Figure 4.2 illustrates this behavior. Although the shape does not look convex at all, this representation is required to depict the topology more accurately.

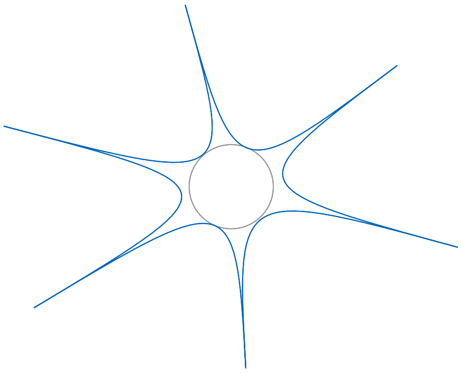


Figure 4.2. Hyperbolic representation of a high-dimensional convex set according to Milman, cf. [Mil98].

We want to clarify this visualization at the example of the ℓ_q-(quasi)-norm balls, which are convex in the case $q \geq 1$. An inscribed Euclidean ball touching the $(N-1)$-dimensional faces of the ℓ_q ball, has radius $N^{1/2-1/q}$. Using the dominant term in expression (4.26) for the asymptotic volume of the respective balls, we deduce

$$\left(\lambda^N(\mathcal{B}_q^N)\right)^{1/N} \sim \left(\lambda^N(\mathcal{B}_2^N(\mathbf{0}, N^{1/2-1/q}))\right)^{1/N} \sim N^{-1/q}. \tag{4.27}$$

Thus, even convex sets such as the ℓ_1 ball gather almost all their volume in a small bulk, while almost volumeless long spikes point into the direction of the coordinate axes.

With this we can now come back to Figure 4.1. Having the special aspects of high-dimensional geometry in mind, it is reasonable to depict also the suitably scaled convex ℓ_1-norm ball using a non-convex shape, as done in Subfigure 4.1(b).
After this brief excursion to high-dimensional geometry let us now turn back to the desirable structural properties of global minimizers of the multi-penalty functional $\mathcal{J}_{\alpha,\beta}^{p,q,R}$.

4.2.3 Sparsity and Low-Rankness of Global Optima

First, we investigate the sparsity structure of the left and right component vectors of the matrix $\widehat{\mathbf{X}}_{\alpha,\beta}^{p,q}$. The proposition below, which is a straightforward adaptation of Lemma 3.3 from [FMN19] establishes effective sparsity in the component vectors.

Proposition 4.4 (Sparsity Control of Component Vectors of Global Minimizers)**.** *Let us assume that* $\mathbf{X}$ *admits a decomposition as in* (4.2) *and let* $\mathbf{y} \in \mathbb{R}^m$*. Moreover, let* $\big((\hat{\mathbf{u}}_{\alpha,\beta}^{p,q})_1, \dots, (\hat{\mathbf{u}}_{\alpha,\beta}^{p,q})_R, (\hat{\mathbf{v}}_{\alpha,\beta}^{p,q})_1, \dots, (\hat{\mathbf{v}}_{\alpha,\beta}^{p,q})_R\big)$ *denote a global minimizer of* $\mathcal{J}_{\alpha,\beta}^{p,q,R}$*. Then,*

(i) if $\big\|(\hat{\mathbf{u}}_{\alpha,\beta}^{p,q})_r\big\|_2^p \geq \|\mathbf{y}\|_2^2/\gamma_1$ *for some* $\gamma_1 > 0$*, it holds*

$$\frac{\big\|\big(\hat{\mathbf{u}}_{\alpha,\beta}^{p,q}\big)_r\big\|_p}{\big\|\big(\hat{\mathbf{u}}_{\alpha,\beta}^{p,q}\big)_r\big\|_2} < \left(\frac{\gamma_1}{\alpha}\right)^{1/p}, \tag{4.28}$$

implying $(\hat{\mathbf{u}}_{\alpha,\beta}^{p,q})_r \in K_{\hat{s}_1}^{p,n_1}$ *with* $\hat{s}_1 = (\gamma_1/\alpha)^{2/(2-p)}$.

(ii) if $\big\|(\hat{\mathbf{v}}_{\alpha,\beta}^{p,q})_r\big\|_2^q \geq \|\mathbf{y}\|_2^2/\gamma_2$ *for some* $\gamma_2 > 0$*, it holds*

$$\frac{\big\|\big(\hat{\mathbf{v}}_{\alpha,\beta}^{p,q}\big)_r\big\|_q}{\big\|\big(\hat{\mathbf{v}}_{\alpha,\beta}^{p,q}\big)_r\big\|_2} < \left(\frac{\gamma_2}{\beta}\right)^{1/q}, \tag{4.29}$$

implying $(\hat{\mathbf{v}}_{\alpha,\beta}^{p,q})_r \in K_{\hat{s}_2}^{q,n_2}$ *with* $\hat{s}_2 = (\gamma_2/\beta)^{2/(2-q)}$.

Proof. Using the definition of $\mathcal{J}_{\alpha,\beta}^{p,q,R}$ and that $\big((\hat{\mathbf{u}}_{\alpha,\beta}^{p,q})_1, \dots, (\hat{\mathbf{v}}_{\alpha,\beta}^{p,q})_R\big)$ is a global minimizer thereof, we observe

$$\begin{aligned}\alpha \sum_{r=1}^{R} \big\|\big(\hat{\mathbf{u}}_{\alpha,\beta}^{p,q}\big)_r\big\|_p^p + \beta \sum_{r=1}^{R} \big\|\big(\hat{\mathbf{v}}_{\alpha,\beta}^{p,q}\big)_r\big\|_q^q &\leq \mathcal{J}_{\alpha,\beta}^{p,q,R}\left(\big(\hat{\mathbf{u}}_{\alpha,\beta}^{p,q}\big)_1, \dots, \big(\hat{\mathbf{v}}_{\alpha,\beta}^{p,q}\big)_R\right) \\ &\leq \mathcal{J}_{\alpha,\beta}^{p,q,R}(\mathbf{0}, \dots, \mathbf{0}) = \|\mathbf{y}\|_2^2.\end{aligned} \tag{4.30}$$

The claim follows by employing the respective assumption. For the strict inequality note that at least two strictly positive terms appear on the left-hand side. □

Concerning the proposition above we want to append a few important comments. Evidently, we showed that the components $(\hat{\mathbf{u}}_{\alpha,\beta}^{p,q})_r$ and $(\hat{\mathbf{v}}_{\alpha,\beta}^{p,q})_r$ are either close to zero or provably effectively sparse with respect to their own sparsity promoting norm. The associated sparsity indicators $\hat{s}_1$ and $\hat{s}_2$ can be controlled by the parameters α and β. The larger they are, the smaller the respective indicator and thus the sparser the component vector is. This formalizes what was already addressed in the discussion after Proposition 4.1, yielding a trade-off between data fidelity and structure in form of sparsity. To control this trade-off, a reasonable and careful choice of the regularization parameters is

indispensable. Moreover, although the previous result only makes a statement about effective sparsity, in most numerical examples exact sparsity is achieved. In fact, a smaller regularizing quasi-norm parameter q effects stronger sparsity. We will give numerical evidence thereof in Figure 5.3 in the following chapter.
It remains to treat low-rankness. By construction, the functional takes into consideration only matrices, or more precisely the $2R$ component vectors of their non-orthogonal rank-R decomposition, which are or result in matrices of at most rank R. This immediately ensures the desired low-rankness.
After having seen that even under minimal assumptions on the measurement process $\mathcal{A}$ global minimizers $\widehat{\mathbf{X}}_{\alpha,\beta}^{p,q}$ of the functional $\mathcal{J}_{\alpha,\beta}^{p,q,R}$ provide reasonable approximations to the simultaneously sparse and low-rank matrix $\mathbf{X}$, we want to answer the question of how to approach such optima in the next section.

4.3 Alternating Ridge and Bridge or ℓ_0-Regression

The numerical algorithm we propose, ARBeR, attempts to find global minimizers of the highly non-convex multi-penalty functional $\mathcal{J}_{\alpha,\beta}^{p,q,R}$. Before providing a detailed formulation of the method a comment on naming is in order. The designation Alternating Ridge and Bridge or ℓ_0-Regression originates from the situation where we want to recover a low-rank matrix $\mathbf{X}$ with merely sparse right component vectors $\mathbf{v}_r$, i.e., $\mathbf{X} \in \mathcal{S}_{n_1,s_2}^R$. In this case we set $p = 2$ and let $0 < q \leq 1$. As a consequence, the intermediate vector-valued optimization problem (4.6) employs Tikhonov regularization, which is also known as ridge regression. In turn, problem (4.7) becomes an ℓ_q-regularized optimization problem, which is known as bridge regression, cf. [Tib96, Section 11], and which was suggested in [FF93] by Frank and Friedman. In the particular case $q = 1$ we obtain a LASSO problem as in (1.67) and ARBeR reduces to ATLAS. Moreover, we also extend our method to the limit case $q = 0$, where (4.7) becomes an ℓ_0-regularized optimization problem.
For the rest of this section let us turn back to the general situation.

4.3.1 A Formulation of the Numerical Algorithm

At first we want to formalize our method and describe how to tackle the occurring update rules for the component vectors. The former is done on the next page in the form of Algorithm 5. It remains to provide a solution strategy for the vector-valued subproblems (4.31) and (4.32) under the different choices of $g_{\mathbf{u}}$ and $g_{\mathbf{v}}$, respectively. Exemplarily, let us look at the update of the rth right component vector, which has the form

$$\hat{\mathbf{v}}_r^k = \underset{\hat{\mathbf{v}} \in \mathbb{R}^{n_2}}{\arg\min} \left\| \tilde{\mathbf{y}} - \tilde{\mathbf{A}} \hat{\mathbf{v}} \right\|_2^2 + g_{\mathbf{v}}(\hat{\mathbf{v}}), \tag{4.33}$$

where $\tilde{\mathbf{y}}$ contains the partial residual and $\tilde{\mathbf{A}} : \mathbb{R}^{n_1} \to \mathbb{R}^m$ parametrizes the action of the operator $\mathcal{A}$ for a fixed opposite, here, left component vector $\hat{\mathbf{u}}_r^k$.
Now, if $g_{\mathbf{v}}(\,\cdot\,) = \beta \|\cdot\|_2^2$, i.e., if we have the situation that no sparsity shall be promoted, the solution of the resulting Tikhonov regularization problem can be given explicitly by $\hat{\mathbf{v}}_r^k = \left(\beta \operatorname{Id} + \tilde{\mathbf{A}}^T \tilde{\mathbf{A}}\right)^{-1} \tilde{\mathbf{A}}^T \tilde{\mathbf{y}}$.
Elsewise, if $g_{\mathbf{v}}(\,\cdot\,) = \beta \|\cdot\|_q^q$ or $g_{\mathbf{v}}(\,\cdot\,) = \beta \|\cdot\|_0$, i.e., if we want to promote sparsity in the respective component, we have recourse to the tools developed in Subsection 1.4.3.

Algorithm 5 Alternating Ridge and Bridge or ℓ_0-Regression (ARBeR)

Input: Measurement operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$, measurements $\mathbf{y} \in \mathbb{R}^m$, rank R, regularizing (quasi)-norm parameters $0 \leq p, q \leq 2$, regularization parameters $\alpha, \beta > 0$ and number of iterations K and L, where K relates to the power factorization iterations and L to the iterations of the employed iterative thresholding method.

Output: Minimizer $\widehat{\mathbf{X}}^{p,q}_{\text{ARBeR}}$.

```
For $r \in [R]$ initialize $\hat{\mathbf{u}}_r^0 \in \mathbb{R}^{n_1}$ and $\hat{\mathbf{v}}_r^0 \in \mathbb{R}^{n_2}$ to be the $r$th leading left and right
singular vector of $\mathcal{A}^*(\mathbf{y})$, respectively. Moreover, set $k = 0$.
while $k \leq K$ and stopping criterion not fulfilled
    Set $k = k + 1$.
    for $r = 1, \dots, R$
        if $s_1 = n_1$ then
            Set the regularizer $g_{\mathbf{u}}$ to be the Tikhonov regularizer, i.e., $g_{\mathbf{u}}(\hat{\mathbf{u}}) = \alpha \|\hat{\mathbf{u}}\|_2^2$.
        else
            if $q = 0$ then
                Set the regularizer $g_{\mathbf{u}}$ to be the $\ell_0$-norm, i.e., $g_{\mathbf{u}}(\hat{\mathbf{u}}) = \alpha \|\hat{\mathbf{u}}\|_0$.
            else
                Set the regularizer $g_{\mathbf{u}}$ to be the $\ell_p$-(quasi)-norm, i.e., $g_{\mathbf{u}}(\hat{\mathbf{u}}) = \alpha \|\hat{\mathbf{u}}\|_p^p$.
            end if
        end if
        Update the $r$th left component vector by solving the appropriately regularized
        optimization problem,
            $$\hat{\mathbf{u}}_r^k = \arg\min_{\hat{\mathbf{u}} \in \mathbb{R}^{n_1}} \Big\| \mathbf{y} - \mathcal{A}\Big( \sum_{\rho < r} \hat{\mathbf{u}}_\rho^k (\hat{\mathbf{v}}_\rho^k)^T \Big) - \mathcal{A}\big( \hat{\mathbf{u}} (\hat{\mathbf{v}}_r^{k-1})^T \big) - \mathcal{A}\Big( \sum_{\rho > r} \hat{\mathbf{u}}_\rho^{k-1} (\hat{\mathbf{v}}_\rho^{k-1})^T \Big) \Big\|_2^2 + g_{\mathbf{u}}(\hat{\mathbf{u}}). \tag{4.31}$$
        if $s_2 = n_2$ then
            Set the regularizer $g_{\mathbf{v}}$ to be the Tikhonov regularizer, i.e., $g_{\mathbf{v}}(\hat{\mathbf{v}}) = \beta \|\hat{\mathbf{v}}\|_2^2$.
        else
            if $q = 0$ then
                Set the regularizer $g_{\mathbf{v}}$ to be the $\ell_0$-norm, i.e., $g_{\mathbf{v}}(\hat{\mathbf{v}}) = \beta \|\hat{\mathbf{v}}\|_0$.
            else
                Set the regularizer $g_{\mathbf{v}}$ to be the $\ell_q$-(quasi)-norm, i.e., $g_{\mathbf{v}}(\hat{\mathbf{v}}) = \beta \|\hat{\mathbf{v}}\|_q^q$.
            end if
        end if
        Update the $r$th right component vector by solving the appropriately regularized
        optimization problem,
            $$\hat{\mathbf{v}}_r^k = \arg\min_{\hat{\mathbf{v}} \in \mathbb{R}^{n_2}} \Big\| \mathbf{y} - \mathcal{A}\Big( \sum_{\rho < r} \hat{\mathbf{u}}_\rho^k (\hat{\mathbf{v}}_\rho^k)^T \Big) - \mathcal{A}\big( \hat{\mathbf{u}}_r^k \hat{\mathbf{v}}^T \big) - \mathcal{A}\Big( \sum_{\rho > r} \hat{\mathbf{u}}_\rho^{k-1} (\hat{\mathbf{v}}_\rho^{k-1})^T \Big) \Big\|_2^2 + g_{\mathbf{v}}(\hat{\mathbf{v}}). \tag{4.32}$$
    end for
end while
Set $\widehat{\mathbf{X}}^{p,q}_{\text{ARBeR}} = \sum_{r=1}^R \hat{\mathbf{u}}_r^k (\hat{\mathbf{v}}_r^k)^T$.
```

In this case, iterative thresholding algorithms can be employed to approximate the solution to vector-valued optimization problem. Iterative soft thresholding (ISTA), cf. Paragraph 1.4.3(1), can be used to approach solutions to the resulting LASSO problem if $q = 1$. In the case $0 < q < 1$ the emerging ℓ_q-regularized non-convex minimization problem can be overcome with non-smooth iterative bridge-q thresholding, cf. Paragraph 1.4.3(2). Lastly, an approximate solution to the ℓ_0-regularized non-convex minimization problem can be found using iterative hard thresholding, cf. Paragraph 1.4.3(3).
Naturally, this works analogously for updating the left component vectors.

4.3.2 On Theoretical Convergence of the Algorithm

Having a numerical procedure at hand which provides reasonable grounds to believe that it can find minimizers of $\mathcal{J}_{\alpha,\beta}^{p,q,R}$ we are curious about its theoretical convergence guarantees. In this subsection we want to sketch how to prove local convergence to global optima and global convergence to stationary points making avail of the framework presented in [ABRS10]. To do so, we slightly modify the update rules (4.6) and (4.7) from the alternating scheme by complementing them with terms, which assure convergence from a theoretical point of view, however, turn out to be empirically unnecessary. This results in

$$\begin{aligned}\hat{\mathbf{u}}_r^k = \operatorname*{arg\,min}_{\hat{\mathbf{u}}\in\mathbb{R}^{n_1}} & \left\| \mathbf{y} - \mathcal{A}\Big(\sum_{\rho<r} \hat{\mathbf{u}}_\rho^k (\hat{\mathbf{v}}_\rho^k)^T\Big) - \mathcal{A}\big(\hat{\mathbf{u}}(\hat{\mathbf{v}}_r^{k-1})^T\big) - \mathcal{A}\Big(\sum_{\rho>r} \hat{\mathbf{u}}_\rho^{k-1} (\hat{\mathbf{v}}_\rho^{k-1})^T\Big) \right\|_2^2 \\ & + \alpha \|\hat{\mathbf{u}}\|_p^p + \tfrac{1}{2\mu_r^k} \left\|\hat{\mathbf{u}} - \hat{\mathbf{u}}_r^{k-1}\right\|_2^2 \end{aligned} \tag{4.34}$$

and

$$\begin{aligned}\hat{\mathbf{v}}_r^k = \operatorname*{arg\,min}_{\hat{\mathbf{v}}\in\mathbb{R}^{n_2}} & \left\| \mathbf{y} - \mathcal{A}\Big(\sum_{\rho<r} \hat{\mathbf{u}}_\rho^k (\hat{\mathbf{v}}_\rho^k)^T\Big) - \mathcal{A}\big(\hat{\mathbf{u}}_r^k \hat{\mathbf{v}}^T\big) - \mathcal{A}\Big(\sum_{\rho>r} \hat{\mathbf{u}}_\rho^{k-1} (\hat{\mathbf{v}}_\rho^{k-1})^T\Big) \right\|_2^2 \\ & + \beta \|\hat{\mathbf{v}}\|_q^q + \tfrac{1}{2\nu_r^k} \left\|\hat{\mathbf{v}} - \hat{\mathbf{v}}_r^{k-1}\right\|_2^2 \end{aligned} \tag{4.35}$$

with suitably chosen positive sequences $(\mu_r^k)_{k\geq 1}$ and $(\nu_r^k)_{k\geq 1}$ for all $r \in [R]$.
A generalization of the setting from [ABRS10], which was also done in [FMN19, Subsection 3.4], now addresses the convergence analysis of an alternating minimization approach to a function L of the form

$$\begin{cases} L(\tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{v}}_R) = Q(\tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{v}}_R) + \sum_{r=1}^R f_r(\tilde{\mathbf{u}}_r) + \sum_{r=1}^R g_r(\tilde{\mathbf{v}}_r), \\ f_r : \mathbb{R}^{n_1} \to \mathbb{R} \cup \{\infty\}, g_r : \mathbb{R}^{n_2} \to \mathbb{R} \cup \{\infty\} \text{ are proper lower semi-continuous for } r \in [R], \\ Q : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_2} \to \mathbb{R} \text{ is continuously differentiable,} \\ \nabla Q \text{ is Lipschitz continuous on bounded subsets of } \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_2}. \end{cases} \tag{4.36}$$

It is straightforward to see how to fit our problem into this framework. Then, applying the adaptations of Theorems 3.2 and 3.3 from [ABRS10], which were conducted in [Mal19, Appendix B.2], yields the following. If for an initial value $(\hat{\mathbf{u}}_1^0, \dots, \hat{\mathbf{v}}_R^0)$ and suitable positive sequences $(\mu_r^k)_{k\geq 1}$ and $(\nu_r^k)_{k\geq 1}$ it holds

$$\begin{cases} \inf L > -\infty, \\ L(\cdot, \hat{\mathbf{u}}_2^0 \dots, \hat{\mathbf{v}}_R^0) \text{ is proper,} \\ \text{for some positive } r_- < r_+ \text{ the sequences } (\mu_1^k)_{k\geq 1}, \dots, (\nu_R^k)_{k\geq 1} \text{ belong to } (r_-, r_+), \end{cases} \tag{4.37}$$

and if L fulfills the so-called Kurdyka-Łojasiewicz property at the global minimum of the function L, the sequence generated by

$$\begin{cases} \hat{\mathbf{u}}_r^k = \arg\min_{\hat{\mathbf{u}}\in\mathbb{R}^{n_1}} L(\hat{\mathbf{u}}_1^k,\dots,\hat{\mathbf{u}}_{r-1}^k,\hat{\mathbf{u}},\hat{\mathbf{u}}_{r+1}^{k-1},\dots,\hat{\mathbf{u}}_R^{k-1},\hat{\mathbf{v}}_1^k,\dots,\hat{\mathbf{v}}_{r-1}^k,\hat{\mathbf{v}}_r^{k-1},\hat{\mathbf{v}}_{r+1}^{k-1},\dots,\hat{\mathbf{v}}_R^{k-1}) \\ \qquad\qquad + \frac{1}{2\mu_r^k}\left\|\hat{\mathbf{u}}-\hat{\mathbf{u}}_r^{k-1}\right\|_2^2, \\ \hat{\mathbf{v}}_r^k = \arg\min_{\hat{\mathbf{v}}\in\mathbb{R}^{n_2}} L(\hat{\mathbf{u}}_1^k,\dots,\hat{\mathbf{u}}_{r-1}^k,\hat{\mathbf{u}}_r^k,\hat{\mathbf{u}}_{r+1}^{k-1},\dots,\hat{\mathbf{u}}_R^{k-1},\hat{\mathbf{v}}_1^k,\dots,\hat{\mathbf{v}}_{r-1}^k,\hat{\mathbf{v}},\hat{\mathbf{v}}_{r+1}^{k-1},\dots,\hat{\mathbf{v}}_R^{k-1}) \\ \qquad\qquad + \frac{1}{2\nu_r^k}\left\|\hat{\mathbf{v}}-\hat{\mathbf{v}}_r^{k-1}\right\|_2^2, \end{cases} \tag{4.38}$$

when iterating over r in an interior and k in an outer loop, converges to the global minimum. Moreover, if the Kurdyka-Łojasiewicz property holds at each point of the domain, the sequence $(\hat{\mathbf{u}}_1^k,\dots,\hat{\mathbf{v}}_R^k)$ either tends to infinity or converges to a stationary point of L.

Roughly speaking, the idea behind the function L having the Kurdyka-Łojasiewicz property at a global minimum is to assure the existence of a continuous concave function φ with which the range can be reparameterized in a way that the composition $\varphi\circ L$ has a kink in the minimum and increases steeply around it. For a precise definition we refer to [ABRS10, Definition 3.2].

In order to apply these results to our setting, i.e., to $L = \mathcal{J}_{\alpha,\beta}^{p,q,R}$, and therefore providing evidence for the assertions about local and global convergence from the beginning, we need to verify that $\mathcal{J}_{\alpha,\beta}^{p,q,R}$ fulfills the requirements, foremost the Kurdyka-Łojasiewicz property. In fact, all but this property can be checked straightforwardly. To show the remaining condition we make use of a result from algebraic geometry, stating that semialgebraic functions have the Kurdyka-Łojasiewicz property [BDLS07]. For this approach, however, we restrict ourselves to the case where the parameters p and q are rational. A function is called semialgebraic if its graph can be written as a finite union of sets of the form

$$\{\mathbf{z}\in\mathbb{R}^d : \rho_s(\mathbf{z})=0,\ \varrho_t(\mathbf{z})>0,\ s\in[S],\ t\in[T]\} \tag{4.39}$$

with real polynomials ρ_s and ϱ_t. It is easy to see that polynomials and the absolute value of one component of a vector, i.e., the mapping $\mathbf{z}\mapsto|z_i|$, are semialgebraic, see, e.g., [FMN19, pages 14–15]. Moreover, if $q=a/b\in\mathbb{Q}$, the function $h:\mathbb{R}_+\to\mathbb{R}_+$, $z\mapsto z^q$ is semialgebraic since

$$\operatorname{graph}(h)=\{(z,r)\in\mathbb{R}\times\mathbb{R} : z^a-r^b=0,\ z>0,\ r>0\}\cup\{(z,r)\in\mathbb{R}\times\mathbb{R} : z=0,\ r=0\}. \tag{4.40}$$

Finally, since also compositions, finite sums and finite products of semialgebraic functions turn out to be semialgebraic, the semialgebraicity of $\mathcal{J}_{\alpha,\beta}^{p,q,R}$ follows as

$$\mathcal{J}_{\alpha,\beta}^{p,q,R}(\tilde{\mathbf{u}}_1,\dots,\tilde{\mathbf{v}}_R)=\sum_{\ell=1}^m\left|y_\ell-\sum_{r=1}^R\left\langle\mathbf{A}_\ell,\tilde{\mathbf{u}}_r\tilde{\mathbf{v}}_r^T\right\rangle_F\right|^2+\alpha\sum_{r=1}^R\sum_{i=1}^{n_1}|\tilde{u}_{ri}|^p+\beta\sum_{r=1}^R\sum_{j=1}^{n_2}|\tilde{v}_{rj}|^q \tag{4.41}$$

can be rewritten as a finite sum, composition and finite product of semialgebraic functions.

4.3.3 Initialization of the Method

Generally speaking, non-convex methods and optimal initialization in combination are a difficult task as already addressed in Section 3.3. For this reason, we only propose a heuristic in this subsection, which does not claim to be ideal.

First of all, note that the seemingly natural initialization with R zero vectors, which is often the first choice in applications, is not advisable, as $(\mathbf{0}, \dots, \mathbf{0})$ is a stationary point of the method.

We, in turn, suggest to initialize ARBeR with the R leading left and right singular vectors of $\mathcal{A}^*(\mathbf{y})$, where $\mathcal{A}^* : \mathbb{R}^m \to \mathbb{R}^{n_1 \times n_2}$ denotes the adjoint of the measurement operator $\mathcal{A}$ and is given by

$$\mathcal{A}^*(\mathbf{y}) = \sum_{\ell=1}^{m} y_\ell \mathbf{A}_\ell. \tag{4.42}$$

Numerical tests, which, on the one hand, give evidence that our proposed initialization is plausible, yet, on the other hand, point out that it is not optimal, will be provided in Subsection 5.1.3. The latter will be verified by comparing the performance to the one obtained when using the left and right component vectors of the solution $\mathbf{X}$ as initialization, which are, needless to say, not accessible in practice.

Lastly, we want to propose a novel multilevel-type strategy, which is in particular designed for very small regularizing quasi-norm parameters p and/or q. The core reasoning of this idea rests on the common belief that initialization becomes more decisive the more severely non-convex the optimization problem is in some sense. As the parameters p and q are readily accessible and provide the sole external way to have an impact on the degree of non-convexity, they can be increased in order to alleviate non-convexity. However, since we are primarily interested in minimizers of the multi-penalty functional $\mathcal{J}_{\alpha,\beta}^{p,q,R}$ with the original parameters p and q, we propose to construct a finite and in both components non-increasing sequence $(p_\lambda, q_\lambda)_{\lambda=1}^{\Lambda}$ such that $(p_\Lambda, q_\Lambda) = (p, q)$ and (p_1, q_1) are both sufficiently large. Starting from $\lambda = 1$, for each tuple (p_λ, q_λ) we perform our algorithm ARBeR with the respective parameters p_λ and q_λ, which are gradually decreased. At each level λ the left and right component vectors of the obtained solution $\widehat{\mathbf{X}}_{\text{ARBeR}}^{p_{\lambda-1},q_{\lambda-1}}$ are used as initialization. The first stage is initialized as described above. This approach will also be investigated numerically in Subsection 5.1.3. It has to be mentioned, however, that this initialization method may be costly, as ARBeR is additionally performed $\Lambda - 1$ times.

4.4 A Restricted Isometry Property

So far, in former Section 4.2, global minimizers $\widehat{\mathbf{X}}_{\alpha,\beta}^{p,q}$ of the multi-penalty functional $\mathcal{J}_{\alpha,\beta}^{p,q,R}$ were analyzed merely indirectly by examining properties such as data fidelity and structure. In the following, however, we want to provide the necessary tools to upper bound the approximation error directly. Therefore, we introduce two particular types of matrix models together with suitably designed restricted isometry properties, which generalize the ones from [FMN19] and are supposed to be the proper ones to be considered when recovering simultaneously (effectively) sparse and low-rank matrices possessing a non-orthogonal low-rank decomposition.

To begin with, we restrict the set $\mathcal{S}_{s_1,s_2}^{R}$ of (s_1, s_2)-sparse rank-R matrices from equation (3.7) such that the norm of their quasi-singular values is bounded by $\Gamma \geq 1$. That means we define the set

$$\mathcal{S}_{s_1,s_2}^{R,\Gamma} = \Big\{ \mathbf{Z} = \sum_{r=1}^{R} \sigma_r \mathbf{u}_r \mathbf{v}_r^T \in \mathcal{S}_{s_1,s_2}^{R} : \ \|\boldsymbol{\sigma}\|_2 \leq \Gamma \Big\}. \tag{4.43}$$

Based upon this matrix model we want to introduce a relaxation by replacing sparsity with effective sparsity. Our analysis will be built upon this set. For $\Gamma \geq 1$, the set $\mathcal{K}_{s_1,s_2}^{p,q,R,\Gamma}$ of (ℓ_p, ℓ_q)-effectively (s_1, s_2)-sparse rank-R matrices is defined as

$$\begin{aligned}\mathcal{K}_{s_1,s_2}^{p,q,R,\Gamma} = \big\{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2} : \exists\ \mathbf{u}_1, \ldots, \mathbf{u}_R \in K_{s_1}^{p,n_1},\ \mathbf{v}_1, \ldots, \mathbf{v}_R \in K_{s_2}^{q,n_2},\ \text{and } \boldsymbol{\sigma} \in \mathbb{R}^R,& \\ \text{s.t. } \mathbf{Z} = \sum_{r=1}^{R} \sigma_r \mathbf{u}_r \mathbf{v}_r^T,& \\ \text{where } \|\mathbf{u}_r\|_2 = \|\mathbf{v}_r\|_2 = 1\ \forall r \in [R], \text{ and } \|\boldsymbol{\sigma}\|_2 \leq \Gamma\big\}.&\end{aligned} \tag{4.44}$$

A decomposition of the form $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ as in (4.44) with $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\sigma})$ is called effectively sparse decomposition of $\mathbf{X} \in \mathcal{K}_{s_1,s_2}^{p,q,R,\Gamma}$.
Regarding the former definition a few remarks are in order. First, the cases $p = 2$ or $q = 2$ are not explicitly excluded, in turn, they can be understood as having no effective sparsity in the respective component vectors. Naturally, the same holds for $s_1 = n_1$ or $s_2 = n_2$. Second, the restriction $\Gamma \geq 1$ is a natural condition as explained in [FMN19, Remark 4.6]. And third, the properties discussed after Definition 4.3 can be transferred immediately to the set $\mathcal{K}_{s_1,s_2}^{p,q,R,\Gamma}$, meaning that we have $\mathcal{S}_{s_1,s_2}^{R,\Gamma} \subset \mathcal{K}_{s_1,s_2}^{p,q,R,\Gamma}$ as well as monotonicity in the components p, q, s_1 and s_2 if the others are fixed. Evidently, we also have monotonicity in the rank R and in Γ.
The key property of this set is closedness under matrix addition to some extent. More precisely, for $\mathbf{Z} = \sum_{r=1}^{R} \sigma_r \mathbf{u}_r \mathbf{v}_r^T \in \mathcal{K}_{s_1,s_2}^{p,q,R,\Gamma}$ and $\widetilde{\mathbf{Z}} = \sum_{r=1}^{\widetilde{R}} \tilde{\sigma}_r \tilde{\mathbf{u}}_r \tilde{\mathbf{v}}_r^T \in \mathcal{K}_{\tilde{s}_1,\tilde{s}_2}^{\tilde{p},\tilde{q},\widetilde{R},\widetilde{\Gamma}}$ it holds

$$\mathbf{Z} - \widetilde{\mathbf{Z}} \in \mathcal{K}_{\max\{s_1,\tilde{s}_1\},\max\{s_2,\tilde{s}_2\}}^{\max\{p,\tilde{p}\},\max\{q,\tilde{q}\},R+\widetilde{R},\sqrt{\Gamma^2+\widetilde{\Gamma}^2}}. \tag{4.45}$$

This can be seen directly by noting that

$$\mathbf{Z} - \widetilde{\mathbf{Z}} = \sum_{r=1}^{R+\widetilde{R}} \sigma_r \mathbf{u}_r \mathbf{v}_r^T \tag{4.46}$$

is an effectively sparse decomposition of $\mathbf{Z} - \widetilde{\mathbf{Z}}$, when setting $\sigma_{R+r} = \tilde{\sigma}_r$, $\mathbf{u}_{R+r} = -\tilde{\mathbf{u}}_r$ and $\mathbf{v}_{R+r} = \tilde{\mathbf{v}}_r$ for all $r \in [\widetilde{R}]$. The crucial point for this to work is that we dispense with orthogonality of the component vectors. Motivated by the framework of sparse principal component analysis [ZHT06], this was put forward in [FMN19] as was a corresponding restricted isometry property, which we propose to generalize as follows.

Definition 4.5 (Additive Low-Rank and (Effectively) Sparse Restricted Isometry Property)**.** *A linear operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ satisfies the additive rank-R and (ℓ_p, ℓ_q)-effectively (s_1, s_2)-sparse Γ-restricted isometry property (RIP$_\Gamma$) with isometry constant $0 < \delta < 1$ if*

$$\big| \|\mathcal{A}(\mathbf{Z})\|_2^2 - \|\mathbf{Z}\|_F^2 \big| \leq \delta \tag{4.47}$$

for all $\mathbf{Z} \in \mathcal{K}_{s_1,s_2}^{p,q,R,\Gamma}$. If equation (4.47) only holds for all $\mathbf{Z} \in \mathcal{S}_{s_1,s_2}^{R,\Gamma}$, the operator $\mathcal{A}$ has the weaker additive rank-R and (s_1, s_2)-sparse RIP$_\Gamma$.

Clearly, our restricted isometry property distinguishes itself from the familiar ones, such as, e.g., Definition 3.7. We emphasize this by calling it additive. The indispensability of this modification can be directly traced back to the non-orthogonality of the sparse decomposition. In oder to reason this let us anticipate what will be developed in the subsequent section, namely that, with high probability, Gaussian measurement ensembles have our additive restricted isometry property, if the number of measurements scales, up to polylogarithmic factors, as $m \gtrsim R(s_1+s_2)$. The hidden constant depends on the diameter of the considered matrix set. If we would have not given up the scaling invariance, this would not be achievable, as we sketch now following Remark 3.14 from [FMN19]. To do so, let us for simplicity consider the set $\mathcal{K}^{2,q,2,\Gamma}_{n_1,s_2}$ of rank-2 matrices with merely sparse right component vectors in the setting $n_1 \approx s_2 \ll n_2$. Moreover, let $\mathbf{u} \in \mathbb{R}^{n_1}$ and $\mathbf{v}_1 \in \mathbb{R}^{n_2}$ denote unit norm vectors with $\|\mathbf{v}_1\|_q \leq {s_2}^{1/q-1/2}/2$ and let us define the vector $\mathbf{v}_2 = -\mathbf{v}_1 + \epsilon\mathbf{w}$ for any $\mathbf{w} \in \mathbb{R}^{n_2}$ and $\epsilon > 0$ small enough such that $\mathbf{v}_2 \in K^{q,n_2}_{s_2}$. Then, for $\Gamma = \max\{1, (1+\|\mathbf{v}_2\|_2)/2\}$, we observe that

$$\mathbf{Z} = \frac{1}{2}\epsilon\mathbf{u}\mathbf{w}^T = \frac{1}{2}\mathbf{u}\mathbf{v}_1^T + \frac{1}{2}\mathbf{u}\mathbf{v}_2^T \in \mathcal{K}^{2,q,2,\Gamma}_{n_1,s_2}, \tag{4.48}$$

which is a rank-1 matrix without any particular further structure. Yet, by being small in size, it admits an effectively rank-2 sparse decomposition. Now, if our restricted isometry property was of the form

$$(1-\delta)\|\mathbf{Z}\|_F^2 \leq \|\mathcal{A}(\mathbf{Z})\|_2^2 \leq (1+\delta)\|\mathbf{Z}\|_F^2, \tag{4.49}$$

i.e., scaling invariant, it would immediately hold for any rank-1 matrix, also the ones without an effectively sparse right component vector. This, however, demands an information theoretic lower bound on the required number of measurements of order $\mathcal{O}(n_2)$, which is substantially worse than $\mathcal{O}(R(n_1+s_2))$.
Let us now continue with the main result of this section, which provides an upper bound on the approximation error $\|\mathbf{X} - \widehat{\mathbf{X}}^{p,q}_{\alpha,\beta}\|_F$ for global minimizers of the multi-penalty functional under the assumption of our restricted isometry property. This generalizes Theorem 3.7 from [FMN19].

Theorem 4.6 (Approximation Quality for Global Minimizers under Additive RIP)**.** *Let us assume that $\mathcal{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ has the additive rank-$2R$ and (ℓ_p, ℓ_q)-effectively $\big(\max\{s_1, (\gamma_1/\alpha)^{2/(2-p)}\}, \max\{s_2, (\gamma_2/\beta)^{2/(2-q)}\}\big)$-sparse*[9] *$(c+1)\Gamma$-restricted isometry property with constant $0 < \delta < 1$ for a fixed choice of $\gamma_1, \gamma_2 > 0$ and $c \geq 1$.*
Then, for any $\mathbf{X} \in \mathcal{K}^{p,q,R,\Gamma}_{s_1,s_2}$, a global minimizer $\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta}$ of $\mathcal{J}^{p,q,R}_{\alpha,\beta}$ with noisy measurements $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\eta}$ is an element of $\mathcal{K}^{p,q,R,c\Gamma}_{\hat{s}_1,\hat{s}_2}$ with[10] *$\hat{s}_1 = (\gamma_1/\alpha)^{2/(2-p)}$ and $\hat{s}_2 = (\gamma_2/\beta)^{2/(2-q)}$ and moreover fulfills*

$$\|\mathbf{X} - \widehat{\mathbf{X}}^{p,q}_{\alpha,\beta}\|_F \lesssim \sqrt{s_1^{\frac{q(2-p)}{2(p+q)}} s_2^{\frac{p(2-q)}{2(p+q)}} R^{1-\frac{pq}{2(p+q)}}} (\alpha^q\beta^p)^{\frac{1}{2(p+q)}} \|\mathbf{X}\|^{\frac{pq}{2(p+q)}}_{\frac{pq}{p+q}} + 2\|\boldsymbol{\eta}\|_2 + \sqrt{\delta}, \tag{4.50}$$

if the following assumptions on the global optimizer hold. First, denoting its rth left and its rth right component vector by $(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_r$ and $(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_r$, respectively, it have to hold

(i) $\big\|\big(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta}\big)_r\big\|_2^p \geq \big(\|\mathbf{X}\|_F + \sqrt{\delta} + \|\boldsymbol{\eta}\|_2\big)^2/\gamma_1$ *and*

[9] In case that $p = 2$ or that $(\gamma_1/\alpha)^{2/(2-p)}$ exceeds n_1, we replace the expression $(\gamma_1/\alpha)^{2/(2-p)}$ by n_1. This is handled analogously for $q = 2$ and $(\gamma_2/\beta)^{2/(2-q)}$, which is replaced by n_2 in case.

[10] See footnote 9.

(ii) $\left\|\left(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta}\right)_r\right\|_2^q \geq \left(\|\mathbf{X}\|_F + \sqrt{\delta} + \|\boldsymbol{\eta}\|_2\right)^2/\gamma_2$

for all $r \in [R]$*. Second, denoting the vector of quasi-singular values by* $\hat{\boldsymbol{\sigma}}^{p,q}_{\alpha,\beta}$*, it has to hold* $\left\|\hat{\boldsymbol{\sigma}}^{p,q}_{\alpha,\beta}\right\|_2 \leq c\Gamma$*.*

Proof. Let us abbreviate $(\gamma_1/\alpha)^{2/(2-p)}$ by $\hat{s}_1$ and $(\gamma_2/\beta)^{2/(2-q)}$ by $\hat{s}_2$ and let us set $\hat{s}_1 = n_1$ if $p = 2$ and $\hat{s}_2 = n_2$ if $q = 2$.

Then, by means of triangle inequality and the additive restricted isometry property, which applies since $\mathbf{X} \in \mathcal{K}^{p,q,R,\Gamma}_{s_1,s_2} \subset \mathcal{K}^{p,q,2R,(c+1)\Gamma}_{\max\{s_1,\hat{s}_1\},\max\{s_2,\hat{s}_2\}}$, we observe

$$\|\mathbf{y}\|_2 \leq \|\mathcal{A}(\mathbf{X})\|_2 + \|\boldsymbol{\eta}\|_2 \leq \sqrt{\|\mathbf{X}\|_F^2 + \delta} + \|\boldsymbol{\eta}\|_2 \leq \|\mathbf{X}\|_F + \sqrt{\delta} + \|\boldsymbol{\eta}\|_2. \tag{4.51}$$

Now, Proposition 4.4 can be applied. Utilizing the assumptions on the component vectors of the global minimizer $\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta}$, the requirements of the proposition can be assured as we have

$$\left\|(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_r\right\|_2^p \geq \left(\|\mathbf{X}\|_F + \sqrt{\delta} + \|\boldsymbol{\eta}\|_2\right)^2/\gamma_1 \geq \|\mathbf{y}\|_2^2/\gamma_1, \tag{4.52}$$

where we made use of (4.51) in the last inequality. $\left\|(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_r\right\|_2^q$ can be bound analogously. Hence, we obtain $(\hat{\mathbf{u}}^{p,q}_{\alpha,\beta})_r \in K^{p,n_1}_{\hat{s}_1}$ and $(\hat{\mathbf{v}}^{p,q}_{\alpha,\beta})_r \in K^{q,n_2}_{\hat{s}_2}$ for all $r \in [R]$. Integrating the assumption on the vector of quasi-singular values of the global minimizer we conclude that $\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta} \in \mathcal{K}^{p,q,R,c\Gamma}_{\hat{s}_1,\hat{s}_2}$, which verifies the last claim.

In combination with $\mathbf{X} \in \mathcal{K}^{p,q,R,\Gamma}_{s_1,s_2}$, the closedness of the matrix set $\mathcal{K}$ under matrix addition from (4.45) shows

$$\mathbf{X} - \widehat{\mathbf{X}}^{p,q}_{\alpha,\beta} \in \mathcal{K}^{p,q,2R,(c+1)\Gamma}_{\max\{s_1,\hat{s}_1\},\max\{s_2,\hat{s}_2\}}. \tag{4.53}$$

Therefore, the additive rank-$2R$ and (ℓ_p, ℓ_q)-effectively $(\max\{s_1,\hat{s}_1\}, \max\{s_2,\hat{s}_2\})$-sparse $(c+1)\Gamma$-restricted isometry property can be applied, yielding

$$\begin{aligned}\left\|\mathbf{X} - \widehat{\mathbf{X}}^{p,q}_{\alpha,\beta}\right\|_F &\leq \sqrt{\left\|\mathcal{A}(\mathbf{X}) - \mathcal{A}\big(\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta}\big)\right\|_2^2 + \delta} \leq \left\|\mathcal{A}(\mathbf{X}) - \mathcal{A}\big(\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta}\big)\right\|_2 + \sqrt{\delta} \\ &\leq \left\|\mathbf{y} - \mathcal{A}\big(\widehat{\mathbf{X}}^{p,q}_{\alpha,\beta}\big)\right\|_2 + \|\boldsymbol{\eta}\|_2 + \sqrt{\delta}.\end{aligned} \tag{4.54}$$

The first term in the last line is the measurement misfit, which can be controlled by Proposition 4.1. This results in

$$\begin{aligned}\left\|\mathbf{X} - \widehat{\mathbf{X}}^{p,q}_{\alpha,\beta}\right\|_F &\leq \sqrt{\|\boldsymbol{\eta}\|_2^2 + C_{pq}\left(\alpha^q\beta^p\right)^{\frac{1}{p+q}} \sum_{r=1}^R \left(\|\mathbf{u}_r\|_p \|\mathbf{v}_r\|_q\right)^{\frac{pq}{p+q}}} + \|\boldsymbol{\eta}\|_2 + \sqrt{\delta} \\ &\leq \sqrt{C_{pq} \sum_{r=1}^R \left(\|\mathbf{u}_r\|_p \|\mathbf{v}_r\|_q\right)^{\frac{pq}{p+q}} \left(\alpha^q\beta^p\right)^{\frac{1}{2(p+q)}}} + 2\|\boldsymbol{\eta}\|_2 + \sqrt{\delta},\end{aligned} \tag{4.55}$$

for which it remains to upper bound $\sum_{r=1}^R \left(\|\mathbf{u}_r\|_p \|\mathbf{v}_r\|_q\right)^{\frac{pq}{p+q}}$. By exploiting that $\mathbf{u}_r \in K^{p,n_1}_{s_1}$ and $\mathbf{v}_r \in K^{q,n_2}_{s_2}$ for all $r \in [R]$ in the first inequality, this can be done as follows. Namely,

$$\begin{aligned}\sum_{r=1}^R \left(\|\mathbf{u}_r\|_p \|\mathbf{v}_r\|_q\right)^{\frac{pq}{p+q}} &\leq s_1^{\frac{q(2-p)}{2(p+q)}} s_2^{\frac{p(2-q)}{2(p+q)}} \sum_{r=1}^R \left(\|\mathbf{u}_r\|_2 \|\mathbf{v}_r\|_2\right)^{\frac{pq}{p+q}} \\ &\leq c_{\underline{\mathbf{U}}}^{-\frac{pq}{p+q}} s_1^{\frac{q(2-p)}{2(p+q)}} s_2^{\frac{p(2-q)}{2(p+q)}} R^{1-\frac{pq}{2(p+q)}} \|\mathbf{X}\|_{\frac{pq}{p+q}}^{\frac{pq}{p+q}},\end{aligned} \tag{4.56}$$

where the last inequality makes use of relation (4.9) noting that $0 < pq/(p+q) < 2$. This concludes the proof. □

Before turning towards the conclusive question of how to design measurement operators fulfilling our additive restricted isometry property and the related question of the required number of measurements, let us spend a few words on the preceding theorem.

First, we observe that by choosing the regularization parameters aligned with the noise-to-signal ratio, i.e., $\alpha = \beta = \|\boldsymbol{\eta}\|_2^2 / \|\mathbf{X}\|_{pq/(p+q)}^{pq/(p+q)}$, we obtain a reduced bound of the form

$$\|\mathbf{X} - \widehat{\mathbf{X}}_{\alpha,\beta}^{p,q}\|_F \lesssim \left(\sqrt{s_1^{\frac{q(2-p)}{2(p+q)}} s_2^{\frac{p(2-q)}{2(p+q)}} R^{1-\frac{pq}{2(p+q)}}} + 2 \right) \|\boldsymbol{\eta}\|_2 + \sqrt{\delta} \tag{4.57}$$

in place of (4.50), which resembles typical compressed sensing bounds.

Second, and returning to the general statement, we identify three individual terms contributing to the approximation error. Two of them are separate terms covering the noise level and the RIP constant. And one depends entirely on structural properties of the signal we want to recover and on the regularizing parameters of the multi-penalty functional. We note that this latter term could be made vanish if we would let the regularization parameters α and β become arbitrarily small. This, however, is not possible for already known reasons, see, for instance, the discussion after Proposition 4.1. The previous theorem adds yet another aspect why small regularization parameters are not practicable. Namely, if α or β is getting too small, the restricted isometry property degenerates in the sense that it is required to hold for an exceedingly large class of matrices. This can be seen by noting that, for decreasing parameters α and β, $\hat{s}_1 = (\gamma_1/\alpha)^{2/(2-p)}$ and $\hat{s}_2 = (\gamma_2/\beta)^{2/(2-q)}$ increase. Obviously, this is independent from one another. According to footnote 9 they are limited by n_1 and n_2, respectively. In this case, however, quantities of the ambient dimension get involved, what leads to not being able to take advantage of any parsimony of the original model with respect to this structure.

Third, let us have a closer look at the restricted isometry property. Limiting ourselves to the situation of matrices with merely sparse right component vectors, we set $p = 2$ and let $0 < q \leq 1$. Moreover, we assume that the singular value decomposition is a sparse decomposition, which makes the prefactor depending on R disappear. This can be seen by comparing the, in this case, identity $\|\mathbf{X}\|_q^q = \sum_{r=1}^R (\|\mathbf{u}_r\|_2 \|\mathbf{v}_r\|_2)^q$ to (4.9). We want to firstly understand the relation between s_2 and $\hat{s}_2$, i.e., between the original and the recovered sparsity parameter. Therefore, to obtain an upper bound of the type $\mathcal{O}(\|\mathbf{X}\| + \|\boldsymbol{\eta}\| + \sqrt{\delta})$ for a suitable (quasi)-norm $\|\cdot\|$, we need to choose $\beta = \mathcal{O}\big(s_2^{(q-2)/(2+q)}\big)$, having made the, in the spirit of Lemma 4.2, beneficial assumption $\alpha = \beta$. With this choice we derive $\hat{s}_2 = \mathcal{O}\big(s_2^{2/(2+q)}\big)$ and are able to conclude that the matrix set for which the requested restricted isometry property has to hold contains only matrices with effectively s_2-sparse right component vectors. A similar computation, however, without the assumption $\alpha = \beta$, can be carried out for sparsity in both components. Suitable choices of the parameters yield $\hat{s}_i = \mathcal{O}(s_i)$ for $i = 1, 2$.

And fourth, forming the transition to the subsequent section, we observe that the respective restricted isometry property is easier to be fulfilled the smaller the parameters p and q are. This goes back to the monotonicity property of the set of (ℓ_p, ℓ_q)-effectively (s_1, s_2)-sparse rank-R matrices. The quantification of this observation is the content of the remainder of this chapter.

Before that, let us also refer to the discussion after the more special version of Theorem 4.6, Theorem 3.7 in [FMN19], for further aspects of this result.

4.5 The Required Number of Measurements

One of the most intriguing questions concerning the recovery of structured matrices with non-orthogonal sparse decomposition of low rank from linear measurements was not answered so far. Namely, how many measurements there are required to ensure provably good approximation and how corresponding measurement operators look like. By means of Theorem 4.6 this boils down to investigate which measurement operators have the additive restricted isometry property from Definition 4.5. We are peculiarly interested in the situations described by the matrix sets $\mathcal{S}_{s_1,s_2}^{R,\Gamma}$ and $\mathcal{S}_{n_1,s_2}^{R,\Gamma}$, i.e., the recovery of low-rank matrices with sparse left and right or merely sparse right component vectors. As an artifact of the proof technique used in former Sections 4.2 and 4.4 we introduced alleviated, i.e., larger, matrix sets, on which the analysis of global minimizers of the multi-penalty functional $\mathcal{J}_{\alpha,\beta}^{p,q,R}$ was built on.
In this section, we restrict ourselves slightly regarding the variability in the regularizing (quasi)-norm parameters. However, from an applied point of view, this causes no severe limitations. Following what was already mentioned in the outline at the end of Section 4.1, in the case of sparsity in both components we set $p = q$ and let $0 < q \leq 1$. Our analysis is then founded on the set $\mathcal{K}_{s_1,s_2}^{q,q,R,\Gamma}$, where $s_1 < n_1$ and $s_2 < n_2$. In turn, if only the right component is sparse we set $p = 2$ and let $0 < q \leq 1$ and consider the set $\mathcal{K}_{n_1,s_2}^{2,q,R,\Gamma}$, for which we note that it coincides with $\mathcal{K}_{n_1,s_2}^{q,q,R,\Gamma}$. Consequently, in analogy to the multi-penalty functional, we can introduce the reduced notation $\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}$ with $s_1 < n_1$ in the first and $s_1 = n_1$ in the second case.
Based on this set we will now work out which measurement operators have the associated additive rank-R and (ℓ_q, ℓ_q)-effectively (s_1, s_2)-sparse Γ-restricted isometry property by establishing the following result, which extends Lemma 3.12 from [FMN19]. For comparison, we provide an analogous result for the smaller set $\mathcal{S}_{s_1,s_2}^{R,\Gamma}$ and its associated weaker additive rank-R and (s_1, s_2)-sparse RIP_Γ.

Theorem 4.7 (Gaussian Measurement Ensembles have the Additive Restricted Isometry Property). *Let $\mathcal{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ be a Gaussian measurement ensemble and assume that*

$$m \geq C\left(\frac{\delta}{\Gamma^2 R}\right)^{-2} R\,(s_1+s_2+1)\log\left(\max\left\{e\sqrt{R}, \frac{en_1}{s_1}, \frac{en_2}{s_2}\right\}\right) \tag{4.58}$$

holds for a constant $C > 0$. Then, with probability at least $1 - 2\exp\big(-d(\delta/(\Gamma^2 R))m\big)$, where $d > 0$ denotes a constant, the operator $\frac{1}{\sqrt{m}}\mathcal{A}$ satisfies the additive rank-R and (s_1, s_2)-sparse Γ-restricted isometry property with isometry constant $0 < \delta < \Gamma^2 R$.
Moreover, let $0 < q \leq 1$ and assume that

$$m \geq C'\left(\frac{\delta}{\Gamma^2 R}\right)^{-2} R\left(\left(\left(\frac{s_1}{n_1}\right)^{2/q-2} s_1 + \left(\frac{s_2}{n_2}\right)^{2/q-2} s_2 + 1\right) + 144^{(2q-2)/(2-q)}(s_1+s_2)\right) \cdot \operatorname{polylog}\left(e\sqrt{R}\max\left\{\frac{n_1}{s_1}, \frac{n_2}{s_2}\right\}^{\max\{1/q-1/2,1\}}\right) \tag{4.59}$$

holds for a constant $C' > 0$. Then, with probability at least $1 - 2\exp\big(-d'(\delta/(\Gamma^2 R))m\big)$, where $d' > 0$ denotes a constant, the operator $\frac{1}{\sqrt{m}}\mathcal{A}$ satisfies the stronger additive rank-R

and (ℓ_q, ℓ_q)*-effectively* (s_1, s_2)*-sparse* Γ*-restricted isometry property with isometry constant* $0 < \delta < \Gamma^2 R$*.*

Before providing a proof thereof, what we will do, divided into several steps, in the remainder of this section, let us discuss this theorem.
Gaussian measurement operators are guaranteed to have the additive rank-R and ((ℓ_q, ℓ_q)-effectively) (s_1, s_2)-sparse Γ-restricted isometry property with isometry constant δ, if, up to polylogarithmic terms and neglecting on q dependent factors, $m \gtrsim \Delta^{-2}\mathcal{O}(R(s_1 + s_2))$ measurements are taken. Here, $0 < \Delta < 1$ denotes $\Delta = \delta/(\Gamma^2 R)$, where $\Gamma^2 R$ is the squared Frobenius radius of the matrix sets $\mathcal{S}^{R,\Gamma}_{s_1,s_2}$ and $\mathcal{K}^{q,R,\Gamma}_{s_1,s_2}$, respectively. That means, the required number of measurements depends essentially linearly on the intrinsic dimension of the signal. Let us now focus on the on q dependent prefactors, which come into play if $0 < q < 1$. First, the terms $(s_1/n_1)^{2/q-2}$ and $(s_2/n_2)^{2/q-2}$ are smaller the smaller the relative sparsities s_1/n_1 and s_2/n_2 are and thus reduce the necessary measurements. This effect is strengthened by smaller q's. Second, the factor $144^{(2q-2)/(2-q)}$ decreases if $q \to 0$, yet, it does not converge to zero. It is bounded from below by $1/144$. Third and contrarily, the polylogarithmic term, which includes first and third powers of the logarithm, is worsened by small effective sparsities and small q's.
Unfortunately, however, to the best of our knowledge, we are not aware of any information theoretical limits on the required number of measurements for our very general class of matrices $\mathcal{K}^{q,R,\Gamma}_{s_1,s_2}$, cf. [FMN19, Remark 3.13]. Information theoretical investigations regarding non-orthogonal multi-structured decompositions provide a direction for future research.

Remark 4.8 (Extension to Sub-Gaussian Measurement Ensemble)**.** The assertion of the previous theorem can be straightforwardly extended to the even larger class of random measurement operators with sub-Gaussian entries. Recall that a random variable ξ is called L-sub-Gaussian if it obeys the tail bound $\mathsf{P}(|\xi| \geq t) \leq 2\exp(-t^2/(2L^2))$ for all $t > 0$, i.e., its distribution is dominated by the distribution of a Gaussian random variable. L is up to an absolute constant equivalent to the sub-Gaussian norm $\|\xi\|_{\psi_2}$ of ξ, which is defined as $\|\xi\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathsf{E}|\xi|^p)^{1/p}$. Then, in the case that $\mathcal{A}$ is a sub-Gaussian measurement ensemble, Theorem 4.7 holds with modified constants d, C, d', and C' that depend on L. In fact, the proof we provide covers this case as well.

We now give a concise outline of the proof of Theorem 4.7, which we eventually give in Subsection 4.5.4. Our main ingredient therefore is a bound on suprema of chaos processes that was presented in [KMR14]. The paper considers random variables of the form

$$\sup_{\mathbf{H}\in\mathcal{H}} \left| \|\mathbf{H}\boldsymbol{\xi}\|_2^2 - \mathsf{E}\|\mathbf{H}\boldsymbol{\xi}\|_2^2 \right|, \tag{4.60}$$

where $\mathcal{H}$ denotes a set of matrices and $\boldsymbol{\xi}$ a random vector. By expanding the norms, one can show that (4.60) is the supremum of an order-2 chaos process. The authors established expectation and deviation bounds on this random variable, which we formulate in Subsection 4.5.3. This involves two complexity measures of the set $\mathcal{H}$. One of them is the radius of the matrix set $\mathcal{H}$ and the other is Talagrand's γ_2-functional, see, e.g., Definition 4.14 below. In order to upper bound the latter quantity, what we do in Subsection 4.5.2, we employ Dudley's inequality, which connects the functional to covering numbers. For this reason we start with computing the metric entropy of the sets $\mathcal{S}^{R,\Gamma}_{s_1,s_2}$ and $\mathcal{K}^{q,R,\Gamma}_{s_1,s_2}$ in the subsequent Subsection 4.5.1. Let us note that this last step is convenient as the covering

number is an elementary geometric quantity, however, it is not optimal. The reason is as follows. The proof of Dudley's inequality relies on a very elementary chaining method, resulting in a logarithmic gap, i.e., the bound is only sharp up to a logarithmic factor. Talagrand's γ_2-functional, in turn, is based on the more sophisticated generic chaining method, which takes the geometry better into account.
Before going into the details of the proof of Theorem 4.7, we want to highlight the analogy of the proof technique to the one of Theorems 1.11 and 2.11. In any case, we want to establish a concentration inequality for an expression of the type (4.60), where $\boldsymbol{\xi}$ parametrizes the measurement operator and $\mathcal{H}$ denotes the set of signals, for which we want to have a near-isometry. Typically, a concentration inequality for an individual signal serves as a starting point. In order to extend it to all signals, it is firstly established on a finite subset of the signal set, usually on a suitable ϵ-net. This step requires some sort of generalization from one signal to finitely many. In the proofs from the first two chapters this involved a union bound. In the following proof, however, we require the more elaborate tool in form of Theorem 4.17, which relies on a chaining argument. Lastly, the extension to the whole set is performed using a perturbation argument.

4.5.1 Metric Entropy of the Matrix Sets $\mathcal{S}_{s_1,s_2}^{R,\Gamma}$ and $\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}$

In order to establish bounds on the covering number and therefore the metric entropy of the matrix sets $\mathcal{S}_{s_1,s_2}^{R,\Gamma}$ and $\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}$, we require control of the covering numbers of the respective component vectors, i.e., the sets Σ_s^N and $K_s^{q,N}$. To this end let us investigate the geometry of the closely related set

$$\widetilde{K}_s^{q,N} = \left\{\mathbf{z} \in \mathbb{R}^N : \|\mathbf{z}\|_2 \le 1 \text{ and } \|\mathbf{z}\|_q \le s^{1/q-1/2}\right\}, \tag{4.61}$$

which was introduced in the case $q = 1$ in [PV13]. We notice that $K_s^{q,N} \cap \mathcal{B}_2^N \subset \widetilde{K}_s^{q,N}$ and define $\widetilde{\Sigma}_s^N = \Sigma_s^N \cap \mathcal{B}_2^N$ correspondingly. Because of this inclusion let us recall the monotonicity property of the covering number. For two sets obeying $K \subset \widetilde{K}$ it holds

$$N(K, \|\cdot\|, \epsilon) \le N(\widetilde{K}, \|\cdot\|, \epsilon/2). \tag{4.62}$$

This relation will help us later on to construct ϵ-nets of $K_s^{q,N} \cap \mathcal{B}_2^N$ from ϵ-nets of $\widetilde{K}_s^{q,N}$. To get there, we firstly show that an ϵ-net of the latter can be built from sparse vectors, whose relative sparsity is only slightly larger than the effective relative sparsity s/N of the set $\widetilde{K}_s^{q,N}$, which we want to cover. With this result we generalize Lemma 3.2 from [PV13]. To do so, we adapt a commonly employed proof technique in compressed sensing, which was already utilized in this form in Theorem 1.17.

Lemma 4.9 (Sparse Net of $\widetilde{K}_s^{q,N}$)**.** *Let $0 < q < 2$. Then, if $s \le t$, the set $\widetilde{\Sigma}_t^N \cap \widetilde{K}_s^{q,N}$ is an $(s/t)^{1/q-1/2}$-net of $\widetilde{K}_s^{q,N}$ with respect to the normed space $(\mathbb{R}^N, \|\cdot\|_2)$.*

Proof. Let $\mathbf{z} \in \widetilde{K}_s^{q,N}$ and define a partition

$$\mathcal{T} = \{T_\ell : |T_\ell| = t \text{ for all } \ell < \lfloor N/t \rfloor\}_{\ell=0}^{\lfloor N/t \rfloor} \tag{4.63}$$

of $[N]$ associated with a nonincreasing rearrangement of $\mathbf{z}$, i.e., for all $\ell \ge 1$ it holds

$$|z_i| \le |z_j| \quad \text{for all } i \in T_\ell \text{ and } j \in T_{\ell-1}. \tag{4.64}$$

By construction, $\mathbf{z}|_{T_0} \in \widetilde{\Sigma}_t^N \cap \widetilde{K}_s^{q,N}$ and $\|\mathbf{z} - \mathbf{z}|_{T_0}\|_2 = \big\|\mathbf{z}|_{\cup_{\ell \geq 1} T_\ell}\big\|_2$, which remains to be upper bounded. Therefore, by definition of the partition $\mathcal{T}$, we observe that for $\ell \geq 1$ it also holds $|z_i|^q \leq |z_j|^q$ for all $i \in T_\ell$ and $j \in T_{\ell-1}$. Firstly, for $\ell \geq 2$, summation over j yields

$$|z_i| \leq t^{-1/q} \big\|\mathbf{z}|_{T_{\ell-1}}\big\|_q \tag{4.65}$$

for all $i \in T_\ell$. Secondly, taking the ℓ_2-norm over $i \in T_\ell$ subsequently shows

$$\|\mathbf{z}|_{T_\ell}\|_2 \leq t^{1/2-1/q} \big\|\mathbf{z}|_{T_{\ell-1}}\big\|_q. \tag{4.66}$$

With this we conclude that

$$\begin{aligned} \|\mathbf{z} - \mathbf{z}|_{T_0}\|_2^q &= \big\|\mathbf{z}|_{\cup_{\ell\geq 1} T_\ell}\big\|_2^q \leq \sum_{\ell \geq 1} \|\mathbf{z}|_{T_\ell}\|_2^q \\ &\leq \sum_{\ell \geq 1} t^{q/2-1} \big\|\mathbf{z}|_{T_{\ell-1}}\big\|_q^q \leq t^{q/2-1} \|\mathbf{z}\|_q^q \leq t^{q/2-1} s^{1-q/2}, \end{aligned} \tag{4.67}$$

where the first inequality follows from A.2(i) with $p = 1$ and $q/2$ instead of q when considering the vector $\big(\|\mathbf{z}|_{T_\ell}\|_2^2\big)_{\ell \geq 1}$. The last one exploits that $\mathbf{z} \in \widetilde{K}_s^{q,N}$. □

Required to apply this result is the metric entropy of the space of sparse vectors. To this end, let us cite the following lemma, which bounds this quantity by exploiting that $\widetilde{\Sigma}_s^N$ is a union of $\binom{N}{s}$ s-dimensional unit balls, a property which is inherited from the space Σ_s^N. For each low-dimensional ball we can then make use of the well-known bound

$$N(\mathcal{B}_2^s, \|\cdot\|_2, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^s, \tag{4.68}$$

which relies on a standard volume comparison argument, see, e.g., [Pis89, Lemma 4.16]. In fact, it can be seen easily by noting that $N(\mathcal{B}_2^s, \|\cdot\|_2, \epsilon) \leq M(\mathcal{B}_2^s, \|\cdot\|_2, \epsilon)$, where $M(K, \|\cdot\|, \epsilon)$ denotes the ϵ-packing number of a set K. It is the largest cardinality of any set $K^{\boxplus} \subset K$ such that for all $\mathbf{z}_1, \mathbf{z}_2 \in K^{\boxplus}$ it holds $\|\mathbf{z}_1 - \mathbf{z}_2\| > \epsilon$. Such a set is called ϵ-packing. Since balls of radius $\epsilon/2$ and centered in the points of a maximal ϵ-packing are mutually disjoint and contained in the Minkowski sum $K + \mathcal{B}_2^s(\mathbf{0}, \epsilon/2)$, we obtain $M(\mathcal{B}_2^s, \|\cdot\|_2, \epsilon) \lambda^s(\mathcal{B}_2^s(\mathbf{0}, \epsilon/2)) \leq \lambda^s(\mathcal{B}_2^s(\mathbf{0}, 1 + \epsilon/2))$, from what we can deduce the result.

Lemma 4.10 (Metric Entropy of $\widetilde{\Sigma}_s^N$, [PV13, Lemma 3.3])**.** *Let $0 < \epsilon < 1$ and $1 \leq s \leq N$. Then, for the metric entropy of $\widetilde{\Sigma}_s^N$ it holds*

$$\log N(\widetilde{\Sigma}_s^N, \|\cdot\|_2, \epsilon) \leq s \log\left(\frac{3eN}{\epsilon s}\right). \tag{4.69}$$

We can now combine the former two lemmas to upper bound the metric entropy of the set $\widetilde{K}_s^{q,N}$. This extends Lemma 3.4 from [PV13] in the following way.

Lemma 4.11 (Metric Entropy of $\widetilde{K}_s^{q,N}$)**.** *Let $0 < q < 2$, $0 < \epsilon < 1$ and $1 \leq s \leq N$. Then, for the metric entropy of $\widetilde{K}_s^{q,N}$ it holds*

$$\begin{aligned} \log N(\widetilde{K}_s^{q,N}, \|\cdot\|_2, \epsilon) &\leq \begin{cases} N \log\left(\frac{5}{\epsilon}\right) & \text{if } \epsilon \in (0, 2\left(\frac{s}{N}\right)^{1/q-1/2}], \\ s\left(\frac{2}{\epsilon}\right)^{2q/(2-q)} \log\left(\frac{6eN}{s}\left(\frac{\epsilon}{2}\right)^{(3q-2)/(2-q)}\right) & \text{else,} \end{cases} \\ &\lesssim s \left(\frac{2}{\epsilon}\right)^{2q/(2-q)} \log\left(\frac{eN}{s}\right), \end{aligned} \tag{4.70}$$

where the hidden constant may depend on q.

Proof. First of all, for any $\epsilon \in (0,1)$ we derive a straightforward upper bound from the covering number of the unit ball $\mathcal{B}_2^N$. As $\widetilde{K}_s^{q,N} \subset \mathcal{B}_2^N$, it holds

$$N(\widetilde{K}_s^{q,N}, \|\cdot\|_2, \epsilon) \leq N(\mathcal{B}_2^N, \|\cdot\|_2, \epsilon/2) \leq \left(1 + \frac{2}{\epsilon/2}\right)^N \leq \left(\frac{5}{\epsilon}\right)^N, \tag{4.71}$$

where the first inequality uses the monotonicity property (4.62) of the covering number and the next-to-last follows from (4.68).

Now, let us assume that $\epsilon \in (2\,(s/N)^{1/q-1/2}, 1)$, which assures that for $t := s\,(2/\epsilon)^{2q/(2-q)}$ it holds $t \leq N$. Moreover, since $t \geq s$, according to Lemma 4.9, $\widetilde{\Sigma}_t^N \cap \widetilde{K}_s^{q,N}$ is an $\epsilon/2$-net of $\widetilde{K}_s^{q,N}$. In combination with the monotonicity property (4.62) of the covering number and Lemma 4.10, applied with $\epsilon/4$ and t, we observe that $\widetilde{\Sigma}_t^N \cap \widetilde{K}_s^{q,N}$ admits an $\epsilon/2$-net $\mathcal{N}$ with $|\mathcal{N}| \leq \left(\frac{12eN}{\epsilon t}\right)^t$. Using triangle inequality, we conclude that $\mathcal{N}$ is an ϵ-net of $\widetilde{K}_s^{q,N}$. □

This lemma shows that the metric entropy of $\widetilde{K}_s^{q,N}$ is equivalent to the one of $\widetilde{\Sigma}_s^N$ regarding the dependency on the intrinsic and ambient dimension.

Building on Lemmas 4.10 and 4.11, we can now derive the metric entropy of the matrix sets $\mathcal{S}_{s_1,s_2}^{R,\Gamma}$ and $\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}$. We start with the former by providing the following result.

Lemma 4.12 (Metric Entropy of $\mathcal{S}_{s_1,s_2}^{R,\Gamma}$, [FMN19, Lemma 4.2])**.** *Let $0 < \epsilon < 6\Gamma\sqrt{R}$ and $1 \leq s_i \leq n_i$ for $i = 1, 2$. Then, for the metric entropy of $\mathcal{S}_{s_1,s_2}^{R,\Gamma}$ it holds*

$$\log N(\mathcal{S}_{s_1,s_2}^{R,\Gamma}, \|\cdot\|_F, \epsilon) \leq R(s_1 + s_2 + 1)\log\left(\frac{18\Gamma R}{\epsilon}\right) + Rs_1 \log\left(\frac{en_1}{s_1}\right) + Rs_2 \log\left(\frac{en_2}{s_2}\right). \tag{4.72}$$

In order to prove this, the cited paper modifies the proof of Lemma 3.1 from [CP11]. We will now use the same proof technique to establish the following extension of Lemma 4.4 from [FMN19]. Actually, by replacing the respective sets and covering numbers, the proof of Lemma 4.12 can be recovered.

Lemma 4.13 (Metric Entropy of $\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}$)**.** *Let $0 < q < 2$, $0 < \epsilon < 6\Gamma\sqrt{R}$ and $1 \leq s_i \leq n_i$ for $i = 1, 2$. Furthermore, without loss of generality assume that $s_2/n_2 \leq s_1/n_1$. Then, for the metric entropy of $\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}$ it holds*

$$\begin{aligned}
&\log N(\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}, \|\cdot\|_F, \epsilon) \\
&\quad \leq R\log N(\widetilde{K}_{s_1}^{q,n_1}, \|\cdot\|_2, \tfrac{\epsilon}{6\Gamma\sqrt{R}}) + \log N(\mathcal{B}_2^R(\mathbf{0},\Gamma), \|\cdot\|_2, \tfrac{\epsilon}{6R}) + R\log N(\widetilde{K}_{s_2}^{q,n_2}, \|\cdot\|_2, \tfrac{\epsilon}{6\Gamma\sqrt{R}}) \\
&\quad \leq \begin{cases}
R(n_1+n_2+1)\log\left(\frac{30\Gamma R}{\epsilon}\right) & \textit{if } \epsilon \in I_{02}, \\
Rs_2\left(\frac{12\Gamma\sqrt{R}}{\epsilon}\right)^{2q/(2-q)} \log\left(\frac{6en_2}{s_2}\left(\frac{\epsilon}{12\Gamma\sqrt{R}}\right)^{(3q-2)/(2-q)}\right) & \\
\quad +R(n_1+1)\log\left(\frac{30\Gamma R}{\epsilon}\right) & \textit{if } \epsilon \in I_{21}, \\
Rs_1\left(\frac{12\Gamma\sqrt{R}}{\epsilon}\right)^{2q/(2-q)} \log\left(\frac{6en_1}{s_1}\left(\frac{\epsilon}{12\Gamma\sqrt{R}}\right)^{(3q-2)/(2-q)}\right) & \\
\quad +Rs_2\left(\frac{12\Gamma\sqrt{R}}{\epsilon}\right)^{2q/(2-q)} \log\left(\frac{6en_2}{s_2}\left(\frac{\epsilon}{12\Gamma\sqrt{R}}\right)^{(3q-2)/(2-q)}\right) & \\
\quad +R\log\left(\frac{30\Gamma R}{\epsilon}\right) & \textit{else,}
\end{cases}
\end{aligned} \tag{4.73}$$

where $I_{02} = 12\Gamma\sqrt{R}(0,(s_2/n_2)^{1/q-1/2}]$ *and* $I_{21} = 12\Gamma\sqrt{R}((s_2/n_2)^{1/q-1/2},(s_1/n_1)^{1/q-1/2}]$ *abbreviate the intervals.*

Proof. Inspired by the relaxation (4.61) of Definition 4.3, let us define the matrix set

$$\begin{aligned}\widetilde{\mathcal{K}}^{q,R,\Gamma}_{s_1,s_2} = \big\{\widetilde{\mathbf{Z}} \in \mathbb{R}^{n_1\times n_2} : \exists\ \tilde{\mathbf{u}}_1,\dots,\tilde{\mathbf{u}}_R \in \widetilde{K}^{q,n_1}_{s_1},\ \tilde{\mathbf{v}}_1,\dots,\tilde{\mathbf{v}}_R \in \widetilde{K}^{q,n_2}_{s_2},\\ \text{and } \widetilde{\mathbf{\Sigma}} \in \mathbb{R}^{R\times R} \text{ diagonal with } \|\widetilde{\mathbf{\Sigma}}\|_F \le \Gamma,\ \text{s.t. } \widetilde{\mathbf{Z}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^T\big\}\end{aligned} \tag{4.74}$$

and note that $\mathcal{K}^{q,R,\Gamma}_{s_1,s_2} \subset \widetilde{\mathcal{K}}^{q,R,\Gamma}_{s_1,s_2}$. Then, by the monotonicity property (4.62) it holds $N(\mathcal{K}^{q,R,\Gamma}_{s_1,s_2},\|\cdot\|_F,\epsilon) \le N(\widetilde{\mathcal{K}}^{q,R,\Gamma}_{s_1,s_2},\|\cdot\|_F,\epsilon/2)$ and it remains to find an $\epsilon/2$-net of $\widetilde{\mathcal{K}}^{q,R,\Gamma}_{s_1,s_2}$. To this end, with respect to $\|\cdot\|_2$, let $(\widetilde{K}^{q,n_1}_{s_1})^\#$ and $(\widetilde{K}^{q,n_2}_{s_2})^\#$ denote minimal $\epsilon/(6\Gamma\sqrt{R})$-nets of $\widetilde{K}^{q,n_1}_{s_1}$ and $\widetilde{K}^{q,n_2}_{s_2}$, respectively. Furthermore, with respect to $\|\cdot\|_F$, let $(\mathcal{D}_\Gamma)^\#$ denote a minimal $\epsilon/(6R)$-net of the set of diagonal $R\times R$ matrices with Frobenius norm bounded by Γ. For the covering number of this set, according to (4.68), it holds $N(\mathcal{D}_\Gamma,\|\cdot\|_F,\epsilon) = N(\mathcal{B}^R_2(\mathbf{0},\Gamma),\|\cdot\|_2,\epsilon) \le (3\Gamma/\epsilon)^R$. To conclude, we show that the set

$$\begin{aligned}\mathcal{K}^\# = \big\{\widetilde{\mathbf{Z}}^\# \in \mathbb{R}^{n_1\times n_2} : \exists\ (\tilde{\mathbf{u}}_1)^\#,\dots,(\tilde{\mathbf{u}}_R)^\# \in \big(\widetilde{K}^{q,n_1}_{s_1}\big)^\#,\ (\tilde{\mathbf{v}}_1)^\#,\dots,(\tilde{\mathbf{v}}_R)^\# \in \big(\widetilde{K}^{q,n_2}_{s_2}\big)^\#,\\ \text{and } \widetilde{\mathbf{\Sigma}}^\# \in \big(\mathcal{D}_\Gamma\big)^\#,\ \text{s.t. } \widetilde{\mathbf{Z}}^\# = \widetilde{\mathbf{U}}^\#\widetilde{\mathbf{\Sigma}}^\#\big(\widetilde{\mathbf{V}}^\#\big)^T\big\}\end{aligned} \tag{4.75}$$

is an $\epsilon/2$-net of $\widetilde{\mathcal{K}}^{q,R,\Gamma}_{s_1,s_2}$. Therefore, let $\widetilde{\mathbf{Z}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^T \in \widetilde{\mathcal{K}}^{q,R,\Gamma}_{s_1,s_2}$. Then, first, for any $r \in [R]$ choose $(\tilde{\mathbf{u}}_r)^\# \in (\widetilde{K}^{q,n_1}_{s_1})^\#$, i.e., such that $\|\tilde{\mathbf{u}}_r-(\tilde{\mathbf{u}}_r)^\#\|_2 \le \epsilon/(6\Gamma\sqrt{R})$ and analogously choose $(\tilde{\mathbf{v}}_r)^\# \in (\widetilde{K}^{q,n_2}_{s_2})^\#$. Furthermore, select $\widetilde{\mathbf{\Sigma}}^\# \in (\mathcal{D}_\Gamma)^\#$ such that $\|\widetilde{\mathbf{\Sigma}}-\widetilde{\mathbf{\Sigma}}^\#\|_F \le \epsilon/(6R)$. With this we obtain

$$\begin{aligned}\|\widetilde{\mathbf{Z}}-\widetilde{\mathbf{Z}}^\#\|_F &\le \|(\widetilde{\mathbf{U}}-\widetilde{\mathbf{U}}^\#)\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^T\|_F+\|\widetilde{\mathbf{U}}^\#(\widetilde{\mathbf{\Sigma}}-\widetilde{\mathbf{\Sigma}}^\#)\widetilde{\mathbf{V}}^T\|_F+\|\widetilde{\mathbf{U}}^\#\widetilde{\mathbf{\Sigma}}^\#(\widetilde{\mathbf{V}}-\widetilde{\mathbf{V}}^\#)^T\|_F\\ &\le \Gamma\|\widetilde{\mathbf{U}}-\widetilde{\mathbf{U}}^\#\|_F+R\|\widetilde{\mathbf{\Sigma}}-\widetilde{\mathbf{\Sigma}}^\#\|_F+\Gamma\|\widetilde{\mathbf{V}}-\widetilde{\mathbf{V}}^\#\|_F\\ &\le \Gamma\epsilon/(6\Gamma)+R\epsilon/(6R)+\Gamma\epsilon/(6\Gamma) = \epsilon/2,\end{aligned} \tag{4.76}$$

having used the submultiplicativity of the Frobenius norm, $\|\widetilde{\mathbf{U}}^\#\|_F^2 = \sum_{r=1}^R\|(\tilde{\mathbf{u}}_r)^\#\|_2^2 \le R$ (analogously for $\widetilde{\mathbf{V}}$) and $\|\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^T\|_F^2 = \sum_{r=1}^R\|(\widetilde{\Sigma})_{rr}\tilde{\mathbf{v}}_r\|_2^2 \le \Gamma^2$ (analogously for $\widetilde{\mathbf{U}}^\#\widetilde{\mathbf{\Sigma}}^\#$) in the second inequality. The last inequality involves the properties of the nets and uses $\|\widetilde{\mathbf{U}}-\widetilde{\mathbf{U}}^\#\|_F^2 = \sum_{r=1}^R\|\tilde{\mathbf{u}}_r-(\tilde{\mathbf{u}}_r)^\#\|_2^2$ (analogously for $\|\widetilde{\mathbf{V}}-\widetilde{\mathbf{V}}^\#\|_F^2$). An application of Lemma 4.11 then yields the claim as

$$N(\widetilde{\mathcal{K}}^{q,R,\Gamma}_{s_1,s_2},\|\cdot\|_F,\epsilon/2) \le \big|\mathcal{K}^\#\big| \le \big|(\widetilde{K}^{q,n_1}_{s_1})^\#\big|^R\big|(\mathcal{D}_\Gamma)^\#\big|\big|(\widetilde{K}^{q,n_2}_{s_2})^\#\big|^R. \tag{4.77}$$

□

4.5.2 An Upper Bound on Talagrand's γ_2-Functional

In order to derive a sharp upper bound on the expectation

$$\mathsf{E}\sup_{\mathbf{H}\in\mathcal{H}}\xi_{\mathbf{H}} \tag{4.78}$$

of the supremum of a stochastic process $(\xi_{\mathbf{H}})_{\mathbf{H}\in\mathcal{H}}$, which is indexed by a general metric space, an improved concept of chaining was expounded in [Tal05]. At the heart of this generic chaining methods stand the following quantity.

Definition 4.14 (Talagrand's γ_2-Functional, [Ver18, Definition 8.5.1])**.** *Let $(\mathcal{H}, d)$ be a metric space. A sequence of subsets $(\mathcal{H}_k)_{k=0}^{\infty}$ of $\mathcal{H}$ is called an admissible sequence if the cardinalities of $\mathcal{H}_k$ satisfy $|\mathcal{H}_0| = 1$ and $|\mathcal{H}_k| = 2^{2^k}$ for all $k \geq 1$. The γ_2-functional of $\mathcal{H}$ is then defined as*

$$\gamma_2 \left(\mathcal{H}, d\right) = \inf_{(\mathcal{H}_k)_k} \sup_{\mathbf{H} \in \mathcal{H}} \sum_{k=0}^{\infty} 2^{k/2} d(\mathbf{H}, \mathcal{H}_k), \tag{4.79}$$

where the infimum is with respect to all admissible sequences.

Let us assume that the random variables $\xi_{\mathbf{H}}$ are mean-zero with sub-Gaussian increments, i.e., $\|\xi_{\mathbf{H}_1} - \xi_{\mathbf{H}_2}\|_{\psi_2} \lesssim d(\mathbf{H}_1, \mathbf{H}_2)$ for all $\mathbf{H}_1, \mathbf{H}_2 \in \mathcal{H}$. Then it can be shown, see, e.g., [Ver18, Theorem 8.5.3], that

$$\mathsf{E} \sup_{\mathbf{H} \in \mathcal{H}} \xi_{\mathbf{H}} \lesssim \gamma_2 \left(\mathcal{H}, d\right). \tag{4.80}$$

This bound is an improvement of Dudley's inequality, which can be recovered by observing that

$$\gamma_2(\mathcal{H}, d) \lesssim \int_0^{\infty} \sqrt{\log N(\mathcal{H}, d, \epsilon)} \, \mathrm{d}\epsilon. \tag{4.81}$$

Note that the upper bound can be replaced by the diameter $\operatorname{diam}(\mathcal{H})$ of $\mathcal{H}$ in the metric d. Moreover, if $\mathcal{H}$ is symmetric, i.e., $\mathcal{H} = -\mathcal{H}$, we can even use the radius, i.e., the quantity $d(\mathcal{H}) = \sup_{\mathbf{H} \in \mathcal{H}} d(\mathbf{H}, \mathbf{0})$, as an upper bound. Regarding notation, if the metric is induced by a norm $\| \cdot \|$, we write $\gamma_2(\mathcal{H}, \| \cdot \|)$ for Talagrand's γ_2-functional. For the radius of the set $\mathcal{H}$ with respect to some norm, we write $d(\mathcal{H})$ and add a subscript to indicate the norm. For instance, we write $d_F(\mathcal{H})$ for the radius in the Frobenius norm and $d_\infty(\mathcal{H})$ for the radius in the spectral norm.

As we face matrix-indexed random variables of the form $\xi_{\mathbf{H}} = \|\mathbf{H}\boldsymbol{\xi}\|_2^2 - \mathsf{E}\|\mathbf{H}\boldsymbol{\xi}\|_2^2$ for the two very special situations associated with the matrix sets $\mathcal{S}_{s_1,s_2}^{R,\Gamma}$ and $\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}$, we require bounds on the quantities mentioned above for these two cases. Note that $\boldsymbol{\xi}$ is an associated random vector containing the randomness, which will consist of i.i.d. mean-zero sub-Gaussian entries of unit variance in our setting.

Let us therefore fit our formulation of the restricted isometry property into the setting of this subsection. To this end, recall that we are interested in the failure probability

$$\mathsf{P} \left(\sup_{\mathbf{Z}} \left| \left\| \frac{1}{\sqrt{m}} \mathcal{A}(\mathbf{Z}) \right\|_2^2 - \|\mathbf{Z}\|_F^2 \right| \geq \delta \right), \tag{4.82}$$

where the supremum is with respect to one of the two sets $\mathcal{S}_{s_1,s_2}^{R,\Gamma}$ and $\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}$. In order to reformulate this properly let us define the random vector $\boldsymbol{\xi}_{\mathcal{A}} \in \mathbb{R}^{mn_1n_2}$ associated with a measurement operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ by

$$\boldsymbol{\xi}_{\mathcal{A}} = \begin{pmatrix} \operatorname{vec}(\mathbf{A}_1) \\ \vdots \\ \operatorname{vec}(\mathbf{A}_m) \end{pmatrix}, \tag{4.83}$$

where the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_m$ are as described after equation (4.1). In turn, for the matrix $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$ we define an $m \times m$-block diagonal matrix $\mathbf{H}_{\mathbf{Z}} \in \mathbb{R}^{m \times mn_1n_2}$ by

$$\mathbf{H}_{\mathbf{Z}} = \frac{1}{\sqrt{m}} \begin{pmatrix} \operatorname{vec}(\mathbf{Z})^T & \mathbf{0} & \cdots \\ & \ddots & \\ \cdots & \mathbf{0} & \operatorname{vec}(\mathbf{Z})^T \end{pmatrix}. \tag{4.84}$$

We observe that $\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{Z}) = \mathbf{H}_{\mathbf{Z}}\boldsymbol{\xi}_{\mathcal{A}}$. Building upon this let us introduce the auxiliary matrix set $\mathcal{H}_{\mathcal{S}^{R,\Gamma}_{s_1,s_2}} = \{\mathbf{H}_{\mathbf{Z}} : \mathbf{Z} \in \mathcal{S}^{R,\Gamma}_{s_1,s_2}\}$ and define the set $\mathcal{H}_{\mathcal{K}^{q,R,\Gamma}_{s_1,s_2}}$ analogously. Since we will formulate statements in the following, which hold for both auxiliary matrix sets, we will write $\mathcal{H}$ abbreviatorily, if we address both situations. The mapping $\mathbf{Z} \mapsto \mathbf{H}_{\mathbf{Z}}$ is an isometric linear bijection and we have $\|\mathbf{H}_{\mathbf{Z}}\|_F = \|\mathbf{Z}\|_F$ and $\|\mathbf{H}_{\mathbf{Z}}\| = \|\mathbf{Z}\|_F/\sqrt{m}$. Furthermore, assuming that the entries of $\boldsymbol{\xi}_{\mathcal{A}}$ are i.i.d. with zero mean and unit variance, we can verify that $\mathsf{E}\|\mathbf{H}_{\mathbf{Z}}\boldsymbol{\xi}_{\mathcal{A}}\|_2^2 = \|\mathbf{H}_{\mathbf{Z}}\|_F^2 = \|\mathbf{Z}\|_F^2$.

From these properties we can directly deduce upper bounds on the two quantities $d_F(\mathcal{H})$ and $d_\infty(\mathcal{H})$. Therefore, let $\mathbf{Z} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ denote an (effectively) sparse decomposition and note that $\|\mathbf{Z}\|_F = \|\mathbf{U}\boldsymbol{\Sigma}\|_F\|\mathbf{V}\|_F \le \Gamma\sqrt{R}$. Then, $d_F(\mathcal{H}) = \sup_{\mathbf{H}_{\mathbf{Z}}\in\mathcal{H}}\|\mathbf{H}_{\mathbf{Z}}\|_F \le \Gamma\sqrt{R}$ and $d_\infty(\mathcal{H}) = \sup_{\mathbf{H}_{\mathbf{Z}}\in\mathcal{H}}\|\mathbf{H}_{\mathbf{Z}}\| \le \Gamma\sqrt{R}/\sqrt{m}$.

The last ingredient before applying the bound on suprema of chaos processes, which we formulate in Theorem 4.17, is an upper bound on Talagrand's γ_2-functional. For this reason we compute Dudley's integral, i.e., the right-hand side of (4.81), for the two cases $\mathcal{H}_{\mathcal{S}^{R,\Gamma}_{s_1,s_2}}$ and $\mathcal{H}_{\mathcal{K}^{q,R,\Gamma}_{s_1,s_2}}$.

Lemma 4.15 (A Bound on Dudley's Integral for $\mathcal{H}_{\mathcal{S}^{R,\Gamma}_{s_1,s_2}}$, [FMN19, Lemma 7.1])**.** *For $\Gamma \ge 1$ it holds*

$$\int_0^{d_{\mathcal{S}}} \sqrt{\log N(\mathcal{H}_{\mathcal{S}^{R,\Gamma}_{s_1,s_2}}, \|\cdot\|, \epsilon)}\,\mathrm{d}\epsilon \lesssim \sqrt{\frac{\Gamma^2 R^2 (s_1+s_2+1)\log\left(\max\left\{e\sqrt{R}, \frac{en_1}{s_1}, \frac{en_2}{s_2}\right\}\right)}{m}}, \tag{4.85}$$

where $d_{\mathcal{S}}$ abbreviates $d_\infty\big(\mathcal{H}_{\mathcal{S}^{R,\Gamma}_{s_1,s_2}}\big)$.

The proof of the next lemma resembles to one of the former, however, it requires much more involved computations, which will be outsourced to Appendix A.2.

Lemma 4.16 (A Bound on Dudley's Integral for $\mathcal{H}_{\mathcal{K}^{q,R,\Gamma}_{s_1,s_2}}$)**.** *Let $0 < q \le 1$. For $\Gamma \ge 1$ it holds*

$$\int_0^{d_{\mathcal{K}}} \sqrt{\log N(\mathcal{H}_{\mathcal{K}^{q,R,\Gamma}_{s_1,s_2}}, \|\cdot\|, \epsilon)}\,\mathrm{d}\epsilon \lesssim \sqrt{\frac{\Gamma^2 R^2\Big(\Big(\big(\frac{s_1}{n_1}\big)^{2/q-2}s_1 + \big(\frac{s_2}{n_2}\big)^{2/q-2}s_2 + 1\Big) + C(q)(s_1+s_2)\Big)}{m}} \cdot \sqrt{\operatorname{polylog}\Big(e\sqrt{R}\max\left\{\frac{n_1}{s_1}, \frac{n_2}{s_2}\right\}^{\max\{1/q-1/2,1\}}\Big)}, \tag{4.86}$$

where $d_{\mathcal{K}}$ abbreviates $d_\infty\big(\mathcal{H}_{\mathcal{K}^{q,R,\Gamma}_{s_1,s_2}}\big)$, $C(q) = 144^{(2q-2)/(2-q)}$ is decreasing in q and the hidden constant is independent of q.

Proof. As the mapping $\mathbf{Z} \mapsto \mathbf{H}_{\mathbf{Z}}$ is bijective and an isometry with $\|\mathbf{H}_{\mathbf{Z}}\| = \|\mathbf{Z}\|_F/\sqrt{m}$, it holds $N(\mathcal{H}_{\mathcal{K}^{q,R,\Gamma}_{s_1,s_2}}, \|\cdot\|, \epsilon) = N(\mathcal{K}^{q,R,\Gamma}_{s_1,s_2}, \|\cdot\|_F, \sqrt{m}\epsilon)$. Thus it remains to bound the integral

$$\int_0^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \sqrt{\log N(\mathcal{K}^{q,R,\Gamma}_{s_1,s_2}, \|\cdot\|_F, \sqrt{m}\epsilon)}\,\mathrm{d}\epsilon, \tag{4.87}$$

whose integrand can be controlled by means of Lemma 4.13. Therefore, let us assume $s_2/n_2 \le s_1/n_1$ for the moment. Moreover, let us include a factor of $\sqrt{m}$ in the integral.

After the change of variables $\epsilon' = \sqrt{m}\epsilon$ we use the bound on the metric entropy of the set $\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}$ together with the basic inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for all $x, y \geq 0$ to obtain the following upper bound

$$
\begin{aligned}
&\int_0^{\frac{\Gamma\sqrt{R}}{\sqrt{m}}} \sqrt{m}\sqrt{\log N(\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}, \|\cdot\|_F, \sqrt{m}\epsilon)}\, \mathrm{d}\epsilon = \int_0^{\Gamma\sqrt{R}} \sqrt{\log N(\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}, \|\cdot\|_F, \epsilon')}\, \mathrm{d}\epsilon' \\
&\quad \leq \int_0^{12\Gamma\sqrt{R}(s_2/n_2)^{1/q-1/2}} \sqrt{R(n_1+n_2+1)\log\big(\tfrac{30\Gamma R}{\epsilon'}\big)}\, \mathrm{d}\epsilon' \\
&\quad + \int_{12\Gamma\sqrt{R}(s_2/n_2)^{1/q-1/2}}^{12\Gamma\sqrt{R}(s_1/n_1)^{1/q-1/2}} \sqrt{Rs_2\big(\tfrac{12\Gamma\sqrt{R}}{\epsilon'}\big)^{2q/(2-q)}\log\Big(\tfrac{6en_2}{s_2}\big(\tfrac{\epsilon'}{12\Gamma\sqrt{R}}\big)^{(3q-2)/(2-q)}\Big)}\, \mathrm{d}\epsilon' \\
&\quad + \int_{12\Gamma\sqrt{R}(s_2/n_2)^{1/q-1/2}}^{12\Gamma\sqrt{R}(s_1/n_1)^{1/q-1/2}} \sqrt{R(n_1+1)\log\big(\tfrac{30\Gamma R}{\epsilon'}\big)}\, \mathrm{d}\epsilon \\
&\quad + \int_{12\Gamma\sqrt{R}(s_1/n_1)^{1/q-1/2}}^{\Gamma\sqrt{R}} \sqrt{Rs_1\big(\tfrac{12\Gamma\sqrt{R}}{\epsilon'}\big)^{2q/(2-q)}\log\Big(\tfrac{6en_1}{s_1}\big(\tfrac{\epsilon'}{12\Gamma\sqrt{R}}\big)^{(3q-2)/(2-q)}\Big)}\, \mathrm{d}\epsilon' \\
&\quad + \int_{12\Gamma\sqrt{R}(s_1/n_1)^{1/q-1/2}}^{\Gamma\sqrt{R}} \sqrt{Rs_2\big(\tfrac{12\Gamma\sqrt{R}}{\epsilon'}\big)^{2q/(2-q)}\log\Big(\tfrac{6en_2}{s_2}\big(\tfrac{\epsilon'}{12\Gamma\sqrt{R}}\big)^{(3q-2)/(2-q)}\Big)}\, \mathrm{d}\epsilon' \\
&\quad + \int_{12\Gamma\sqrt{R}(s_1/n_1)^{1/q-1/2}}^{\Gamma\sqrt{R}} \sqrt{R\log\big(\tfrac{30\Gamma R}{\epsilon'}\big)}\, \mathrm{d}\epsilon' \\
&\quad =: I_1 + I_2 + I_3 + I_4 + I_5 + I_6 =: I,
\end{aligned}
\tag{4.88}
$$

which we elaborate in Appendix A.2. From the computations there we obtain

$$
\begin{aligned}
I \lesssim \Bigg(&\Gamma^2 R^2\bigg(\bigg(\Big(\frac{s_1}{n_1}\Big)^{2/q-2} s_1 + \Big(\frac{s_2}{n_2}\Big)^{2/q-2} s_2 + 1\bigg) + 144^{(2q-2)/(2-q)}(s_1+s_2)\bigg) \\
&\cdot \operatorname{polylog}\bigg(e\sqrt{R}\max\Big\{\frac{n_1}{s_1}, \frac{n_2}{s_2}\Big\}^{\max\{1/q-1/2,1\}}\bigg)\Bigg)^{1/2},
\end{aligned}
\tag{4.89}
$$

where the hidden constant is an absolute constant. Due to the symmetry of this bound with respect to s_1 and s_2 and analogously with respect to n_1 and n_2 and combinations thereof, the preliminarily made assumption that $s_2/n_2 \leq s_1/n_1$ can be dropped. □

4.5.3 A Probabilistic Bound on Suprema of Chaos Processes

In this short subsection we present a probabilistic deviation and tail bound for the supremum of a chaos process, which is the main tool for the proof of Theorem 4.7.

Theorem 4.17 (A Bound on Suprema of Chaos Processes, Corollary of [KMR14, Theorem 3.1])**.** *Let $\mathcal{H}$ be a symmetric set of matrices and let $\boldsymbol{\xi}$ be a random vector whose entries ξ_i are independent mean-zero L-sub-Gaussian random variables of unit variance. Set*

$$
\begin{aligned}
E &= \gamma_2(\mathcal{H}, \|\cdot\|)\big(\gamma_2(\mathcal{H}, \|\cdot\|) + d_F(\mathcal{H})\big), \\
V &= d_\infty(\mathcal{H})\big(\gamma_2(\mathcal{H}, \|\cdot\|) + d_F(\mathcal{H})\big) \ \textit{and} \\
U &= d_\infty^2(\mathcal{H}).
\end{aligned}
\tag{4.90}
$$

Then, for $t > 0$,

$$\mathsf{P}\left(\sup_{\mathbf{H}\in\mathcal{H}}\left|\|\mathbf{H}\boldsymbol{\xi}\|_2^2 - \mathsf{E}\|\mathbf{H}\boldsymbol{\xi}\|_2^2\right| \geq cE + t\right) \leq 2\exp\left(-d\min\left\{\frac{t^2}{V^2}, \frac{t}{U}\right\}\right), \tag{4.91}$$

where the constants c and d depend only on L.

Remark 4.18. Theorem 4.17 follows immediately from Theorem 3.1 in [KMR14] by noting that the symmetry, i.e., $\mathcal{H} = -\mathcal{H}$, implies $d_\infty(\mathcal{H}) \leq \gamma_2(\mathcal{H}, \|\cdot\|)$, what can be seen directly from the definition of Talagrand's γ_2-functional.

4.5.4 Proof of the Main Result

We now have all necessary ingredients to conclude with a proof of Theorem 4.7. We give the proof for the set $\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}$. The adaption for the smaller set $\mathcal{S}_{s_1,s_2}^{R,\Gamma}$ is straightforward and can be found, for instance, on page 19 in [FMN19].

Proof of Theorem 4.7. Let us abbreviate $\mathcal{H}_{\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}}$ by $\mathcal{H}_{\mathcal{K}}$. Making use of the auxiliary quantities introduced in equations (4.83) and (4.84), we observe

$$\mathsf{P}\left(\sup_{\mathbf{Z}\in\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}}\left|\left\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{Z})\right\|_2^2 - \|\mathbf{Z}\|_F^2\right| \geq \delta\right) = \mathsf{P}\left(\sup_{\mathbf{H}_\mathbf{Z}\in\mathcal{H}_\mathcal{K}}\left|\|\mathbf{H}_\mathbf{Z}\boldsymbol{\xi}_\mathcal{A}\|_2^2 - \mathsf{E}\|\mathbf{H}_\mathbf{Z}\boldsymbol{\xi}_\mathcal{A}\|_2^2\right| \geq \delta\right). \tag{4.92}$$

Let us now assume that the number of measurements m is sufficiently large in the sense that it obeys (4.59), i.e., for some $0 < \Delta < 1$ it holds

$$\begin{aligned} m \gtrsim \Delta^{-2}R\Bigg(\bigg(\Big(\frac{s_1}{n_1}\Big)^{2/q-2}s_1 + \Big(\frac{s_2}{n_2}\Big)^{2/q-2}s_2 + 1\bigg) + 144^{(2q-2)/(2-q)}(s_1+s_2)\Bigg) \\ \cdot \operatorname{polylog}\left(e\sqrt{R}\max\left\{\frac{n_1}{s_1}, \frac{n_2}{s_2}\right\}^{\max\{1/q-1/2,1\}}\right), \end{aligned} \tag{4.93}$$

where the hidden constant shall be the one from Lemma 4.16. Denoting the bound on Dudley's integral for the set $\mathcal{H}_\mathcal{K}$, which was established in this lemma, by $\mathcal{D}_\mathcal{K}$, we are now able to control the quantities E, V and U from Theorem 4.17. More precisely, it hold

$$E \leq \mathcal{D}_\mathcal{K}^2 + \Gamma\sqrt{R}\mathcal{D}_\mathcal{K}, \quad V \leq \frac{\Gamma\sqrt{R}\mathcal{D}_\mathcal{K} + \Gamma^2 R}{\sqrt{m}} \quad \text{and} \quad U \leq \frac{\Gamma^2 R}{m}. \tag{4.94}$$

We can verify that $\mathcal{D}_\mathcal{K} \leq \Delta\Gamma\sqrt{R} \leq \Gamma\sqrt{R}$ and $\mathcal{D}_\mathcal{K}^2 + \Gamma\sqrt{R}\mathcal{D}_\mathcal{K} \leq \Gamma^2 R(\Delta^2+\Delta) \leq 2\Gamma^2 R\Delta$, and thus $E \leq 2\Gamma^2 R\Delta$ and $V \leq 2\Gamma^2 R/\sqrt{m}$. Now, let $c > 0$. Then, for $\delta \geq 3c\Gamma^2 R\Delta$, which assures $\delta \geq c(E + \Gamma^2 R\Delta)$, we can apply Theorem 4.17 to obtain

$$\begin{aligned} \mathsf{P}\left(\sup_{\mathbf{H}_\mathbf{Z}\in\mathcal{H}_\mathcal{K}}\left|\|\mathbf{H}_\mathbf{Z}\boldsymbol{\xi}_\mathcal{A}\|_2^2 - \mathsf{E}\|\mathbf{H}_\mathbf{Z}\boldsymbol{\xi}_\mathcal{A}\|_2^2\right| \geq \delta\right) &\leq \mathsf{P}\left(\sup_{\mathbf{H}_\mathbf{Z}\in\mathcal{H}_\mathcal{K}}\left|\|\mathbf{H}_\mathbf{Z}\boldsymbol{\xi}_\mathcal{A}\|_2^2 - \mathsf{E}\|\mathbf{H}_\mathbf{Z}\boldsymbol{\xi}_\mathcal{A}\|_2^2\right| \geq c(E+\Gamma^2 R\Delta)\right) \\ &\leq 2\exp\left(-d\min\left\{\frac{c^2\Gamma^4 R^2\Delta^2}{(2\Gamma^2 R)^2}m, \frac{c\Gamma^2 R\Delta}{\Gamma^2 R}m\right\}\right) \\ &\leq 2\exp\left(-d'\Delta^2 m\right), \end{aligned} \tag{4.95}$$

showing the claim. □

Chapter 5

Numerical Experiments

As already announced in the preface of the former chapter, in this final chapter we provide numerical evidence for the theoretical results we presented. We start with the familiar setting, meaning that we aim at recovering simultaneously sparse and low-rank matrices. At first, properties such as the empirical recovery probability, the mean approximation error of the numerical solution and the sparsity of its component vectors are addressed. Afterwards, we spend a few words on the role of the hyperparameters, before we test a multilevel-type initialization strategy. Eventually, the setting of the numerical experiments is modified by making a further structural assumption, namely positivity of the sparse component vectors. We analyze numerically whether one can benefit from this additional structure.

For our numerical simulations the LRZ Linux-Cluster CoolMUC-2 was used. The computations were performed on one single compute node with an Intel Xeon E5-2690 v3 with 28 cores and 64GB RAM using a MATLAB R2019a implementation.

5.1 Numerical Analysis of ARBeR

As described in Sections 3.1 and 4.1, the objective of our numerical method ARBeR, as given in Algorithm 5, is the efficient and robust recovery of a low-rank matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ which admits a non-orthogonal sparse decomposition of the form (4.2).

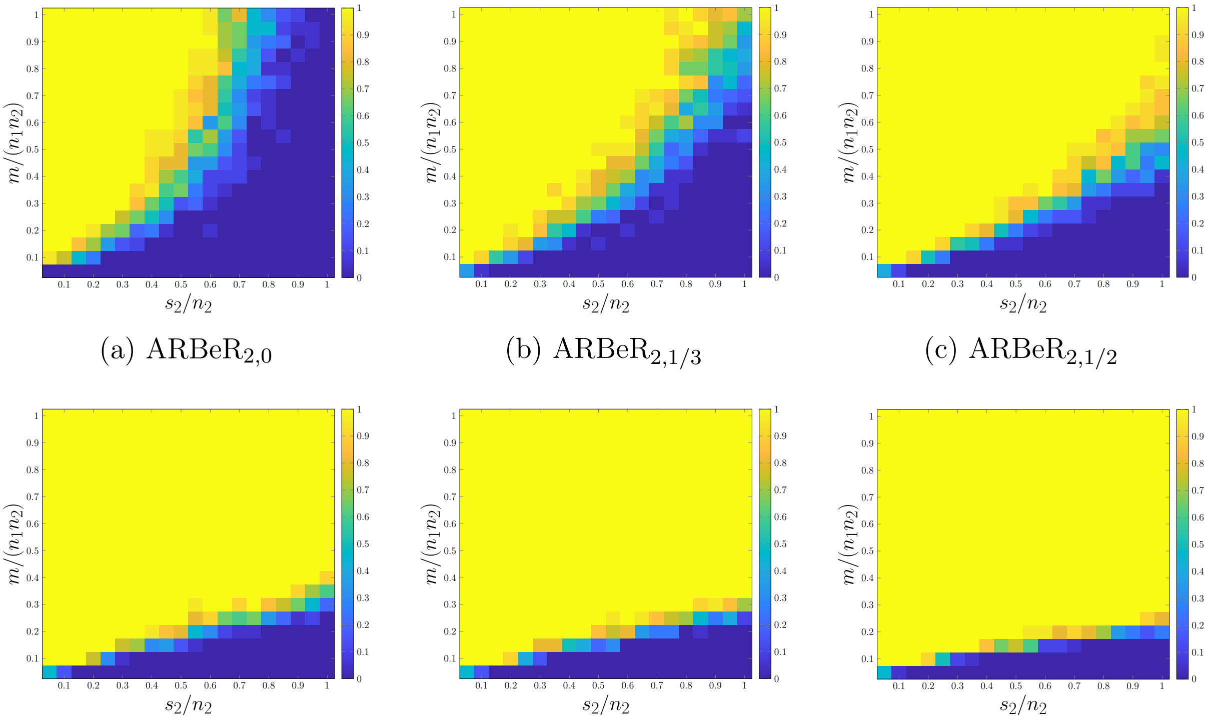
Throughout this and also the next section, we consider low-rank-R matrices with merely sparse right component vectors, i.e., we have $1 \leq s_2 \leq n_2$, but no sparsity in the left component, i.e., $s_1 = n_1$. As a consequence, the usual dimensional setting requires $n_1 \ll n_2$, since we could not benefit from the sparsity in the right component vectors elsewise. This means that we consider the multi-penalty functional $\mathcal{J}_{\alpha,\beta}^{2,q,R}$ with $0 \leq q \leq 1$. Hence, our algorithmic approach employs ridge regression for the left and bridge-q regression for the right component vectors. We denote this compactly by $\text{ARBeR}_{2,q}$

A vector $\mathbf{y} \in \mathbb{R}^m$ of m inaccurate and incomplete measurements is obtained from the matrix $\mathbf{X}$ according to (4.1). The involved measurement operator $\mathcal{A}$ in this and the subsequent section is a suitably scaled Gaussian measurement ensemble. However, numerical experiments showing comparable results were also performed using operators with, e.g., Bernoulli random variables. Furthermore, the noise we assume is ineliminable and of considerable magnitude. To model this, we set the noise level to $\eta = \|\boldsymbol{\eta}\|_2 = 0.3\|\mathbf{X}\|_F$.

In the following, we focus mainly on the comparison of different values for the regularizing (quasi)-norm parameter q. For appropriate numerical results with the state-of-the-art algorithm SPF, we direct the reader to Section 5 from [FMN19].

5.1.1 On the Numerical Solution of the Algorithm

The first and maybe already the most intriguing issue, which was addressed extensively from the theoretical perspective in Section 4.5, deals with the number of required measurements to ensure, with high probability, successful recovery of the simultaneously structured matrix. In order to investigate this question numerically, we analyze the recovery of 20 randomly drawn rank-1 matrices $\mathbf{X} \in \mathbb{R}^{8\times 128}$ with Frobenius norm 10.
Therefore, for different regularizing (quasi)-norm parameters $q \in \{0, 1/3, 1/2, 2/3, 3/4, 1\}$, we visualize the empirical recovery probability achieved by ARBeR$_{2,q}$ for various relative sparsities $0 < s_2/n_2 \leq 1$ and various numbers of measurements m in Figure 5.1. The latter is also plotted relative to the ambient dimension of the matrix space, i.e., relative to $n_1 n_2$. We call a recovery successful, if for the relative approximation error it holds $\left\|\mathbf{X} - \widehat{\mathbf{X}}_{\mathrm{ARBeR}_{2,q}}\right\|_F / \|\mathbf{X}\|_F \leq 0.4$. ARBeR$_{2,q}$ employs iterative bridge-q thresholding to find an approximate solution to the vector-valued subproblem (4.32), which is responsible for the recovery of the sparse component vector.



(a) ARBeR$_{2,0}$

(b) ARBeR$_{2,1/3}$

(c) ARBeR$_{2,1/2}$

(d) ARBeR$_{2,2/3}$

(e) ARBeR$_{2,3/4}$

(f) ATLAS, resp. ARBeR$_{2,1}$

Figure 5.1. Phase transition diagrams depicting the recovery success of ARBeR$_{2,q}$ for different values of q and various sparsities and numbers of measurements. The empirical recovery probability is depicted by color from zero (blue) to one (yellow).

Regarding the internal parameters of the employed numerical method, we fix the regularization parameters $\alpha = \beta = 0.5$ and for the number of outer iterations we choose $K = 10$. Apart from that, the number of iterations for the involved intermediate iterative thresholding algorithms tackling the multiple (non)-convex optimization problems (4.32) is limited by $L = 10^6$. The iterative method may break earlier if two consecutive iterations are closer than a prescribed tolerance of 10^{-8}. Moreover, the method is initialized with the leading singular vectors of $\mathcal{A}^*(\mathbf{y})$.

The phase transition diagrams in Figure 5.1 contain various information. It stands out that ARBeR$_{2,q}$ with a parameter q close to one performs best concerning the recovery success for various relative sparsities. This is most visible for less sparse right component vectors and is not surprising as small q's naturally promote sparsity stronger than larger ones, highlighting the well-known trade-off between data fidelity and sparsity, whose illustration will be complemented by Figure 5.3. For small sparsities ($s_2 \leq 0.05 \cdot n_2$), the methods perform approximately equally well. However, we do not observe any noticeable improvement for smaller values of q, which conflicts the theoretical results from Section 4.5 and is most likely due to the significant non-convexity and the linked difficulty to find a good initialization. Moreover, considering the diagrams individually, each variant of ARBeR associated with a different value of q experiences a relatively sharp transition of the empirical recovery probability as the number of measurements increases, i.e., the recovery success shifts within a small interval. This is a typical behavior when reconstructing signals with some sort of parsimony. We furthermore observe that the transition margin is wider the denser the right component is.

In order to quantify the former observations more explicitly and provide numerical evidence also for the recovery of matrices with higher but still low rank we conduct the experiment visualized in Figure 5.2. Based on 20 randomly drawn rank-5 matrices $\mathbf{X} \in \mathbb{R}^{16\times 100}$, which admit a non-orthogonal decomposition with 10-sparse right component vectors and have Frobenius norm 10, we compare the empirical recovery probabilities and the relative average approximation errors of ARBeR$_{2,q}$ for the same values of q as previously. In this case, a relative approximation error of 0.5 suffices for a recovery to be called successful.

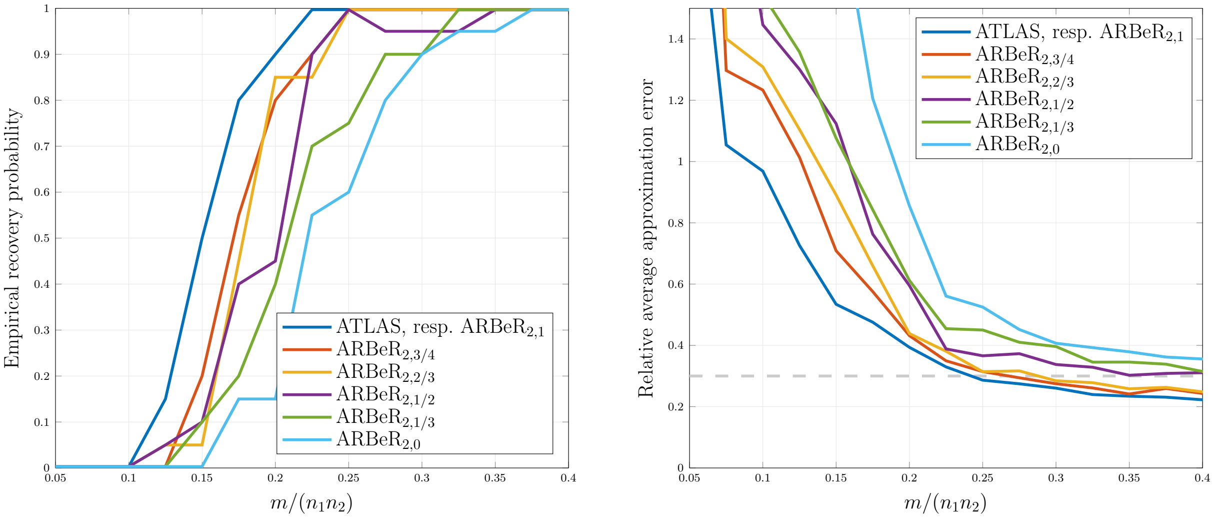


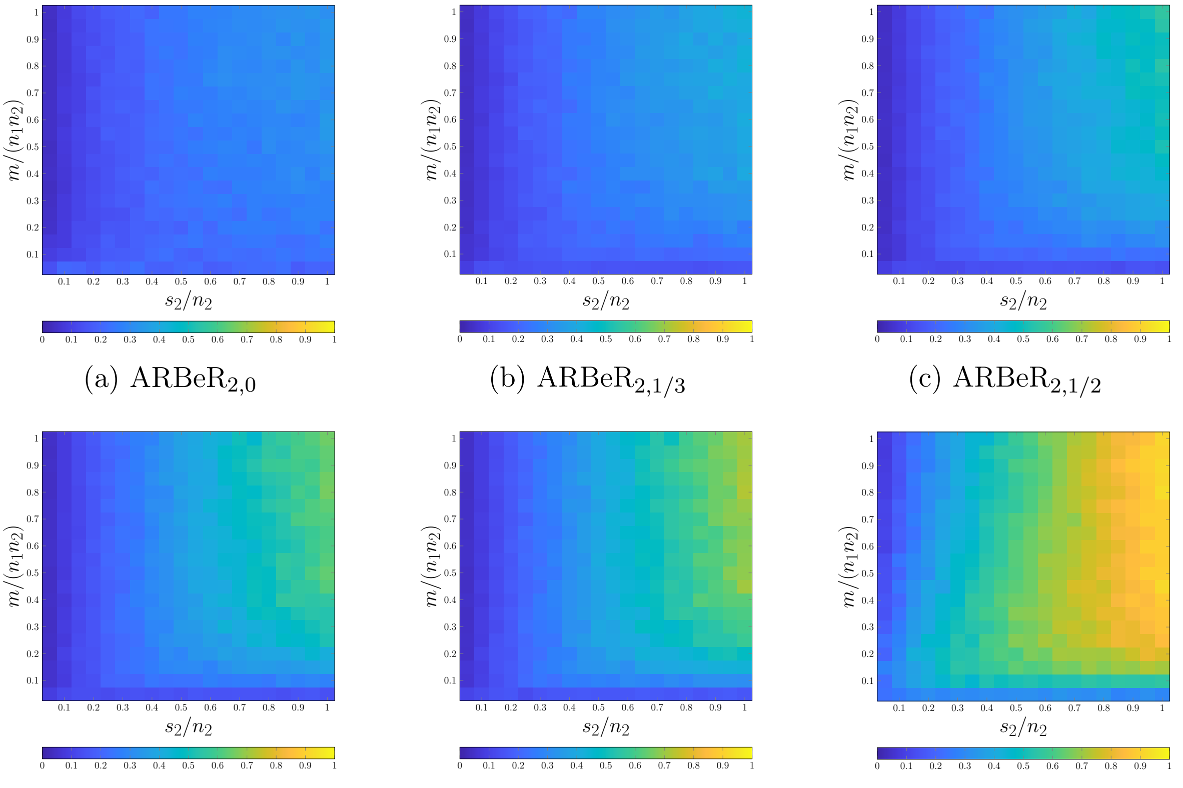
Figure 5.2. Comparison of the empirical recovery probabilities (left) and the relative average approximation errors (right) of ARBeR$_{2,q}$ for different values of q and various numbers of measurements. For reference, the noise level is appended in the right figure as a dashed line.

The internal parameters are chosen as before, we only use $K = 50$ outer iterations and initialize the algorithm with the leading five singular vector pairs of $\mathcal{A}^*(\mathbf{y})$.

The figure on the left-hand side, which depicts the empirical recovery probability for a varying number of measurements, confirms the sharp transition boundaries and the superiority of ATLAS. However, also the other methods show convincing results, in particular when bringing to mind that the measurements are highly corrupted and the initialization is certainly not optimal. Beyond that, in the figure on the right we directly compare the corresponding relative average approximation errors. We observe that the relative error is at the noise level as soon as the number of measurements exceeds the phase transition. Moreover, we note the familiar relation between the different versions of ARBeR.

The, after these two figures, noticeable superiority of ATLAS does not come without a cost. To see this, let us have a closer look at one particular matrix recovery and investigate the structure of the recovered components. The number of available measurements is set to $m = 0.3 \cdot n_1 n_2$. We recall that the interesting components of our signal $\mathbf{X}$ are the 10-sparse right components, whose ambient dimension is 100. The by ATLAS proposed reconstruction is a rank-4 matrix with one 41-sparse, two 35-sparse and one 15-sparse component. In turn, ARBeR$_{2,0}$ suggests a rank-5 matrix with one 6-sparse, two 4-sparse and two 1-sparse components. The other variants balance this by finding sparser components than ATLAS does but less sparse components than ARBeR$_{2,0}$.

To quantify this effect, we return to the setting of Figure 5.1. For different values of the parameter q, we depict the relative sparsity of the by ARBeR$_{2,q}$ recovered right component vector for various relative sparsities and various numbers of measurements.



(a) ARBeR$_{2,0}$

(b) ARBeR$_{2,1/3}$

(c) ARBeR$_{2,1/2}$

(d) ARBeR$_{2,2/3}$

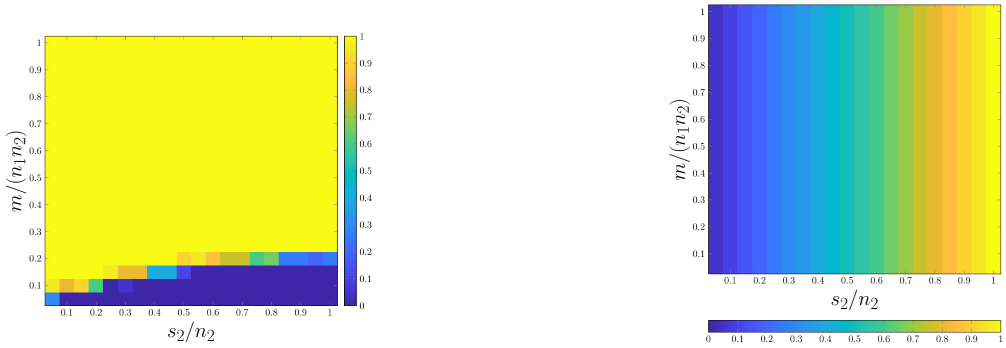
(e) ARBeR$_{2,3/4}$

(f) ATLAS, resp. ARBeR$_{2,1}$

Figure 5.3. Color gradient diagrams picturing the relative sparsity of the by ARBeR$_{2,q}$ recovered right component vector for different values of q and various sparsities and numbers of measurements. The recovered averaged relative sparsity is depicted by color from zero (blue) to one (yellow).

Interpreting the diagrams separately, we observe that ARBeR tends to return rather sparse components as long as too few measurements are available, i.e., if the number of available measurements is below the in Figure 5.1 observed phase transition. This is a consequence of the structure of our multi-penalty functional. If the accessible information is too little to ensure data fidelity, rather sparse components which have at least some positive effect on the residual term are a reasonable choice to achieve a small value for the functional. A comparison of the diagrams among each other convincingly demonstrates that a small choice of the parameter q promotes sparsity. However, as can be seen in Subfigure 5.3(a), ARBeR$_{2,0}$ also enforces comparably strong sparsity in regions, where the true sparsity of the component vectors is much weaker, i.e., the components are denser. This hinders the recovery as shown in the corresponding Subfigure 5.1(a) and makes small q's unsuitable for matrices with less sparse components. Conversely, for significantly sparse components, ATLAS proposes too dense vectors as can be seen when having a close look at Subfigure 5.3(f). Even though this has no negative effect on the recovery success as shown in Subfigure 5.1(f), it is certainly undesirable, as the recovery of the structures is a central point and their lack would result in a loss of interpretability. Remarkably, let us note that, e.g., ARBeR$_{2,2/3}$ accomplishes successful recoverability in regions of larger sparsity by returning an approximate solution with sparser component vectors. This is particularly beneficial if defects in the sparsity are present in the original signal.

At the end of this subsection, let us give analogous figures for the performance in case of ARBeR$_{2,0}$ when replacing the ℓ_0-regularized optimization problem (4.32) by its ℓ_0-constrained version, cf. Lemma A.4. For our numerical method this means that the iterative hard thresholding method is replaced by an iterative best s_2-term approximation method, which were both described in Paragraph 1.4.3(3). Of course, this requires prior knowledge of the sparsity level s_2.



(a) Recovery success

(b) Relative sparsity

Figure 5.4. Phase transition diagram for the recovery success of ARBeR (left) and color gradient diagram for the relative sparsity of the reconstructed right component vector (right), when utilizing the best s_2-term approximation algorithm to tackle the ℓ_0-constrained version of the regularized problem (4.32). In both cases, the respective quantity is depicted by color from zero (blue) to one (yellow).

As one expects, by construction, this variant of ARBeR$_{2,0}$ always recovers the correct sparsity, see Subfigure 5.4(b). Furthermore the recovery success of this method is comparable to the one of ATLAS in the range of large sparsities and to the standard variant of ARBeR$_{2,0}$ in case of small sparsities.

At the end of this subsection we want to show that the different variants of ARBeR associated with different values of the parameter q perform approximately equally well in regards to the dependency of the relative approximation error on the noise-to-signal ratio $\|\boldsymbol{\eta}\|_2/\|\mathbf{X}\|_F$. For this experiment, whose results are depicted in Figure 5.5, the setting of Figure 5.2 in the case of $m = 0.3 \cdot n_1 n_2$ measurements is reused. This includes the choice $\alpha = \beta = 0.5$ for the regularization parameters. The data, though, is only founded on 8 randomly drawn matrices $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$. The theoretical upper bounds are derived from Theorem 4.6 and rescaled with a factor of 1/2. To obtain a bound on the quantity $\|\mathbf{X}\|_{pq/(p+q)}$ we employ Lemma A.2(ii), see, e.g., the first inequality in (4.10). Note that a theoretical bound in case of ARBeR$_{2,0}$ is not available.

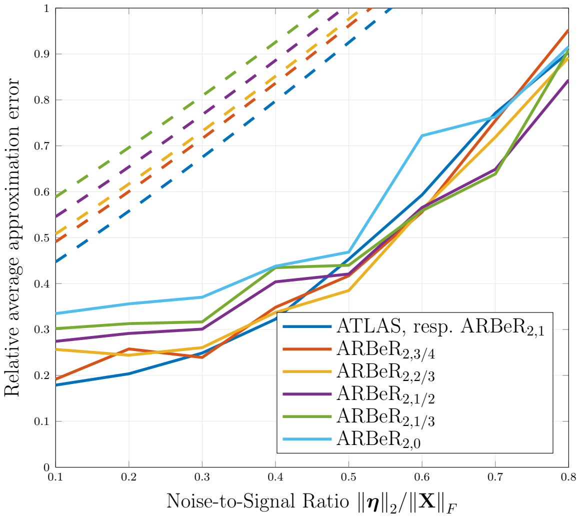


Figure 5.5. Dependency of the relative average approximation errors of ARBeR$_{2,q}$ for different values of q on the noise-to-signal ratio. For reference, theoretical upper bounds are appended in the left figure as dashed lines.

5.1.2 Hyperparameter Tuning

Having observed that the regularizing (quasi)-norm parameter q can be utilized to steer the sparsity of the recovered right component vectors, in this subsection we investigate which impact the regularization parameters α and β have on the performance of the algorithm and the structure of the recovered matrix. For the sake of simplicity, we assume $\alpha = \beta$. Complementary numerical experiments testing also other relations between α and β in the case of ATLAS can be found in Subsection 5.1 of [FMN19]. There, however, rank-1 matrices of the same dimension were used. One observation we will not make here, but transfers directly to the versions of ARBeR is that small values of the parameter α result in a small relative approximation error. This is also compatible with the theory in form of Theorem 4.6.

In Figure 5.6, we numerically analyze the dependency of the relative average approximation error and the relative average sparsity of the right component vectors with respect to the choice of the regularization parameters α and β. Except for two modifications, the setting of Figure 5.2 for a fixed number of $m = 0.3 \cdot n_1 n_2$ measurements is maintained. The major change is that the noise level is set to zero. Moreover, as already done in the last experiment of the previous subsection, we only draw 8 matrices at random. We also derive theoretical upper bounds on the relative approximation error from Theorem 4.6. As before, we rescale them by a factor of 1/2.

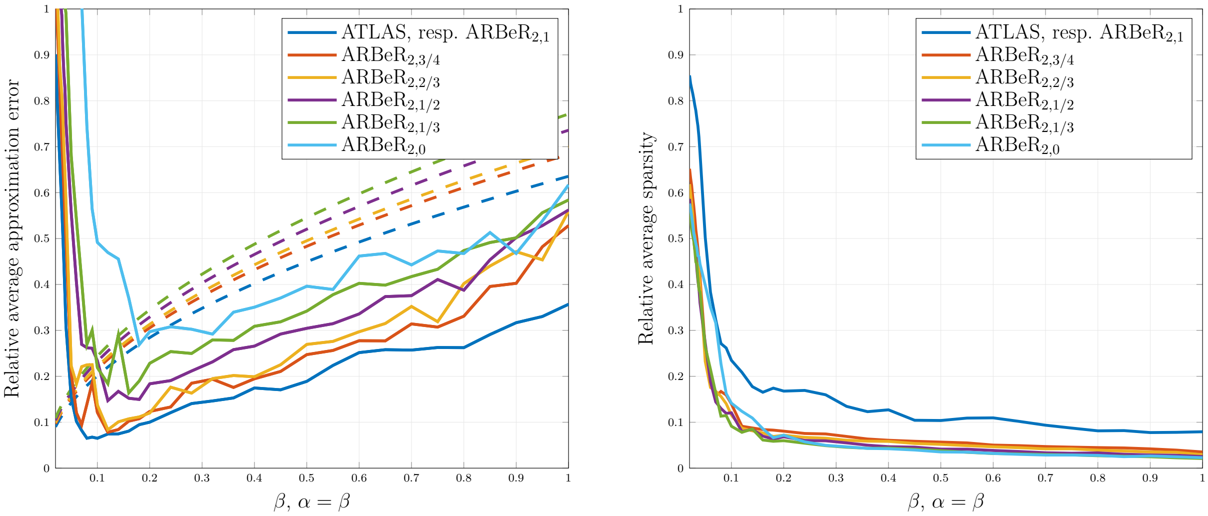


Figure 5.6. Dependency of the relative average approximation errors of ARBeR$_{2,q}$ (left) and the relative average sparsities of the reconstructed right component vectors (right) for different values of q on the regularization parameter β under the assumption $\alpha = \beta$. For reference, theoretical upper bounds are appended in the left figure as dashed lines.

The first immediate observation, concerning the figure on the left, is a decrease of the relative approximation errors as the regularization parameters α and β tend to zero. This decay follows the predicted ones given by the theoretical bounds. However, an interesting failure of recovery occurs if β becomes too small. More precisely, located between $\beta = 0.08$ and $\beta = 0.25$ for the different values of q, the error abruptly explodes. This can be traced back to a breakdown of our restricted isometry property, cf. Theorem 4.6. Small values of β force the measurement operator $\mathcal{A}$ to fulfill a restricted isometry property which requires a nearly isometric embedding for matrices with less, than just ℓ_q-effectively s_2-sparsely, structured right components. Since the number of measurements is chosen close to the limit sufficient to assure the restricted isometry property for the smaller set $\mathcal{K}^{q,R,\Gamma}_{n_1,s_2}$, the requirements on $\mathcal{A}$ to ensure recoverability in this case are too strong.
The figure on the right-hand side shows that the choice of the parameter β has an effect on the sparsity. As one expects, sparser components can be obtained by choosing the regularization parameter β sufficiently large. We want to conclude this experiment with a note. The attentive reader may wonder why, for $\beta = 0.5$, ATLAS proposes right components which approximately have the desired relative average sparsity of 0.1. This seemingly stands in conflict with the exemplary recovery ahead of Figure 5.3. But, the crucial point is that the measurements here are noise-free. As soon as noise comes into play, ATLAS returns significantly denser components.

5.1.3 The Influence of the Initialization

The criticality of the initialization was already mentioned in Subsection 4.3.3, where we proposed the initialization with the R leading singular vector pairs of $\mathcal{A}^*(\mathbf{y})$. This was also the one used so far. In this subsection we compare different initializations to investigate their influence on the recovery success. In the setting of Figure 5.2 we run $\text{ARBeR}_{2,0}$ for the three different types of initialization we describe in the following. The first, indicated by initialization with $\mathbf{X}$, utilizes the R leading singular vector pairs of the desired signal $\mathbf{X}$. This serves as reference and is not available in practice. The second, labeled multilevel initialization, employs the novel multilevel-type strategy described in the last passage of Subsection 4.3.3. The used $\Lambda = 6$ levels are $\{(2,1),(2,3/4),(2,2/3),(2,1/2),(2,1/3),(2,0)\}$. And finally, the last is the well-known initialization with the adjoint.

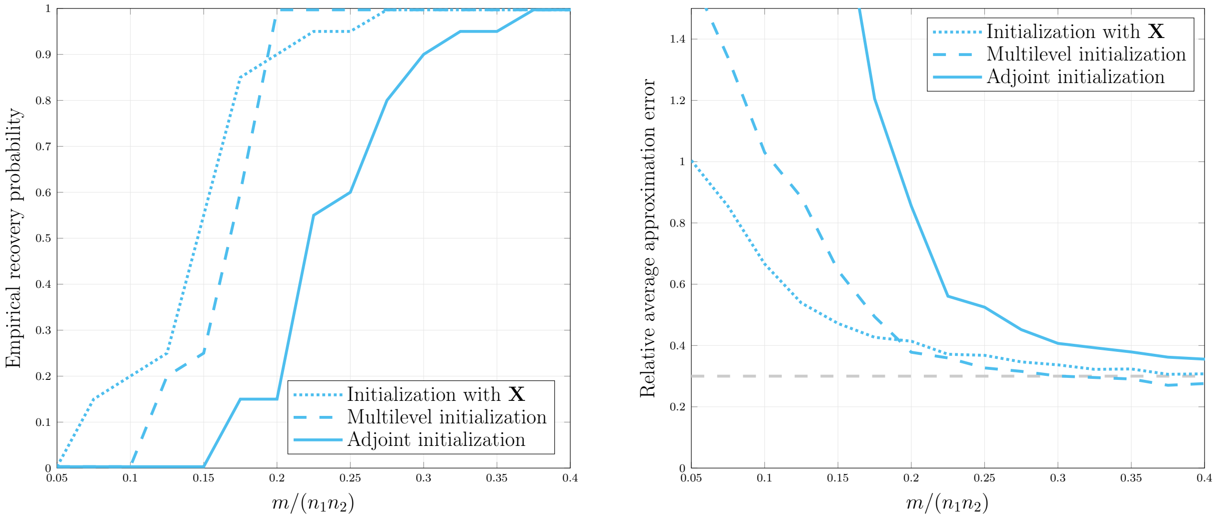


Figure 5.7. Comparison of the empirical recovery probabilities (left) and the relative average approximation errors (right) of $\text{ARBeR}_{2,0}$ for different initializations and various numbers of measurements. For reference, the noise level is appended in the right figure as a gray line.

As was to be expected, the initialization with $\mathbf{X}$ performs significantly better than the adjoint initialization. Successful recovery can be observed already for a very low number of measurements. Also the relative average approximation error is smaller and approaches noise level earlier. Let us now turn to the performance of our novel multilevel initialization strategy. It achieves a considerable improvement of the recovery probability and the relative approximation error compared to the adjoint initialization. In comparison to Figure 5.2 we realize that recovery sets in at the same time as for ATLAS, which is due to the fact that the multilevel version starts therewith. From this we conjecture that the convergence radius of the resulting multilevel version of $\text{ARBeR}_{2,0}$ was enlarged towards the one of ATLAS. By noting that the respective recovery probabilities are about as good as the ones of ATLAS, we provided proof that the bottleneck for the severely non-convex versions of ARBeR is the initialization. Moreover, we presented a heuristic how a promising initialization can be computed, for which, however, one may not forget about the involved computational cost.

There is one more feature worthy to be mentioned. For a sufficiently large number of measurements, our multilevel initialization leads to a smaller approximation error than the initialization with the signal of interest itself. We suppose that the reason therefore is that alternating minimization searches minimizers to the multi-penalty functional (4.3), which are, particularly in the noisy case, different from $\mathbf{X}$. In this case, initialization with a related minimizer for a larger q is more beneficial than initialization with $\mathbf{X}$.

5.2 Adding Non-Negativity to the Matrix Decomposition

In the introductory example of the grocery store from Section 3.1 we motivated additional structure of the right component vectors. Besides their sparsity, in order to give them a real-world meaning, non-negativity of their entries was reasoned. After a suitable rescaling this allows to interpret a component $\mathbf{v}_r$ as a discrete probability vector.
As this structural assumption restricts the space of eligible signals, it is inevitable to come up with the question whether this is reflected by the number of necessary measurements. In contrast to a structure such as low-rankness or sparsity, though, one cannot expect a further reduction of the order of the measurements as the dimension of the space remains unchanged. Whereas an s-sparse vector in dimension N is located in a union of $\binom{N}{s}$ s-dimensional subspaces, a non-negativity assumption restricts the set to exactly this number of s-dimensional non-negative orthants embedded into the larger ambient space.
Detached from the question of required measurements for successful recovery, there is one desirable feature one would like to have implemented in a numerical method. Supposing that one is a priori aware of this kind of structure, one would like to have ensured that the by our algorithm proposed matrix possesses this structural property. It is not clear if ATLAS, respectively ARBeR$_{2,1}$, is capable of detecting[11] this structure.
However, by a minor modification we are able to enforce the non-negativity of the right component vectors directly in our framework. By restricting the multi-penalty functional $\mathcal{J}_{\alpha,\beta}^{2,1,R}$ to the set $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_1} \times \mathbb{R}_+^{n_2} \times \cdots \times \mathbb{R}_+^{n_2}$ and setting it to infinity outside there, minimizers thereof possess the desired structure. Formally, this can be done by employing the regularizer

$$\|\mathbf{z}\|_1^+ = \begin{cases} \|\mathbf{z}\|_1 & \text{if } \mathbf{z} \geq \mathbf{0}, \\ \infty & \text{else,} \end{cases} \tag{5.1}$$

where "$\geq$" is understood component-wise. Of course, an analogous adaption can be done for the more general functional $\mathcal{J}_{\alpha,\beta}^{p,q,R}$. In the following we limit ourselves to the case $q = 1$. The natural alternating minimization approach for this modified functional replaces the vector-valued optimization problem (4.7) by

$$\hat{\mathbf{v}}_r^k = \underset{\hat{\mathbf{v}} \in \mathbb{R}^{n_2}, \hat{\mathbf{v}} \geq \mathbf{0}}{\arg\min} \left\| \mathbf{y} - \mathcal{A}\Big(\sum_{\rho < r} \hat{\mathbf{u}}_\rho^k (\hat{\mathbf{v}}_\rho^k)^T \Big) - \mathcal{A}(\hat{\mathbf{u}}_r^k \hat{\mathbf{v}}^T) - \mathcal{A}\Big(\sum_{\rho > r} \hat{\mathbf{u}}_\rho^{k-1} (\hat{\mathbf{v}}_\rho^{k-1})^T \Big) \right\|_2^2 + \beta \|\hat{\mathbf{v}}\|_1. \tag{5.2}$$

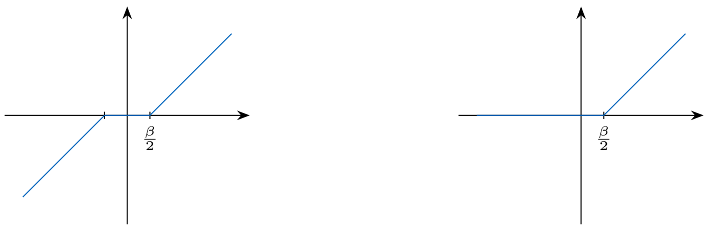
In order to derive an iterative algorithm finding an approximate solution to this problem, the tools from Section 1.4.3 can be reused. The only missing piece is a suitable thresholding operator, which takes the non-negativity into account.
For this reason, let us introduce a non-linear function, which gained tremendous attention as an activation function of artificial neurons in the theory and application of artificial neural networks in recent years. The so-called rectified linear unit (ReLU) is defined as

$$\mathrm{ReLU}(z) = \max\{0, z\}. \tag{5.3}$$

Its popularity is due to being a realistic model of biological neurons and enabling efficient training of neural networks [GBB11]. This partially closed a modeling gap between computational neurosciences and machine learning research. For notational ease let us write

[11] Note that the component vector pairs are only determined up to a sign change modifying this question to whether all entries have the same sign.

ReLU_β for the function $z \mapsto \mathrm{ReLU}(z - \beta/2)$. This function can be regarded as an activation function with threshold $\beta/2$, i.e., the associated neuron fires if the joint stimulus is above this value. Together with the scalar soft thresholding operator S_β this function is sketched in Figure 5.8 below.



(a) S_β

(b) ReLU_β

Figure 5.8. Scalar soft thresholding operator (left) and shifted ReLU (right).

It remains to see how this non-linear function is linked to the optimization problem (5.2). To this end, in the subsequent lemma we show that a suitably shifted rectified linear unit can be regarded as the proximal mapping of the regularizer $\|\mathbf{z}\|_1^+$, when applied component-wise.

Lemma 5.1. *The shifted rectified linear unit* $\mathrm{ReLU}_{2\beta}$ *is the solution of the scalar-valued optimization problem*

$$\min_{v \geq 0} \beta|v| + \frac{1}{2}(v - z)^2. \tag{5.4}$$

Proof. Due to the restriction of the optimization problem to positive v's, the absolute value in (5.4) can be dropped. The resulting optimality condition reads $\beta + (v - z) = 0$ and thus yields $v = z - \beta$, which is a valid choice as long as $z \geq \beta$. In the case that $z < \beta$, by noting that $\beta v + \frac{1}{2}(v - z)^2 = (\beta - z)v + \frac{1}{2}(v^2 + z^2)$, we immediately conclude that $v = 0$ is the optimal solution. □

Thus, in consequence, by replacing the soft thresholding operator $\mathbb{S}_\beta$ by a vectorized form of the rectified linear unit ReLU_β with threshold $\beta/2$ we are able to iteratively tackle the optimization problem

$$\hat{\mathbf{v}}_r^k = \operatorname*{arg\,min}_{\hat{\mathbf{v}} \in \mathbb{R}^{n_1}, \hat{\mathbf{v}} \geq \mathbf{0}} \left\|\tilde{\mathbf{y}} - \tilde{\mathbf{A}}\hat{\mathbf{v}}\right\|_2^2 + \beta\|\hat{\mathbf{v}}\|_1, \tag{5.5}$$

of whose form (5.2) is. This leads to a non-negative formulation of the algorithm ATLAS. Before presenting numerical results investigating the performance of the method, a word on initialization is in order. It turns out that the standard initialization with the R leading singular vector pairs of $\mathcal{A}^*(\mathbf{y})$ fails in situations where most of the leading right singular vectors are orientated oppositely to the right component vectors of $\mathbf{X}$[12]. A remedy therefore is to replace the right component vectors of the initialization with the entry-wise absolute value of the vector. This will be also used to initialize ATLAS, where it helps to identify the non-negativity structure.

Taking up the experimental setting of Figure 5.2, in the figure below we compare the empirical recovery probabilities and the relative average approximation errors of ATLAS

[12] This is based on the numerical observation that recovery saturates at probability $\approx 1-(1/2)^R$ without allowing further measurements to improve.

and its non-negative version, which employs the ReLU instead of soft thresholding, when recovering 20 randomly drawn rank-5 matrices $\mathbf{X} \in \mathbb{R}^{16\times 100}$, which admit a non-orthogonal decomposition with 10-sparse non-negative right component vectors.

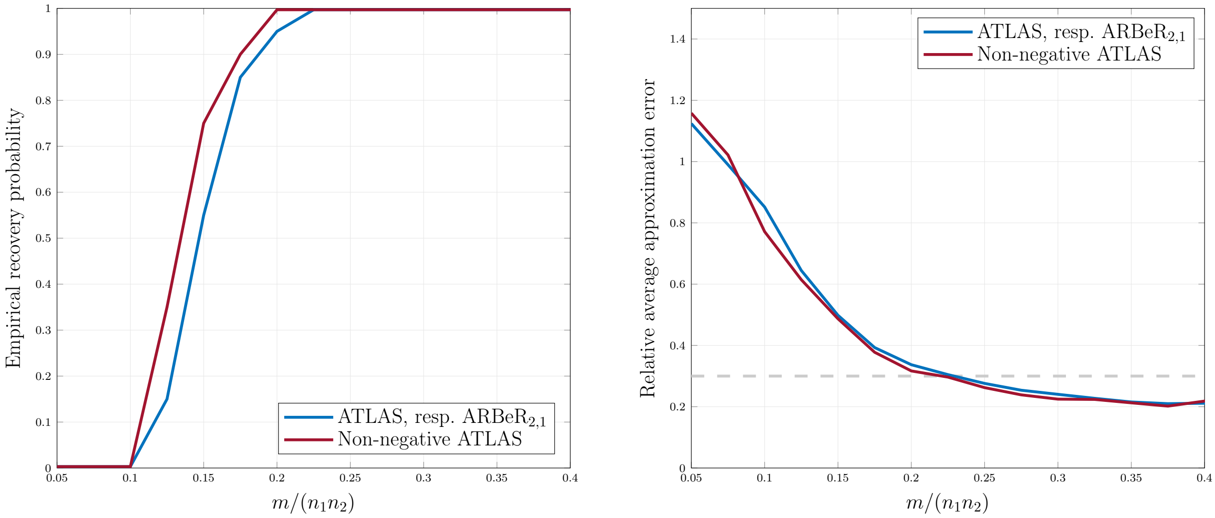


Figure 5.9. Comparison of the empirical recovery probabilities (left) and the relative average approximation errors (right) of ATLAS and its non-negative version for various numbers of measurements. For reference, the noise level is appended in the right figure.

We witness a slight improvement of the recovery probability when employing the non-negative version of ATLAS. This means that the number of required measurements indeed benefits from the additional information about the structure of the decomposition.
A further much more significant advantage of employing the ReLU is illustrated in the following figure, where we analyze the non-negativity structure of the recovered components.

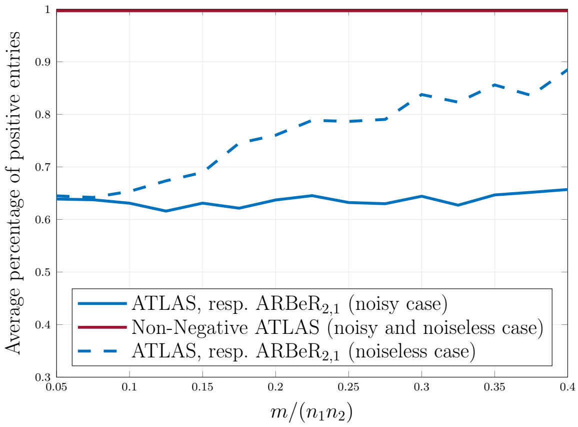


Figure 5.10. Comparison of the average percentage of positive entries in the by ATLAS and its non-negative version recovered right component vectors for various numbers of measurements. The noisy and noiseless case are considered separately.

By construction, non-negative ATLAS ensures right component vectors with merely non-negative entries. ATLAS, however, is only partially capable of recovering this structure from a reasonable amount of measurements. Especially in the case of significant measurement noise an exceedingly large number of measurements would be necessary.

Conclusions and Outlook

This thesis was concerned with the efficient and robust recovery of signals which admit multiple structures simultaneously. Inaccurate and incomplete information about the signal was given in terms of few linear perturbed measurements. In particular, we considered matrices with a low-dimensional intrinsic complexity, which was effected by low-rankness of the matrix, sparsity and positivity in the component vectors of a non-orthogonal low-rank decomposition. Therefore, we analyzed both theoretically and numerically a highly non-convex approach based on alternating minimization of a suitable multi-penalty functional. Advantageously, the emerging individual vector-valued optimization problems we obtained were numerically tractable and involved discontinuous iterative thresholding based algorithms, which are well-understood and investigated.
In order to embed this approach into the wide field of compressed sensing, where it undoubtedly belongs to, we started by outlining the most important concepts, ideas, proofs and numerical algorithms of elementary vector-valued compressed sensing in Chapter 1. Despite providing a general overview of the field, our first chapter was tailored to comply the special requirements of the final theoretical Chapter 4, which led to the study of the NP-hard non-convex ℓ_q-minimization problem. After that, in Chapter 2, we turned towards the recovery of matrices with a single structure, namely low-rankness. Observing that such matrices can be recovered, with high probability, from an optimal number of measurements, we raised the question, whether one could profit from further, additional structure. Before we addressed this question from a purely theoretical and partially abstract point of view in Chapter 3, we laid the foundation for our numerical approach by presenting the power factorization method, an alternating minimization based approach. From an applied perspective, it transpired that non-convex approaches may be worth to be considered as they arise naturally from the nature of the matrix decomposition and are not a priori intractable. Theoretically, this was made more rigorous in the already brought up third chapter, where we explained why approaches relying on convex optimization are limited and cannot make use of the full available structure. In turn, we emphasized the chances non-convex methods offer, for what we also gave theoretical evidence. Eventually, in the last two chapters, once from a theoretical point of view and once from a numerical perspective, we analyzed the proposed highly non-convex approach for the compressed sensing and robust recovery of simultaneously structured matrices from inaccurate and incomplete linear measurements. In Chapter 4 we first investigated global optimizers of the multi-penalty functional, which comprises data fidelity, low-rankness and sparsity in the component vectors of the non-orthogonal matrix decomposition. This included a brief excursion to high-dimensional geometry revealing both stunning and singular features of high-dimensional shapes. Afterwards, our algorithm which we dubbed Alternating Ridge and Bridge or ℓ_0-Regression was stated and convergence was established. Lastly, we introduced a very general class of simultaneously (effectively) sparse and low-rank matrices

and a consistent restricted isometry property, which was shown to be sufficient to guarantee some sort of approximation. Subsequently, using tools from the theory of stochastic processes, we proved that random measurements operators fulfill our additive restricted isometry property with high probability provided that the number of measurements scales, up to a polylogarithmic factor, linearly in the intrinsic dimension of the signal. Moreover, we saw that this number can be further improved by using "more non-convex" regularizers. However, this improvement reflects itself only in prefactors depending on the relative sparsity of the sparse component vectors and constant factors. By means of Chapter 5 we supported the former chapter with numerical verification of the theoretical results. At the very end, we even made a third structural assumption besides low-rankness and sparsity, namely positivity of the vector components.

In retrospect, we see several possible directions for further research, some of which we want to address in the following.

As for all non-convex algorithms, initialization is a delicate task, both decisive and challenging. We did not establish any provably correct recovery guarantees or convergence guarantees for our or a different type of initialization. This remains an open problem. Moreover, there are two further quantities associated with the multi-penalty functional and, in consequence, our numerical method for which we require some sort of prior knowledge. First, the rank of the matrix was hard-coded into the formulation, raising the question if one could determine it in advance if it is unknown. Regarding this, however, we saw that our algorithm does not enforce the rank but sees it as an upper bound. And even if it underestimates the rank it still yields a suitable approximation, meaning that very generous estimates of the rank would be sufficient for practical applications. Second, the regularization parameters α and β were fixed in advance and need to be chosen cautiously, which is a well-known drawback of multi-penalty approaches in general. Furthermore, we focussed exclusively on measurement operators constructed using (unstructured) sub-Gaussian measurement ensembles. In several applications, however, one does not have this freedom as one may be bound to a limited amount of randomness. Examples include structured random measurements, such as random partial Fourier matrices or partial random circulant matrices. Moreover, whereas Gaussian measurement ensembles have full rank with probability one, in certain instances one may face rank-one measurements. While having received much attention in the standard compressed sensing framework, an adaption to our setting remains to be done. A further open, yet, central question to be answered is the question of optimality. This issue is closely linked to information theoretical bounds for our introduced matrix sets of $((\ell_p, \ell_q)$-effectively) (s_1, s_2)-sparse rank-R matrices. We are currently not aware of any such bounds, for which reason we only showed that our established bound on the necessary number of measurements for reconstructing such matrices improves upon previous results.

Eventually, as the final contribution of this thesis, we want to outline a beautiful connection to deep feedforward neural networks.

Machine learning, and the broader field of artificial intelligence, gained tremendous attention in recent years. Due to their superior performance in applications such as face and handwriting recognition, natural language processing and strategy games as well as their progress in autonomous driving and medical diagnosis, they experienced enormous research interest. Machine learning typically comes into play when it is too complicated to code a computer program directly and when one disposes of large amounts of data. In supervised learning, a branch of machine learning, labeled training data is used to find patterns in order to generalize beyond the input data, i.e., be able to predict labels of unknown instances. Artificial neural networks, among others, such as support vector machines, provide a way to represent data and are capable of revealing complex models on the basis of unstructured data. An illustration of a three-layered feedforward neural network is given below.

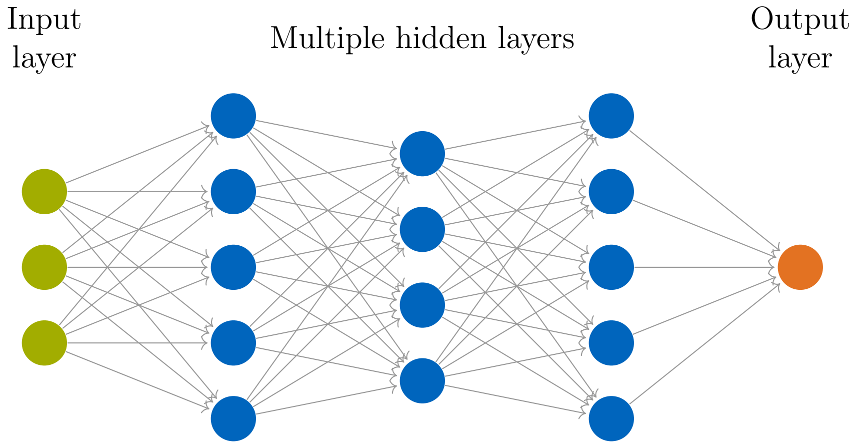


Figure: Structure of a fully connected feedforward neural network with three input neurons, three hidden layers with five, four and five neurons and one single output neuron.

Each blue node of the previous figure represents one neuron and processes its input data $\mathbf{u}$ by evaluating $\sigma\big(\theta + \sum_{i=1}^{m} a_i u_i\big)$, where $\mathbf{a}$ is a weight vector, θ a scalar-valued threshold and $\sigma : \mathbb{R} \to \mathbb{R}$ a non-linear function, the so-called activation function. Note that the former two quantities are individual for each neuron, whereas the latter is, in general, fixed for the whole network. The same holds for the orange node(s), however, sometimes, the activation function is skipped in the output neuron(s). Training such neural networks is an extremely difficult task, as non-convex optimization is involved. Moreover, due to the large number of parameters, interpretability is also an issue. This draws parallels to sparse principal component analysis.

Let us now have a closer look at the two central subproblems of our numerical method ARBeR, the optimization problems (4.31) and (4.32). As already observed, they are essentially of the form (4.33) and can be tackled by employing iterative thresholding, cf. Subsection 1.4.3. The respective update rules were given in (1.82), (1.90) and (1.94) taking the general form $\hat{\mathbf{x}}^k = \sigma(\mathbf{W}\hat{\mathbf{x}}^{k-1} + \mathbf{w})$, where σ denotes the respective scalar thresholding operator and is applied to each entry of a vector if its input is a vector. We note that one iteration of the thresholding algorithm can be described by one layer of a neural network. Thus, the update of one single component of our non-orthogonal matrix decomposition can be computed by a feedforward neural network, whose number of hidden layers plus the output layer matches the number of iterations. Each layer contains the same number of neurons, coinciding with the dimension of the component vector. As our approach ARBeR alternates between $\mathbf{u}$ and $\mathbf{v}$ components, this is followed by a further feedforward neural network of the same type. Only the number of neurons per layer as well as the activation function change. This alternating procedure is repeated over all component vectors and the outer loop of our numerical method, yielding a very deep neural network. Eventually, the desired component vectors are distributed over the output layers of the last $2R$ neural sub-networks.

Appendix

This additional part provides a collection of further material referred to in the thesis but skipped there for reasons of clarity and brevity.

A.1 Auxiliary Results

We start with some auxiliary tools, which supplement assertions from the thesis and complement certain steps in proofs.

Lemma A.2 (Relations of ℓ^q-Norms and ℓ^q-Quasi-Norms)**.** *Let $0 < q \leq p < \infty$. Then,*

(i) for $\mathbf{z} \in \ell^q(\mathbb{R})$ it holds $\|\mathbf{z}\|_p \leq \|\mathbf{z}\|_q$, i.e., $\ell^q(\mathbb{R}) \subset \ell^p(\mathbb{R})$, and

(ii) for $\mathbf{z} \in \mathbb{R}^N$ it also holds $\|\mathbf{z}\|_q \leq N^{1/q-1/p}\|\mathbf{z}\|_p$.

Proof. Firstly, for (i), we assume without loss of generality that $\|\mathbf{z}\|_q = 1$ so that in particular $|z_i| \leq 1$ for all $k \in \mathbb{N}$. Since $p \geq q$ we have $|z_i|^p = |z_i|^q |z_i|^{p-q} \leq |z_i|^q$ and $1/p \leq 1/q$, which is used in the first and second inequality of

$$\|\mathbf{z}\|_p = \left(\sum_{i=1}^{\infty} |z_i|^p\right)^{1/p} \leq \left(\sum_{i=1}^{\infty} |z_i|^q\right)^{1/p} \leq \left(\sum_{i=1}^{\infty} |z_i|^q\right)^{1/q} = \|\mathbf{z}\|_q = 1, \tag{A.6}$$

respectively. For a general non-zero $\mathbf{z} \in \ell^q(\mathbb{R})$ consider $\mathbf{z}/\|\mathbf{z}\|_q$.
Secondly, assertion (ii), follows by applying Hölder's inequality with $\tilde{p} := p/q \in (1, \infty)$ and $\tilde{q} = \tilde{p}/(\tilde{p}-1) = p/(p-q)$, as

$$\|\mathbf{z}\|_q^q = \sum_{i=1}^{N} (|z_i|^q \cdot 1) \leq \left(\sum_{i=1}^{N} (|z_i|^q)^{p/q}\right)^{q/p} \cdot N^{(p-q)/p} = \|\mathbf{z}\|_p^q \cdot N^{(p-q)/p}. \tag{A.7}$$

□

Remark A.3. If $\mathbf{z} \in \mathbb{R}^N$ is s-sparse, the statement from Lemma A.2(ii) can be sharpened by replacing N with s, i.e., for $\mathbf{z} \in \Sigma_s^N$ we have $\|\mathbf{z}\|_q \leq s^{1/q-1/p}\|\mathbf{z}\|_p$ for $0 < q \leq p < \infty$.

Lemma A.4. *Let $\mathbf{A} \in \mathbb{R}^{m \times N}$ and $\mathbf{y} \in \mathbb{R}^m$.*

(i) Let $\eta \geq 0$. If $\hat{\mathbf{z}}$ is a unique minimizer of (1.65), then there exists a parameter $s \in \mathbb{N}_0$ such that $\hat{\mathbf{z}}$ is also a unique minimizer of (1.75).

(ii) Conversely, let $s \in \mathbb{N}_0$. If $\hat{\mathbf{z}}$ is a unique minimizer of (1.75), there exists a parameter $\eta \geq 0$ such that $\hat{\mathbf{z}}$ is also a unique minimizer of (1.65).

Proof. Firstly, for (i), let us set $s = \|\hat{\mathbf{z}}\|_0 \in \mathbb{N}_0$ and consider all $\mathbf{z} \in \mathbb{R}^N$ with $\mathbf{z} \neq \hat{\mathbf{z}}$ and $\|\mathbf{z}\|_0 \leq s$. As $\hat{\mathbf{z}}$ is a unique minimizer of (1.65), for all such $\mathbf{z}$ it necessarily has to hold $\|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 > \eta \geq \|\mathbf{A}\hat{\mathbf{z}} - \mathbf{y}\|_2$. Thus, $\hat{\mathbf{z}}$ is the unique minimizer of (1.75).
Secondly, for (ii), let us set $\eta = \|\mathbf{A}\hat{\mathbf{z}} - \mathbf{y}\|_2$ and consider all $\mathbf{z} \in \mathbb{R}^N$ with $\mathbf{z} \neq \hat{\mathbf{z}}$ and $\|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta$. As $\hat{\mathbf{z}}$ is a unique minimizer of (1.75), for all such $\mathbf{z}$ it necessarily has to hold $\|\mathbf{z}\|_0 > s \geq \|\hat{\mathbf{z}}\|_0$. Thus, $\hat{\mathbf{z}}$ is the unique minimizer of (1.65). □

Lemma A.5 (Schur Complement)**.** *Let $\mathbf{M} \in \mathbb{R}^{(\nu_1+\nu_2)\times(\nu_1+\nu_2)}$ denote a symmetric matrix of the form*

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix} \tag{A.8}$$

with symmetric $\mathbf{A} \in \mathbb{R}^{\nu_1\times\nu_1}$ and $\mathbf{C} \in \mathbb{R}^{\nu_2\times\nu_2}$ and arbitrary $\mathbf{B} \in \mathbb{R}^{\nu_1\times\nu_2}$. Let $\mathbf{A}$ be additionally invertible. Then

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix} = \begin{pmatrix} \mathbf{Id}_{\nu_1} & \mathbf{0} \\ \mathbf{B}^T\mathbf{A}^{-1} & \mathbf{Id}_{\nu_2} \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B} \end{pmatrix} \begin{pmatrix} \mathbf{Id}_{\nu_1} & \mathbf{A}^{-1}\mathbf{B} \\ \mathbf{0} & \mathbf{Id}_{\nu_2} \end{pmatrix}. \tag{A.9}$$

Moreover, $\mathbf{M} \succeq 0$ if and only if $\mathbf{A} \succ 0$ and $\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B} \succeq 0$.

Proof. The matrix equality obviously holds. The second statement follows since a block diagonal matrix is positive (semi)definite if and only if each of its diagonal blocks is positive (semi)definite. □

Lemma A.6 (Norm Equivalence for a Matrix with a Non-Orthogonal Decomposition)**.** *Let $\mathbf{Z} \in \mathbb{R}^{n_1\times n_2}$ possess a decomposition of the form*

$$\mathbf{Z} = \underline{\mathbf{U}}\mathbf{\Sigma}\underline{\mathbf{V}}^T = \sum_{r=1}^{R} \sigma_r \underline{\mathbf{u}}_r \left(\underline{\mathbf{v}}_r\right)^T, \tag{A.10}$$

where $\{\underline{\mathbf{u}}_r\}_{r=1}^R \subset \mathbb{R}^{n_1}$ denotes a set of linearly independent, yet, maybe non-orthogonal, vectors of unit norm, $\{\underline{\mathbf{v}}_r\}_{r=1}^R \subset \mathbb{R}^{n_2}$ a set of vectors of unit norm and $\{\sigma_r\}_{r=1}^R \subset \mathbb{R}_+$ a set of positive scalars. Then,

$$c_{\underline{\mathbf{U}}}\|\mathbf{\Sigma}\|_F \leq \|\mathbf{Z}\|_F \leq C_{\underline{\mathbf{U}}}\|\mathbf{\Sigma}\|_F \tag{A.11}$$

with constants $c_{\underline{\mathbf{U}}} = \sqrt{\lambda_{\min}(\underline{\mathbf{U}}^T\underline{\mathbf{U}})}$ and $C_{\underline{\mathbf{U}}} = \sqrt{\lambda_{\max}(\underline{\mathbf{U}}^T\underline{\mathbf{U}})}$.

Proof. By rewriting the squared Frobenius norm of $\mathbf{Z}$ we observe that

$$\|\mathbf{Z}\|_F^2 = \sum_{j=1}^{n_2}\sum_{i=1}^{n_1}\left|\sum_{r=1}^{R}\sigma_r \underline{u}_{ri}\underline{v}_{rj}\right|^2 = \sum_{j=1}^{n_2}\left\|\sum_{r=1}^{R}\sigma_r\underline{v}_{rj}\underline{\mathbf{u}}_r\right\|_2^2 = \sum_{j=1}^{n_2}\left\|\underline{\mathbf{U}}\begin{pmatrix}\sigma_1\underline{v}_{1j}\\ \vdots \\ \sigma_R\underline{v}_{Rj}\end{pmatrix}\right\|_2^2. \tag{A.12}$$

Let us now define the auxiliary vector $\mathbf{w}_j = (\sigma_1\underline{v}_{1j},\ldots,\sigma_R\underline{v}_{Rj})^T \in \mathbb{R}^R$ and note that for fixed j it holds $\lambda_{\min}(\underline{\mathbf{U}}^T\underline{\mathbf{U}})\|\mathbf{w}_j\|_2^2 \leq \|\underline{\mathbf{U}}\mathbf{w}_j\|_2^2 \leq \lambda_{\max}(\underline{\mathbf{U}}^T\underline{\mathbf{U}})\|\mathbf{w}_j\|_2^2$. With this we conclude

$$\|\mathbf{Z}\|_F^2 = \sum_{j=1}^{n_2}\|\underline{\mathbf{U}}\mathbf{w}_j\|_2^2 \simeq \sum_{j=1}^{n_2}\|\mathbf{w}_j\|_2^2 = \sum_{j=1}^{n_2}\sum_{r=1}^{R}|\sigma_r\underline{v}_{rj}|^2 = \sum_{r=1}^{R}|\sigma_r|^2\|\underline{\mathbf{v}}_r\|^2 = \sum_{r=1}^{R}|\sigma_r|^2 = \|\mathbf{\Sigma}\|_F^2. \tag{A.13}$$

for hidden constants $c_{\underline{\mathbf{U}}}^2 = \lambda_{\min}(\underline{\mathbf{U}}^T\underline{\mathbf{U}})$ and $C_{\underline{\mathbf{U}}}^2 = \lambda_{\max}(\underline{\mathbf{U}}^T\underline{\mathbf{U}})$. □

Remark A.7. If the decomposition in (A.10) is a singular value decomposition of $\mathbf{Z}$, the statement of Lemma A.6 reduces to the well-known fact $\|\mathbf{Z}\|_F = \|\boldsymbol{\Sigma}\|_F$.

Lemma A.8 ([FMN19, Lemma 4.1])**.** *Let $\alpha, \beta, a, b, p, q > 0$. Then*

$$f : \mathbb{R}_+ \to \mathbb{R}, \quad \lambda \mapsto f(\lambda) = \alpha\lambda^p a + \beta\lambda^{-q} b \tag{A.14}$$

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

$$\min f = f(\tilde{\lambda}) = C_{pq}(\alpha a)^{\frac{q}{p+q}}(\beta b)^{\frac{p}{p+q}}, \tag{A.15}$$

where $C_{pq} = \left(\frac{q}{p}\right)^{\frac{p}{p+q}} + \left(\frac{p}{q}\right)^{\frac{q}{p+q}}$.

Proof. Differentiating f shows $f'(\lambda) = \alpha p \lambda^{p-1} a - \beta q \lambda^{-q-1} b$, yielding the critical point when set equal to zero. By considering the limits for $\lambda \to 0^+$ and $\lambda \to \infty$, it is a minimizer. The minimal value can be computed immediately. □

Lemma A.9 (Monotonicity of $K_s^{q,N}$ in q)**.** *Let $0 < q_1, q_2 \leq 2$. If $q_1 \leq q_2$, then it holds $K_s^{q_1,N} \subset K_s^{q_2,N}$.*

Proof. Let us firstly exclude the two cases $q_2 = 2$ and $q_1 = q_2$, in which the statement is obvious. Thus, we can restrict ourselves to the situation $0 < q_1 < q_2 < 2$. Now, let $\mathbf{z} \in K_s^{q_1,N}$ and without loss of generality assume that $\|\mathbf{z}\|_2 = 1$.
Since $q_2 < 2$ we have $\tilde{p} := (2-q_1)/(q_2-q_1) \in (1,\infty)$ and $\tilde{q} = \tilde{p}/(\tilde{p}-1) = (2-q_1)/(2-q_2)$. Applying Hölder's inequality in the last step of the first line yields

$$\begin{aligned}
\|\mathbf{z}\|_{q_2}^{q_2} &= \sum_{i=1}^N |z_i|^{q_2} = \sum_{i=1}^N |z_i|^{2/\tilde{p}} |z_i|^{q_2-2/\tilde{p}} \leq \left(\sum_{i=1}^N |z_i|^2\right)^{1/\tilde{p}} \left(\sum_{i=1}^N |z_i|^{\tilde{q}(q_2-2/\tilde{p})}\right)^{1/\tilde{q}} \\
&= \|\mathbf{z}\|_2^{2/\tilde{p}} \left(\sum_{i=1}^N |z_i|^{q_1}\right)^{(2-q_2)/(2-q_1)} \qquad = \|\mathbf{z}\|_{q_1}^{q_1(2-q_2)/(2-q_1)} \\
&\leq \left(s^{1/q_1-1/2}\right)^{q_1(2-q_2)/(2-q_1)} = s^{1-q_2/2},
\end{aligned} \tag{A.16}$$

where the next-to-last step involves the assumption that $\mathbf{z} \in K_s^{q_1,N}$, i.e., $\|\mathbf{z}\|_{q_1}^{q_1} \leq s^{1-q_1/2}$. For a general non-zero $\mathbf{z} \in K_s^{q_1,N}$ consider $\mathbf{z}/\|\mathbf{z}\|_2$. □

A.2 An Upper Bound on Dudley's Integral for $\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}$

In this technical section we compute bounds for the individual integrals $I_1, \ldots, I_6$ appearing in inequality (4.88). To be precise, we consider the integrals I_1, I_3, I_6 and $I_2 + I_5, I_4$, noting that two different types are apparent. To keep the notation compact, we make use of the abbreviation $I_{25} = I_2 + I_5$
Let us estimate the integrals I_1, I_3 and I_6 first. We start with a second change of variables, namely $\epsilon'' = \epsilon'/(30\Gamma R)$, which yields the first equality in the following computation.

Furthermore, we employ Cauchy-Schwarz inequality to obtain the second line. After an evaluation of the remaining integral we have

$$\begin{aligned}
I_1 &= 30\Gamma R\sqrt{R(n_1+n_2+1)}\int_0^{\frac{2}{5\sqrt{R}}(s_2/n_2)^{1/q-1/2}} \sqrt{\log\left(\frac{1}{\epsilon''}\right)}\,\mathrm{d}\epsilon'' \\
&\leq 30\Gamma R\sqrt{R(n_1+n_2+1)}\sqrt{\int_0^{\frac{2}{5\sqrt{R}}(s_2/n_2)^{1/q-1/2}} 1^2\,\mathrm{d}\epsilon'' \int_0^{\frac{2}{5\sqrt{R}}(s_2/n_2)^{1/q-1/2}} \log\left(\frac{1}{\epsilon''}\right)\mathrm{d}\epsilon''} \\
&= 30\Gamma R\sqrt{R(n_1+n_2+1)}\sqrt{\frac{2}{5\sqrt{R}}\left(\frac{s_2}{n_2}\right)^{1/q-1/2}\epsilon''\left(1+\log\left(\frac{1}{\epsilon''}\right)\right)\Bigg|_0^{\frac{2}{5\sqrt{R}}(s_2/n_2)^{1/q-1/2}}} \\
&= 12\Gamma R\sqrt{n_1+n_2+1}\,\left(\frac{s_2}{n_2}\right)^{1/q-1/2}\sqrt{1+\log\left(\frac{5\sqrt{R}}{2}\left(\frac{s_2}{n_2}\right)^{1/2-1/q}\right)} \\
&\leq \begin{cases} \left(432\Gamma^2 R^2\left(\frac{s_2}{n_2}\right)^{2/q-2} s_2 \log\left(\frac{5e\sqrt{R}}{2}\left(\frac{s_2}{n_2}\right)^{1/2-1/q}\right)\right)^{1/2} & \text{if } n_2 \geq n_1, \\ \left(288\Gamma^2 R^2\left(\frac{s_1}{n_1}\right)^{2/q-2} s_1 \log\left(\frac{5e\sqrt{R}}{2}\left(\frac{s_2}{n_2}\right)^{1/2-1/q}\right)\right)^{1/2} & \text{if } n_2 < n_1, \end{cases}
\end{aligned} \tag{A.17}$$

where the last inequality makes use of the assumption $s_2/n_2 \leq s_1/n_1$ for the case $n_2 < n_1$. Apart from the assumption $s_2/n_2 \leq s_1/n_1$ entering additionally in the fourth line, we proceed similarly for I_3, which results in

$$\begin{aligned}
I_3 &= 30\Gamma R\sqrt{R(n_1+1)}\int_{\frac{2}{5\sqrt{R}}(s_2/n_2)^{1/q-1/2}}^{\frac{2}{5\sqrt{R}}(s_1/n_1)^{1/q-1/2}} \sqrt{\log\left(\frac{1}{\epsilon''}\right)}\mathrm{d}\epsilon'' \\
&\leq 30\Gamma R\sqrt{R(n_1+1)}\sqrt{\int_{\frac{2}{5\sqrt{R}}(s_2/n_2)^{1/q-1/2}}^{\frac{2}{5\sqrt{R}}(s_1/n_1)^{1/q-1/2}} 1^2\,\mathrm{d}\epsilon'' \int_{\frac{2}{5\sqrt{R}}(s_2/n_2)^{1/q-1/2}}^{\frac{2}{5\sqrt{R}}(s_1/n_1)^{1/q-1/2}} \log\left(\frac{1}{\epsilon''}\right)\mathrm{d}\epsilon''} \\
&= 30\Gamma R\sqrt{R(n_1+1)}\sqrt{\frac{2}{5\sqrt{R}}\left(\left(\frac{s_1}{n_1}\right)^{1/q-1/2}-\left(\frac{s_2}{n_2}\right)^{1/q-1/2}\right)\epsilon''\left(1+\log\left(\frac{1}{\epsilon''}\right)\right)\Bigg|_{\frac{2}{5\sqrt{R}}(s_2/n_2)^{1/q-1/2}}^{\frac{2}{5\sqrt{R}}(s_1/n_1)^{1/q-1/2}}} \\
&\leq 12\Gamma R\sqrt{n_1+1}\left(\left(\frac{s_1}{n_1}\right)^{1/q-1/2}-\left(\frac{s_2}{n_2}\right)^{1/q-1/2}\right)\sqrt{1+\log\left(\frac{5\sqrt{R}}{2}\left(\frac{s_2}{n_2}\right)^{1/2-1/q}\right)} \\
&\leq \left(144\Gamma^2 R^2(n_1+1)\left(\left(\frac{s_1}{n_1}\right)^{2/q-1}+\left(\frac{s_2}{n_2}\right)^{2/q-1}\right)\left(1+\log\left(\frac{5\sqrt{R}}{2}\left(\frac{s_2}{n_2}\right)^{1/2-1/q}\right)\right)\right)^{1/2} \\
&\leq \begin{cases} \left(288\Gamma^2 R^2\left(\left(\frac{s_1}{n_1}\right)^{2/q-2}s_1+\left(\frac{s_2}{n_2}\right)^{2/q-2}s_2\right)\log\left(\frac{5e\sqrt{R}}{2}\left(\frac{s_2}{n_2}\right)^{1/2-1/q}\right)\right)^{1/2} & \text{if } n_2 \geq n_1, \\ \left(432\Gamma^2 R^2\left(\frac{s_1}{n_1}\right)^{2/q-2}s_1\log\left(\frac{5e\sqrt{R}}{2}\left(\frac{s_2}{n_2}\right)^{1/2-1/q}\right)\right)^{1/2} & \text{if } n_2 < n_1, \end{cases}
\end{aligned} \tag{A.18}$$

where the next-to-last inequality follows as $(x-y)^2 \leq x^2+y^2$ for all $x, y \geq 0$.

Upper bounding the last integral of this type, namely I_6, essentially follows the lines of the former I_3 and uses $s_1/n_1 \leq 1$ in the last inequality. Thus we get

$$\begin{aligned}
I_6 &= 30\Gamma R\sqrt{R}\int_{\frac{2}{5\sqrt{R}}(s_1/n_1)^{1/q-1/2}}^{\frac{1}{30\sqrt{R}}}\sqrt{\log\left(\frac{1}{\epsilon''}\right)}\,\mathrm{d}\epsilon'' \\
&\leq 30\Gamma R\sqrt{R}\sqrt{\int_{\frac{2}{5\sqrt{R}}(s_1/n_1)^{1/q-1/2}}^{\frac{1}{30\sqrt{R}}}1^2\,\mathrm{d}\epsilon''\int_{\frac{2}{5\sqrt{R}}(s_1/n_1)^{1/q-1/2}}^{\frac{1}{30\sqrt{R}}}\log\left(\frac{1}{\epsilon''}\right)\mathrm{d}\epsilon''} \\
&= 30\Gamma R\sqrt{R}\sqrt{\frac{1}{30\sqrt{R}}\left(1-12\left(\frac{s_1}{n_1}\right)^{1/q-1/2}\right)\epsilon''\left(1+\log\left(\frac{1}{\epsilon''}\right)\right)\Bigg|_{\frac{2}{5\sqrt{R}}(s_1/n_1)^{1/q-1/2}}^{\frac{1}{30\sqrt{R}}}} \qquad \text{(A.19)}\\
&\leq \Gamma R\left(1-12\left(\frac{s_1}{n_1}\right)^{1/q-1/2}\right)\sqrt{1+\log\left(\frac{5\sqrt{R}}{2}\left(\frac{s_1}{n_1}\right)^{1/2-1/q}\right)} \\
&\leq \left(\Gamma^2R^2\left(1+144\left(\frac{s_1}{n_1}\right)^{2/q-1}\right)\left(1+\log\left(\frac{5\sqrt{R}}{2}\left(\frac{s_1}{n_1}\right)^{1/2-1/q}\right)\right)\right)^{1/2} \\
&\leq \left(145\Gamma^2R^2\log\left(\frac{5e\sqrt{R}}{2}\left(\frac{s_1}{n_1}\right)^{1/2-1/q}\right)\right)^{1/2}.
\end{aligned}$$

Now, let us turn to the remaining integrals and derive upper bounds for $I_{25} = I_2 + I_5$ and I_4. The change of variables $\epsilon'' = \epsilon'/(12\Gamma\sqrt{R})$ yields

$$I_{25} = 12\Gamma R\sqrt{s_2}\int_{(s_2/n_2)^{1/q-1/2}}^{1/12}\left(\frac{1}{\epsilon''}\right)^{q/(2-q)}\sqrt{\log\left(\frac{6en_2}{s_2}\epsilon''^{(3q-2)/(2-q)}\right)}\,\mathrm{d}\epsilon'', \qquad \text{(A.20)}$$

and

$$I_4 = 12\Gamma R\sqrt{s_1}\int_{(s_1/n_1)^{1/q-1/2}}^{1/12}\left(\frac{1}{\epsilon''}\right)^{q/(2-q)}\sqrt{\log\left(\frac{6en_1}{s_1}\epsilon''^{(3q-2)/(2-q)}\right)}\,\mathrm{d}\epsilon'', \qquad \text{(A.21)}$$

respectively. Due to the analogy of both integrands and their integration bounds it suffices to estimate the first one and substitute s_2 by s_1 and n_2 by n_1 in order to obtain a bound for the latter one as well.

Unfortunately, for $q \neq 2/3$ or $q \neq 1$, the integrand in (A.20) admits no elementary antiderivative as it would involve the error function. This in turn necessitates upper bounding the integrands beforehand. Therefore, let $\mu, \nu \in \mathbb{R}$ and consider

$$\left(\frac{1}{\epsilon''}\right)^{q/(2-q)-\mu}\left(\frac{1}{\epsilon''}\right)^{\mu}\sqrt{\log\left(\frac{6en_2}{s_2}\epsilon''^{(3q-2)/(2-q)-\nu}\epsilon''^{\nu}\right)}, \qquad \text{(A.22)}$$

where the gray terms highlight the ones to be bounded in advance. We note that for $(\mu,\nu) \in \big\{\{1\}\times\mathbb{R}\big\} \cup \big\{\mathbb{R}\times\{0\}\big\}$ we can find elementary antiderivatives.

Let us firstly cover the special case $q = 2/3$. A straightforward computation shows

$$\begin{aligned} I_{25} &= 12\Gamma R\sqrt{s_2}\int_{s_2/n_2}^{1/12}\left(\frac{1}{\epsilon''}\right)^{1/2}\sqrt{\log\left(\frac{6en_2}{s_2}\right)}\,\mathrm{d}\epsilon'' = 24\Gamma R\sqrt{s_2\log\left(\frac{6en_2}{s_2}\right)}\sqrt{\epsilon''}\Big|_{s_2/n_2}^{1/12} \\ &= 24\Gamma R\sqrt{s_2\log\left(\frac{6en_2}{s_2}\right)}\left(\sqrt{\frac{1}{12}}-\sqrt{\frac{s_2}{n_2}}\right) \le 24\Gamma R\sqrt{\frac{1}{12}}\sqrt{s_2\log\left(\frac{6en_2}{s_2}\right)} \\ &= \left(48\Gamma^2R^2s_2\log\left(\frac{6en_2}{s_2}\right)\right)^{1/2}. \end{aligned} \tag{A.23}$$

For $0 < q < 2/3$, it turns out that among the feasible parameters μ and ν the optimal upper bound is obtained by adapting the computations from the case $q = 2/3$. More precisely, choosing $\mu = q/(2-q)$ and $\nu = 0$ yields

$$\begin{aligned} I_{25} &\le 12\Gamma R\sqrt{s_2}\int_{(s_2/n_2)^{1/q-1/2}}^{1/12}\left(\frac{1}{\epsilon''}\right)^{q/(2-q)}\sqrt{\log\left(6e\left(\frac{s_2}{n_2}\right)^{(q-2)/(2q)}\right)}\,\mathrm{d}\epsilon'' \\ &= 12\Gamma R\sqrt{s_2\log\left(6e\left(\frac{s_2}{n_2}\right)^{(q-2)/(2q)}\right)}\left(\frac{2-q}{2-2q}\right)(\epsilon'')^{(2-2q)/(2-q)}\Big|_{(s_2/n_2)^{1/q-1/2}}^{1/12} \\ &= 12\Gamma R\left(\frac{2-q}{2-2q}\right)\sqrt{s_2\log\left(6e\left(\frac{s_2}{n_2}\right)^{(q-2)/(2q)}\right)}\left(\left(\frac{1}{12}\right)^{(2-2q)/(2-q)}-\left(\frac{s_2}{n_2}\right)^{1/q-1}\right) \\ &\le 12\Gamma R\left(\frac{2-q}{2-2q}\right)\left(\frac{1}{12}\right)^{(2-2q)/(2-q)}\sqrt{s_2\log\left(6e\left(\frac{s_2}{n_2}\right)^{(q-2)/(2q)}\right)} \\ &= \left(144^{q/(2-q)}\left(\frac{2-q}{2-2q}\right)^2\Gamma^2R^2s_2\log\left(6e\left(\frac{s_2}{n_2}\right)^{(q-2)/(2q)}\right)\right)^{1/2}. \end{aligned} \tag{A.24}$$

In contrast, in the case $2/3 < q < 1$, the same methodology yields the bound

$$\begin{aligned} I_{25} &\le 12\Gamma R\sqrt{s_2}\int_{(s_2/n_2)^{1/q-1/2}}^{1/12}\left(\frac{1}{\epsilon''}\right)^{q/(2-q)}\sqrt{\log\left(\frac{6en_2}{s_2}\left(\frac{1}{12}\right)^{(3q-2)/(2-q)}\right)}\,\mathrm{d}\epsilon'' \\ &= 12\Gamma R\left(\frac{2-q}{2-2q}\right)\sqrt{s_2\log\left(\frac{6en_2}{s_2}\left(\frac{1}{12}\right)^{(3q-2)/(2-q)}\right)}\left(\left(\frac{1}{12}\right)^{(2-2q)/(2-q)}-\left(\frac{s_2}{n_2}\right)^{1/q-1}\right) \\ &\le 12\Gamma R\left(\frac{2-q}{2-2q}\right)\left(\frac{1}{12}\right)^{(2-2q)/(2-q)}\sqrt{s_2\log\left(\frac{6en_2}{s_2}\left(\frac{1}{12}\right)^{(3q-2)/(2-q)}\right)} \\ &= \left(144^{q/(2-q)}\left(\frac{2-q}{2-2q}\right)^2\Gamma^2R^2s_2\log\left(\frac{6en_2}{s_2}\left(\frac{1}{12}\right)^{(3q-2)/(2-q)}\right)\right)^{1/2}, \end{aligned} \tag{A.25}$$

which is unfortunately not optimal for the whole range of q due to the singularity in $q = 1$ in (A.25). For q sufficiently close to 1, a better bound can be obtained by choosing $\mu = 1$

and $\nu = (3q-2)/(2-q)$, which results in

$$\begin{aligned}
I_{25} &\le 12\Gamma R\sqrt{s_2}\int_{(s_2/n_2)^{1/q-1/2}}^{1/12}\left(\frac{1}{12}\right)^{(2-2q)/(2-q)}\left(\frac{1}{\epsilon''}\right)\sqrt{\log\left(\frac{6en_2}{s_2}\epsilon''^{(3q-2)/(2-q)}\right)}\,\mathrm{d}\epsilon''\\
&= 12\Gamma R\left(\frac{1}{12}\right)^{(2-2q)/(2-q)}\sqrt{s_2}\left(\frac{2}{3}\frac{2-q}{3q-2}\right)\log^{3/2}\left(\frac{6en_2}{s_2}\epsilon''^{(3q-2)/(2-q)}\right)\Bigg|_{(s_2/n_2)^{1/q-1/2}}^{1/12}\\
&= 8\Gamma R\left(\frac{2-q}{3q-2}\right)\left(\frac{1}{12}\right)^{(2-2q)/(2-q)}\sqrt{s_2}\left(\log^{3/2}\left(\frac{6en_2}{s_2}\left(\frac{1}{12}\right)^{(3q-2)/(2-q)}\right)-\log^{3/2}\left(6e\left(\frac{s_2}{n_2}\right)^{1/2-1/q}\right)\right)\\
&\le 8\Gamma R\left(\frac{2-q}{3q-2}\right)\left(\frac{1}{12}\right)^{(2-2q)/(2-q)}\sqrt{s_2}\log^{3/2}\left(\frac{6en_2}{s_2}\left(\frac{1}{12}\right)^{(3q-2)/(2-q)}\right)\\
&= \left(144^{q/(2-q)}\left(\frac{2}{3}\frac{2-q}{3q-2}\right)^2\Gamma^2R^2s_2\log^3\left(\frac{6en_2}{s_2}\left(\frac{1}{12}\right)^{(3q-2)/(2-q)}\right)\right)^{1/2}.
\end{aligned}\tag{A.26}$$

Note that by construction, for $q = 2/3$, the bounds (A.24) and (A.25) coincide with (A.23).

Let us now turn to the second special case $q = 1$ and reproduce the computations from [FMN19, Lemma 7.1, Integrals I_2 and I_4]. We get

$$\begin{aligned}
I_{25} &= 12\Gamma R\sqrt{s_2}\int_{(s_2/n_2)^{1/2}}^{1/12}\left(\frac{1}{\epsilon''}\right)\sqrt{\log\left(\frac{6en_2}{s_2}\epsilon''\right)}\,\mathrm{d}\epsilon'' = 8\Gamma R\sqrt{s_2}\log^{3/2}\left(\frac{6en_2}{s_2}\epsilon''\right)\Bigg|_{(s_2/n_2)^{1/2}}^{1/12}\\
&= 8\Gamma R\sqrt{s_2}\left(\log^{3/2}\left(\frac{en_2}{2s_2}\right)-\log^{3/2}\left(6e\left(\frac{s_2}{n_2}\right)^{-1/2}\right)\right) \le 8\Gamma R\sqrt{s_2}\log^{3/2}\left(\frac{en_2}{2s_2}\right)\\
&= \left(64\Gamma^2R^2s_2\log^3\left(\frac{en_2}{2s_2}\right)\right)^{1/2}.
\end{aligned}\tag{A.27}$$

Note that for $q \to 1$ the bound (A.26) converges pointwise in s_2 to the bound in (A.27). As already mentioned previously, bounds on I_4 are obtained analogously.

To conclude, we need to put the obtained individual bounds together by making use of the basic inequality $x+y \le \sqrt{2(x^2+y^2)}$ for all $x, y \in \mathbb{R}$. For the first type of integral we obtain

$$\begin{aligned}
I_1 + I_3 + I_6 \le \Bigg(CR^2&\left(\left(\frac{s_1}{n_1}\right)^{2/q-2}s_1+\left(\frac{s_2}{n_2}\right)^{2/q-2}s_2+1\right)\\
&\cdot\log\left(\frac{5e\sqrt{R}}{2}\max\left\{\left(\frac{s_1}{n_1}\right)^{1/2-1/q},\left(\frac{s_2}{n_2}\right)^{1/2-1/q}\right\}\right)\Bigg)^{1/2},
\end{aligned}\tag{A.28}$$

where $C = 12\cdot 144\Gamma^2$ denotes an absolute constant.

Similarly, for the second type of integral we get

$$I_2 + I_4 + I_5 \\ \leq \begin{cases} \big(R^2\big(s_1 \cdot \mathbb{1}_{S_1(q)}(s_1) + s_2 \cdot \mathbb{1}_{S_2(q)}(s_2)\big) L_q^1\big(\tfrac{s_1}{n_1}, \tfrac{s_2}{n_2}\big)\big)^{1/2} & \text{if } q \in (0, 2/3], \\ \big(R^2\big(s_1 \cdot \mathbb{1}_{S_1(q)}(s_1) + s_2 \cdot \mathbb{1}_{S_2(q)}(s_2)\big)\min\big\{L_q^1\big(\tfrac{s_1}{n_1}, \tfrac{s_2}{n_2}\big), L_q^3\big(\tfrac{s_1}{n_1}, \tfrac{s_2}{n_2}\big)\big\}\big)^{1/2} & \text{if } q \in (2/3, 1), \\ \big(R^2\big(s_1 \cdot \mathbb{1}_{S_1(q)}(s_1) + s_2 \cdot \mathbb{1}_{S_2(q)}(s_2)\big) L_1^3\big(\tfrac{s_1}{n_1}, \tfrac{s_2}{n_2}\big)\big)^{1/2} & \text{if } q = 1, \end{cases} \tag{A.29}$$

where

$$S_i(q) = \Big\{s_i < \big(\tfrac{1}{12}\big)^{2q/(2-q)} n_i\Big\} \tag{A.30}$$

denotes the set of the indicator $\mathbb{1}_{S_i(q)}(s_i)$ for $i = 1, 2$. This indicator function comes into play since the integrals I_4 and I_{25} are only active if the sparsities s_1 and s_2 are sufficiently small. We furthermore collect constants, which may depend on q, together with the logarithmic terms in the function

$$L_q^\iota\Big(\frac{s_1}{n_1}, \frac{s_2}{n_2}\Big) = C_q^\iota\Big(L_q\Big(\frac{s_1}{n_1}, \frac{s_2}{n_2}\Big)\Big)^\iota, \tag{A.31}$$

where

$$C_q^\iota = \begin{cases} 2 \cdot 144^{q/(2-q)}\big(\tfrac{2-q}{2-2q}\big)^2 \Gamma^2 & \text{if } \iota = 1, \\ 2 \cdot 144^{q/(2-q)}\big(\tfrac{2}{3}\tfrac{2-q}{3q-2}\big)^2 \Gamma^2 & \text{if } \iota = 3, \end{cases} \tag{A.32}$$

and

$$L_q\Big(\frac{s_1}{n_1}, \frac{s_2}{n_2}\Big) = \begin{cases} \log\Big(6e \max\Big\{\big(\tfrac{s_1}{n_1}\big)^{(q-2)/(2q)}, \big(\tfrac{s_2}{n_2}\big)^{(q-2)/(2q)}\Big\}\Big) & \text{if } q \in (0, 2/3], \\ \log\Big(6e\big(\tfrac{1}{12}\big)^{(3q-2)/(2-q)} \max\Big\{\big(\tfrac{s_1}{n_1}\big)^{-1}, \big(\tfrac{s_2}{n_2}\big)^{-1}\Big\}\Big) & \text{if } q \in [2/3, 1]. \end{cases} \tag{A.33}$$

Note that for $0 < q \leq 2/3$ we have $(2-q)/(2-2q) \leq 2$ and that for $2/3 < q \leq 1$ we have $\min\{(2-q)/(2-2q), (2(2-q))/(3(3q-2))\} \leq 2\sqrt{2}$.

Neglecting constants and summarizing all logarithmic terms in one polylogarithmic term we arrive at

$$\begin{aligned} I &\lesssim \Bigg(\Gamma^2 R^2 \Bigg(\Bigg(\Big(\frac{s_1}{n_1}\Big)^{2/q-2} s_1 + \Big(\frac{s_2}{n_2}\Big)^{2/q-2} s_2 + 1\Bigg) + 144^{(2q-2)/(2-q)}(s_1 + s_2)\Bigg) \\ &\qquad \cdot \operatorname{polylog}\Bigg(e\sqrt{R}\max\Bigg\{\Big(\frac{s_1}{n_1}\Big)^{-\max\{1/q-1/2,1\}}, \Big(\frac{s_2}{n_2}\Big)^{-\max\{1/q-1/2,1\}}\Bigg\}\Bigg)\Bigg)^{1/2} \\ &= \Bigg(\Gamma^2 R^2 \Bigg(\Bigg(\Big(\frac{s_1}{n_1}\Big)^{2/q-2} s_1 + \Big(\frac{s_2}{n_2}\Big)^{2/q-2} s_2 + 1\Bigg) + 144^{(2q-2)/(2-q)}(s_1 + s_2)\Bigg) \\ &\qquad \cdot \operatorname{polylog}\Bigg(e\sqrt{R}\max\Big\{\frac{n_1}{s_1}, \frac{n_2}{s_2}\Big\}^{\max\{1/q-1/2,1\}}\Bigg)\Bigg)^{1/2}. \end{aligned} \tag{A.34}$$

Let us emphasize that the hidden constant is independent of the parameter q.

Remark A.10. The prefactor $\Gamma^2 R^2$ should be read as $\Gamma^2 R \cdot R$. The latter factor, R, belongs to the lower bound on the information theoretic complexity of the matrix set $\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}$, which is of order $\mathcal{O}(R(s_1 + s_2))$, see, e.g., the computation after equation (3.6). The former factor, $\Gamma^2 R$, is the squared Frobenius radius $d_F\big(\mathcal{H}_{\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}}\big)$ of the auxiliary matrix set $\mathcal{H}_{\mathcal{K}_{s_1,s_2}^{q,R,\Gamma}}$.

Remark A.11. We want to note that the two cases $q = 2/3$ and $q = 1$ seem to be special in the sense that the integrands can be integrated directly without further bounding. Furthermore, recall that $q = 2/3$ and $q = 1$ were also cases where the thresholding operators could be determined explicitly, cf. Paragraphs 1.4.3(1) and 1.4.3(2). However, despite the latter being true also for $q = 1/2$, the antiderivative of (A.20) involves the error function in this case.

List of Figures

Bibliography

[ABRS10] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. *Math. Oper. Res.*, 35(2):438–457, 2010.

[ABS13] Hédy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Program., Ser. A*, 137(1-2):91–129, 2013.

[Ach03] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003. Special issue on PODS 2001 (Santa Barbara, CA).

[ALMT13] Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: a geometric theory of phase transitions in convex optimization. arXiv:1303.6672, 2013.

[BD08] Thomas Blumensath and Mike E. Davies. Iterative thresholding for sparse approximations. *J. Fourier Anal. Appl.*, 14(5-6):629–654, 2008.

[BD09] Thomas Blumensath and Mike E. Davies. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.*, 27(3):265–274, 2009.

[BDDW08] Richard Baraniuk, Mark A. Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28(3):253–263, 2008.

[BDLS07] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM J. Optim.*, 18(2):556–572, 2007.

[BL07] James Bennett and Stan Lanning. The Netflix prize. In *Proceedings of KDD cup and workshop*, pages 3–6, 2007.

[BLR15] Kristian Bredies, Dirk A. Lorenz, and Stefan Reiterer. Minimization of non-smooth, non-convex functionals by iterative thresholding. *J. Optim. Theory Appl.*, 165(1):78–112, 2015.

[Bre95] Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, 2004.

[Can08] Emmanuel J. Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris*, 346(9-10):589–592, 2008.

[CDD09] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22(1):211–231, 2009.

[CDS98] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[CLMW11] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):Art. 11, 37 pages, 2011.

[CP11] Emmanuel J. Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inform. Theory*, 57(4):2342–2359, 2011.

[CR09] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[CRT06a] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[CRT06b] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[CRTV05] Emmanuel J. Candès, Mark Rudelson, Terence Tao, and Roman Vershynin. Error correction via linear programming. In *46th Annual IEEE Symposium on Foundations of Computer Science*, FOCS'05, pages 668–681, 2005.

[CS08] Rick Chartrand and Valentina Staneva. Restricted isometry properties and nonconvex compressive sensing. *Inverse Probl.*, 24(3):035020, 14 pages, 2008.

[CSX13] Wenfei Cao, Jian Sun, and Zongben Xu. Fast image deconvolution using closed-form thresholding formulas of $L_q(q = \frac{1}{2}, \frac{2}{3})$ regularization. *J. Vis. Commun. Image Represent.*, 24(1):31–41, 2013.

[CT05] Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005.

[CT06] Emmanuel J. Candès and Terence Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[CT10] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010.

[DDDM04] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pure Appl. Math.*, 57(11):1413–1457, 2004.

[DDT$^+$08] Marco F. Duarte, Mark A. Davenport, Dharmpal Takhar, Jason N. Laska, Ting Sun, Kevin F. Kelly, and Richard G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Process. Mag.*, 25(2):83–91, 2008.

[DeV98] Ronald A. DeVore. Nonlinear approximation. In *Acta Numerica*, volume 7, pages 51–150. Cambridge University Press, Cambridge, 1998.

[Don06] David L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[DR16] Mark A. Davenport and Justin K. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE J. Sel. Topics Signal Process.*, 10(4):608–622, 2016.

[DW10] Mark A. Davenport and Michael B. Wakin. Analysis of orthogonal matching pursuit using the restricted isometry property. *IEEE Trans. Inform. Theory*, 56(9):4395–4401, 2010.

[Eco10] The Economist. Data, data everywhere: a special report on managing information. 14 pages, 2010.

[EHJT04] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Stat.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors.

[Faz02] Maryam Fazel. *Matrix rank minimization with applications.* PhD thesis, Stanford University, 2002.

[FCRP08] Maryam Fazel, Emmanuel J. Candès, Benjamin Recht, and Pablo A. Parrilo. Compressed sensing and robust recovery of low rank matrices. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1043–1047, 2008.

[FF93] Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[FMN19] Massimo Fornasier, Johannes Maly, and Valeriya Naumova. Robust recovery of low-rank matrices with non-orthogonal sparse decomposition from incomplete measurements. arXiv:1801.06240, 2019.

[For10] Massimo Fornasier. Numerical methods for sparse recovery. In Massimo Fornasier, editor, *Theoretical foundations and numerical methods for sparse recovery*, volume 9 of *Radon Series on Computational and Applied Mathematics*, pages 93–200. Walter de Gruyter, Berlin, 2010.

[FPRU10] Simon Foucart, Alain Pajor, Holger Rauhut, and Tino Ullrich. The Gelfand widths of ℓ_p-balls for $0 < p \leq 1$. *J. Complex.*, 26(6):629–640, 2010.

[FR08] Massimo Fornasier and Holger Rauhut. Iterative thresholding algorithms. *Appl. Comput. Harmon. Anal.*, 25(2):187–208, 2008.

[FR13] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing.* Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

[FR15] Massimo Fornasier and Holger Rauhut. Compressive sensing. In Otmar Scherzer, editor, *Handbook of mathematical methods in imaging. Vol. 1, 2, 3*, pages 205–256. Springer, New York, 2015.

[FVD19] Massimo Fornasier, Jan Vybíral, and Ingrid Daubechies. Robust and resource efficient identification of shallow neural networks by fewest samples. arXiv:1804.01592, 2019.

[FW10] Massimo Fornasier and Rachel Ward. Iterative thresholding meets free-discontinuity problems. *Found. Comput. Math.*, 10(5):527–567, 2010.

[GB97] Hong-Ye Gao and Andrew G. Bruce. WaveShrink with firm shrinkage. *Stat. Sin.*, 7(4):855–874, 1997.

[GBB11] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, 2011.

[GG84] Andrej Yu. Garnaev and Efim D. Gluskin. The widths of a Euclidean ball. *Dokl. Akad. Nauk SSSR*, 277(5):1048–1052, 1984.

[GJY11] Dongdong Ge, Xiaoye Jiang, and Yinyu Ye. A note on the complexity of L_p minimization. *Math. Program., Ser. B*, 129(2):285–299, 2011.

[GLF$^+$10] David Gross, Yi-Kai Liu, Steven T. Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Phys. Rev. Lett.*, 105(15):150401, 4 pages, 2010.

[Glu84] Efim D. Gluskin. Norms of random matrices and widths of finite-dimensional sets. *Math. USSR Sb.*, 48(1):173–182, 1984.

[GNOT92] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, 1992.

[GPYZ15] Yi Gao, Jigen Peng, Shigang Yue, and Yuan Zhao. On the null space property of ℓ_q-minimization for $0 < q \leq 1$ in compressed sensing. *J. Funct. Spaces*, Art. ID 579853, 10 pages, 2015.

[Gro11] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 57(3):1548–1566, 2011.

[GSB14] Tom Goldstein, Christoph Studer, and Richard Baraniuk. A field guide to forward-backward splitting with a FASTA implementation. arXiv:1411.3406, 2014.

[HJ13] Roger A. Horn and Charles R. Johnson. *Matrix analysis.* Cambridge University Press, Cambridge, second edition, 2013.

[HMT11] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, 2011.

[JK17] Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Found. Trends in Mach. Learn.*, 10(3-4):142–336, 2017.

[JL84] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability (New Haven, Conn., 1982)*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, Providence, RI, 1984.

[JMD10] Prateek Jain, Raghu Meka, and Inderjit S. Dhillon. Guaranteed rank minimization via singular value projection. In John D. Lafferty, Chris K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 937–945. Curran Associates, Inc., 2010.

[JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *STOC'13—Proceedings of the 2013 ACM Symposium on Theory of Computing*, pages 665–674. ACM, New York, 2013.

[Jol02] Ian T. Jolliffe. *Principal component analysis.* Springer Series in Statistics. Springer-Verlag, New York, second edition, 2002.

[KF09] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Chris K. I. Williams, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1033–1041. Curran Associates, Inc., 2009.

[KMR14] Felix Krahmer, Shahar Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. *Commun. Pure Appl. Math.*, 67(11):1877–1904, 2014.

[LDSP08] Michael Lustig, David L. Donoho, Juan M. Santos, and John M. Pauly. Compressed sensing MRI. *IEEE Signal Process. Mag.*, 25(2):72–82, 2008.

[Lor04] Dirk A. Lorenz. *Wavelet shrinkage in signal & image processing: an investigation of relations and equivalences.* PhD thesis, Universität Bremen, 2004.

[LV09] Zhang Liu and Lieven Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM J. Matrix Anal. Appl.*, 31(3):1235–1256, 2009.

[LWB18] Kiryung Lee, Yihong Wu, and Yoram Bresler. Near-optimal compressed sensing of a class of sparse low-rank matrices via sparse power factorization. *IEEE Trans. Inform. Theory*, 64(3):1666–1698, 2018.

[Mac09] David Mackenzie. Compressed sensing makes every pixel count. *What's Happening in the Mathematical Sciences*, 7:114–127, 2009.

[Mal19] Johannes Maly. *Recovery algorithms for quantized compressed sensing*. PhD thesis, Technische Universität München, 2019.

[MHWG14] Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 73–81, 2014.

[Mil98] Vitali Milman. Surprising geometric phenomena in high-dimensional convexity theory. In *European Congress of Mathematics, Vol. II (Budapest, 1996)*, volume 169 of *Progress in Mathematics*, pages 73–91. Birkhäuser, Basel, 1998.

[Moo11] Robert J. Moore. Eric Schmidt's "5 exabytes" quote is a load of crap. *The Data Point*. https://blog.rjmetrics.com/2011/02/07/eric-schmidts-5-exabytes-quote-is-a-load-of-crap/, 2011. Accessed on September 29, 2019.

[Nat95] Balas K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.

[NP14] Valeriya Naumova and Steffen Peter. Minimization of multi-penalty functionals by alternating iterative thresholding and optimal parameter choices. *Inverse Probl.*, 30(12):125003, 34 pages, 2014.

[NT09] Deanna Needell and Joel A. Tropp. CoSaMP: iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.*, 26(3):301–321, 2009.

[ODBP15] Peter Ochs, Alexey Dosovitskiy, Thomas Brox, and Thomas Pock. On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision. *SIAM J. Imaging Sci.*, 8(1):331–372, 2015.

[OJF$^+$15] Samet Oymak, Amin Jalali, Maryam Fazel, Yonina C. Eldar, and Babak Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Trans. Inform. Theory*, 61(5):2886–2908, 2015.

[Pea01] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philos. Mag., Ser. 6*, 2(11):559–572, 1901.

[Pet16] Stefen Peter. *Algorithms for robust and fast sparse recovery: new approaches towards the noise folding problem and the big data challenge*. PhD thesis, Technische Universität München, 2016.

[PHH09] Justin P. Haldar and Diego Hernando. Rank-constrained solutions to linear matrix equations using powerfactorization. *IEEE Signal Process. Lett.*, 16(7):584–587, 2009.

[Pis89] Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics.* Cambridge University Press, Cambridge, 1989.

[PV13] Yaniv Plan and Roman Vershynin. One-bit compressed sensing by linear programming. *Commun. Pure Appl. Math.*, 66(8):1275–1297, 2013.

[Rau10] Holger Rauhut. Compressive sensing and structured random matrices. In Massimo Fornasier, editor, *Theoretical foundations and numerical methods for sparse recovery*, volume 9 of *Radon Series on Computational and Applied Mathematics*, pages 1–92. Walter de Gruyter, Berlin, 2010.

[Rec11] Benjamin Recht. A simpler approach to matrix completion. *J. Mach. Learn. Res.*, 12:3413–3430, 2011.

[RFP10] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

[RGR18] David Reinsel, John Gantz, and John Rydning. The digitization of the world: from edge to core. *IDC White Paper*, Doc#US44413318, 28 pages, 2018.

[Roc70] R. Tyrrell Rockafellar. *Convex analysis.* Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[ROV14] Emile Richard, Guillaume R Obozinski, and Jean-Philippe Vert. Tight convex relaxations for sparse matrix factorization. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3284–3292. Curran Associates, Inc., 2014.

[SCOY08] Rayan Saab, Rick Chartrand, and Özgür Yilmaz. Stable sparse approximations via nonconvex optimization. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3885–3888, 2008.

[SY10] Rayan Saab and Özgür Yilmaz. Sparse recovery by non-convex optimization—instance optimality. *Appl. Comput. Harmon. Anal.*, 29(1):30–48, 2010.

[Tal05] Michel Talagrand. *The generic chaining: upper and lower bounds of stochastic processes.* Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005.

[TG07] Joel A. Tropp and Anna C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inform. Theory*, 53(12):4655–4666, 2007.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc., Ser. B*, 58(1):267–288, 1996.

[Tro04] Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

[Van13] Jeff Vance. Big data analytics overview. *Datamination.* https://www.datamation.com/applications/big-data-analytics-overview.html, 2013. Accessed on September 29, 2019.

[Ver12] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed sensing: theory and applications*, pages 210–268. Cambridge University Press, Cambridge, 2012.

[Ver15] Roman Vershynin. Estimation in high dimensions: a geometric perspective. In Götz E. Pfander, editor, *Sampling theory, a renaissance: compressive sensing and other developments*, pages 3–66. Birkhäuser, Cham, 2015.

[Ver18] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, with a foreword by Sara van de Geer.

[WGMM13] John Wright, Arvind Ganesh, Kerui Min, and Yi Ma. Compressive principal component pursuit. *Inf. Inference*, 2(1):32–68, 2013.

[XCXZ12] Zongben Xu, Xiangyu Chang, Fengmin Xu, and Hai Zhang. $L_{1/2}$ regularization: a thresholding representation theory and a fast solver. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(7):1013–1027, 2012.

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. Royal Stat. Soc., Ser. B*, 67(2):301–320, 2005.

[ZHT06] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *J. Comput. Graph. Stat.*, 15(2):265–286, 2006.