Technische Universität München

Fakultät für Mathematik

Understanding and Enhancing Data Recovery Algorithms

From Noise-Blind Sparse Recovery to Reweighted Methods for Low-Rank Matrix Optimization

Christian Erwin Kümmerle

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Prof. Dr. Michael Ulbrich
Prüfer der Dissertation:	1.	Prof. Dr. Felix Krahmer
	2.	Prof. Dr. Daniel Potts Technische Universität Chemnitz
	3.	Prof. Dr. Rayan Saab (schriftliche Beurteilung) University of California, San Diego

Die Dissertation wurde am 17.10.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 06.12.2019 angenommen.

Abstract

We prove new results about the robustness of noise-blind decoders for the problem of reconstructing a sparse vector from underdetermined linear measurements. Our results imply provable robustness of equality-constrained ℓ_1-minimization for random measurements with heavy-tailed distributions and, furthermore, correspond to a generalization of bounds on inscribed bodies of random polytopes.

We further propose a new algorithm for the reconstruction of low-rank matrices from few linear observations or from missing data, based on the iterative minimization of well-designed quadratic models of a non-convex objective. Our method, which is an instance of Iteratively Reweighted Least Squares (IRLS), is the first of its kind to combine data efficiency with computational scalability and fast local convergence rates. We show that the method attains a superlinear local convergence rate under near-optimal assumptions on the sample complexity for several random observation models. These theoretical statements are supported by computational experiments which suggest an improved data efficiency compared to the state-of-the-art. We provide an implementation of the proposed IRLS algorithm that solves computational issues of previous related methods.

Finally, we extend the framework to the completion of structured low-rank matrices such as low-rank Hankel or Toeplitz matrices. For this more flexible model, we propose an IRLS algorithm with quadratic local convergence rate under weak assumptions on the number and distribution of provided samples.

Zusammenfassung

Für das Problem der Rekonstruktion dünnbesetzter Lösungen unterbestimmter linearer Gleichungssysteme zeigen wir neue Resultate, die die Robustheit von Dekodern betreffen, welche keine Schätzung des Messfehlers benötigen. Sie implizieren eine beweisbare Robustheit von ℓ_1-Minimierung mit Gleichheitsnebenbedingung bei zufälligen Messungen mit schweren Verteilungsenden. Zudem generalisieren wir Schranken, die die Geometrie von zufälligen Polytopen beschreiben.

Die Arbeit entwickelt zudem einen neuen Algorithmus für die Identifikation von Niedrigrangmatrizen anhand von wenigen Beobachtungen. Der Algorithmus basiert auf der iterativen Minimierung von quadratischen Modellen nicht-konvexer Zielfunktionen und kann als „Iteratively Reweighted Least Squares“-Methode (IRLS) betrachtet werden. Er ist der erste seiner Art, der eine sehr gute Dateneffizienz mit numerischer Skalierbarkeit sowie schnellen lokalen Konvergenzraten kombiniert. Dabei beweisen wir, dass der Algorithmus eine superlineare lokale Konvergenzrate aufweist unter schwachen Annahmen an den Informationsgehalt von Messdaten, denen gewisse randomisierte Beobachtungsregeln zugrunde liegen. Numerische Experimente illustrieren die Dateneffizienz unseres Ansatzes, für welchen wir eine effiziente Implementierung beschreiben. Dies löst numerische Stabilitätsprobleme, die verwandte Methoden aufweisen.

Wir zeigen auf, wie sich diese Ideen auf die Vervollständigung von struktuierten Niedrigrangmatrizen, wie etwa Hankel- oder Toeplitzmatrizen niedrigen Ranges, übertragen lassen. Wir stellen einen IRLS-Algorithmus für dieses flexiblere Modell vor, der eine quadratische Konvergenzrate aufweist, und zwar beweisbar unter schwachen Annahmen an die Anzahl und die Verteilung der vorliegenden Einträge der zu rekonstruierenden Matrix.

Contents

Acknowledgements

First and foremost I would like thank Felix Krahmer, who advised me throughout the last four years and during this time, he has been a great role model for me. I appreciate it very much that he gave me a lot of freedom to find and pursue problems in an independent manner, and that he encouraged me to establish meaningful collaborations, which enabled me to develop two different lines of research through out my Ph.D. studies. I have been always impressed by the exceptional mathematical intuition and technical skills he shared with me whenever we worked together more closely. I surely learned a lot from him.

I would like to thank Massimo Fornasier for his strong support and care of me, which started already when I wrote my master thesis under his supervision. My interactions with him shaped and developed heavily my interest for the research I have been conducting, and I learned from him how to write research articles.

Throughout my Ph.D., I was fortunate to collaborate closely with a number of people, whose insights and suggestions I value very much, in particular Olivier Guédon, Tomáš Masák, Holger Rauhut, Shahar Mendelson, Juliane Sigl and Claudio Verdun.

My research was supported by the German-Israeli Foundation (G-1266-304.6/2015), the German Research Foundation (DFG) within the context of the Emmy Noether Junior Research Group “Randomized Sensing and Quantization of Signals and Images” (KR 4512/1-1), and the European Research Council (Starting Grant HDSPCONTR - 306274).

During the first year of my Ph.D. studies, I was fortunate to enjoy the support and hospitality of the Hausdorff Research Institute for Mathematics (HIM) in Bonn during the Trimester Program “Mathematics of Signal Processing”. The three months in 2016 provided me with a great starting point of my research, as I was exposed to a lot of recent trends in the community.

I would like to thank Jianwei Ma for his hospitality during my research visit at the Harbin Institute of Technology. I learned a lot about mathematical techniques in geophysics, and this experience motivated the research project on the recovery of structured low-rank matrices.

Furthermore, I would like to thank Stefan Kunis and Ivan Markovsky for their valuable feedback during the visits at their research groups. I am thankful to Rayan Saab and Daniel Potts for evaluating my dissertation in my examination committee, and to Michael Ulbrich for chairing the committee.

Neither would have been my work possible nor enjoyable without the many friends and companions I shared many moments with at the TUM in the Department of Mathematics. Beyond the people already mentioned, I would like to thank Giacomo Albi, Stefano Almi, Carlos Améndola, Mattia Bongini, Ricardo Acevedo Cabra, Rufat Badal, Stefan Bamberger, Sandro Belz, Joe-Mei Feng, Tim Fuchs, Sebastian Garreis, Marinus Gottschau, Olga Graf, Martina Gschwendtner, Markus Hansen, Margret Heinze, Sara Krause-Solberg, Johannes Maly, Oleh Melnyk, Marco Morandotti, Steffen Peter, Simon Plazotta, Michael Rauchensteiner, Michael Sandbichler, Kristof Schröder, Katharina Selig, Korbinian Singhammer, Nada Sissouno, Dominik Stöger, Philippe Sünnen, Mario Teixeira Parente and Laure Vuaille for many fun and interesting discussions during coffee and lunch breaks, and beyond.

Finally, my deepest gratitude goes to my family: my parents Ute and Erwin, to whom I owe my existence and a happy childhood; my brother Harald, who spurred my interest in mathematics; and my lovely wife Di, to whom I dedicate this thesis and with whom I share my life.

Introduction

At the end of this decade, the advent of *data-driven* inference and prediction methods that combine computational tools with the presence of unprecedented swaths of information easily available in digital formats has been one of the more important driving forces of innovations for many businesses. These methods are expected to provide have an important impact on societies and individuals in the future, as they have been seen to perform on par with human or better than humans on tasks that had been known to be notoriously hard: Image classification [KSH12], playing Go [Sil16] and language translation [BCB14] are just a few famous examples of tasks that can be solved by computer programs considerably well by today.

Many of these advances have been achieved by computer scientists using tools like *deep neural networks* [LBH15], which can be seen as classes of functions that concatenate linearities with non-linearities and that allow for efficient computations even in the case of millions of high-dimensional data points. In the last five years, it has been a very active field of research to improve empirical performances by finding the most adequate network architecture for a specific task, and to extend the applicability of the techniques to new problems. In this process, experimentation and simultations have played a crucial role.

One might wonder, what is the role of mathematics and mathematicians in all of this?

First, it is clear that the theories of linear algebra, calculus, and optimization provide the building blocks to even just write down what one of these *machine learning* systems are actually doing: Backprogation, for example, is an efficient algorithm to compute the gradient of a *loss function* defined on the network with respect to changes in the so-called *weights* which define the linearities of the network, but can be interpreted as a smart application of the conventional chain rule in calculus [GBC16]. The loss function is typically then optimized based on the available data using a variant of *stochastic gradient descent* [RM51, KY03], which is well studied in optimization.

Thus, one contribution a mathematician might be well-suited to make is to come up with a new building block such as an optimization algorithm or a computational "trick" to be used to in the more complex machine learning architecture.

Another, possibly even more relevant contribution of mathematics can be *clarity*. Among the thousands of possible architectures, which one should we use on which type of data? Even when real datasets are messy and complicated, proving certain optimality properties or error bounds for a network architecture given a certain, simplified type of data can be more convincing than the mere empirical success in experiments. This is due to the fact that experiments are always finite and limited, and it is not always easy for other researchers to reproduce the exact hyperparameters and algorithm specifications that have been used in published works [DST09, HIB+18].

This thesis does not discuss neural network architectures, but somewhat different problems: We study recovery problems involving data that is intrinsically low-dimensional, but embedded in a very high-dimensional space, and which is observed not directly, but through *linear measurements*. The low-dimensional structures we are interested in consist of *sparse* *vectors* and *low-rank matrices*. As we will see, this modelling is flexible enough to tackle various problems in medical imaging, computational physics, seismology, recommender systems and control theory.

In our investigations, the mindset outlined above will be, however, our guiding principle. On one the hand, we provide new theory shedding light on the performance of algorithms that had been previously proposed. On the other hand, we will develop new algorithms that improve on previous methods in important aspects, while providing theory validating the scope of these improvements. This dissertation consists of the three following chapters:

- In **Chapter 1**, we discuss the robustness to noise of classical algorithms designed to recover *sparse vectors* from underdetermined, linear measurements. This problem is also called *compressed sensing*, as by using well-designed measurements, the number of samples to identify the original sparse vector can be chosen to be surprisingly small. We show that reconstruction with small errors is possible using classical convex decoders under the presence of noise even if the noise level information is not used in the decoder, for a considerably larger class of measurements than it had been known before.

- In **Chapter 2**, we focus on the related, but different problem of recovery of a *low-rank matrix* from incomplete linear measurements. After discussing shortcomings of previous methods for this widely studied problem with applications in machine learning and computational physics, we propose a new algorithm based on iterative quadratic majorization of non-convex surrogates of the rank function. We embed the derivation of this algorithm, which we call `MatrixIRLS` as it is an instance of *Iteratively Reweighted Least Squares*, into a comprehensive review of that class of algorithms, explaining why our approach improves considerably previously proposed IRLS algorithms. We provide a local convergence theory that includes results about fast local convergence under quite weak assumptions, and discuss why our implementation allows for a scalability of `MatrixIRLS` to problems involving very high-dimensional data. We provide numerical experiments illustrating that our algorithm can beat the state-of-the-art in terms of data efficiency.

- In **Chapter 3**, we consider the more general problem of completing *structured* low-rank matrix from few entries. We motivate this modelling by illustrating how it naturally arises in signal processing and in time series analysis, for example in harmonic retrieval problems. We modify the IRLS framework of Chapter 2 to be applicable to this problem and propose the algorithm `StrucIRLS`. We show that the low intrinsic dimension of the model can be indeed utilized to reduce the number of optimization variables to the number of degrees of freedom, while using fast matrix-vector multiplication if the structured matrix is of Hankel or Toeplitz type. In terms of convergence analysis, we obtain similar results as in Chapter 2, including a local quadratic convergence rate of the algorithm under weak assumptions. Again, numerical experiments show an improved data efficiency compared to competing methods.

Chapters 1 to 3 all contain original work the author of this dissertation had been conducting during his Ph.D. studies, partially jointly with collaborators. In this context, I acknowledge in the respective chapters the contribution of my collaborators. The chapters are largely self-contained. However, for a full understanding of Chapter 3, the first sections of Chapter 2 can be helpful.

Chapter 1

Robustness of Noise-Blind Compressed Sensing

There are many engineerings problems in signal processing and medical imaging which are in fact *data acquisition* problems. In situations where the *acquisition cost* is high, it is of interest to find ways to reduce this cost while still obtaining the gist of the data up to a high accuracy. A considerable amount of research on efficient algorithms for this problem and on their understanding has been conducted, and the common framework to tackle this problem is often called *compressed sensing*.

Indifferently on what this framework is about precisely, one might want to ask: What happens if the data to be processed is corrupted by noise? Do the algorithms that are used in the processing break down, or are they are able to cope with this situation, maybe even provably so?

A positive answer to this question is desired for any algorithm that hopes to be used for realistic engineering problems. In many cases it is possible to use prior knowledge of the magnitude and type of the noise as an algorithmic parameter to ensure optimal reconstruction or optimal processing.

The goal of this first chapter of the thesis is to present and develop theory for the understanding of an *inherent noise-robustness* of efficient optimization approaches for recovering sparse vectors from incomplete linear measurements—which is how the *compressed sensing* or *sparse recovery* problem can be defined mathematically. We show that a large class of linear system matrices—usually called *measurement matrices* in the compressed sensing setting—allows for robust recovery by these efficient methods even if no or little prior knowledge about the noise is present, a setting that we call *noise-blind compressed sensing*.

We also shed light on the connection of our results to the geometry of certain *random polytopes* that are spanned by the columns of respective measurement matrices.

The work presented in this chapter is based on the preprint [KKR18] co-authored with Felix Krahmer and Holger Rauhut and a subsequent generalization, for which the author of this dissertation was the leading author. Parts of this work will also appear in [GKK+19], jointly co-authored with Olivier Guédon, Felix Krahmer, Shahar Mendelson and Holger Rauhut, which further generalizes and strengthens the obtained results, focussing on the geometrical property of random polytopes that underly the implications for compressed sensing presented here.

1.1 Introduction: Compressed Sensing

Consider the following mathematical problem: Let $\mathbf{x}_0 \in \mathbb{R}^N$ be a high-dimensional vector and $\mathbf{A} \in \mathbb{R}^{m \times N}$ be a matrix with fewer rows than columns, i.e., with $m < N$.

Is it possible to reconstruct $\mathbf{x}_0$ from its linear image $\mathbf{y} = \mathbf{A}\mathbf{x}_0 \in \mathbb{R}^m$ if $\mathbf{y}$ and $\mathbf{A}$ are given? Basic linear algebra answers this in the negative, as there is a $N - m$-dimensional affine space of vectors $\{\mathbf{x} \in \mathbb{R}^N : \mathbf{A}\mathbf{x} = \mathbf{y}\}$ that is compatible with the underdetermined linear system.

However, it has been observed since the 1960's that in certain situations, it is possible to *invert* the system

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

perfectly and solve it for $\mathbf{x}_0$, even in a computationally efficient way: Work by Logan [Log65] about frequency estimation and of geophysicists [TBM79] suggested for many *measurement matrices* $\mathbf{A}$, the vector $\mathbf{x}_0 \in \mathbb{R}^N$ can be reconstruted by solving the *ℓ_1-norm minimization problem*

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

under the additional assumption that $\mathbf{x}_0$ is *sparse* enough. In particular, this minimization problem can be solved by standard methods from convex optimization as it is even equivalent to a linear program.

Definition 1.1.1. *Let $s, N \in \mathbb{N}$ such that $s \leq N$.*
We say that the vector $\mathbf{x} \in \mathbb{R}^N$ is s-sparse *if $\mathbf{x} \in \Sigma_s$ and*

$$\Sigma_s := \{\mathbf{z} \in \mathbb{R}^N : \#\operatorname{supp}(\mathbf{z}) \leq s\} = \left\{\mathbf{z} \in \mathbb{R}^N : \mathbf{z}_i = 0 \textit{ for at least } N - s \textit{ indices } i \in [N]\right\}.$$

*We define the ℓ_0-*norm *(which is not a norm in the mathematical sense) of x as*

$$\|\mathbf{x}\|_0 = \#\operatorname{supp}(\mathbf{x}) = \#\left\{i \in [N] : \mathbf{x}_i \neq 0\right\}.$$

The *sparsity* assumption became more and more useful with the advent of wavelets [Dau88, Dau92] and other sparsifying transforms [SCD02], as they can be used to represent many natural images and signals by approximately sparse, high-dimensional vectors.

In the past 15 years, seminal papers [CT05, CRT06a, CRT06b, Don06] and subsequent progress have provided a quite thorough understanding of why and when ℓ_1-minimization can indeed find the sparsest vector compatible with underdetermined linear measurements, see [FR13] for an overview.

In particular, by using *randomization* in the measurement process, i.e., in the design of A, it has been shown that an almost linear scaling of linear measurements m with respect to the signal sparsity s is sufficient to guarantee efficient reconstruction of sparse vectors, or more precisely, if

$$m \sim s \log(\mathrm{e} N/s),$$

with high probability.

The implications of this theory to science and medicine have been already manifold, as it has, for example, led to shorter *magnetic resonance imaging (MRI)* scanning times in the newest generation of commercially used MRI scanners [LDSP08, FWHS16]. The idea behind this acceleration is the realization that an intelligently designed *sensing* process can be used to obtain *compressed* versions of high-resolution MRI images, reducing the amount of data traditionally needed to obtain images of similar quality. By reducing costs and new, previously infeasible MR imaging variants with new diagnostic possibilities, the resulting gains of efficiency have been a main talking point in the June 2017 briefing of the recent Gauss Prize winner David Donoho in front of the United States Congress with respect to the benefits of foundational mathematical research [Don18].

Another, quite different successful application of the compressed sensing framework has been seen in the context of exploration seismology [Her10], which aims to detect gas or oil fields from large seismic imaging data volumes. These data volumes are traditionally collected by deploying a large amount of sources and receivers in a uniform sampling pattern across a vast area, which can be a land or sea area. Compressed sensing-type sampling schemes and reconstruction techniques enable seismic explorations to be carried out more efficiently, also due to their reduced requirements on equipment and crews. Furthermore, these sampling schemes are also more compatible with external, geographic constraints or unpredictable weather situations [MLJ+17, BML+17].

1.1.1 Stability under Model Misfit

A key factor contributing to why compressed sensing is applicable in these fields is its ability to cope with *approximate models* and *errors in the measurement process*. In particular, in a slight generalization of the setting above, we seek to recover an approximately sparse vector $\mathbf{x}_0 \in \mathbb{R}^N$ from noisy, underdetermined measurements

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{w} \tag{1.1}$$

with $\mathbf{A} \in \mathbb{R}^{m\times N}$, $m < N$, and where $\mathbf{w} \in \mathbb{R}^n$ is an unknown noise vector. *Approximate sparsity* can be measured, for example, in terms of a *best s-term approximation error*

$$\sigma_s(\mathbf{x}_0)_1 := \inf\{\|\mathbf{x}_0 - \mathbf{z}\|_1 : \mathbf{z} \in \mathbb{R}^N, \|\mathbf{z}\|_0 \leq s\},$$

the ℓ_1-distance of $\mathbf{x}_0$ to the set of s-sparse vectors. This quantity is equal to zero for any s-sparse vector $\mathbf{x}$, i.e., if $\|\mathbf{x}_0\|_0 \leq s$.

Under the absence of noise, corresponding to $\mathbf{w} = 0$, we can hope to recover $\mathbf{x}_0$ by solving the equality-constrained ℓ_1-minimization program

$$\Delta_1(\mathbf{y}) := \operatorname*{arg\,min}_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{z}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}. \tag{1.2}$$

In fact, for simplicity, if we assume that $\mathbf{A}$ is an $m \times N$ matrix with independent standard Gaussian $\mathcal{N}(0, 1)$ random variables and

$$m \geq Cs\log(\mathrm{e}\, N/s) \tag{1.3}$$

for a small constant C, it can be shown that the minimizer[1] $\Delta_1(\mathbf{y})$ of (1.2) then coincides with $\mathbf{x}_0$ if $\|\mathbf{x}_0\|_0 \leq s$ and more generally [CRT06a, CT06, FR13],

$$\|\mathbf{x}_0 - \Delta_1(\mathbf{y})\|_1 \leq C\sigma_1(\mathbf{x}_0)_s \quad \text{and} \quad \|\mathbf{x}_0 - \Delta_1(\mathbf{y})\|_2 \leq C\frac{\sigma_1(\mathbf{x}_0)_s}{\sqrt{s}}, \tag{1.4}$$

with high probability, where C is again a constant. This quantifies the approximation quality of (1.2) in presence of an imperfect sparsity model.

1.1.2 Robustness and Noise-Aware Sparse Recovery

In the case of noisy measurements (1.1) with $\mathbf{w} \neq 0$, the optimization problem can be altered to take prior knowledge on $\mathbf{w}$ into account. In particular, if a *bound* η on the *noise level* is known such that $\|\mathbf{w}\|_2 \leq \eta$, one commonly considers the *quadratically constrained ℓ_1-minimization* program

$$\Delta_{1,\eta}(\mathbf{y}) := \operatorname*{arg\,min}_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{z}\|_1 \quad \text{subject to} \quad \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta, \tag{1.5}$$

which is, as (1.2), a convex optimization problem. Then it can be shown that for a Gaussian $m \times N$ matrix $\mathbf{A}$ with $m \geq Cs\log(\mathrm{e}\, N/s)$ with high probability, the minimizer $\Delta_{1,\eta}(\mathbf{y})$ of (1.5)

[1] The uniqueness of the minimizer is not a trivial statement, but we omit the proof here.

satisfies

$$\|\mathbf{x}_0 - \Delta_{1,\eta}(\mathbf{y})\|_1 \leq C\sigma_1(\mathbf{x}_0)_s + D\sqrt{s}\eta \quad \text{and} \quad \|\mathbf{x}_0 - \Delta_{1,\eta}(\mathbf{y})\|_2 \leq C\frac{\sigma_1(\mathbf{x}_0)_s}{\sqrt{s}} + D\eta, \tag{1.6}$$

where C and D are absolute constants [CRT06b, FR13].

The bounds (1.6) are strong results, as they show that reconstruction of signals is possible with quantifiable bounds from severely imcomplete, as potentially $m \ll N$, measurements, while assuring *stability*, as vectors need to be only approximately sparse, and furthermore *robustness* to inaccurate observations such as $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}$ with $\mathbf{w} \neq 0$.

The assumption on $\mathbf{A}$ to be a Gaussian matrix with i.i.d. entries enabling guarantees of the form (1.6) can be relaxed to matrices with sub-Gaussian entries, for example by establishing a so-called *restricted isometry property* [BDDW08, Can08] for $\mathbf{A}$ with high probability, which is likewise possible for the almost linear scaling (1.3) of measurements m with respect to signal sparsity s.

In some applications, it is typically impossible to design measurement matrices with such a high degree of randomness, since structural constraints are prescribed. In magnetic resonance imaging, measurement matrices resemble randomly subsampled partial Fourier matrices [LDSP08], for which (1.6) has been shown [CT06, RV08, HR17] only for logarithmically smaller sparsity levels than in (1.3). In radar imaging, subsampled random circulant matrices play an important role [Rom09], the existing theory establishes results similar to the ones known for partial Fourier matrices [RRT12, KMR14].

In order to gauge the adequacy of existing setups and to enable the design of improved measurement designs, the extension of robust and stable recovery guarantees such as (1.6) to broad classes of matrices has been intensely studied by different authors.

More precisely, the above results could be extended to more structured matrices with independent rows [MPTJ08] or columns [Ver12, Section 5.6.1], if the distribution of the rows resp. columns obeys strong concentration properties, as it is the case for sub-Gaussian random vectors.

A further generalization of the validatity of the guarantees (1.6) has been shown by Mendelson and Lecué [ML17], see also [DLR18]. The authors generalize the results to random matrices with independent, possibly quite heavy-tailed entries whose distributions have only $\log(N)$ finite moments or, alternatively, to random matrices with independent columns, whose linear forms have $\log(N)$ finite moments, which additionally fulfill a *small-ball assumption*. The proofs depart from the classical method of establishing a restriced isometry property for the corresponding matrices, as it can be shown that it simply does not hold for heavy-tailed distributions in the regime of optimal scaling of measurements m vs. sparsity s [ALPTJ11]. Similar results for a smaller class of distributions with exponential tail behavior have already been shown in [Fou14, Kol11].

We note that all the results described in this section address an algorithmic setting that is *noise-aware*: The decoder (1.5) based on quadratically constrained ℓ_1-minimization requires a correct estimate η of the noise level such that $\|\mathbf{w}.\|_2 \leq \eta$ to establish (1.6).

This noise-awareness can be problematic in two cases:

- **Underestimated noise level:** The above results do not contain any statement on what to expect if the noise level is *underestimated* such that $\eta < \|\mathbf{w}\|_2$.
- **Overestimated noise level:** For a choice of η that corresponds to an *overestimation* of the noise level, i.e., if $\|\mathbf{w}\|_2 \ll \eta$, the bounds (1.6) may be very pessimistic as they depend on η rather on the true noise level $\|\mathbf{w}\|_2$.

Remark 1.1.1. *Ideas very similar to the principles of compressed sensing can be used for high-dimensional* variable selection *in a statistical context. In particular, consider the regression problem where an outcome variable y_i is observed m times such that for each $i \in [m]$, a set of N predictor variables $(a_{i1}, \ldots, a_{iN})$ is associated. If it is known that* only $s \ll N$ variables *play an important role in the prediction, it has been proposed to use the* least absolute shrinkage and selection operator (LASSO) [Tib96, TWH15], *which can be formulated in its Lagrangian form as the solution of*

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^N} \|\mathbf{A}\beta - \mathbf{y}\|_2^2 + \lambda \|\beta\|_1. \tag{1.7}$$

In (1.7), $\mathbf{A} = (a_{ij})_{i \in [m], j \in [N]}$ *is the predictor variable matrix and λ is a* tuning parameter, *and the selected variables correspond to the non-zero coordinates of the parameter vector $\widehat{\beta} \in \mathbb{R}^N$.*

The tuning parameter λ plays a similar role as η in (1.5) *and is to be chosen based on the expected observation error. If little knowledge about the observation error is available, it is not clear how to optimally chose λ* [Wai19, Section 7.3.2], *and it is popular to rely on data-dependent strategies such as cross validation* [TWH15], *which are not well-understood.*

1.2 Noise-Blind Sparse Recovery

In many realistic application scenarios of compressed sensing, it might be difficult to obtain accurate prior knowledge on the noise vector w in (1.1), which makes the choice of η in (1.5) difficult. Indeed, multiple, qualitatively different error sources can contribute to the noise w:

For example, in magnetic resonance imaging applications, modelling the sensing process as a *linear sensing process* such as (1.1) is already a simplification, as the underlying Fourier transform only approximates more complicated physical effects. The discretization from a continuous linear model to a discrete one is similarly hard to quantify. These error sources should not be neglected, as they need to be considered in addition to the conceptually simpler physical noise on the measurements, for which a simple statistical model as a Gaussian one might be appropriate.

This raises the question of whether there are decoders that can be used in *noise-blind* situations.

1.2.1 Noise-Blind Guarantees for ℓ_1-Minimization

This question has been already studied by P. Wojtaszczyk [Woj10, DPW09], who observed that conventional *equality-constrained ℓ_1-minimization* (1.2), which is equivalent to setting $\eta = 0$ in (1.5), exhibits an inherent noise-robustness in certain cases.

This noise-robustness of (1.2) is not implied by, for example, a restricted isometry property of the measurement matrix $\mathbf{A}$ and neither by weaker properties such as robust null space properties [FR13, Chapter 4] or restricted eigenvalue properties [Wai19], despite the fact that the latter are almost equivalent to the success of ℓ_1-minimization in the noiseless case.

However, by using so-called *(ℓ_1-)quotient properties*, Wojtaszczyk provided a robustness analysis for (1.2) in the case of Gaussian [Woj10] measurement matrices $\mathbf{A}$, which has been later adapted to sub-Gaussian matrices [DPW09]. These quotient properties are related to geometric properties of random polytopes associated to the random matrix $\mathbf{A}$, which will be pointed out in Section 1.3.

The precise forms of the resulting error bounds depend on the distribution, and can be expressed such that

$$\|\mathbf{x}_0 - \Delta_1(\mathbf{y})\|_1 \leq C\sigma_1(\mathbf{x}_0)_s + D\sqrt{s}|\!|\!|\mathbf{w}|\!|\!| \quad \text{and} \quad \|\mathbf{x}_0 - \Delta_1(\mathbf{y})\|_2 \leq C\frac{\sigma_1(\mathbf{x}_0)_s}{\sqrt{s}} + D|\!|\!|\mathbf{w}|\!|\!|, \tag{1.8}$$

where the norm $\|\!|\cdot|\!\|$ is such that

- $\|\!|\cdot|\!\| = \|\cdot\|_2$ is the Euclidean norm if A is *Gaussian*, and
- $\|\!|\cdot|\!\| = \max\{\|\cdot\|_2, \sqrt{\log(\mathrm{e}\,N/m)}\|\cdot\|_\infty\}$ is an interpolation norm between Euclidean norm $\|\cdot\|_2$ and supremum norm $\|\cdot\|_\infty$, if $\mathbf{A}$ is a matrix with *i.i.d. sub-Gaussian entries*.

In [Fou14], it was shown that matrices with *i.i.d. Weibull entries* exhibit the same behavior as Gaussian matrices. By considering, for example, Bernoulli matrices $\mathbf{A}$ with symmetric i.i.d. $\{\pm 1\}$ entries, it can be seen that the weaker bound quantified by $\max\{\|\cdot\|_2, \sqrt{\log(\mathrm{e}\,N/m)}\|\cdot\|_\infty\}$ is indeed necessary for some sub-Gaussian matrices [FR13, Section 11.3], and the bounds (1.8) cannot be expected to hold for $\|\!|\cdot|\!\| = \|\cdot\|_2$ in these cases.

Our Contribution

The presented noise-blind error bounds cover only specific cases of measurement matrices A, which are much more restricted than the state-of-the-art of noise-aware results as presented in Section 1.1.2. In particular, bounds as in (1.8) are only available for very few matrices with heavy-tailed distribution or structured random matrices [FR13, Chapter 11].

In this chapter, we extend the noise-blind bounds (1.8) to measurement matrices without concentration properties, but fulfilling only weak moment assumptions. This matches even the weakest possible conditions for sample complexity-optimal recovery in the noiseless or noise-aware case [ML17, DLR18]. Furthermore, we weaken the assumptions on the distribution of A for the robustness regimes both with respect to the Euclidean norm $\|\cdot\|_2$ and with respect to the $\max\{\|\cdot\|_2, \sqrt{\log(\mathrm{e}\,N/m)}\|\cdot\|_\infty\}$. Our results provide a better understanding on the transition between the two regimes, and remarkably, our findings suggest that a *super-Gaussian* property of distributions is instrumental for achieving the tighter robustness bound (1.8) with respect to $\|\!|\cdot|\!\| = \|\cdot\|_2$. We present the results in Section 1.4. Our proofs differ from previous proofs of quotient properties, which crucially relied on concentration properties of the random measurement matrix. More precisely, our proofs rely on lower bounds on certain empirical processes, and can be seen as applications of Mendelson's small ball method [Men15, KM15].

In [BA18], quotient properties were used to analyze the robustness of quadratically constrained ℓ_1-minimization (1.5) for underestimated noise levels such that $\|\mathbf{w}\|_2 > \eta$. We formulate our results such that underestimated noise levels are also covered. The analysis of quadratically constrained ℓ_1-minimization for underestimated noise levels if $\mathbf{A}$ fulfills a quotient property relative to $\max\{\|\cdot\|_2, \sqrt{\log(\mathrm{e}\,N/m)}\|\cdot\|_\infty\}$ is new to the best of our knowledge.

1.2.2 Noise-Robustness of Other Decoders

We note that instead of convex decoders minimizing ℓ_1-norms, there are also many other efficient algorithmic approaches to reconstruct sparse vectors from noisy, incomplete measurements. For some of these algorithms, guarantees as (1.8) can be shown. In particular, for the greedy algorithms Compressive Sampling Matching Pursuit (CoSaMP) [NT09] and Orthogonal Matching Pursuit (OMP) [PRK93, Zha11], it can be shown that the iterates converge to a vector $\mathbf{x}^\sharp$ such that

$$\|\mathbf{x}_0 - \mathbf{x}^\sharp\|_1 \leq C\sigma_1(\mathbf{x}_0)_s + D\sqrt{s}\|\mathbf{w}\|_2 \quad \text{and} \quad \|\mathbf{x}_0 - \mathbf{x}^\sharp\|_2 \leq C\frac{\sigma_1(\mathbf{x}_0)_s}{\sqrt{s}} + D\|\mathbf{w}\|_2$$

if the matrix $\mathbf{A}$ fulfills a restricted isometry property of appropriate order [FR13, Section 6.4], and a similar statement holds true for iterative hard thresholding algorithms [BD09], [FR13, Section 6.3].

However, while these algorithms are *noise-blind* in the a sense that prior knowledge about the error $\mathbf{w}$ is not an algorithmic parameter, they need another type of prior knowledge about the target sparsity s. Furthermore, the strongest available guarantees for these types of iterative methods are all based on the restricted isometry property, which does not hold for heavy-tailed measurement matrices $\mathbf{A}$. An extension of these guarantees to non-RIP matrices has not been successful yet [DLR18, Section VI]. While some guarantees in the optimal parameter regime are available for distribution-dependent versions of iterative algorithms, it is conjectured that these alterations are necessary for heavy-tailed measurement matrices [FL17].

We further note that there is an ℓ_1-type decoder which comes with a certain, intrinsic noise-blindness: The tuning parameter λ of the *square-root LASSO* [BCW11, ABB19, TLT18]

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^N} \|\mathbf{A}\beta - \mathbf{y}\|_2 + \lambda\|\beta\|_1$$

can be chosen with prior knowledge about the noise. Still, even this decoder needs prior knowledge about the model, as λ depends [ABB19] on the sparsity level s.

1.3 Geometry of Random Polytopes

In this section, we explore a connection between the conditions on the measurement matrix $\mathbf{A}$ that implies robust recovery guarantees of noise-aware compressed sensing and the properties of a geometric object associated to $\mathbf{A}$.

Consider a matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$ with columns $\mathbf{A}_1, \ldots, \mathbf{A}_N \in \mathbb{R}^m$.

Definition 1.3.1. *For a finite subset $S \subset \mathbb{R}^m$, we denote the* convex hull $\operatorname{conv}(S)$ *of S by*

$$\operatorname{conv}(S) = \left\{ \sum_{i=1}^{|S|} \alpha_i S_i : S_1, \ldots, S_{|S|} \in S, \alpha_1, \ldots, \alpha_{|S|} \geq 0 \textit{ and } \sum_{i=1}^{|S|} \alpha_i = 1 \right\},$$

and the absolutely symmetric convex hull $\operatorname{absconv}(S)$ *of S by*

$$\operatorname{absconv}(S) = \left\{ \sum_{i=1}^{|S|} \alpha_i S_i : S_1, \ldots, S_{|S|} \in S \textit{ and } \sum_{i=1}^{|S|} |\alpha_i| \leq 1 \right\}.$$

Then we note that the image $\mathbf{A}\mathbf{B}_1^N$ of the ℓ_1-unit ball $\mathbf{B}_1^N = \{\mathbf{x} \in \mathbb{R}^N : \|\mathbf{x}\|_1 = 1\}$ with respect to the matrix $\mathbf{A}$ can be written as

$$\operatorname{conv}(\pm\mathbf{A}_1, ..., \pm\mathbf{A}_N) = \operatorname{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N) = \mathbf{A}\mathbf{B}_1^N, \tag{1.9}$$

or in other words, as a polytope spanned by the columns $\mathbf{A}_i$ and their negative counterparts $-\mathbf{A}_i$ [Zie95]. This polytope is *random* if $\mathbf{A}$ is a random matrix, as it is the case for measurement matrices typically considered in the context of compressed sensing.

1.3.1 The ℓ_1-Quotient Property and Inscribed Bodies of Random Polytopes

The first noise-blind robustness guarantees for ℓ_1-minimization outlined in Section 1.2.1 were proved by Wojtasczyk in [Woj10, DPW09], who established a connection to results about random polytopes $\operatorname{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N)$ known in asymptotic geometric analysis, see [AAGM15] for an overview of this field of study.

We formulate a few notable results that were the starting point of [Woj10, DPW09].

Theorem 1.1 ([Glu89, Kas83, Woj10]). *Let $\mathbf{A}_1, \ldots, \mathbf{A}_N$ be independent vectors whose entries are independent standard Gaussian random variables and $0 < \beta < 1$. Then there exist constants*

$\widetilde{C}(\beta)$,$D(\beta)$ and C such that if $N \geq \widetilde{C}(\beta)m$, then

$$\operatorname{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N) \supset \frac{\sqrt{\log(eN/m)}}{D(\beta)}\mathbf{B}_2^m \tag{1.10}$$

with probability at least $1 - 2\exp(-CN^{1-\beta}m^{\beta})$.

This statement corresponds to a lower bound on the *inradius* of $\operatorname{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N) = \mathbf{A}\mathbf{B}_1^N$, i.e., on the radius of the largest Euclidean ball that is contained in the random polytope.

A slightly weaker statement can be deduced for random polytopes generated by vectors whose entrywise distribution has a tail that is below the one of a Gaussian—so-called *sub-Gaussian* distributions; see for example [Ver18] for an extensive treatment.

Definition 1.3.2. *We say that a random variable ξ is L-*sub-Gaussian *if*

$$\mathbb{P}(|\xi| \geq \gamma) \leq 2\exp\left(-\frac{\gamma^2}{L^2}\right) \textit{ for all } \gamma \geq 0.$$

Theorem 1.2 ([LPRTJ05])**.** *Let $\mathbf{A}_1, \ldots, \mathbf{A}_N$ be independent vectors whose entries are independent, mean-zero, of variance 1 and L-sub-Gaussian, let $0 < \beta < 1$. Then there exist constants $\widetilde{C}(\beta, L)$, $D(\beta, L)$ and C such that if $N \geq c_0(\beta, L)n$, then*

$$\operatorname{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N) \supset \frac{1}{D(\beta, L)}\left(\sqrt{\log(eN/m)}B_2^m \cap B_\infty^m\right) \tag{1.11}$$

with probability at least $1 - 2\exp(-CN^{1-\beta}n^{\beta})$.

This result basically says that while a random polytope $\operatorname{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N)$ defined by $\mathbf{A}_i$ with sub-Gaussian entries might not always contain a large Euclidean ball, it generically contains a slightly different regular convex body, namely, the intersection of an Euclidean ball in the scaling of (1.10) and of an ℓ_∞-ball. Theorem 1.2 has been generalized to isotropic, log-concave vectors [DGT09], [BGVV14, Chapter 11].

In [Woj10, DPW09], the authors used the results on the geometric properties of Theorem 1.1 and Theorem 1.2 to show guarantees for stable and robust recovery for equality-constrained ℓ_1-minimization from noisy measurements. They restated (1.10) and (1.11) using the following definition of *quotient properties*.

Definition 1.3.3 ([Woj10, Fou14, BA18])**.** *A matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ is said to possess the ℓ_1-*quotient property *with constant c relative to a norm $\|\cdot\|$ on $\mathbb{R}^m$ if, for all $\mathbf{w} \in \mathbb{R}^m$, there exists $\mathbf{z} \in \mathbb{R}^N$ such that*

$$\mathbf{A}\mathbf{z} = \mathbf{w} \quad \textit{and} \quad \|\mathbf{z}\|_1 \leq c s_*^{1/2}\|\mathbf{w}\|, \tag{1.12}$$

with $s_* = m/\log(e\,N/m)$.

The precise relationship between the ℓ_1-quotient property and the geometric inclusions (1.10) and (1.11) can be specified in Proposition 1.3.1, whose proof we provide for completeness.

We note that the name *quotient property* is motivated by the fact that (1.12) can be reformulated using a quotient norm of the set $[\mathbf{w}] = \mathbf{z} + \ker\mathbf{A}$ of preimages of a vector $\mathbf{w} \in \mathbb{R}^m$ with $\mathbf{w} = \mathbf{A}\mathbf{z}$. More precisely, it holds that

$$\|[\mathbf{w}]\|_{\ell_1/\ker\mathbf{A}} \leq c s_*^{1/2}\|\mathbf{w}\| \text{ for all } \mathbf{w} \in \mathbb{R}^m$$

is equivalent to (1.12), if $\|[\mathbf{w}]\|_{\ell_1/\ker \mathbf{A}}$ is defined such that

$$\|[\mathbf{w}]\|_{\ell_1/\ker \mathbf{A}} = \inf\{\|\mathbf{z}\|_1 : \mathbf{A}\mathbf{z} = \mathbf{w}\}.$$

Proposition 1.3.1 ([Woj10],[BA18, Proposition II.15])**.** *The following holds for any matrix* $\mathbf{A} \in \mathbb{R}^{m\times N}$ *with columns* $\mathbf{A}_1, \dots, \mathbf{A}_N \in \mathbb{R}^m$.

1. $\frac{1}{\sqrt{m}}$ *satisfies the* ℓ_1*-quotient property relative to* $|||\cdot||| = \|\cdot\|_2$ *with constant* c^{-1} *if and only if*

$$\mathrm{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N) \supset c\sqrt{\log(eN/m)}\mathbf{B}_2^m.$$

2. $\frac{1}{\sqrt{m}}A$ *satisfies the* ℓ_1*-quotient property relative to* $|||\cdot||| = \max(\|\cdot\|_2, \sqrt{\log(\mathrm{e}\,N/m)}\|\cdot\|_\infty)$ *with constant* c^{-1} *if and only if*

$$\mathrm{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N) \supset c\left(\mathbf{B}_\infty^m \cap \sqrt{\log(eN/m)}\mathbf{B}_2^m\right).$$

Proof. Assume that $\frac{1}{\sqrt{m}}\mathbf{A}$ satisfies the ℓ_1-quotient property relative to $|||\cdot|||$ with constant c^{-1}. Then for any $\mathbf{w} \in c\sqrt{\log(eN/m)}\mathbf{B}_{|||\cdot|||} := \{\mathbf{w} \in \mathbb{R}^m : |||w||| \le c\sqrt{\log(eN/m)}\}$, there exists a vector $\mathbf{u} \in \mathbb{R}^N$ fulfilling $\frac{1}{\sqrt{m}}\mathbf{A}\mathbf{u} = \frac{1}{\sqrt{m}}\mathbf{w}$ and $\|\mathbf{u}\|_1 \le \frac{\sqrt{m}}{c\sqrt{\log(eN/m)}}|||\frac{\mathbf{w}}{\sqrt{m}}||| = \frac{|||\mathbf{w}|||}{c\sqrt{\log(eN/m)}} \le 1$, which implies that $\mathbf{w} \in \mathbf{A}\mathbf{B}_1^N = \mathrm{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N)$.

Conversely, assume that $\sqrt{\log(eN/m)}\mathbf{B}_{|||\cdot|||} \subset \mathrm{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N)$. Given $\mathbf{w} \in \mathbb{R}^m$, we define $\hat{\mathbf{w}} = \frac{c\sqrt{\log(eN/m)}\mathbf{w}}{|||\mathbf{w}|||}$. Then $|||\hat{\mathbf{w}}||| \le c\sqrt{\log(eN/m)}$ and therefore $\hat{\mathbf{w}} \in \sqrt{\log(eN/m)}\mathbf{B}_{|||\cdot|||}$. Due to the assumption, there exists $\hat{\mathbf{z}} \in \mathbb{R}^N$ such that $\mathbf{A}\hat{\mathbf{z}} = \hat{\mathbf{w}}$ and $\|\hat{\mathbf{z}}\|_1 \le 1$. Using the definition of $\hat{\mathbf{w}}$, we obtain that $\mathbf{z} = \frac{\sqrt{m}|||\mathbf{w}|||}{c\sqrt{\log(eN/m)}}$ fulfills $\frac{1}{\sqrt{m}}\mathbf{A}\mathbf{z} = \mathbf{w}$ and $\|\mathbf{z}\|_1 = \|\hat{\mathbf{z}}\|_1\frac{\sqrt{m}}{c\sqrt{\log(eN/m)}}|||\mathbf{w}||| \le \frac{\sqrt{m}}{c\sqrt{\log(eN/m)}}|||\mathbf{w}|||$, which shows that $\frac{1}{\sqrt{m}}\mathbf{A}$ fulfills the ℓ_1-quotient property with constant c^{-1} relative to $|||\cdot|||$.

The two statements then follow by specifying that $\mathbf{B}_{|||\cdot|||} = \mathbf{B}_2^m$ for $|||\cdot||| = \|\cdot\|_2$ and $\mathbf{B}_{|||\cdot|||} = \left(\sqrt{\log(eN/m)}^{-1}\mathbf{B}_\infty^m \cap \mathbf{B}_2^m\right)$ for $|||\cdot||| = \max(\|\cdot\|_2, \sqrt{\log(\mathrm{e}\,N/m)}\|\cdot\|_\infty)$. □

We note that results such as (1.10), (1.11) (or in a different language, involving the ℓ_1-quotient property) are statements of relevance beyond just the scope of understanding sparse recovery decoders. In particular, following the arguments of [LPRTJ05], these geometric properties have implications on bounds of other geometric quantities of the corresponding random polytopes such as their volume and their mean width. Lower bounds on the volume of random polytopes have also been used in the context of differential privacy [HT10].

Our Contribution

In this chapter, we establish statements about inclusions such as (1.10) and (1.11) for a significantly enlarged class of random matrices A. In particular, this class includes *heavy-tailed* entry-wise distributions which do not fulfill strong concentration properties. We refer to Section 1.4.1 for the details.

1.3.2 Polytope Geometry and Sparse Recovery Conditions

Polytope geometry can also be used to understand the success of ℓ_1-minimization for sparse recovery itself regardless of robustness questions. Indeed, in [Don06], the equivalence of the solution of the equality-constrained ℓ_1-minimization program

$$\min_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{z}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}, \tag{1.13}$$

with the solution of the ℓ_0-minimization problem with $\|\mathbf{z}\|_0 := \#\operatorname{supp}(\mathbf{z})$ such that

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

was characterized by the following geometric property of the polytope $\operatorname{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N) = \mathbf{A}B_1^N$.

Definition 1.3.4 ([Grü03, Don05])**.** *A centrally symmetric polytope P is called* (centrally) *s-*neighborly *if for all $k+1$ non-antipodal vertices $v_1, \dots, v_{k+1} \in \{\pm\mathbf{A}_1, \dots, \pm\mathbf{A}_N\}$, the convex hull* $\operatorname{conv}(v_1, \dots, v_{k+1})$ *is a k-face of P.*

A k-face is a set (on the boundary of P) that can be written as an intersection of P with a k-dimensional hyperplane. The neighborliness of $\operatorname{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N)$ gives rise to the most well-known geometrical approach to the understanding of the success of compressed sensing.

Proposition 1.3.2 ([Don05, Theorem 1])**.** *Every s-sparse vector $\mathbf{x}$ is the unique solution of (1.13) with $\mathbf{y} = \mathbf{A}\mathbf{x}$ if and only if* $\operatorname{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N)$ *is s-neighborly.*

By using Proposition 1.3.2, understanding the number of k-faces of $\mathbf{A}B_1^N$ can be used to calculate sharp phase transitions that correspond to the empirically observed behavior of ℓ_1-minimization for sparse recovery [DT09].

The result of Proposition 1.3.2 can be equivalently formulated in terms of the *null space property* of a matrix $\mathbf{A}$. This is a slightly weaker property than the robust sparse recovery conditions for noise-aware ℓ_1-minimization mentioned in Section 1.1.2. In this sense, the results of [Fou14, ML17, DLR18] also imply the s-neighborliness of $\operatorname{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N)$ with high probability for $m \sim s\log(\mathrm{e}N/s)$.

1.4 Main Results

In this section, we present the main results of this chapter.

In Section 1.4.1, we prove more general statement than Theorem 1.2, requiring only distributions with logarithmically many well-behaved moments (Theorem 1.3). Furthermore, we establish a generalized version of Theorem 1.1 as compared to the ones for Gaussian [Glu89] or Weibull [Fou14] distributions; namely, our result only requires (independent) matrix entries to be super-Gaussian (for the precise meaning of this concept, we refer to Definition 1.4.2), see Theorem 1.4.

Using these results, we provide in Section 1.4.2 new robustness guarantees for noise-blind ℓ_1-minimization for measurement matrices with quite general entrywise distributions in the regime of optimal sample complexity $m \approx Cs\log(\mathrm{e}N/s)$ in Theorem 1.5. Our result covers both the case of equality-constrained ℓ_1-minimization (cf. Remark 1.4.3) and the case of quadratically constrained ℓ_1-minimization with underestimated noise level, as studied in [BA18].

Notably, we provide a unified proof strategy for our results, which covers all previous results for matrices with independent entries, both on Gluskin-type inclusions and on the robustness of noise-blind ℓ_1-minimization. The proofs of our results can be found in Section 1.6.

In Section 1.5, our results are complemented by numerical experiments, confirming the robustness of noise-blind ℓ_1-minimization for certain heavy-tailed measurement scenarios and exploring the recovery properties for different types of noise.

1.4.1 Geometry of Random Polytopes from General Distributions

We obtain results that depend on fine properties of the distribution of the vertices $A_1, \dots, A_N$ of the random polytope absconv$(A_1, ..., A_N)$. The first notion we need is the one of *isotropic* random vectors, which is a generalization of random variables with unit L_2-norm.

Definition 1.4.1. *We say that an m-dimensional random vector* $\mathbf{X} = (x_1, \dots, x_m)$ *is*

1. isotropic *if* $\mathbb{E}[\mathbf{X}\mathbf{X}^T] = \mathbf{I}_m$.
2. centered *if* $\mathbb{E}[\mathbf{X}] = 0$.
3. unconditional *if for every* $(\varepsilon_i)_{i=1}^m \in \{-1, 1\}^m$, $(x_1, x_2, \dots, x_m)$ *has the same distribution as* $(\varepsilon_1 x_1, \varepsilon_2 x_2, \dots, \varepsilon_m x_m)$.

A helpful tool will be the following mild notion of stochastic domination.

Definition 1.4.2 (Stochastic domination)**.** *Let* $\mathbf{X}, \mathbf{Y}$ *be centered random vectors in* $\mathbb{R}^n$*. We say that* $\mathbf{X}$ (stochastically) dominates $\mathbf{Y}$ *with constants* λ_1 *and* λ_2 *if for every* $\mathbf{t} \in \mathbb{R}^n$ *and every* $\gamma > 0$,

$$\mathbb{P}\left(|\langle \mathbf{X}, \mathbf{t}\rangle| \geq \gamma\right) \geq \lambda_1 \, \mathbb{P}\left(|\langle \mathbf{Y}, \mathbf{t}\rangle|s \geq \lambda_2 \gamma\right).$$

Accordingly, we say that a centered random variable x (stochastically) dominates *a random variable* y *with constants* λ_1 *and* λ_2 *if for every* $\gamma > 0$,

$$\mathbb{P}\left(|x| \geq \gamma\right) \geq \lambda_1 \, \mathbb{P}\left(|y| \geq \lambda_2 \gamma\right).$$

Definition 1.4.3 (Weak moment growth condition)**.** *Let* x *be a centered random variable with unit variance. We say that* x fulfills the weak moment assumption of order k with constants κ and $\alpha \geq 1/2$ *if*

$$\|x\|_{L_p} \leq \kappa p^\alpha \quad \text{for all } 4 \leq p \leq k.$$

We say that a centered, isotropic random vector X fulfills the weak moment assumption of order k with constants κ and $\alpha \geq 1/2$ *if*

$$\|\langle \mathbf{X}, \mathbf{t}\rangle\|_{L_p} \leq \kappa p^\alpha \quad \text{for all } 4 \leq p \leq k$$

for all $\mathbf{t} \in S^{m-1} = \{\mathbf{y} \in \mathbb{R}^m : \|\mathbf{y}\|_2 = 1\}$.

Theorem 1.3 (ℓ_1-quotient property for matrices with general distribution)**.** *Let* $\mathbf{A} = \left(\mathbf{A}_1, \dots, \mathbf{A}_N\right)$ *be an* $(m \times N)$ *random matrix with independent, isotropic columns* $\mathbf{A}_i \sim X$ *for all* $i \in [N]$*, let* $0 < \beta \leq 1$*. Assume that one of the following two conditions is fulfilled:*

(a) $\mathbf{X}$ *is unconditional and there exist constants* $c_1 > 1$ *and* $0 < c_0 \leq 1$ *such that*

$$\|\langle \mathbf{X}, \mathbf{e}_i\rangle\|_{L_1} \geq c_0 \|\langle \mathbf{X}, \mathbf{e}_i\rangle\|_{L_2} = c_0$$

for all m standard basis vector $\mathbf{e}_i \in \mathbb{R}^m$*, and such that the first* $2q_0$ *moments* $\|\langle \mathbf{X}, \mathbf{t}\rangle\|_{L_q}$ *exist for* $q_0 = \frac{\beta}{4}\frac{\log(N/m)}{\log(c_1/c_0)}$ *and for all* $\mathbf{t} \in S^{m-1}$*, and fulfill*

$$\|\langle \mathbf{X}, \mathbf{t}\rangle\|_{L_{2q}} \leq c_1 \|\langle \mathbf{X}, \mathbf{t}\rangle\|_{L_q}$$

for all $q \leq q_0$ *and for all* $\mathbf{t} \in S^{m-1}$.

(b) *The coordinates of* $\mathbf{X} = (x_1, \ldots, x_m)$ *are independent and fulfill the weak moment assumption of order* $q_0 = \max\left\{4, \frac{\beta}{4}\frac{\log(N/m)}{\log(\frac{32}{9}\kappa^4 4^{4\alpha})}\right\}$ *with constants* κ *and* $\alpha \geq 1/2$.

There exist constants $\widetilde{C}, D > 0$ *such that in either case, if* $N \geq \widetilde{C}m$,

$$\operatorname{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N) \supset \frac{1}{D}\big(\mathbf{B}_\infty^m \cap \sqrt{\log(eN/m)}\mathbf{B}_2^m\big)$$

and $\frac{1}{\sqrt{m}}\mathbf{A}$ *fulfills the* ℓ_1*-quotient property with constant* D *relative to the* clipped ℓ_2-norm

$$\|\cdot\|^{(\sqrt{\log(\mathrm{e}N/m)})} := \max\{\|\cdot\|_2, \sqrt{\log(\mathrm{e}N/m)}\|\cdot\|_\infty\},$$

both with probability at least $1 - 2\exp(-2N^{1-\beta}m^\beta)$.

Our second result addresses vectors with i.i.d. entries having tails that are *heavier* than the one of a Gaussian random variable.

Theorem 1.4 (ℓ_1-quotient property for super-Gaussian matrices)**.** *Let* $\mathbf{A} = \big(\mathbf{A}_1, \ldots, \mathbf{A}_N\big)$ *be an* $(m \times N)$ *random matrix with independent, isotropic columns* $\mathbf{A}_i \sim X$ *for all* $i \in [N]$, *let* $0 < \beta \leq 1$. *Assume that one of the following two conditions is fulfilled:*

(a) $\mathbf{X}$ *dominates a standard Gaussian vector* $\mathbf{G}$ *with constants* $1/2$ *and* $\lambda_2 \geq 1$.

(b) *The coordinates of* $\mathbf{X} = (x_1, \ldots, x_m)$ *are independent and dominate the standard Gaussian random variable* g *with constants* 1 *and* $\lambda_2 \geq 1$.

There exist constants $\widetilde{C}, D > 0$ *such that in either case, if* $N \geq \widetilde{C}m$,

$$\operatorname{absconv}(\mathbf{A}_1, ..., \mathbf{A}_N) \supset \frac{1}{D}\sqrt{\log(eN/m)}\mathbf{B}_2^m$$

and $\frac{1}{\sqrt{m}}\mathbf{A}$ *fulfills the* ℓ_1*-quotient property with constant* D *relative to the Euclidean norm* $\|\cdot\|_2$, *both with probability at least* $1 - 2\exp(-2N^{1-\beta}m^\beta)$.

In some sense, it can be said that the conditions (a) and (b) are fulfilled by random vectors $\mathbf{X}$ that do not exhibit a *sub-Gaussian* behavior as the random variables from Definition 1.3.2, but, on the contrary, rather a *super-Gaussian* behavior.

Remark 1.4.1. 1. *In the statement of Theorem 1.3, the constants* $\widetilde{C}$ *and* D *depend on* β *and* c_0, c_1 *and* κ, α, *respectively. In particular,* D *can be chosen such that*

$$D = \begin{cases} \dfrac{8\sqrt{2}C\exp(\beta/2)}{c_0\sqrt{\beta\log(c_1/c_0)}} \sim \dfrac{\exp(\beta/2)}{c_0\sqrt{\beta\log(c_1/c_0)}}, & \textit{in the case of condition (a)}, \\ \dfrac{16C\exp(\beta/2)}{c\sqrt{\beta\log(\frac{32}{9}\kappa^4 4^{4\alpha})}} \sim \dfrac{\exp(\beta/2)}{\sqrt{\beta(\log(\kappa)+4\alpha)}}, & \textit{in the case of condition (b)}, \end{cases}$$

and $\widetilde{C}$ *such that* $\widetilde{C} = \left(\frac{72}{c_0}\right)^{4/\beta} \sim (c_0^{-4})^{\frac{1}{\beta}}$, *and* $\widetilde{C} = \left(\frac{18}{\sqrt{2}c}\right)^{4/\beta} \sim (c^{-4})^{\frac{1}{\beta}}$, *respectively.*

2. *The constants of Theorem 1.4 depend on* β *and* λ_1 *and can be chosen as*

$$D = \frac{4\sqrt{2}\,\mathrm{e}^{\beta/2}\,\lambda_2}{\sqrt{\beta}} \sim \frac{\exp(\beta/2)\lambda_2}{\sqrt{\beta}},$$

and $\widetilde{C} = \mathrm{e}\left(\frac{2\sqrt{\pi\beta}+16\sqrt{2\pi}\lambda_2}{4\beta}\right)^{\frac{4}{\beta}} \sim \left(\beta^{-1/2} + \lambda_2\beta^{-1}\right)^{\frac{4}{\beta}}$.

We note that in both cases, it was not our objective to find the best possible constants $\widetilde{C}$ and D. By considering the case of $cm < N < \widetilde{C}m$ with much smaller c than $\widetilde{C}$ separately and analyzing the smallest singular value of A, the range of validity of the theorem can be extended considerably, cf. also [FR13, Theorem 11.19]. *For lower bounding the least singular values under the present random models, results as in* [KM15] *are useful tools.*

We note the inclusion of the statement of Theorem 1.3

$$\mathbf{A}\mathbf{B}_1^N \supset \frac{1}{D}\left(\sqrt{\log(eN/m)}\mathbf{B}_2^m \cap \mathbf{B}_\infty^m\right) \tag{1.14}$$

can be expressed equivalently in terms of L_q-*centroid bodies*, as introduced by [LZ97] and studied by Paouris [Pa006].

Definition 1.4.4. *Let K be a convex body in $\mathbb{R}^m$ with volume one, let $q \geq 1$ and μ be the uniform measure of K. We call the set $Z_q(K)$, defined implicitly via its support function $h_{Z_q(K)}$ such that*

$$h_{Z_q(K)}(t) = \|\langle\cdot,\mathbf{t}\rangle\|_{L_q(K)} = \left(\int_K |\langle\mathbf{X},\mathbf{t}\rangle|^q d\mu(\mathbf{X})\right)^{1/q}$$

for all $\mathbf{t} \in S^{m-1}$, the L_q-centroid body of K.

As pointed out in [DGT09], it can be seen that (1.14) is equivalent to

$$\mathbf{A}\mathbf{B}_1^N \supset \frac{1}{D}c_1\mathbf{Z}_{\log(eN/m)}\left(\frac{1}{2}\mathbf{B}_\infty^m\right),$$

since the norms $\|\langle\cdot,\mathbf{t}\rangle\|_{L_{\log(eN/m)}(\frac{1}{2}\mathbf{B}_\infty^m)}$ and $\|\mathbf{t}\|^{(\sqrt{\log(\mathrm{e}N/m)})} := \max\{\|\mathbf{t}\|_2, \sqrt{\log(\mathrm{e}N/m)}\|\mathbf{t}\|_\infty\}$ are equivalent.

Remark 1.4.2. *After a version of the results presented in Theorem 1.3 and Theorem 1.4 was published on arXiv* [KKR18] *in June 2018, the authors of* [GLT19] *and* [Men19] *presented different proofs to generalize Theorem 1.3 further. In particular,* [GLT19] *weakens assumption (b) in Theorem 1.3 to random variables variables fulfilling a small ball assumption (and with unit L_2-norm).* [GLT19] *does not consider dependencies between entries of X, unlike case (a) of Theorem 1.3. A similar result was obtained in* [Men19], *but with a simpler proof.*

1.4.2 Implications for Robustness of Noise-Blind Compressed Sensing

As pointed out in Section 1.2.1, stable reconstruction guarantees can be shown for equality-constrained ℓ_1-minimization

$$\Delta_1(\mathbf{y}) := \underset{\mathbf{z}\in\mathbb{R}^N}{\arg\min} \|\mathbf{z}\|_1 \text{ subject to } \mathbf{A}\mathbf{z} = \mathbf{y}, \tag{1.15}$$

for a wide range of possibly heavy-tailed measurement matrices A, as it is the case for robust and stable guarantees for quadratically constrained ℓ_1-minimization

$$\Delta_{1,\eta}(\mathbf{y}) := \underset{\mathbf{z}\in\mathbb{R}^N}{\arg\min} \|\mathbf{z}\|_1 \text{ subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta \tag{1.16}$$

in a noise-aware setting. These results can be achieved by establishing the following property for $\mathbf{A}$ with high probability [ML17, DLR18].

Definition 1.4.5 (Robust null space property, [FR13, Section 4.3])**.** *A matrix* $\mathbf{A} \in \mathbb{R}^{m\times N}$ *is said to satisfy the* robust null space property *of order s (with respect to* $|||\cdot|||$*) with constants* $0 < \rho < 1$ *and* $\tau > 0$ *if*

$$\|\mathbf{v}_S\|_1 \le \rho\|\mathbf{v}_{\overline{S}}\|_1 + \tau|||\mathbf{A}\mathbf{v}|||$$

for all $\mathbf{v} \in \mathbb{R}^N$ *and all* $S \subset [N]$ *with* $|S| \le s$*.*

We use Theorem 1.3 and Theorem 1.4 together with the results of [ML17, DLR18] about robust null space properties as defined in Definition 1.4.5 to obtain robust and stable guarantees for (1.15) and (1.16) in noise-blind scenarios with high probability, if the measurement matrix $\mathbf{A}$ is drawn from a wide range of i.i.d. entrywise distributions. The precise formulation of the result is as follows.

Theorem 1.5. *Let* $\mathbf{B} = (b_{ji})$ *be an* $m \times N$ *random matrix with independent symmetric entries* $b_{ji} \sim b$ *for all* $j \in [m]$*,* $i \in [N]$ *and* $\mathbf{A} := \frac{1}{\sqrt{m}}\mathbf{B}$*. For* $s \in \mathbb{N}$*, let* $\sigma_s(\mathbf{x})_1 := \inf\{\|\mathbf{x}-\mathbf{z}\|_1 : \mathbf{z} \in \mathbb{R}^N, \|\mathbf{z}\|_0 \le s\}$ *be the* ℓ_1*-error of the best s-term approximation of* $\mathbf{x} \in \mathbb{R}^N$*. Assume* $\eta \ge 0$*.*

1. *Assume that that b satisfies the weak moment assumption from Definition 1.4.3 of order* $\max(4, \log(N))$ *with constants* κ *and* $\alpha \ge 1/2$*. Then there exist constants* $\widetilde{c}_1, \widetilde{c}_2, \widetilde{C}, C, D, E > 0$ *depending only on* κ *and* α *such that if*

$$N \ge \widetilde{C}m, \quad m \ge \log^{2\alpha-1}(N) \textit{ and } s \le \widetilde{c}_2 s_* := \widetilde{c}_2\frac{m}{\log(\mathrm{e}\,N/m)},$$

   with probability at least $1 - 3\exp(-\widetilde{c}_1 m)$*, the solution* $\Delta_{1,\eta}(\mathbf{A}\mathbf{x}+\mathbf{w})$ *of quadratically constrained* ℓ_1*-minimization (1.5) given the measurement matrix* $\mathbf{A}$ *and data vector* $\mathbf{y} = \mathbf{A}\mathbf{x}+\mathbf{w}$ *fulfills the* ℓ_p*-error estimates*

$$\begin{aligned}&\|\mathbf{x}-\Delta_{1,\eta}(\mathbf{A}\mathbf{x}+\mathbf{w})\|_p \\ \le&\frac{C}{s^{1-1/p}}\sigma_s(\mathbf{x})_1 + s_*^{1/p-1/2}\Big(D\eta + E\frac{\|\mathbf{w}\|^{(\sqrt{\log \mathrm{e}\,N/m})}}{\|\mathbf{w}\|_2}\max\{\|\mathbf{w}\|_2-\eta, 0\}\Big)\end{aligned} \tag{1.17}$$

   for $1 \le p \le 2$*, for all* $\mathbf{x} \in \mathbb{R}^N$ *and all* $\mathbf{w} \in \mathbb{R}^m$*, where* $\|\cdot\|^{(\sqrt{\log(\mathrm{e}\,N/m)})} := \max\{\|\cdot\|_2, \sqrt{\log(\mathrm{e}\,N/m)}\|\cdot\|_\infty\}$*.*

2. *If the random variable b is as in (a) with the additional assumption that b dominates a standard Gaussian random variable g with constants* 1 *and* $\lambda_2 \ge 1$*, there exist constants* $\widetilde{c}_1, \widetilde{c}_2, \widetilde{C}, C, D, E > 0$ *depending on* κ*,* α *and* λ_2 *such that if*

$$N \ge \widetilde{C}m, \quad m \ge \log^{2\alpha-1}(N) \textit{ and } s \le \widetilde{c}_2 s_* := \widetilde{c}_2\frac{m}{\log(\mathrm{e}\,N/m)},$$

   with probability at least $1 - 3\exp(-\widetilde{c}_1 m)$*, the solution* $\Delta_1(\mathbf{A}\mathbf{x}+\mathbf{w})$ *of quadratically constrained* ℓ_1*-minimization (1.5) fulfills the* ℓ_p*-error estimates*

$$\|\mathbf{x}-\Delta_{1,\eta}(\mathbf{A}\mathbf{x}+\mathbf{w})\|_p \le \frac{C}{s^{1-1/p}}\sigma_s(\mathbf{x})_1 + s_*^{1/p-1/2}\big(D\eta + E\max\{\|\mathbf{w}\|_2-\eta, 0\}\big) \tag{1.18}$$

   for $1 \le p \le 2$*, and for all* $\mathbf{x} \in \mathbb{R}^N$ *and* $\mathbf{w} \in \mathbb{R}^m$*.*

For the proof of Theorem 1.5, we refer the reader to Section 1.6.3.

Remark 1.4.3. *We point out that robust recovery guarantees of Theorem 1.5 can be specified for the noise-blind* equality-constrained *ℓ_1-minimization program (1.15),*

$$\Delta_{1,0}(\mathbf{y}) = \Delta_1(\mathbf{y}) = \operatorname*{arg\,min}_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{z}\|_1 \text{ subject to } \mathbf{A}\mathbf{z} = \mathbf{y},$$

such that for all $1 \leq p \leq 2$,

$$\|\mathbf{x} - \Delta_1(\mathbf{A}\mathbf{x} + \mathbf{w})\|_p \leq \frac{C}{s^{1-1/p}}\sigma_s(\mathbf{x})_1 + s_*^{1/p-1/2} E\|\mathbf{w}\|^{(\sqrt{\log \mathrm{e} N/m})}$$

and

$$\|\mathbf{x} - \Delta_1(\mathbf{A}\mathbf{x} + \mathbf{w})\|_p \leq \frac{C}{s^{1-1/p}}\sigma_s(\mathbf{x})_1 + s_*^{1/p-1/2} E\|\mathbf{w}\|_2$$

for all $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{w} \in \mathbb{R}^m$ in the first and second part of Theorem 1.5, respectively.

This means that the existing robust recovery guarantees for this decoder using matrices with i.i.d. sub-Gaussian [FR13, Theorem 11.10], *Gaussian* [FR13, Theorem 11.9] *and Weibull* [Fou14, Theorem 11] *random variables can be considered as special cases of the theorem.*

Next, we illustrate the generality of the assumptions of Theorem 1.4 and Theorem 1.5 by enumerating random models which are covered by our theory, but which mostly have not been covered by the robustness analyses of [LPRTJ05, Woj10, Fou14].

Example 1.4.1. *Let A be a real random matrix with i.i.d. entries $a_{ji} = \frac{1}{\sqrt{m}} \frac{b_{ji}}{\|b_{ji}\|_{L_2}}$, $j \in [n]$, $i \in [N]$.*

1. *Assume that the b_{ji} are distributed as b_γ, where b_γ is a $\psi_{\frac{1}{\gamma}}$-random variable with the same distribution as $\operatorname{sign}(g)|g|^{2\gamma}$, where g is a standard normal variable and $\gamma > 0$. Then, b_γ is of exponential type, i.e., has a probability density function of $p(x) = c_1 \mathrm{e}^{-\frac{|x|^{\frac{1}{\gamma}}}{c_2}}$. If $1/2 \leq \gamma \leq 1$, the assumptions of the second part of Theorem 1.5 apply, cf.* [DLR18, Example V.4]. *In particular, the $b_{ji}/\|b_{ji}\|_{L_2}$ dominate a standard Gaussian g with parameters 1 and $1 \leq \lambda_2 \leq 2$. We note that the special cases $\gamma = \frac{1}{2}$ and $\gamma = 1$ have been covered already by the existing theory, in the latter case by* [Fou14].

   For $\gamma > 1$, the theorem still applies as b_γ keeps dominating a standard Gaussian random variable with parameters 1 and $\lambda_2 \leq 2$, but this is only true assuming a more restrictive upper bound $\widetilde{c}_2(\gamma)\frac{m}{\log(\mathrm{e} N/m)}$ on the sparsity, where $\widetilde{c}_2(\gamma)$ depends on γ, cf. [DLR18, Example V.4].

2. *Let $d \in \mathbb{N}$. If the b_{ji} are distributed as Student-t variables b_d with d degrees of freedom, they dominate (after normalization) the standard Gaussian g with parameters 1 and some $1 \leq \lambda_2 \leq 2$ as well, so that Theorem 1.5.2 applies if enough "small" moments are provided, i.e., if additionally $d \geq \log(N)$.*

3. *If the b_{ji} are distributed as symmetric Weibull variables b_r with exponent $1 \leq r \leq 2$, recovery guarantees or equality-constrained ℓ_1-minimization that are robust relative to the ℓ_2-norm have been shown in the optimal regime of m already in* [Fou14]. *Since the normalized symmetric Weibull variables $b_{ji}/\|b_{ji}\|_{L_2}$ dominate the standard Gaussian g with parameters 1 and some $1 \leq \lambda_2 \leq 2$ also here, Theorem 1.5.2 also applies in the setting of* [Fou14].

A comparison of the two parts of Theorem 1.5 suggests the existence of qualitatively different robustness regimes of equality-constrained ℓ_1-minimization (1.15), depending on subtle properties of the measurement matrix A—a phenomenon that is not present for noise-aware

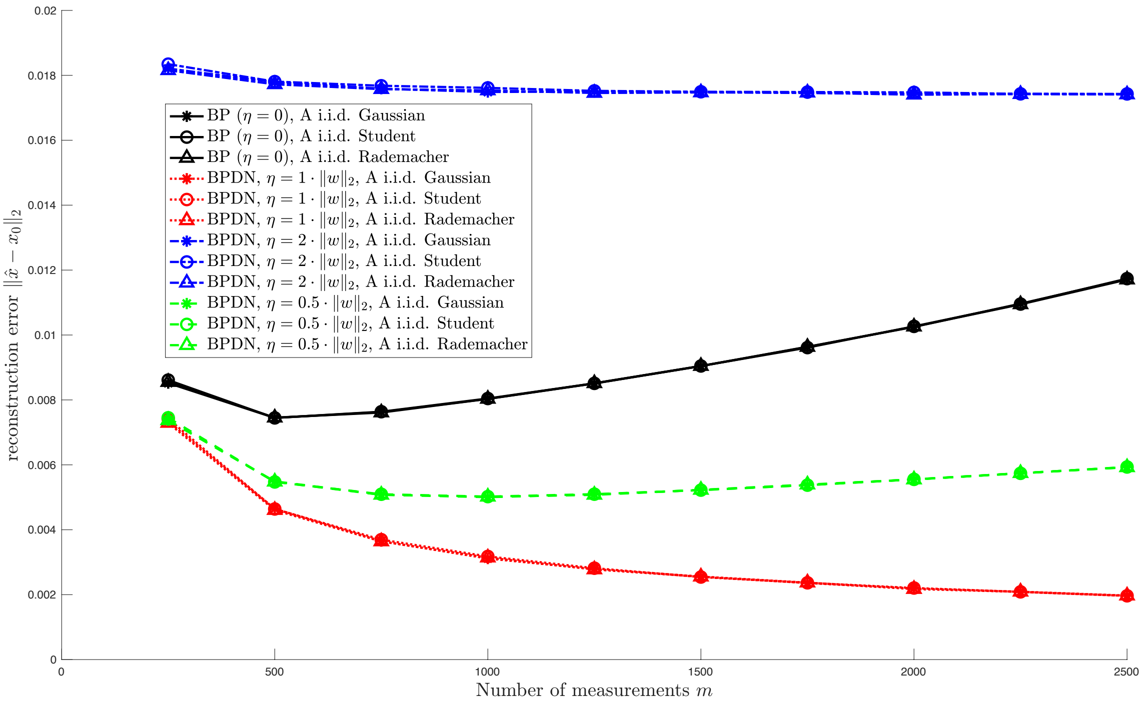

Figure 1.1 – y-axis: Reconstruction errors $\|\widehat{x} - x_0\|_2$ of the solutions $\widehat{x}$ of (1.15) and (1.16) from measurements $y = Ax_0 + w$, where w is a random spherical noise vector with $\|w\|_2 = 10^{-2}$, for different measurement matrices A. x-axis: number of rows m.

formulations of ℓ_1-minimization. In particular, the analysis suggests that measurement matrices with entries drawn from certain "super-Gaussian" distributions come with an *asymptotically more robust behavior* than those whose entries are drawn from certain sub-Gaussian, bounded distributions as the Rademacher distribution of random signs.

1.5 Numerical Experiments

In this section, we show in a case study that the results of Theorem 1.5 provide an appropriate explanation for the empirical robustness behavior of different measurement matrices. In particular, we consider three types of measurement matrices: random matrices with i.i.d. Gaussian, Bernoulli and Student-t entries. The presented numerical experiments have been conducted using MATLAB R2017b on a MacBook Pro with a 2.3 GHz Intel Core i5 processor. The convex optimization problems of our experiments are solved using the CVX package [GB14].

1.5.1 Behavior under Spherical Noise

In our first experiment, we perform simulations for the reconstruction of a s-sparse vector $x_0 \in \mathbb{R}^N$ with $\|x_0\|_2 = 1$ from measurements $y = Ax_0 + w$ which are perturbed by a random vector $w \in \mathbb{R}^m$ that is drawn from the uniform distribution on the sphere of radius $\|w\|_2 = 10^{-2}$. To obtain our reconstruction result $\widehat{x}$, we use equality-constrained ℓ_1-minimization (1.15) as defined by $\Delta_1(y)$, also called *basis pursuit (BP)* and quadratically constrained ℓ_1-minimization (1.16), also called *basis pursuit denoising (BPDN)*

$$\Delta_{1,\eta}(y) = \arg\min_{z \in \mathbb{R}^N} \|z\|_1 \text{ subject to } \|Az - y\|_2 \leq \eta,$$

where the noise level estimate η is chosen such that $\eta \in \{\|w\|_2, 2\|w\|_2, 0.5\|w\|_2\}$, i.e., the noise level $\|w\|_2$ is either estimated accurately or over- or underestimated by a factor of two. The support S of x_0 is drawn uniformly among the $\binom{N}{s}$ possibilities, and the non-zero coordinates are drawn uniformly on the sphere $\mathcal{S}^{S-1} = \{x \in \mathbb{R}^N : \|x\|_2 = 1, \operatorname{supp}(x) \subset S\}$.

In Figure 1.1, the resulting recovery ℓ_2-errors $\|\widehat{x} - x_0\|_2$ can be observed for the three different random models (in case of Student-t measurements, $k = 9$ degrees of freedoms were used) for the measurement matrix A mentioned above, where the parameters were chosen as $N = 5000$, $s = 10$ and $m \in \{\lceil kN/20 \rceil, k = 1, \dots, 14\}$. The reported errors are averaged over 500 runs of the simulation.

We notice that in the experiment, the recovery error of the equality-constrained formulation (1.15) is comparable to the one of quadratically constrained ℓ_1-minimization (1.16) with correctly estimated or underestimated noise level $\eta \in \{\|w\|_2, 0.5\|w\|_2\}$, if A has a small number of rows $m \leq 500$. For larger m, the robustness of (1.16) improves further if $\eta = \|w\|_2$, whereas it stagnates for underestimated noise of $\eta = 0.5\|w\|_2$ and it deteriorates slightly for (1.15).

It can be also observed that an overestimation of the noise level such that $\eta = 2\|w\|_2$ in (1.16) leads to a significantly worse reconstruction error $\|\widehat{x} - x_0\|_2$ than all the other methods, for all the considered number of measurements m.

Importantly, we observe that the robustness behavior of the algorithms does not depend on the choice of Gaussian, Bernoulli or Student-t measurement matrices in this case of presence of spherical noise.

This is precisely in accordance to the result of Theorem 1.5: Bernoulli variables b fulfill the assumptions of the first part of the theorem, but not of the second part, since they are sub-Gaussian. On the other hand, Gaussian and Student-t variables (with a sufficient number of degrees of freedom) fulfill the assumptions for the stronger statement of Theorem 1.5.2. In general, Bernoulli measurement matrices entail the weaker statement predicting a reconstruction error of

$$\|x - \Delta_1(y)\|_2 \leq D\|w\|^{(\sqrt{\log \mathrm{e} N/m})}$$

with a constant D for equality-constrained ℓ_1-minimization. For spherical noise, though, this coincides with the statement of Theorem 1.5.2, since $\|w\|^{(\sqrt{\log \mathrm{e} N/m})} = \|w\|_2$ with high probability under this noise model.

1.5.2 Behavior under Heavy-Tailed Noise

Next, instead of uniform spherical noise, we consider more heavy-tailed noise such that $w = 10^{-2} \frac{\widetilde{w}}{\|\widetilde{w}\|_2} \in \mathbb{R}^m$, where $(\widetilde{w})_i$ are i.i.d. ψ_α random variables for the parameter $\alpha = 0.2$, cf. also 1.4.1.1. Such a noise has most of its mass in just few coordinates, and the size of its largest entry $\|w\|_\infty$ is comparable to its ℓ_2-norm $\|w\|_2$, i.e. $\|w\|_\infty \approx \|w\|_2 = 10^{-2}$ with high probability.

In this case, the conclusions about the recovery accuracy of equality-constrained ℓ_1-minimization (1.15) that can be drawn from Theorem 1.5.1 and Theorem 1.5.2 predict a better behavior of Gaussian and Student-t measurements than for Bernoulli measurements, in particular for $m \ll N$: Since then

$$\|w\|^{(\sqrt{\log \mathrm{e} N/m})} = \sqrt{\log \mathrm{e} N/m} \cdot \|w\|_\infty \approx \sqrt{\log \mathrm{e} N/m} \cdot \|w\|_2 = \sqrt{\log \mathrm{e} N/m} \cdot 10^{-2}$$

with high probability, Theorem 1.5 predicts a reconstruction error of

$$\|x - \Delta_1(y)\|_2 \leq D \cdot \sqrt{\log \frac{\mathrm{e} N}{n}} \cdot 10^{-2}$$

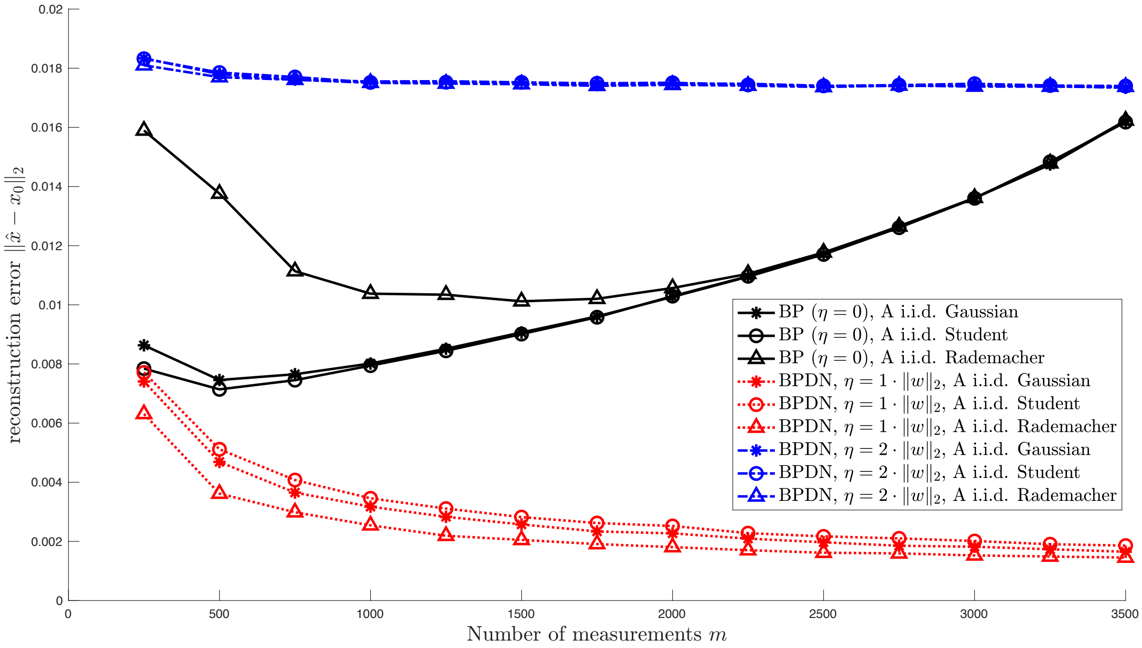

Figure 1.2 – y-axis: Reconstruction errors $\|\widehat{x}-x_0\|_2$ of the solutions $\widehat{x}$ of (1.15) and (1.16) from measurements $y = Ax_0 + w$, $w = 10^{-2}\cdot\frac{\widetilde{w}}{\|\widetilde{w}\|_2}$ where the entries of $\widetilde{w}$ are i.i.d. $\psi_{0.2}$ random variables, for different measurement matrices A. x-axis: number of rows of m.

for Bernoulli measurements, but a reconstruction error of

$$\|x-\Delta_1(y)\|_2 \leq D\cdot 10^{-2}$$

for the two other, more heavy-tailed measurement models (here, D is some constant).

These predictions can be well confirmed in the experiment illustrated in Figure 1.2, repeating the experiment from Section 1.5.1 for this different, heavy-tailed noise model: Unlike before, the reconstruction error of equality-constrained ℓ_1-minimization (1.15) for Bernoulli matrices is now consistently worse than for the Gaussian and Student-t measurement matrices if $m \ll N$, i.e., if $m = 250,\ldots,2000$. It is interesting to note that equality-constrained ℓ_1-minimization with Student-t matrices (with $k = \lceil\log(N)\rceil = 9$ degrees of freedom) is even slightly more robust than in the case of Gaussian matrices, especially if m is small.

On the other hand, the relative performance of Student-t measurements is worse than the one of Gaussian measurements if the noise-aware quadratically constrained ℓ_1-minimization (1.16) is used as a reconstruction algorithm.

As for spherical noise, we also note here that overestimating the noise level by a factor of two ($\eta = 2\cdot\|w\|_2$) in (1.16) leads to worse reconstructions than the noise-blind usage of (1.15).

We want to stress two conclusions from these experiments:

- The noise-blind reconstruction algorithm (1.15) is at least as robust in presence of certain heavy-tailed measurement matrices as in the case of Gaussian measurement matrices, especially if the measurement matrix has few rows m.
- While a very precise choice in the noise level estimate η of (1.16) leads to better reconstructions than using the noise-blind variant (1.15), the reconstructions deteriorate quickly once η is chosen as an overestimate of the actual noise level. In this sense, it

is preferred to choose an underestimated η or even $\eta = 0$ (resulting again in (1.15)) in situations where there is little a priori knowledge about the noise w.

1.6 Proofs

In this section, we provide proofs of Theorem 1.3, Theorem 1.4 and Theorem 1.5.

A starting point for our proofs is Lemma 1.6.1, which provides a necessary and sufficient condition for the ℓ_1-quotient property relative to a norm $\|\!| \cdot \|\!|$. We define the *dual norm* $\|\!| \cdot \|\!|_*$ of a norm $\|\!| \cdot \|\!|$ such that

$$\|\!| w \|\!|_* = \max\{|\langle y, w\rangle| : \|\!| y \|\!| \leq 1\}.$$

Lemma 1.6.1 ([FR13, Lemma 11.17]). *A matrix $A \in \mathbb{R}^{m\times N}$ has the ℓ_1-quotient property with constant d relative to $\|\!| \cdot \|\!|$ if and only if*

$$\|\!| w \|\!|_* \leq d s_*^{1/2} \|A^* w\|_\infty \quad \textit{for all } w \in \mathbb{R}^m,$$

where $s_* = m/\log(\mathrm{e}\, N/m)$.

The general proof strategy uses the framework of *Mendelson's small ball method* [Men15, KM15, DLR18]. The quantity $\|A^* w\|_\infty$ is lower bounded by a non-negative empirical process, which then will be lower bounded with high probability. This is formalized in the following lemma.

Lemma 1.6.2. *Let $A = \begin{pmatrix} A_1, & \dots, & A_N \end{pmatrix} \in \mathbb{R}^{m\times N}$ be a matrix with columns $A_i \in \mathbb{R}^m$, $i \in [N]$. If*

$$\inf_{w\in S^{\|\!|\cdot\|\!|_*}} \left(\frac{1}{N} \sum_{i=1}^N |\langle A_i, w\rangle|^{\log(N)} \right)^{\frac{1}{\log(N)}} \geq \frac{1}{D}\sqrt{\log\left(\frac{\mathrm{e}\, N}{m}\right)}, \tag{1.19}$$

where $S^{\|\!|\cdot\|\!|_*} = \{w \in \mathbb{R}^m : \|\!| w \|\!|_* = 1\}$ is the unit sphere of the dual norm $\|\!| \cdot \|\!|_*$ of some norm $\|\!| \cdot \|\!|$, then $\frac{1}{\sqrt{m}}A$ fulfills the ℓ_1-quotient property with constant D relative to the norm $\|\!| \cdot \|\!|$.

Proof. Let $w \in \mathbb{R}^m$. Then

$$\begin{aligned}\left\| \left(\frac{1}{\sqrt{m}}A\right)^* w \right\|_\infty &\geq \frac{1}{\mathrm{e}\sqrt{m}}\|A^* e\|_{\log N} = \frac{1}{\mathrm{e}\,\sqrt{m}}\left(\sum_{i=1}^N |\langle A_i, w\rangle|^{\log(N)}\right)^{\frac{1}{\log(N)}} \\ &= \frac{1}{\sqrt{m}}\left(\frac{1}{N}\sum_{i=1}^N |\langle A_i, w\rangle|^{\log(N)}\right)^{\frac{1}{\log(N)}},\end{aligned}$$

as $N^{1/q} = N^{1/\log(N)} = \mathrm{e}$. It follows from Lemma 1.6.1 that the ℓ_1-quotient property of $\frac{1}{\sqrt{m}}A$ with constant D relative to the norm $\|\cdot\|$ is implied by

$$\|A^* w\|_\infty \geq \frac{1}{D}\sqrt{\log(\mathrm{e}\, N/m)}\|\!| w \|\!|_* \quad \text{for all } w \in \mathbb{R}^m,$$

where $\|\cdot\|_*$ is the dual norm of the norm $\|\cdot\|$. This implies that

$$\left(\frac{1}{N}\sum_{i=1}^N |\langle A_i, w\rangle|^{\log(N)}\right)^{\frac{1}{\log(N)}} \geq \frac{1}{D}\sqrt{\log\left(\frac{\mathrm{e}\, N}{m}\right)}\|\!| w \|\!|_* \quad \text{for all } w \in \mathbb{R}^m$$

is a sufficient condition for the assertion of the lemma. □

The following result due to [KM15, Theorem 1.5] will be used to show the sufficient condition of Lemma 1.6.2.

Lemma 1.6.3 ([KM15, Theorem 1.5],[DLR18, Lemma III.1])**.** *Let $1 \leq q < \infty$, let $S \subset \mathbb{R}^m$ be a set and $A_1, \dots, A_N$ be i.i.d. copies of a random vector X in $\mathbb{R}^m$. For $u > 0$, define*

$$Q_S(u) := \inf_{w \in S} \mathbb{P}(|\langle X, w \rangle| \geq u)$$

and

$$\mathcal{R}_N(S) := \mathbb{E}\left[\sup_{w \in S} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \langle A_i, w \rangle \right| \right], \tag{1.20}$$

where $(\epsilon_i)_{i \in [N]}$ is a Rademacher sequence that is independent from $A_1, \dots, A_N$. Then, for $t > 0$, with probability at least $1 - 2e^{-2t^2}$,

$$\inf_{w \in S} \frac{1}{N} \sum_{i=1}^{N} |\langle A_i, w \rangle|^q \geq u^q \left(Q_S(2u) - \frac{4}{u} \mathcal{R}_N(S) - \frac{t}{\sqrt{N}} \right). \tag{1.21}$$

1.6.1 Proof of Theorem 1.3

For the proof of Theorem 1.3, we use some facts about L_q-centroid bodies [LZ97,Pa006], which can be related to the clipped ℓ_2-norm, which is defined for all $r > 0$ such that

$$\|t\|^{(\sqrt{r})} := \max\{\|t\|_2, \sqrt{r}\|t\|_\infty\}$$

for all $t \in \mathbb{R}^m$. We recall some definitions and facts about L_q*-centroid bodies*.

Definition 1.6.1. *Let K be a symmetric convex body in $\mathbb{R}^m$ with volume one.*

1. *The support function $h_K : \mathbb{R}^m \longrightarrow \mathbb{R}$ of K is defined as*

   $$h_K(x) := \max\{\langle x, y \rangle, y \in K\}.$$

   We note that K can be implicity defined by providing the support function h_K.

2. *Let $q \geq 1$ and μ the uniform measure of K. We call the symmetric convex body $Z_q(K)$ implicitly defined via its support function $h_{Z_q(K)}$ such that*

   $$h_{Z_q(K)}(t) = \|\langle \cdot, t \rangle\|_{L_q(K)} = \left(\int_K |\langle X, t \rangle|^q d\mu(X) \right)^{1/q}$$

   *for all $\theta \in \mathbb{R}^m$ the L_q-*centroid body of *K.*

Remark 1.6.1. *We note that if K is a symmetric convex body, K coincides with the unit ball with respect to the norm*

$$\|x\|_K := \inf\{\lambda \geq 0 : x \in \lambda K\}$$

defined by its Minkowski functional. In this case, the support function h_K is the dual norm *$\|\cdot\|_{K_*}$ of $\|\cdot\|_K$,*

$$h_K(x) = \|x\|_{K_*} = \|x\|_{K^\circ}$$

and furthermore, the polar body *K° of K defined as*

$$K^\circ := \{y \in \mathbb{R}^m : \langle x, y \rangle \leq 1, y \in K\}$$

is the unit ball with respect to that norm $\|\cdot\|_{K_*}$.

The proof of our main result will crucially use the following norm equivalence (see also [DGT09]).

Lemma 1.6.4 (Corollary of (4) of [BGMN05]). *There exist constants* $0 < c < C$ *such that for all* $r \geq 1$*,*

$$c\|\langle\cdot,t\rangle\|_{L_r(B_\infty^m)} \leq \sqrt{r}\|t\|_*^{(\sqrt{r})} \leq C\|\langle\cdot,t\rangle\|_{L_r(B_\infty^m)}$$

and

$$\frac{\sqrt{r}}{C}\|t\|_{Z_r(B_\infty^m)} \leq \|t\|^{(\sqrt{r})} \leq \frac{\sqrt{r}}{c}\|t\|_{Z_r(B_\infty^m)},$$

where $\|t\|^{(\sqrt{r})} := \max\{\|t\|_2, \sqrt{r}\|t\|_\infty\}$ *is the norm of* t *defined by the symmetric convex body*

$$K := \left(B_2^m \cap \frac{1}{\sqrt{r}}B_\infty^m\right)$$

and $\|\cdot\|_*^{(\sqrt{r})}$ *is such that*

$$\|t\|_*^{(\sqrt{r})} = \inf_{u\in\mathbb{R}^m}\left\{\|u\|_2 + \frac{1}{\sqrt{r}}\|t-u\|_1\right\}$$

for $t \in \mathbb{R}^m$ *is the dual norm of* $\|\cdot\|^{(\sqrt{r})}$*. In particular, this implies that*

$$\frac{1}{C}Z_r(B_\infty^m) \subset \sqrt{r}B_2^m \cap B_\infty^m \subset \frac{1}{c}Z_r(B_\infty^m)$$

for all $r \geq 1$.

Proof. It follows from (4) of [BGMN05] that there exist constants $0 < c < C$ such that

$$c\|\langle\cdot,t\rangle\|_{L_r(B_\infty^m)} \leq \sqrt{r}\|t\|_*^{(\sqrt{r})} \leq C\|\langle\cdot,t\rangle\|_{L_r(B_\infty^m)}.$$

For the fact that $\|t\|_*^{(\sqrt{r})} = \inf_{u\in\mathbb{R}^m}\left\{\|u\|_2 + \frac{1}{\sqrt{r}}\|t-u\|_1\right\}$ is the dual norm of $\|\cdot\|^{(\sqrt{r})}$ at t we refer the reader to the proof of [FR13, Theorem 11.9]. The other conclusion follows from properties of dual norms as $\|t\|^{(\sqrt{r})} = (\|t\|_*^{(\sqrt{r})})_*$. □

Another lemma about the dual norm $\|t\|_*^{(\sqrt{r})}$ that will be used is as follows.

Lemma 1.6.5 ([FR13, Lemma 11.22]). *Assume that* r *is an integer. Then the dual norm* $\|\cdot\|_*^{(\sqrt{r})}$ *of* $\|\cdot\|^{(\sqrt{r})}$ *is comparable with the norm* $\|\cdot\|_{r,\dagger}$ *defined by*

$$\|t\|_{r,\dagger} := \max\left\{\sum_{\ell=1}^{r}\|t_{B_\ell}\|_2 : B_1,\ldots,B_r \text{ form a partition of } [m]\right\} \tag{1.22}$$

in the sense that

$$\frac{1}{\sqrt{r}}\|t\|_{r,\dagger} \leq \|t\|_*^{(\sqrt{r})} \leq \frac{\sqrt{2}}{\sqrt{r}}\|t\|_{r,\dagger}$$

for all $t \in \mathbb{R}^m$.

We proceed with a lemma bounding the *Rademacher complexity* of L_r-unit balls.

Lemma 1.6.6. *Let $A_1, \dots, A_N$ be i.i.d. isotropic random vectors in $\mathbb{R}^m$ such that $A_i \sim X$ for all $i \in [N]$. Let $S^{L_r(B_\infty^m)} := \{y \in \mathbb{R}^m : \|\langle \cdot, y\rangle\|_{L_r(B_\infty^m)} = 1\}$ be the unit sphere of $L_r(B_\infty^m)$ for some $r \geq 2$.*

Then, for all $r \geq 2$, the Rademacher complexity parameter *of Lemma 1.6.3*

$$\mathcal{R}_N(S^{L_r(B_\infty^m)}) := \mathbb{E}\left[\sup_{w \in S^{\|\cdot\|_*}} \left|\frac{1}{N}\sum_{i=1}^N \epsilon_i \langle A_i, w\rangle\right|\right],$$

where $(\epsilon_i)_{i=1}^N$ is a Rademacher sequence independent of the $(A_i)_{i=1}^N$, fulfills

$$\mathcal{R}_N(S^{L_r(B_\infty^m)}) \leq \sqrt{\frac{m}{N}}.$$

Proof. Let $h := \frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i A_i$, where $(\epsilon_i)_{i=1}^N$ is a Rademacher sequence independent of $A_1, \dots, A_N$.

Then we obtain

$$\begin{aligned}\mathcal{R}_N(S^{L_r(B_\infty^m)}) &= \mathbb{E}\left[\sup_{w \in S^{L_r(B_\infty^m)}} \left|\left\langle w, \frac{1}{N}\sum_{i=1}^N \epsilon_i A_i\right\rangle\right|\right] = \frac{1}{\sqrt{N}}\mathbb{E}\Big[\sup_{\substack{w \in \mathbb{R}^m \\ \|\langle\cdot,w\rangle\|_{L_r(B_\infty^m)} \leq 1}} \langle w, h\rangle\Big] \\ &\leq \frac{1}{\sqrt{N}}\mathbb{E}\Big[\Big(\sup_{\substack{w \in \mathbb{R}^m \\ \|\langle\cdot,w\rangle\|_{L_r(\mu)} \leq 1}} \|w\|_2\Big)\|h\|_2\Big] \leq \frac{1}{\sqrt{N}}\mathbb{E}\left[\|h\|_2\right],\end{aligned} \tag{1.23}$$

by using Hölder's inequality and the isotropicity of μ, i.e.,

$$\|w\|_2 = \|\langle\cdot, w\rangle\|_{L_2(B_\infty^m)} \leq \|\langle\cdot, w\rangle\|_{L_r(B_\infty^m)} = 1$$

for all $w \in S^{L_r(B_\infty^m)}$ as $q \geq 2$. Furthermore, using Jensen's inequality, we estimate

$$\begin{aligned}\mathbb{E}\big[\|h\|_2\big] &= \frac{1}{\sqrt{N}}\mathbb{E}\Big[\|\sum_{i=1}^N \epsilon_i A_i\|_2\Big] \leq \frac{1}{\sqrt{N}}\mathbb{E}\Big[\|\sum_{i=1}^N \epsilon_i A_i\|_2^2\Big]^{1/2} \\ &= \frac{1}{\sqrt{N}}\Big(\mathbb{E}_A\left[\mathbb{E}_\epsilon\left[\sum_{i,k=1}^N\sum_{j=1}^m \epsilon_i\epsilon_k (A_i)_j (A_k)_j\right]\right]\Big)^{1/2} = \frac{1}{\sqrt{N}}\Big(\mathbb{E}_b\left[\sum_{i=1}^N\sum_{j=1}^m (A_i)_j (A_i)_j\right]\Big)^{1/2} \\ &= \frac{1}{\sqrt{N}}\Big(\sum_{i=1}^N \mathbb{E}\left[\|A_i\|_2^2\right]\Big)^{1/2} = \frac{1}{\sqrt{N}}(N \cdot m)^{1/2} = \sqrt{m},\end{aligned}$$

using again that $A_i \sim X$ is isotropic by assumption. Putting the two estimates together concludes the proof. □

As a preparation for the proof of Theorem 1.3, we provide a lemma which goes back to an argument of Rafał Latała, and which appeared in [DGT09, Section 2.1].

Lemma 1.6.7. *Let $r \geq 1$ and $X = (x_1, \dots, x_m)$ an m-dimensional random vector with unconditional, isotropic distribution, and let $0 < c_0 \leq 1$ be a constant such that*

$$\mathbb{E}[|\langle X, e_i\rangle|] = \|\langle X, e_i\rangle\|_{L_1} \geq c_0\|\langle X, e_i\rangle\|_{L_2} = c_0$$

for all $i \in [m]$, where e_i is the i-th standard basis vector of $\mathbb{R}^m$.

Then, for all $\theta \in S^{\|\cdot\|_2}$,

$$\|\langle X, \theta\rangle\|_{L_r} \geq c_0 \|\langle X, \theta\rangle\|_{L_r(B_\infty^m)}.$$

Proof. Let $\epsilon_1, \dots, \epsilon_m$ be independent Rademacher random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{Q})$, i.e., $\mathbb{Q}(\epsilon_i = 1) = \mathbb{Q}(\epsilon_i = -1) = 1/2$ for all $i \in [m]$. Since X is unconditional, we have that

$$\begin{aligned} \|\langle X, \theta\rangle\|_{L_r} &= \left(\int_{\mathbb{R}^m} \left|\sum_{i=1}^m x_i \theta_i\right|^r d\mu(X)\right)^{1/r} = \|\langle X, \theta\rangle\|_{L_r} = \left(\int_\Omega \int_{\mathbb{R}^m} \left|\sum_{i=1}^m \epsilon_i \theta_i |x_i|\right|^r d\mu(X) d\mathbb{Q}(\epsilon)\right)^{1/r} \\ &\geq \left(\int_\Omega \left|\sum_{i=1}^m \epsilon_i \theta_i \int_{\mathbb{R}^m} |\langle X, e_i\rangle| d\mu(X)\right|^r d\mathbb{Q}(\epsilon)\right)^{1/r} \geq c_0 \left(\int_\Omega \left|\sum_{i=1}^m \epsilon_i \theta_i\right|^r d\mathbb{Q}(\epsilon)\right)^{1/r}. \end{aligned} \tag{1.24}$$

By the contraction principle (cf. [FR13, Theorem 8.3]), it follows that

$$\begin{aligned} \left(\int_\Omega \left|\sum_{i=1}^m \epsilon_i \theta_i\right|^r d\mathbb{Q}(\epsilon)\right)^{1/r} &\geq \left(\int_{B_\infty^m} \int_\Omega \left|\sum_{i=1}^m \epsilon_i y_i \theta_i\right|^r d\mathbb{Q}(\epsilon) dy\right)^{1/r} = \left(\int_\Omega \int_{B_\infty^m} \left|\sum_{i=1}^m \epsilon_i y_i \theta_i\right|^r dy d\mathbb{Q}(\epsilon)\right)^{1/r} \\ &= \left(\int_{B_\infty^m} \left|\sum_{i=1}^m y_i \theta_i\right|^r dy\right)^{1/r} = \|\langle X, \theta\rangle\|_{L_r(B_\infty^m)}, \end{aligned}$$

as the uniform distribution in $B_\infty^m = \{y \in \mathbb{R}^m : \|y\|_\infty \leq 1\}$ is unconditional. Inserting this into (1.24) finishes the proof of the lemma. □

Lemma 1.6.8. *Assume that X is an m-dimensional random vector with unconditional, isotropic distribution, and let $0 < c_0 \leq 1$ be a constant such that*

$$\|\langle X, e_i\rangle\|_{L_1} \geq c_0 \|\langle X, e_i\rangle\|_{L_2}$$

for all m standard basis vectors $e_i \in \mathbb{R}^m$. Furthermore, assume that there is a constant $c_1 > 1$ such that the first $2q$ moments $\|\langle X, w\rangle\|_{L_r(\mu)}$ exist for all $w \in S^{m-1}$ and fulfill

$$\|\langle X, w\rangle\|_{L_{2r}} \leq c_1 \|\langle X, w\rangle\|_{L_r} \tag{1.25}$$

for all $w \in S^{m-1}$ and for all $r \leq q$. Then, it holds that:

1. *For all $2 \leq r \leq q$,*

$$\mathbb{P}\left(|\langle X, w\rangle| \geq \frac{c_0}{2} \|\langle \cdot, w\rangle\|_{L_r(B_\infty^m)}\right) \geq \frac{1}{4} \left(\frac{c_0^2}{c_1^2}\right)^r.$$

2. *Consider the the unit ball $S^{L_q(B_\infty^m)} = \{w \in \mathbb{R}^m : \|\langle \cdot, w\rangle\|_{L_q(B_\infty^m)} = 1\}$. If $q = \frac{\beta}{4} \frac{\log(N/m)}{\log(c_1/c_0)}$, then*

$$Q_{S^{L_q(B_\infty^m)}}(u) := \inf_{w \in S^{L_q(B_\infty^m)}} \mathbb{P}\left(|\langle X, w\rangle| \geq u\right) \geq \frac{1}{4} \left(\frac{m}{N}\right)^{\frac{\beta}{4}}$$

for $u = \frac{c_0}{2}$. If furthermore $N \geq K^{4/\beta} m$, then

$$Q_{S^{L_q(B_\infty^m)}}(u) \geq \frac{K}{4} \left(\frac{m}{N}\right)^{\frac{\beta}{2}}.$$

Proof. 1. The proof of the first statement follows ideas of [DGT09, Lemma 2.7.].

Let $r \geq 1$ and $w \in \mathbb{R}^m$. By the Paley-Zygmund inequality (see, e.g., [DlPG12, Chapter 3.3]) it follows that

$$\begin{aligned}\mathbb{P}\Big(|\langle X, w\rangle| \geq \frac{c_0}{2}\|\langle\cdot, w\rangle\|_{L_r(B_\infty^m)}\Big) &= \mathbb{P}\Big(|\langle X, w\rangle|^r \geq \frac{c_0^r}{2^r}\|\langle\cdot, w\rangle\|_{L_r(B_\infty^m)}^r\Big)\\ \geq \frac{\Big(\mathbb{E}\big[|\langle X, w\rangle|^r\big] - \frac{c_0^r}{2^r}\|\langle\cdot, w\rangle\|_{L_r(B_\infty^m)}^r\Big)^2}{\mathbb{E}\big[|\langle X, w\rangle|^{2r}\big]} &= \frac{\Big(\|\langle X, w\rangle\|_{L_r}^r - \frac{c_0^r}{2^r}\|\langle\cdot, w\rangle\|_{L_r(B_\infty^m)}^r\Big)^2}{\|\langle X, w\rangle\|_{L_{2r}}^{2r}}.\end{aligned} \tag{1.26}$$

From Lemma 1.6.7, it follows that

$$c_0^r\|\langle\cdot, w\rangle\|_{L_r(B_\infty^m)}^r \leq \|\langle\cdot, w\rangle\|_{L_r}^r,$$

and inserting this into (1.26) yields, together with assumption (1.25),

$$\mathbb{P}\Big(|\langle X, w\rangle| \geq \frac{c_0}{2}\|\langle\cdot, w\rangle\|_{L_r(B_\infty^m)}\Big) \geq c_0^{2r}\left(1 - \frac{1}{2^r}\right)^2 \frac{\|\langle X, w\rangle\|_{L_r}^{2r}}{\|\langle X, w\rangle\|_{L_{2r}}^{2r}} \geq \frac{1}{4}c_0^{2r}\frac{1}{c_1^{2r}} = \frac{1}{4}\left(\frac{c_0^2}{c_1^2}\right)^r.$$

2. Let $q = \frac{\beta}{4}\frac{\log(N/m)}{\log(c_1/c_0)}$. From the statement of 1. it follows that for all $w \in S^{L_q(B_\infty^m)}$,

$$\mathbb{P}\Big(|\langle X, w\rangle| \geq \frac{c_0}{2}\Big) \geq \frac{1}{4}\left(\frac{c_0^2}{c_1^2}\right)^q = \frac{1}{4}\left(\frac{c_0^2}{c_1^2}\right)^{\frac{\beta}{4}\frac{\log(N/m)}{\log(c_1^2/c_0^2)}} = \frac{1}{4}\left(\frac{m}{N}\right)^{\frac{\beta}{4}}.$$

Since $N \geq K^{4/\beta}m$ is equivalent to $\left(\frac{m}{N}\right)^{\beta/4} \leq K\left(\frac{m}{N}\right)^{\beta/2}$, the last statement follows. □

Proof of Theorem 1.3. Consider condition (a).

Using Lemma 1.6.3 for the choice of $S = S^{L_{q_0}(B_\infty^m)}$, $u = c_0/4$ and $t = N^{\frac{1-\beta}{2}}m^{\frac{\beta}{2}}$, we obtain with probability at least $1 - 2\exp(-2N^{1-\beta}m^\beta)$ that

$$\inf_{w \in S^{L_{q_0}(B_\infty^m)}} \frac{1}{N}\sum_{i=1}^N |\langle A_i, w\rangle|^{\log N} \geq \left(\frac{c_0}{4}\right)^{\log N}\left(Q_{S^{L_{q_0}(B_\infty^m)}}\left(\frac{c_0}{2}\right) - \frac{16}{c_0}\mathcal{R}_N(S^{L_{q_0}(B_\infty^m)}) - 1\left(\frac{m}{N}\right)^{\frac{\beta}{2}}\right) \tag{1.27}$$

Plugging in the estimates

$$Q_{S^{L_{q_0}(B_\infty^m)}}\left(\frac{c_0}{2}\right) \geq \frac{K}{4}\left(\frac{m}{N}\right)^{\frac{\beta}{2}}$$

for $N \geq K^{4/\beta}m$ from Lemma 1.6.8 and

$$\mathcal{R}_N(S^{L_{q_0}(B_\infty^m)}) \leq \sqrt{\frac{m}{N}} \leq \left(\frac{m}{N}\right)^{\frac{\beta}{2}},$$

from Lemma 1.6.6 and $0 < \beta \leq 1$, we obtain the bound

$$\inf_{w \in S^{L_{q_0}(B_\infty^m)}} \frac{1}{N}\sum_{i=1}^N |\langle A_i, w\rangle|^{\log N} \geq \left(\frac{c_0}{4}\right)^{\log N}\left(\frac{K}{4} - \frac{16}{c_0} - 1\right)\left(\frac{m}{N}\right)^{\frac{\beta}{2}} \geq \left(\frac{c_0}{4}\right)^{\log N}\left(\frac{K}{4} - \frac{17}{c_0}\right)\left(\frac{m}{N}\right)^{\frac{\beta}{2}}$$

for (1.27). Choosing $K = 72/c_0$ and therefore $K/4 - 17/c_0 = 1/c_0$, we see that this is equivalent to

$$\inf_{w\in S^{L_{q_0}(B_\infty^m)}}\left(\frac{1}{N}\sum_{i=1}^N|\langle A_i,w\rangle|^{\log N}\right)^{1/\log N}\geq\frac{c_0}{4}\left(\frac{1}{c_0}\right)^{\frac{1}{\log(N)}}\left(\frac{m}{N}\right)^{\frac{\beta}{2\log N}}\geq\frac{c_0}{4}\cdot 1\cdot \mathrm{e}^{-\beta/2}\geq\frac{c_0}{4\,\mathrm{e}^{\beta/2}}.$$

By Lemma 1.6.2, this implies that $\frac{1}{\sqrt{m}}A$ fulfills the ℓ_1-quotient property with constant $D=\frac{4\,\mathrm{e}^{\beta/2}}{c_0}\sqrt{\log(\mathrm{e}\,N/m)}$ relative to the norm $\|\cdot\|_{Z_{q_0}(B_\infty^m)}$, as $S^{L_{q_0}(B_\infty^m)}$ is the unit sphere of the dual norm of $\|\cdot\|_{Z_{q_0}(B_\infty^m)}$. Thus, by Definition 1.3.3, for all $w\in\mathbb{R}^m$, there exists a vector $z\in\mathbb{R}^N$ such that $\frac{1}{\sqrt{m}}Az=w$ and

$$\|z\|_1\leq\frac{4\exp(\beta/2)}{c_0}\sqrt{\log(\mathrm{e}\,N/m)}s_*^{1/2}\|w\|_{Z_{q_0}(B_\infty^m)}\leq\frac{4\exp(\beta/2)}{c_0}\sqrt{\log(\mathrm{e}\,N/m)}s_*^{1/2}\frac{C}{\sqrt{q_0}}\|w\|^{(\sqrt{q_0})},$$

with $s_*=\frac{m}{\log(\mathrm{e}\,N/m)}$ and C being the constant from Lemma 1.6.4. Using that

$$\frac{\sqrt{\log(\mathrm{e}\,N/m)}}{\sqrt{q_0}}\leq\sqrt{\frac{8}{\beta\log\left(\frac{c_1}{c_0}\right)}}$$

and the fact that $\sqrt{q_0}=\sqrt{\frac{\beta}{4}\frac{\log(N/m)}{\log(c_1/c_0)}}\leq\sqrt{\log(\mathrm{e}\,N/m)}$, if follows from this that $\frac{1}{\sqrt{m}}A$ fulfills the ℓ_1-quotient property with constant $D=\frac{8\sqrt{2}C\exp(\beta/2)}{c_0\sqrt{\beta\log(c_1/c_0)}}$ relative to the norm $\|\cdot\|^{(\sqrt{\log(\mathrm{e}\,N/m)})}=\max\{\|\cdot\|_2,\sqrt{\log(\mathrm{e}\,N/m)}\|\cdot\|_\infty\}$, which shows the statement of Theorem 1.3 under the set of conditions (a).

Assume now condition (b).

Let $w\in\mathbb{R}^m$ and $X=(x_1,\dots,x_m)$ be a random vector fulfilling the weak moment assumption of order $q_0=\dots$ with constants κ and $\alpha\geq 1/2$.

It is clear from the definition of $\|\cdot\|^{(\sqrt{r})}$ that $r_1\leq r_2$ implies $\|w\|^{(\sqrt{r_1})}\leq\|w\|^{(\sqrt{r_2})}$ for all $w\in\mathbb{R}^m$. It follows by duality that $r_1\leq r_2$ implies $\|w\|_*^{(\sqrt{r_1})}\geq\|w\|_*^{(\sqrt{r_2})}$. We note that due to Lemma 1.6.4 and the upper inequality of Lemma 1.6.5, there exists a constant $c>0$ such that

$$\|\langle\cdot,w\rangle\|_{L_{q_0}(B_\infty^m)}\leq\frac{\sqrt{q_0}}{c}\|w\|_*^{(\sqrt{q_0})}\leq\frac{\sqrt{2}}{c}\|w\|_{q_0,\dagger},$$

where $\|w\|_{q_0,\dagger}=\max\left\{\sum_{\ell=1}^{q_0}\|t_{B_\ell}\|_2, B_1,\dots,B_{q_0}\text{ form a partition of }[m]\right\}$. Therefore,

$$\mathbb{P}\left(|\langle X,w\rangle|\geq u\|\langle\cdot,w\rangle\|_{L_{q_0}(B_\infty^m)}\right)\geq\mathbb{P}\left(|\langle X,w\rangle|\geq\frac{\sqrt{2}u}{c}\|w\|_{q_0,\dagger}\right).\tag{1.28}$$

Let now $B_1,\dots,B_{\lfloor r\rfloor}$ be a partition of $[m]$ such that $\|w\|_{q_0,\dagger}=\sum_{\ell=1}^{q_0}\|w_{B_\ell}\|_2$, cf. (1.22). Then

$$\begin{aligned}
\mathbb{P}\Big(|\langle X,w\rangle|\geq\frac{\sqrt{2}u}{c}\|w\|_{q_0,\dagger}\Big)&=\mathbb{P}\Big(|\langle X,w\rangle|\geq\frac{\sqrt{2}u}{c}\sum_{\ell=1}^{q_0}\|w_{B_\ell}\|_2\Big)\\
&\geq\mathbb{P}\Big(\sum_{j\in B_\ell}x_jw_j\geq\frac{\sqrt{2}u}{c}\|w_{B_\ell}\|_2\text{ for all }\ell\in[q_0]\Big)\\
&=\prod_{\ell=1}^{q_0}\mathbb{P}\Big(\sum_{j\in B_\ell}x_jw_j\geq\frac{\sqrt{2}u}{c}\|w_{B_\ell}\|_2\Big),
\end{aligned}$$

where we used that the coordinates $x_1, \dots x_m$ of X are independent. Combining this with (1.28), we obtain

$$\mathbb{P}\left(|\langle X, w\rangle| \ge u\|\langle \cdot, w\rangle\|_{L_{q_0}(B_\infty^m)}\right) \ge \prod_{\ell=1}^{q_0} \mathbb{P}\Big(\sum_{j\in B_\ell} x_j w_j \ge \frac{\sqrt{2}u}{c}\|w_{B_\ell}\|_2\Big). \tag{1.29}$$

Using the symmetry of the distribution of the x_j and choosing $u = \frac{c}{2\sqrt{2}}$, it follows that

$$\begin{aligned}\mathbb{P}\Big(\sum_{j\in B_\ell} x_j w_j \ge \frac{\sqrt{2}u}{c}\|w_{B_\ell}\|_2\Big) &= \frac{1}{2}\mathbb{P}\Big(\Big|\sum_{j\in B_\ell} x_j w_j\Big| \ge \frac{\sqrt{2}u}{c}\|w_{B_\ell}\|_2\Big) = \frac{1}{2}\mathbb{P}\Big(\Big|\sum_{j\in B_\ell} x_j w_j\Big| \ge \frac{1}{2}\|w_{B_\ell}\|_2\Big)\\ &\ge \frac{1}{2}\frac{\left(1-(1/2)^2)\right)^2}{\|x\|_{L_4}^4} \ge \frac{9}{32\kappa^4 4^{4\alpha}},\end{aligned}$$

using the Paley-Zygmund inequality [DlPG12, Chapter 3.3], [FR13, Lemma 7.17] and the weak moment assumption of order q_0 on X in the last two inqualities. Combining this lower bound with (1.29), we obtain

$$\begin{aligned}\mathbb{P}\left(|\langle X, w\rangle| \ge \frac{c}{2\sqrt{2}}\|\langle \cdot, w\rangle\|_{L_{q_0}(B_\infty^m)}\right) &\ge \left(\frac{9}{32\kappa^4 4^{4\alpha}}\right)^{q_0} = \left(\frac{9}{32\kappa^4 4^{4\alpha}}\right)^{\frac{\beta\log(N/m)}{4\log(\frac{32}{9}\kappa^4 4^{4\alpha})}} = \left(\frac{m}{N}\right)^{\frac{\beta}{4}}\\ &\ge K\left(\frac{m}{N}\right)^{\frac{\beta}{2}}\end{aligned}$$

for $N \ge K^{4/\beta}m$. Using this, we can proceed as under condition (a) and use Lemma 1.6.6 and Lemma 1.6.3 with $S^{L_{q_0}(B_\infty^m)}$, $u = \frac{c}{4\sqrt{2}}$ and $t = N^{\frac{1-\beta}{2}}m^{\frac{\beta}{2}}$, we obtain with probability at least $1 - 2\exp(-2N^{1-\beta}m^\beta)$ that

$$\begin{aligned}\inf_{w\in S^{L_{q_0}(B_\infty^m)}} \frac{1}{N}\sum_{i=1}^N |\langle A_i, w\rangle|^{\log N} &\ge \left(\frac{c}{4\sqrt{2}}\right)^{\log N}\left(Q_{S^{L_{q_0}(B_\infty^m)}}\left(\frac{c}{2\sqrt{2}}\right) - \frac{16\sqrt{2}}{c}\mathcal{R}_N(S^{L_{q_0}(B_\infty^m)}) - \left(\frac{m}{N}\right)^{\frac{\beta}{2}}\right)\\ &\ge \left(\frac{c}{4\sqrt{2}}\right)^{\log N}\left(K - \frac{16\sqrt{2}}{c} - 1\right)\left(\frac{m}{N}\right)^{\frac{\beta}{2}}\\ &\ge \left(\frac{c}{4\sqrt{2}}\right)^{\log N}\left(K - \frac{17\sqrt{2}}{c}\right)\left(\frac{m}{N}\right)^{\frac{\beta}{2}}.\end{aligned}$$

Choosing $K = 18\sqrt{2}/c$, we see that this implies that if $N \ge (18\sqrt{2}/c)^{4/\beta}m$, then with probability at least $1 - 2\exp(-2N^{1-\beta}m^\beta)$,

$$\inf_{w\in S^{L_{q_0}(B_\infty^m)}}\left(\frac{1}{N}\sum_{i=1}^N |\langle A_i, w\rangle|^{\log N}\right)^{1/\log N} \ge \frac{c}{4\sqrt{2}}\left(\frac{1}{c}\right)^{\frac{1}{\log N}}\left(\frac{m}{N}\right)^{\frac{\beta}{2\log N}} \ge \frac{c}{4\sqrt{2}}\mathrm{e}^{-\beta/2} = \frac{c}{4\sqrt{2}\,\mathrm{e}^{\beta/2}}.$$

Arguing as above we conclude that with probability at least $1 - 2\exp(-2N^{1-\beta}m^\beta)$, $\frac{1}{\sqrt{m}}A$ fulfills the ℓ_1-quotient property with constant $D = \frac{16C\exp(\beta/2)}{c\sqrt{\beta\log(\frac{32}{9}\kappa^4 4^{4\alpha})}}$ relative to the norm $\|\cdot\|^{(\sqrt{\log(\mathrm{e}N/m)})} = \max\{\|\cdot\|_2, \sqrt{\log(\mathrm{e}N/m)}\|\cdot\|_\infty\}$, which concludes the proof. □

1.6.2 Proof of Theorem 1.4

To prove the statement of Theorem 1.4 about matrices with ℓ_1-quotient property relative to the Euclidean norm $\|\cdot\|_2$, we again use Lemma 1.6.3. However, instead of applying with respect to unit balls $S^{L_q(B_\infty^m)}$, we apply it with respect to the Euclidean unit ball $S^{\|\cdot\|_2} = \{y \in \mathbb{R}^m : \|y\|_2 = 1\}$. As a preparation, we show the following lemma.

Lemma 1.6.9. *Let $A_1, \dots, A_N$ be i.i.d. isotropic random vectors in $\mathbb{R}^m$ such that $A_i \sim X$ for all $i \in [N]$. Let $S^{\|\cdot\|_2} = \{y \in \mathbb{R}^m : \|y\|_2 = 1\}$ be the Euclidean unit sphere and $(\epsilon_i)_{i=1}^N$ be a Rademacher sequence independent of the $(A_i)_{i=1}^N$. Then the* Rademacher complexity parameter *of Lemma 1.6.3 fulfills*

$$\mathcal{R}_N(S^{\|\cdot\|_2}) := \mathbb{E}\left[\sup_{w \in S^{\|\cdot\|_2}} \left|\frac{1}{N}\sum_{i=1}^N \epsilon_i \langle A_i, w\rangle\right|\right] \leq \sqrt{\frac{m}{N}}.$$

Proof. Let $h := \frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i A_i$, where $(\epsilon_i)_{i=1}^m$ is a Rademacher sequence independent of $A_1, \dots, A_N$. Then we obtain

$$\mathcal{R}_N(S^{\|\cdot\|_2}) = \mathbb{E}\left[\sup_{w \in S^{\|\cdot\|_2}} \left|\left\langle w, \frac{1}{N}\sum_{i=1}^N \epsilon_i A_i\right\rangle\right|\right] = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \epsilon_i A_i\right\|_2\right] = \frac{1}{\sqrt{N}}\mathbb{E}[\|h\|_2],$$

since the dual norm of the Euclidean norm $\|\cdot\|_2$ coincides with the Euclidean norm.

As in the proof of Lemma 1.6.6, it can be shown that $\mathbb{E}\big[\|h\|_2\big] \leq \sqrt{m}$, using the isotropicity of the A_i, which concludes the proof. □

To achieve a meaningful lower bound of the tail parameter $Q_{S^{\|\cdot\|_2}}(u)$ in Lemma 1.6.3, we will a powerful result about tensorization of stochastic domination as defined in Definition 1.4.2.

Lemma 1.6.10 ([KW92, Theorem 3.2.1])**.** *Let $x_1, \dots, x_m$ and $y_1, \dots, y_m$ be independent random variables such that for all $i \in [m]$, x_i stochastically dominates y_i with constants λ_1 and λ_2. Then, the random vector $X = (x_1, \dots, x_m)$ dominates $Y = (y_1, \dots, y_m)$ with constants $\frac{1}{2\lceil\frac{1}{\lambda_1}\rceil}$ and $\lambda_2\lceil\frac{1}{\lambda_1}\rceil$, or more explicitly,*

$$\mathbb{P}\left(|\langle X, t\rangle| > \gamma\right) \geq \frac{1}{2\lceil\frac{1}{\lambda_1}\rceil}\mathbb{P}\left(|\langle Y, t\rangle| > \lambda_2\left\lceil\frac{1}{\lambda_1}\right\rceil\gamma\right)$$

for any $t \in \mathbb{R}^m$ and $\gamma > 0$.

We proceed to the proof of Theorem 1.4.

Proof of Theorem 1.4. We first show that condition (b) of Theorem 1.4 actually implies condition (a): Let $X = (x_1, \dots, x_m)$ be an isotropic random vector with independent coordinates that dominate standard Gaussian variables g with constants 1 and $\lambda_2 \geq 1$. Then Lemma 1.6.10 implies that X dominates a standard Gaussian vector $G = (g_1, \dots, g_m)$ with constants $1/2$ and λ_2, which corresponds to condition (a).

Without loss of generality, we therefore assume condition (a). Using Lemma 1.6.3 for $S = S^{\|\cdot\|_2}$, $u = \log(\mathrm{e}N/m)$... and $t = N^{\frac{1-\beta}{2}} m^{\frac{\beta}{2}}$, it follows that with probability at least $1 -$

$2\exp(-2N^{1-\beta}m^{\beta})$,

$$\begin{aligned}\inf_{w\in S^{\|\cdot\|_2}}\left(\frac{1}{N}\sum_{i=1}^{N}|\langle A_i,w\rangle|^{\log N}\right)^{\frac{1}{\log N}} &\geq u\left(Q_{S^{\|\cdot\|_2}}(2u)-\frac{4}{u}\mathcal{R}_N(S^{\|\cdot\|_2})-\left(\frac{m}{N}\right)^{\frac{\beta}{2}}\right)^{\frac{1}{\log N}}\\ &\geq u\left(Q_{S^{\|\cdot\|_2}}(2u)-\left(\frac{4}{u}+1\right)\left(\frac{m}{N}\right)^{\frac{\beta}{2}}\right)^{\frac{1}{\log N}},\end{aligned}\tag{1.30}$$

where we used Lemma 1.6.9 in the last inequality.

Let $G=(g_1,\dots,g_m)$ be a standard Gaussian vector. As we have $\|w\|_2=1$ for $w\in S^{\|\cdot\|_2}$, $\langle G,w\rangle$ is a standard Gaussian random variable in this case and it follows from Lemma 1.6.10 that

$$\begin{aligned}&\mathbb{P}(|\langle X,w\rangle|\geq 2u)\geq\frac{1}{2}\,\mathbb{P}(|\langle G,w\rangle|\geq 2\lambda_2 u)=\frac{1}{2}\,\mathbb{P}(|g_1|\geq 2\lambda_2 u)\\ &=\frac{1}{\sqrt{2\pi}}\int_{2\lambda_2 u}^{\infty}\mathrm{e}^{-z^2/2}\,dz\geq\frac{1}{\sqrt{2\pi}}\int_{2\lambda_2 u}^{4\lambda_2 u}\mathrm{e}^{-z^2/2}\,dz\geq\frac{1}{\sqrt{2\pi}}2\lambda_2 u\exp(-8\lambda_2^2u^2).\end{aligned}$$

Inserting that $u=\frac{\sqrt{\beta\log(\mathrm{e}\,N/m)}}{4\sqrt{2}\lambda_2}$, we obtain

$$\begin{aligned}\mathbb{P}\left(|\langle X,w\rangle|\geq\frac{\sqrt{\beta\log(\mathrm{e}\,N/m)}}{2\sqrt{2}\lambda_2}\right)&\geq\frac{4}{\sqrt{\pi}}\sqrt{\beta\log(\mathrm{e}\,N/m)}\exp(-8\lambda_2^2u^2)=\frac{4}{\sqrt{\pi}}\sqrt{\beta\log(\mathrm{e}\,N/m)}\left(\frac{m}{\mathrm{e}\,N}\right)^{\frac{\beta}{4}}\\ &\geq\frac{4\sqrt{\beta}}{\sqrt{\pi}}K\left(\frac{m}{N}\right)^{\frac{\beta}{2}}\end{aligned}$$

if $K>0$ and $N\geq\mathrm{e}\,K^{\frac{4}{\beta}}m$. If K is chosen such that $K=\frac{2\sqrt{\pi\beta}+16\sqrt{2\pi}\lambda_2}{4\beta}$, it follows from that and (1.30) that with probability at least $1-2\exp(-2N^{1-\beta}m^{\beta})$,

$$\begin{aligned}\inf_{w\in S^{\|\cdot\|_2}}\left(\frac{1}{N}\sum_{i=1}^{N}|\langle A_i,w\rangle|^{\log N}\right)^{\frac{1}{\log N}}&\geq\frac{\sqrt{\beta\log(\frac{\mathrm{e}N}{m})}}{4\sqrt{2}\lambda_2}\left(Q_{S^{\|\cdot\|_2}}\left(\frac{\sqrt{\beta\log(\frac{\mathrm{e}N}{m})}}{2\sqrt{2}\lambda_2}\right)-\left(\frac{16\sqrt{2}\lambda_2}{\sqrt{\beta\log(\frac{\mathrm{e}N}{m})}}+1\right)\left(\frac{m}{N}\right)^{\frac{\beta}{2}}\right)^{\frac{1}{\log N}}\\ &\geq\frac{\sqrt{\beta\log(\frac{\mathrm{e}N}{m})}}{4\sqrt{2}\lambda_2}\left(\left(\frac{4\sqrt{\beta}}{\sqrt{\pi}}K-\frac{16\sqrt{2}\lambda_2}{\sqrt{\beta\log(\frac{\mathrm{e}N}{m})}}-1\right)\left(\frac{m}{N}\right)^{\frac{\beta}{2}}\right)^{\frac{1}{\log N}}\\ &\geq\frac{\sqrt{\beta\log(\frac{\mathrm{e}N}{m})}}{4\sqrt{2}\lambda_2}\,\mathrm{e}^{-\beta/2}\end{aligned}$$

if $N\geq\mathrm{e}\,K^{\frac{4}{\beta}}m$. Using Lemma 1.6.2, we conclude that with probability at least $1-2\exp(-2N^{1-\beta}m^{\beta})$, $\frac{1}{\sqrt{m}}A$ fulfills the ℓ_1-quotient property with constant

$$D=\frac{4\sqrt{2}\,\mathrm{e}^{\beta/2}\,\lambda_2}{\sqrt{\beta}}$$

relative to the Euclidean norm $\|\cdot\|_2$ if $N \geq \widetilde{C}m$ with

$$\widetilde{C} = \mathrm{e}\left(\frac{2\sqrt{\pi\beta} + 16\sqrt{2\pi}\lambda_2}{4\beta}\right)^{\frac{4}{\beta}}.$$

□

1.6.3 Proof of Theorem 1.5

We continue with the proof of Theorem 1.5.

Proof. 1. Assume first that b is a random variable with unit variance which fulfills the weak moment assumption of Definition 1.4.3 of order $\max\{4, \log(N)\}$ with constants κ and $\alpha \geq 1/2$. We first note that then, with probability at least $1 - \exp(-c_1(\alpha,\kappa)m)$, A fulfills the ℓ_2-robust null space property of order $\widetilde{c}_2 s_*$, where $s_* = m/\log(\mathrm{e}N/m)$ and $0 < \widetilde{c}_2 < 1$, and constants $\rho = 0.9$, $\tau > 0$ relative to the ℓ_2-norm, i.e.,

$$\|x_S\|_2 \leq \frac{\rho}{s^{1/2}}\|x_{S^c}\|_1 + \tau\|Ax\|_2 \tag{1.31}$$

for all $x \in \mathbb{R}^N$ and $S \subset [N]$ with $s = |S| \leq \widetilde{c}_2 s_*$ and $S^c = [N] \setminus S$ if $m \geq \log^{2\alpha-1}(N)$, where $c_1 = c_1(\alpha,\kappa), \widetilde{c}_2$ and τ depend on κ and α. Indeed, this follows from [DLR18, Theorem V.1] resp. [DLR18, Corollary V.3], as for a row vector $b^{\rightarrow} = (b_1, \dots, b_N)$ with i.i.d. entries $b_i \sim b$ and a vector $t \in \mathbb{R}^N$ with $\|t\|_2 = 1$

$$\mathbb{P}\left(|\langle b^{\rightarrow}, t\rangle| \geq 2^{-1/2}\right) \geq \frac{(1-1/2)^2}{\mathbb{E}[|b|^4]} \geq \frac{1}{4\kappa^4 4^{4\alpha}} =: \delta > 0, \tag{1.32}$$

which holds due to the Paley-Zygmund inequality [DlPG12], [FR13, Lemma 7.17], and furthermore since

$$\frac{\kappa^2 e^{4\alpha-2}}{(2^{-1/2})^2\delta^2} \leq 2\kappa^{10} e^{2\log(2)(8\alpha+1)+4\alpha-2}$$

We call the event on which (1.31) holds $\mathscr{E}_{\mathrm{NSP}}$.

It then follows from the classical result [FR13, Theorem 4.22] that there exist constants $C' > 1$, $D' > 0$ such that if $\|w\|_2 \leq \eta$, on the event $\mathscr{E}_{\mathrm{NSP}}$, it holds that

$$\|x - \Delta_{1,\eta}(Ax + w)\|_p \leq \frac{C'}{s^{1-1/p}}\sigma_s(x)_1 + s_*^{1/p-1/2}D'\eta$$

for $s = \widetilde{c}_2 s_*$, and this is true for all $1 \leq p \leq 2$ and all $x \in \mathbb{R}^N$. Since $s \mapsto \sigma_s(x)_1/s^{1-1/p}$ is non-increasing, we may replace s also by a smaller value $s' < s = \widetilde{c}_2 s_*$, which implies inequality (1.17) for the case of $\|w\|_2 \leq \eta$.

Consider now the case $\|w\|_2 > \eta$ and let $z \in \mathbb{R}^N$ such that $\|Az - w\|_2 \leq \eta$. It follows again by [FR13, Theorem 4.22] that on the event $\mathscr{E}_{\mathrm{NSP}}$, for $1 \leq p \leq 2$, there exist constants

$C' > 1$, $D' > 0$ such that

$$\begin{aligned}
&\|x - \Delta_{1,\eta}(Ax+w)\|_p = \left\|(x+z) - \Delta_{1,\eta}\big(A(x+z) + w - Az\big) - z\right\|_p \\
&\le \left\|(x+z) - \Delta_{1,\eta}\big(A(x+z) + (w - Az)\big)\right\|_p + \|z\|_p \\
&\le \frac{C'}{s^{1-1/p}}\sigma_s(x+z)_1 + s^{1/p-1/2}D'\eta + \|z\|_p \\
&\le \frac{C'}{s^{1-1/p}}\sigma_s(x)_1 + s^{1/p-1/2}D'\eta + \frac{C'}{s^{1-1/p}}\|z\|_1 + \|z\|_p \\
&\le \frac{C'}{s^{1-1/p}}\sigma_s(x)_1 + s^{1/p-1/2}D'\eta + C'\left[\frac{\|z\|_1}{s^{1-1/p}} + \|z\|_p\right] \\
&\le \frac{C'}{s^{1-1/p}}\sigma_s(x)_1 + s^{1/p-1/2}D'\eta + Cs^{1/p-1/2}\left[\frac{\|z\|_1}{s^{1/2}} + s^{1/2-1/p}\|z\|_p\right]. \qquad (1.33)
\end{aligned}$$

Moreover, we note that since $\frac{\beta}{4}\frac{\log(N/m)}{\log(\frac{32}{9}\kappa^4 4^{4\alpha})} \le \log(N)$, the assumptions of Theorem 1.3(b) are fulfilled for $\beta = 1$, and therefore there exist constants $\widetilde{C}$ and D (depending on κ and α) such that with probability at least $1 - 2\exp(-2m)$, $A = \frac{1}{\sqrt{m}}B$ fulfills the ℓ_1-quotient property relative to the norm $\|\cdot\|^{(\sqrt{\log(\mathrm{e}\,N/m)})} := \max\{\|\cdot\|_2, \sqrt{\log(\mathrm{e}\,N/m)}\|\cdot\|_\infty\}$ with constant $\widetilde{D}$ if $N \ge \widetilde{C}m$, and we call the corresponding event $\mathscr{E}_{\text{QP-clipped}}$.

Consider now on the event $\mathscr{E}_{\text{QP-clipped}} \cap \mathscr{E}_{\text{NSP}}$, which occurs with probability at least $1 - 3\exp(-\min\{2, c_1(\alpha,\kappa)\}m)$, a vector $\tilde{z} \in \mathbb{R}^N$ such that $A\tilde{z} = w$ and

$$\|\tilde{z}\|_1 s_*^{-1/2} \le \widetilde{D}\|w\|^{(\sqrt{\log(eN/m)})}, \qquad (1.34)$$

which exists due to the ℓ_1-quotient property relative to the norm $\|\cdot\|^{(\sqrt{\log(eN/m)})}$. If $z \in \mathbb{R}^N$ is chosen such that $z := (1 - \eta/\|w\|_2)\tilde{z}$, it holds that

$$\|Az - w\|_2 = \left\|A\tilde{z} - w - A\tilde{z}\frac{\eta}{\|w\|_2}\right\|_2 = \frac{\eta}{\|w\|_2}\|-A\tilde{z}\|_2 = \eta.$$

Choose S as an index set of s largest absolute coefficients of $\tilde{z}$. It is known [FR13, Chapter 4.3] that the ℓ_2-null space property (1.31) implies the ℓ_p-null space property for $1 \le p \le 2$ in the form

$$\|\tilde{z}_S\|_p \le \frac{\rho}{s^{1-1/p}}\|\tilde{z}_{S^c}\|_p + \tau s^{1/p-1/2}\|A\tilde{z}\|_2.$$

Together with Stechkin's estimate (see, e.g., [FR13, Proposition 2.3]) this gives

$$\begin{aligned}
\|\tilde{z}\|_p &\le \|\tilde{z}_S\|_p + \|\tilde{z}_{S^c}\|_p \le \frac{\rho}{s^{1-1/p}}\|\tilde{z}_{S^c}\|_1 + \tau s^{1/p-1/2}\|A\tilde{z}\|_2 + \sigma_s(\tilde{z})_p \\
&\le \frac{\rho}{s^{1-1/p}}\|\tilde{z}\|_1 + \frac{1}{s^{1-1/p}}\|\tilde{z}\|_1 + \tau s^{1/p-1/2}\|A\tilde{z}\|_2 = \frac{1+\rho}{s^{1-1/p}}\|\tilde{z}\|_1 + \tau s^{1/p-1/2}\|w\|_2 \\
&\le \frac{1+\rho}{s^{1-1/p}}\|\tilde{z}\|_1 + \tau s^{1/p-1/2}\|w\|^{(\sqrt{\log(eN/m)})}.
\end{aligned}$$

(1.35)

Using (1.33) and $s = \widetilde{c}_2 s_*$ we obtain

$$
\begin{aligned}
&\|x - \Delta_{1,\eta}(Ax + w)\|_p \\
&\leq \frac{C'}{s^{1-1/p}}\sigma_s(x)_1 + (\widetilde{c}_2 s_*)^{1/p-1/2}\Big[D'\eta + C'\big(\frac{\|\tilde{z}\|_1}{(\widetilde{c}_2 s_*)^{1/2}} + (\widetilde{c}_2 s_*)^{1/2-1/p}\|\tilde{z}\|_p\big)(1 - \frac{\eta}{\|w\|_2})\Big] \\
&\leq \frac{C'}{s^{1-1/p}}\sigma_s(x)_1 + (\widetilde{c}_2 s_*)^{1/p-1/2}\Big[D'\eta + C'\big(\frac{(\rho + 2)\widetilde{D}}{\widetilde{c}_2^{1/2}} + \tau\big)\|w\|^{(\sqrt{\log(eN/m)})}(1 - \frac{\eta}{\|w\|_2})\Big] \\
&= \frac{C'}{s^{1-\frac{1}{p}}}\sigma_s(x)_1 + s_*^{\frac{1}{p}-\frac{1}{2}}\left[\frac{D'}{\widetilde{c}_2^{\frac{1}{p}-\frac{1}{2}}}\eta + \frac{C'\big((\rho + 2)\widetilde{D} + \widetilde{c}_2^{1/2}\tau\big)}{\widetilde{c}_2^{1-\frac{1}{p}}}\frac{\|w\|^{(\sqrt{\log(eN/m)})}}{\|w\|_2}(\|w\|_2 - \eta)\right],
\end{aligned}
$$

where we used (1.34) and (1.35) in the second inequality. Since $s \mapsto \sigma_s(x)_1/s^{1-1/p}$ is non-increasing, again we may replace s by a smaller value $s' < s = \widetilde{c}_2 s_*$. This concludes the proof of (1.17), as the constants can be defined as $C = C'$, $D = D'\widetilde{c}_2^{\frac{1}{2}-\frac{1}{p}}$ and $E = \widetilde{c}_2^{\frac{1}{p}-1}C'\big((\rho + 2)\widetilde{D} + \widetilde{c}_2^{1/2}\tau\big)$.

2. Suppose now that b is as above, but additionally dominates a standard Gaussian random variable g with constants 1 and $\lambda_2 \geq 1$. As in the first statement, with probability at least $1 - \exp(-c_1(\alpha, \kappa)m)$, A fulfills the ℓ_2-robust null space property (1.31) of order $\widetilde{c}_2 s_*$ with constants $\rho = 0.9$ and $\tau > 0$ relative to the ℓ_2-norm if $m \geq \log^{2\alpha-1}(N)$, defining the event $\mathscr{E}_{\mathrm{NSP}}$.

   Conditional on $\mathscr{E}_{\mathrm{NSP}}$, it follows from [BA18, Theorem II.22] that there exist constants $C > 1$, $D > 0$ that depend only on κ and α such that

$$
\|x - \Delta_{1,\eta}(Ax + w)\|_p \leq \frac{C}{s^{1-1/p}}\sigma_s(x)_1 + s_*^{1/p-1/2}\big(D\eta + C(\mathbb{Q}_{s_*}(A)_1 + \tau)\max\{\|w\|_2 - \eta, 0\}\big) \tag{1.36}
$$

   for all $1 \leq p \leq 2$ and $s \leq \widetilde{c}_2 s_*$, where $s_* = m/\log(\mathrm{e}\,N/m)$ and

$$
\mathbb{Q}_{s_*}(A)_1 := \sup_{w \in \mathbb{R}^m \setminus \{0\}} \min_{u \in \mathbb{R}^N, Au = w} \frac{\|u\|_1}{\sqrt{s_*}\|w\|_2}. \tag{1.37}
$$

   Furthermore, we note that the assumptions on the random matrix B are such that Theorem 1.4 can be applied for $\beta = 1$, which implies that there exist constants $\widetilde{C}$ and D (depending on the parameter λ_2 which quantifies the stochastic domination of the Gaussian) such that with probability at least $1 - 2\exp(-2m)$, $A = \frac{1}{\sqrt{m}}B$ fulfills the ℓ_1-quotient property relative to the Euclidean norm $\|\cdot\|_2$ with constant D if $N \geq \widetilde{C}m$. We call the event on this ℓ_1-quotient property holds $\mathscr{E}_{\text{QP-}\ell_2}$. Recalling Definition 1.3.3, the definition of the ℓ_1-quotient property, we note that in this case

$$
\mathbb{Q}_{s_*}(A)_1 \leq D.
$$

   Thus, combining this with (1.36), we conclude that on the event $\mathscr{E}_{\mathrm{NSP}} \cap \mathscr{E}_{\text{QP-}\ell_2}$, which holds with probability at least $1 - 3\exp(-\min\{2, c_1(\alpha, \kappa)\}m)$, (1.18) is true for all $s \leq \widetilde{c}_2 s_*$ if $N \geq \widetilde{C}n$ and $m \geq \log^{2\alpha-1}(N)$ with the constant $E := CD + \tau$. This concludes the proof. □

1.7 Outlook and Open Problems

In this chapter of the thesis, we studied the robustness of so-called *noise-blind* compressed sensing settings. We showed that in fact, the guarantees that have been established in the literature for measurement matrices whose entries are drawn independently from certain specific distributions or families of distribution like Gaussian and sub-Gaussian are generalizable to the weakest known assumptions provably sufficient (that are known to be almost necessary) to imply compressed sensing recovery guarantees in the noiseless case [ML17, DLR18]. These noise-blind guarantees come in two different versions, depending on whether the entrywise distributions exhibit a *super-Gaussian* behavior or not. The underlying key properties are so-called ℓ_1-quotient properties, which are equivalent to certain geometric inclusions of random polytopes spanned by the column vectors of the measurement matrix.

An interesting problem that remains open is whether statements such as Theorem 1.3 and Theorem 1.4 can be also shown for *structured random matrices*, such as subsampled random circulant matrices [Rom09], time-frequency structured random matrices [PRT13, KMR14] or random partial Fourier matrices [CT06, HR17], because there are hardly any results present in the literature for these measurement matrices of great relevance in applications. Our techniques do not carry over to these matrices, as the columns of the measurement matrix are not independent in these cases, which was an important ingredient in our proof strategy.

In [BA18], the authors derive results about noise-blind compressed sensing using measurement matrices sampled from bounded orthonormal systems, which is a class of random matrices that includes random partial Fourier matrices as a special case. The proof derives a statement about an ℓ_1-quotient property relative to the Euclidean norm indirectly by using lower bounds on the smallest singular value of the measurement matrix. This comes at a price of additional logarithmic factors in a bound [BA18, Theorem IV.12] of the type (1.18). A removal of these logarithmic factors or an analysis involving noise robustness relative to a clipped ℓ_2-norm would be interesting, but is left to future work.

We note that there also exist approaches which modify the measurement matrix A if A does not fulfill any ℓ_1-quotient property. By adding certain columns to A, it might be possible to establish recovery guarantees for noise-blind ℓ_1-decoders which use that modified measurement matrix [Woj12, HLSJ18]. One advantage of this modification is that the acquisition process is not altered, as the modification just involves the decoder. However, the sparse recovery quality of the altered measurement matrix might be compromised by this due to worse restricted isometry or null space properties. Therefore, finding optimal modifications is not yet well-understood.

Finally, we leave it to a future investigation to generalize the conditions on the measurement matrix available in the literature [SY10] for ℓ_p-quotient properties corresponding to ℓ_p-quasinorms with $0 < p < 1$. These are relevant for the noise robustness analysis of the non-convex decoders

$$\Delta_p(y) = \arg\min_{z \in \mathbb{R}^N} \|z\|_p^p \quad \text{subject to } Az = y,$$

which exhibit some advantageous properties compared to convex decoders [Cha07, ZMW⁺17].

Chapter 2

Tight Quadratic Bounds: Making Iteratively Reweighted Least Squares Work For Low-Rank Promoting Objectives

The entirety of this chapter is dedicated to the following basic problem: Assume there is a matrix $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ with d_1 rows and d_2 columns that is known to be of a low rank $\text{rank}(\mathbf{X}_0) \ll \min(d_1, d_2) =: d$. To put it differently, we assume that there are matrices $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$ with $r \ll d$ such that

$$\mathbf{X}_0 = \mathbf{U}\mathbf{V}^*.$$

However, this matrix is not provided directly, but just through *underdetermined linear measurements or observations*, which means through its image

$$\mathbf{y} = \Phi(\mathbf{X}_0)$$

with respect to a linear operator $\Phi : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ with non-trivial null-space, i.e., with

$$m \ll d_1 d_2.$$

Is it possible to identify or recover $\mathbf{X}_0$ just from the knowledge of $\mathbf{y}$ and Φ and if yes, how? This problem is usually called the *low-rank matrix recovery* problem [DR16].

This problem has gained considerable attention in the last few years, as it turns out that *low-rank models* are ubiquitous and arise in many ways in physics, engineering, data science and applied mathematics. In particular, they arise in areas as diverse as system identification [LHV13, LV10], signal processing [AR15], quantum tomography [GLF+10, Gro11], phase retrieval [CSV13, CESV13, GKK15], and *machine learning* [SRJ05]. An instance of this problem of particular importance, e.g., in recommender systems [SRJ05, GNOT92, KBV09, CR09], is the *matrix completion* problem, where the measurements correspond to entries of the matrix to be recovered. In one form or the other, many problems in the mentioned fields can be considered as instances of the low-rank matrix recovery problem.

There is a certain analogy to the compressed sensing problem detailed in Chapter 1, where the goal was to identify sparse, high-dimensional vectors from underdetermined linear measurements. As with the ℓ_0-minimization problem for the sparse vector recovery of Chapter 1, we can formulate the low-rank matrix recovery problem in an optimization framework such that

$$\min_{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}} \text{rank}(\mathbf{X}) \quad \text{subject to } \Phi(\mathbf{X}) = \mathbf{y}, \tag{2.1}$$

a formulation that is also called *affine rank minimization* [RFP10].

If the matrix X_0 used to define (2.1) is diagonal, this problem reduces to the ℓ_0-minimization problem, which is known to be NP-hard in general [Nat95]. However, within the last decade, a large body of research has been dedicated to develop tractable algorithms that provably solve (2.1) in many important cases. The *nuclear norm minimization* (NNM) approach [Faz02, CR09], which solves a surrogate semidefinite program, is particularly well-understood. For NNM, recovery guarantees have been shown for a number of measurements on the order of the information theoretical lower bound $r(d_1 + d_2 - r)$, if r denotes the rank of a $d_1 \times d_2$-matrix [RFP10, CR09]; i.e., for a number of measurements $m \geq \rho r(d_1+d_2-r)$ with some oversampling constant $\rho \geq 1$. Even though NNM is solvable in polynomial time, it can be computationally very demanding if the problem dimensions are large, which is the case in many potential applications. Another issue is that although the number of measurements necessary for successful recovery by nuclear norm minimization is of *optimal order*, it is not *optimal*. More precisely, it turns out that the oversampling factor ρ of nuclear norm minimization *needs to be much larger than the oversampling factor of some other, non-convex algorithmic approaches* [ZL16a, TW13].

These limitations of convex relaxation approaches have led to a rapidly growing line of research discussing the advantages of non-convex optimization for the low-rank matrix recovery problem [JMD10, TW13, HH09, JNS13, WYZ12, TW16, Van13, WCCL16a, TBS+16]. For several of these non-convex algorithmic approaches, recovery guarantees comparable to those of NNM have been derived [CLS15, TBS+16, ZL16a, SL16]. Their advantage is a higher empirical recovery rate and an often more efficient implementation. While there are some results about global convergence of first-order methods minimizing a non-convex objective [GLM16, BNS16] so that a success of the method might not depend on a particular initialization, the assumptions of these results are not always optimal, e.g., in the scaling of the numbers of measurements m in the rank r [GLM16, Theorem 5.3]. For survey articles on the recent progress on the theoretical properties of many non-convex approaches, we refer to [CW18, CLC19]. We note that there is currently no clear picture about which algorithm is *best* to solve low-rank matrix recovery problems.

The material of this chapter is partially based on the paper [KS18], where a similar algorithm for Schatten-p minimization was proposed and analyzed. This paper was jointly conceived and written by the author of this dissertation and Juliane Sigl, who both contributed equally to the paper. The results of the paper [KS18] also appeared in the Ph.D. dissertation [Sig18], and a preliminary version of [KS18] was presented at the 12th International Conference on Sampling Theory and Applications in Tallinn, Estonia, in July 2017 [KS17]. Parts of the analysis contained in this chapter, in particular Section 2.3.2 and Section 2.3.3, take a very different view point than the paper [KS18] by providing a *constructive* approach to design the weight operator. We note that Section 2.3.3 is based on discussions and joint work with Felix Krahmer, Tomáš Masák and Claudio Verdun. Section 2.3.2 and Section 2.3.3 motivate the definition of the algorithm MatrixIRLS, which has many similar properties as the harmonic mean weight operator of [KS18], but the latter work used variational tools in its analysis, in contrast to our constructive approach. The proof of the locally superlinear convergence presented in Section 2.4.2 follows similar ideas as the on in [KS18], but covers also the case of $p = 0$. The results of Section 2.5.3 and Section 2.6 have not been published before writing this dissertation.

2.1 Introduction

The goal of this chapter is to improve solution strategies for matrix optimization problems such as (2.1). We provide a new, improved approach based on a framework usually called

Iteratively Reweighted Least Squares (IRLS), which is a conceptually simple class of algorithms that has been previously studied in many other contexts. We will argue why this approach combines a few desirable properties, since it is

- *data efficient* in the sense that the required amount of inputs is close to the information theoretic limit,
- *scalable* to large problems.

Our investigations will also provide provable *algorithmic guarantees* for our strategy and quantify the local rate of convergence to global minimizers of (2.1) under certain assumptions on the measurements.

In the following, we argue why these properties are desirable in the first place.

2.1.1 Data Efficiency

What is meant by *data efficiency* in the context of the problems we are interested in? In fact, this concept is related to the *uniqueness* of the affine rank minimization problem (2.1), i.e., conditions when (2.1) has not more than one solution of rank $r = \operatorname{rank}(X_0)$.

Proposition 2.1.1 ([Van13, Lee13]). *The set*

$$\mathcal{M}_r = \{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2} : \operatorname{rank}(\mathbf{X}) = r\}$$

is a smooth submanifold *of* $\mathbb{R}^{d_1\times d_2}$ *of dimension* $r(d_1 + d_2 - r)$*.*

If $\mathbf{X} = \mathbf{U\Sigma V}^*$ *is a singular value decomposition of the rank-r matrix* $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$ *with* $\mathbf{\Sigma} = \operatorname{diag}(\sigma_i(\mathbf{X}))_{i=}^r)$ *and matrices* $\mathbf{U} \in \mathbb{R}^{d_1\times r}$ *and* $\mathbf{V} \in \mathbb{R}^{d_2\times r}$ *with orthonormal columns, then the* tangent space *of* $\mathcal{M}_r$ *at* $\mathbf{X}$ *is given by*

$$\begin{aligned} T_X(\mathcal{M}_r) &= \{\mathbf{UM}_1\mathbf{V}^* + \mathbf{UM}_2^* + \mathbf{M}_3\mathbf{V}^* : \mathbf{M}_1 \in \mathbb{R}^{r\times r}, \mathbf{M}_2 \in \mathbb{R}^{d_2\times r}, \mathbf{M}_2^*\mathbf{V} = 0, \mathbf{M}_3 \in \mathbb{R}^{d_1\times r}, \mathbf{M}_3^*\mathbf{U} = 0\} \\ &= \{\mathbf{U}\widetilde{\mathbf{M}}_1^* + \widetilde{\mathbf{M}}_2\mathbf{V}^* : \widetilde{\mathbf{M}}_1 \in \mathbb{R}^{d_2\times r}, \widetilde{\mathbf{M}}_2 \in \mathbb{R}^{d_1\times r}\}. \end{aligned}$$

The proof of this proposition uses tools from differential geometry which are beyond the scope of this thesis. However, understanding its interplay with the linear operator $\Phi : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$, which is determined by matrices $\mathbf{A}_1, \ldots, \mathbf{A}_m \in \mathbb{R}^{d_1\times d_2}$ such that

$$\Phi(\mathbf{X})_\ell = \operatorname{tr}(\mathbf{A}_\ell^*\mathbf{X}) \quad \text{for all } \ell \in [m], \tag{2.2}$$

of (2.1) is a key for obtaining necessary and sufficient conditions for the uniqueness of the solution (2.1). In particular, it was shown in [ENP12] that if the $\mathbf{A}_1, \ldots, \mathbf{A}_m$ are independent Gaussian matrices and $d_1 = d_2$, (2.1) recovers a fixed rank-r matrix $\mathbf{X}_0 \in \mathbb{R}^{d_1\times d_2}$ with probability 1 ("non-uniform recovery") under the condition that

$$m \geq r(d_1 + d_2 - r) + 1,$$

and *all* rank-r matrices ("uniform recovery") under the condition that

$$m \geq 2r(d_1 + d_2 - 2r).$$

Using arguments from algebraic geometry, these identifiability conditions can be generalized to operators Φ associated to matrices $\mathbf{A}_1, \ldots, \mathbf{A}_m$ that are *generic*, i.e., for example drawn from a continuous probability distribution [Xu18, CRWX18]. Furthermore, [CRWX18] showed that for generic measurement matrices $\mathbf{A}_1, \dots, \mathbf{A}_m$, if

$$m < r(d_1 + d_2 - r),$$

even a non-uniform recovery property for (2.1) cannot hold.

From an information theoretic point of view, the dimension

$$\deg_f = r(d_1 + d_2 - r)$$

of $\mathcal{M}_r$ can be also seen as the number of *degrees of freedom* of an $(d_1 \times d_2)$-dimensional rank-r matrix. In fact, the number $\deg_f$ coincides [CR09] with the number of parameters used to determine the singular values $\sigma_1, \dots, \sigma_r$, and left and right singular vectors $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$ of a $(d_1 \times d_2)$-dimensional rank-r matrix $\mathbf{X} = \mathbf{U}\,\mathrm{diag}(\sigma_i)_{i=i}^r \mathbf{V}^*$: The number of free parameters of the matrix $\mathbf{U}$ with orthonormal columns is $(d_1 - 1) + (d_1 - 2) + \dots + (d_2 - (r-1)) = d_1 r - \frac{r(r+1)}{2}$, as can be seen from the normalization, i.e., each columns has at most $d_1 - 1$ free parameters, and by successively imposing the orthogonality relations. By an analogous argument we see that there are $d_2 r - \frac{r(r+1)}{2}$ free parameters in a description of V, and therefore

$$r + d_1 r - \frac{r(r+1)}{2} + d_2 r - \frac{r(r+1)}{2} = r + r(d_1 + d_2) - r(r+1) = r(d_1 + d_2 - r) = \deg_f .$$

These considerations suggest that if the rank r of a matrix $\mathbf{X}_0$ is low such that $r \ll \min(d_1, d_2)$, it is often possible to identify it using (2.1) if the dimension m of the range space of Φ fulfills

$$m \approx \deg_f = r(d_1 + d_2 - r).$$

As there is little hope to solve (2.1) directly in a *computationally* efficient manner, it is of interest to develop algorithms which are possibly iterative, but which have similar *data efficiency* properties as (2.1) in its ability to recover low-rank matrices, i.e., which are successful already if m is about as small as $\deg_f$.

To see why data efficiency is relevant in the the context of applications of our modelling, we consider the *recommender system* setting where low-rank matrix modelling has been used with quite some success.

For a company using *online marketing* of their products, an important question to consider is how to guess what their potential and actual customers might like, and to target personalized products *recommended by an algorithm* to a specific customer based on what they know about the customer, e.g., based on a user's usage history of their website.

With the goal of improving their internally used algorithm, the company *Netflix* announced a competition in 2006 [BL07]: They published an (anonymized) dataset of many *ratings ranging from one to five* provided by the a large collection of users of their DVD rental service, quantifying how much specific users liked certain movies they had watched. The idea of the competition was to award $1,000,000 to any person or team that could improve the predicitions of user preferences by at least 10% compared to the algorithm used at the time by the company.

This competition about winning the *Netflix Prize* spurred a large research activity in the emerging field of *machine learning*, culminating in involved algorithms that combined many statistical models and algorithms *tuned* to the data, and a final conclusion of the competition in 2009 by reaching the 10% improvement [Kor09].

However, one of the most influential ideas and contributing models to the winning solution [Fun06, KBV09] was the idea to consider the *user-movie data matrix* $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ for d_1 users and

d_2 movies, whose (i,j)-th entry $\mathbf{X}_{ij}$ is given by the i-th user's rating of movie j. Of course, this matrix is in large parts unknown. While this matrix contains a lot of unknown values, it was argued that, if all users' movie preferences were known, i.e., the matrix contained no unknown entries, the resulting matrix $\mathbf{X}_0$ would be (approximately) *of low rank* $r \ll d = \min(d_1, d_2)$ [RFP10], as any given user could be modelled as a *linear combination* of just *few* typical *model users* or *preference types*.

Based on this modelling, many variants of methods were proposed and successfully used in the competition [Fun06, KBV09] which would be called *non-convex optimization approaches based on matrix factorization* [CLC19] in the recent literature, optimizing a function $F : \mathbb{R}^{d_1 \times r} \times \mathbb{R}^{d_1 \times r} \to \mathbb{R}$ for some tuning parameter λ and some estimate of the rank r, defined for example such that

$$F(\mathbf{U},\mathbf{V}) := \|\Phi(\mathbf{U}\mathbf{V}^*) - \mathbf{y}\|_F^2 + \lambda \left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 \right) \tag{2.3}$$

by an alternating minimization or (stochastic) gradient descent scheme, where Φ from (2.2) is such that

$$\Phi(\mathbf{X})_\ell = \mathbf{X}_{i_\ell j_\ell}$$

with $(i_\ell, j_\ell) \in [d_1] \times [d_2]$ for $\ell = 1, \ldots, m$, and m is the total number of given ratings in the dataset. These schemes attempt, in fact, to solve (2.1), while using knowledge about the rank r of $\mathbf{X}$.

Assuming that the low-rank model is reasonable, it can be said that for the purpose of building a recommender system, it is important that an algorithm of choice of this type is able to identify the *best low-rank model compatible with the training data*, based on the limited training data of m of user's movie ratings that are available. This means that it will be helpful if an algorithm returns a feasible point of (2.1) with a rank that this as low as possible, or, in the formulation of (2.3), matrices $\mathbf{U}$ and $\mathbf{V}$ of a rank as low as possible, while still obtaining a measurement error $\|\Phi(\mathbf{U}\mathbf{V}^*) - \mathbf{y}\|_F^2$ that is zero or as small as possible.

2.1.2 Amenability to Rigorous Analysis

Important steps towards the understanding of the low-rank matrix recovery problem were achieved in [CR09, CT10, RFP10, Rec11, Che15], when it proved that in many cases, a number of measurements

$$m \approx CrD \tag{2.4}$$

or

$$m \approx CrD \log^2(D) \tag{2.5}$$

where $D = \max(d_1, d_2)$ and C is an absolute constant, suffices to reconstruct a rank-r solution $\mathbf{X}_0$ of (2.1) by the solving convex optimization formulation

$$\min_{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}} \|\mathbf{X}\|_* \text{ subject to } \Phi(\mathbf{X}) = \mathbf{y}, \tag{2.6}$$

where $\|\mathbf{X}\|_* = \|\mathbf{X}\|_{S_1}$ is the so-called *nuclear norm*.

The nuclear norm is an instance of a class of quasi-norms that plays an important role in this chapter. For $p > 0$, We call

$$\|\mathbf{X}\|_{S_p} := \begin{cases} \left(\sum_{i=1}^{d} \sigma_i^p(\mathbf{X}) \right)^{1/p}, & \text{for } p < \infty, \\ \sigma_1(\mathbf{X}), & \text{for } p = \infty \end{cases}$$

the *Schatten-p quasi-norm* for the matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$.

The great advantage of the *nuclear norm minimization* formulation (2.6) is that, unlike (2.1), it is solvable in *polynomial time complexity* as it belongs to the class of *convex optimization* algorithms [Ber99, BV04], since the the Schatten-p quasinorm is actually a norm for $p \geq 1$, and (2.6) corresponds to $p = 1$.

As (2.6) can be solved directly, and due to the *guarantee* that its solution corresponds to the minimum-rank solution of (2.1), this gives us a certain level of reliance that even for problem data we have not seen yet, the algorithm will work reliably, and furthermore, we know that there cannot be a *lower* rank solution than the one returned by nuclear norm minimization, if the conditions of the theorems of [CT10, RFP10, Rec11] are fulfilled. In particular for more critical applications as quantum tomography [GLF⁺10, Gro11, BK10], this *reliability* is important.

Similar assumptions, often stronger than (2.4) and (2.5) by logarithmic factors $\log(D)$, powers of the rank r or by the dependence on the condition number κ of the matrix to be recovered, have been shown to provably imply the convergence of *non-convex* optimization methods, such as for various methods optimizing an objective such as (2.3), to low-rank solutions, see for example the works [KMO10, CLS15, GLM16, SL16, MWCC19] and the survey [CLC19].

On the other hand, we note that for approximate message passing algorithms [PSC14a, PSC14b, KKM⁺16] or methods motivated from a Bayesian framework [BLMK12, XW15], very good low-rank recovery properties have been reported—in some sense even better ones. For these methods, rigorous sample complexity bounds as in the works mentioned above have *not* been proven, and, perhaps consequently, these methods have not gained the attention they could have if it was just for their empirical performance.

This could be regarded as an extra-mathematical motivation for actually *proving* algorithmic guarantees. Therefore, one important focus of this chapter will also be the establishment of theoretical guarantees of the proposed algorithm.

2.1.3 Scalability

In many applications of the low-rank recovery setting, the problem dimensions are quite large: The Netflix prize competition data set, for example, contains data about $d_1 \approx 480000$ users and $d_2 \approx 17000$ with $m \approx 10^8$ observations in the training dataset [ZWSP08]. In this case, working with *full* or *dense* matrices $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, storing $d_1 d_2$ variables in memory, is computationally almost prohibitive.

The nuclear norm formulation of (2.6) has been well-studied and well-understood, as we mentioned above. However, solvers need to work with the full matrices in memory, and furthermore, while (2.6) being provably equivalent to a semidefinite program [FHB03], standard semidefinite program solvers as SDPT3 [TTT99] of cubically in the problem size of $d_1 d_2$. To mitigate this problem, a lot of work has been done to make solvers of (2.6) somehow more scalable, for example by using first order optimization methods involving soft thresholding operators [CCS10, MHT10, YK15]. Most of these first-order methods are, however, rather designed to solve the unconstrained variant of (2.6)

$$\min_{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}} \|\Phi(\mathbf{X}) - \mathbf{y}\|_2^2 + \lambda \|\mathbf{X}\|_*$$

for some $\lambda > 0$, which, while being also of large interest, is not quite the same. Other approaches [CC18] for solving nuclear-norm type problems are based on conditional gradients [RSW15] or the Frank-Wolfe method [FGM17].

Not only, but also because of the scalability problems of convex optimization approaches as the nuclear norm formulation, in large-scale applications, methods based on *matrix factorization* as described above (2.3) have been more popular [KBV09], due to their memory requirement of $O(rD)$ and also lower demands in their time complexity [CC18].

This motivates the derivation of a method that has comparable space and time requirements as the *matrix factorization* approach, rather than working with full $(d_1 \times d_2)$-dimensional matrices.

2.1.4 Outline

We give a brief overview of the structure of this chapter. In Section 2.2, we introduce *low-rank promoting surrogate objectives* of the rank function, laying out the theory of the functions to be optimized by our framework. In Section 2.3, we introduce the concept of *Iteratively Reweighted Least Squares* (IRLS) algorithms, discussing prior literature and deriving our algorithmic approach, which we call `MatrixIRLS`. This section contains our first main theoretical result Theorem 2.4, which is the basis for an interpretation of IRLS in terms of an objective function.

In Section 2.4, we show and state results about the convergence behavior of `MatrixIRLS`, concluding the section with the second main result of this chapter, Theorem 2.7, characterizing fast local convergence of `MatrixIRLS`.

While the conditions of Theorem 2.7 are stated in terms of an abstract assumption, we verify this assumption for various measurement models in Section 2.5, including sub-Gaussian operators, Gaussian rank-one measurements and entry-wise measurements as present for the *matrix completion* problem.

Section 2.6 contains computational considerations, elaborating on the scalability of our method, and Section 2.7 contains numerical experiments with artificial data. Section 2.8 contains proofs that have beed added for completeness, corresponding to properties that have been know already for previous IRLS algorithms for low-rank matrix recovery.

Finally, in Section 2.9, we discuss and summarize our contributions, and provide an outlook to remaining open questions in Section 2.10.

2.2 Optimization of Surrogate Objectives Promoting Low-Rank Solutions

In this section, we pave the ground for `MatrixIRLS`, a computational approach for the affine rank minimization problem

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \operatorname{rank}(\mathbf{X}) \quad \text{subject to } \Phi(\mathbf{X}) = \mathbf{y},$$

of (2.1) which is the main focus of this chapter.

While (2.1) is the problem we would like to solve, its objective $\operatorname{rank}(\mathbf{X})$ is particularly difficult to handle from an optimization point of view due to two properties: its *non-convexity* and its *non-smoothness*. As it was discussed in the introduction, replacing $\operatorname{rank}(\mathbf{X})$ by the nuclear norm $\|\mathbf{X}\|_* = \sum_{i=1}^{d} \sigma_i(\mathbf{X})$, which is the norm associated to the convex hull of the set of (Frobenius norm-normalized) rank-one matrices [CRPW12], gives rise to an optimization problem whose minimizer solves (2.1) in many cases, opening the door for the mature theory and methods for *convex optimization* to deal with the non-smoothness of $\|\mathbf{X}\|_*$.

However, putting even its computational challenges aside for the moment, it can be seen that its *data efficiency* properties are not optimal, i.e., the nuclear norm minimizer under the affine constraint might not coincide with the (unique) minimizer $\mathbf{X}_0$ of (2.1) if the dimension m of the range space of Φ is just slightly larger than the number of the degrees of freedom of $\mathbf{X}_0$ (see also Section 2.1.1) [RFP10, p. 497].

Already in the beginning of the 2000s when the minimization of the nuclear norm was proposed as a computationally tractable “heuristic” [Faz02] promoting low-rank solutions of linear matrix equations, the minimization of the *logarithm of the determinant (log-det)* objective

$$\min_{\mathbf{X}\in\mathbb{R}^{d\times d}} \log\det(\mathbf{X}+\epsilon\mathbf{I}) \quad \text{subject to } \Phi(\mathbf{X})=\mathbf{y}, \mathbf{X}\succeq 0, \tag{2.7}$$

was proposed, where $\epsilon > 0$ is a small regularization parameter for low-rank matrix recovery problems for which the ground truth matrix $\mathbf{X}_0 \in \mathbb{R}^{d_1\times d_2}$ is additionally known to be square (i.e., $d_1 = d_2 = d$) and positive semidefinite [FHB03]. The rationale behind $\log\det(\mathbf{X}+\epsilon\mathbf{I})$ is that due to its concavity and just logarithmic growth, its minimizers might be closer to the minimizers of a rank objective than the minimizers of the convex nuclear norm $\|\mathbf{X}\|_*$ [RFP10], which coincides with the *trace* of $\mathbf{X}$ for positive semidefinite matrices since

$$\|\mathbf{X}\|_* = \sum_{i=1}^{d}\sigma_i(\mathbf{X}) = \sum_{i=1}^{d}\lambda_i(\mathbf{X}) = \operatorname{tr}(\mathbf{X}).$$

Iterative algorithms for the solution of (2.7) had been proposed early [Faz02, FHB03], and can be based on the following observation: Since $F_\epsilon(\mathbf{X}) = \log\det(\mathbf{X}+\epsilon\mathbf{I})$ is concave in $\mathbf{X}$ and smooth with derivative $\nabla F_\epsilon(\mathbf{X}) = (\mathbf{X}+\epsilon\mathbf{I})^{-1}$, for any positive semidefinite $\mathbf{X}^{(k)} \in \mathbb{R}^{d\times d}$, we have that

$$\begin{aligned}\log\det(\mathbf{X}+\epsilon\mathbf{I}) &\leq F_\epsilon(\mathbf{X}^{(k)}) + \langle \nabla F_\epsilon(\mathbf{X}^{(k)}), \mathbf{X}-\mathbf{X}^{(k)}\rangle \\ &= \log\det(\mathbf{X}^{(k)}+\epsilon\mathbf{I}) + \operatorname{tr}((\mathbf{X}^{(k)}+\epsilon\mathbf{I})^{-1}\mathbf{X}) - \operatorname{tr}((\mathbf{X}^{(k)}+\epsilon\mathbf{I})^{-1}\mathbf{X}^{(k)}),\end{aligned}$$

which motivates an iterative algorithm of a sequence of *weighted trace minimizations* such that

$$\mathbf{X}^{(k+1)} = \operatorname*{arg\,min}_{\mathbf{X}\succeq 0, \Phi(\mathbf{X})=\mathbf{y}} \operatorname{tr}((\mathbf{X}^{(k)}+\epsilon\mathbf{I})^{-1}\mathbf{X}) \quad \text{for } k \geq 0. \tag{2.8}$$

This approach has been also more recently proposed and analyzed for the *phase retrieval* problem [CESV13], which can be formulated as a rank-1 matrix recovery problem on the cone of positive semidefinite matrices.

Remark 2.2.1. *We note that well-known positive semidefinite embedding of general rectangular matrices $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$, an analogous approach can be derived for the general affine rank minimization problem (2.1) without positive semidefinite constraint* [FHB03]*. In particular, as for (2.7), iterative strategies can be derived from*

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}, \mathbf{Z}_1\in\mathbb{R}^{d_1\times d_1}, \mathbf{Z}_2\in\mathbb{R}^{d_2\times d_2}} \log\det(\operatorname{diag}(\mathbf{Z}_1,\mathbf{Z}_2)+\epsilon\mathbf{I}) \quad \textit{subject to } \Phi(\mathbf{X})=\mathbf{y}, \begin{pmatrix}\mathbf{Z}_1 & \mathbf{X}\\ \mathbf{X}^* & \mathbf{Z}_2\end{pmatrix} \succeq 0,$$

since $\operatorname{diag}(\mathbf{Z}_1,\mathbf{Z}_2)$ *is positive semidefinite if the constraint* $\begin{pmatrix}\mathbf{Z}_1 & \mathbf{X}\\ \mathbf{X}^* & \mathbf{Z}_2\end{pmatrix} \succeq 0$ *is fulfilled. One computational disadvantage of this formulation we mention already here is that by the semidefinite lifting, the optimization variable has become an $((d_1+d_2)\times(d_1+d_2))$ matrix instead of a $(d_1\times d_2)$ matrix, which might be wasteful.*

Despite the good empirical behavior of simple iterative procedures as (2.8) for solving (2.7) [FHB03], there are many theoretical questions related to non-convex rank surrogates as in (2.7) which are poorly understood, and practical issues deterring a wide popularity of corresponding optimization approaches to large-scale applications of low-rank matrix recovery.

2.2.1 Connections Between Convex and Non-Convex Optimization Programs and Affine Rank Minimization

In this section, we review theory that characterizes when the affine rank minimization problem (2.1) can be solved exactly by convex or non-convex reformulations such as nuclear norm minimization (2.6) or the minimization of a logarithm of a determinant (2.7).

In a certain analogy to the (robust) null space properties (see Definition 1.4.5) of measurement matrices A which imply that equality-constrained ℓ_1-minimization (1.2) recovers sparse vectors, null space properties can also be formulated for measurement operators $\Phi : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ defining low-rank matrix recovery problems, for example as in the following definition.

Definition 2.2.1 ([RXH11, MBS14, Fou18]). *Let $d = \min(d_1, d_2)$ and $f : [0,\infty) \to [0,\infty)$ be a function. We say that $\Phi : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ fulfills the* rank null space property with respect f of order r *if*

$$\sum_{i=1}^{r} f(\sigma_i(\mathbf{H})) < \sum_{i=r+1}^{d} f(\sigma_i(\mathbf{H})) \quad \textit{for all } \mathbf{H} \in \ker(\Phi) \setminus \{0\},$$

where $\sigma(\mathbf{H}) = (\sigma_1(\mathbf{H}), \ldots, \sigma_d(\mathbf{H})) \in \mathbb{R}^d$ is the vector of singular values of $\mathbf{H}$ ordered in a non-increasing way.

These null space properties provide necessary and sufficient conditions for the equivalence of the optimization problems

$$\min_{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}} F(\mathbf{X}) \quad \text{subject to } \Phi(\mathbf{X}) = \mathbf{y}, \tag{2.9}$$

where $\sigma(\mathbf{X}) \in \mathbb{R}^d$ is the vector of ordered singular values of $\mathbf{X}$ and F the spectral function such that $F(\mathbf{X}) = \sum_{i=1}^{d} f(\sigma_i(\mathbf{X}))$, with affine rank minimization (2.1). This connection is formalized in the following theorem, whose proof we provide as it has been omitted in [Fou18]. It generalizes the proof of [FR13, Theorem 4.40], which covers the case of f being the identity map.

Theorem 2.1 ([Fou18, Theorem 2]). *Let $f : [0,\infty) \to [0,\infty)$ be a concave function satisfying $f(0) = 0$. Given a linear operator $\Phi : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$, every matrix $X \in \mathbb{R}^{d_1 \times d_2}$ of rank at most r fulfilling $\mathbf{y} = \Phi(\mathbf{X})$ is the unique solution of (2.9) if and only if Φ fulfills the rank null space property with respect to f of order r.*

Proof. We first assume that Φ is such that every $\mathbf{X}' \in \mathbb{R}^{d_1 \times d_2}$ of rank at most r is the unique solution of (2.9) with $\mathbf{y} = \Phi(\mathbf{X}')$. Consider a matrix $\mathbf{H} \in \ker(\Phi) \setminus \{0\}$ and its singular value decomposition

$$\mathbf{H} = \begin{pmatrix} \mathbf{U} & \mathbf{U}_\perp \end{pmatrix} \begin{pmatrix} \operatorname{diag}(\sigma_1(\mathbf{H}), \ldots, \sigma_r(\mathbf{H})) & 0 \\ 0 & \operatorname{dg}(\sigma_{r+1}(\mathbf{H}), \ldots, \sigma_d(\mathbf{H})) \end{pmatrix} \begin{pmatrix} \mathbf{V}^* \\ \mathbf{V}_\perp^* \end{pmatrix},$$

where $\mathbf{U} \in \mathbb{R}^{d_1 \times r}, \mathbf{U}_\perp \in \mathbb{R}^{d_1 \times (d_1 - r)}, \mathbf{V} \in \mathbb{R}^{d_2 \times r}$ and $\mathbf{V}_\perp \in \mathbb{R}^{d_2 \times (d_2 - r)}$ are matrices with orthonormal columns and $\mathbf{U}^*\mathbf{U}_\perp = 0$, $\mathbf{V}^*\mathbf{V}_\perp = 0$. Then $\mathbf{H}_1 = \mathbf{U}\operatorname{diag}(\sigma_1(\mathbf{H}), \ldots, \sigma_r(\mathbf{H}))\mathbf{V}^*$ and $\mathbf{H}_2 = \mathbf{U}_\perp \operatorname{dg}(\sigma_{r+1}(\mathbf{H}), \ldots, \sigma_d(\mathbf{H}))\mathbf{V}_\perp^*$ fulfill $\mathbf{H} = \mathbf{H}_1 + \mathbf{H}_2$ and as $\mathbf{H} \in \ker(\Phi)$, $\Phi(\mathbf{H}_1) = \Phi(-\mathbf{H}_2)$. Since $\mathbf{H}_1$ is of rank at most r by construction, using the assumption on the unique minimizer of (2.9) with $\mathbf{y} = \Phi(\mathbf{H}_1)$ implies that

$$\sum_{i=1}^{r} f(\sigma_i(\mathbf{H})) = F(\sigma(\mathbf{H}_1)) < F(\sigma(-\mathbf{H}_2)) = F(\sigma(\mathbf{H}_2)) = \sum_{i=r+1}^{d} f(\sigma_i(\mathbf{H})).$$

Conversely, assume that Φ is such that for all $\mathbf{H} \in \ker(\Phi) \setminus \{0\}$, $\sum_{i=1}^{r} f(\sigma_i(\mathbf{H})) < \sum_{i=r+1}^{d} f(\sigma_i(\mathbf{H}))$ if $\sigma(\mathbf{H}) = (\sigma_1(\mathbf{H}), \ldots, \sigma_d(\mathbf{H}))$ are the ordered singular values of $\mathbf{H}$. Consider a matrix $\mathbf{X}' \in \mathbb{R}^{d_1 \times d_2}$ of rank at most r, and let $\mathbf{X}$ be an arbitrary matrix $\mathbf{X} \neq \mathbf{X}'$ such that $\Phi(\mathbf{X}) = \Phi(\mathbf{X}')$. Due to the concave Mirksy inequality [Fou18, Theorem 1], which states that for any concave function $f : [0, \infty) \to [0, \infty)$ with $f(0) = 0$,

$$\sum_{i=1}^{d} |f(\sigma_i(\mathbf{X})) - f(\sigma_i(\mathbf{Y}))| \leq \sum_{i=1}^{d} f(\sigma_i(\mathbf{X} - \mathbf{Y})) \quad \text{for all } \mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d_1 \times d_2}, \tag{2.10}$$

and which is a generalization of a classical inequality in matrix analysis due to Mirksy [Mir60], [SS92, Theorem 4.11], we have that the singular values of $\mathbf{H} = \mathbf{X}' - \mathbf{X}$, $\mathbf{X}'$ and $\mathbf{X}$ fulfill

$$F(\mathbf{X}) = F(\mathbf{X}' - \mathbf{H}) = \sum_{i=1}^{d} f(\sigma_i(\mathbf{X}' - \mathbf{H})) \geq \sum_{i=1}^{d} |f(\sigma_i(\mathbf{X}')) - f(\sigma_i(\mathbf{H}))|.$$

Since $f(\sigma_i(\mathbf{X}')) = \sigma_i(\mathbf{X}') =$ for $i > r$, we conclude that

$$F(\mathbf{X}) \geq \sum_{i=1}^{r} \{f(\sigma_i(\mathbf{X}')) - f(\sigma_i(\mathbf{H}))\} + \sum_{i=r+1}^{d} f(\sigma_i(\mathbf{H})) > \sum_{i=1}^{r} f(\sigma_i(\mathbf{X}')) + 0 = F(\mathbf{X}'),$$

where we used the rank null space property assumption in the last inequality. This concludes the proof as $\mathbf{X}'$ is therefore the unique minimizer of (2.9). □

From Theorem 2.1, we can see that the rank null space property with respect to the function $f = \mathrm{id}$ characterizes the success of nuclear norm minimization. This property with respect to the function $f = \mathrm{id}$ of order r holds with high probability for many random linear operators $\Phi : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ if m is such that

$$m \geq Cr(d_1 + d_2), \tag{2.11}$$

where $C \geq 1$ is a dimensionless constant, example for random operators Φ whose matrix representation consists of i.i.d. standard Gaussians [CP11, RXH11] or which are associated to Gaussian rank-one measurements [KKRT16]. We note that (2.11) is optimal, at least up to a constant factor, as can be seen in Section 2.1.1.

The next statement tells us that from a theoretical point of view, optimization problems with spectral objectives F associated to concave functions f tend to have *better low-rank recovery properties* than the those with convex objectives such as the nuclear norm. It is a generalization of [Fou18, Corollary 3], see also [Aud14].

Lemma 2.2.1. *Let $r \in \mathbb{N}$, let $f_1, g : [0, \infty) \to [0, \infty)$ be functions such that f_1 is monotonously increasing and g is concave with $g(0) = 0$. Then the success of rank-r recovery via* (2.9) *where the associated function of F is f_1, implies the success of rank-r recovery via* (2.9) *where the associated function of F is $f_2 := g \circ f_1$.*

Proof. By assumption it holds the linear operator $\Phi : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ is such that every matrix

$\mathbf{X}' \in \mathbb{R}^{d_1 \times d_2}$ of rank at most r is the unique solution of

$$\min_{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}} \sum_{i=1}^{d} f_1(\sigma_i(\mathbf{X})) \quad \text{subject to } \Phi(\mathbf{X}) = \mathbf{y}$$

with $\mathbf{y} = \Phi(\mathbf{X}')$. By Theorem 2.1, this implies that Φ fulfills the rank null space property with respect to f_1 of order r, i.e.,

$$\sum_{i=1}^{r} f_1(\sigma_i(\mathbf{H})) < \sum_{i=r+1}^{d} f_1(\sigma_i(\mathbf{H})) \quad \text{for all } \mathbf{H} \in \ker(\Phi) \setminus \{0\},$$

and by this very theorem, it remains to show that

$$\sum_{i=1}^{r} f_2(\sigma_i(\mathbf{H})) < \sum_{i=r+1}^{d} f_2(\sigma_i(\mathbf{H})) \quad \text{for all } \mathbf{H} \in \ker(\Phi) \setminus \{0\}$$

for f_2 with $f_2(x) = g(f_1(x))$ for any $x \geq 0$, which is equivalent to

$$\sum_{i=1}^{r} \frac{1}{\sum_{j=r+1}^{d} g(f_1(\sigma_j(\mathbf{H})))/g(f_1(\sigma_i(\mathbf{H})))} < 1 \quad \text{for all } \mathbf{H} \in \ker(\Phi) \setminus \{0\}. \tag{2.12}$$

Supposing that

$$\frac{g(f_1(\sigma_j(fH)))}{g(f_1(\sigma_i(\mathbf{H})))} \geq \frac{f_1(\sigma_j(\mathbf{H}))}{f_1(\sigma_i(\mathbf{H}))} \tag{2.13}$$

for all $i \in [r]$ and $j \in \{r+1, \ldots, d\}$ if $\mathbf{H} \in \ker(\Phi) \setminus \{0\}$, the conclusion of the lemma follows, as

$$\sum_{i=1}^{r} \frac{1}{\sum_{j=r+1}^{d} g(f_1(\sigma_j(\mathbf{H})))/g(f_1(\sigma_i(\mathbf{H})))} \leq \sum_{i=1}^{r} \frac{1}{\sum_{j=r+1}^{d} f_1(\sigma_j(\mathbf{H})))/f_1(\sigma_i(\mathbf{H}))}$$

in this case and therefore (2.12) is implied by the rank null space property with respect to f_1 of order r.

It remains to show (2.13). For this, we note that for any $\mathbf{H} \in \ker(\Phi) \setminus \{0\}$, $\sigma_j(\mathbf{H}) \leq \sigma_i(\mathbf{H})$ if $i \in [r]$ and $j \in \{r+1, \ldots, d\}$, and furthermore $f_1(\sigma_j(\mathbf{H})) \leq f_1(\sigma_i(\mathbf{H}))$ due to the monotonicity of f_1. By the concavity of g, as $f_1(\sigma_j(\mathbf{H}) = (1-\lambda) \cdot 0 + \lambda \cdot f_1(\sigma_i(\mathbf{H}))$ if $\lambda = f_1(\sigma_j(\mathbf{H}))/f_1(\sigma_i(\mathbf{H}))$, it follows that

$$g(f_1(\sigma_j(\mathbf{H}))) \geq (1-\lambda) \cdot g(0) + \lambda \cdot g(f_1(\sigma_i(\mathbf{H}))) = \lambda \cdot g(f_1(\sigma_i(\mathbf{H}))) = \frac{f_1(\sigma_j(\mathbf{H}))g(f_1(\sigma_i(\mathbf{H})))}{f_1(\sigma_i(\mathbf{H}))},$$

since $g(0) = 0$, which concludes the proof. □

From Lemma 2.2.1, we can infer that ϵ-regularized Schatten-p quasinorm minimization has stronger low-rank recovery properties than nuclear norm minimization if $0 < p < 1$, and that a ϵ-regularized log-det minimization has stronger low-rank recovery properties than ϵ-regularized Schatten-p quasinorm minimization for any $0 < p \leq 1$.

Corollary 2.2.1. *Let $f_{p,\epsilon} : [0, \infty) \to [0, \infty)$ be the function defined such that $f_{p,\epsilon}(x) = (x+\epsilon)^p - \epsilon^p$ for $\epsilon > 0$, $0 < p \leq 1$, and $f_{0,\epsilon} : [0, \infty) \to [0, \infty)$ be such that $f_{0,\epsilon}(x) = \log(x + \epsilon) - \log(\epsilon)$ for $\epsilon > 0$. Then, the following statements hold.*

1. *If a linear operator* $\Phi : \mathbb{R}^{d_1\times d_2} \longrightarrow \mathbb{R}^m$ *is such that for all matrices* $\mathbf{X}' \in \mathbb{R}^{d_1\times d_2}$ *of rank at most* r *and* $fy = \Phi(\mathbf{X}')$*, the unique minimizer of the nuclear norm minimization problem*

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \|\mathbf{X}\|_* \quad \textit{subject to} \quad \Phi(\mathbf{X}) = \mathbf{y} \tag{2.14}$$

   coincides with $\mathbf{X}'$*, then the minimizers of the optimization problems*

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \sum_{i=1}^{d} f_{p,\epsilon}(\sigma_i(\mathbf{X})) \quad \textit{subject to} \quad \Phi(\mathbf{X}) = \mathbf{y} \tag{2.15}$$

   are unique and coincide with $\mathbf{X}'$*, for all* $p \in [0, 1)$ *and all* $\epsilon > 0$.

2. *Let* $\epsilon_1 > \epsilon_2 > 0$ *and* $0 < p < 1$ *be fixed. If a linear operator* $\Phi : \mathbb{R}^{d_1\times d_2} \longrightarrow \mathbb{R}^m$ *is such that for all matrices* $\mathbf{X}' \in \mathbb{R}^{d_1\times d_2}$ *of rank at most* r *and* $\mathbf{y} = \Phi(\mathbf{X}')$*, the unique minimizer of the* ϵ_1-regularized Schatten-p minimization problem

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \sum_{i=1}^{d} (\sigma_i(\mathbf{X}) + \epsilon_1)^p \quad \textit{subject to} \quad \Phi(\mathbf{X}) = \mathbf{y}$$

   coincides with $\mathbf{X}'$*, then the minimizer of the of the* ϵ_2-regularized Schatten-p minimization problem

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \sum_{i=1}^{d} (\sigma_i(\mathbf{X}) + \epsilon_2)^p \quad \textit{subject to} \quad \Phi(\mathbf{X}) = \mathbf{y}$$

   is unique and coincides with $\mathbf{X}'$*. Furthermore, this is not true if* $0 < \epsilon_1 < \epsilon_2$.

3. *Let* $\epsilon_1 > \epsilon_2 > 0$*. If a linear operator* $\Phi : \mathbb{R}^{d_1\times d_2} \longrightarrow \mathbb{R}^m$ *is such that for all matrices* $\mathbf{X}' \in \mathbb{R}^{d_1\times d_2}$ *of rank at most* r *and* $\mathbf{y} = \Phi(\mathbf{X}')$*, the unique minimizer of the* ϵ_1-regularized log-det minimization problem

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \sum_{i=1}^{d} \log(\sigma_i(\mathbf{X}) + \epsilon_1) \quad \textit{subject to} \quad \Phi(\mathbf{X}) = \mathbf{y}$$

   coincides with $\mathbf{X}'$*, then the minimizer of the of the* ϵ_2-regularized log-det minimization problem

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \sum_{i=1}^{d} \log(\sigma_i(\mathbf{X}) + \epsilon_2) \quad \textit{subject to} \quad \Phi(\mathbf{X}) = \mathbf{y}$$

   is unique and coincides with $\mathbf{X}'$*. Furthermore, this is not true if* $0 < \epsilon_1 < \epsilon_2$.

4. *Let* $\epsilon > 0$ *and* $0 \leq p < q \leq 1$*. If a linear operator* $\Phi : \mathbb{R}^{d_1\times d_2} \longrightarrow \mathbb{R}^m$ *is such that for all matrices* $\mathbf{X}' \in \mathbb{R}^{d_1\times d_2}$ *of rank at most* r *and* $\mathbf{y} = \Phi(\mathbf{X}')$*, the unique minimizer of*

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \sum_{i=1}^{d} f_{q,\epsilon}(\sigma_i(\mathbf{X})) \quad \textit{subject to} \quad \Phi(\mathbf{X}) = \mathbf{y}$$

   coincides with $\mathbf{X}'$*, then this also holds for the optimization problem* (2.15).

Proof. 1. In view of Lemma 2.2.1, we can define $f_1(x) := x$ and $g(x) := f_2(x) = (x+\epsilon)^p - \epsilon^p$. $g : [0, \infty) \longrightarrow [0, \infty)$ fulfills then $g(0) = 0$, is twice differentiable on its domain with

$g''(x) = p(p-1)(x+\epsilon)^{p-2}$. As a differentiable function with non-positive second derivative, g is concave for $p < 1$, and applying Lemma 2.2.1 implies the statement.

2. Let $\epsilon_1 > \epsilon_2 > 0$. Since adding and subtracting constants ϵ_1^p and ϵ_2^p does not change the minimizer of an objective function, we can define $f_1(x) := (x+\epsilon_1)^p - \epsilon^p$ and $f_2(x) := (x+\epsilon_2)^p - \epsilon^p$. Then we have $g(f_1(x)) = f_2(x)$ if we define $g : [0,\infty) \to [0,\infty)$ such that
$$g(y) := \left[(y+\epsilon_1^p)^{1/p} - \epsilon_1 + \epsilon_2\right]^p - \epsilon_2^p,$$
and it holds that $g(0) = 0$. Since g is furthermore twice differentiable with
$$g'(y) = \frac{(y+\epsilon_1^p)^{1/p-1}}{\left((y+\epsilon_1^p)^{1/p} + \epsilon_2 - \epsilon_1\right)^{1-p}}$$
and
$$g''(y) = \left(\frac{1}{p} - 1\right)\left(\frac{(y+\epsilon_1^p)^{1/p-2}}{\left((y+\epsilon_1^p)^{1/p} + \epsilon_2 - \epsilon_1\right)^{1-p}} - \frac{(y+\epsilon_1^p)^{2/p-2}}{\left((y+\epsilon_1^p)^{1/p} + \epsilon_2 - \epsilon_1\right)^{2-p}}\right),$$
which is non-positive on the domain and thus, g is concave. The result follows from Lemma 2.2.1.

3. This follows from the same argument as in 2., defining the function $g : [0,\infty) \to [0,\infty)$ such that
$$g(y) := \log\left(\exp(y+\epsilon_1) - \epsilon_1 + \epsilon_2\right) - \log(\epsilon_2).$$

4. This follows from defining $f_1(x) := (x+\epsilon)^q - \epsilon^q$, $f_2(x) := (x+\epsilon)^p - \epsilon^p$ and $g(y) := (y+\epsilon^q)^{p/q} - \epsilon^p$, and from applying Lemma 2.2.1 along the same lines.

□

Disregarding the potential intractability related to the minimization of the non-convex functions $\sum_{i=1}^d f_{p,\epsilon}(\sigma_i(\mathbf{X}))$ for $p < 1$ for the moment, we summarize the conclusions that can be drawn from Corollary 2.2.1:

- Minimizing non-convex ϵ-regularized Schatten-p quasinorms related to $p < 1$ recovers low-rank solutions to underdetermined linear systems always when nuclear norm minimization (2.14) recovers the low-rank solution. There might be cases in which nuclear norm minimization does not recover the low-rank solution, but ϵ-regularized Schatten-p quasinorm does.

- If $\epsilon_2 < \epsilon_1$, it can be seen that ϵ_2-regularized Schatten-p quasinorm minimization recovers low-rank solutions *in potentially more cases* than ϵ_1-regularized Schatten-p quasinorm minimization, and an analogue statement is true for ϵ_2- and ϵ_1-regularized log-det minimization.

- For fixed regularization parameter $\epsilon > 0$, ϵ-regularized logdet minimization
$$\min_{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}} \sum_{i=1}^d \log(\sigma_i(\mathbf{X}) + \epsilon) \quad \text{subject to} \quad \Phi(\mathbf{X}) = \mathbf{y} \tag{2.16}$$
recovers low-rank solutions *in potentially more cases* than ϵ-regularized Schatten-p quasinorm minimization for any $0 < p < 1$, or than nuclear norm minimization.

2.2.2 Issues and Open Questions

The discussion of Section 2.2.1 suggests that minimizing non-convex objectives such as (2.15) or (2.16) under the linear constraint could be promising, as solving these problems will return us the actual solution of (2.1) in many cases, while the functions being now at least *continuous*, unlike the rank funciton $\text{rank}(\mathbf{X})$.

However, due to their non-convexity, (2.15) or (2.16) do not possess the property of *convex optimization* problems that local minima are actually global, which means that algorithms such as variants of gradient descent, which converge in general to *(first order) stationary points* [Ber99], do not necessarily solve the problem, but might find non-global minimizers or saddle points instead of global minimizers.

One important class of algorithms that has been proposed to solve these type of problems has been mentioned at the beginning of this section: *Reweighted algorithms* such as (2.8) [Faz02, FHB03] which–in the general case of rectangular matrices– boil often down to *reweighted nuclear norm* algorithms [MF10b].

For these methods, excellent empirical results have been reported in the literature [LTYL15, YM18, GZZF14, GXM+17], but results about conditions for convergence to a low-rank matrix have not been obtained. The situation is somewhat comparable to what is known about *(re-)weighted ℓ_1-minimization* methods for sparse recovery [CWB08, WN10, ZL12], which achieve good empirical results, but which have not been understood completely on a theoretical level.

A major issue of reweighted nuclear norm methods is that from a computational viewpoint, at each (outer) iteration, a problem similar to nuclear norm minimization (2.6) needs to be solved, which results in the scalability issues described above.

Lastly, the question of the *choice of the ϵ-smoothing* parameter is an issue, as it is not clear how to choose it and there is little theory supporting any specific rule. For a discussion of smoothing parameters in the sparse recovery setting, we refer to [WN10, Section V].

As a summary, we conclude that while the modelling with non-convex rank objectives appears promising based on empirical results obtained for a few methods, many open questions remain, which is probably why they have been mainly ignored by the most influential literature in the field [DR16, CLC19].

2.3 Iteratively Reweighted Least Squares for Rank Surrogates

In this section of the thesis, we discuss a specific class of optimization approaches to minimize rank surrogate objectives as (2.9), (2.7), (2.6) under linear constraints.

These approaches belong to the algorithmic framework of *Iteratively Reweighted Least Squares (IRLS)*, a framework which goes back to the 1930s [Wei37].

The basic strategy of *IRLS* is to mimic the minimization of an objective F (a spectral function F as in (2.9)) by solving a sequence of weighted least squares problems.

In particular, for some $\epsilon > 0$, assume that $F_\epsilon : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ is a continuously differentiable function that can be considered as a *smoothing* of an objective F to minimized as in (2.9),

$$\min_{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}} F(\mathbf{X}) \quad \text{subject to } \Phi(\mathbf{X}) = \mathbf{Y}.$$

The idea of *Iteratively Reweighted Least Squares* now consists of an iterative algorithm the performs the following *three main conceptual steps* in each iteration k:

1. **Define tight quadratic upper bound:** Given the current iterate $\mathbf{X}^{(k)}$ and the smoothing

parameter $\epsilon_k > 0$, we define a quadratic function $Q_{\epsilon_k}(\cdot|\mathbf{X}^{(k)})$ such that

$$Q_{\epsilon_k}(\mathbf{X}|\mathbf{X}^{(k)}) = F_{\epsilon_k}(\mathbf{X}^{(k)}) + \langle \nabla F_{\epsilon_k}(\mathbf{X}^{(k)}), \mathbf{X} - \mathbf{X}^{(k)} \rangle + \frac{1}{2}\langle \mathbf{X} - \mathbf{X}^{(k)}, W^{(k)}(\mathbf{X} - \mathbf{X}^{(k)})\rangle \tag{2.17}$$

with the so-called *weight matrix* (or *weight operator*) $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ fulfilling

$$\nabla F_{\epsilon_k}(\mathbf{X}^{(k)}) = W^{(k)}(\mathbf{X}^{(k)}) \tag{2.18}$$

and which is a **global upper bound** or **global majorizer** of F_{ϵ_k} such that

$$Q_{\epsilon_k}(\mathbf{X}|\mathbf{X}^{(k)}) \geq F_{\epsilon_k}(\mathbf{X}) \quad \text{for any } \mathbf{X} \in \mathbb{C}^{d_1 \times d_2}. \tag{2.19}$$

By construction, Q_{ϵ_k} fulfills further

$$Q_{\epsilon_k}(\mathbf{X}^{(k)}|\mathbf{X}^{(k)}) = F_{\epsilon_k}(\mathbf{X}^{(k)}). \tag{2.20}$$

2. **Solve weighted least squares problem:** The new iterate $\mathbf{X}^{(k)}$ is defined as

$$\mathbf{X}^{(k+1)} = \underset{\mathbf{X}:\Phi(\mathbf{X})=\mathbf{Y}}{\arg\min} \langle \mathbf{X}, W^{(k)}(\mathbf{X})\rangle, \tag{2.21}$$

which is actually a minimizer of $Q_{\epsilon_k}(\cdot|\mathbf{X}^{(k)})$ under the linear data constraint described by the linear operator Φ and data vector $y \in \mathbb{C}^{\Omega}$ from (2.9). This is due to property (2.20), as therefore

$$Q_{\epsilon_k}(\mathbf{X}|\mathbf{X}^{(k)}) = F_{\epsilon_k}(\mathbf{X}^{(k)}) + \frac{1}{2}\left(\langle \mathbf{X}, W^{(k)}(\mathbf{X})\rangle - \langle \mathbf{X}^{(k)}, W^{(k)}(\mathbf{X}^{(k)})\rangle\right)$$

for any $\mathbf{X} \in \mathbb{C}^{d_1 \times d_2}$.

It is well-known (see [NW06, Chapter 16.1]) that the solution of (2.21) is given explicitly by the solution of a linear system, as the weight operator $W^{(k)}$ is always chosen to be positive definite.

It is clear from (2.19) and (2.20) that $F_{\epsilon_k}(\mathbf{X}^{(k+1)}) \leq F_{\epsilon_k}(\mathbf{X}^{(k)})$, i.e., we obtain a decrease in the objective F_{ϵ_k} by the calculation of the new iterate $\mathbf{X}^{(k+1)}$.

3. **Update smoothing:** The smoothing parameter ϵ_k controls both the steepness of the optimization landscape of $F_{\epsilon_k}(\cdot)$ and the conditioning of the system matrix of the linear system to be solved in (2.21). Therefore its choice plays an important role in this framework.

Typically, a non-decreasing rule for ϵ_{k+1} is chosen. Precise rules will be discussed later in this chapter.

While the focus of this chapter is the application of the IRLS framework to the affine rank minimization problem where F is a low-rank promoting spectral function with matrix-valued optimization variable $\mathbf{X}$, IRLS can be used for a range of quite different problems. In Section 2.3.1, we provide an overview of the most relevant literature about IRLS methods, also discussing some prior work where IRLS has been already used for low-rank matrix optimization problems.

By establishing connections to the second-order derivative structure of low-rank promoting spectral functions, we discuss in Section 2.3.2 choices of quadratic bounds Q_ϵ that are much tighter than the ones proposed previously in the literature. In Section 2.3.1, we show that the proposed choice of quadratic functions Q_ϵ, which a priori majorizes the smoothed spectral function F_ϵ just locally, is indeed a global majorizer of F_ϵ. This corresponds to the first main result of the chapter.

Based on this, we formulate an improved class of IRLS algorithms for Schatten-p quasi-norm minimization and for logdet-minimization in Section 2.3.4, which we call `MatrixIRLS`.

2.3.1 Existing Approaches

Possibly the very first instance of a method that contains the main idea of IRLS was proposed by E. Weiszfeld [Wei37, WP09] in 1937 for the *Fermat-Weber problem* of finding a vector $\hat{\mathbf{x}} \in \mathbb{R}^d$ that minimizes the sum of (weighted) Euclidean distances to the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^d$,

$$\hat{\mathbf{x}} = \min_{\mathbf{x} \in \mathbb{R}^d} \sum_{i=1}^{m} \omega_i \|\mathbf{x} - \mathbf{a}_i\|_2 \tag{2.22}$$

with weights $\omega_1, \ldots, \omega_m > 0$. Unlike the IRLS framework above, Weiszfeld's algorithm does not involve an objective with updated smoothing – for convex objectives, the smoothing is less crucial than for non-convex objectives. A modern exposition including many historical remarks can be found in [BS15]. We note that in statistics, the objective of (2.22) (with $\omega_i = 1$ for any $i \in [m]$) is called *spatial median* [DM87] or L_1-median [VZ00] of the data points $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^m$.

The name *Iteratively Reweighted Least Squares* comes from *regression* problems: While in classical *linear* regression the regression coefficients $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of a linear model with explanatory variables in the columns of an (overdetermined) matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m > n$ and dependent variables $\mathbf{y}_1, \ldots, \mathbf{y}_m$ are calculated by solving the *least squares* problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2, \tag{2.23}$$

regression estimates that are *more robust to outliers* from the statistical model can be obtained by solving an ℓ_p-regression problem [Nyq83] for $1 \leq p \leq 2$ such as

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_p^p. \tag{2.24}$$

Unlike (2.23), its solution cannot be obtained by solving just a linear system if $p < 2$, but a simple computational strategy [Sch73, MS74, BT74, HW77] with remarkable properties to obtain the solution of ℓ_p-regression consists of iterating

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \arg\min_{\mathbf{x} \in \mathbb{R}^n} \sum_{i=1}^{m} w_i^{(k)} \left((\mathbf{A}\mathbf{x})_i - \mathbf{y}_i\right)^2, \\ w_i^{(k+1)} &= \left|(\mathbf{A}\mathbf{x}^{(k+1)})_i - \mathbf{y}_i\right|^{p-2} \quad \text{for any } i \in [m]. \end{aligned} \tag{2.25}$$

Osborne [Osb85, Chapter 5.4] showed that this IRLS algorithm is locally convergent to the solution of (2.24) with a linear rate for $1 \leq p < 3$[1]. ℓ_p-regression problems are of interested also beyond the classical statistical context, e.g., in semi-supervised learning using graph p-Laplacians [BH09, EACR+16], also for $p > 2$. In approximation theory, similar algorithms for L_∞- and L_p-norm approximation with orthogonal polynomials had already been proposed in

[1] For $p = 1$, the additional assumption that the solution of (2.24) is unique is needed.

the 1960s [Law61, RU68]. A historical account of some of the history of IRLS methods in the approximation theory community can be found in [Wat01].

In the image processing community, the idea of IRLS is usually known under the name of *half-quadratic minimization* [AV97, NN05], interpreting the procedure as an alternating minimization of a variational functional that is *quadratic* in one of its two variables. The first papers to introduce this idea to obtain efficient algorithms for minimizing total variation [ROF92] type of regularizations are [GR92, GY95]. In [ODBP15], IRLS algorithms are considered as a part of a general framework of iteratively reweighted algorithms to minimize non-convex objectives for applications in computer vision.

The IRLS framework was formulated and extended to general convex and coercive functionals defined on a convex domain in [BDMS09], allowing for a general class of continuously differentiable smoothing functions F_ϵ and respective quadratic approximations Q_ϵ. In [Bec15], asymptotic sublinear rate of convergence to the minimizer of a convex functional F was shown for an IRLS scheme corresponding to a fixed smoothing parameter ϵ.

In Chapter 1, the problem of finding sparse solutions of underdetermined linear systems with system matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$, $m \ll n$, and measurement vector $\mathbf{y} \in \mathbb{R}^m$, was already intensively presented and studied. An IRLS algorithm for the *log-sum* objective, the vector analogue of the *log-det* objective from (2.7), and for ℓ_p-minimization

$$\min_{A\mathbf{x}=\mathbf{y}} \|\mathbf{x}\|_p^p \tag{2.26}$$

for $0 < p \leq 1$ has been proposed in [GR97, RK99] under the name FOCUSS. An additional challenge compared to the problems tackled by the IRLS algorithms above is here the potential, severe non-convexity of the objective, as in (2.26) for $p < 1$. This renders the convergence to non-global, local minima that are far from the optimal function value to be an issue [GR97, WN10].

To mitigate this, IRLS algorithms with smoothing strategies for (2.26) were proposed in [CY08, DDFG10, WN10, Vor12, VD17], and a non-trivial theoretical analysis of the algorithm's convergence behavior was provided in [DDFG10]. In particular, it was shown that, under the assumption that A fulfills an ℓ_p*-null space property* [FR13, Chapter 4] and that there is a sparse solution $\mathbf{x}^0$ compatible with the measurements, the IRLS strategy of [DDFG10]

- converges globally to the sparse solution (convex case of $p = 1$), and that
- for any $0 < p < 1$, the corresponding IRLS strategy converges to the sparse solution if the iterations begin with a vector sufficient close to $\mathbf{x}^0$.
- Furthermore, a local convergence analysis showed *locally linear convergence* for $p = 1$ and, interestingly, *locally superlinear convergence* with a rate of $2 - p$ for $0 < p < 1$.

A result about a local rate of convergence of order $2 - p$ can be already found also for FOCUSS [GR97, RK99], which, unlike the IRLS of [DDFG10], does not use any smoothing of the objective. The result even holds for the log-sum case, which can be seen as the limit case $p \to 0$ in a certain sense, where a *quadratic* local convergence rate is available. However, FOCUSS has worse sparse recovery properties than the smoothed IRLS variants [CY08], and unlike the local convergence analysis of [DDFG10], [GR97, Theorem 3] is of asymptotic nature.

The works most related to this chapter of the thesis are the papers [MF10a, MF12, FRW11], where Iteratively Reweighted Least Squares algorithms for the low-rank recovery problem (2.1) were proposed and analyzed. In [FRW11], an algorithm called IRLS-M was conceived to minimize the smoothed nuclear norm

$$F_\epsilon(\mathbf{X}) = \sum_{i=1}^{d_1} f_{1,\epsilon}(\sigma_i(\mathbf{X})) \tag{2.27}$$

with

$$f_{1,\epsilon}(\sigma_i(\mathbf{X})) = \begin{cases} \sigma_i(\mathbf{X}), & \text{if } \sigma_i(\mathbf{X}) > \epsilon, \\ \frac{1}{2}\left(\frac{\sigma_i^2(\mathbf{X})}{\epsilon} + \epsilon\right), & \text{if } \sigma_i(\mathbf{X}) \leq \epsilon, \end{cases}$$

of a matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$. This is done by using quadratic bounds $Q_{\epsilon_k}(\cdot|\mathbf{X}^{(k)})$ as defined in (2.17) corresponding to the weight operator $W^{(k)}$ defined by

$$W^{(k)}(\mathbf{X}) = (\mathbf{X}^{(k)}\mathbf{X}^{(k)*})_{\epsilon_k}^{-1/2}\mathbf{X}, \tag{2.28}$$

where $(\mathbf{X}^{(k)}\mathbf{X}^{(k)*})_{\epsilon_k}$ is the matrix resulting from replacing its singular values $\left(\sigma_i^2(\mathbf{X}^{(k)})\right)_{i=1}^d$ by $\left(\max(\epsilon_k^2, \sigma_i^2(\mathbf{X}^{(k)})\right)_{i=1}^d$. Furthermore, the smoothing ϵ_k is updated according to the rule that for some $c > 0$,

$$\epsilon_{k+1} = \max\left(\epsilon_k, c\sigma_{r+1}(\mathbf{X}^{(k)})\right), \tag{2.29}$$

for any iteration $k \in \mathbb{N}$, where r is a guess of the rank of the matrix X^0 of lowest rank compatible with the measurements such that $\Phi(X_0) = Y$ in (2.1).

The authors of [FRW11] show that the accumulation points of the sequence of iterates $(X^{(k)})_{k\in\mathbb{N}}$ are minimizers of F_{ϵ^*} under the linear constraint defined by the measurements (Φ, Y) if $\epsilon^* := \lim_{k\to\infty}\epsilon_k > 0$. Furthermore, it was shown that the smoothing parameters eventually convergence to 0, i.e., $\epsilon^* = 0$, and the iterates converge to the nuclear norm minimizer if Φ fulfills an appropriate *restricted isometry* or *strong null space property* (see Section 2.2.1 for a discussion).

In this sense, it can be said that the analysis for IRLS-M has generalized the corresponding property of the IRLS strategy of [DDFG10] for sparse vector recovery to low-rank matrix recovery as far as the first bullet point above is concerned, but a generalization of the other two could not be achieved.

Around the same time, but independently, the papers [MF10a, MF12] proposed a family of IRLS algorithms for low-rank matrix recovery to minimize, for a given parameter $0 \leq p \leq 1$, the smoothed functionals $F_\epsilon : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}$ given by

$$F_\epsilon(\mathbf{X}) = \begin{cases} \sum_{i=1}^{d_2} \left(\sigma_i^2(\mathbf{X}) + \epsilon^2\right)^{\frac{p}{2}} & \text{for } 0 < p \leq 1, \\ \sum_{i=1}^{d_2} \log\left(\sigma_i^2(\mathbf{X}) + \epsilon^2\right) & \text{for } p = 0, \end{cases} \tag{2.30}$$

corresponding to smoothed *Schatten-p* quasi-norms for $p > 0$ and to a smoothed *logdet* objective for $p = 0$. At iteration k, their algorithms use quadratic bounds $Q_{\epsilon_k}(\cdot|\mathbf{X}^{(k)})$ corresponding to the weight operator

$$W^{(k)}(\mathbf{X}) = \mathbf{X}(\mathbf{X}^{(k)*}\mathbf{X}^{(k)} + \epsilon_k^2\mathbf{I}_{d_2})^{\frac{p}{2}-1}, \tag{2.31}$$

while using the smoothing update rule

$$\epsilon_{k+1} = \frac{\epsilon_k}{\eta} \tag{2.32}$$

for any $k \in \mathbb{N}$ and some given parameters $\epsilon_0 > 0$, $\eta > 1$.

For this IRLS algorithm, [MF10a, MF12] obtain a similar (but slightly weaker) convergence statement as [FRW11] in the nuclear norm case corresponding to $p = 1$. For the non-convex cases $p < 1$, they also show that the accumulation points of the iterate sequence $(X^{(k)})_{k\in\mathbb{N}}$ are stationary points of the smoothed functionals (2.30) under the linear constraint. Unlike in the sparse vector recovery case [DDFG10], local convergence statements or rates of convergence have not been obtained for $p < 1$.

In [LXY13], the authors presented an analysis of extensions of the IRLS algorithms of [DDFG10] and of [MF12, FRW11] to *unconstrained* ℓ_p-minimization and Schatten-p minimization, that is, for IRLS algorithms designed to solve

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_p^p + \frac{1}{2\lambda}\|A\mathbf{x} - \mathbf{y}\|_2^2$$

or

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \|\mathbf{X}\|_{S_p}^p + \frac{1}{2\lambda}\|\Phi(\mathbf{X} - \mathbf{Y}\|_F^2$$

for a given regularization parameter $\lambda > 0$. For their algorithms, the authors prove analogue results to the ones of [DDFG10] and of [MF12, FRW11]. While the authors claim that the local convergence rate analysis of their sparse recovery IRLS [LXY13, Theorem 2.9] will extend to the low-rank matrix recovery case and is left to be shown by the interested reader [LXY13, p.950], no numerical evidence is presented to support this claim.

Finally, a few recent works were dedicated to a more refined convergence analysis of IRLS algorithms without updated smoothing: For continuously differentiable objectives and not necessarily convex F_ϵ, the authors of [RZ15] show convergence of the sequence of iterates to the set of stationary points and provides a qualitative analysis of basin of attractions of local minima. [RYZ18] extends these results to the case of inexactly solved weighted least squares problems, for example by conjugate gradient methods, and strengths the analysis of [RZ15] by showing that the iterates' trajectory has finite length with convergence to stationary points. A similar statement follows from [BP16, Theorem 6].

2.3.2 The Quest for Optimal Quadratic Bounds: Second-Order Structure of Spectral Functions

In Section 2.3.1, we saw that it has been possible to derive IRLS algorithms for non-convex ℓ_p-minimization problems (2.26) with local convergence guarantees under appropriate assumptions on the null space property of the linear operator defining the problem, including guarantees indicating a local rate of convergence of order $2 - p$ [DDFG10]. It is also well-documented that this superlinear convergence rate often can be observed already after quite few iterations, for example in [DDFG10, Section 8.1], making IRLS a simple and scalable, viable algorithmic option for ℓ_p-quasi norm minimization [ZMW+17].

However, in the low-rank matrix recovery context, the IRLS algorithms of [MF12, FRW11] for Schatten-p quasi-norm minimization

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \sum_{i=1}^{d} \sigma_i^p(\mathbf{X}) \quad \text{subject to } \Phi(\mathbf{X}) = \mathbf{y},$$

with $0 < p < 1$ have not been known to possess local convergence guarantees of a similar type, despite the similar nature of the two problems, and neither has any superlinear rate been observed empirically [MF12, Figures 1-2].[2]

[2] This is also the case for the natural extension of `IRLS-M` of [FRW11] to $p < 1$, adapting the exponent of the weight matrix in its weight operator $W^{(k)}$.

This raises the following questions:

1. **Why do the quasi-norm variants of the IRLS algorithms** [MF12, FRW11] **not exhibit a fast convergence behavior?** Is it due to a fundamental difference between the structure of sparse and low-rank recovery problems?

2. **Is it possible to define improved IRLS algorithms for non-convex rank surrogates which combine fast local convergence rates with good global convergence behavior?**

In the rest of this section, we shed light on both questions.

From an algorithmic point of view, it can be said that the algorithms of [MF12] and [FRW11] differ in basically two main details (apart from implementation details): The choice of the smoothing, including the choice of ϵ_k (see (2.29) and (2.32)) and the fact that the *weight operator* $W^{(k)}$ of [FRW11] acts through left-multiplication on $\mathbf{X}$, whereas $W^{(k)}$ acts through right-multiplication on $\mathbf{X}$ in the IRLS proposed by [MF12].

Recalling the results for sparse recovery, it can be conjectured that the local convergence rate does not depend on the exact choice of the smoothing, or at least, that any fast convergence rate would be rather deterred than promoted by any sort of smoothing, as even the unsmoothed IRLS variant FOCUSS [GR97, RK99] exhibited the fast convergence rate of order $2 - p$.

For this reason, leaving details in the smoothing aside, and taking the version of [MF12] as a starting point, it is straightforward to conceive an IRLS algorithm based on the minimization of weight least squares problems

$$\mathbf{X}^{(k+1)} = \underset{\mathbf{X}:\Phi(\mathbf{X})=\mathbf{y}}{\arg\min} \langle \mathbf{X}, W^{(k)}(\mathbf{X})\rangle \tag{2.33}$$

with weight operator $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$,

$$W^{(k)}(\mathbf{X}) = (\mathbf{X}^{(k)}\mathbf{X}^{(k)*} + \epsilon_k^2 \mathbf{I}_{d_1})^{\frac{p}{2}-1}\mathbf{X}, \tag{2.34}$$

which we call IRLS-col as the matrix $(\mathbf{X}^{(k)}\mathbf{X}^{(k)*} + \epsilon_k^2 \mathbf{I}_{d_1})^{\frac{p}{2}-1}$ acts on the *columns* of $\mathbf{X}$ through left-multiplication. Following up on this convention, we call the IRLS algorithms defined by the weight operator (2.31) in the subsequent discussion IRLS-row, as the weight matrix $(\mathbf{X}^{(k)*}\mathbf{X}^{(k)} + \epsilon_k^2 \mathbf{I}_{d_2})^{\frac{p}{2}-1}$ acts on $\mathbf{X}$ through right-multiplication, corresponding to a transformation of the *rows* of $\mathbf{X}$.

It is important to observe that the quadratic forms of (2.33) are quite *different*, even in the case of rectangular matrices $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ with $d_1 = d_2$, depending on whether IRLS-row or IRLS-col is used: If $\mathbf{X}^{(k)} = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^*$ is a singular value decomposition of a rectangluar matrix $\mathbf{X}^{(k)}$, then

$$\mathbf{W}_{\text{col}}^{(k)} := (\mathbf{X}^{(k)}\mathbf{X}^{(k)*} + \epsilon_k^2 \mathbf{I}_{d_1})^{\frac{p}{2}-1} = \mathbf{U}_k \operatorname{diag}(\widetilde{\sigma}_i^{(k)})_{i=1}^d \mathbf{U}_k^*,$$

where

$$\widetilde{\sigma}_i^{(k)} = \left((\Sigma_k)_{ii}^2 + \epsilon_k^2\right)^{\frac{p-2}{2}} \tag{2.35}$$

for $i \in [d]$, but

$$\mathbf{W}_{\text{row}}^{(k)} := (\mathbf{X}^{(k)*}\mathbf{X}^{(k)} + \epsilon_k^2 \mathbf{I}_{d_2})^{\frac{p}{2}-1} = \mathbf{V}_k \operatorname{diag}(\widetilde{\sigma}_i^{(k)})_{i=1}^d \mathbf{V}_k^*,$$

so that, for example, if $\mathbf{U}_k = \mathbf{I}_{d_1}$,

$$\begin{aligned}\langle \mathbf{X}, W^{(k)}(\mathbf{X})\rangle = \operatorname{tr}(\mathbf{X}^* W^{(k)}(\mathbf{X})) = \operatorname{tr}\left(\mathbf{X}^*\mathbf{U}_k \operatorname{diag}(\widetilde{\sigma}_i^{(k)})_{i=1}^d \mathbf{U}_k^*\mathbf{X}\right) = \operatorname{tr}\left(\mathbf{X}^* \operatorname{diag}(\widetilde{\sigma}_i)_{i=1}^d \mathbf{X}\right)\\ = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} \widetilde{\sigma}_j^{(k)} \mathbf{X}_{ij}^2\end{aligned}$$

for `IRLS-col`, but

$$\langle \mathbf{X}, W^{(k)}(\mathbf{X})\rangle = \operatorname{tr}(\mathbf{X}^* W^{(k)}(\mathbf{X})) = \operatorname{tr}\left((\mathbf{X}\mathbf{V}_k)^*\mathbf{X}\mathbf{V}_k \operatorname{diag}(\widetilde{\sigma}_i^{(k)})_{i=1}^d\right) = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} \widetilde{\sigma}_i^{(k)} (\mathbf{X}^{(k)}\mathbf{V}_k)_{ij}^2$$

if we decide to use `IRLS-row`, so that the landscape of the quadratic forms is very much different, for example, if $\mathbf{V}_k$ is far from the identity.

In fact, it turns out that the IRLS algorithms of [MF12] and [FRW11] only use **either** information about the column space **or** about the row space of the current iterate $X^{(k)}$ to define the quadratic forms to be minimized in the next iteration, but **not** information about both of them. This seems to be quite unnatural from an information-theoretic point of view, as a certain low-rank (rank-r) matrix $\mathbf{X}^0$ is characterized by the fact that it is an element of the intersection of its (r-dimensional) row space and its (r-dimensional) column space.

The theoretical convergence properties of the algorithms in the papers that have been our main focus of discussion, [DDFG10, MF12, FRW11], are derived using a variational approach, interpreting the main steps (2.17) and (2.21) of the IRLS algorithm as an *alternating minimization* of a functional depending on both $\mathbf{X}$ and the weight operator W. While this approach serves to a certain extent its purpose, it is not a *constructive approach* to derive more suitable weight operator rules, as the precise form of the weight operator is already implicit in the design of the variational functional (which is called $\mathcal{J}$ in the three papers).

First-order derivation of IRLS algorithms

To understand from where a potential deficiency of `IRLS-row` and `IRLS-col` arises, let us recall the following, simple argument to derive IRLS algorithms, which goes back to [GR92], see also [CBAB97, Theorem 1]:

For the recovery of a sparse, n-dimensional vector $\mathbf{x}$, let $0 < p \leq 2$ and consider the ϵ_k-smoothed ℓ_p quasi-norm $F_{\epsilon_k}(\mathbf{x}) := \frac{2}{p}\sum_{i=1}^n (|x_i|^2 + \epsilon_k^2)^{p/2}$. Then the function $G_{\epsilon_k} : \mathbb{R}^n \to \mathbb{R}$, $G_{\epsilon_k}(\mathbf{x}) := F_{\epsilon_k}(\sqrt{|\mathbf{x}|})$ is *concave* in its argument and therefore (see, e.g., [BV04, Chapter 3.1])

$$\begin{aligned}F_{\epsilon_k}(\mathbf{x}) = G_{\epsilon_k}((\mathbf{x})^2) \leq G_{\epsilon_k}((\mathbf{x}^{(k)})^2) + \langle \nabla G_{\epsilon_k}((\mathbf{x}^{(k)})^2), \mathbf{x}^2 - (\mathbf{x}^{(k)})^2\rangle\\ = F_{\epsilon_k}(\mathbf{x}^{(k)}) + \sum_{i=1}^d \frac{1}{((\mathbf{x}_i^{(k)})^2 + \epsilon_k^2)^{1-p/2}}\left(\mathbf{x}_i^2 - (\mathbf{x}_i^{(k)})^2\right) =: Q_{\epsilon_k}(\mathbf{x}|\mathbf{x}^{(k)}),\end{aligned}$$

where $\nabla G_{\epsilon_k} : \mathbb{R}^n \to \mathbb{R}^n$ is the *gradient* of G_{ϵ_k} fulfilling

$$F_{\epsilon_k}(\mathbf{x}) = \nabla G_{\epsilon_k}(\mathbf{y}) = \operatorname{diag}\left(\frac{1}{(y_i + \epsilon_k^2)^{1-p/2}}\right)_{i=1}^n,$$

and $\mathbf{x}^2$ and $(\mathbf{x}^{(k)})^2$ denote the vectors of squared coordinates of $\mathbf{x}$ and $\mathbf{x}^{(k)}$, respectively.

This shows already that we have obtained with $Q_{\epsilon_k}(\cdot|\mathbf{x}^{(k)})$ a valid, global quadratic bound

(2.19) of F_{ϵ_k} which coincides with F_{ϵ_k} at $\mathbf{x} = \mathbf{x}^{(k)}$, i.e., (2.20) holds. This gives rise to precisely the algorithm of [DDFG10], as $Q_{\epsilon_k}(\cdot|\mathbf{x}^{(k)})$ can now be minimized under the the linear constraint to obtain a descent in the function value of $F_{\epsilon_k}(\mathbf{x})$.

An analogue argument can be used to derive the algorithms `IRLS-col` and `IRLS-row`: Indeed, if $F_{\epsilon_k} : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}$ is the ϵ_k-smoothed Schatten-p quasi-norm such that

$$F_{\epsilon_k}(\mathbf{X}) = \frac{2}{p} \operatorname{tr} \left(\mathbf{X}\mathbf{X}^* + \epsilon_k^2 \mathbf{I}_{d_1}\right)^{\frac{p}{2}} \tag{2.36}$$

and if $G_{\epsilon_k} : \mathbb{R}^{d_1 \times d_1} \longrightarrow \mathbb{R}$ is defined such that $G_{\epsilon_k}(\mathbf{Y}) = \frac{2}{p} \operatorname{tr}\left(\mathbf{Y} + \epsilon_k^2 \mathbf{I}_{d_1}\right)^{\frac{p}{2}}$, we will make below rigorous that G_{ϵ_k} is concave for $0 < p \leq 2$ and therefore

$$\begin{aligned} F_{\epsilon_k}(\mathbf{X}) = G_{\epsilon_k}(\mathbf{X}\mathbf{X}^*) &\leq G_{\epsilon_k}(\mathbf{X}^{(k)}\mathbf{X}^{(k)*}) + \langle \nabla G_{\epsilon_k}(\mathbf{X}^{(k)}\mathbf{X}^{(k)*}), \mathbf{X}\mathbf{X}^* - \mathbf{X}^{(k)}\mathbf{X}^{(k)*} \rangle \\ &= F_{\epsilon_k}(\mathbf{X}^{(k)}) + \operatorname{tr}\left(\left(\mathbf{X}^{(k)}\mathbf{X}^{(k)*} + \epsilon_k^2 \mathbf{I}_{d_1}\right)^{\frac{p}{2}-1} \left(\mathbf{X}\mathbf{X}^* - \mathbf{X}^{(k)}\mathbf{X}^{(k)*}\right)\right) =: Q_{\epsilon_k}(\mathbf{X}|\mathbf{X}^{(k)}), \end{aligned} \tag{2.37}$$

as the gradient ∇G_{ϵ_k} of G_{ϵ_k} at $\mathbf{X}^{(k)}\mathbf{X}^{(k)*}$ fulfills $\nabla G_{\epsilon_k}(\mathbf{X}^{(k)}\mathbf{X}^{(k)*}) = \left(\mathbf{X}^{(k)}\mathbf{X}^{(k)*} + \epsilon_k^2 \mathbf{I}_{d_1}\right)^{\frac{p}{2}-1}$. Comparing this to (2.17) and (2.34) quadratic bound $Q_{\epsilon_k}(\cdot|\mathbf{X}^{(k)})$, we see that we have basically derived `IRLS-col` with this estimate.

However, if we just slightly change F_{ϵ_k} such that $F_{\epsilon_k}(\mathbf{X}) = \frac{2}{p} \operatorname{tr}\left(\mathbf{X}\mathbf{X}^* + \epsilon_k^2 \mathbf{I}_{d_2}\right)^{\frac{p}{2}}$, and set G_{ϵ_k} as $G_{\epsilon_k}(\mathbf{Y}) = \frac{2}{p} \operatorname{tr}\left(\mathbf{Y} + \epsilon_k^2 \mathbf{I}_{d_2}\right)^{\frac{p}{2}}$, we obtain

$$\begin{aligned} F_{\epsilon_k}(\mathbf{X}) = G_{\epsilon_k}(\mathbf{X}^*\mathbf{X}) &\leq G_{\epsilon_k}(\mathbf{X}^{(k)*}\mathbf{X}^{(k)}) + \langle \nabla G_{\epsilon_k}(\mathbf{X}^{(k)*}\mathbf{X}^{(k)}), \mathbf{X}^*\mathbf{X} - \mathbf{X}^{(k)*}\mathbf{X}^{(k)} \rangle \\ &= F_{\epsilon_k}(\mathbf{X}^{(k)}) + \operatorname{tr}\left(\left(\mathbf{X}^{(k)*}\mathbf{X}^{(k)} + \epsilon_k^2 \mathbf{I}_{d_2}\right)^{\frac{p}{2}-1} \left(\mathbf{X}^*\mathbf{X} - \mathbf{X}^{(k)*}\mathbf{X}^{(k)}\right)\right) =: Q_{\epsilon_k}(\mathbf{X}|\mathbf{X}^{(k)}), \end{aligned} \tag{2.38}$$

which corresponds to `IRLS-row` due to (2.31).

The fact that this standard derivation of IRLS algorithms does not lead to a unique quadratic upper bound of F_ϵ in the rank surrogate case suggests that it is worthwhile to look for a deeper understanding of the problem structure.

In [KS18], we analyze an IRLS algorithm called *Harmonic Mean Iteratively Reweighted Least Squares* (`HM-IRLS`), which comes to its name as it corresponds to choosing the quadratic bounds $Q_{\epsilon_k}(\cdot|\mathbf{X}^{(k)})$ and the associated weight operator $W^{(k)}$ as the *harmonic mean* of the corresponding ones used in `IRLS-col` and `IRLS-row` (or, to put differently, the harmonic mean of the right hand sides of (2.37) and (2.38)). More precisely, instead of (2.31) or (2.34), the harmonic mean weight operator $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ is defined such that for any $\mathbf{X}^{d_1 \times d_2}$,

$$W^{(k)}(\mathbf{X}) = \mathbf{U}_k \left[\mathbf{H}^{(k)} \circ \left(\mathbf{U}_k^* \mathbf{X} \mathbf{V}_k\right) \right] \mathbf{V}_k^*, \tag{2.39}$$

where $\mathbf{U}_k \in \mathbb{R}^{d_1 \times d_1}$ and $\mathbf{V}_k \in \mathbb{R}^{d_2 \times d_2}$ are again left and right singular vector matrices of $\mathbf{X}^{(k)}$, $\mathbf{A} \circ \mathbf{B}$ denotes the entrywise (Hadamard) product of the matrices $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$ and $\mathbf{B} \in \mathbb{R}^{d_1 \times d_2}$ and $\mathbf{H}^{(k)}$ is defined as

$$(\mathbf{H}^{(k)})_{ij} = \frac{2}{(\widetilde{\sigma_i}^{(k)})^{-1} + (\widetilde{\sigma_j}^{(k)})^{-1}}$$

for any $i \in [d_1], j \in [d_2]$ with $\widetilde{\sigma_i}^{(k)}$ as in (2.35).

The analysis of [KS18] shows that this choice of $W^{(k)}$ leads to considerably *tighter* quadratic bounds than `IRLS-col` or `IRLS-row`, achieving fast local convergence properties of order $2-p$, comparable to what has been obtained for the non-convex variants of [DDFG10] for sparse recovery. On the other hand, it was shown that simpler weight operator choices as the *arithmetic mean* of the weight operators of `IRLS-col` and `IRLS-row` do not possess these favorable properties.

But is there a way to justify the success of the harmonic mean weight operator (2.39) in a more constructive way? Are there maybe weight operators corresponding to even *tighter* quadratic bounds than the harmonic mean ones?

To understand this, the *derivative* structure of smoothed spectral functions F_ϵ will be crucial. As a preparation, we collect a few results on first and second (generalized) derivatives of spectral functions.

Definition 2.3.1 (Spectral functions over $\mathbb{R}^{d_1 \times d_2}$, [Fri81, Lew95, Bec17, LS05a, LS05b])**.** *A function $F : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}$ is called a* spectral function *over $\mathbb{R}^{d_1 \times d_2}$ if there exists a function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ for which $F = f \circ \sigma$, where*

$$\sigma : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^d, \mathbf{X} \mapsto \sigma(\mathbf{X}) = (\sigma_1(\mathbf{X}), \ldots, \sigma_d(\mathbf{X}))$$

is the function mapping matrices in $\mathbb{R}^{d_1 \times d_2}$ to its singular value vector $\sigma(\mathbf{X})$. In this case, we call f the associated function *of F.*

An important property of spectral functions F is that they are, the singular values $\sigma(\mathbf{X})$, *unitarily invariant* in the sense that for any $\mathbf{U} \in \mathbb{O}^{d_1} = \{\mathbf{U} \in \mathbb{R}^{d_1 \times d_1} : \mathbf{U}\mathbf{U}^* = \mathbf{U}^*\mathbf{U} = \mathbf{I}_{d_1}\}$, $\mathbf{V} \in \mathbb{O}^{d_2} = \{\mathbf{V} \in \mathbb{R}^{d_2 \times d_2} : \mathbf{V}\mathbf{V}^* = \mathbf{V}^*\mathbf{V} = \mathbf{I}_{d_2}\}$ and $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$,

$$F(\mathbf{U}^*\mathbf{X}\mathbf{V}) = F(\mathbf{X}).$$

Remark 2.3.1. *We note that in the matrix calculus literature, the term* spectral function *more commonly refers to functions that are defined on the space of symmetric matrices $\mathbb{S}^d = \{\mathbf{X} \in \mathbb{R}^{d \times d} : \mathbf{X}^* = \mathbf{X}\}$ as a concatenation of the function $\lambda : \mathbb{S}^d \longrightarrow \mathbb{R}^d, X \longrightarrow \lambda(\mathbf{X}) = (\lambda_1(\mathbf{X}), \ldots, \lambda_d(\mathbf{X}))$ mapping matrices to their* eigenvalues *with a real-valued function f* [Lew96, LS01, LS02, MS18]*, while functions of* singular values *as in Definition 2.3.1 are sometimes called* singular value functions [LS05a, LS05b]*. We follow the more general Definition 2.3.1, used also in* [Bec17, Chapter 7]*, throughout this thesis.*

Definition 2.3.2 (Absolutely permutation symmetric functions)**.**

1. *Let $\mathbf{x} \in \mathbb{R}^d$. We call $r(\mathbf{x}) \in \mathbb{R}^d$ the* non-increasing rearrangement *of $\mathbf{x}$ if it holds that*

$$r(\mathbf{x})_1 \geq r(\mathbf{x})_2 \geq \ldots \geq r(\mathbf{x})_d$$

and there is a permutation matrix $\mathbf{P} \in \mathbb{P}^d$ such that $r(\mathbf{x})_i = (\mathbf{P}\mathbf{x})_i$ for all $i \in [d]$.

2. *We call a matrix $\mathbf{P}$* signed permutation matrix *if there exists a permutation matrix $\mathbf{P}_0 \in \mathbb{P}^d$ and a diagonal matrix $\mathbf{D} \in \mathbb{R}^{d \times d}$ with only values of -1 and 1 on the diagonal such that $\mathbf{P} = \mathbf{P}_0\mathbf{D}$. We denote the corresponding set by $\pm\mathbb{P}^d$.*

3. *We say that a function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is* absolutely permutation symmetric *if*

$$f(\mathbf{x}) = f(r(|\mathbf{x}|)) \tag{2.40}$$

for any $x \in \mathbb{R}^d$.

We note that (2.40) is equivalent to

$$f(\mathbf{x}) = f(\mathbf{P}\mathbf{x})$$

for all signed permutation matrices $\mathbf{P} \in \pm\mathbb{P}^d$.

In the following, whenever it is clear that $d = \min(d_1, d_2)$ refers to the minimum of two dimension parameters d_1 and d_2, for $v \in \mathbb{R}^d$, we use the notation $\mathrm{dg}(v) \in \mathbb{R}^{d_1 \times d_2}$ for the rectangular diagonal matrix such that for any $i \in [d_1]$, $j \in [d_2]$,

$$\mathrm{dg}(v)_{ij} = \begin{cases} v_i, & \text{if } i = j, \\ 0, & \text{else.} \end{cases}$$

We proceed with the following results due to Lewis and Sendov [LS05a] about first derivatives and convexity of spectral functions.

Proposition 2.3.1. *Let* $F : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ *be a spectral function* $F = f \circ \sigma$ *with an associated function* $f : \mathbb{R}^d \to \mathbb{R}$ *that is absolutely permutation symmetric. Then,* F *is differentiable at* $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ *if and only if* f *is differentiable at* $\sigma(\mathbf{X}) \in \mathbb{R}^d$.

In this case, the gradient ∇F *of* F *at* $\mathbf{X}$ *is given by*

$$\nabla F(\mathbf{X}) = \mathbf{U}\,\mathrm{dg}\left(\nabla f(\sigma(\mathbf{X})\right)\mathbf{V}^*$$

if $\mathbf{X} = \mathbf{U}\,\mathrm{dg}\left(\sigma(\mathbf{X})\right)\mathbf{V}^*$ *for unitary matrices* $\mathbf{U} \in \mathbb{O}^{d_1}$ *and* $\mathbf{V} \in \mathbb{O}^{d_2}$.

Proof. The statement about differentiability follows directly from [LS05a, Corollary 7.4]. The expression for the gradient $\nabla F_\epsilon(X)$ follows by specifying the formula of [LS05a, Theorem 7.1] to the case of singleton subgradients. □

Proposition 2.3.2 ([LS05a, Proposition 6.1])**.** *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be convex and absolutely permutation symmetric. Then the spectral function* $f \circ \sigma$ *is convex on* $\mathbb{R}^{d_1 \times d_2}$ *if and only if* f *is convex.*

Proposition 2.3.1 and Proposition 2.3.2 can be used to make the calculations of (2.37) and (2.38) rigorous, as, first, the concavity of the functions G_{ϵ_k} follows from the concavity of the function $g_{\epsilon_k} : \mathbb{R}^d \to \mathbb{R}$, $g_{\epsilon_k}(\mathbf{x}) = \frac{2}{p}\sum_{i=1}^{d_1}(|y_i| + \epsilon_k^2)^{p/2}$ in $\mathbf{x}$ and Proposition 2.3.2, and, second, if $\mathbf{X}^{(k)} = \mathbf{U}_k\,\mathrm{dg}(\sigma(\mathbf{X}^{(k)}))\mathbf{V}_k$ with unitary $\mathbf{U}_k \in \mathbb{O}^{d_1}$ and $\mathbf{V}_k \in \mathbb{O}^{d_2}$, then $Y_k = \mathbf{X}^{(k)}\mathbf{X}^{(k)*} = \mathbf{U}_k\,\mathrm{dg}(\sigma^2(\mathbf{X}^{(k)}))\mathbf{U}_k^*$, and therefore

$$\nabla G_{\epsilon_k}(\mathbf{X}^{(k)}\mathbf{X}^{(k)*}) = \nabla G_{\epsilon_k}(Y^{(k)}) = \mathbf{U}_k\,\mathrm{dg}\left(((\sigma_i^2(\mathbf{X}^{(k)}) + \epsilon_k^2)^{p/2-1})_{i=1}^{d_1}\right)\mathbf{U}_k^* = \left(\mathbf{X}^{(k)}\mathbf{X}^{(k)*} + \epsilon_k^2\mathbf{I}_{d_1}\right)^{\frac{p}{2}-1},$$

which justifies (2.37) and `IRLS-col`. An analogue argument can be made to justify (2.38) and `IRLS-row`.

Related to the presented derivation of IRLS algorithms based on a change of variables are certain *recipes* that can be found in the literature as a guideline to derive IRLS algorithms [ODBP15, Algorithm 6], [RZ15, (2.2), (2.3) and Proposition 2.6].

In fact, these recipes are based on property (2.18), which required that the spectral function F_{ϵ_k} and weight operator $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ fulfill

$$\nabla F_{\epsilon_k}(\mathbf{X}^{(k)}) = W^{(k)}(\mathbf{X}^{(k)}).$$

This property is needed to ensure that the quadratic bound $Q_{\epsilon_k}(\mathbf{X}|\mathbf{X}^{(k)})$ of (2.17) is *quadratic* in $\mathbf{X}$ without a term that is *linear* in $\mathbf{X}$. The purely quadratic structure of $Q_{\epsilon_k}(\cdot|\mathbf{X}^{(k)})$ can be justified by the desire to obtain quadratic bounds with a similar geometry as the function F_{ϵ_k} to be majorized, which also attains its global minimum at the origin (if no constraints are present) in all cases we are interested in.

Indeed, recalling the spectral function F_{ϵ_k} with

$$F_{\epsilon_k}(\mathbf{X}) = \frac{2}{p} \operatorname{tr}\left(\mathbf{X}\mathbf{X}^* + \epsilon_k^2 \mathbf{I}_{d_1}\right)^{\frac{p}{2}} = \frac{2}{p} \sum_{i=1}^{d_1} \left(\sigma_i^2(\mathbf{X}) + \epsilon_k^2\right)^{\frac{p}{2}}$$

used for `IRLS-col` above, we use Proposition 2.3.1 to see that

$$\nabla F_{\epsilon_k}(\mathbf{X}) = \mathbf{U} \operatorname{dg}\left(\frac{\sigma_i(\mathbf{X})}{(\sigma_i^2(\mathbf{X}) + \epsilon_k^2)^{1-\frac{p}{2}}}\right)_{i=1}^{d} \mathbf{V}^*$$

if $\mathbf{U} \in \mathbb{O}^{d_1}$ and $\mathbf{V} \in \mathbb{O}^{d_2}$ are unitary such that $\mathbf{X} = \mathbf{U} \operatorname{dg}(\sigma(\mathbf{X}))\mathbf{V}^*$. This verifies that (2.18) holds, since furthermore

$$\begin{aligned} W^{(k)}(\mathbf{X}^{(k)}) &= \mathbf{W}_{\text{col}}^{(k)}\mathbf{X}^{(k)} = (\mathbf{X}^{(k)}\mathbf{X}^{(k)*} + \epsilon_k^2 \mathbf{I}_{d_1})^{\frac{p}{2}-1}\mathbf{X}^{(k)} \\ &= \mathbf{U}_k \operatorname{diag}((\sigma_i^2(\mathbf{X}^{(k)}) + \epsilon_k^2)^{\frac{p}{2}-1})_{i=1}^{d} \mathbf{U}_k^* \mathbf{U}_k \operatorname{dg}(\sigma(\mathbf{X}^{(k)}))\mathbf{V}_k^* \\ &= \mathbf{U}_k \operatorname{dg}\left(\frac{\sigma_i(\mathbf{X}^{(k)})}{(\sigma_i^2(\mathbf{X}^{(k)}) + \epsilon_k^2)^{1-\frac{p}{2}}}\right)_{i=1}^{d} \mathbf{V}_k^* = \nabla F_{\epsilon_k}(\mathbf{X}^{(k)}). \end{aligned}$$

Analoguously, (2.18) can be easily verified for `IRLS-col`, and also for the variant `HM-IRLS` [KS18].

This illustrates why the constructive rules [ODBP15, Algorithm 6], [RZ15, (2.2), (2.3) and Proposition 2.6], which are adapted to functions F_ϵ that are separable in the coordinates of their argument, are not sufficient to obtain a unique formulation of IRLS for low-rank matrix recovery.

Our approach: Second-order derivation of IRLS

We now would like to take a different point of view on the problem, which involves *(generalized) second derivatives* of spectral functions.

A relationship between IRLS algorithms and *second-order* methods as *Newton's method* was already observed in [Wat77], where Watson argued that IRLS for ℓ_p-regression (2.24) with $1 < p < 2$ [Sch73, MS74] corresponds to taking a step of Newton's method with a step length of $p-1$ times the original Newton step length, see also [Li93]. In [GR97, Section V.C], it was observed that for non-convex sparsity surrogates, the unsmoothed IRLS variant `FOCUSS` correspond to a modified Newton's method where the modification is such that the negative eigenvalues of the *Hessian* of the objective a replaced by the positive ones at an appropriate scaling.

In general, Newton-type algorithms exhibit often quadratic or superlinear convergence rates if they are used in a neighborhood of a desired solution, however, it is not clear whether they actually provide a *descent* in the objective at each iteration.

A step of *Newton's method* (see, e.g., [NW06, Chapter 3]), also called *Newton-Raphson method* [Ber99], to minimize a spectral function F_{ϵ_k} with evaluation point $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ under the linear constraint $\Phi(\mathbf{X}) = \mathbf{Y}$ would correspond to solving

$$\mathbf{X}^{(k+1)} = \underset{\mathbf{X}:\Phi(\mathbf{X})=\mathbf{Y}}{\arg\min} F_{\epsilon_k}(\mathbf{X}^{(k)}) + \langle \nabla F_{\epsilon_k}(\mathbf{X}^{(k)}), \mathbf{X} - \mathbf{X}^{(k)} \rangle + \frac{1}{2}\langle \mathbf{X} - \mathbf{X}^{(k)}, \nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})(\mathbf{X} - \mathbf{X}^{(k)}) \rangle \quad (2.41)$$

if F_{ϵ_k} is a twice continuously differentiable function, where the *Hessian* of F_{ϵ_k} at $\mathbf{X}^{(k)}$, $\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})$, is a linear operator such that $\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)}) : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$.

However, if F_{ϵ_k} is an ϵ_k-smoothed Schatten-p quasi-norm with $p < 1$ as in (2.36), we expect the Hessian $\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})$ to be *indefinite* in general, since far from the zero matrix $0 \in \mathbb{R}^{d_1 \times d_2}$, the ϵ_k-smoothing will not play any role and F_{ϵ_k} will be locally concave, whereas close to the zero matrix $0 \in \mathbb{R}^{d_1 \times d_2}$, (2.36) is basically quadratic, and therefore convex.

Thus, we can say that the quadratic model of F_{ϵ_k} used by (2.41) will not fulfill a majorization property as (2.19), not even locally around $\mathbf{X}^{(k)}$. While there might be *line search* techniques or other step-size rules that can be applied to ensure a descent in the objective, this is not the avenue we want to pursue, also since evaluations of spectral functions as F_{ϵ_k} will be computationally expensive.

In [BL88], a general framework for quadratic approximation was presented which ensures the majorization property (2.19) at each iteration for a scheme that, translated to our matrix function setting, replaces the Hessian $\nabla F_{\epsilon_k}(\mathbf{X}^{(k)})$ at each iteration by a matrix operator $B : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ that is a global upper bound such that

$$B \succeq \nabla^2 F_{\epsilon_k}(\mathbf{X})$$

for any $X \in \mathbb{R}^{d_1 \times d_2}$. Here, as usual, we use the symbol $\succeq$ in the sense of *Loewner order* such that two matrix operators $A, B : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ fulfill $A \succeq B$ if and only if

$$\langle \mathbf{Z}, A(\mathbf{Z}) \rangle \geq \langle \mathbf{Z}, B(\mathbf{Z}) \rangle$$

for all matrices $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$.

The framework of [BL88] does not quite correspond to what our intended IRLS scheme is supposed to do, as the weight operators $W^{(k)}$ of (2.17) depend on the matrix $\mathbf{X}^{(k)}$, but it gives us the intuition that

$$W^{(k)} \succeq \nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)}) \quad (2.42)$$

will be at least a *necessary* condition we will need to fulfill for any $X^{(k)}$ to have any chance to ensure the majorization property (2.19).

To ensure a property as (2.42), it is crucial to understand *second derivatives* of spectral functions.

Second derivatives of spectral functions have been considered in the literature almost exclusively for the case of functions depending on the *eigenvalues* of a symmetric matrix [LS01, LS02, QY03, DK15], but, to the best of the author's knowledge, not of spectral functions as defined in Definition 2.3.1.

However, it is possible to derive results for second derivatives of functions depending on the singular values of a matrix by combining Proposition 2.3.1 with results on the derivatives of certain *spectral operators* [DSST18], so-called non-Hermitian Loewner operators [Yan09, Theorem 2.2.6], [Nof17, Corollary 3.10].

As a preparation for the next result, we define the *symmetrization operator* $S : \mathbb{R}^{d \times d} \longrightarrow$

$\mathbb{R}^{d\times d}$ that maps any $\mathbf{Z} \in \mathbb{R}^{d\times d}$ such that

$$S(\mathbf{Z}) = \frac{1}{2}(\mathbf{Z} + \mathbf{Z}^*), \tag{2.43}$$

and the *antisymmetrization operator* $T : \mathbb{R}^{d\times d} \longrightarrow \mathbb{R}^{d\times d}$ such that

$$T(\mathbf{Z}) = \frac{1}{2}(\mathbf{Z} - \mathbf{Z}^*) \tag{2.44}$$

for any $\mathbf{Z} \in \mathbb{R}^{d\times d}$.

Theorem 2.2. *Let $f : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}$ be differentiable function with L-Lipschitz first derivative f' such that f' is right differentiable at 0 and $f'(0) = 0$.*

Then the spectral function $F : \mathbb{R}^{d_1\times d_2} \longrightarrow \mathbb{R}$, $F(\mathbf{X}) = \sum_{i=1}^{d} f(\sigma_i(\mathbf{X})) = \sum_{i=1}^{d} f(\sigma_i)$ is differentiable with L-Lipschitz gradients ∇F and furthermore almost everywhere twice differentiable.

In particular, F is twice differentiable at $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$ if and only if f is twice differentiable at all $\sigma_1, \ldots, \sigma_d$. In that case, if additionally $d_1 \leq d_2$, the Hessian $\nabla^2 F(\mathbf{X})$ of F at $\mathbf{X}$ is given by

$$\nabla^2 F(\mathbf{X})(\mathbf{Z}) = \mathbf{U}\begin{bmatrix}\mathbf{H}_1 \circ S(\mathbf{U}^*\mathbf{Z}\mathbf{V}_1) + \mathbf{H}_2 \circ T(\mathbf{U}^*\mathbf{Z}\mathbf{V}_1) & \mathbf{H}_3 \circ (\mathbf{U}^*\mathbf{Z}\mathbf{V}_2)\end{bmatrix}\mathbf{V}^* \tag{2.45}$$

for any $\mathbf{Z} \in \mathbb{R}^{d_1\times d_2}$. In the notation of (2.45), $\mathbf{X}$ has the singular value decomposition $\mathbf{X} = \mathbf{U}\,\mathrm{dg}(\sigma)\mathbf{V}^*$ with $\mathbf{U} \in \mathbb{O}^{d_1}$, $\mathbf{V} = \begin{bmatrix}\mathbf{V}_1 & \mathbf{V}_2\end{bmatrix} \in \mathbb{O}^{d_2}$, $\mathbf{V}_1 \in \mathbb{R}^{d_2\times d}$, $\mathbf{V}_2 \in \mathbb{R}^{d_2\times(d_2-d)}$, T and S are as in (2.43) and (2.44), and $\mathbf{H}_1 \in \mathbb{R}^{d\times d}$ is such that for $i, j \in [d]$,

$$(\mathbf{H}_1)_{ij} = \begin{cases}\frac{f'(\sigma_i)-f'(\sigma_j)}{\sigma_i-\sigma_j} & \textit{if } \sigma_i \neq \sigma_j,\\ f''(\sigma_i) & \textit{if } \sigma_i = \sigma_j,\end{cases}$$

the matrix $\mathbf{H}_2 \in \mathbb{R}^{d\times d}$ is such that for $i, j \in [d]$,

$$(\mathbf{H}_2)_{ij} = \begin{cases}\frac{f'(\sigma_i)+f'(\sigma_j)}{\sigma_i+\sigma_j} & \textit{if } \sigma_i + \sigma_j \neq 0,\\ f''(\sigma_i) & \textit{if } \sigma_i = \sigma_j = 0,\end{cases}$$

and $\mathbf{H}_3 \in \mathbb{R}^{d_1\times(d_2-d)}$ is such that for $i \in [d_1], j \in [d_2 - d]$,

$$(\mathbf{H}_3)_{ij} = \begin{cases}f'(\sigma_i)/\sigma_i & \textit{if } \sigma_i \neq 0,\\ f''(\sigma_i) & \textit{if } \sigma_i = 0.\end{cases}$$

Proof. Since f is differentiable, it follows from Proposition 2.3.1 that the spectral function F is differentiable with gradient

$$\nabla F(\mathbf{X}) = \mathbf{U}\,\mathrm{dg}\left(\nabla f(\sigma(\mathbf{X})\right)\mathbf{V}^* = \mathbf{U}\,\mathrm{dg}\left(\left(f'(\sigma_i)\right)_{i=1}^{d}\right)\mathbf{V}^*$$

at any $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$. Functions as $\mathbf{X} \mapsto \nabla F(\mathbf{X})$ are also called *non-Hermitian Löwner operators* [Löw34, SS08, DSST18] or *generalized matrix functions* [HBI73, Nof17], and [ACP16, Theorem 1.1] implies that since f' is L-Lipschitz, $\mathbf{X} \mapsto \nabla F(\mathbf{X})$ is L-Lipschitz with respect to the Frobenius norm. Rademacher's theorem then implies that ∇F is almost everywhere differentiable (with respect to the Lebesgue measure).

Furthermore, it follows from [Yan09, Theorem 2.2.6] that f is twice differentiable at $\sigma = \sigma(\mathbf{X})$ if and only if F is twice differentiable at $\mathbf{X}$, and the formula for the Hessian (2.45) at the points of twice differntiablity $\mathbf{X}$ is due to [Yan09, Theorem 2.2.6] and [Nof17, Corollary 3.10]. □

In the remainder of this chapter, we focus on a specific family of spectral functions which are *smoothings* of (scaled) Schatten-p quasi-norms and logdet objectives and generalizations of (2.27) [FRW11]. This family is given by $\{F_{p,\epsilon} : 0 \leq p \leq 1, \epsilon > 0\}$ in Lemma 2.3.1.

Lemma 2.3.1. *Let $0 \leq p \leq 1$ and $\epsilon > 0$. Define the function $f_{p,\epsilon} : \mathbb{R}_{\geq 0} \to \mathbb{R}$ such that*

$$f_{p,\epsilon}(\sigma) = \begin{cases} \frac{1}{p}\sigma^p, & \text{if } \sigma > \epsilon, \\ \frac{1}{2}\frac{\sigma^2}{\epsilon^{2-p}} + \left(\frac{1}{p} - \frac{1}{2}\right)\epsilon^p, & \text{if } 0 \leq \sigma \leq \epsilon \end{cases} \tag{2.46}$$

for $0 < p \leq 1$, and

$$f_{p,\epsilon}(\sigma) = \begin{cases} \log(\sigma), & \text{if } \sigma > \epsilon, \\ \frac{1}{2}\frac{\sigma^2}{\epsilon^2} + \frac{1}{2}\left(2\log(\epsilon) - 1\right), & \text{if } 0 \leq \sigma \leq \epsilon \end{cases} \tag{2.47}$$

for $p = 0$.

Then the spectral function $F_{p,\epsilon} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ defined by

$$F_{p,\epsilon}(\mathbf{X}) = \sum_{i=1}^{d} f_{p,\epsilon}(\sigma_i(\mathbf{X})) = \begin{cases} \sum_{i=1}^d \left[\log\left(\max(\sigma_i(\mathbf{X}), \epsilon)\right) + \frac{1}{2}\left(\frac{\min(\sigma_i(\mathbf{X}),\epsilon)^2}{\epsilon^2} - 1\right)\right], & \text{if } p = 0, \\ \sum_{i=1}^d \left[\frac{1}{p}\max(\sigma_i(\mathbf{X}), \epsilon)^p + \frac{1}{2}\left(\frac{\min(\sigma_i(\mathbf{X}),\epsilon)^2}{\epsilon^{2-p}} - \epsilon^p\right)\right], & \text{if } 0 < p \leq 1, \end{cases} \tag{2.48}$$

is differentiable[3] *with ϵ^{p-2}-Lipschitz gradient $\nabla F_{p,\epsilon}$, which is given by*

$$\nabla F_{p,\epsilon}(\mathbf{X}) = \mathbf{U}\,\mathrm{dg}\left(\left(\frac{\sigma_i(\mathbf{X})}{\max(\sigma_i(\mathbf{X}), \epsilon)^{2-p}}\right)_{i=1}^{d}\right)\mathbf{V}^*$$

for any $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, where $\mathbf{X} = \mathbf{U}\,\mathrm{dg}\left(\sigma(\mathbf{X})\right)\mathbf{V}^* = \mathbf{U}\,\mathrm{dg}\left(\sigma\right)\mathbf{V}^*$ is a singular value decomposition with unitary matrices $\mathbf{U} \in \mathbb{O}^{d_1}$ and $\mathbf{V} \in \mathbb{O}^{d_2}$.

Furthermore, $F_{p,\epsilon}$ is twice differentiable for all $\mathbf{X} \in \mathcal{D}_\epsilon := \left\{\mathbf{X} : \sigma_i(\mathbf{X}) \neq \epsilon \text{ for all } i \in [d]\right\}$ and, defining $R := |\{i \in [d] : \sigma_i(\mathbf{X}) > \epsilon\}|$ and if additionally $d_1 \leq d_2$, its Hessian $\nabla^2 F_{p,\epsilon}(\mathbf{X}) : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ at $\mathbf{X} \in \mathcal{D}_\epsilon$ is given by

$$\nabla^2 F_{p,\epsilon}(\mathbf{X})(\mathbf{Z}) = \mathbf{U}\left[\mathbf{H}_1 \circ S(\mathbf{U}^*\mathbf{Z}\mathbf{V}_1) + \mathbf{H}_2 \circ T(\mathbf{U}^*\mathbf{Z}\mathbf{V}_1) \quad \mathbf{H}_3 \circ (\mathbf{U}^*\mathbf{Z}\mathbf{V}_2)\right]\mathbf{V}^*, \tag{2.49}$$

where T and S are as in (2.43) and (2.44), $\begin{bmatrix}\mathbf{V}_1 & \mathbf{V}_2\end{bmatrix} = \mathbf{V}$, $\mathbf{V}_1 \in \mathbb{R}^{d_2 \times d}$, $\mathbf{V}_2 \in \mathbb{R}^{d_2 \times (d_2 - d)}$, and $\mathbf{H}_1 \in \mathbb{R}^{d_1 \times d_2}$ is such that

$$(\mathbf{H}_1)_{ij} = \begin{cases} (p-1)\sigma_i^{p-2} & \text{if } i, j \leq R \text{ and } \sigma_i = \sigma_j, \\ \frac{\sigma_i^{p-1} - \sigma_j^{p-1}}{\sigma_i - \sigma_j} = \frac{\sigma_j^{1-p} - \sigma_i^{1-p}}{\sigma_i^{1-p}\sigma_j^{1-p}(\sigma_i - \sigma_j)} & \text{if } i, j \leq R \text{ and } \sigma_i \neq \sigma_j, \\ \frac{\sigma_i^{p-1} - \frac{\sigma_j}{\epsilon^{2-p}}}{\sigma_i - \sigma_j} & \text{if } i \leq R, R < j, \\ \frac{\sigma_j^{p-1} - \frac{\sigma_i}{\epsilon^{2-p}}}{\sigma_j - \sigma_i} & \text{if } j \leq R, R < i, \\ \epsilon^{p-2} & \text{if } R < i, j, \end{cases} \tag{2.50}$$

[3] Whenever there is no ambiguity about p, we will call $f_{p,\epsilon}$ and $F_{p,\epsilon}$ also f_ϵ and F_ϵ, respectively.

the matrix $\mathbf{H}_2 \in \mathbb{R}^{d\times d}$ *is such that for* $i, j \in [d]$,

$$(\mathbf{H}_2)_{ij} = \begin{cases} \sigma_i^{p-2} & \textit{if } i,j \leq R \textit{ and } \sigma_i = \sigma_j, \\ \frac{\sigma_i^{p-1}+\sigma_j^{p-1}}{\sigma_i+\sigma_j} = \frac{\sigma_j^{1-p}+\sigma_i^{1-p}}{\sigma_i^{1-p}\sigma_j^{1-p}(\sigma_i+\sigma_j)} & \textit{if } i,j \leq R \textit{ and } \sigma_i \neq \sigma_j, \\ \frac{\sigma_i^{p-1}+\frac{\sigma_j}{\epsilon^{2-p}}}{\sigma_i+\sigma_j} & \textit{if } i \leq R, R < j, \\ \frac{\sigma_j^{p-1}+\frac{\sigma_i}{\epsilon^{2-p}}}{\sigma_j+\sigma_i} & \textit{if } j \leq R, R < i, \\ \epsilon^{p-2} & \textit{if } R < i, j, \end{cases} \tag{2.51}$$

and $\mathbf{H}_3 \in \mathbb{R}^{d_1\times(d_2-d)}$ *is such that for* $i \in [d_1], j \in [d_2 - d]$,

$$(\mathbf{H}_3)_{ij} = \begin{cases} \sigma_i^{p-2} & \textit{if } i \leq R, \\ \epsilon^{p-2} & \textit{if } R < i. \end{cases}$$

Proof. Considering first the one-dimensional functions $f_{p,\epsilon}$ from (2.46) and (2.47), we note that $f_{p,\epsilon}$ is differentiable on $\mathbb{R}_{>0}$ with derivative

$$f'_{p,\epsilon}(\sigma) = \begin{cases} \sigma^{p-1} = \frac{\sigma}{\sigma^{2-p}}, & \text{if } \sigma > \epsilon, p > 0 \\ \frac{\sigma}{\epsilon^{2-p}}, & \text{if } 0 \leq \sigma < \epsilon, p > 0, \\ \frac{1}{\sigma} = \frac{\sigma}{\sigma^2}, & \text{if } \sigma > \epsilon, p = 0 \\ \frac{\sigma}{\epsilon^2}, & \text{if } 0 \leq \sigma < \epsilon, p = 0 \end{cases} = \frac{\sigma}{\max(\sigma,\epsilon)^{2-p}} \text{ for } \sigma \neq \epsilon.$$

For the case $\sigma = \epsilon$, we note that

$$\lim_{h>0,h\to 0} \frac{f_{p,\epsilon}(\epsilon+h) - f_{p,\epsilon}(\epsilon)}{h} = \lim_{h>0,h\to 0} \frac{\frac{1}{p}(\epsilon+h)^p - \frac{1}{p}\epsilon^p}{h} = \epsilon^{p-1}$$

and

$$\lim_{h<0,h\to 0} \frac{f_{p,\epsilon}(\epsilon+h) - f_{p,\epsilon}(\epsilon)}{h} = \lim_{h>0,h\to 0} \frac{\frac{1}{2}\frac{(\epsilon+h)^2}{\epsilon^{2-p}} + (\frac{1}{p}-\frac{1}{2})\epsilon^p - \frac{1}{p}\epsilon^p}{h} = \epsilon^{p-2} \lim_{h>0,h\to 0} \frac{\frac{1}{2}(\epsilon+h)^2 - \frac{1}{2}\epsilon^2}{h} = \epsilon^{p-1}$$

for $p > 0$, which shows that

$$f'_{p,\epsilon}(\sigma) = \frac{\sigma}{\max(\sigma,\epsilon)^{2-p}} \tag{2.52}$$

for any $\sigma > 0$ and $0 \leq p \leq 1$, as an analogous calculation shows the existence and value of $f'_{p,\epsilon}(\epsilon)$ for $p = 0$.

Furthermore, $f'_{p,\epsilon}$ is Lipschitz continuous with Lipschitz constant $L = \frac{1}{\epsilon^{2-p}}$, as its (generalized) derivative is globally bounded by $\frac{1}{\epsilon^{2-p}}$. $f'_{p,\epsilon}$ is also right differentiable at 0 and $f'_{p,\epsilon}(0) = 0$. In fact, it can be seen that $f'_{p,\epsilon}$ is differentiable almost everywhere except from at $\sigma = \epsilon$, and

$$f''_{p,\epsilon}(\sigma) = \begin{cases} (p-1)\sigma^{p-2}, & \text{if } \sigma > \epsilon, \\ \epsilon^{p-2}, & \text{if } \sigma < \epsilon. \end{cases} \tag{2.53}$$

The statements of Lemma 2.3.1 then follow directly from Proposition 2.3.1 and Theorem 2.2 by inserting the formulas (2.52) and (2.53). □

Based on Lemma 2.3.1, we can now state necessary conditions for *quadratic* matrix functions $Q_\epsilon(\cdot|\mathbf{X}^{(k)})$ to be *local* upper bounds onto the smoothed spectral functions F_ϵ at a given $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1\times d_2}$.

Theorem 2.3 (Necessary condition for quadratic upper bounds for smoothed rank surrogates)**.** *Let $d_1 \leq d_2$ and $X^{(k)} \in \mathscr{D}_\epsilon := \left\{\mathbf{X} \in \mathbb{R}^{d_1\times d_2} : \sigma_i(\mathbf{X}) \neq \epsilon \text{ for all } i \in [d]\right\}$ be a matrix where the spectral function F_ϵ from (2.48) is twice differentiable, let $\mathbf{X}^{(k)} = \mathbf{U}_k \operatorname{dg}(\sigma^{(k)})\mathbf{V}_k^*$ be a singular value decomposition of $\mathbf{X}^{(k)}$ with $\mathbf{U}_k \in \mathbb{O}^{d_1}$, $\mathbf{V}_k = \begin{bmatrix}\mathbf{V}_1^{(k)} & \mathbf{V}_2^{(k)}\end{bmatrix} \in \mathbb{O}^{d_2}$, $\mathbf{V}_1^{(k)} \in \mathbb{R}^{d_2\times d}$, $\mathbf{V}_2^{(k)} \in \mathbb{R}^{d_2\times(d_2-d)}$, and let $R = |\{i \in [d] : \sigma_i^{(k)} > \epsilon\}|$ be the number of singular values of $\mathbf{X}^{(k)}$ larger than ϵ.*

Let $\mathbf{H}_1^{(k)} \in \mathbb{R}^{d_1\times d_1}$, $\mathbf{H}_2^{(k)} \in \mathbb{R}^{d_1\times d_1}$ and $\mathbf{H}_3^{(k)} \in \mathbb{R}^{d_1\times(d_2-d_1)}$ be matrices and

$$(\mathbf{H}_1^{(k)})_{ij} \geq (p-1)(\sigma_i^{(k)})^{p-2} \text{ and } (\mathbf{H}_2^{(k)})_{ij} \geq (\sigma_i^{(k)})^{p-2} \text{ for all } i,j \leq R \text{ with } \sigma_i^{(k)} = \sigma_j^{(k)}, \tag{C1}$$

$$(\mathbf{H}_1^{(k)})_{ij} \geq \frac{(\sigma_i^{(k)})^{p-1}-(\sigma_j^{(k)})^{p-1}}{\sigma_i^{(k)}-\sigma_j^{(k)}} \text{ and } (\mathbf{H}_2^{(k)})_{ij} \geq \frac{(\sigma_i^{(k)})^{p-1}+(\sigma_j^{(k)})^{p-1}}{\sigma_i^{(k)}+\sigma_j^{(k)}} \text{ for all } i,j \leq R \text{ with } \sigma_i^{(k)} \neq \sigma_j^{(k)}, \tag{C2}$$

$$(\mathbf{H}_1^{(k)})_{ij} \geq \frac{(\sigma_i^{(k)})^{p-1}-\frac{\sigma_j^{(k)}}{\epsilon^{2-p}}}{\sigma_i^{(k)}-\sigma_j^{(k)}} \qquad \text{and } (\mathbf{H}_2^{(k)})_{ij} \geq \frac{(\sigma_i^{(k)})^{p-1}+\frac{\sigma_j^{(k)}}{\epsilon^{2-p}}}{\sigma_i^{(k)}+\sigma_j^{(k)}} \text{ for all } i \leq R, R < j, \tag{C3}$$

$$(\mathbf{H}_1^{(k)})_{ij} \geq \frac{(\sigma_j^{(k)})^{p-1}-\frac{\sigma_i^{(k)}}{\epsilon^{2-p}}}{\sigma_j^{(k)}-\sigma_i^{(k)}} \qquad \text{and } (\mathbf{H}_2^{(k)})_{ij} \geq \frac{(\sigma_j^{(k)})^{p-1}+\frac{\sigma_i^{(k)}}{\epsilon^{2-p}}}{\sigma_j^{(k)}+\sigma_i^{(k)}} \text{ for all } j \leq R, R < i, \tag{C4}$$

$$(\mathbf{H}_1^{(k)})_{ij} \geq \epsilon^{p-2} \qquad \text{and } (\mathbf{H}_2^{(k)})_{ij} \geq \epsilon^{p-2} \text{ for all } j \leq R, R < i, \tag{C5}$$

$$(\mathbf{H}_3^{(k)})_{i,j-d_1} \geq (\sigma_i^{(k)})^{p-2} \qquad \text{for all } i \leq R \text{ and } j > d_1, \tag{C6}$$

$$(\mathbf{H}_3^{(k)})_{i,j-d_1} \geq \epsilon^{p-2} \qquad \text{for all } R < i \text{ and } j > d_1. \tag{C7}$$

Then the conditions (C1)–(C7) are necessary *for the quadratic matrix function (see also (2.17)) $Q_\epsilon(\cdot|\mathbf{X}^{(k)} : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ defined by*

$$Q_\epsilon(\mathbf{X}|\mathbf{X}^{(k)}) = F_\epsilon(\mathbf{X}^{(k)}) + \langle \nabla F_\epsilon(\mathbf{X}^{(k)}), \mathbf{X}-\mathbf{X}^{(k)}\rangle + \frac{1}{2}\langle \mathbf{X}-\mathbf{X}^{(k)}, W^{(k)}(\mathbf{X}-\mathbf{X}^{(k)})\rangle$$

with weight operator *$W^{(k)} : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ such that*

$$W^{(k)}(\mathbf{Z}) = \mathbf{U}_k \begin{bmatrix}\mathbf{H}_1^{(k)} \circ S(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_1^{(k)}) + \mathbf{H}_2^{(k)} \circ T(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_1^{(k)}) & \mathbf{H}_3^{(k)} \circ (\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_2^{(k)})\end{bmatrix} \mathbf{V}_k^*,$$

*to be a **local upper bound of F_ϵ in a neighborhood of $\mathbf{X}^{(k)}$**, i.e., to fulfill*

$$Q_\epsilon(\mathbf{X}|\mathbf{X}^{(k)}) \geq F_\epsilon(\mathbf{X})$$

for all $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$ in an open neighborhood of $\mathbf{X}^{(k)}$.

Proof. The result of Theorem 2.3 follows from Lemma 2.3.1 and Taylor's theorem: Indeed, by Lemma 2.3.1, if any of the conditions (C1)–(C7) is not fulfilled, there exists a pair $(i,j) \in [d_1]\times[d_2]$ and a matrix $\mathbf{Z} \in \mathbb{R}^{d_1\times d_2}$ such that $\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k = \frac{1}{\sqrt{2}}(e_ie_j^* + e_je_i^*)$, $\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k = \frac{1}{\sqrt{2}}(e_ie_j^* - e_je_i^*)$ (if $j \leq d_1$) or $\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k = e_ie_j^*$ (if $j > d_1$). In the first case, it follows then that

$$\langle \mathbf{Z}, W^{(k)}(\mathbf{Z})\rangle = \langle \mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k, \mathbf{H}_1^{(k)} \circ (\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\rangle = (\mathbf{H}_1^{(k)})_{ij} < \langle \mathbf{Z}, \nabla^2 F_{p,\epsilon}(\mathbf{X}^{(k)})(\mathbf{Z})\rangle,$$

and the two other cases are analogous.

By Taylor's theorem, since $\nabla^2 F_\epsilon(\mathbf{X}^{(k)})$ is continuous in a neighborhood of $\mathbf{X}^{(k)}$, for any $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ that is close enough to $\mathbf{X}^{(k)}$, there exists a matrix $\Theta_{\mathbf{X}^{(k)},\mathbf{X}}$ such that

$$\langle \mathbf{Z}, W^{(k)}(\mathbf{Z})\rangle < \langle \mathbf{Z}, \nabla^2 F_\epsilon\left(\Theta_{\mathbf{X}^{(k)},\mathbf{X}}\right)(\mathbf{Z})\rangle,$$

and therefore

$$\begin{aligned} F_\epsilon(\mathbf{X}) &= F_\epsilon(\mathbf{X}^{(k)}) + \langle \nabla F_\epsilon(\mathbf{X}^{(k)}), \mathbf{X} - \mathbf{X}^{(k)}\rangle + \frac{1}{2}\langle \mathbf{X} - \mathbf{X}^{(k)}, \nabla^2 F_\epsilon\left(\Theta_{\mathbf{X}^{(k)},\mathbf{X}}\right)(\mathbf{X} - \mathbf{X}^{(k)})\rangle \\ &> F_\epsilon(\mathbf{X}^{(k)}) + \langle \nabla F_\epsilon(\mathbf{X}^{(k)}), \mathbf{X} - \mathbf{X}^{(k)}\rangle + \frac{1}{2}\langle \mathbf{X} - \mathbf{X}^{(k)}, W^{(k)}(\mathbf{X} - \mathbf{X}^{(k)})\rangle = Q_\epsilon(\mathbf{X}|\mathbf{X}^{(k)}), \end{aligned}$$

if $\mathbf{X}$ is chosen such that there exists $\theta > 0$ with $\mathbf{Z} = \theta \cdot (\mathbf{X} - \mathbf{X}^{(k)})$. □

It can be verified that the conditions of Theorem 2.3 are fulfilled for weights chosen as in the papers [FRW11] and [MF12] (adapted to our specific smoothing), or in other words, chosen as `IRLS-col` or `IRLS-row`.

2.3.3 Global Majorization of Rank Surrogate by Tight Quadratic Bounds

The analysis of Section 2.3.2 provides *local tight quadratic bounds* onto smoothed rank surrogates as F_ϵ from (2.48), for example, by replacing the inequalities in conditions (C1)–(C7) by *strict* inequalities.

However, a *global majorization* property of the quadratic function $Q_\epsilon(\cdot|\mathbf{X}^{(k)})$ as (2.19) would be highly desirable for an IRLS scheme as outlined in the beginning of Section 2.3.

In this section, we propose a specific choice of the weight operator $W^{(k)}$ that **combines tight local majorization** with **global majorization** of F_ϵ on the entirety of its domain.

Without loss of generality, we will assume $d_1 \leq d_2$; the definition can be adapted to cover the case of $d_1 > d_2$ by considering transposed matrices.

To give the precise definition, we define, similar to (2.43) and (2.44), the *extended symmetrization operator* $\widetilde{S} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ that maps any $\mathbf{Z} = \begin{bmatrix}\mathbf{Z}_1 & \mathbf{Z}_2\end{bmatrix} \in \mathbb{R}^{d_1 \times d_2}$, $\mathbf{Z}_1 \in \mathbb{R}^{d_1 \times d_1}$, $\mathbf{Z}_2 \in \mathbb{R}^{d_1 \times (d_2 - d_1)}$ such that

$$\widetilde{S}(\mathbf{Z}) = \begin{bmatrix}\frac{1}{2}(\mathbf{Z}_1 + \mathbf{Z}_1^*) & \mathbf{Z}_2\end{bmatrix} \tag{2.54}$$

and the *extended antisymmetrization operator* $\widetilde{T} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ such that

$$\widetilde{T}(\mathbf{Z}) = \begin{bmatrix}\frac{1}{2}(\mathbf{Z}_1 - \mathbf{Z}_1^*) & \mathbf{0}_{d_1 \times (d_2 - d_1)}.\end{bmatrix} \tag{2.55}$$

Furthermore, we use the following definition.

Definition 2.3.3 (Power mean). *Let $-\infty \leq q \leq \infty$. We call*

$$M_q(a,b) = \begin{cases} \min(a,b), & \textit{if } q = -\infty, \\ \left(\frac{a^q + b^q}{2}\right)^{\frac{1}{q}}, & \textit{if } q \in (-\infty, 0) \cup (0, \infty), \\ \sqrt{ab}, & \textit{if } q = 0, \\ \max(a,b) & \textit{if } q = \infty \end{cases} \tag{2.56}$$

*the q-*power mean [Bul03, Section III.1, Definition 1] *of the numbers $a, b > 0$.*

In Definition 2.3.4, we define the weight operator $W^{(k)}$ to be used and analyzed in the rest of this chapter.

Definition 2.3.4 (Weight operator of MatrixIRLS)**.** *Let $d_1 \leq d_2$, let $0 \leq p \leq 1$ and $\epsilon > 0$. Let $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ be a matrix with singular value decomposition $\mathbf{X}^{(k)} = \mathbf{U}_k \operatorname{dg}(\sigma^{(k)})\mathbf{V}_k^*$ be a singular value decomposition of $\mathbf{X}^{(k)}$ with $\mathbf{U}_k \in \mathbb{O}^{d_1}$, $\mathbf{V}_k \in \mathbb{O}^{d_2}$, and $R_k = |\{i \in [d] : \sigma_i^{(k)} > \epsilon\}|$ be the number of singular values of $\mathbf{X}^{(k)}$ larger than ϵ. For $\mathbf{X}^{(k)}$, we define the* weight operator *$W^{(k)} : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$* of MatrixIRLS *such that*

$$W^{(k)}(\mathbf{Z}) = \mathbf{U}_k \left[\mathbf{H}_1^{(k)} \circ \widetilde{S}(\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k) + \mathbf{H}_2^{(k)} \circ \widetilde{T}(\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k) \right] \mathbf{V}_k^*, \tag{2.57}$$

where $\widetilde{S}$ and $\widetilde{T}$ are as in (2.54) and (2.55) and $\mathbf{H}_1^{(k)}, \mathbf{H}_2^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ are matrices with positive entries such that

$$(\mathbf{H}_1^{(k)})_{ij} = M_{\frac{p}{p-2}}(\widetilde{\sigma}_i^{(k)}, \widetilde{\sigma}_i^{(k)}) = \begin{cases} \frac{1}{\max(\sigma_i^{(k)},\epsilon)\max(\sigma_j^{(k)},\epsilon)}, & \text{if } p = 0, \\ \frac{\left(\max(\sigma_i^{(k)},\epsilon)^p + \max(\sigma_j^{(k)},\epsilon)^p\right)^{1-2/p}}{2^{1-2/p}}, & \text{if } 0 < p \leq 1, \end{cases} \tag{2.58}$$

where $M_{\frac{p}{p-2}}$ is the power mean (2.56) of power $q = \frac{p}{p-2}$, $\widetilde{\sigma}_i^{(k)} = \max(\sigma_j^{(k)}, \epsilon)^{p-2}$ for all i and

$$(\mathbf{H}_2^{(k)})_{ij} = \frac{\left(\max(\sigma_i^{(k)}, \epsilon)\right)^{p-1} + \left(\max(\sigma_j^{(k)}, \epsilon)\right)^{p-1}}{\max(\sigma_i^{(k)}, \epsilon) + \max(\sigma_j^{(k)}, \epsilon)} = \begin{cases} \frac{(\sigma_i^{(k)})^{p-1} + (\sigma_j^{(k)})^{p-1}}{\sigma_i^{(k)} + \sigma_j^{(k)}} & \text{if } i, j \leq R_k, \\ \frac{(\sigma_i^{(k)})^{p-1} + \epsilon^{p-1}}{\sigma_i^{(k)} + \epsilon} & \text{if } i \leq R_k, R_k < j, \\ \frac{(\sigma_j^{(k)})^{p-1} + \epsilon^{p-1}}{\sigma_j^{(k)} + \epsilon} & \text{if } j \leq R_k, R_k < i, \\ \epsilon^{p-2} & \text{if } R_k < i, j. \end{cases} \tag{2.59}$$

Remark 2.3.2. *In the definition of the weights on the symmetric part $\mathbf{H}_1^{(k)}$, we note that their entries correspond to the*

- geometric mean *of $\widetilde{\sigma}_i^{(k)}$ and $\widetilde{\sigma}_j^{(k)}$ for $p = 0$, and to the*
- harmonic mean *for $p = 1$, while*
- *interpolating between these to means by the $\frac{p}{p-2}$-power mean for $0 < p < 1$.*

Furthermore, for the case of $p = 0$, we see that $\mathbf{H}_1^{(k)}$ and $\mathbf{H}_2^{(k)}$ actually coincide as then

$$\begin{aligned} (\mathbf{H}_2^{(k)})_{ij} &= \frac{\left(\sigma_i^{(k)} \vee \epsilon\right)^{-1} + \left(\sigma_j^{(k)} \vee \epsilon\right)^{-1}}{\sigma_i^{(k)} \vee \epsilon + \sigma_j^{(k)} \vee \epsilon} = \frac{\sigma_i^{(k)} \vee \epsilon + \sigma_j^{(k)} \vee \epsilon}{(\sigma_i^{(k)} \vee \epsilon)(\sigma_j^{(k)} \vee \epsilon)\left(\sigma_i^{(k)} \vee \epsilon + \sigma_j^{(k)} \vee \epsilon\right)} \\ &= \frac{1}{(\sigma_i^{(k)} \vee \epsilon)(\sigma_j^{(k)} \vee \epsilon)} = (\mathbf{H}_1^{(k)})_{ij}. \end{aligned}$$

This means that in the case of $p = 0$, which corresponds to the weight operator of a logdet-objective (2.48), $W^{(k)}$ of Definition 2.3.4 can be written as mapping any $Z \in \mathbb{R}^{d_1 \times d_2}$ to

$$W^{(k)}(\mathbf{Z}) = \mathbf{U}_k \left[\mathbf{H}^{(k)} \circ (\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k) \right] \mathbf{V}_k^*$$

with $\mathbf{H}^{(k)} = \mathbf{H}_1^{(k)} = \mathbf{H}_2^{(k)}$.

We note as it is the case for the Hessian $\nabla^2 F_\epsilon(\mathbf{X}^{(k)})$ from (2.49) in section 2.3.2, $W^{(k)}$ from Definition 2.3.4 is a self-adjoint operator: Indeed, by the cyclicity of the trace, for any $\mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$,

$$\begin{aligned}
\langle \mathbf{Y}, W^{(k)}(\mathbf{Z})\rangle &= \operatorname{tr}\left(\mathbf{Y}^*\mathbf{U}_k\left[\mathbf{H}_1^{(k)} \circ \widetilde{S}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k) + \mathbf{H}_2^{(k)} \circ \widetilde{T}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\right]\mathbf{V}_k^*\right) \\
&= \operatorname{tr}\left((\mathbf{U}_k^*\mathbf{Y}\mathbf{V}_k)^*\left[\mathbf{H}_1^{(k)} \circ \widetilde{S}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k) + \mathbf{H}_2^{(k)} \circ \widetilde{T}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\right]\right) \\
&= \operatorname{tr}\left(\left(\widetilde{S}(\mathbf{U}_k^*\mathbf{Y}\mathbf{V}_k) + \widetilde{T}(\mathbf{U}_k^*\mathbf{Y}\mathbf{V}_k)\right)^*\left[\mathbf{H}_1^{(k)} \circ \widetilde{S}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k) + \mathbf{H}_2^{(k)} \circ \widetilde{T}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\right]\right) \\
&= \operatorname{tr}\left((\widetilde{S}(\mathbf{U}_k^*\mathbf{Y}\mathbf{V}_k))^*\left[\mathbf{H}_1^{(k)} \circ \widetilde{S}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\right]\right) + \operatorname{tr}\left((\widetilde{T}(\mathbf{U}_k^*\mathbf{Y}\mathbf{V}_k))^*\left[\mathbf{H}_2^{(k)} \circ \widetilde{T}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\right]\right) \\
&= \operatorname{tr}\left((\mathbf{H}_1^{(k)} \circ \widetilde{S}(\mathbf{U}_k^*\mathbf{Y}\mathbf{V}_k))^*\widetilde{S}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\right) + \operatorname{tr}\left((\mathbf{H}_2^{(k)} \circ \widetilde{T}(\mathbf{U}_k^*\mathbf{Y}\mathbf{V}_k))^*\widetilde{T}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\right) \\
&= \operatorname{tr}\left((\mathbf{H}_1^{(k)} \circ \widetilde{S}(\mathbf{U}_k^*\mathbf{Y}\mathbf{V}_k) + \mathbf{H}_2^{(k)} \circ \widetilde{T}(\mathbf{U}_k^*\mathbf{Y}\mathbf{V}_k))^*\left[\widetilde{S}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k) + \widetilde{T}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\right]\right) \\
&= \operatorname{tr}\left((\mathbf{H}_1^{(k)} \circ \widetilde{S}(\mathbf{U}_k^*\mathbf{Y}\mathbf{V}_k) + \mathbf{H}_2^{(k)} \circ \widetilde{T}(\mathbf{U}_k^*\mathbf{Y}\mathbf{V}_k))^*(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\right) = \langle W^{(k)}(\mathbf{Y}), \mathbf{Z}\rangle,
\end{aligned}$$

where we used the fact that

$$\mathbb{R}^{d_1 \times d_2} = \operatorname{span}\left(\frac{1}{\sqrt{2}}(e_i e_j^* + e_j e_i^*), e_i e_k^* : i, j \in [d_1], k \in [d_2 - d_1]\right) \oplus \operatorname{span}\left(\frac{1}{\sqrt{2}}(e_i e_j^* - e_j e_i^*) : i, j \in [d_1]\right)$$

and that $\widetilde{T}$ and $\widetilde{S}$ are the respective orthogonal projection operators in the fourth and sixth equality, and $\operatorname{tr}(\mathbf{C}^*(\mathbf{B} \circ \mathbf{A})) = \operatorname{tr}((\mathbf{C} \circ \mathbf{B})^*\mathbf{A})$ for any matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ [Ber09, Fact 7.6.10] in the fifth equality.

Furthermore, with this argument we can diagonalize $W^{(k)}$ and we see that the eigenvalues of the operator are

$$\lambda(W^{(k)}) = \left\{(\mathbf{H}_1^{(k)})_{ij} : i \in [d_1], j \in [d_2], i \leq j\right\} \cup \left\{(\mathbf{H}_2^{(k)})_{ij} : i \in [d_1], j \in [d_1], i < j\right\}. \tag{2.60}$$

In particular, since all the $(\mathbf{H}_1^{(k)})_{ij}$ and $(\mathbf{H}_2^{(k)})_{ij}$ are positive, this means that $W^{(k)}$ from Definition 2.3.4 is *positive definite*, whereas the second derivative $\nabla^2 F_\epsilon(\mathbf{X}^{(k)})$ of F_ϵ at $X^{(k)}$ always has also negative eigenvalues (as long as $R \leq 1$), compare, e.g., the values

$$(\mathbf{H}_1)_{ii} = (p-1)(\sigma_i^{(k)})^{p-2}$$

of the diagonal of the matrix $\mathbf{H}_1$ from (2.49).

We now obtain the following global majorization result of the quadratic defined $Q_\epsilon(\cdot|X^{(k)})$ by $W^{(k)}$ from Definition 2.3.4.

Theorem 2.4 (Global majorization of smoothed rank surrogate by `MatrixIRLS` quadratic). *Let $0 \leq p \leq 1$ and $\epsilon > 0$, let $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ and $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ be as in Definition 2.3.4, and let $F_\epsilon : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}$ be the ϵ-smoothed rank surrogate function (2.48). Then the property (2.18) is fulfilled such that*

$$\nabla F_\epsilon(\mathbf{X}^{(k)}) = W^{(k)}(\mathbf{X}^{(k)}),$$

and the quadratic function $Q_\epsilon(\cdot|\mathbf{X}^{(k)}) : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}$ defined by

$$Q_\epsilon(\mathbf{X}|\mathbf{X}^{(k)}) = F_\epsilon(\mathbf{X}^{(k)}) + \langle \nabla F_\epsilon(\mathbf{X}^{(k)}), \mathbf{X} - \mathbf{X}^{(k)}\rangle + \frac{1}{2}\langle \mathbf{X} - \mathbf{X}^{(k)}, W^{(k)}(\mathbf{X} - \mathbf{X}^{(k)})\rangle \tag{2.61}$$

globally majorizes F_ϵ (cf. (2.19)) such that

$$Q_\epsilon(\mathbf{X}|\mathbf{X}^{(k)}) \geq F_\epsilon(\mathbf{X})$$

for any $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$.

To prove the second statement of Theorem 2.4, we will pursue an analytical approach involving optimality conditions of the function $Q_\epsilon(\cdot|\mathbf{X}^{(k)}) - F_\epsilon(\cdot)$. However, as $F_\epsilon(\cdot)$ is not a $\mathscr{C}^2$ function on the entirety of its domain, but only differentiable with Lipschitz continuous gradient, we introduce the following *generalized* notion of derivatives.

Definition 2.3.5. *Let $F : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ be a differentiable function with whose gradient ∇F is Lipschitz continuous in a neighborhood of $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, and let $\mathscr{D}_F := \{\mathbf{X} : F \text{ is twice differntiable at } \mathbf{X}\}$. Then we define the* generalized Hessian *at $\mathbf{X}$ as the set*

$$\partial^2 F(\mathbf{X}) = \operatorname{conv}\left(\{M \in \{G : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2} \text{ linear}\} : \exists (\mathbf{X}^\ell)_\ell \subset \mathscr{D}_F \text{ s. t. } \mathbf{X}^\ell \to \mathbf{X}, \nabla^2 F(\mathbf{X}^\ell) \to M\}\right).$$

Remark 2.3.3. *Differentiable functions with locally Lipschitz continuous gradient are also called $\mathscr{C}^{1,1}$ functions [HUSN84, KT88, CC90, YJ92] or $LC1$ functions [Qi94, Fac95] in the literature. Definition 2.3.5 goes back to [HU77, HUSN84] and is based on the notion of* generalized Jacobian *as defined by Clarke [Cla75, Cla90] — the generalized Hessian of Definition 2.3.5 is just the generalized Jacobian of the gradient ∇F of F.*

With Definition 2.3.5, we can formulate necessary second-order conditions for minima of F_ϵ.

Theorem 2.5 ([HUSN84, Theorem 3.1])**.** *Let $F : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ be a differentiable function with whose gradient ∇F is Lipschitz continuous on $\mathbb{R}^{d_1 \times d_2}$.*

If $\mathbf{X}^\mathbf{0} \in \mathbb{R}^{d_1 \times d_2}$ is a local minimum of F, then $\nabla F(\mathbf{X}) = 0$ and for each $\Delta \in \mathbb{R}^{d_1 \times d_2}$, there exists a matrix $\mathbf{M} \in \partial^2 F(\mathbf{X})$ such that

$$\langle \Delta, M\Delta \rangle \geq 0.$$

Proof of Theorem 2.4. By Lemma 2.3.1, F_ϵ is differentiable with Lipschitz gradient

$$\nabla F_\epsilon(\mathbf{X}) = \mathbf{U} \operatorname{dg}\left(\left(\frac{\sigma_i(\mathbf{X})}{\max(\sigma_i(\mathbf{X}), \epsilon)^{2-p}}\right)_{i=1}^{d}\right) \mathbf{V}^* \tag{2.62}$$

at $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, where $\mathbf{X} = \mathbf{U} \operatorname{dg}\left(\sigma(\mathbf{X})\right)\mathbf{V}^* = \mathbf{U} \operatorname{dg}\left(\sigma\right)\mathbf{V}^*$ is a singular value decomposition with unitary matrices $\mathbf{U} \in \mathbb{O}^{d_1}$ and $\mathbf{V} \in \mathbb{O}^{d_2}$.

Inserting $\mathbf{X}^{(k)} = \mathbf{U}_k \operatorname{dg}\left(\sigma(\mathbf{X}^{(k)})\right)\mathbf{V}_k^*$ into the definition (2.57) of the weight operator $W^{(k)}$, we see that

$$W^{(k)}(\mathbf{X}^{(k)} = \mathbf{U}_k \left[\mathbf{H}_1^{(k)} \circ \operatorname{dg}\left(\sigma(\mathbf{X}^{(k)})\right)\right] \mathbf{V}_k^* = \mathbf{U}_k \operatorname{dg}\left(\left(\frac{\sigma_i(\mathbf{X}^{(k)})}{\max(\sigma_i(\mathbf{X}^{(k)}), \epsilon)^{2-p}}\right)_{i=1}^{d}\right) \mathbf{V}_k^*,$$

which shows the first statement of Theorem 2.4.

For the second statement of the theorem, let $G_\epsilon : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ be the function such that

$$G_\epsilon(\mathbf{X}) := Q_\epsilon(\mathbf{X}|\mathbf{X}^{(k)}) - F_\epsilon(\mathbf{X}) \tag{2.63}$$

or any $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$. It suffices to show that

$$G_\epsilon(\mathbf{X}) \geq 0 \tag{2.64}$$

for any $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$.

As $Q_\epsilon(\cdot|\mathbf{X}^{(k)})$ is quadratic, it is in $\mathscr{C}^\infty$, and since $F_\epsilon(\cdot)$ is differentiable with Lipschitz gradient, also G_ϵ is differentiable with Lipschitz continuous gradient.

Furthermore, G_ϵ is *coercive* as a difference of a quadratic function and the function $F_\epsilon(\cdot)$, which is p-homogenous function ($0 < p \leq 1$) or has logarithmic growth ($p = 0$). As G_ϵ is continuous, (2.64) is shown therefore if all local minima of G_ϵ are non-negative, and this again can be shown by showing that all points stationary points $\mathbf{X}$ satisfy $G_\epsilon(\mathbf{X}) \geq 0$.

Let now $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ be a stationary point of $G_\epsilon(\cdot)$, i.e., let $\mathbf{X}$ be such that $\nabla G_\epsilon(X) = W^{(k)}(\mathbf{X}) - \nabla F_\epsilon(\mathbf{X}) = 0$ or, equivalently, such that

$$W^{(k)}(\mathbf{X}) = \nabla F_\epsilon(\mathbf{X}). \tag{2.65}$$

Thus, inserting (2.62) and the formula for $W^{(k)}$ from Definition 2.3.4 into (2.65), we see that (2.65) is equivalent to

$$\mathbf{U}^*\mathbf{U}^{(k)} \left[\mathbf{H}_1^{(k)} \circ \widetilde{S}(\mathbf{U}_k^*\mathbf{U}\Sigma\mathbf{V}^*\mathbf{V}_k) + \mathbf{H}_2^{(k)} \circ \widetilde{T}(\mathbf{U}_k^*\mathbf{U}\Sigma\mathbf{V}^*\mathbf{V}_k)\right] \mathbf{V}_k^*\mathbf{V} = \mathrm{dg}\left(\left(\frac{\sigma_i(\mathbf{X})}{\max(\sigma_i(\mathbf{X}), \epsilon)^{2-p}}\right)_{i=1}^d\right),$$

which can be written as

$$\widetilde{\mathbf{U}} \left[\mathbf{H}_1^{(k)} \circ \widetilde{S}(\widetilde{\mathbf{U}}^*\Sigma\widetilde{\mathbf{V}}) + \mathbf{H}_2^{(k)} \circ \widetilde{T}(\widetilde{\mathbf{U}}^*\Sigma\widetilde{\mathbf{V}})\right] \widetilde{\mathbf{V}}^* = \mathrm{dg}\left(\left(\frac{\sigma_i(\mathbf{X})}{\max(\sigma_i(\mathbf{X}), \epsilon)^{2-p}}\right)_{i=1}^d\right), \tag{2.66}$$

if we define $\widetilde{\mathbf{U}} = \left(\widetilde{u}_1, \ \ldots, \ \widetilde{u}_{d_1}\right) = \mathbf{U}^*\mathbf{U}_k$ and $\widetilde{\mathbf{V}} = \left(\widetilde{v}_1, \ \ldots, \ \widetilde{v}_{d_2}\right) = \mathbf{V}^*\mathbf{V}_k$. As $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ are orthogonal matrices, their columns $(\widetilde{u}_i)_{i=1}^{d_1}$ and $(\widetilde{v}_i)_{i=1}^{d_2}$ constitute bases of $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively, and $(\widetilde{u}_i\widetilde{v}_j^*)_{i=1,j=1}^{d_1,d_2}$ constitutes a basis of $\mathbb{R}^{d_1 \times d_2}$ as an outer product of two bases.

For the next proof step, we define the vectorization operator

$$\mathrm{vec} : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 d_2}, \mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \ldots & \mathbf{X}_{d_2}\end{bmatrix} \mapsto \mathbf{X}_{\mathrm{vec}} = \begin{pmatrix}\mathbf{X}_1 \\ \vdots \\ \mathbf{X}_{d_2}\end{pmatrix}$$

that stacks the columns of a matrix $\mathbf{X}$, and use the notation $B = \widetilde{\mathbf{V}} \otimes \widetilde{\mathbf{U}} \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ for the *Kronecker product* of $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$. Furthermore, we write $Q \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ for the matricization of the matrix operator that describes the basis transformation mapping from the coefficients of the orthonormal basis

$$\left\{\frac{1}{\sqrt{2}}(e_i e_j^* + e_j e_i^*), e_i e_k^* : i, j \in [d_1], k \in [d_2 - d_1]\right\} \cup \left\{\frac{1}{\sqrt{2}}(e_i e_j^* - e_j e_i^*) : i, j \in [d_1]\right\} \tag{2.67}$$

to the coefficients of the standard basis

$$\left\{e_i e_j^* : i, j \in [d_1], k \in [d_2]\right\}.$$

With this notation, by applying vec to both sides of (2.66), it can be seen that (2.66) is equivalent to

$$BQ\begin{bmatrix}\mathrm{diag}\left(((\mathbf{H}_1^{(k)})_{i\le j})_{\mathrm{vec}}\right) & \\ & \mathrm{diag}\left(((\mathbf{H}_2^{(k)})_{i>j})_{\mathrm{vec}}\right)\end{bmatrix} Q^* B^* \Sigma_{\mathrm{vec}} = (\mathbf{D}_{\mathbf{X}})_{\mathrm{vec}}$$

with the notation $\mathbf{D}_{\mathbf{X}} = \mathrm{dg}\left(\frac{\sigma_i(\mathbf{X})}{\max(\sigma_i(\mathbf{X}),\epsilon)^{2-p}}\right)_{i=1}^{d}$.

Let r be the rank of $\mathbf{X}$, i.e., $r = \mathrm{rank}(\mathbf{X})$. By considering the restriction of the latter equation to the entries corresponding to the the double index set $I = \{(k,l) \in [d_1] \times [d_2] : k \neq l \text{ or } (k = l \text{ and } \sigma_k(X) = 0)\}$, we learn that

$$\left(BQ\begin{bmatrix}\mathrm{diag}\left(((\mathbf{H}_1^{(k)})_{i\le j})_{\mathrm{vec}}\right) & \\ & \mathrm{diag}\left(((\mathbf{H}_2^{(k)})_{i>j})_{\mathrm{vec}}\right)\end{bmatrix} Q^* B^* \Sigma_{\mathrm{vec}}\right)_I = 0, \tag{2.68}$$

where $(\mathbf{v})_I$ is the restriction of the vector $\mathbf{v} \in \mathbb{R}^{d_1 d_2}$ to the elements indexed by I, and $|I| = d_1 d_2 - r$. Since B, Q, Q^* and the diagonal matrix with entries of $\mathbf{H}_1$ and $\mathbf{H}_2$ are all of full rank, their products are, and also

$$M_{I,:} := \left(BQ\begin{bmatrix}\mathrm{diag}\left(((\mathbf{H}_1^{(k)})_{i\le j})_{\mathrm{vec}}\right) & \\ & \mathrm{diag}\left(((\mathbf{H}_2^{(k)})_{i>j})_{\mathrm{vec}}\right)\end{bmatrix} Q^*\right)_{I,:} \tag{2.69}$$

is of full rank $d_1 d_2 - r$. Thus, there exist $d_1 d_2 - r$ columns of the matrix $M_{I,:}$ of (2.69) that are linear independent. Let $J \subset [d_1] \times [d_2]$ be the double index set corresponding to this set of $d_1 d_2 - r$ columns.

Then, $M_{I,J}$ is of full rank $d_1 d_2 - r$, and in particular, it has an empty null space. As the right hand side of (2.68) is zero, this entails that

$$(B^* \Sigma_{\mathrm{vec}})_J = 0,$$

and since $(B^* \Sigma_{\mathrm{vec}})_J = \left(\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}\right)_J$, it follows that

$$\left(\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}\right)_{ij} = 0$$

for all $(i,j) \in J$.

This shows that there can be at most r non-zero entries in the $(d_1 \times d_2)$-matrix $\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}$.

Furthermore, it holds that $\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}$ has, in fact, exactly r non-zero entries, which are distributed across r non-zero columns and r non-zero rows. Indeed, if this was not the case, (2.66) could not be fulfilled, as the matrix $\mathbf{D}_{\mathbf{X}}$ on the right hand side has rank r (as $\mathrm{rank}(\mathbf{X}) = r$), and as $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ are orthogonal matrices and therefore of full rank.

This means that there exists an $(r \times r)$ submatrix $\mathbf{M}_1$ of $\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}$ such that

$$\mathbf{M}_1 = \mathbf{P}\,\mathrm{diag}\left(\sigma_\ell(\mathbf{X})\right)_{\ell=1}^{r} \tag{2.70}$$

and $\mathbf{M}_1$ as exactly one non-zero entry in each row and in each column, and where $\mathbf{P} \in \mathbb{R}^{r \times r}$ is a *signed permutation matrix*, cf. Definition 2.3.2. Since $\mathbf{D}_{\mathbf{X}}$ is non-zero at precisely the locations where $\Sigma = \mathrm{dg}(\sigma(\mathbf{X}))$ in non-zero, $\widetilde{\mathbf{U}}^* \mathbf{D}_{\mathbf{X}} \widetilde{\mathbf{V}}$ has the same r locations of non-zeros as $\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}$, and there exists an an $(r \times r)$ submatrix $\mathbf{M}_2$ of $\widetilde{\mathbf{U}}^* \mathbf{D_X} \widetilde{\mathbf{V}}$ such that

$$\mathbf{M}_2 = \mathbf{P}\,\mathrm{diag}\left(\frac{\sigma_\ell(\mathbf{X})}{\max(\sigma_\ell(\mathbf{X}), \epsilon)^{2-p}}\right)_{\ell=1}^{r} \tag{2.71}$$

with the same matrix $\mathbf{P}$ as above.

Taking this into account, we can determine the singular values $\sigma_1(\mathbf{X}), \dots, \sigma_r(\mathbf{X})$ of the stationary point X. As (2.66) is equivalent to

$$\left[\mathbf{H}_1^{(k)} \circ \widetilde{S}(\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}) + \mathbf{H}_2^{(k)} \circ \widetilde{T}(\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}})\right] = \widetilde{\mathbf{U}}^* \mathbf{D_X} \widetilde{\mathbf{V}}, \tag{2.72}$$

we obtain the restricted matrix equations

$$\mathbf{H}_1^{(k)} \circ \widetilde{S}(\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}) = \widetilde{S}(\widetilde{\mathbf{U}}^* \mathbf{D_X} \widetilde{\mathbf{V}}), \tag{2.73}$$

and

$$\mathbf{H}_2^{(k)} \circ \widetilde{T}(\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}) = \widetilde{T}(\widetilde{\mathbf{U}}^* \mathbf{D_X} \widetilde{\mathbf{V}}). \tag{2.74}$$

Since $\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}$ and $\widetilde{\mathbf{U}}^* \mathbf{D_X} \widetilde{\mathbf{V}}$ have r non-zero entries with r non-zero rows and columns at the same locations, and due to (2.70) and (2.71), there exist r index pairs $\{(i_\ell, j_\ell)_{\ell=1}^r\} \subset [d_1] \times [d_2]$ such that $(i_\ell, j_\ell)_{\ell=1}^r = [d_1] \times [d_2] \setminus J =: J^c$, and it holds that

$$(\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}})_{i,j} = \begin{cases} (\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}})_{i_\ell, j_\ell} = \pm \sigma_\ell(\mathbf{X}), \\ 0, \end{cases} \text{and } (\widetilde{\mathbf{U}}^* \mathbf{D_X} \widetilde{\mathbf{V}})_{i,j} = \begin{cases} (\widetilde{\mathbf{U}}^* \mathbf{D_X} \widetilde{\mathbf{V}})_{i_\ell, j_\ell} = \pm (\mathbf{D_X})_{\ell,\ell}, & \text{if } (i_\ell, j_\ell) \in J^c, \\ 0, & \text{otherwise}, \end{cases} \tag{2.75}$$

where the sign of $(\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}})_{i_\ell, j_\ell}$ and $(\widetilde{\mathbf{U}}^* \mathbf{D_X} \widetilde{\mathbf{V}})_{i_\ell, j_\ell}$ always coincides.

Thus, for each $\ell \in [r]$, we obtain up to two equations from (2.73) and (2.74), which can be stated adequately if we define a function $\omega : [r] \to [r] \cup \emptyset$ such that for any $\ell \in [r]$, ω maps to

$$\omega(\ell) = \begin{cases} \ell' \in [r] & \text{if } (\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}})_{i_\ell, j_\ell} = \pm \sigma_\ell(\mathbf{X}) \text{ and there exists } \ell' \in [r] \text{ s.t. } (\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}})_{j_\ell, i_\ell} = \pm \sigma_{\ell'}(\mathbf{X}), \\ \emptyset & \text{otherwise.} \end{cases}$$

In particular, let now $\ell \in [r]$.

We consider now three cases: The case that $\omega(\ell) = \emptyset$, the case that $\omega(\ell)\ell' \in [r]$ such that $\mathrm{sgn}\left(\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}\right)_{i_\ell, j_\ell} = \mathrm{sgn}\left(\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}\right)_{i_{\omega(\ell)}, j_{\omega(\ell)}}$, and the third case such that $\omega(\ell) \neq \emptyset$ and also $\mathrm{sgn}\left(\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}\right)_{i_\ell, j_\ell} \neq \mathrm{sgn}\left(\widetilde{\mathbf{U}}^* \Sigma \widetilde{\mathbf{V}}\right)_{i_{\omega(\ell)}, j_{\omega(\ell)}}$.

Case 1: Suppose $\omega(\ell) = \emptyset$.

In this case (2.73) implies that

$$\frac{1}{2}(\mathbf{H}_1^{(k)})_{i_\ell, j_\ell}\,(\sigma_\ell(\mathbf{X}) + 0) = \frac{1}{2}\frac{\sigma_\ell(\mathbf{X})}{\max(\sigma_\ell(\mathbf{X}), \epsilon)^{2-p}} + 0$$

and (2.74) implies that

$$\frac{1}{2}(\mathbf{H}_2^{(k)})_{i_\ell, j_\ell}\,(\sigma_\ell(\mathbf{X}) - 0) = \frac{1}{2}\frac{\sigma_\ell(\mathbf{X})}{\max(\sigma_\ell(\mathbf{X}), \epsilon)^{2-p}} - 0,$$

which can only be fulfilled for $\sigma_\ell(\mathbf{X}) \neq 0$ if (i_ℓ, j_ℓ) are such that

$$\max(\sigma_\ell(\mathbf{X}), \epsilon) = (\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}} = (\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}}.$$

If $p > 0$, the (i,j)-th entry of $(\mathbf{H}_1^{(k)})$ and $(\mathbf{H}_2^{(k)})$ only coincide if (i,j) is a double index corresponding to the diagonal, or if both $i_\ell > R_k$ and $j > R_k$ due to their definitions (2.58) and (2.59). As diagonal indices (i_ℓ, j_ℓ) are excluded due to $\omega(\ell) = \emptyset$, $(\mathbf{H}_1^{(k)})_{i_\ell,j_\ell} = (\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}$ can be only true if

$$\sigma_\ell(\mathbf{X}) = \left[(\epsilon^p)^{1-2/p}\right]^{-\frac{1}{2-p}} = \epsilon^{-\frac{p-2}{2-p}} = \epsilon.$$

If $p = 0$, it follows that

$$\sigma_\ell(\mathbf{X}) = \left[(\sigma_i^{(k)} \vee \epsilon)(\sigma_j^{(k)} \vee \epsilon)\right]^{-\frac{1}{2}} = \begin{cases} \sqrt{\sigma_i^{(k)}\sigma_j^{(k)}} & \text{if } i,j \leq R_k, \\ \sqrt{\sigma_i^{(k)}\epsilon} & \text{if } i \leq R_k, R_k < j, \\ \sqrt{\sigma_j^{(k)}\epsilon} & \text{if } j \leq R_k, R_k < i, \\ \epsilon & \text{if } R_k < i, j. \end{cases}$$

Case 2: Suppose $\omega(\ell) = \ell' \in [r]$ and $\operatorname{sgn}\left(\widetilde{\mathbf{U}}^*\Sigma\widetilde{\mathbf{V}}\right)_{i_\ell,j_\ell} = \operatorname{sgn}\left(\widetilde{\mathbf{U}}^*\Sigma\widetilde{\mathbf{V}}\right)_{i_{\omega(\ell)},j_{\omega(\ell)}}$.
In this case we obtain from (2.73) and (2.74) the equations

$$\frac{1}{2}(\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}\left(\sigma_\ell(\mathbf{X}) + \sigma_{\omega(\ell)}(\mathbf{X})\right) = \frac{1}{2}\left(\frac{\sigma_\ell(\mathbf{X})}{\max(\sigma_\ell(\mathbf{X}),\epsilon)^{2-p}} + \frac{\sigma_{\omega(\ell)}(\mathbf{X})}{\max(\sigma_{\omega(\ell)}(\mathbf{X}),\epsilon)^{2-p}}\right) \tag{2.76}$$

and

$$\frac{1}{2}(\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}\left(\sigma_\ell(\mathbf{X}) - \sigma_{\omega(\ell)}(\mathbf{X})\right) = \frac{1}{2}\left(\frac{\sigma_\ell(\mathbf{X})}{\max(\sigma_\ell(\mathbf{X}),\epsilon)^{2-p}} - \frac{\sigma_{\omega(\ell)}(\mathbf{X})}{\max(\sigma_{\omega(\ell)}(\mathbf{X}),\epsilon)^{2-p}}\right). \tag{2.77}$$

Due to the symmetry of these equations, we assume that $\omega(\ell) \leq \ell$ without loss of generality. If (i_ℓ, j_ℓ) is such that $(\mathbf{H}_2^{(k)})_{i_\ell,j_\ell} = \frac{1}{\epsilon^{2-p}}$, (2.76) and (2.77) can be only fulfilled if $\sigma_\ell(\mathbf{X}), \sigma_{\omega(\ell)}(\mathbf{X}) \leq \epsilon$.

Now let $(\mathbf{H}_2^{(k)})_{i_\ell,j_\ell} < \frac{1}{\epsilon^{2-p}}$.
By rearranging (2.77) we obtain

$$(\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}\sigma_\ell(\mathbf{X}) - \frac{\sigma_\ell(\mathbf{X})}{\max(\sigma_\ell(\mathbf{X}),\epsilon)^{2-p}} = (\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}\sigma_{\omega(\ell)}(\mathbf{X}) - \frac{\sigma_{\omega(\ell)}(\mathbf{X})}{\max(\sigma_{\omega(\ell)}(\mathbf{X}),\epsilon)^{2-p}}. \tag{2.78}$$

As the function $h : (0,\infty) \longrightarrow \mathbb{R}$,

$$\sigma \longmapsto h(\sigma) := (\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}\sigma - \frac{\sigma}{\max(\sigma,\epsilon)^{2-p}}$$

is only negative in the interval $\left(0, (\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}}\right)$, and has negative derivative for all $\sigma \in (0,\epsilon)$, and positive derivative for $\sigma \in \left(\epsilon, (\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}}\right)$, we can infer from (2.78) that either

(a) $\sigma_{\omega(\ell)}(\mathbf{X}) = \sigma_\ell(\mathbf{X})$ or

(b) $\sigma_{\omega(\ell)}(\mathbf{X}) < \epsilon < \sigma_\ell(\mathbf{X}) \leq (\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}}$.

In the case of (a), we can rearrange (2.76) to obtain

$$\max(\sigma_{\omega(\ell)}(\mathbf{X}), \epsilon) = \max(\sigma_\ell(\mathbf{X}), \epsilon) = (\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}}.$$

For (b), we know that $\frac{\sigma_{\omega(\ell)}(\mathbf{X})}{\max(\sigma_{\omega(\ell)}(\mathbf{X}),\epsilon)^{2-p}} = \frac{\sigma_{\omega(\ell)}(\mathbf{X})}{\epsilon^{2-p}}$, and therefore substracting (2.77) from (2.76) yields the equation

$$\frac{1}{2}(\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}\left(\sigma_\ell + \sigma_{\omega(\ell)}\right) - \frac{1}{2}(\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}\left(\sigma_\ell - \sigma_{\omega(\ell)}\right) = \frac{\sigma_{\omega(\ell)}}{\epsilon^{2-p}},$$

$\sigma_\ell := \sigma_\ell(\mathbf{X})$ and $\sigma_{\omega(\ell)} := \sigma_{\omega(\ell)}(\mathbf{X})$, which can be rearranged such that

$$\frac{1}{2}\left((\mathbf{H}_1^{(k)})_{i_\ell,j_\ell} - (\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}\right)\sigma_\ell = \left(\frac{1}{\epsilon^{2-p}} - \frac{1}{2}\left((\mathbf{H}_1^{(k)})_{i_\ell,j_\ell} + (\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}\right)\right)\sigma_{\omega(\ell)}.$$

The left hand side of the latter equation is smaller or equal to 0 since

$$(\mathbf{H}_2^{(k)})_{i_\ell,j_\ell} - (\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}$$

for any (i_ℓ, j_ℓ), whereas the right hand side is strictly positive due to the assumption that

$$(\mathbf{H}_2^{(k)})_{i_\ell,j_\ell} < \frac{1}{\epsilon^{2-p}}.$$

This is a contradiction, which means that case (b) cannot occur.

To sum up, the only remaining possibilities for values of σ_ℓ and $\sigma_{\omega(\ell)}$ in Case 2 are

$$\sigma_\ell \leq \epsilon \text{ and } \sigma_{\omega(\ell)} \leq \epsilon$$

(if $i_\ell > R_k$ and $j_\ell > R_k$) or

$$\sigma_{\omega(\ell)} = \sigma_\ell = (\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}}.$$

Case 3: Suppose $\omega(\ell) = \ell' \in [r]$ and $\operatorname{sgn}\left(\widetilde{\mathbf{U}}^*\Sigma\widetilde{\mathbf{V}}\right)_{i_\ell,j_\ell} \neq \operatorname{sgn}\left(\widetilde{\mathbf{U}}^*\Sigma\widetilde{\mathbf{V}}\right)_{i_{\omega(\ell)},j_{\omega(\ell)}}$.

In this case, we obtain from (2.73) and (2.74) the equations

$$\frac{1}{2}(\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}\left(\sigma_\ell - \sigma_{\omega(\ell)}\right) = \frac{1}{2}\left(\frac{\sigma_\ell}{\max(\sigma_\ell,\epsilon)^{2-p}} - \frac{\sigma_{\omega(\ell)}}{\max(\sigma_{\omega(\ell)},\epsilon)^{2-p}}\right) \tag{2.79}$$

and

$$\frac{1}{2}(\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}\left(\sigma_\ell + \sigma_{\omega(\ell)}\right) = \frac{1}{2}\left(\frac{\sigma_\ell}{\max(\sigma_\ell,\epsilon)^{2-p}} + \frac{\sigma_{\omega(\ell)}}{\max(\sigma_{\omega(\ell)},\epsilon)^{2-p}}\right). \tag{2.80}$$

As in Case 2, we can assume without loss of generality that $\sigma_{\omega(\ell)} \leq \sigma_\ell$.

If $\sigma_{\omega(\ell)} = \sigma_\ell$, (2.79) is fulfilled, and we obtain

$$\max(\sigma_{\omega(\ell)}(\mathbf{X}), \epsilon) = \max(\sigma_\ell(\mathbf{X}), \epsilon) = (\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}}$$

from (2.80) implying that $\sigma_{\omega(\ell)} = \sigma_\ell \leq \epsilon$ if $i_\ell > R_k$ and $j_\ell > R_k$, and that $\sigma_{\omega(\ell)} = \sigma_\ell = (\mathbf{H}_2^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}}$ if (i_ℓ, j_ℓ) is such that $i_\ell \leq R_k$ or $j_\ell \leq R_k$.

Let now $\sigma_{\omega(\ell)} < \sigma_\ell$. Rearranging (2.79), we obtain

$$(\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}\sigma_\ell - \frac{\sigma_\ell}{\max(\sigma_\ell,\epsilon)^{2-p}} = (\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}\sigma_{\omega(\ell)} - \frac{\sigma_{\omega(\ell)}}{\max(\sigma_{\omega(\ell)},\epsilon)^{2-p}}, \tag{2.81}$$

and as above, the monotonicity of $\sigma \mapsto h(\sigma) := (\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}\sigma - \frac{\sigma}{\max(\sigma,\epsilon)^{2-p}}$ implies that

$$\sigma_{\omega(\ell)} < \epsilon < \sigma_\ell \leq (\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}},$$

since

$$h'(\sigma) = \begin{cases} (\mathbf{H}_1^{(k)})_{i_\ell,j_\ell} - \frac{1}{\epsilon^{2-p}}, & \text{if } \sigma \leq \epsilon, \\ (\mathbf{H}_1^{(k)})_{i_\ell,j_\ell} + (1-p)\sigma^{p-2}, & \text{if } \sigma > \epsilon \end{cases}$$

and $h(\sigma) < 0$ if and only if $\sigma \in \left(0, (\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}}\right)$.

We now have considered all possible cases, and we conclude, as a summary, that for each $\ell \in [r]$, one of the following three statements is true:

$$\sigma_{\omega(\ell)}(\mathbf{X}) = \sigma_\ell(\mathbf{X}) = (\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}}, \tag{2.82}$$

$$\sigma_{\omega(\ell)}(\mathbf{X}) < \epsilon < \sigma_\ell(\mathbf{X}) \leq (\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}} \tag{2.83}$$

(if $(\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}} > \epsilon$) or

$$\sigma_\ell(\mathbf{X}) \leq \epsilon \text{ and } \sigma_{\omega(\ell)}(\mathbf{X}) \leq \epsilon \tag{2.84}$$

(if $(\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}} = \epsilon$).

As the last part of the proof of Theorem 2.4, we now provide an alternative expression of the function value of $G_\epsilon(\cdot)$ at stationary points $\mathbf{X}$. Let $\mathbf{X}$ again be a stationary point of $G_\epsilon(\cdot)$ of rank r. Then

$$G_\epsilon(\mathbf{X}) = F_\epsilon(\mathbf{X}^{(k)}) + \frac{1}{2}\left(\langle \mathbf{X}, W^{(k)}(\mathbf{X})\rangle - \langle \mathbf{X}^{(k)}, W^{(k)}(\mathbf{X}^{(k)})\rangle\right) - F_\epsilon(\mathbf{X}) = \widetilde{F}_\epsilon(\mathbf{X}) - \widetilde{F}_\epsilon(\mathbf{X}^{(k)}) \tag{2.85}$$

with the definition that $\widetilde{F}_\epsilon(\mathbf{X}) := \frac{1}{2}\langle \mathbf{X}, \widetilde{W}^{(k)}(\mathbf{X})\rangle - F_\epsilon(\mathbf{X})$. At stationary points $\mathbf{X}$, $\widetilde{F}_\epsilon$ can be rewritten as a function of depending exclusively on the singular values of $\mathbf{X}$. Recall that $R = |\{\ell \in [d] : \sigma_\ell(\mathbf{X}) > \epsilon\}|$. Indeed, we note that then, using (2.72) and the notation $\sigma_\ell = \sigma_\ell(\mathbf{X})$,

$$\begin{aligned} \frac{1}{2}\langle \mathbf{X}, \widetilde{W}^{(k)}(\mathbf{X})\rangle &= \frac{1}{2}\langle \widetilde{\mathbf{U}}^*\Sigma\widetilde{\mathbf{V}}, \left[\mathbf{H}_1^{(k)} \circ \widetilde{S}(\widetilde{\mathbf{U}}^*\Sigma\widetilde{\mathbf{V}}) + \mathbf{H}_2^{(k)} \circ \widetilde{T}(\widetilde{\mathbf{U}}^*\Sigma\widetilde{\mathbf{V}})\right]\rangle = \frac{1}{2}\langle \widetilde{\mathbf{U}}^*\Sigma\widetilde{\mathbf{V}}, \widetilde{\mathbf{U}}^*\mathbf{D}_{\mathbf{X}}\widetilde{\mathbf{V}}\rangle \\ &= \frac{1}{2}\langle \Sigma, \mathbf{D}_{\mathbf{X}}\rangle = \begin{cases} \frac{1}{2}\left(R + \sum\limits_{\ell:\sigma_\ell \leq \epsilon} \frac{\sigma_\ell^2}{\epsilon^2}\right), & \text{if } p = 0, \\ \frac{1}{2}\left(\sum\limits_{\ell:\sigma_\ell > \epsilon} \sigma_\ell^p + \sum\limits_{\ell:\sigma_\ell \leq \epsilon} \frac{\sigma_\ell^2}{\epsilon^{2-p}}\right), & \text{if } 0 \leq p \leq 1, \end{cases} \end{aligned}$$

and the value of the smoothed objective $F_\epsilon(\mathbf{X})$ at $\mathbf{X}$ is

$$F_\epsilon(\mathbf{X}) = \begin{cases} \sum_{\ell=1}^d \left[\log(\max(\sigma_\ell,\epsilon)) + \frac{1}{2}\left(\frac{\min(\sigma_\ell,\epsilon)^2}{\epsilon^2} - 1\right)\right], & \text{if } p = 0, \\ \sum_{\ell=1}^d \left[\frac{1}{p}\max(\sigma_\ell,\epsilon)^p + \frac{1}{2}\left(\frac{\min(\sigma_\ell,\epsilon)^2}{\epsilon^{2-p}} - \epsilon^p\right)\right], & \text{if } 0 \leq p \leq 1. \end{cases}$$

This means that for $p = 0$, we obtain

$$\widetilde{F}_\epsilon(\mathbf{X}) = \frac{1}{2}\left(R - (d-R)\right) - \sum_{\ell=1}^{d} \log\left(\max(\sigma_\ell, \epsilon)\right) = -\frac{d}{2} - \sum_{\ell=1}^{d} \log\left(\max(\sigma_\ell, \epsilon)\right) \tag{2.86}$$

and

$$\widetilde{F}_\epsilon(\mathbf{X}) = -\left(\frac{1}{p} - \frac{1}{2}\right)\sum_{\ell=1}^{d} \max(\sigma_\ell, \epsilon)^p \tag{2.87}$$

for $p > 0$. Inserting this back in (2.85), this means that

$$G_\epsilon(\mathbf{X}) = \begin{cases} \sum_{\ell=1}^{d} \log\left(\max(\sigma_\ell(\mathbf{X}^{(k)}), \epsilon)\right) - \sum_{\ell=1}^{d} \log\left(\max(\sigma_\ell(\mathbf{X}), \epsilon)\right) & \text{if } p = 0, \\ \left(\frac{1}{p} - \frac{1}{2}\right)\left[\sum_{\ell=1}^{d} \max(\sigma_\ell(\mathbf{X}^{(k)}), \epsilon)^p - \sum_{\ell=1}^{d} \max(\sigma_\ell(\mathbf{X}), \epsilon)^p\right] & \text{if } 0 < p \leq 1, \end{cases} \tag{2.88}$$

as $\mathbf{X}^{(k)}$ is also a stationary point of G_ϵ.

Let now again $\mathbf{X}$ be an arbitrary stationary point of G_ϵ of rank r, and recall the double index set $J^c = \{(i_\ell, j_\ell)_{\ell=1}^r\} \subset [d_1] \times [d_2]$ of (2.75).

In the case of $p = 0$, by (2.82)–(2.84), we see that

$$\begin{aligned}
\sum_{\ell=1}^{d} \log\left(\max(\sigma_\ell(\mathbf{X}), \epsilon)\right) &= \sum_{\ell=1}^{r} \log\left(\max(\sigma_\ell(\mathbf{X}), \epsilon)\right) + (d-r)\log(\epsilon) \\
&\leq \sum_{\ell=1}^{r} \log\left((\mathbf{H}_1^{(k)})^{\frac{-1}{2}}_{i_\ell, j_\ell}\right) + (d-r)\log(\epsilon) \\
&= \sum_{\ell=1}^{r} \log\left(\sqrt{\max(\sigma_{i_\ell}^{(k)}, \epsilon)}\sqrt{\max(\sigma_{j_\ell}^{(k)}, \epsilon)}\right) + (d-r)\log(\epsilon) \\
&= \frac{1}{2}\left(\sum_{\ell=1}^{r} \log(\max(\sigma_{i_\ell}^{(k)}, \epsilon)) + \sum_{\ell=1}^{r} \log(\max(\sigma_{j_\ell}^{(k)}, \epsilon))\right) + (d-r)\log(\epsilon) \\
&\leq \frac{1}{2}\left(\sum_{i=1}^{r} \log(\max(\sigma_i(\mathbf{X}^{(k)}), \epsilon)) + \sum_{j=1}^{r} \log(\max(\sigma_j(\mathbf{X}^{(k)}), \epsilon))\right) + (d-r)\log(\epsilon) \\
&\leq \sum_{i=1}^{d} \log(\max(\sigma_i(\mathbf{X}^{(k)}), \epsilon)),
\end{aligned}$$

where we used the definition (2.58) of $\mathbf{H}_1^{(k)}$ in the second equality and the fact that, due to the specific structure of the double index set J^c (see also (2.70) and (2.71)), each $i_\ell \in [d_1]$ and $j_\ell \in [d_2]$ can only occur once in the second inequality.

Similarly, for $p > 0$, we calculate that

$$\begin{aligned}\sum_{\ell=1}^{d}(\max(\sigma_\ell(\mathbf{X}),\epsilon)^p &= \sum_{\ell=1}^{r}(\max(\sigma_\ell(\mathbf{X}),\epsilon)^p + (d-r)\epsilon^p \leq \sum_{\ell=1}^{r}\left((\mathbf{H}_1^{(k)})_{i_\ell,j_\ell}^{\frac{-1}{2-p}}\right)^p + (d-r)\epsilon^p \\ &= \sum_{\ell=1}^{r}\left(\frac{\max(\sigma_i^{(k)},\epsilon)^p + \max(\sigma_j^{(k)},\epsilon)^p}{2}\right)^{-\frac{p-2}{2-p}} + ((d-r)\epsilon^p \\ &= \frac{1}{2}\left(\sum_{\ell=1}^{r}\max(\sigma_{i_\ell}^{(k)},\epsilon)^p + \sum_{\ell=1}^{r}\max(\sigma_{j_\ell}^{(k)},\epsilon)^p\right) + (d-r)\epsilon^p \\ &\leq \frac{1}{2}\left(\sum_{i=1}^{r}\max(\sigma_i(\mathbf{X}^{(k)}),\epsilon)^p + \sum_{j=1}^{r}\max(\sigma_j(\mathbf{X}^{(k)}),\epsilon)^p\right) + (d-r)\epsilon^p \\ &\leq \sum_{i=1}^{d}\max(\sigma_i(\mathbf{X}^{(k)}),\epsilon)^p,\end{aligned}$$

using the same arguments as above.

Using this in (2.88), we conclude that

$$G_\epsilon(\mathbf{X}) \geq G_\epsilon(\mathbf{X}^{(k)}) = 0$$

for any stationary point $\mathbf{X}$ of G_ϵ.

By Theorem 2.5, it follows that there is no local minimum of G_ϵ with negative objective value. Thus, $G_\epsilon \geq 0$ globally as G_ϵ is coercive, and this concludes the proof of Theorem 2.4.

□

2.3.4 Formulation of `MatrixIRLS`

Theorem 2.4 paves the ground for a new Iteratively Reweighted Least Squares algorithms for low-rank matrix recovery, which we call `MatrixIRLS`. They optimize logdet or Schatten-p based objectives (2.48), and they are based precisely on the three basic steps of IRLS outlined in the beginning of Section 2.3, where the quadratic upper bounds $Q_{\epsilon_k}(\cdot|\mathbf{X}^{(k)})$ are defined with the weight operators $W^{(k)}$ (2.57) from Definition 2.3.4.

Before we describe `MatrixIRLS` in a schematic way, we recall the original problem, affine rank minimization (2.1), we are interested in: For a given linear operator $\Phi : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$ with non-trivial nullspace and observations $\mathbf{Y} \in \mathbb{R}^m$, we are looking for the matrix $\mathbf{X}_0$ of lowest rank such that $\Phi(\mathbf{X}_0) = \mathbf{Y}$.

We introduce the notation that the *best rank-r* approximation $\mathcal{T}_r(\mathbf{X})$ of a matrix $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$ with singular value decomposition $\mathbf{X} = \begin{bmatrix}\mathbf{U}^{(k)} & \mathbf{U}_\perp^{(k)}\end{bmatrix}\begin{bmatrix}\text{diag}(\sigma_i(\mathbf{X}))_{i=1}^r & 0 \\ 0 & \text{diag}(\sigma_i(\mathbf{X}))_{r=+1}^d\end{bmatrix}\begin{bmatrix}\mathbf{V}^{(k)} \\ \mathbf{V}_\perp^{(k)*}\end{bmatrix}$, where $\mathbf{U}^{(k)} \in \mathbb{R}^{d_1\times r}$ and $\mathbf{V}^{(k)} \in \mathbb{R}^{d_2\times r}$ are matrices with orthonormal columns, is a matrix that fulfills

$$\mathcal{T}_r(\mathbf{X}) = \mathbf{U}^{(k)}\,\text{diag}(\sigma_i(\mathbf{X}))_{i=1}^r\mathbf{V}^{(k)*} = \underset{\mathbf{Z}:\text{rank}(\mathbf{Z})\leq r}{\arg\min}\ \|\mathbf{Z}-\mathbf{X}\|, \tag{2.89}$$

where $\|\cdot\|$ can be any unitarily invariant norm. We note that the last inequality is nothing but the famous Eckardt-Young-Mirsky theorem [EY36, Mir60].

Algorithm 1 `MatrixIRLS` for low-rank matrix recovery

```
Input: Linear operator $\Phi : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^m$, observations $\mathbf{y} \in \mathbb{R}^m$, rank estimate $\widetilde{r}$, concavity
       parameter $0 \leq p \leq 1$.
Output: Sequences $(\mathbf{X}^{(k)})_{k \geq 1} \subset \mathbb{R}^{d_1 \times d_2}$ and $(\epsilon_k)_{k \geq 1}$.
Initialize $k = 0$, $\epsilon^{(0)} = \infty$ and $W^{(0)} = \mathrm{id} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^{d_1 \times d_2}$.
for $k = 1, 2, \ldots$ do
    1. Use a conjugate gradient method to solve linearly constrained quadratic program
           $\mathbf{X}^{(k)} = \arg\min_{\Phi(\mathbf{X}) = \mathbf{y}} \langle \mathbf{X}, W^{(k-1)}(\mathbf{X}) \rangle.$        (2.90)
    2. Find best rank-$(\widetilde{r}+1)$ approximation of $\mathbf{X}^{(k)}$ to obtain $\mathscr{T}_{\widetilde{r}}(\mathbf{X}^{(k)}) = \mathbf{U}^{(k)} \operatorname{diag}(\sigma_i^{(k)})_{i=1}^{\widetilde{r}} \mathbf{V}^{(k)*}$
       and $\sigma_{\widetilde{r}+1}^{(k)}$, update smoothing:
           $\epsilon_k = \min\left(\epsilon_{k-1}, \sigma_{\widetilde{r}+1}(\mathbf{X}^{(k)})\right)$        (2.91)
    3. Update $W^{(k)}$ as in (2.57), using parameters $\epsilon = \epsilon_k$ and $p$ in (2.58) and (2.59), and the
       information $\mathbf{U}^{(k)}, \mathbf{V}^{(k)}$ and $\sigma_1^{(k)}, \ldots, \sigma_{\widetilde{r}+1}$ from item 2.
```

Algorithm 1 is defined in a way that has a few computational advantages compared to other IRLS algorithms as [DDFG10, MF12, FRW11].

We defer implementation details and several computational aspects to Section 2.6.

The choice of the rule (2.91) for the update of the smoothing parameter ϵ plays an important role in the convergence analysis and also has major implications on the geometry of the optimization landscape. On the other hand, there will be several rules that work well. We refer to [Vor12, VD17, FPRW16] for smoothing parameter rules for IRLS algorithm that *do not* depend on estimates of the model order $\widetilde{r}$, but still would allow for a similar analysis.

2.4 Convergence Analysis of `MatrixIRLS`

In this section, we provide a theoretical convergence analysis of `MatrixIRLS`. Furthermore, these convergence properties are put into a relation with existing convergence results about other IRLS algorithms for low-rank matrix recovery such as `IRLS-col` and `IRLS-row` [FRW11, MF12].

In Section 2.4.1, we first state some results that are the outcome of a majorize-minimization interpretation of `MatrixIRLS`, which are enabled by the global majorization result of Theorem 2.4. These results hold for any linear measurement operator $\Phi : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^m$.

We then proceed to a more refined local convergence analysis in Section 2.4.2, which is based on a *(local) restricted isometry property* or *null space property* of the measurement operator Φ.

2.4.1 Basic Properties and Minimization of Smoothed Objective

The statements of the following theorem guarantee that `MatrixIRLS` is actually a valid strategy to minimize the auxiliary objectives F_ϵ, as they never increase the objective F_ϵ. Recall the definition (2.48) of the smoothed logdet and smoothed Schatten-p objective $F_\epsilon : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}$, which was defined such that

$$F_\epsilon(\mathbf{X}) = \begin{cases} \sum_{i=1}^d \left[\log\left(\max(\sigma_i(\mathbf{X}),\epsilon)\right) + \frac{1}{2}\left(\frac{\min(\sigma_i(\mathbf{X}),\epsilon)^2}{\epsilon^2} - 1\right)\right], & \text{if } p = 0, \\ \sum_{i=1}^d \left[\frac{1}{p}\max(\sigma_i(\mathbf{X}),\epsilon)^p + \frac{1}{2}\left(\frac{\min(\sigma_i(\mathbf{X}),\epsilon)^2}{\epsilon^{2-p}} - \epsilon^p\right)\right], & \text{if } 0 < p \leq 1. \end{cases} \tag{2.92}$$

Theorem 2.6 (Monotonicity & Stationary Points Are Accumulation Points)**.** *Let $\Phi : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$, $\mathbf{Y} \in \mathbb{R}^m$, $0 \leq p \leq 1$ and $\widetilde{r} \in \mathbb{N}$. Let $(\mathbf{X}^{(k)}, \epsilon_k)_{k\geq 1}$ a sequence of outputs of* `MatrixIRLS` *corresponding to these parameters. Then the following holds.*

1. *The sequence $\left(F_{\epsilon_k}(\mathbf{X}^{(k)})\right)_{k\geq 1}$ is non-increasing, i.e., $F_{\epsilon_k}(\mathbf{X}^{(k)}) \leq F_{\epsilon_{k-1}}(\mathbf{X}^{(k-1)})$ for any $k \in \mathbb{N}$.*

2. *For any $k \in \mathbb{N}$,*
$$F_{\epsilon_k}(\mathbf{X}^{(k)}) \geq \begin{cases} \sum_{i=1}^d \log\left((\sigma_i(\mathbf{X}^{(k)})\right) & \textit{if } p = 0, \\ \frac{1}{p}\|\mathbf{X}^{(k)}\|_{S_p}^p & \textit{if } 0 < p \leq 1, \end{cases}$$
where $\mathbf{X}^{(k)} = \mathbf{U}_k \operatorname{dg}(\sigma_i(\mathbf{X}))\mathbf{V}_k^*$ is a singular value decomposition of $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1\times d_2}$.

3. *If $p > 0$ or if $\bar{\epsilon} := \lim_{k\to\infty} \epsilon_k > 0$, then the iterates $\mathbf{X}^{(k)}$ and $\mathbf{X}^{(k+1)}$ fulfill $\lim_{k\to\infty} \|\mathbf{X}^{(k)} - \mathbf{X}^{(k+1)}\|_F = 0$.*

4. *If $\bar{\epsilon} := \lim_{k\to\infty} \epsilon_k > 0$, then each subsequence of $(\mathbf{X}^{(k)})_k$ has a convergent subsequence, and each accumulation point $\bar{\mathbf{X}}$ of $(\mathbf{X}^{(k)})_k$ is a stationary point of $F_{\bar{\epsilon}}$ subject to the linear constraint $\{\mathbf{X} : \Phi(\mathbf{X}) = \mathbf{Y}\}$.*

We provide a proof for the statements of Theorem 2.6 in Section 2.8.

Remark 2.4.1.

1. *We note that the statements of Theorem 2.6 had been shown already for the IRLS algorithms of* [DDFG10, FRW11, MF12]*. However, we in particular state this as the majorization property* (2.19) *is not trivial for the tight quadratic bounds defined by the $W^{(k)}$ from* (2.57)*. In the variational framework of* [DDFG10, FRW11, MF12]*,* (2.19) *is an outcome of the minimization of the variational functional $\mathcal{J}$ (see, e.g.,* [DDFG10, (7.9)]*), an argument that is much harder to make for the more involved structure of the weight operator of* `MatrixIRLS`.

2. *The assumption of statement 3. and 4. that $\bar{\epsilon} := \lim_{k\to\infty} \epsilon_k > 0$ is not really a restriction in a practical setting, at least if we know that a matrix $fX_0 \in \mathbb{R}^{d_1\times d_2}$ of rank r is compatible with the measurements and if we used this knowledge algorithmitcally by chosing $\widetilde{r} = r$ in* `MatrixIRLS`*: In this case, $\bar{\epsilon} = 0$ would just imply that iterates $\mathbf{X}^{(k)}$ are of rank at most r in the limit case $k \to \infty$, which is a desirable result, as we are interested in finding low-rank solutions of the linear system.*

The next lemma provides an explicit formula for the calculation of the new iterate $\mathbf{X}^{(k)}$ of `MatrixIRLS` and its characterization by optimality conditions. It is well-known in the IRLS literature ([DDFG10, Eq. (1.9) and Lemma 5.2], [FRW11, Lemma 5.1]. We provide a proof in Section 2.8 for completeness.

Lemma 2.4.1. *Let $0 \leq p \leq 1$ be arbitrary, let $\Phi : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^m$. Let $W^{(k-1)} : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ be the weight operator of Definition 2.3.4 defined based on $\mathbf{X}^{(k-1)} \in \mathbb{R}^{d_1\times d_2}$. Then the solution of* (2.90) *is unique and*

$$\mathbf{X}^{(k)} = \underset{\Phi(\mathbf{X})=\mathbf{y}}{\arg\min}\langle \mathbf{X}, W^{(k-1)}(\mathbf{X})\rangle = (W^{(k-1)})^{-1}\left(\Phi^*\left(\Phi(W^{(k-1)})^{-1}\Phi^*\right)^{-1}(\mathbf{y})\right), \tag{2.93}$$

where $(W^{(k-1)})^{-1} : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ *is the inverse matrix operator of* $W^{(k-1)}$*.*

Moreover, a matrix $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ *coincides with the one of* (2.93) *if and only if*

$$\langle W^{(k-1)}(\mathbf{X}^{(k)}), \eta \rangle = 0 \ \textit{ for all } \ \eta \in \ker \Phi \quad \textit{and} \quad \Phi(\mathbf{X}^{(k)}) = \mathbf{y}. \tag{2.94}$$

While (2.93) provides an explicit formula for the calculation of a new iterate $\mathbf{X}^{(k)}$ of `MatrixIRLS`, using this formula by calculating full matrix operator inversions can be quickly prohibitive from a computational perspective for large data dimensions if the problem structure is not fully used. We refer to Section 2.6 for a detailed discussion of these computational aspects.

2.4.2 Local Convergence Theory: Superlinear Convergence Rates

In the following section, we state our main theoretical results about convergence properties of the algorithm `MatrixIRLS` of Section 2.3.4. Furthermore, we discuss why they improve considerably on what has been obtained for the algorithms `IRLS-col` and `IRLS-row` [FRW11, MF12].

As discussed in Section 2.1.1, it cannot be expected that a low-rank matrix recovery algorithm like `MatrixIRLS` succeeds to converge to a low-rank matrix without any assumptions on the measurement operator $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ that defines the recovery problem (2.1), i.e., the problem

$$\min_{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}} \operatorname{rank}(\mathbf{X}) \quad \text{subject to } \Phi(\mathbf{X}) = \mathbf{y},$$

together with the data vector $\mathbf{y} \in \mathbb{R}^m$.

Apart from the fact that the dimensionality of the range space of Φ needs to be *larger* than the number of degrees of freedom $\deg_f(\mathbf{X}_0) = r(d_1 + d_2 - r)$ of the rank-r matrix $\mathbf{X}_0$ to be recovered, i.e.,

$$m \geq \deg_f(\mathbf{X_0}) = r(d_1 + d_2 - r),$$

it is desirable that Φ fulfills a certain genericity property, avoiding that there are more rank-r solutions than $\mathbf{X}_0$ and ensuring that (2.1) is a well-conditioned problem.

Genericity assumptions of this type are, for example, *rank null space properties* as discussed in Section 2.2.1, or the following *restricted isometry property*, which is a sufficient condition for rank null space properties.

Definition 2.4.1 (Restricted isometry property (RIP))**.** *The* restricted isometry constant $\delta_r > 0$ of order r *of the linear map* $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ *is defined as the smallest number such that*

$$(1 - \delta_r)\|\mathbf{X}\|_F^2 \leq \|\Phi(\mathbf{X})\|_{\ell_2}^2 \leq (1 + \delta_r)\|\mathbf{X}\|_F^2$$

for all matrices $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ *of rank at most* r*.*

If Φ *has a restricted isometry constant* $\delta_r < 1$ *for some* r*, we say that* Φ fulfills the restricted isometry property (RIP) *of order* r *with constant* δ_r*.*

It can be shown (see proof of [CDK15, Theorem 4.1]) that the restricted isometry property order $2r$ with a constant $\delta_{2r} < \frac{2}{\sqrt{2}+3} \approx 0.4531$ implies the following *strong* rank null space properties [FRW11, OMFH11, FR13], which are slightly stronger than the rank null space properties discussed in Section 2.2.1.

Definition 2.4.2 (Strong Schatten-p null space property)**.** *Let* $0 < p \leq 1$*. We say that a linear map* $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ *fulfills the* strong Schatten-p null space property (Schatten-p NSP) *of* *order r with constant $0 < \gamma_r \leq 1$ if*

$$\left(\sum_{i=1}^{r} \sigma_i^2(\mathbf{X})\right)^{p/2} < \frac{\gamma_r}{r^{1-\frac{p}{2}}} \left(\sum_{i=r+1}^{d} \sigma_i^p(\mathbf{X})\right) \tag{2.95}$$

for all $\mathbf{X} \in \ker(\Phi) \setminus \{0\}$.

More precisely, it can be seen that $\delta_{2r} < \frac{2}{\sqrt{2}+3}$ of Definition 2.4.1 implies that the strong Schatten-p NSP (2.95) of order r holds with the constant $\gamma_r = \frac{(\sqrt{2}+1)^p}{2^p}\frac{\delta_{2r}^p}{(1-\delta_{2r})^p}$.

Linear maps that are instances drawn from certain random models are known to fulfill the restricted isometry property with high probability if the number of measurements is sufficiently large [DR16], and, a fortiori, the Schatten-p null space property. In particular, this is true for *sub-Gaussian* linear measurement maps $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ whose matrix representation is such that

$$\frac{1}{\sqrt{m}}\widetilde{\Phi} \in \mathbb{R}^{m \times d_1 d_2}, \text{ where } \widetilde{\Phi} \text{ has i.i.d. standard sub-Gaussian entries,} \tag{2.96}$$

as it is summarized in the following lemma that is well-known in the compressed sensing literature.

Proposition 2.4.1 ([CP11, Theorem 2.3]). *Let $r \in \mathbb{N}$ and $0 < \delta < 1$. Then there exists constants $C(\delta)$ and C_1 such that if $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ is a sub-Gaussian random operator (e.g. as defined in (2.96)) and if*

$$m \geq C(\delta) r(d_1 + d_2),$$

then Φ fulfills the restricted isometry property of order r with constant $\delta_r = \delta$, with probability at least $1 - e^{-C_2 m}$.

However, we note that there are also important linear operators Φ for which the properties Definition 2.4.1 and Definition 2.4.2 do not hold, but which are generic enough to still allow for a local convergence analysis of `MatrixIRLS` (or other low-rank matrix recovery algorithms). In particular, this is case for the entrywise sampling operators $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ such that

$$\Phi(\mathbf{X})_\ell = \langle e_{i_\ell} e_{j_\ell}^*, \mathbf{X} \rangle = \mathbf{X}_{i_\ell, j_\ell} \qquad \text{for } \ell = 1, \ldots, m, \tag{2.97}$$

for any $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, where $(i_\ell, j_\ell) \in [d_1] \times [d_2]$ are double index pairs for all $\ell \in [m]$. Such a Φ can be used to define the well-known *low-rank matrix completion* problem [SRJ05, GNOT92, CR09, Rec11]. For Φ as in (2.97), it can be easily shown that the properties from Definition 2.4.1 and Definition 2.4.2 cannot hold if $m < d_1 d_2$ [CR09].

To cover a class of linear operators as large as possible in our convergence analysis, we will use Assumption 2.1 below. To state this assumption, we use the notion of the *tangent space* $T = T_{\mathbf{Z}}$ of the manifold of rank-r matrices at a given matrix

$$\mathbf{Z} = \begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma} & 0 \\ 0 & \mathbf{\Sigma}_\perp \end{bmatrix} \begin{bmatrix} \mathbf{V}^* \\ \mathbf{V}_\perp^* \end{bmatrix}, \tag{2.98}$$

where $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$ are matrices with orthonormal bases of the first r left resp. right singular vectors of $\mathbf{X}$ their columns, and $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ and $\mathbf{\Sigma}_\perp \in \mathbb{R}^{(d_1 - r) \times (d_2 - r)}$ being diagonal matrices with the (ordered) singular values of $\mathbf{X}$ on their diagonals. In particular, we define

$T = T_{\mathbf{Z}}$ as the linear subspace of $\mathbb{R}^{d_1 \times d_2}$ of dimension $r(d_1 + d_2 - r)$ such that

$$T_{\mathbf{Z}} := \{\mathbf{U}\mathbf{M}^* + \widetilde{\mathbf{M}}\mathbf{V}^* : \ \mathbf{M} \in \mathbb{R}^{d_2 \times r}, \ \widetilde{\mathbf{M}} \in \mathbb{R}^{d_1 \times r}\}. \tag{2.99}$$

Tangent spaces as defined in (2.99) play an important role in *Riemannian optimization* approaches for low-rank matrix recovery [Van13, WCCL16a], but also in the analysis of *convex optimization* approaches [CT10, Rec11].

Assumption 2.1. *There exists a constant* $c(r, d_1, d_2)$ *depending on* $r \in \mathbb{N}$, d_1 *and* d_2 *and a radius* $\xi > 0$ *such that the linear operator* $\Phi : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ *fulfills for a given matrix* $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ *that*

$$\|\eta\|_F \leq c(r, d_1, d_2)\|\mathcal{P}_{T_{\mathbf{Z}}^{\perp}}\eta\|_F \qquad \textit{for all } \eta \in \ker \Phi, \tag{2.100}$$

for any $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$ *such that*

$$\|\mathbf{Z} - \mathbf{X}\|_{S_\infty} \leq \xi\sigma_r(\mathbf{X}).$$

In (2.100), $\mathcal{P}_{T_{\mathbf{Z}}^{\perp}}$ *is the projection onto the orthogonal complement* $T_{\mathbf{Z}}^{\perp}$ *of the tangent space* $T_{\mathbf{Z}}$ *of the rank-r manifold at* $\mathbf{Z}$ *as defined in* (2.99).

We note that explicit formulas can be given for the action of $\mathcal{P}_{T_{\mathbf{Z}}}$ (the projection onto the tangent space T_Z) and $\mathcal{P}_{T_{\mathbf{Z}}^{\perp}}$ [Rec11, Section 2]: It holds that for any $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$,

$$\mathcal{P}_{T_{\mathbf{Z}}}(\mathbf{X}) = \mathbf{U}\mathbf{U}^*\mathbf{X} + \mathbf{X}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{X}\mathbf{V}\mathbf{V}^*$$

and

$$\mathcal{P}_{T_{\mathbf{Z}}^{\perp}}(\mathbf{X}) = (\mathbf{I} - \mathcal{P}_{T_{\mathbf{Z}}})(\mathbf{X}) = (\mathbf{I} - \mathbf{U}\mathbf{U}^*)\mathbf{X}(\mathbf{I} - \mathbf{V}\mathbf{V}^*).$$

Assumption 2.1 will be discussed and verified for measurement operators corresponding to the *matrix completion* setting and others in Section 2.5.

In the following theorem, we provide a local convergence statement for `MatrixIRLS` in the non-convex setting of $p < 1$ under Assumption 2.1, stating that the algorithm recovers a rank-r matrix $\mathbf{X}_0$ if the k-th iterate obtain an iterate $\mathbf{X}^{(k)}$ is close enough to $\mathbf{X}_0$.

Theorem 2.7 (Local Convergence with Superlinear Rate of order $2 - p$)**.** *Let* $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ *be a matrix of rank r, and let* $\Phi : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ *be a linear operator that fulfills Assumption 2.1 for* $\mathbf{X}_0$ *with constant* $c(r, d_1, d_2)$ *and radius* $\xi > 0$. *If the output matrix* $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ *of the k-th iteration of* `MatrixIRLS` *with inputs* Φ, $\mathbf{y} = \Phi(\mathbf{X}_0)$, $p \in [0, 1)$ *and* $\widetilde{r} = r$ *fulfills*

$$\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \leq \min(\xi, \zeta)\sigma_r(\mathbf{X}_0) \tag{2.101}$$

for some $0 < \zeta < 1$ *(which might depend on d and r), then*

$$\mathbf{X}^{(k)} \xrightarrow{k \to \infty} \mathbf{X}_0$$

if also the condition number $\kappa = \frac{\sigma_1(\mathbf{X}_0)}{\sigma_r(\mathbf{X}_0)}$ *of* X_0 *and* ζ *are sufficiently small, in particular, if* μ *from* (2.119) *fulfills*

$$\mu\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty}^{1-p} < 1. \tag{2.102}$$

Furthermore, the local convergence rate is of order $2 - p$ *in the sense that* μ *from* (2.119) *(which depends on d,* ζ *and* κ*) is such that*

$$\|\mathbf{X}^{(k+1)} - \mathbf{X}_0\|_{S_\infty} \leq \mu\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty}^{2-p}. \tag{2.103}$$

The second part of Theorem 2.7 states that in a neighborhood of a low-rank matrix $\mathbf{X}_0$ that defining the problem data $\mathbf{Y} = \Phi(\mathbf{X}_0)$, the algorithm `MatrixIRLS` converges to $\mathbf{X}_0$ with a *convergence rate* that is *superlinear of the order* $2 - p$, if Φ and $\mathbf{X}_0$ fulfill Assumption 2.1. This means that for the very non-convex logdet-objective corresponding to $p = 0$, the local convergence rate is even *quadratic*.

Remark 2.4.2. *It is interesting to compare Theorem 2.7 with a related result for an IRLS algorithm for the sparse vector recovery problem in* [DDFG10, Theorem 7.9]*. We observe that while the statement describes the observed rates of convergence very accurately (cf. Section 2.7.2), the experiments are much more optimistic about the size of the neighborhood enabling a fast rate of convergence than condition* (2.102)*. We leave a proof of a potentially larger radius of fast convergence to future work.*

In the rest of this section, we prove Theorem 2.7. In the proof of Theorem 2.7, we will use the following bound on perturbations of the singular value decomposition, which is originally due to [Wed72]. The result bounds the alignment of the subspaces spanned by the singular vectors of two matrices by their norm distance, given a gap between the first singular values of the one matrix and the last singular values of the other matrix that is sufficiently pronounced.

Lemma 2.4.2 (Wedin's bound [Ste06])**.** *Let* $\mathbf{X}$ *and* $\bar{\mathbf{X}}$ *be two matrices of the same size and their singular value decompositions*

$$\mathbf{X} = \begin{pmatrix} \mathbf{U} & \mathbf{U}_\perp \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma} & 0 \\ 0 & \mathbf{\Sigma}_\perp \end{pmatrix} \begin{pmatrix} \mathbf{V}^* \\ \mathbf{V}_\perp^* \end{pmatrix} \quad \textit{and} \quad \bar{\mathbf{D}} = \begin{pmatrix} \bar{\mathbf{U}} & \bar{\mathbf{U}}_\perp \end{pmatrix} \begin{pmatrix} \bar{\mathbf{\Sigma}} & 0 \\ 0 & \bar{\mathbf{\Sigma}}_\perp \end{pmatrix} \begin{pmatrix} \bar{\mathbf{V}}^* \\ \bar{\mathbf{V}}_\perp^* \end{pmatrix},$$

where the submatrices have the sizes of corresponding dimensions. Suppose that δ, α *satisfying* $0 < \delta \leq \alpha$ *are such that* $\alpha \leq \sigma_{\min}(\Sigma)$ *and* $\sigma_{\max}(\bar{\Sigma}_\perp) < \alpha - \delta$*. Then*

$$\|\bar{\mathbf{U}}_\perp^* \mathbf{U}\|_{S_\infty} \leq \sqrt{2}\frac{\|\mathbf{X} - \bar{\mathbf{X}}\|_{S_\infty}}{\delta} \textit{ and } \|\bar{\mathbf{V}}_\perp^* \mathbf{V}\|_{S_\infty} \leq \sqrt{2}\frac{\|\mathbf{X} - \bar{\mathbf{X}}\|_{S_\infty}}{\delta}. \tag{2.104}$$

As a first step towards the proof of Theorem 2.7, we show the following lemma.

Lemma 2.4.3. *Let* $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ *be a matrix of rank* r*, let* $(\mathbf{X}^{(k)}, \epsilon_k = \sigma_{r+1}(\mathbf{X}^{(k)}))$ *be the output sequence of Algorithm 1 for parameters* $\Phi, \mathbf{y} = \Phi(\mathbf{X}_0), r$ *and* $0 \leq p \leq 1$ *at the* k*-th iteration, and let* (2.100) *be fulfilled for* $\mathbf{X}^{(k)}$ *and the corresponding tangent space onto the manifold of rank-*r *matrices* $T_k := T_{\mathbf{X}^{(k)}}$ *with constant* $c(r, d_1, d_2)$*. Then*

$$\|\mathbf{X}^{(k+1)} - \mathbf{X}_0\|_{S_\infty} \leq c(r, d_1, d_2)^2 \epsilon_k^{2-p} \|W^{(k)}(\mathbf{X}_0)\|_{S_1}, \tag{2.105}$$

where $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ *is the weight operator* (2.57) *corresponding to* $\mathbf{X}^{(k)}$*.*

Proof of Lemma 2.4.3. Let $\eta^{(k+1)} := \mathbf{X}^{(k+1)} - \mathbf{X}_0$. From (2.100), it follows that

$$\|\eta^{(k+1)}\|_{S_\infty}^2 \leq \|\eta^{(k+1)}\|_F^2 \leq c(r, d_1, d_2)^2 \|\mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\|_F^2 \tag{2.106}$$

We now recall the definition of the weight operator $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ from Definition 2.3.4, and if

$$\mathbf{X}^{(k)} = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^* = \begin{bmatrix} \mathbf{U}^{(k)} & \mathbf{U}_\perp^{(k)} \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}^{(k)} & 0 \\ 0 & \mathbf{\Sigma}_\perp^{(k)} \end{bmatrix} \begin{bmatrix} \mathbf{V}^{(k)*} \\ \mathbf{V}_\perp^{(k)*} \end{bmatrix} \tag{2.107}$$

is a singular value decomposition with $\mathbf{U}^{(k)} \in \mathbb{R}^{d_1 \times r}$, $\mathbf{U}_\perp^{(k)} \in \mathbb{R}^{d_1 \times (d_1 - r)}$, $\mathbf{V}^{(k)} \in \mathbb{R}^{d_1 \times r}$, $\mathbf{V}_\perp^{(k)} \in$

$\mathbb{R}^{d_2\times(d_2-r)}$, we have that

$$\begin{aligned}\langle \mathbf{Z}, W^{(k)}(\mathbf{Z})\rangle &= \langle \mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k, \mathbf{H}_1^{(k)}\circ \widetilde{S}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k) + \mathbf{H}_2^{(k)}\circ \widetilde{T}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\rangle \\ &= \langle \widetilde{S}\circ(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k), \mathbf{H}_1^{(k)}\circ\widetilde{S}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\rangle + \langle \widetilde{T}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k), \mathbf{H}_2^{(k)}\circ\widetilde{T}(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\rangle\end{aligned} \tag{2.108}$$

where $\widetilde{S}$, $\widetilde{T}$, $\mathbf{H}_1^{(k)} \in \mathbb{R}^{d_1\times d_2}$ and $\mathbf{H}_2^{(k)} \in \mathbb{R}^{d_1\times d_2}$ are as in Definition 2.3.4.

If $\mathbf{Z} = \mathcal{P}_{T_k^\perp}(\eta^{(k+1)}) \in T_k^\perp$, we know that $\mathbf{U}^{(k)*}\mathbf{Z} = 0$ and $\mathbf{Z}\mathbf{V}^{(k)} = 0$, and therefore

$$\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k = \begin{bmatrix}\mathbf{U}^{(k)*} \\ \mathbf{U}_\perp^{(k)*}\end{bmatrix} \mathbf{Z} \begin{bmatrix}\mathbf{V}^{(k)} & \mathbf{V}_\perp^{(k)}\end{bmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{U}_\perp^{(k)*}\mathbf{Z}\mathbf{V}_\perp^{(k)}\end{pmatrix}$$

with $\mathbf{U}_\perp^{(k)*}\mathbf{Z}\mathbf{V}_\perp^{(k)} \in \mathbb{R}^{(d_1-r)\times(d_2-r)}$.

By assumption of Lemma 2.4.3, we know that $\epsilon_k = \sigma_{r+1}(\mathbf{X}^{(k)})$, which means that $R_k := |\{i \in [d] : \sigma_i(\mathbf{X}^{(k)}) > \epsilon\}| \leq r$, and therefore $(\mathbf{H}_1^{(k)})_{ij} = (\mathbf{H}_2^{(k)})_{ij} = \epsilon_k^{p-2}$ for all $i, j > r$. This entails with (2.108) that

$$\begin{aligned}&\langle \mathcal{P}_{T_k^\perp}(\eta^{(k+1)}), W^{(k)}(\mathcal{P}_{T_k^\perp}(\eta^{(k+1)}))\rangle \\ &= \epsilon_k^{p-2}\left(\left\langle\widetilde{S}\left(\mathbf{U}_k^*\mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\mathbf{V}_k\right), \widetilde{S}\left(\mathbf{U}_k^*\mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\mathbf{V}_k\right)\right\rangle + \left\langle\widetilde{T}\left(\mathbf{U}_k^*\mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\mathbf{V}_k\right), \widetilde{T}\left(\mathbf{U}_k^*\mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\mathbf{V}_k\right)\right\rangle\right) \\ &= \epsilon_k^{p-2}\langle \mathbf{U}_k^*\mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\mathbf{V}_k, \mathbf{U}_k^*\mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\mathbf{V}_k\rangle \\ &= \epsilon_k^{p-2}\langle \mathcal{P}_{T_k^\perp}(\eta^{(k+1)}), \mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\rangle = \epsilon_k^{p-2}\|\mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\|_F^2,\end{aligned}$$

using the cyclicity of the trace and the fact that $\mathbf{U}_k$ and $\mathbf{V}_k$ are orthonormal matrices.

Inserting this into (2.106), we obtain

$$\begin{aligned}\|\eta^{(k+1)}\|_{S_\infty}^2 &\leq c(r,d_1,d_2)^2\epsilon_k^{2-p}\left\langle \mathcal{P}_{T_k^\perp}(\eta^{(k+1)}), W^{(k)}(\mathcal{P}_{T_k^\perp}(\eta^{(k+1)}))\right\rangle \\ &\leq c(r,d_1,d_2)^2\epsilon_k^{2-p}\left\langle \eta^{(k+1)}, W^{(k)}(\eta^{(k+1)})\right\rangle,\end{aligned} \tag{2.109}$$

where the last inequality holds since $W^{(k)}$ is positive definite and since $\left\langle\mathcal{P}_{T_k^\perp}(\eta^{(k+1)}), W^{(k)}(\mathcal{P}_{T_k}(\eta^{(k+1)}))\right\rangle = 0$ due to the orthogonality of T_k and $T_k^\perp$. Due to Lemma 2.4.1, we know that the new iterate $\mathbf{X}^{(k+1)}$ fulfills

$$0 = \langle W^{(k)}(\mathbf{X}^{(k+1)}), \eta^{(k+1)}\rangle = \langle \widetilde{W}^{(k)}(\eta^{(k+1)} + \mathbf{X}_0), \eta^{(k+1)}\rangle,$$

and therefore

$$\left\langle \eta^{(k+1)}, W^{(k)}(\eta^{(k+1)})\right\rangle = -\left\langle \widetilde{W}^{(k)}(\mathbf{X}_0), \eta^{(k+1)}\right\rangle \leq \|W^{(k)}(\mathbf{X}_0)\|_{S_1}\|\eta^{(k+1)}\|_{S_\infty},$$

using Hölder's inequality for Schatten-p (quasi-)norms [GGK00, Theorem 11.2]. Dividing (2.109) by $\|\eta^{(k+1)}\|_{S_\infty}$ concludes the proof of Lemma 2.4.3. □

In order to obtain a fast local convergence rate, it is crucial to bound $\|W^{(k)}(\mathbf{X}_0)\|_{S_1}$. For this, we split $\|W^{(k)}(\mathbf{X}_0)\|_{S_1}$ into three parts and estimate the parts separately by using the classical singular subspace perturbation result of Lemma 2.4.2.

Lemma 2.4.4. *Let $W^{(k)} : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ be the weight operator (2.57) corresponding to $\mathbf{X}^{(k)}$,* *let* $\epsilon_k = \sigma_{r+1}(\mathbf{X}^{(k)}) = \sigma_r^{(k)}$ *and* $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ *be a rank-r matrix. Then*

$$\left\| W^{(k)}(\mathbf{X}_0) \right\|_{S_1} \leq r(1-\zeta)^{p-2}\sigma_r(\mathbf{X}_0)^{p-1}\left(1 + 4\frac{\|\eta^{(k)}\|_{S_\infty}}{\epsilon_k}\frac{\sigma_1(\mathbf{X}_0)}{\sigma_r(\mathbf{X}_0)} + 2\frac{\|\eta^{(k)}\|_{S_\infty}^2}{\sigma_r(\mathbf{X}_0)^p\epsilon_k^{2-p}}\frac{\sigma_1(\mathbf{X}_0)}{\sigma_r(\mathbf{X}_0)}\right).$$

Proof. Indeed, using the decompositions

$$\mathbf{H}_1^{(k)} = \begin{bmatrix} (\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}} & (\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}_\perp} \\ (\mathbf{H}_1^{(k)})_{\mathbf{U}_\perp,\mathbf{V}} & (\mathbf{H}_1^{(k)})_{\mathbf{U}_\perp,\mathbf{V}_\perp} \end{bmatrix}$$

and

$$\mathbf{H}_2^{(k)} = \begin{bmatrix} (\mathbf{H}_2^{(k)})_{\mathbf{U},\mathbf{V}} & (\mathbf{H}_2^{(k)})_{\mathbf{U},\mathbf{V}_\perp} \\ (\mathbf{H}_2^{(k)})_{\mathbf{U}_\perp,\mathbf{V}} & (\mathbf{H}_2^{(k)})_{\mathbf{U}_\perp,\mathbf{V}_\perp} \end{bmatrix}$$

of the weighting factor matrices of Definition 2.3.4 with $(\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}}, (\mathbf{H}_2^{(k)})_{\mathbf{U},\mathbf{V}} \in \mathbb{R}^{r \times r}$, we note that

$$\begin{aligned} \max_{i\in[r],j\in[r]}((\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}})_{ij} &= \max_{i\in[r],j\in[r]}(\mathbf{H}_1^{(k)})_{ij} \leq \begin{cases} \frac{1}{\sigma_r^{(k)}\cdot\sigma_r^{(k)}} & \text{if } p = 0, \\ \left(\frac{(\sigma_r^{(k)})^p + (\sigma_r^{(k)})^p}{2}\right)^{1-2/p} & \text{if } p > 0, \end{cases} = (\sigma_r^{(k)})^{p-2}, \\ \max_{i\in[r],j\in[r]}((\mathbf{H}_2^{(k)})_{\mathbf{U},\mathbf{V}})_{ij} &= \max_{i\in[r],j\in[r]}(\mathbf{H}_2^{(k)})_{ij} \leq \frac{(\sigma_r^{(k)})^{p-1} + (\sigma_r^{(k)})^{p-1}}{\sigma_r^{(k)} + \sigma_r^{(k)}} = (\sigma_r^{(k)})^{p-2}, \end{aligned} \tag{2.110}$$

and

$$\begin{aligned} &\max\left(\max_{i,j}((\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}_\perp})_{ij}, \max_{i,j}((\mathbf{H}_1^{(k)})_{\mathbf{U}_\perp,\mathbf{V}})_{ij}\right) = \max_{i\in[r],r+1\leq j\leq d_2}(\mathbf{H}_1^{(k)})_{ij} \leq \left(\frac{(\sigma_r^{(k)})^p + \epsilon_k^p}{2}\right)^{1-2/p} \\ &\leq \sqrt{(\sigma_r^{(k)})^{p-2}\epsilon_k^{p-2}} \leq 2(\sigma_r^{(k)})^{p/2-1}\epsilon_k^{p/2-1}, \end{aligned} \tag{2.111}$$

using the the fact that $\left(\frac{(\sigma_r^{(k)})^p + \epsilon_k^p}{2}\right)^{1-2/p}$ is the $q = \frac{p}{p-2}$-th power mean of $(\sigma_r^{(k)})^{p-2}$ and ϵ_k^{p-2}, $\sqrt{(\sigma_r^{(k)})^{p-2}\epsilon_k^{p-2}}$ the geometric mean ($q = 0$) and the *power mean inequality* [Bul03, Section III.3, Theorem 1] in the second inequality. Furthermore,

$$\begin{aligned} &\max\left(\max_{i,j}((\mathbf{H}_2^{(k)})_{\mathbf{U},\mathbf{V}_\perp})_{ij}, \max_{i,j}((\mathbf{H}_2^{(k)})_{\mathbf{U}_\perp,\mathbf{V}})_{ij}\right) = \max_{i\in[r],r+1\leq j\leq d_2}(\mathbf{H}_2^{(k)})_{ij} \leq \frac{(\sigma_r^{(k)})^{p-1} + \epsilon_k^{p-1}}{\sigma_r^{(k)} + \epsilon_k} \\ &\leq \frac{2\epsilon_k^{p-1}}{\sigma_r^{(k)}} \leq 2(\sigma_r^{(k)})^{p/2-1}\epsilon^{p/2-1} \end{aligned} \tag{2.112}$$

and

$$\max_{i,j}((\mathbf{H}_1^{(k)})_{\mathbf{U}_\perp,\mathbf{V}_\perp})_{ij} = \max_{i,j}((\mathbf{H}_2^{(k)})_{\mathbf{U}_\perp,\mathbf{V}_\perp})_{ij} = \epsilon_k^{p-2}. \tag{2.113}$$

In view of these entrywise bounds on the submatrices of $\mathbf{H}_1^{(k)}$ and $\mathbf{H}_2^{(k)}$, we calculate, using the definition of the weight operator (2.57) and (2.107), that

$$\begin{aligned}&\left\|W^{(k)}(\mathbf{X}_0)\right\|_{S_1} \leq \left\|\mathbf{U}^{(k)}[(\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}} \circ \widetilde{S}(\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)}) + (\mathbf{H}_2^{(k)})_{\mathbf{U},\mathbf{V}} \circ \widetilde{T}(\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)})]\mathbf{V}^{(k)*}\right\|_{S_1} \\ &+ \left\|\mathbf{U}_k\left[\mathbf{H}_1^{(k)} \circ \widetilde{S}\left(\begin{bmatrix} 0 & \mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)} \\ \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)} & 0 \end{bmatrix}\right) + \mathbf{H}_2^{(k)} \circ \widetilde{T}\left(\begin{bmatrix} 0 & \mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}) \\ \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)} & 0 \end{bmatrix}\right)\right]\mathbf{V}_k^*\right\|_{S_1} \\ &+ \left\|\mathbf{U}_\perp^{(k)}[(\mathbf{H}_1^{(k)})_{\mathbf{U}_\perp,\mathbf{V}_\perp} \circ (\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)})]\mathbf{V}_\perp^{(k)*}\right\|_{S_1} =: (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}).\end{aligned}$$

We now bound the terms (I), (II) and (III) separately.

First, we see that

$$\begin{aligned}(\mathrm{I}) &= \left\|(\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}} \circ \widetilde{S}(\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)}) + (\mathbf{H}_2^{(k)})_{\mathbf{U},\mathbf{V}} \circ \widetilde{T}(\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)})\right\|_{S_1} \\ &\leq \sqrt{r}\left\|(\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}} \circ \widetilde{S}(\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)}) + (\mathbf{H}_2^{(k)})_{\mathbf{U},\mathbf{V}} \circ \widetilde{T}(\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)})\right\|_F \\ &\leq \sqrt{r}\left\|(\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}} \circ \widetilde{S}(\mathbf{U}^{(k)*}\mathbf{X}^{(k)}\mathbf{V}^{(k)}) + (\mathbf{H}_2^{(k)})_{\mathbf{U},\mathbf{V}} \circ \widetilde{T}(\mathbf{U}^{(k)*}\mathbf{X}^{(k)}\mathbf{V}^{(k)})\right\|_F \\ &+ \sqrt{r}\left\|(\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}} \circ \widetilde{S}(\mathbf{U}^{(k)*}\eta^{(k)}\mathbf{V}^{(k)}) + (\mathbf{H}_2^{(k)})_{\mathbf{U},\mathbf{V}} \circ \widetilde{T}(\mathbf{U}^{(k)*}\eta^{(k)}\mathbf{V}^{(k)})\right\|_F \\ &\leq \sqrt{r}\left\|(\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}} \circ \mathbf{\Sigma}^{(k)}\right\|_F + \sqrt{r}(\sigma_r^{(k)})^{p-2}\|\mathbf{U}^{(k)*}\eta^{(k)}\mathbf{V}^{(k)}\|_F,\end{aligned}$$

where we used the Cauchy-Schwartz inequality in the first inequality, the notation $\eta^{(k)} = \mathbf{X}^{(k)} - \mathbf{X}_0$ and the triangle inequality in the second inequality and (2.110) in the third inequality.

Since

$$\left\|(\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}} \circ \mathbf{\Sigma}^{(k)}\right\|_F = \left(\sum_{i=1}^r (\sigma_i^{(k)})^{2p-2}\right)^{1/2} \leq \sqrt{r}(\sigma_r^{(k)})^{p-1}$$

and

$$\|\mathbf{U}^{(k)*}\eta^{(k)}\mathbf{V}^{(k)}\|_F \leq \sqrt{r}\|\mathbf{U}^{(k)*}\eta^{(k)}\mathbf{V}^{(k)}\|_{S_\infty} \leq \sqrt{r}\|\eta^{(k)}\|_{S_\infty} \leq \sqrt{r}\zeta\sigma_r(\mathbf{X}_0)$$

from assumption (2.138), it follows then that

$$(\mathrm{I}) \leq r(\sigma_r^{(k)})^{p-2}\left(\sigma_r^{(k)} + \zeta\sigma_r(\mathbf{X}_0)\right).$$

We can use the proximity assumption (2.138) further to get rid of the dependence on k in the bound, as

$$\sigma_r(\mathbf{X}_0) = \sigma_r(\mathbf{X}^{(k)} - \eta^{(k)}) \leq \sigma_r^{(k)} + \sigma_1(\eta^{(k)}) = \sigma_r^{(k)} + \|\eta^{(k)}\|_{S_\infty} \leq \sigma_r^{(k)} + \zeta\sigma_r(\mathbf{X}_0),$$

using $\sigma_{i+j-1}(\mathbf{A}) \leq \sigma_i(\mathbf{A}+\mathbf{B}) + \sigma_j(\mathbf{B})$ for any i, j (cf. [HJ91, Theorem 3.3.16]) with $\mathbf{A}+\mathbf{B} = X^{(k)} - \eta^{(k)}$ and $\mathbf{B} = \eta^k$ so that

$$\sigma_r^{(k)} \geq (1-\zeta)\sigma_r(\mathbf{X}_0), \tag{2.114}$$

and hence

$$(\mathrm{I}) \leq r(\sigma_r(\mathbf{X}_0))^{p-2}(1-\zeta)^{p-2}\left(\sigma_r(\mathbf{X}_0)(1-\zeta) + \zeta\sigma_r(\mathbf{X}_0)\right) = r(1-\zeta)^{p-2}\sigma_r(\mathbf{X}_0)^{p-1}. \tag{2.115}$$

For the term (II), we calculate that

$$\begin{aligned}
\text{(II)} &\leq \sqrt{2r}\left\|\mathbf{H}_1^{(k)} \circ \widetilde{S}\left(\begin{bmatrix} 0 & \mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)} \\ \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)} & 0 \end{bmatrix}\right) + \mathbf{H}_2^{(k)} \circ \widetilde{T}\left(\begin{bmatrix} 0 & \mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}) \\ \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)} & 0 \end{bmatrix}\right)\right\|_F \\
&\leq 2\sqrt{2r}(\sigma_r^{(k)})^{p/2-1}\epsilon_k^{p/2-1}\left(\left\|\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\right\|_F + \left\|\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)}\right\|_F\right) \\
&\leq 2\sqrt{2r}(\sigma_r^{(k)})^{p/2-1}\epsilon_k^{p/2-1}\left(\|\mathbf{U}^{(k)*}\mathbf{U}_0\mathbf{\Sigma}_0\|_F\|\mathbf{V}_0^*\mathbf{V}_\perp^{(k)}\|_{S_\infty} + \|\mathbf{U}_\perp^{(k)*}\mathbf{U}_0\|_{S_\infty}\|\mathbf{\Sigma}_0\mathbf{V}_0^*\mathbf{V}^{(k)}\|_F\right),
\end{aligned}$$

using the singular value decomposition $\mathbf{X}_0 = \mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^*$ of the rank-r matrix $\mathbf{X}_0$ with $\mathbf{U}_0 \in \mathbb{R}^{d_1 \times r}$, $\mathbf{V}_0 \in \mathbb{R}^{d_2 \times r}$. This allows us to use the singular subspace perturbation result of Lemma 2.4.2, so that $\|\mathbf{V}_0^*\mathbf{V}_\perp^{(k)}\|_{S_\infty}$ and $\|\mathbf{U}_\perp^{(k)*}\mathbf{U}_0\|_{S_\infty}$ can compensate for the negative power of the ϵ_k, avoiding a blow-up of term (II): Indeed, using Lemma 2.4.2 with $\mathbf{X} = \mathbf{X}_0$, $\bar{\mathbf{X}} = \mathbf{X}^{(k)}$, $\alpha = \sigma_r(\mathbf{X}_0)$ and $\delta = (1-\zeta)\sigma_r(\mathbf{X}_0)$ results in

$$\max(\|\mathbf{V}_0^*\mathbf{V}_\perp^{(k)}\|_{S_\infty}, \|\mathbf{U}_\perp^{(k)*}\mathbf{U}_0\|_{S_\infty}) \leq \frac{\sqrt{2}\|\eta^{(k)}\|_{S_\infty}}{(1-\zeta)\sigma_r(\mathbf{X}_0)},$$

and since $\|\mathbf{U}^{(k)*}\mathbf{U}_0\mathbf{\Sigma}_0\|_F \leq \|\mathbf{\Sigma}_0\|_F \leq \sqrt{r}\sigma_1(\mathbf{X}_0)$, $\|\mathbf{\Sigma}_0\mathbf{V}_0^*\mathbf{V}^{(k)}\|_F \leq \sqrt{r}\sigma_1(\mathbf{X}_0)$, and $\epsilon_k^{p/2} \leq (\sigma_r^{(k)})^{p/2}$, we obtain with (2.114) that

$$\text{(II)} \leq 4r(1-\zeta)^{p-2}\sigma_r(\mathbf{X}_0)^{p-1}\frac{\|\eta^{(k)}\|_{S_\infty}}{\epsilon_k}\frac{\sigma_1(\mathbf{X}_0)}{\sigma_r(\mathbf{X}_0)}. \tag{2.116}$$

It remains to bound the last term (III). For (III), we can use subspace perturbation lemma *twice* in the same summand and (2.113), such that

$$\begin{aligned}
\text{(III)} &= \left\|(\mathbf{H}_1^{(k)})_{\mathbf{U}_\perp,\mathbf{V}_\perp} \circ (\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)})\right\|_{S_1} = \epsilon_k^{p-2}\|\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\|_{S_1} \leq \sqrt{r}\epsilon_k^{p-2}\|\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\|_F \\
&\leq \sqrt{r}\epsilon_k^{p-2}\|\mathbf{U}_\perp^{(k)*}\mathbf{U}_0\|_{S_\infty}\|\mathbf{\Sigma}_0\|_F\|\mathbf{V}_0^*\mathbf{V}_\perp^{(k)}\|_{S_\infty} \\
&\leq \sqrt{r}\epsilon_k^{p-2}\frac{\sqrt{2}\|\eta^{(k)}\|_{S_\infty}}{(1-\zeta)\sigma_r(\mathbf{X}_0)}\sqrt{r}\sigma_1(\mathbf{X}_0)\frac{\sqrt{2}\|\eta^{(k)}\|_{S_\infty}}{(1-\zeta)\sigma_r(\mathbf{X}_0)} = 2r(1-\zeta)^{-2}\epsilon_k^p\sigma_r(\mathbf{X}_0)^{-1}\frac{\|\eta^{(k)}\|_{S_\infty}^2}{\epsilon_k^2}\frac{\sigma_1(\mathbf{X}_0)}{\sigma_r(\mathbf{X}_0)}.
\end{aligned} \tag{2.117}$$

Combining (2.115)–(2.117) finally yields the statement of Lemma 2.4.4. □

We can now put Lemma 2.4.3 and Lemma 2.4.4 together to prove the local convergence statement of Theorem 2.7.

Proof of Theorem 2.7. Let $k = k_0$ and $\mathbf{X}^{(k)}$ be the k-th iterate of `MatrixIRLS` with the parameter stated in Theorem 2.7. Then, by Lemmas 2.4.3 and 2.4.4,

$$\|\mathbf{X}^{(k+1)} - \mathbf{X}_0\|_{S_\infty} \leq c(r, d_1, d_2)^2\epsilon_k^{2-p}r(1-\zeta)^{p-2}\sigma_r(\mathbf{X}_0)^{p-1}\left(1 + 4\frac{\|\eta^{(k)}\|_{S_\infty}}{\epsilon_k}\frac{\sigma_1(\mathbf{X}_0)}{\sigma_r(\mathbf{X}_0)} + 2\frac{\|\eta^{(k)}\|_{S_\infty}^2}{\sigma_r(\mathbf{X}_0)^p\epsilon_k^{2-p}}\frac{\sigma_1(\mathbf{X}_0)}{\sigma_r(\mathbf{X}_0)}\right),$$

and, since $\|\eta^{(k)}\|_{S_\infty} \leq \zeta\sigma_r(\mathbf{X}_0)$ due to (2.138) and since

$$\epsilon_k \leq \sigma_{r+1}(\mathbf{X}^{(k)}) = \|\mathbf{X}^{(k)} - \mathbf{X}_r^{(k)}\|_{S_\infty} \leq \|\mathbf{X}^{(k)} - \mathbf{X}_0^{(k)}\|_{S_\infty} = \|\eta^{(k)}\|_{S_\infty},$$

where $\mathbf{X}_r^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ denotes the best rank-r approximation of $\mathbf{X}^{(k)}$ in any unitarily invariant norm, we obtain

$$\|\mathbf{X}^{(k+1)} - \mathbf{X}_0\|_{S_\infty} \leq \frac{c(r,d_1,d_2)^2 r\,(1+4\kappa+2\zeta^p\kappa))}{(1-\zeta)^{2-p}\sigma_r(\mathbf{X}_0)^{1-p}}\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty}^{2-p}, \tag{2.118}$$

using the notation $\kappa := \sigma_1(\mathbf{X}_0)/\sigma_r(\mathbf{X}_0)$.

Thus, the statement of Theorem 2.7 is shown with

$$\mu := \frac{c(r,d_1,d_2)^2 r\,(1+4\kappa+2\zeta^p\kappa)}{(1-\zeta)^{2-p}\sigma_r(\mathbf{X}_0)^{1-p}}. \tag{2.119}$$

□

As already mentioned in the discussion of existing approaches in Section 2.3.1, a statement about superlinear convergence of the IRLS algorithms low-rank matrix recovery variants of [MF12, FRW11, LXY13] as Theorem 2.7 has not been shown, and in fact, is not observed numerically, as will be investigated in Section 2.7.2. For algorithm of [KS18] which combines the weights of `IRLS-col` and `IRLS-row` as their harmonic mean (2.39), a comparable statement as Theorem 2.7 was shown, relying a stronger Schatten-p null space property as defined inDefinition 2.4.2 instead of Assumption 2.1.

Remark 2.4.3. *We note that the weight operators of the previously studied IRLS approaches* `IRLS-col` *and* `IRLS-row` [FRW11, MF12] *expressed in a notation similar to* (2.39) *or Definition 2.3.4 such that*

$$W^{(k)}(\mathbf{X}) = \mathbf{U}_k\left[\mathbf{H}^{(k)} \circ (\mathbf{U}_k^*\mathbf{X}\mathbf{V}_k)\right]\mathbf{V}_k^*$$

where $\mathbf{H}^{(k)} \in \mathbb{R}^{d_1\times d_2}$ is such that

$$(\mathbf{H}^{(k)})_{ij} = \left(\sigma_i(\mathbf{X}^{(k)})^2 + \epsilon_k^2\right)^{p/2-1}, \textit{ for any } i \in [d_1], j \in [d_2] \tag{2.120}$$

for `IRLS-col` *and such that*

$$(\mathbf{H}^{(k)})_{ij} = \left(\sigma_j(\mathbf{X}^{(k)})^2 + \epsilon_k^2\right)^{p/2-1}, \textit{ for any } i \in [d_1], j \in [d_2] \tag{2.121}$$

for `IRLS-row`.

It turns out that for both `IRLS-col` *and* `IRLS-row`, *there are some entries of $\mathbf{H}^{(k)}$ that are* too large *to allow for superlinear convergence rates. In fact, if we consider the tangent space*

$$T_k = \{\mathbf{U}^{(k)}\mathbf{M}^* + \widetilde{\mathbf{M}}\mathbf{V}^{(k)*} : \mathbf{M} \in \mathbb{R}^{d_2\times r}, \widetilde{\mathbf{M}} \in \mathbb{R}^{d_1\times r}\}$$

onto the rank-r manifold at iterate $\mathbf{X}^{(k)}$ ($\mathbf{U}^{(k)}$ and $\mathbf{V}^{(k)}$ are, as above, the matrices of first r left resp. right singular vectors) or, which is, from a different point of view, the direct sum of the spaces spanned by the first r left singular vectors and the first r right singular vectors, there are indices of $\mathbf{H}^{(k)}$ corresponding to some parts of T_k with simply too large *entries resp. weights.*

To be precise, if we tried to replicate the proof of Lemma 2.4.4 by a decomposition of the term $\|W^{(k)}(\mathbf{X}_0)\|_{S_1}$ such that

$$\left\|W^{(k)}(\mathbf{X}_0)\right\|_{S_1} \leq \left\|\mathbf{U}^{(k)}[(\mathbf{H}^{(k)})_{\mathbf{U},\mathbf{V}} \circ (\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)})]\mathbf{V}^{(k)*}\right\|_{S_1} + \left\|\mathbf{U}_k\left[\mathbf{H}^{(k)} \circ \begin{pmatrix} 0 & \mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)} \\ \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)} & 0\end{pmatrix}\right]\mathbf{V}_k^*\right\|_{S_1}$$
$$+ \left\|\mathbf{U}_\perp^{(k)}[(\mathbf{H}^{(k)})_{\mathbf{U}_\perp,\mathbf{V}_\perp} \circ (\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)})]\mathbf{V}_\perp^{(k)*}\right\|_{S_1} =: \textit{(I)} + \textit{(II)} + \textit{(III)},$$

some of the entries of $\mathbf{H}^{(k)}$ *active in term (II) are too large, if* $\mathbf{H}^{(k)}$ *is as in* (2.120) *or* (2.121)*. A similar situation would arise if* $\mathbf{H}_1^{(k)}$ *of* (2.58) *was defined as any q-power mean of* $\widetilde{\sigma}_i^{(k)} = \max(\sigma_i^{(k)}, \epsilon)^{p-2}$ *and* $\widetilde{\sigma}_j^{(k)} = \max(\sigma_j^{(k)}, \epsilon)^{p-2}$ of power $q > 0$*, i.e., of any power mean that is larger than the geometric mean. In particular, by this argument, it can be seen that the straightforward choice of* arithmetic mean *of* $\widetilde{\sigma}_i^{(k)}$ *and* $\widetilde{\sigma}_j^{(k)}$ *would* ***not*** *lead to a superlinear convergence rate for* $p < 1$.

2.5 Measurement Models for `MatrixIRLS` with Guarantees From an Optimal Number of Samples

In this section, we will verify Assumption 2.1 for several measurement models that arise in well-studied applications, and corresponding linear operators $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$. This makes sure that Theorem 2.7 is applicable for these measurement models.

By doing this, in particular, we show that Theorem 2.7 holds for

- operators Φ with dense sub-Gaussian matrices with independent entries as covered by Proposition 2.4.1,
- symmetric rank-one measurements as in the *phase retrieval* setting,
- entry-wise measurements that correspond to a *matrix completion* setting,

for a sample complexity or number of measurements m that is *optimal* up to dimension-free constants.

2.5.1 Sub-Gaussian Measurements

The in some sense *most generic* linear operators $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ for our purposes are dense, so called *sub-Gaussian* operators.

Definition 2.5.1 ([Ver18, Chapter 2.5])**.** *We call a linear operator* $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ sub-Gaussian *if its action on matrices* $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ *is such that*

$$\Phi(\mathbf{X})_\ell = \langle \mathbf{A}_\ell, \mathbf{X} \rangle$$

for all $\ell \in [m]$*, where* $(\mathbf{A}_\ell)_{\ell=1}^m$ *are independent* $(d_1 \times d_2)$*-dimensional random matrices whose entries* $((\mathbf{A}_\ell)_{ij})_{ij}$*,* $i \in [d_1], j \in [d_2]$*, are independent and fulfill*

$$\mathbb{P}(|(\mathbf{A}_\ell)_{ij}| \geq t) \leq 2\exp(-t^2/K^2) \tag{2.122}$$

for some constant $K > 0$.

We note that an important special case of sub-Gaussian measurement operators are *Gaussian* operators, whose entrywise random variables $(\mathbf{A}_\ell)_{ij}$ are centered Gaussian random variables. Low-rank recovery problems with measurement operators Φ as in Definition 2.5.1 have been widely studied in the literature [RFP10, JMD10, TBS+16, PKCS18, ZLTW18].

They have been mentioned already in (2.96) in Section 2.4.2, and they are a popular object of study because they fulfill a *restricted isometry property* (see Definition 2.4.1), with high probability, already for a number measurements m that is

$$m \approx Cr(d_1 + d_2),$$

for some constant $C > 1$, which is close to the number of *degrees of freedom* $\deg_f$ of a rank-r matrix of dimension $(d_1 \times d_2)$, see the discussion of Section 2.1.1. That statement corresponds to Proposition 2.4.1.

In the following, we verify that for sub-Gaussian operators Φ, Assumption 2.1 is fulfilled with high probability for a number of measurements m that are at least a constant multiple of the number of degrees of freedom of the low-rank model.

Proposition 2.5.1. *Let $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ be a sub-Gaussian operator as defined in Definition 2.5.1, let $r \in \mathbb{N}$. Then there exists a constant $C > 1$ such that if*

$$m \geq Cr(d_1 + d_2),$$

the following statements hold with high probability.

1. *Φ fulfills the strong Schatten-1 null space property of Definition 2.4.2 of order r with a constant $0 < \gamma_r \leq 1$.*
2. *Φ fulfills (2.100) of Assumption 2.1 for all tangent spaces $T = T_{\mathbf{Z}}$ onto the rank-r manifold with constant $c(r, d_1, d_2)$ such that*

$$c(r, d_1, d_2) \leq \sqrt{\frac{2d}{r}}.$$

Proof. Fixing $\delta < \frac{4}{\sqrt{41}}$, we know from Proposition 2.4.1 that if

$$m \geq 2C(\delta)r(d_1 + d_2),$$

Φ fulfills the restricted isometry property of Definition 2.4.1 of order $2r$ with restricted isometry constant $\delta_{2r} = \delta$ with high probability. By applying an adaptation of [FR13, Theorem 6.13] to the rank case, it follows that then

$$\|\eta_r\|_F < \frac{\gamma_r}{\sqrt{r}} \|\eta_c\|_{S_1} \tag{2.123}$$

for all matrices $\eta \in \ker \Phi$, if η_r is a best rank-r approximation of $\eta \in \mathbb{R}^{d_1 \times d_2}$ and $\eta_c = \eta - \eta_r$, with a constant of $\gamma_r = \frac{\delta}{\sqrt{1-\delta^2} - \delta/4} < 1$, i.e., Φ fulfills the strong Schatten-1 null space property of order r with this constant $\gamma_r < 1$.[4] This shows the first statement.

Furthermore, we can then calculate that

$$\begin{aligned} \|\eta\|_F &\leq \|\eta_r\|_F + \|\eta_c\|_F \leq \frac{\gamma_r}{\sqrt{r}} \|\eta_c\|_{S_1} + \|\eta_c\|_F \\ &\leq \left(\frac{\gamma_r \sqrt{d-r} + \sqrt{r}}{\sqrt{r}} \right) \|\eta_c\|_F \leq \sqrt{\frac{2d}{r}} \|\eta_c\|_F, \end{aligned}$$

where we used (2.123) in the second inequality and the Cauchy-Schwarz inequality on the vector of singular values of η_c in the third inequality.

Finally, we note that

$$\eta_c = \mathcal{P}_{T_{\eta^\perp}}(\eta) = (\mathbf{I} - \mathbf{U}_\eta \mathbf{U}_\eta^*) \eta (\mathbf{I} - \mathbf{V}_\eta \mathbf{V}_\eta^*)$$

with $\mathbf{U}_\eta \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V}_\eta \in \mathbb{R}^{d_2 \times r}$ being the matrices with the first r left resp. right singular

[4] Alternatively, while arriving at slightly different constants, following the proof of [CDK15, Theorem 4.1] results in qualitatively the same statement.

vectors of η in their columns. By [Dax10, Corollary 31], it then follows that

$$\|\eta_c\|_F = \|(\mathbf{I}-\mathbf{U}_\eta\mathbf{U}_\eta^*)\eta(\mathbf{I}-\mathbf{V}_\eta\mathbf{V}_\eta^*)\|_F \leq \|\mathcal{P}_{T^\perp}(\eta)\|_F$$

for *any* tangent space T (2.99) of the rank-r manifold. This shows that (2.100) is fulfilled with $c(r,d_1,d_2) = \sqrt{\frac{2d}{r}}$.

□

With Proposition 2.5.1, we can specify Theorem 2.7 for the case of sub-Gaussian measurement operators Φ.

Corollary 2.5.1 (Local convergence statement for sub-Gaussian operators). *Let $\mathbf{X}_0 \in \mathbb{R}^{d_1\times d_2}$ be a matrix of rank r, and let $\Phi : \mathbb{R}^{d_1\times d_2} \longrightarrow \mathbb{R}^m$ be a sub-Gaussian operator as in Definition 2.5.1. There exists a constant $C \geq 1$ such that if*

$$m \geq Cr(d_1+d_2),$$

with high probability, the following holds: If the output matrix $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1\times d_2}$ of the k-th iteration of `MatrixIRLS` *with inputs Φ, $\mathbf{y} = \Phi(\mathbf{X}_0)$, $p \in [0,1)$ and $\widetilde{r} = r$ fulfills*

$$\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \leq \frac{1}{\left(2^{3-p}(1+6\kappa)d\right)^{\frac{1}{1-p}}}\sigma_r(\mathbf{X}_0), \tag{2.124}$$

then

$$\mathbf{X}^{(k)} \xrightarrow{k\to\infty} \mathbf{X}_0$$

and furthermore, the local convergence rate *is of order $2-p$ in the sense that*

$$\|\mathbf{X}^{(k+1)} - \mathbf{X}_0\|_{S_\infty} \leq \mu\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty}^{2-p} \tag{2.125}$$

with

$$\mu := \frac{2^{3-p}d(1+6\kappa)}{\sigma_r(\mathbf{X}_0)^{1-p}}.$$

Proof. We first note that by Proposition 2.5.1, Assumption 2.1 is fulfilled in the sense that (2.100) holds, with high probability, for all matrices $\mathbf{Z} \in \mathbb{R}^{d_1\times d_2}$ with $c(r,d_1,d_2) = \sqrt{\frac{2d}{r}}$. Thus, ξ of Theorem 2.7 can be chosen such that $\xi = \infty$, which does not impose any restriction on the statement.

Now, we quantify when the condition

$$\mu\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty}^{1-p} < 1 \tag{2.126}$$

is fulfilled: With $\widetilde{\mu}$ as in (2.119) and $\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \leq \zeta\sigma_r(\mathbf{X}_0)$, we see that

$$\begin{aligned}\widetilde{\mu}\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty}^{1-p} &= \frac{\frac{2d}{r}r\left(1+4\kappa+2\zeta^p\kappa\right)}{(1-\zeta)^{2-p}\sigma_r(\mathbf{X}_0)^{1-p}}\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty}^{1-p} \leq \frac{2\cdot 2^{2-p}\left(1+4\kappa+2\kappa\right)}{\sigma_r(\mathbf{X}_0)^{1-p}}\zeta^{1-p}\sigma_r(\mathbf{X}_0)^{1-p}\\ &= 2^{3-p}(1+6\kappa)\zeta^{1-p} < 1,\end{aligned}$$

if

$$\zeta < \frac{1}{2^{(3-p)/(1-p)}(1+6\kappa)^{1/(1-p)}d^{\frac{1}{1-p}}}.$$

This shows that (2.126) is fulfilled if

$$\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} < \frac{1}{2^{(3-p)/(1-p)}(1+6\kappa)^{1/(1-p)} d^{\frac{1}{1-p}}} \sigma_r(\mathbf{X}_0),$$

and the statement of Corollary 2.5.2 is shown for the choice of

$$\mu = \frac{2^{3-p} d(1+6\kappa)}{\sigma_r(\mathbf{X}_0)^{1-p}}.$$

□

2.5.2 Phase Retrieval with Gaussian Measurements

While sub-Gaussian measurement operators as described above have nice theoretical properties as they exhibit restricted isometry properties already if the number of measurements is quite small, direct applications of these operators in the low-rank matrix recovery setting are scarce. An additional issue that limits their applicability in large-scale data analysis settings is the fact that they pose computational challenges for many algorithms:

If $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ is sub-Gaussian as in Section 2.5.1, the time complexity of calculating $\Phi(\mathbf{X})$ for an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ is $O(md_1d_2)$, which cannot be smaller than

$$O(rdD^2)$$

even if $m \sim r(d_1 + d_2)$ (recall the notation $d = \min(d_1, d_2)$ and $D = \max(d_1, d_2)$).

In the following, we briefly explain the applicability of `MatrixIRLS` for the so-called *phase retrieval* problem [GS72, Fie82, CESV13, CSV13, CLS15], which arises in astronomy [FD87], microscopy [TLRW14] and X-ray crystallography [Mil90].

If $\mathbf{x}_0 \in \mathbb{C}^d$ is a complex-valued object (often a vectorization of a 2D image), in these applications, what happens mathematically is that the only knowledge about the object itself is provided through *m phaseless linear measurements*

$$(y_\ell)_{\ell=1}^m = \mathbf{y} = |\mathbf{A}\mathbf{x}_0|^2 = (|\langle \mathbf{a}_\ell, \mathbf{x}_0 \rangle|^2)_{\ell=1}^m, \tag{2.127}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^* \\ \mathbf{a}_2^* \\ \vdots \\ \mathbf{a}_m^* \end{bmatrix}$$

is a complex $(m \times d)$ matrix with rows $\mathbf{a}_\ell^*$. In particular, the *phase* (or in a real setting, the sign of) the linear measurements $\langle \mathbf{a}_\ell, \mathbf{x}_0 \rangle$, $\ell = 1, \dots, m$ is not available for the decoder or reconstruction algorithm.

If the number of measurements $m \geq d$ is large enough, however, it is possible to show that $\mathbf{x}$ can be reconstructed uniquely from the phaseless measurements [BCE06, EM14] for many sets of measurement vectors $(\mathbf{a}_\ell)_{\ell=1^m}$. In many cases, it is possible to obtain reconstructions by tractable algorithms in these cases [GS72, Fie82, CESV13, CSV13, CLS15].

In [CESV13, CSV13, CL14], it was demonstrated that the phase retrieval problem can actually be reformulated as a *recovery problem of a rank-one matrix of size* $(d \times d)$. In fact, if we define the matrix $\mathbf{X}_0 := \mathbf{x}_0\mathbf{x}_0^* \in \mathbb{C}^{d \times d}$, the measurements (2.127) which are *quadratic* in $\mathbf{x}_0$

$$y_\ell = |\langle \mathbf{a}_\ell, \mathbf{x}_0 \rangle|^2 = \mathbf{a}_\ell^* \mathbf{x}_0 \mathbf{x}_0^* \mathbf{a}_\ell = \mathbf{a}_\ell^* \mathbf{X}_0 \mathbf{a}_\ell$$

become *linear* in $\mathbf{X}_0$, and we know that the rank-one and positive definite matrix $\mathbf{X}_0$ is a solution to the linear system

$$\Phi(\mathbf{X}) = \mathbf{y}$$

with linear operator $\Phi : \mathbb{C}^{d\times d} \longrightarrow \mathbb{C}^m$ such that

$$\Phi(\mathbf{X})_\ell = \mathbf{a}_\ell^* \mathbf{X} \mathbf{a}_\ell = \operatorname{tr}(\mathbf{a}_\ell^* \mathbf{X} \mathbf{a}_\ell) = \operatorname{tr}(\mathbf{a}_\ell f a_\ell^* \mathbf{X}) = \langle \mathbf{a}_\ell \mathbf{a}_\ell^*, \mathbf{X}\rangle \tag{2.128}$$

for all $\ell \in [m]$. Algorithmically, it was proposed in [CESV13, CSV13] that $\mathbf{X}_0$ could be recovered by solving the nuclear norm minimization problem

$$\min_{\mathbf{X}\in\mathbb{C}^{d\times d}, \mathbf{X}\succeq 0} \|\mathbf{X}\|_{S_1} \quad \text{subject to} \quad \Phi(\mathbf{X}) = \mathbf{y} \tag{2.129}$$

with Φ as in (2.128), and, once $\mathbf{X}_0$ is calculated, the vector $\mathbf{x}_0 \in \mathbb{C}^d$ with correct phase can be easily calculated as it is the only eigenvector of $\mathbf{X}_0$ with non-zero eigenvalue.

From a theoretical point of view, the works [CESV13, CSV13, KKRT16] established that with high probability, (2.129) successfully recovers any vector $\mathbf{x}_0 \in \mathbb{C}^d$ from corresponding measurements $\mathbf{y} = \Phi(\mathbf{x}_0\mathbf{x}_0^*)$, if the $\mathbf{a}_\ell$ are random complex Gaussian vectors and if

$$m \geq cd$$

for a fixed constant $c > 1$.

The Gaussian assumption on the measurement vector $\mathbf{a}_\ell$ may not fully correspond to the structure of the phaseless measurements in applications of phase retrieval in astronomy or microscopy, where some type of windowed or modulated Fourier measurements are more common [CSV13], but they are already more realistic than the sub-Gaussian operators described in Definition 2.5.1.

While it is more difficult to incorporate the positive definiteness constraint $\mathbf{X} \succeq 0$ of (2.129) into the IRLS framework explicitly, it is natural to apply `MatrixIRLS` to optimize

$$\min_{\mathbf{X}\in\mathbb{C}^{d\times d}:\mathbf{X}^*=\mathbf{X}} F_{p,\epsilon}(\mathbf{X}) \quad \text{subject to} \quad \Phi(\mathbf{X}) = \mathbf{y}$$

with $F_{p,\epsilon}(\mathbf{X})$ as in (2.48) using Algorithm 1 with, for example, $p = 0$ and $\widetilde{r} = 1$. It is to be expected that in case enough measurements m are provided, there is not only a unique rank-one positive definite matrix compatible with the measurements, this solution is also unique among all *Hermitian* rank-ones matrices [KKRT16].

The theory of this chapter mainly addressed the case of *real* low-rank matrices $\mathbf{X}$, but the entire theory can be also translated to the complex domain without many changes.

The fact that the Hermitian constraint $\mathbf{X}^* = \mathbf{X}$ can be incorporated into `MatrixIRLS` quite easily comes from the fact that for Hermitian matrices $\mathbf{X}^{(k)}$, the subspaces spanned by the first r left and right singular vectors are the same, so setting $\mathbf{V}_k = \mathbf{U}_k$ in the definition of the weight operator $W^{(k)}$ (see again Definition 2.3.4) of `MatrixIRLS` allows us to define the weight operator such that

$$W^{(k)}(\mathbf{X}) = \mathbf{U}_k \left[\mathbf{H}_1^{(k)} \circ (\mathbf{U}_k^* \mathbf{X} \mathbf{U}_k)\right] \mathbf{U}_k^* \tag{2.130}$$

for $\mathbf{X} \in \mathbb{C}^{d\times d}$, where $(\mathbf{H}_1^{(k)}) \in \mathbb{R}^{d\times d}$ is defined as in (2.58). In (2.130), the weights corresponding to $(\mathbf{H}_2^{(k)})$ do not appear, since the projection onto the anti-Hermitian part $\widetilde{T} : \mathbb{C}^{d\times d} \longrightarrow \mathbb{C}^{d\times d}$, $\widetilde{T}(\mathbf{X}) := \frac{1}{2}(\mathbf{X} - \mathbf{X}^*)$ (see (2.55)) is zero if applied to a Hermitian matrix $\mathbf{X}$.

Using the weight operator from (2.130), we obtain the following proposition.

Proposition 2.5.2 (Preservation of Hermiticity by weight operator). *Let $\mathbf{X}^{(k)} \in \mathbb{C}^{d\times d}$ be a Hermitian matrix with singular value decomposition $\mathbf{X}^{(k)} = \mathbf{U}_k \operatorname{dg}(\sigma^{(k)})\mathbf{V}_k^*$. Then the weight operator $W^{(k)} : \mathbb{C}^{d\times d} \to \mathbb{C}^{d\times d}$ of (2.130) preserves Hermiticity, i.e., for any $\mathbf{X} \in \mathbb{C}^{d\times d}$ with $\mathbf{X}^* = \mathbf{X}$, it holds that*

$$\left(W^{(k)}(\mathbf{X})\right)^* = W^{(k)}(\mathbf{X}).$$

Proof. By its definition, the matrix $(\mathbf{H}_1^{(k)}) \in \mathbb{R}^{d\times d}$ of (2.58) is real symmetric as $d = d_1 = d_2$, so it is Hermitian as a matrix over the complex domain. Thus,

$$\begin{aligned}\left(W^{(k)}(\mathbf{X})\right)^* &= \left(\mathbf{U}_k \left[\mathbf{H}_1^{(k)} \circ (\mathbf{U}_k^*\mathbf{X}\mathbf{U}_k)\right] \mathbf{U}_k^*\right)^* \\ &= \mathbf{U}_k \left[\mathbf{H}_1^{(k)*} \circ (\mathbf{U}_k^*\mathbf{X}^*\mathbf{U}_k)\right] \mathbf{U}_k^* = \mathbf{U}_k \left[\mathbf{H}_1^{(k)} \circ (\mathbf{U}_k^*\mathbf{X}\mathbf{U}_k)\right] \mathbf{U}_k^* \\ &= W^{(k)}(\mathbf{X}),\end{aligned}$$

using that $\mathbf{X}^* = \mathbf{X}$. □

Using results from [KKRT16] about null space properties for measurement operators $\Phi : \mathbb{C}^{d\times d} \to \mathbb{C}^m$ with rank-one measurements (2.127) that are *complex Gaussian*, we can specify Theorem 2.7 for the phase retrieval setting.

Corollary 2.5.2 (Local convergence of `MatrixIRLS` for Phase Retrieval). *Let $fx_0 \in \mathbb{C}^d$, and assume the measurement operator $\Phi : \mathbb{C}^{d\times d} \to \mathbb{C}^m$ consists of independent complex Gaussian rank-one measurements $(\mathbf{a}_\ell)_{\ell=1}^m$ as in (2.128), i.e., the d-dimensional random vectors $(\mathbf{a}_\ell)_{\ell=1}^m$ are independent with independent complex standard Gaussian entries $(\mathbf{a}_\ell)_j \sim \mathcal{N}(0, 1/\sqrt{2}) + i\mathcal{N}(0, 1/\sqrt{2})$, $j \in [d]$. There exists a constant $c \geq 1$ such that if*

$$m \geq cd, \tag{2.131}$$

with high probability, the following holds: If the output matrix $\mathbf{X}^{(k)} \in \mathbb{R}$ of the k-th iteration of `MatrixIRLS` *with inputs Φ, $\mathbf{y} = \Phi(\mathbf{x}_0\mathbf{x}_0^*)$, $p \in [0, 1)$ and $\widetilde{r} = 1$ fulfills*

$$\|\mathbf{X}^{(k)} - \mathbf{x}_0\mathbf{x}_0^*\|_{S_\infty} \leq \frac{1}{\left(2^{3-p}(1+6\kappa)d\right)^{\frac{1}{1-p}}}\|\mathbf{x}_0\|_2^2, \tag{2.132}$$

then

$$\mathbf{X}^{(k)} \xrightarrow{k\to\infty} \mathbf{x}_0\mathbf{x}_0^*$$

and furthermore, the local convergence rate is of order $2 - p$ *in the sense that*

$$\|\mathbf{X}^{(k+1)} - \mathbf{x}_0\mathbf{x}_0^*\|_{S_\infty} \leq \mu\|\mathbf{X}^{(k)} - \mathbf{x}_0\mathbf{x}_0^*\|_{S_\infty}^{2-p} \tag{2.133}$$

with

$$\mu := \frac{2^{3-p}d(1+6\kappa)}{\|\mathbf{x}_0\|_2^{2-2p}}.$$

Proof. Let $0 < \gamma < 1$. From the statement and proof of [KKRT16, Theorem 1.2], it follows that there exists a constant $C(\gamma) \geq 1$ such that if (2.131) is fulfilled, then Φ fulfills with high probability the strong Schatten-1 null space property of Definition 2.4.2 of order 1 with constant $\gamma = \gamma_1$. Following the proof of Proposition 2.5.1.2, we see that Φ fulfills condition (2.100) for all tangent spaces $T = T_Z$ onto the rank-one manifold with constant $c(d) = \sqrt{2d}$.

The statement of corollary 2.5.2 follows then from the proof of Corollary 2.5.2, using that

$$\sigma_1(\mathbf{x}_0\mathbf{x}_0^*) = \|\mathbf{x}_0\mathbf{x}_0^*\|_{S_\infty} = \|\mathbf{x}_0\|_2^2.$$

□

2.5.3 Matrix Completion

In Section 2.4.2, it has already been mentioned that a strong Schatten-p null space property cannot hold for *low-rank matrix completion*, which is the class of low-rank matrix recovery problems that has perhaps been most widely studied in the literature [CR09, CT10, Rec11, Che15, DR16, CLC19, MWCC19]. Recall that in this setting, the linear measurement operator $\Phi : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$ corresponds to a set of double indices $(i_\ell, j_\ell)_{\ell=1}^m \subset [d_1]\times[d_2]$ such that

$$\Phi(\mathbf{X})_\ell = \langle e_{i_\ell} e_{j_\ell}^*, \mathbf{X}\rangle = \mathbf{X}_{i_\ell, j_\ell} \tag{2.134}$$

for any $\ell \in [m]$, for arbitrary $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$. Here, e_{i_ℓ} and e_{j_ℓ} denote the i_ℓ-th and j_ℓ-th standard basis vector of $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively.

In fact, if $(\bar{i}, \bar{j}) \in [d_1]\times[d_2]$ is an index pair such that $(\bar{i}, \bar{j}) \neq (i_\ell, j_\ell)_{\ell=1}^m$, the matrix $\eta = e_{\bar{i}} e_{\bar{j}}^*$ is of rank $r = 1$, and $P_{T_\eta}(\eta) = \eta$ if T_η is the tangent space onto the rank-1 matrix manifold at η, which implies that

$$\|\eta\|_F > c(r, d_1, d_2)\|\mathcal{P}_{T_\eta^\perp}(\eta)\|_F = 0,$$

for any constant $c(r, d_1, d_2) > 0$, as $P_{T_\eta^\perp}(\eta) = 0$, despite η being an element of $\ker\Phi$. This means that condition (2.100) of Assumption 2.1 cannot be fulfilled for all tangent spaces $T_\mathbf{Z}$, unlike it was the case if Φ fulfilled a restricted isometry property or a strong Schatten-p null space property.

In the rest of this section, we will establish that Assumption 2.1 holds, on the other hand, at least locally around a fixed matrix $\mathbf{X}_0 \in \mathbb{R}^{d_1\times d_2}$ of rank r, if enough entries are provided by Φ and if this matrix fulfills a suitable *incoherence* condition.

Incoherence conditions on the low-rank matrix $\mathbf{X}_0$ to be completed prevent that the information of the subspace in which $\mathbf{X}_0$ lies in is too concentrated in few columns or rows, which would cause an image of $\mathbf{X}_0$ with respect to an entrywise operator as Φ to contain too little information to *invert* Φ by any reconstruction algorithm. Similar conditions of this type have been used in the literature [CR09, Rec11, Che15], establishing, for example, recovery guarantees for nuclear norm minimization in the low-rank matrix completion setting.

We will use the following definition of *incoherence*, which quantifies the alignment of the standard basis $(e_i e_j^*)_{i=1,j=1}^{d_1,d_2}$ of $\mathbb{R}^{d_1\times d_2}$ with the *tangent space* onto the rank-r manifold at rank-r matrix $\mathbf{X}_0 \in \mathbb{R}^{d_1\times d_2}$.

Definition 2.5.2. *We say that a rank-r matrix $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$ with singular value decomposition $\mathbf{X} = \mathbf{U\Sigma V}^*$, $\mathbf{U} \in \mathbb{R}^{d_1\times r}$, $\mathbf{V} \in \mathbb{R}^{d_2\times r}$, is* μ_0-incoherent *if there exists a constant $\mu_0 \geq 1$ such that*

$$\max_{1\leq i\leq d_1, 1\leq j\leq d_2} \|\mathcal{P}_T(e_i e_j^*)\|_F \leq \sqrt{\mu_0 r \frac{d_1+d_2}{d_1 d_2}}, \tag{2.135}$$

where $T = T_\mathbf{X} = \{\mathbf{UM}^* + \widetilde{\mathbf{M}}\mathbf{V}^* : \mathbf{M} \in \mathbb{R}^{d_2\times r}, \widetilde{\mathbf{M}} \in \mathbb{R}^{d_1\times r}\}$ is the tangent space onto the rank-r manifold at $\mathbf{X}$.

Remark 2.5.1. *We note that the assumption that a rank-r matrix $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$ is μ_0-incoherent according to Definition 3.3.1* is weaker *than similar assumptions* [CR09, Definition 1.2, Ao and

A1], [Rec11, Definition 1, Theorem 2], *and even than the assumption* [Che15, (2)], *which is the weakest available incoherence condition in the literature that is used for showing successful completion by nuclear norm minimization. More precisely,* [Che15] *calls a matrix* $\mathbf{X}$ μ_0*-incoherent if*

$$\max_{1\leq i\leq d_1} \|\mathbf{U}^* e_i\|_2 \leq \sqrt{\frac{\mu_0 r}{d_1}} \quad \textit{and} \quad \max_{1\leq j\leq d_2} \|\mathbf{V}^* e_j\|_F \leq \sqrt{\frac{\mu_0 r}{d_2}}. \tag{2.136}$$

In fact, condition (2.136) *is* weaker *than* (2.135)*. This can be seen from the calculation that*

$$\begin{aligned}\|\mathscr{P}_T(e_i e_j^*)\|_F^2 &= \|\mathbf{U}\mathbf{U}^* e_i e_j^* + e_i e_j^* \mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^* e_i e_j^* \mathbf{V}\mathbf{V}^*\|_F^2 = \|\mathbf{U}\mathbf{U}^* e_i e_j^*(\mathbf{I} - \mathbf{V}\mathbf{V}^*) + e_i e_j^* \mathbf{V}\mathbf{V}^*\|_F^2 \\ &= \|\mathbf{U}\mathbf{U}^* e_i e_j^*(\mathbf{I} - \mathbf{V}\mathbf{V}^*)\|_F^2 + \|e_i e_j^* \mathbf{V}\mathbf{V}^*\|_F^2 \leq \|\mathbf{U}\mathbf{U}^* e_i e_j^*\|_F^2 \|\mathbf{I} - \mathbf{V}\mathbf{V}^*\|^2 + \|e_i e_j^* \mathbf{V}\mathbf{V}^*\|_F^2 \\ &\leq \|\mathbf{U}^* e_i e_j^*\|_F^2 + \|e_i e_j^* \mathbf{V}\|_F^2 = \|\mathbf{U}^* e_i\|_2^2 + \|\mathbf{V}^* e_j\|_2^2 \leq \frac{\mu_0 r}{d_1} + \frac{\mu_0 r}{d_2} \leq \frac{\mu_0 r(d_1 + d_2)}{d_1 d_2}\end{aligned}$$

for any $i \in [d_1]$*,* $j \in [d_2]$*, if* (2.136) *is fulfilled, which holds since*

$$\|\mathbf{U}^* e_i e_j^*\|_F^2 = \operatorname{tr}(e_j e_i^* \mathbf{U}\mathbf{U}^* e_i e_j^*) = \operatorname{tr}(e_i^* \mathbf{U}\mathbf{U}^* e_i) = e_i^* \mathbf{U}\mathbf{U}^* e_i = \|\mathbf{U}^* e_i\|_2^2$$

and similarly $\|e_i e_j^* \mathbf{V}\|_F^2 = \|\mathbf{V}^* e_j\|_2^2$*.*

To establish Assumption 2.1 of the measurement operator Φ around a μ_0-incoherent matrix $\mathbf{X}_0$ for a number of samples m as small as possible, we resort to *randomness* in the distribution of the indices (i_ℓ, j_ℓ) of known entries in (2.134), along the same line of reasoning as in the case of sub-Gaussian operators or the operator of random rank-one measurements.

Our model for the locations $(i_\ell, j_\ell)_{\ell=1}^m$ of the entries of $\mathbf{X}$ provided by $\Phi(\mathbf{X})$ assumes that are they are *distributed uniformly without replacement* among all possible locations $[d_1] \times [d_2]$. We obtain the following local convergence statement for `MatrixIRLS`.

Corollary 2.5.3 (Local convergence of `MatrixIRLS` for Matrix Completion)**.** *Let* $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ *be a matrix of rank* r *that is* μ_0*-incoherent, and let* $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ *consist of entry-wise measurements* (2.134) *for a set of locations* $(i_\ell, j_\ell)_{\ell=1}^m$ *that is drawn uniformly without replacement. There exists a constant* $C \geq 1$ *such that if*

$$m \geq C\mu_0 r(d_1 + d_2)\log(d_1 + d_2). \tag{2.137}$$

with high probability, the following holds: If the output matrix $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ *of the k-th iteration of* `MatrixIRLS` *with inputs* Φ*,* $\mathbf{y} = \Phi(\mathbf{X}_0)$*,* $p \in [0, 1)$ *and* $\widetilde{r} = r$ *fulfills*

$$\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \leq \min\left(C_1\sqrt{\frac{\mu_0 r}{d}}, \left(C_2 \frac{\mu_0}{d\log(D)\kappa}\right)^{\frac{1}{1-p}}\right)\sigma_r(\mathbf{X}_0), \tag{2.138}$$

where C_1 *and* C_2 *are absolute constants, then*

$$\mathbf{X}^{(k)} \xrightarrow{k\to\infty} \mathbf{X}_0$$

and furthermore, the local convergence rate *is of order* $2 - p$ *in the sense that*

$$\|\mathbf{X}^{(k+1)} - \mathbf{X}_0\|_{S_\infty} \leq \mu\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty}^{2-p} \tag{2.139}$$

with

$$\mu := 2^{2-p}\frac{\widetilde{C}d\log(D)}{\mu_0 r\sigma_r(\mathbf{X}_0)^{1-p}}r(1+6\kappa).$$

In the proof of Corollary 2.5.3, we will actually work with a sampling model of the the locations $(i_\ell, j_\ell)_{\ell=1}^m$ that is one of *independent sampling with replacement*. By the argument of [Rec11, Proposition 3], the statement then carries over to the above model of sampling without replacement.

Due to the model of sampling with replacement, it is possible to sample locations $(i,j) \in [d_1]\times[d_2]$ more than once, but the following proposition due to [Rec11] bounds the number of repetitions of each location.

Lemma 2.5.1 ([Rec11, Proposition 5]). *Let $D = \max(d_1, d_2)$ and $\beta > 1$, let $\Omega = (i_\ell, j_\ell)_{\ell=1}^m$ be a multiset of double indices from $[d_1]\times[d_2]$ fulfilling $m < d_1 d_2$ that are sampled independently with replacement. Then with probability at least $1 - D^{2-2\beta}$, the maximum number of repetitions of any entry in Ω is less than $\frac{8}{3}\beta\log(D)$ for $D \geq 9$ and $\beta > 1$. Consequently, we have that with probability of at least $1 - D^{2-2\beta}$, the operator $\mathscr{R}_\Omega : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ defined such that*

$$\mathscr{R}_\Omega(\mathbf{X}) := \Phi^*(\Phi(\mathbf{X})) = \sum_{\ell=1}^{m}\langle e_{i_\ell}e_{j_\ell}^*, \mathbf{X}\rangle e_{i_\ell}e_{j_\ell}^* \tag{2.140}$$

fulfills

$$\|\mathscr{R}_\Omega\|_{S_\infty} \leq \frac{8}{3}\beta\log(D).$$

Next, we use a lemma of [Rec11] that can be seen as a result of a *local restricted isometry property*, and which will help us to ensure Assumption 2.1. We detail its proof since we use the weaker incoherence definition of Definition 3.3.1 instead of the incoherence notions of [Rec11, Che15].

Lemma 2.5.2 ([Rec11, Theorem 6]). *Let $0 < \epsilon \leq \frac{1}{2}$, let $\mathbf{X}_0 \in \mathbb{R}^{d_1\times d_2}$ be a μ_0-incoherent matrix whose tangent space $T_0 = T_{\mathbf{X}_0}$ onto the rank-r manifold (2.99) fulfills (2.135) and $\mathbf{R}_\Omega : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ be defined as in (2.140) from m independent uniformly sampled locations. Then*

$$\left\|\frac{d_1 d_2}{m}\mathscr{P}_{T_0}\mathscr{R}_\Omega\mathscr{P}_{T_0} - \mathscr{P}_{T_0}\right\|_{S_\infty} \leq \varepsilon \tag{2.141}$$

holds with probability at least $1-(d_1+d_2)^{-2}$ provided that

$$m \geq \frac{7}{\varepsilon^2}\mu_0 r(d_1+d_2)\log(d_1+d_2). \tag{2.142}$$

Proof. First we define the family of operators $\mathscr{Z}_\ell, \widetilde{\mathscr{Z}}_\ell : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ such that for $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$,

$$\mathscr{Z}_\ell(\mathbf{X}) := \frac{d_1 d_2}{m}\langle e_{i_\ell}e_{j_\ell}^*, \mathscr{P}_{T_0}(\mathbf{X})\rangle\mathscr{P}_{T_0}(e_{i_\ell}e_{j_\ell}^*) - \frac{1}{m}\mathscr{P}_{T_0} := \frac{d_1 d_2}{m}\widetilde{\mathscr{Z}}_\ell(\mathbf{X}) - \frac{1}{m}\mathscr{P}_{T_0}(\mathbf{X})$$

for any $\ell \in [m]$. Then

$$\mathbb{E}[\mathscr{Z}_\ell] = \frac{1}{d_1 d_2}\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\frac{d_1 d_2}{m}\langle e_i e_j^*, \mathscr{P}_{T_0}(\cdot)\rangle\mathscr{P}_{T_0}(e_i e_j^*) - \frac{1}{m}\mathscr{P}_{T_0} = \frac{1}{d_1 d_2}\frac{d_1 d_2}{m}\mathscr{P}_{T_0}\mathbf{I}\mathscr{P}_{T_0} - \frac{1}{m}\mathscr{P}_{T_0} = 0. \tag{2.143}$$

Since for $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$

$$\langle e_{i_\ell} e_{j_\ell}^*, \mathscr{P}_{T_0}(\mathbf{X}) \rangle \mathscr{P}_{T_0}(e_{i_\ell} e_{j_\ell}^*) = \langle \mathscr{P}_{T_0}(e_{i_\ell} e_{j_\ell}^*), \mathbf{X} \rangle \mathscr{P}_{T_0}(e_{i_\ell} e_{j_\ell}^*)$$

we obtain

$$\| \langle e_{i_\ell} e_{j_\ell}^*, \mathscr{P}_{T_0}(\mathbf{X}) \rangle \mathscr{P}_{T_0}(e_{i_\ell} e_{j_\ell}^*) \|_F \leq \left| \langle \mathscr{P}_{T_0}(e_{i_\ell} e_{j_\ell}^*), \mathbf{X} \rangle \right| \| \mathscr{P}_{T_0}(e_{i_\ell} e_{j_\ell}^*) \|_F \leq \| \mathscr{P}_{T_0}(e_{i_\ell} e_{j_\ell}^*) \|_F^2 \| X \|_F$$

by Cauchy-Schwartz, and thus the norm bound

$$\begin{aligned} \frac{d_1 d_2}{m} \left\| \widetilde{\mathfrak{X}}_\ell \right\|_{S_\infty} &\leq \frac{d_1 d_2}{m} \| \mathscr{P}_T(e_{i_\ell} e_{j_\ell}^*) \|_F^2 \leq \frac{d_1 d_2}{m} \max_{i \in [d_1], j \in [d_2]} \| \mathscr{P}_{T_0}(e_{i_\ell} e_{j_\ell}^*) \|_F^2 \\ &\leq \frac{d_1 d_2}{m} \frac{\mu_0 r (d_1 + d_2)}{d_1 d_2} = \frac{\mu_0 r (d_1 + d_2)}{m} \end{aligned} \tag{2.144}$$

using the incoherence assumption (2.135) in the last inequality. Similarly,

$$\left\| \frac{1}{m} \mathscr{P}_{T_0} \right\|_{S_\infty} = \left\| \frac{1}{m} \mathscr{P}_{T_0} \mathbf{I} \mathscr{P}_{T_0} \right\|_{S_\infty} \leq \frac{1}{m} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \left\| \langle \mathscr{P}_{T_0}(e_i e_j^*), (\cdot) \rangle \mathscr{P}_{T_0}(e_i e_j^*) \right\|_{S_\infty} \leq \frac{\mu_0 r (d_1 + d_2)}{m}. \tag{2.145}$$

We note that if operators $\mathscr{A}$ and $\mathscr{B}$ are positive semidefinite, then $\| \mathscr{A} - \mathscr{B} \|_{S_\infty} \leq \max(\| \mathscr{A} \|_{S_\infty}, \| \mathscr{B} \|_{S_\infty})$, and as both $\widetilde{\mathfrak{X}}_\ell$ and $\mathscr{P}_{T_0}$ are positive semidefinite,

$$\| \mathfrak{X}_\ell \|_{S_\infty} \leq \max \left(\frac{d_1 d_2}{m} \left\| \widetilde{\mathfrak{X}}_\ell \right\|_{S_\infty}, \frac{1}{m} \left\| \mathscr{P}_{T_0} \right\|_{S_\infty} \right) = \frac{\mu_0 r (d_1 + d_2)}{m}$$

for all $\ell \in [m]$. For the expectation of the squares of $\mathfrak{X}_\ell$, we obtain

$$\begin{aligned} \mathbb{E} \mathfrak{X}_\ell \mathfrak{X}_\ell^* &= \frac{(d_1 d_2)^2}{m^2} \mathbb{E} \left[(\widetilde{\mathfrak{X}}_\ell)^* \widetilde{\mathfrak{X}}_\ell \right] - \frac{d_1 d_2}{m} \mathbb{E} \left[\widetilde{\mathfrak{X}}_\ell \right] \mathscr{P}_{T_0} - \frac{d_1 d_2}{m} \mathscr{P}_{T_0} \mathbb{E} \left[\widetilde{\mathfrak{X}}_\ell \right] + \frac{1}{m^2} \mathscr{P}_{T_0} \\ &= \frac{(d_1 d_2)^2}{m^2} \mathbb{E} \left[(\widetilde{\mathfrak{X}}_\ell)^* \widetilde{\mathfrak{X}}_\ell \right] + (1 - 2) \frac{1}{m^2} \mathscr{P}_{T_0}, \end{aligned}$$

as $\mathscr{P}_{T_0}^2 = \mathscr{P}_{T_0}$ and $\mathbb{E}[\widetilde{\mathfrak{X}}_\ell] = \frac{1}{d_1 d_2} \mathscr{P}_{T_0}$. Thus,

$$\begin{aligned} \left\| \sum_{\ell=1}^m \mathbb{E} \mathfrak{X}_\ell \mathfrak{X}_\ell^* \right\|_{S_\infty} &\leq \sum_{\ell=1}^m \| \mathbb{E} \mathfrak{X}_\ell \mathfrak{X}_\ell^* \|_{S_\infty} = \sum_{\ell=1}^m \left\| \frac{(d_1 d_2)^2}{m^2} \mathbb{E} \left[(\widetilde{\mathfrak{X}}_\ell)^2 \right] - \frac{1}{m^2} \mathscr{P}_{T_0} \right\|_{S_\infty} \\ &\leq \sum_{\ell=1}^m \max \left(\frac{(d_1 d_2)^2}{m^2} \left\| \mathbb{E} \left[(\widetilde{\mathfrak{X}}_\ell)^2 \right] \right\|_{S_\infty}, \frac{1}{m^2} \left\| \mathscr{P}_{T_0} \right\|_{S_\infty} \right) \\ &\leq \sum_{\ell=1}^m \max \left(\frac{(d_1 d_2)^2}{m^2} \left\| \mathbb{E} \left[\| \mathscr{P}_{T_0}(e_{i_\ell} e_{j_\ell}^*) \|_F^2 \widetilde{\mathfrak{X}}_\ell \right] \right\|_{S_\infty}, \frac{1}{m^2} \right) \\ &\leq \sum_{\ell=1}^m \max \left(\frac{(d_1 d_2)(d_1 + d_2) \mu_0 r}{m^2} \left\| \mathbb{E} \widetilde{\mathfrak{X}}_\ell \right\|_{S_\infty}, \frac{1}{m^2} \right) \\ &\leq \sum_{\ell=1}^m \max \left(\frac{(d_1 + d_2) \mu_0 r}{m^2}, \frac{1}{m^2} \right) = \frac{\mu_0 r (d_1 + d_2)}{m}, \end{aligned}$$

where we used that $\| \mathscr{P}_{T_0} \|_2 \leq 1$ since $\mathscr{P}_{T_0}$ is a projection in the third inequality, the definition of μ_0 in the fourth and the fact that $\mathbb{E}\widetilde{\mathcal{Z}}_\ell = \frac{1}{d_1 d_2}\mathcal{P}_{T_0}$ (see (2.143)) in the fifth. As the $\mathcal{Z}_\ell$ are Hermitian, it follows by the matrix Bernstein inequality (see, e.g., [Ver18, Theorem 5.4.1]) that

$$\begin{aligned}&\mathbb{P}\left(\left\|\frac{d_1 d_2}{m}\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0} - \mathcal{P}_{T_0}\right\|_{S_\infty} \geq \varepsilon\right) \leq (d_1+d_2)\exp\left(-\frac{m\varepsilon^2/2}{\mu_0 r(d_1+d_2) + \mu_0 r(d_1+d_2)\epsilon/3}\right)\\ &\leq (d_1+d_2)\exp\left(-\frac{m\varepsilon^2}{2\mu_0 r(d_1+d_2) + \mu_0 r(d_1+d_2)/3}\right),\end{aligned} \tag{2.146}$$

using that $\varepsilon \leq \frac{1}{2}$ in the last inequality.

Furthermore, if (2.142) is fulfilled, then

$$(d_1+d_2)\exp\left(-\frac{m\varepsilon^2}{\frac{7}{3}\mu_0 r(d_1+d_2)}\right) \leq (d_1+d_2)^{-2},$$

which shows that (3.37) holds with a probability of at least $1-(d_1+d_2)^{-2}$. □

The goal is now to establish the local restricted isometry statement of (3.37) for tangent spaces $T_{\mathbf{X}}$ corresponding to matrices $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$ that are close to $\mathbf{X}_0$.

We show the following refinement of [WCCL16b, Lemma 4.2].

Lemma 2.5.3. *Let $\mathbf{X}_0, \mathbf{X} \in \mathbb{R}^{d_1\times d_2}$ be matrices and assume that $0 < \varepsilon < 1$ and that the following three conditions hold:*

(a) For $\mathcal{R}_\Omega : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ as in (2.140),

$$\|\mathcal{R}_\Omega\|_{S_\infty} \leq \frac{16}{3}\log(D).$$

(b) The tangent space $T_0 = T_{\mathbf{X}_0}$ onto the rank-r manifold at $\mathbf{X}_0$ fulfills

$$\left\|\frac{d_1 d_2}{m}\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0} - \mathcal{P}_{T_0}\right\|_{S_\infty} \leq \varepsilon.$$

(c) The spectral norm distance between $\mathbf{X}$ and $\mathbf{X}_0$ fulfills

$$\|\mathbf{X}-\mathbf{X}_0\|_{S_\infty} \leq \frac{\sqrt{3}}{32\sqrt{\log(D)}\sqrt{(1+\varepsilon)}}\varepsilon\sqrt{\frac{m}{d_1 d_2}}\sigma_r(\mathbf{X}_0).$$

Then the tangent space $T = T_{\mathbf{X}}$ onto the rank-r manifold at $\mathbf{X}$ fulfills

$$\left\|\frac{d_1 d_2}{m}\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_T - \mathcal{P}_T\right\|_{S_\infty} \leq 4\varepsilon. \tag{2.147}$$

Proof. For any $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$, we have

$$\begin{aligned}
\left\|\mathcal{R}_\Omega \mathcal{P}_{T_0}(\mathbf{Z})\right\|_F^2 &= \langle \mathcal{R}_\Omega \mathcal{P}_T(\mathbf{Z}), \mathcal{R}_\Omega \mathcal{P}_T(\mathbf{Z}) \rangle \leq \frac{16}{3} \log(D) \left\langle \mathcal{P}_{T_0}(\mathbf{Z}), \mathcal{R}_\Omega \mathcal{P}_{T_0}(\mathbf{Z}) \right\rangle \\
&= \frac{16}{3} \log(D) \left\langle \mathcal{P}_{T_0}(\mathbf{Z}), \mathcal{P}_{T_0} \mathcal{R}_\Omega \mathcal{P}_{T_0}(\mathbf{Z}) \right\rangle \\
&= \frac{16}{3} \log(D) \left(\left\langle \mathcal{P}_{T_0}(\mathbf{Z}), \frac{m}{d_1 d_2} \mathcal{P}_{T_0}(\mathbf{Z}) \right\rangle + \left\langle \mathcal{P}_{T_0}(\mathbf{Z}), \left(\mathcal{P}_{T_0} \mathcal{R}_\Omega \mathcal{P}_{T_0}(\mathbf{Z}) - \frac{m}{d_1 d_2} \mathcal{P}_{T_0}(\mathbf{Z}) \right) \right\rangle \right) \\
&\leq \frac{16}{3} \log(D) \left(\frac{m}{d_1 d_2} + \varepsilon \frac{m}{d_1 d_2} \right) \left\|\mathcal{P}_{T_0}(\mathbf{Z})\right\|_F^2 \leq \frac{16}{3} \log(D)(1+\varepsilon) \frac{m}{d_1 d_2} \|\mathbf{Z}\|_F^2 ,
\end{aligned}$$

where the first inequality follows from condition (a) and the second one from condition (b). It follows that

$$\left\|\mathcal{R}_\Omega \mathcal{P}_{T_0}\right\| \leq \sqrt{\frac{16}{3} \log(D)(1+\varepsilon) \frac{m}{d_1 d_2}}. \tag{2.148}$$

Furthermore, if $\mathbf{U}, \mathbf{U}_0 \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V}, \mathbf{V}_0 \in \mathbb{R}^{d_2 \times r}$ are the matrices of first r left and right singular vectors of $\mathbf{X}$ and $\mathbf{X}_0$, respectively, it holds that for any $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$,

$$\begin{aligned}
(\mathcal{P}_T - \mathcal{P}_{T_0})(\mathbf{Z}) &= \mathbf{U}\mathbf{U}^*\mathbf{Z} + \mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}_0\mathbf{U}_0^*\mathbf{Z} - \mathbf{Z}\mathbf{V}_0\mathbf{V}_0^* + \mathbf{U}_0\mathbf{U}_0^*\mathbf{Z}\mathbf{V}_0\mathbf{V}_0^* \\
&= \left(\mathbf{U}\mathbf{U}^* - \mathbf{U}_0\mathbf{U}_0^*\right) \mathbf{Z}(\mathbf{I} - \mathbf{V}_0\mathbf{V}_0^*) + (\mathbf{I} - \mathbf{U}\mathbf{U}^*)\mathbf{Z}(\mathbf{V}\mathbf{V}^* - \mathbf{V}_0\mathbf{V}_0^*),
\end{aligned}$$

which we use to estimate

$$\begin{aligned}
\|(\mathcal{P}_T - \mathcal{P}_{T_0})(\mathbf{Z})\|_F &\leq \|\mathbf{U}\mathbf{U}^* - \mathbf{U}_0\mathbf{U}_0^*\|_{S_\infty} \|\mathbf{Z}\|_F \|\mathbf{I} - \mathbf{V}_0\mathbf{V}_0^*\|_{S_\infty} + \|\mathbf{I} - \mathbf{U}\mathbf{U}^*\|_{S_\infty} \|\mathbf{Z}\|_F \|\mathbf{V}\mathbf{V}^* - \mathbf{V}_0\mathbf{V}_0^*\|_{S_\infty} \\
&\leq \frac{\|\mathcal{T}_r(\mathbf{X}) - \mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)} \|\mathbf{Z}\|_F \cdot 1 + 1 \cdot \|\mathbf{Z}\|_F \frac{\|\mathcal{T}_r(\mathbf{X}) - \mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)} \\
&\leq 2 \frac{\|\mathcal{T}_r(\mathbf{X}) - \mathbf{X}\|_{S_\infty} + \|\mathbf{X} - \mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)} \|\mathbf{Z}\|_F,
\end{aligned}$$

where $\mathcal{T}_r(\mathbf{X})$ is the best rank-r approximation (3.21). Here, we used the results

$$\|\mathbf{U}\mathbf{U}^* - \mathbf{U}_0\mathbf{U}_0^*\|_{S_\infty} \leq \frac{\|\mathcal{T}_r(\mathbf{X}) - \mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)}$$

and

$$\|\mathbf{V}\mathbf{V}^* - \mathbf{V}_0\mathbf{V}_0^*\|_{S_\infty} \leq \frac{\|\mathcal{T}_r(\mathbf{X}) - \mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)}$$

of [WCCL16a, Lemma 4.2, ineq. (4.3)], which bound the distance between the projections onto the left and right singular subspaces of $\mathbf{X}$ and $\mathbf{X}_0$.

From the Eckardt-Young-Mirsky theorem (3.21), it then follows that

$$\|(\mathcal{P}_T - \mathcal{P}_{T_0})\|_{S_\infty} \leq \frac{4\|\mathbf{X} - \mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)}. \tag{2.149}$$

With this, we further bound

$$\begin{aligned}
\|\mathcal{R}_\Omega\mathcal{P}_T\|_{S_\infty} &\leq \left\|\mathcal{P}_\Omega(\mathcal{P}_T - \mathcal{P}_{T_0})\right\|_{S_\infty} + \left\|\mathcal{R}_\Omega\mathcal{P}_{T_0}\right\|_{S_\infty} \\
&\leq \frac{16}{3}\log(D)\frac{4\,\|\mathbf{X}-\mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)} + \left\|\mathcal{R}_\Omega\mathcal{P}_{T_0}\right\|_{S_\infty} \\
&\leq \frac{16}{3}\log(D)\frac{\sqrt{3}}{8\sqrt{\log(D)}\sqrt{(1+\varepsilon)}}\varepsilon\sqrt{\frac{m}{d_1d_2}} + \sqrt{\frac{16}{3}\log(D)(1+\varepsilon)\frac{m}{d_1d_2}} \\
&= \frac{2}{\sqrt{3}}\sqrt{\log(D)}\frac{1}{\sqrt{(1+\varepsilon)}}\varepsilon\sqrt{\frac{m}{d_1d_2}} + \sqrt{\frac{16}{3}\log(D)(1+\varepsilon)\frac{m}{d_1d_2}} \\
&\leq 2\sqrt{3}\log(D)\sqrt{1+\varepsilon}\sqrt{\frac{m}{d_1d_2}},
\end{aligned} \tag{2.150}$$

where the second inequality follows from (2.149) and the third from condition (c). To prove the statement (2.147), we calculate

$$\begin{aligned}
\left\|\frac{d_1d_2}{m}\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T - \mathcal{P}_T\right\|_{S_\infty} &\leq \left\|\mathcal{P}_T - \mathcal{P}_{T_0}\right\|_{S_\infty} + \frac{d_1d_2}{m}\left\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_T - \mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_{T_0}\right\|_{S_\infty} \\
&\quad + \frac{d_1d_2}{m}\left\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_{T_0} - \mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0}\right\|_{S_\infty} + \left\|\mathcal{P}_{T_0} - \frac{d_1d_2}{m}\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0}\right\|_{S_\infty} \\
&\leq \left\|\mathcal{P}_T - \mathcal{P}_{T_0}\right\|_{S_\infty} + \frac{d_1d_2}{m}\left\|\mathcal{R}_\Omega\mathcal{P}_T\right\|_{S_\infty}\left\|\mathcal{P}_T - \mathcal{P}_{T_0}\right\|_{S_\infty} \\
&\quad + \frac{d_1d_2}{m}\left\|\mathcal{R}_\Omega\mathcal{P}_{T_0}\right\|_{S_\infty}\left\|\mathcal{P}_T - \mathcal{P}_{T_0}\right\|_{S_\infty} + \left\|\mathcal{P}_{T_0} - \frac{d_1d_2}{m}\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0}\right\|_{S_\infty} \\
&\leq \frac{4\,\|\mathbf{X}-\mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)} + \frac{d_1d_2}{m}\left\|\mathcal{R}_\Omega\mathcal{P}_T\right\|_{S_\infty}\frac{4\,\|\mathbf{X}-\mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)} \\
&\quad + \frac{d_1d_2}{m}\left\|\mathcal{R}_\Omega\mathcal{P}_{T_0}\right\|_{S_\infty}\frac{4\,\|\mathbf{X}-\mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)} + \left\|\mathcal{P}_{T_0} - \frac{d_1d_2}{m}\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0}\right\|_{S_\infty} \\
&\leq 4\varepsilon_0
\end{aligned}$$

where in the second inequality, we utilized the fact $\mathcal{P}_\Omega^* = \mathcal{P}_\Omega$ so that $\|\mathcal{P}_T\mathcal{P}_\Omega\|_{S_\infty} = \|\mathcal{P}_\Omega\mathcal{P}_T\|_{S_\infty}$. The very last estimate follows from conditions (b) and (c) and the bounds (2.148) and (2.150) for $\|\mathcal{P}_\Omega\mathcal{P}_T\|_{S_\infty}$ and $\left\|\mathcal{P}_\Omega\mathcal{P}_{T_0}\right\|_{S_\infty}$. □

We conclude this section with the proof of Corollary 2.5.3.

Proof of Corollary 2.5.3. Assume that there are m locations $\Omega = (i_\ell, j_\ell)_{\ell=1}^m$ in $[d_1]\times[d_2]$ sampled independently uniformly *with replacement*, where m fulfills (2.137) with $C := 7/\varepsilon^2$ and $\varepsilon = 0.1$. By Lemma 3.7.1, it follows that the corresponding operator $\mathcal{R}_\Omega : \mathbb{R}^{d_1\times d_2} \longrightarrow \mathbb{R}^{d_1\times d_2}$ from (2.140) fulfills

$$\|\mathcal{R}_\Omega\|_{S_\infty} \leq \frac{16}{3}\log(D) \tag{2.151}$$

on an event called E_Ω, which occurs with a probability of at least $1 - D^{-2}$, and by Lemma 3.7.2, the tangent space $T_0 = T_{fX_0}$ corresponding to the μ_0-incoherent rank-r matrix $\mathbf{X}_0$ fulfills

$$\left\|\frac{d_1d_2}{m}\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0} - \mathcal{P}_{T_0}\right\|_{S_\infty} \leq \varepsilon$$

on an event called E_{Ω,T_0}, which occurs with a probability of at least $1 - D^{-2}$. Let $\varepsilon = \frac{1}{10}$. If $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ is such that $\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \leq \widetilde{\xi}\sigma_r(\mathbf{X}_0)$ with

$$\widetilde{\xi} = \frac{\sqrt{3}}{32} \frac{\epsilon}{\sqrt{\log(D)}(1+\epsilon)} \sqrt{\frac{m}{d_1 d_2}} = \frac{\sqrt{3}}{32} \frac{1}{10\sqrt{\log(D)}(11/10)} \sqrt{\frac{m}{d_1 d_2}}, \tag{2.152}$$

it follows by Lemma 2.5.3 that on the event $E_\Omega \cap E_{\Omega,T_0}$, the tangent space $T_k := \mathbf{X}^{(k)}$ onto the rank-r manifold at $\mathbf{X}^{(k)}$ fulfills

$$\left\| \frac{d_1 d_2}{m} \mathscr{P}_{T_k} \mathscr{R}_\Omega \mathscr{P}_{T_k} - \mathscr{P}_{T_k} \right\|_{S_\infty} \leq 4\varepsilon. \tag{2.153}$$

Next, we claim that on the event $E_\Omega \cap E_{\Omega,T_0}$,

$$\|\eta\|_F \leq \sqrt{\frac{\widetilde{C} d \log(D)}{\mu_0 r}} \|\mathscr{P}_{T_k^\perp}(\eta)\|_F. \tag{2.154}$$

for any $\eta \in \ker \Phi$ with some constant $\widetilde{C}$, where $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ is the measurement operator (2.134) associated to $\Omega = (i_\ell, j_\ell)_{\ell=1}^m$.

Indeed, to show this claim, we first note that $\eta \in \ker \Phi$ if and only if $\eta \in \ker \mathscr{R}_\Omega$. Let $\eta \in \ker \mathscr{R}_\Omega$. Then

$$\begin{aligned}
\|\mathscr{P}_{T_k}(\eta)\|_F^2 &= \langle \mathscr{P}_{T_k}(\eta), \mathscr{P}_{T_k}(\eta) \rangle \\
&= \left\langle \mathscr{P}_{T_k}(\eta), \frac{d_1 d_2}{m} \mathscr{P}_{T_k} \mathscr{R}_\Omega \mathscr{P}_{T_k}(\eta) \right\rangle + \left\langle \mathscr{P}_{T_k}(\eta), \mathscr{P}_{T_k}(\eta) - \frac{d_1 d_2}{m} \mathscr{P}_{T_k} \mathscr{R}_\Omega \mathscr{P}_{T_k}(\eta) \right\rangle \\
&\leq \left\langle \mathscr{P}_{T_k}(\eta), \frac{d_1 d_2}{m} \mathscr{P}_{T_k} \mathscr{R}_\Omega \mathscr{P}_{T_k}(\eta) \right\rangle + \|\mathscr{P}_{T_k}(\eta)\|_F \left\| \mathscr{P}_{T_k} - \frac{d_1 d_2}{m} \mathscr{P}_{T_k} \mathscr{R}_\Omega \mathscr{P}_{T_k} \right\|_{S_\infty} \|\mathscr{P}_{T_k}(\eta)\|_F \\
&\leq \left\langle \mathscr{P}_{T_k}(\eta), \frac{d_1 d_2}{m} \mathscr{P}_{T_k} \mathscr{R}_\Omega \mathscr{P}_{T_k}(\eta) \right\rangle + 4\epsilon \|\mathscr{P}_{T_k}(\eta)\|_F^2,
\end{aligned}$$

using (2.153) in the last inequality, which implies that

$$\begin{aligned}
\|\mathscr{P}_{T_k}(\eta)\|_F^2 &\leq \frac{1}{1-4\epsilon} \frac{d_1 d_2}{m} \langle \mathscr{P}_{T_k}(\eta), \mathscr{P}_{T_k} \mathscr{R}_\Omega^2 \mathscr{P}_{T_k}(\eta) \rangle = \frac{1}{1-4\epsilon} \frac{d_1 d_2}{m} \|\mathscr{R}_\Omega \mathscr{P}_{T_k}(\eta)\|_F^2 \\
&\leq \frac{2 d_1 d_2}{m} \|\mathscr{R}_\Omega \mathscr{P}_{T_k}(\eta)\|_F^2
\end{aligned}$$

using the fact that $\mathscr{R}_\Omega : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ is positive semidefinite and has eigenvalues that are 0 or larger or equal than 1 only. Furthermore, we used that $\epsilon \leq \frac{1}{10}$ in the last inequality.

Since $\eta \in \ker \mathscr{R}_\Omega$, it holds that

$$0 = \|\mathscr{R}_\Omega(\eta)\|_F = \left\| \mathscr{R}_\Omega \left(\mathscr{P}_{T_k}(\eta) + \mathscr{P}_{T_k^\perp}(\eta) \right) \right\|_F \geq \|\mathscr{R}_\Omega \mathscr{P}_{T_k}(\eta)\|_F - \|\mathscr{R}_\Omega \mathscr{P}_{T_k^\perp}(\eta)\|_F$$

so that

$$\|\mathscr{R}_\Omega \mathscr{P}_{T_k}(\eta)\|_F \leq \|\mathscr{R}_\Omega \mathscr{P}_{T_k^\perp}(\eta)\|_F \leq \frac{16}{3} \log(D) \|\mathscr{P}_{T_k^\perp}(\eta)\|_F,$$

where we used (2.151) in the last inequality. Inserting this above, we obtain

$$\begin{aligned}
\|\eta\|_F^2 = \|\mathcal{P}_{T_k}(\eta)\|_F^2 + \|\mathcal{P}_{T_k^\perp}(\eta)\|_F^2 &\leq \left(\frac{2d_1d_2}{m}\frac{16^2}{3^2}\log(D)^2 + 1\right)\|\mathcal{P}_{T_k^\perp}(\eta)\|_F^2 \\
&\leq \left(\frac{2d_1d_2}{C\mu_0 r(d_1+d_2)\log(d_1+d_2)}\frac{16^2}{3^2}\log(D)^2 + 1\right)\|\mathcal{P}_{T_k^\perp}(\eta)\|_F^2 \\
&\leq \frac{\widetilde{C}d\log(D)}{\mu_0 r}\|\mathcal{P}_{T_k^\perp}(\eta)\|_F^2,
\end{aligned}$$

where we used the sample complexity condition (2.137) in the second inequality and the definition

$$\widetilde{C} := \frac{416^2}{C\cdot 3^2}$$

for the constant $\widetilde{C}$. This shows the claim (2.154). In fact, (2.154) is shown with this proof not only for T_k, but for all tangent spaces $T_{\mathbf{Z}}$ onto the rank-r manifold associated to matrices $\mathbf{Z} \in \mathbb{R}^{d_1\times d_2}$ fulfilling

$$\|\mathbf{Z} - \mathbf{X}_0\|_{S_\infty} \leq \widetilde{\xi}\sigma_r(\mathbf{X}_0)$$

with $\widetilde{\xi}$ as in (2.152). This means that Assumption 2.1 is shown on the event $E_\Omega \cap E_{\Omega,T_0}$ for the matrix $\mathbf{X}_0$ with radius

$$\begin{aligned}
\xi &:= \frac{C}{320}\sqrt{\frac{30}{11}}\sqrt{\frac{\mu_0 r}{d}} \leq \frac{\sqrt{3}}{32}\frac{1}{10\sqrt{\log(D)(11/10)}}\sqrt{\frac{C\mu_0 r(d_1+d_2)\log(d_1+d_2)}{d_1d_2}} \\
&\leq \widetilde{\xi} = \frac{\sqrt{3}}{32}\frac{1}{10\sqrt{\log(D)(11/10)}}\sqrt{\frac{m}{d_1d_2}}
\end{aligned}$$

and constant

$$c(r,d_1,d_2) = \sqrt{\frac{\widetilde{C}d\log(D)}{\mu_0 r}}.$$

To finish the proof of Corollary 2.5.3, we need to make sure that Theorem 2.7 is applicable, which means that $\widetilde{\mu}$ from (2.119), i.e.,

$$\widetilde{\mu} = \frac{c(r,d_1,d_2)^2 r\,(1+4\kappa+2\zeta^p\kappa)}{(1-\zeta)^{2-p}\sigma_r(\mathbf{X}_0)^{1-p}}$$

with $\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \leq \zeta\sigma_r(\mathbf{X}_0)$ is small enough such that

$$\widetilde{\mu}\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty}^{1-p} < 1.$$

If $\zeta \leq 1/2$, then

$$\begin{aligned}
\widetilde{\mu}\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty}^{1-p} &\leq 2^{2-p}c(r,d_1,d_2)^2 r\,(1+4\kappa+2\zeta^p\kappa)\,\zeta^{1-p} \\
&\leq 2^{2-p}\frac{\widetilde{C}d\log(D)}{\mu_0 r}r(1+6\kappa)\zeta^{1-p} = 2^{2-p}\frac{\widetilde{C}d\log(D)}{\mu_0}(1+6\kappa)\zeta^{1-p} < 1,
\end{aligned}$$

where the last inequality holds if

$$\zeta < \left(\frac{\mu_0}{2^{2-p}\widetilde{C}d\log(D)(1+6\kappa)}\right)^{\frac{1}{1-p}}.$$

This means that the statement of Corollary 2.5.3 is true with constant

$$\mu = 2^{2-p}\frac{\widetilde{C}d\log(D)}{\mu_0 r\sigma_r(\mathbf{X}_0)^{1-p}}r(1+6\kappa)$$

if

$$\begin{aligned}\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} &\leq \min\left(C_1\sqrt{\frac{\mu_0 r}{d}}, \left(C_2\frac{\mu_0}{d\log(D)\kappa}\right)^{\frac{1}{1-p}}\right)\sigma_r(\mathbf{X}_0) \\ &\leq \min(\zeta,\xi)\sigma_r(\mathbf{X}_0) = \min\left(\frac{C}{320}\sqrt{\frac{30}{11}}\sqrt{\frac{\mu_0 r}{d}}, \left(\frac{\mu_0}{2^{2-p}\widetilde{C}d\log(D)(1+6\kappa)}\right)^{\frac{1}{1-p}}\right)\sigma_r(\mathbf{X}_0)\end{aligned}$$

for constants $C_1 = \frac{C}{320}\sqrt{\frac{30}{11}}$ and $C_2 = \frac{1}{12\cdot 2^{2-p}\widetilde{C}}$, with probability at least $1 - 2D^{-2}$ (which is a lower bound for the probability that the event $E_\Omega \cap E_{\Omega,T_0}$ occurs), if the sampling model of Ω is one of uniformly sampling with replacement. By the argument of [CRT06a, Section II.C], [Rec11, Proposition 3], the result extends to the sampling model of independent locations without replacement with the same probability bound. This concludes the proof of Corollary 2.5.3. □

2.6 Computational Considerations

In this chapter so far we presented an algorithm, called `MatrixIRLS`, and analyzed its local convergence suggesting superlinear convergence rates of the iterates to a low-rank solution under certain assumptions. To make the approach scalable to real-world applications involving not only very small data sets, it is not sufficient that the algorithm needs only few iterations to obtain a high precision solution. In fact, it is also desirable that the iterations are computationally efficient in terms of time and space complexity—one iteration should not take too many basic arithmetic operators, and the iterates should not be stored in a way that creates unnecessary memory overhead or that uses more parameter to describe the data than necessary.

What does this mean for low-rank matrices and their recovery? In Section 2.1.1, it became already evident that a $(d_1 \times d_2)$-dimensional matrix $\mathbf{X}$ of rank r can be described such that

$$\mathbf{X} = \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \dots & \mathbf{u}_r \\ | & & | \end{pmatrix} \cdot \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} \cdot \begin{pmatrix} -- & \mathbf{v}_1^* & -- \\ & \vdots & \\ -- & \mathbf{v}_r^* & -- \end{pmatrix} = \mathbf{U\Sigma V}^*$$

using its singular value decomposition with a $(d_1 \times r)$-matrix $\mathbf{U}$, a $(d_2 \times r)$-matrix $\mathbf{V}$, and a diagonal $(r \times r)$-matrix $\mathbf{\Sigma}$. While the added number of entries of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{\Sigma}$ amounts to

$$d_1 r + d_2 r + r^2 = r(d_1 + d_2 + r),$$

we saw already that its actual *number of degrees of freedom* is slightly lower with

$$\deg_f = r(d_1 + d_2 - r),$$

as the orthogonality and normalization of the columns of $\mathbf{U}$ and $\mathbf{V}$ can be taken into account, and the knowledge about diagonality of $\mathbf{\Sigma}$ also indicates that it has, in fact, only r non-zero values to be stored due to its diagonal structure.

If $D = \max(d_1, d_2)$, both quantities are of order

$$O(Dr).$$

In most applications where low-rank matrix recovery can be used, for example, to build a model for a recommender system [KBV09, BL07] (matrix completion) or to reconstruct a 2D image from phaseless measurements in ptychography [TLRW14] (phase retrieval), good models correspond already to very low ranks

$$r \ll d = \min(d_1, d_2)$$

compared to the ambient dimensions d_1 and d_2. The *Netflix prize dataset* [BL07] for example, which is already more than a decade old and might not represent a typical size of training data sets for contemporary commercial recommender systems, consists of around $d_1 = 480000$ users and $d_2 = 17000$ movies, but models with ranks r in the dozens returned already very competitive results [KBV09, RR13]. In ptychography, typical test 2D test images are of the size 500×500, resulting in $(d \times d)$-dimensional recovery problems of matrices of rank $r = 1$ with $d = 250000$ [XSH⁺18].

From these examples it becomes clear that there are many applications where it is *prohibitive* that any algorithm of choice, iterative or non-iterative,

- stores full $(d_1 \times d_2)$-dimensional matrices $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$,
- uses any operations that scale *quadratically* in the matrix dimensions d_1 and/or d_2,

as the resulting time or space complexity of

$$\Omega(d_1 d_2)$$

is by orders of magnitudes worse than what do expect from algorithms that have complexities that are *basically linear* in D and *polynomial* in r.

While we do not provide a *global* complexity analysis of `MatrixIRLS` due to the *local* nature of our convergence analysis, we detail in this section why `MatrixIRLS` avoids any quadratic dependency on the matrix dimensions in its memory and iteration costs.

2.6.1 Low-Rank Structure of Weight Operator

We recall that the step of calculating the new iterate $\mathbf{X}^{(k+1)} \in \mathbb{R}^{d_1 \times d_2}$ as the solution of the weighted least squares problem

$$\min_{\Phi(\mathbf{X})=\mathbf{y}} \langle \mathbf{X}, W^{(k)}(\mathbf{X}) \rangle.$$

defined by the weight operator $W^k : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ (2.57) associated to the last iterate $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ and the problem data (measurement operator $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ and measurement vector $\mathbf{y} \in \mathbb{R}^m$) can be calculated such that

$$\mathbf{X}^{(k+1)} = (W^{(k)})^{-1} \left(\Phi^* \left(\Phi (W^{(k)})^{-1} \Phi^* \right)^{-1} (\mathbf{y}) \right),$$

as it was stated in Lemma 2.4.1. Here, $(W^{(k)})^{-1} : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ is the *inverse* of the matrix operator $W^{(k)}$–the inverse exists as $W^{(k)}$ is positive definite with eigenvalues (see (2.60))

$$\lambda(W^{(k)}) = \left\{ (\mathbf{H}_1^{(k)})_{ij} : i \in [d_1], j \in [d_2], i \leq j \right\} \cup \left\{ (\mathbf{H}_2^{(k)})_{ij} : i \in [d_1], j \in [d_1], i < j \right\}.$$

It is a priori not clear why there is hope to calculate $\mathbf{X}^{(k+1)}$ in a time complexity that is quadratic or sub-quadratic in the dimension of $\mathbf{X}^{(k+1)}$, even if iterative solvers are used to solve, for example, the linear system

$$\left(\Phi (W^{(k)})^{-1} \Phi^* \right) \mathbf{z} = \mathbf{y}$$

for $\mathbf{z} \in \mathbb{R}^m$ before $\mathbf{X}^{(k+1)}$ can be calculated such that $\mathbf{X}^{(k+1)} = (W^{(k)})^{-1}(\Phi^*(\mathbf{z}))$.

In particular, just the application of a matrix operator mapping from $\mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ to a $(d_1 \times d_2)$-matrix can cost as much as

$$O((d_1 d_2) \cdot (d_1 d_2)) = O(d^2 D^2).$$

However, as a preliminary result, we state the following result about cost of the application of any power of a matrix operator $W^{(k)}$.

Lemma 2.6.1 (Time complexity of the application of weight operators). *Let $k, r \in \mathbb{N}$, $\nu \in \mathbb{R}$ and $\epsilon_k > 0$, let $M^{(k)} : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ be the ν-th power of the weight operator $W^{(k)}$ of* `MatrixIRLS` *as defined in (2.57) of Definition 2.3.4 associated to the matrix $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ that has at most r singular values larger than $\epsilon_k > 0$ with singular value decomposition*

$$\mathbf{X}^{(k)} = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^* = \begin{bmatrix} \mathbf{U}^{(k)} & \mathbf{U}_\perp^{(k)} \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}^{(k)} & 0 \\ 0 & \mathbf{\Sigma}_\perp^{(k)} \end{bmatrix} \begin{bmatrix} \mathbf{V}^{(k)*} \\ \mathbf{V}_\perp^{(k)*} \end{bmatrix},$$

where $\mathbf{U}_k \in \mathbb{O}^{d_1}$, $\mathbf{V}_k \in \mathbb{O}^{d_2}$, $\mathbf{U}^{(k)} \in \mathbb{R}^{d_1 \times r}$, $\mathbf{U}_\perp^{(k)} \in \mathbb{R}^{d_1 \times (d_1 - r)}$, $\mathbf{V}^{(k)} \in \mathbb{R}^{d_2 \times r}$, $\mathbf{V}_\perp^{(k)} \in \mathbb{R}^{d_2 \times (d_2 - r)}$, $\mathbf{\Sigma}^{(k)} = \mathrm{diag}(\sigma_1(\mathbf{X}^{(k)}), \ldots \sigma_r(\mathbf{X}^{(k)}))$ and $\mathbf{\Sigma}_\perp^{(k)} = \mathrm{dg}(\sigma_{r+1}(\mathbf{X}^{(k)}), \ldots \sigma_d(\mathbf{X}^{(k)}))$.

Then for any $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$,

$$\begin{aligned} M^{(k)}(\mathbf{Z}) = {} & \mathbf{U}^{(k)}[(\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}}^{\nu} \circ \widetilde{S}(\mathbf{U}^{(k)*}\mathbf{Z}\mathbf{V}^{(k)}) + (\mathbf{H}_2^{(k)})_{\mathbf{U},\mathbf{V}}^{\nu} \circ \widetilde{T}(\mathbf{U}^{(k)*}\mathbf{Z}\mathbf{V}^{(k)})]\mathbf{V}^{(k)*} \\ & + \mathbf{U}^{(k)} \left((\mathbf{D}_1^{(k)})^{\nu} + (\mathbf{D}_2^{(k)})^{\nu} \right) \mathbf{U}^{(k)*}\mathbf{Z}(\mathbf{I} - \mathbf{V}^{(k)}\mathbf{V}^{(k)*}) \\ & + (\mathbf{I} - \mathbf{U}^{(k)}\mathbf{U}^{(k)*})\mathbf{Z}\mathbf{V}^{(k)} \left((\mathbf{D}_1^{(k)})^{\nu} + (\mathbf{D}_2^{(k)})^{\nu} \right) \mathbf{V}^{(k)*} \\ & + \epsilon_k^{\nu(p-2)} \left(\mathbf{I} - \mathbf{U}^{(k)}\mathbf{U}^{(k)*} \right) \mathbf{Z} \left(\mathbf{I} - \mathbf{V}^{(k)}\mathbf{V}^{(k)*} \right), \end{aligned} \tag{2.155}$$

where $\mathbf{H}_1^{(k)}, \mathbf{H}_2^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ are the matrices from (2.58) and (2.59), $(\mathbf{H}_1^{(k)})^{\nu}, (\mathbf{H}_2^{(k)})^{\nu} \in \mathbb{R}^{d_1 \times d_2}$ are *their* entrywise *ν-th powers with decompositions*[5]

$$(\mathbf{H}_1^{(k)})^\nu = \begin{bmatrix} (\mathbf{H}_1^{(k)})^\nu_{\mathbf{U},\mathbf{V}} & (\mathbf{H}_1^{(k)})^\nu_{\mathbf{U},\mathbf{V}_\perp} \\ (\mathbf{H}_1^{(k)})^\nu_{\mathbf{U}_\perp,\mathbf{V}} & (\mathbf{H}_1^{(k)})^\nu_{\mathbf{U}_\perp,\mathbf{V}_\perp} \end{bmatrix}, \qquad (\mathbf{H}_2^{(k)})^\nu = \begin{bmatrix} (\mathbf{H}_2^{(k)})^\nu_{\mathbf{U},\mathbf{V}} & (\mathbf{H}_2^{(k)})^\nu_{\mathbf{U},\mathbf{V}_\perp} \\ (\mathbf{H}_2^{(k)})^\nu_{\mathbf{U}_\perp,\mathbf{V}} & (\mathbf{H}_2^{(k)})^\nu_{\mathbf{U}_\perp,\mathbf{V}_\perp} \end{bmatrix},$$

and where $(\mathbf{D}_1^{(k)})^\nu$, $(\mathbf{D}_1^{(k)})^\nu$ are $(r \times r)$ diagonal matrices such that

$$\left((\mathbf{D}_\ell^{(k)})^\nu\right)_{ii} = \left((\mathbf{H}_\ell^{(k)})^\nu_{\mathbf{U},\mathbf{V}_\perp}\right)_{i1}$$

for $\ell = 1, 2$ for all $i \in [r]$.

In particular, this means that the image of $\mathbf{Z}$ with respect to $M^{(k)}$, $M^{(k)}(\mathbf{Z})$, can be computed in a time complexity of

$$O(d_1 d_2 r).$$

Proof. Since the eigenvalues of $W^{(k)}$ are as stated above and $W^{(k)}$ is diagonalized by the outer product $\mathbf{V}^{(k)} \otimes \mathbf{U}^{(k)}$, it holds that for any $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$,

$$\begin{aligned}
&M^{(k)}(\mathbf{Z}) = \mathbf{U}_k \left[(\mathbf{H}_1^{(k)})^\nu \circ \widetilde{S}(\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k) + (\mathbf{H}_2^{(k)})^\nu \circ \widetilde{T}(\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k)\right] \mathbf{V}_k^* = \\
&\begin{pmatrix} \mathbf{U}^{(k)} & \mathbf{U}_\perp^{(k)} \end{pmatrix} \left[(\mathbf{H}_1^{(k)})^\nu \circ \widetilde{S}\left(\begin{bmatrix} \mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)} & \mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)} \\ \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)} & \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)} \end{bmatrix}\right) + (\mathbf{H}_2^{(k)})^\nu \circ \widetilde{T}\left(\begin{bmatrix} \mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)} & \mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)} \\ \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)} & \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)} \end{bmatrix}\right)\right] \begin{pmatrix} \mathbf{V}^{(k)*} \\ \mathbf{V}_\perp^{(k)*} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{U}^{(k)} & \mathbf{U}_\perp^{(k)} \end{pmatrix} \Bigg[\begin{bmatrix} (\mathbf{H}_1^{(k)})^\nu_{\mathbf{U},\mathbf{V}} \circ \widetilde{S}\left(\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)}\right) + (\mathbf{H}_2^{(k)})^\nu_{U,V} \circ \widetilde{T}\left(\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)}\right) & 0 \\ 0 & 0 \end{bmatrix} \\
&+ \begin{bmatrix} & (\mathbf{H}_1^{(k)})^\nu_{\mathbf{U},\mathbf{V}_\perp} \\ (\mathbf{H}_1^{(k)})^\nu_{\mathbf{U}_\perp,\mathbf{V}} & \end{bmatrix} \circ \widetilde{S}\left(\begin{bmatrix} 0 & \mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)} \\ \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)} & 0 \end{bmatrix}\right) \\
&+ \begin{bmatrix} & (\mathbf{H}_2^{(k)})^\nu_{\mathbf{U},\mathbf{V}_\perp} \\ (\mathbf{H}_2^{(k)})^\nu_{\mathbf{U}_\perp,\mathbf{V}} & \end{bmatrix} \circ \widetilde{T}\left(\begin{bmatrix} 0 & \mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)} \\ \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)} & 0 \end{bmatrix}\right) \\
&+ \epsilon_k^{\nu(p-2)} \begin{bmatrix} 0 & 0 \\ 0 & \widetilde{S}\left(\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\right) + \widetilde{T}\left(\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\right) \end{bmatrix} \Bigg] \begin{pmatrix} \mathbf{V}^{(k)*} \\ \mathbf{V}_\perp^{(k)*} \end{pmatrix} \\
&=: \begin{pmatrix} \mathbf{U}^{(k)} & \mathbf{U}_\perp^{(k)} \end{pmatrix} [(\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}) + (\mathrm{IV})] \begin{pmatrix} \mathbf{V}^{(k)*} \\ \mathbf{V}_\perp^{(k)*} \end{pmatrix}.
\end{aligned}$$

Here, we used crucially that due to the definition of $\mathbf{H}_1^{(k)}$ and $\mathbf{H}_2^{(k)}$, all the entries of $(\mathbf{H}_1^{(k)})_{\mathbf{U}_\perp,\mathbf{V}_\perp}$ and $(\mathbf{H}_1^{(k)})_{\mathbf{U}_\perp,\mathbf{V}_\perp}$ are equal to ϵ^{p-2}, and therefore, the entries for their ν-th entrywise power are equal to $\epsilon_k^{\nu(p-2)}$.

From this calculation, we see that the first summand corresponds to $\mathbf{U}^{(k)}[(\mathbf{H}_1^{(k)})^\nu_{\mathbf{U},\mathbf{V}} \circ \widetilde{S}(\mathbf{U}^{(k)*}\mathbf{Z}\mathbf{V}^{(k)}) + (\mathbf{H}_2^{(k)})^\nu_{\mathbf{U},\mathbf{V}} \circ \widetilde{T}(\mathbf{U}^{(k)*}\mathbf{Z}\mathbf{V}^{(k)})]\mathbf{V}^{(k)*}$.

For the fourth summand, since

$$\widetilde{S}\left(\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\right) + \widetilde{T}\left(\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\right) = \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)},$$

[5] See the proof of Lemma 2.4.4 for the same decompositions.

we have that

$$\begin{aligned}&\begin{pmatrix}\mathbf{U}^{(k)} & \mathbf{U}_{\perp}^{(k)}\end{pmatrix} \epsilon_k^{\nu(p-2)} \begin{bmatrix}0 & 0 \\ 0 & \widetilde{S}\left(\mathbf{U}_{\perp}^{(k)*}\mathbf{X}_0\mathbf{V}_{\perp}^{(k)}\right) + \widetilde{T}\left(\mathbf{U}_{\perp}^{(k)*}\mathbf{X}_0\mathbf{V}_{\perp}^{(k)}\right)\end{bmatrix} \begin{pmatrix}\mathbf{V}^{(k)*} \\ \mathbf{V}_{\perp}^{(k)*}\end{pmatrix} \\ &= \epsilon_k^{\nu(p-2)}\mathbf{U}_{\perp}^{(k)}\mathbf{U}_{\perp}^{(k)*}\mathbf{X}_0\mathbf{V}_{\perp}^{(k)}\mathbf{V}_{\perp}^{(k)*} = \epsilon_k^{\nu(p-2)}\left(\mathbf{I} - \mathbf{U}^{(k)}\mathbf{U}^{(k)*}\right)\mathbf{X}_0\left(\mathbf{I} - \mathbf{V}^{(k)}\mathbf{V}^{(k)*}\right),\end{aligned}$$

using that

$$\mathbf{U}^{(k)}\mathbf{U}^{(k)*} + \mathbf{U}_{\perp}^{(k)}\mathbf{U}_{\perp}^{(k)*} = \mathbf{I} \quad \text{and} \quad \mathbf{V}^{(k)}\mathbf{V}^{(k)*} + \mathbf{V}_{\perp}^{(k)}\mathbf{V}_{\perp}^{(k)*} = \mathbf{I} \tag{2.156}$$

in the last equality.

To cope with the terms $\mathbf{U}_k\left[(\mathrm{II}) + (\mathrm{III})\right]\mathbf{V}_k^*$, we note that the matrices

$$(\mathbf{H}_{\ell}^{(k)})_{\mathbf{U},\mathbf{V}_{\perp}}^{\nu}$$

have constant rows for $\ell = 1, 2$ (same for the columns of $(\mathbf{H}_{\ell}^{(k)})_{\mathbf{U}_{\perp},\mathbf{V}}^{\nu}$), so that their action by the Schur product $\circ$ can be realized by corresponding multiplication with the matrices $(D_{\ell}^{(k)})^{\nu}$. Also here, we use (2.156), which shows formula (2.155).

To obtain the bound on the number of basic arithmetic operations for the calculation of $M^{(k)}(\mathbf{X})$, we can calculate first $\mathbf{U}^{(k)*}\mathbf{Z} \in \mathbb{R}^{r \times d_2}$ and $\mathbf{Z}\mathbf{V}^{(k)} \in \mathbb{R}^{d_1 timesr}$, and these matrix multiplications are of order $O(rd_1d_2)$ using the standard algorithm. Calculating then the $(r \times r)$ matrix $\mathbf{U}^{(k)*}\mathbf{Z}\mathbf{V}^{(k)}$ costs $O(r^2 d)$, and all operations needed to calculate

$$\widetilde{X}_1 := (\mathbf{H}_1^{(k)})_{\mathbf{U},\mathbf{V}}^{\nu} \circ \widetilde{S}(\mathbf{U}^{(k)*}\mathbf{Z}\mathbf{V}^{(k)}) + (\mathbf{H}_2^{(k)})_{\mathbf{U},\mathbf{V}}^{\nu} \circ \widetilde{T}(\mathbf{U}^{(k)*}\mathbf{Z}\mathbf{V}^{(k)}) \in \mathbb{R}^{r \times r}$$

are $O(r^2)$. Furthermore, the diagonal multiplications needed to obtain

$$\widetilde{X}_2 := \left((\mathbf{D}_1^{(k)})^{\nu} + (\mathbf{D}_2^{(k)})^{\nu}\right)\mathbf{U}^{(k)*}\mathbf{Z} \in \mathbb{R}^{r \times d_2}$$

and

$$\widetilde{X}_3 := \mathbf{Z}\mathbf{V}^{(k)}\left((\mathbf{D}_1^{(k)})^{\nu} + (\mathbf{D}_2^{(k)})^{\nu}\right) \in \mathbb{R}^{d_1 \times r}$$

are $O(rD)$. The remaining operations consist then of additions and subtractions, and of multiplications of $\widetilde{X}_1, \widetilde{X}_2$ and $\widetilde{X}_3$ with $\mathbf{U}^{(k)}$ from the left and $\mathbf{V}^{(k)*}$ from the right, which cost again $O(rd_1d_2)$. □

In fact, a close look at the formula of $M^{(k)}$ from Lemma 2.6.1 reveals an interesting structure that relates its action to the tangent space onto the rank-r manifold at $\mathbf{X}^{(k)}$,

$$T_k := \{\mathbf{U}\mathbf{M}^* + \widetilde{\mathbf{M}}V^* : \ \mathbf{M} \in \mathbb{R}^{d_2 \times r}, \ \widetilde{\mathbf{M}} \in \mathbb{R}^{d_1 \times r}\},$$

which can be also parametrized such that $\mathbf{Z} \in T_k$ if and only if there exist matrices $\mathbf{Z}_1 \in \mathbb{R}^{r \times r}$, $\mathbf{Z}_2 \in \mathbb{R}^{r \times d_1}$, $\mathbf{Z}_3 \in \mathbb{R}^{d_1 \times r}$ such that

$$\mathbf{Z} = \mathbf{U}^{(k)}\mathbf{Z}_1\mathbf{V}^{(k)*} + \mathbf{U}^{(k)}\mathbf{Z}_2(\mathbf{I} - \mathbf{V}^{(k)}\mathbf{V}^{(k)*}) + (\mathbf{I} - \mathbf{U}^{(k)}\mathbf{U}^{(k)*})\mathbf{Z}_3\mathbf{V}^{(k)*}. \tag{2.157}$$

Comparing this parametrization to (2.155), we observe that $M^{(k)}$ acts on $\mathbf{Z} = \mathscr{P}_{T_k}(\mathbf{Z}) + \mathscr{P}_{T_k^{perp}}(\mathbf{Z})$

such that

$$\begin{aligned} M^{(k)}(\mathbf{Z}) &= P_{T_k}\mathscr{D}^{\nu}_{H^{(k)}}P^*_{T_k}(\mathbf{Z}) + \epsilon_k^{\nu(p-2)}\mathscr{P}_{T_k^\perp}(\mathbf{Z}) \\ &= P_{T_k}\left(\mathscr{D}^{\nu}_{H^{(k)}} - \epsilon_k^{\nu(p-2)}\mathbf{I}_{T_k}\right)P^*_{T_k}(\mathbf{Z}) + \epsilon_k^{\nu(p-2)}\mathbf{Z}, \end{aligned} \tag{2.158}$$

where $P_{T_k} : T_k \longrightarrow \mathbb{R}^{d_1\times d_2}$ is a linear operator corresponding to an orthonormal basis of T_k, so that we have for the projection $\mathscr{P}_{T_k} : \mathbb{R}^{d_1\times d_2} \longrightarrow \mathbb{R}^{d_1\times d_2}$ that

$$\mathscr{P}_{T_k} = P_{T_k}P^*_{T_k},$$

and $\mathscr{D}^{\nu}_{H^{(k)}} : T_k \longrightarrow T_k$ is an (almost) diagonal operator that maps any $\mathbf{Z} \in T_k$ written as

$$\mathbf{Z} = \mathbf{U}^{(k)}\mathbf{Z}_1\mathbf{V}^{(k)*} + \mathbf{U}^{(k)}\mathbf{Z}_2(\mathbf{I} - \mathbf{V}^{(k)}\mathbf{V}^{(k)*}) + (\mathbf{I} - \mathbf{U}^{(k)}\mathbf{U}^{(k)*})\mathbf{Z}_3\mathbf{V}^{(k)*}$$

to

$$\begin{aligned} \mathscr{D}^{\nu}_{H^{(k)}}(\mathbf{Z}) = \mathbf{U}^{(k)}&\left[(\mathbf{H}_1^{(k)})^{\nu}_{\mathbf{U},\mathbf{V}} \circ \widetilde{S}(\mathbf{Z}_1) + (\mathbf{H}_2^{(k)})^{\nu}_{\mathbf{U},\mathbf{V}} \circ \widetilde{T}(\mathbf{Z}_1)\right]\mathbf{V}^{(k)*} \\ &+ \mathbf{U}^{(k)}\left((\mathbf{D}_1^{(k)})^{\nu} + (\mathbf{D}_2^{(k)})^{\nu}\right)\mathbf{Z}_2(\mathbf{I} - \mathbf{V}^{(k)}\mathbf{V}^{(k)*}) + (\mathbf{I} - \mathbf{U}^{(k)}\mathbf{U}^{(k)*})\mathbf{Z}_3\left((\mathbf{D}_1^{(k)})^{\nu} + (\mathbf{D}_2^{(k)})^{\nu}\right)\mathbf{V}^{(k)*}. \end{aligned} \tag{2.159}$$

We formalize the computational implication of this observation in the following lemma.

Lemma 2.6.2. *The operator $M^{(k)} : \mathbb{R}^{d_1\times d_2} \longrightarrow \mathbb{R}^{d_1\times d_2}$ from Lemma 2.6.1 can be written such that*

$$M^{(k)} = P_{T_k}\left(\mathscr{D}^{\nu}_{H^{(k)}} - \epsilon_k^{\nu(p-2)}\mathbf{I}_{T_k}\right)P^*_{T_k} + \epsilon_k^{\nu(p-2)}\mathbf{I},$$

where $\mathscr{D}^{\nu}_{H^{(k)}} : T_k \longrightarrow T_k$ is the operator from (2.159).

If elements of T_k are appropriately parametrized, the action of $\mathscr{D}^{\nu}_{H^{(k)}}$ has a time complexity of

$$O(rD).$$

2.6.2 Harnessing Low-Rank Structure for Updates with Sub-Quadratic Complexity

In this section, we explain how the observation of Lemma 2.6.2 can be used for efficient computations in `MatrixIRLS`. Using the definition of $M^{(k)}$ from Lemma 2.6.1 with exponent $\nu = -1$, we rewrite formula (2.93) for the new iterate $\mathbf{X}^{(k+1)}$ such that

$$\mathbf{X}^{(k+1)} = (W^{(k)})^{-1}\left(\Phi^*\left(\Phi(W^{(k)})^{-1}\Phi^*\right)^{-1}(\mathbf{y})\right) = M^{(k)}\left(\Phi^*\left(\Phi M^{(k)}\Phi^*\right)^{-1}(\mathbf{y})\right).$$

As above, if we define $\mathbf{z} := \left(\Phi M^{(k)}\Phi^*\right)^{-1}(\mathbf{y}) \in \mathbb{R}^m$ as the solution of

$$\left(\Phi M^{(k)}\Phi^*\right)\mathbf{z} = \mathbf{y},$$

we know that

$$\mathbf{X}^{(k+1)} = M^{(k)}\Phi^*(\mathbf{z}). \tag{2.160}$$

Using Lemma 2.6.2, we know that

$$\left(\Phi M^{(k)}\Phi^*\right)^{-1} = \left(\Phi P_{T_k}\left(\mathscr{D}^{-1}_{H^{(k)}} - \epsilon_k^{2-p}\mathbf{I}_{T_k}\right)P^*_{T_k}\Phi^* + \epsilon_k^{2-p}\Phi\Phi^*\right)^{-1}.$$

Next, we use the *Sherman-Morrison-Woodbury* formula [Woo50, FRW11], [HJ12, (0.7.4.1)], which states that for any invertible $\mathbf{B} \in \mathbb{R}^{m\times m}$, $\mathbf{C} \in \mathbb{R}^{\ell}$ and any matrices $\mathbf{E}, \mathbf{F} \in \mathbb{R}^{m\times \ell}$,

$$(\mathbf{B} + \mathbf{E}\mathbf{C}\mathbf{F}^*)^{-1} = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{E}(\mathbf{C}^{-1} + \mathbf{F}^*\mathbf{B}^{-1}\mathbf{E})^{-1}\mathbf{F}^*\mathbf{B}^{-1}. \tag{2.161}$$

Using (2.161) with $\mathbf{B} = \epsilon_k^{2-p}\Phi\Phi^*$, $\mathbf{C} = \left(\mathscr{D}_{H^{(k)}}^{-1} - \epsilon_k^{2-p}\mathbf{I}_{T_k}\right)$ and $\mathbf{E} = \mathbf{F} = \Phi P_{T_k}$, we obtain that

$$\begin{aligned}\mathbf{z} &= \left(\Phi M^{(k)}\Phi^*\right)^{-1}(\mathbf{y}) \\ &= \left[\epsilon_k^{p-2}\mathbf{I} - (\epsilon_k^{2-p}\Phi\Phi^*)^{-1}\Phi P_{T_k}\left(\epsilon_k^{2-p}\mathbf{C}^{-1} + P_{T_k}^*\Phi^*(\Phi\Phi^*)^{-1}\Phi P_{T_k}\right)^{-1} P_{T_k}^*\Phi^*\right](\Phi\Phi^*)^{-1}(\mathbf{y}).\end{aligned} \tag{2.162}$$

Furthermore, inserting the formula for $M^{(k)}$ from Lemma 2.6.2 in (2.160), we obtain

$$\mathbf{X}^{(k+1)} = \left[P_{T_k}\left(\mathscr{D}_{H^{(k)}}^{-1} - \epsilon_k^{2-p}\mathbf{I}_{T_k}\right)P_{T_k}^*\Phi^* + \epsilon_k^{2-p}\Phi^*\right](\mathbf{z}). \tag{2.163}$$

If we now apply the projector $P_{T_k}^*$ to this equality, restricting the equation to T_k, it follows that

$$\begin{aligned}\mathbf{X}_{T_k}^{(k+1)} &:= P_{T_k}^*(\mathbf{X}^{(k+1)}) = P_{T_k}^*P_{T_k}\mathbf{C}P_{T_k}^*\Phi^*(\mathbf{y}) + \epsilon_k^{2-p}P_{T_k}^*\Phi^*(\mathbf{z}) \\ &= \mathbf{C}P_{T_k}^*\Phi^*(\mathbf{y}) + \epsilon_k^{2-p}P_{T_k}^*\Phi^*(\mathbf{z}) = \left(\mathbf{C} + \epsilon_k^{2-p}\mathbf{I}\right)P_{T_k}^*\Phi^*(\mathbf{y}) \\ &= \mathscr{D}_{H^{(k)}}^{-1}P_{T_k}^*\Phi^*(\mathbf{z}).\end{aligned}$$

Inserting now (2.162) for $\mathbf{z}$, we obtain

$$\begin{aligned}\mathbf{X}_{T_k}^{(k+1)} &= \mathscr{D}_{H^{(k)}}^{-1}\epsilon_k^{p-2}\left[P_{T_k}^*\Phi^* - P_{T_k}^*\Phi^*(\Phi\Phi^*)^{-1}\Phi P_{T_k}\left(\epsilon_k^{2-p}\mathbf{C}^{-1} + P_{T_k}^*\Phi^*(\Phi\Phi^*)^{-1}\Phi P_{T_k}\right)^{-1} P_{T_k}^*\Phi^*\right](\Phi\Phi^*)^{-1}(\mathbf{y}) \\ &= \mathscr{D}_{H^{(k)}}^{-1}\epsilon_k^{p-2}\left[\mathbf{I} - P_{T_k}^*\Phi^*(\Phi\Phi^*)^{-1}\Phi P_{T_k}\left(\epsilon_k^{2-p}\mathbf{C}^{-1} + P_{T_k}^*\Phi^*(\Phi\Phi^*)^{-1}\Phi P_{T_k}\right)^{-1}\right]P_{T_k}^*\Phi^*(\Phi\Phi^*)^{-1}(\mathbf{y}) \\ &= \mathscr{D}_{H^{(k)}}^{-1}\epsilon_k^{p-2}\left[\epsilon_k^{2-p}\mathbf{C}^{-1}\left(\epsilon_k^{2-p}\mathbf{C}^{-1} + P_{T_k}^*\Phi^*(\Phi\Phi^*)^{-1}\Phi P_{T_k}\right)^{-1}\right]P_{T_k}^*\Phi^*(\Phi\Phi^*)^{-1}(\mathbf{y}) \\ &= \left(\mathbf{I} + \epsilon_k^{2-p}\mathbf{C}^{-1}\right)\left(\epsilon_k^{2-p}\mathbf{C}^{-1} + P_{T_k}^*\Phi^*(\Phi\Phi^*)^{-1}\Phi P_{T_k}\right)^{-1} P_{T_k}^*\Phi^*(\Phi\Phi^*)^{-1}(\mathbf{y}) \\ &=: \left(\mathbf{I} + \epsilon_k^{2-p}\mathbf{C}^{-1}\right)\gamma_k,\end{aligned} \tag{2.164}$$

where γ_k is the solution of the $\dim(T_k) \times \dim(T_k) \approx O(rD) \times O(rD)$ positive definite linear system

$$\left(\epsilon_k^{2-p}\mathbf{C}^{-1} + P_{T_k}^*\Phi^*(\Phi\Phi^*)^{-1}\Phi P_{T_k}\right)\gamma_k = P_{T_k}^*\Phi^*(\Phi\Phi^*)^{-1}(\mathbf{y}), \tag{2.165}$$

which is now the main cost in the calculation of $\mathbf{X}_{T_k}^{(k+1)}$, as the application of the operator $\left(\mathbf{I} + \epsilon_k^{2-p}\mathbf{C}^{-1}\right)$ needs just $O(rD)$ operations–the opertor $\mathbf{C} = \left(\mathscr{D}_{H^{(k)}}^{-1} - \epsilon_k^{2-p}\mathbf{I}_{T_k}\right)$ from above can be inverted easily.

Furthermore, for the part $\mathbf{X}_{T_k^\perp}^{(k+1)}$ of $\mathbf{X}^{(k+1)}$ that is orthogonal to the subspace T_k, if we define the linear operator $P_{T_k^\perp} : T_k^\perp \to \mathbb{R}^{d_1\times d_2}$ analogously to the one of T_k, i.e., such that

$$\mathscr{P}_{T_k^\perp} = P_{T_k^\perp}P_{T_k^\perp}^*.$$

Using this notation, it follows from (2.163) and (2.162) that

$$\begin{aligned}\mathbf{X}^{(k+1)}_{T_k^\perp} &= P^*_{T_k^\perp}(\mathbf{X}^{(k+1)}) = \epsilon_k^{2-p}\mathscr{P}_{T_k^\perp}\Phi^*(\mathbf{z}) \\ &= P^*_{T_k^\perp}\Phi^*\left[\mathbf{I} - (\Phi\Phi^*)^{-1}\Phi P_{T_k}\left(\epsilon_k^{2-p}\mathbf{C}^{-1} + P^*_{T_k}\Phi^*(\Phi\Phi^*)^{-1}\Phi P_{T_k}\right)^{-1} P^*_{T_k}\Phi^*\right](\Phi\Phi^*)^{-1}(\mathbf{y}) \\ &= P^*_{T_k^\perp}\Phi^*(\Phi\Phi^*)^{-1}\left[\mathbf{y} - \Phi P_{T_k}(\gamma_k)\right],\end{aligned} \tag{2.166}$$

also inserting the definition of γ_k from (2.165) in the last equality.

Now, we as we will verify shortly, we can use (2.164) and (2.166) to obtain an exact representation of $\mathbf{X}^{(k+1)}$ that is

- *memory-efficient* in the sense that it only needs $O(rD+m)$ parameters, and
- (potentially) *computationally efficient* in the sense that the computational cost is for the calculation of the update of $\mathbf{X}^{(k+1)}$ is dominated by the cost of the inner iterations of a iterative linear system solver for (2.165), which can be as low as $O(r^2D+mr)$ depending on the measurement operator $\Phi:\mathbb{R}^{d_1\times d_2}\longrightarrow\mathbb{R}^m$.

Indeed, we have that

$$\begin{aligned}\mathbf{X}^{(k+1)} &= P_{T_k}(\mathbf{X}_{T_k}) + P_{T_k^\perp}(\mathbf{X}_{T_k^\perp}) \\ &= P_{T_k}\left(\mathbf{I}+\epsilon_k^{2-p}\mathbf{C}^{-1}\right)\gamma_k + P_{T_k^\perp}P^*_{T_k^\perp}\Phi^*(\Phi\Phi^*)^{-1}\left[\mathbf{y}-\Phi P_{T_k}(\gamma_k)\right] \\ &= P_{T_k}\left(\mathbf{I}+\epsilon_k^{2-p}\mathbf{C}^{-1}\right)\gamma_k + \left(\mathbf{I}-P_{T_k}P^*_{T_k}\right)\Phi^*(\Phi\Phi^*)^{-1}\left[\mathbf{y}-\Phi P_{T_k}(\gamma_k)\right] \\ &= \Phi^*(\Phi\Phi^*)^{-1}\left(\mathbf{y}-\Phi P_{T_k}(\gamma_k)\right) + P_{T_k}\left(\left(\mathbf{I}+\epsilon_k^{2-p}\mathbf{C}^{-1}\right)\gamma_k - P^*_{T_k}\Phi^*(\Phi\Phi^*)^{-1}\left(\mathbf{y}-\Phi P_{T_k}(\gamma_k)\right)\right) \\ &=: \Phi^*(\Phi\Phi^*)^{-1}\mathbf{r}_k + P_{T_k}\left(\left(\mathbf{I}+\epsilon_k^{2-p}\mathbf{C}^{-1}\right)\gamma_k - P^*_{T_k}\Phi^*(\Phi\Phi^*)^{-1}\mathbf{r}_k\right) \\ &=: \Phi^*(\Phi\Phi^*)^{-1}\mathbf{r}_k + P_{T_k}\widetilde{\gamma_\mathbf{k}}\end{aligned}$$

with the definition of the *residual* $\mathbf{r}_k := \mathbf{y}-\Phi P_{T_k}(\gamma_k)\in\mathbb{R}^m$ and the vector $\widetilde{\gamma}\in T_k$ that lives in the tangent space T_k, which is a linear space of dimension $r(d_1+d_2-r)$.

The pair $(\mathbf{r}_k,\widetilde{\gamma_k})$ can be used as a *memory-efficient* representation for $\mathbf{X}^{(k+1)}$.

This fact alone should not be a surprise, as already (2.160) hints at a representation of $\mathbf{X}^{(k+1)}$ that is $O(m)$ by using the vector $\mathbf{z}\in\mathbb{R}^m$.

However, the representation we derived by $(\mathbf{r}_k,\widetilde{\gamma_k})$ has the advantage that for many measurement operators $\Phi:\mathbb{R}^{d_1\times d_2}\longrightarrow\mathbb{R}^m$, all computational steps are sub-quadratic in d, and furthermore, the linear system (2.165) to be solved is *well-conditioned* at least in the limit. The latter aspect will be discussed in Section 2.6.3.

For the first aspect, we recall that it has been already briefly mentioned that *sub-Gaussian* operators Φ as discussed in Section 2.5.1 are computationally problematic, as images $\Phi(\mathbf{X})$ of arbitrary $(d_1\times d_2)$ matrices $\mathbf{X}$ need $O(md_1d_2)$ basic arithmetic operations. The situation is similar for the Gaussian rank-one measurements as in the phase retrieval setting Section 2.5.2 (even if the complexity of these measurements is smaller if $\mathbf{X}$ is provided, for example, in a factorized form).

For entrywise measurements Φ as in the *matrix completion* setting of Section 2.5.3, however, the situation is considerably better, as obtaining the images $\Phi(\mathbf{X})\in\mathbb{R}^m$ does not involve computations (except from entrywise access to the entries of $\mathbf{X}$)), and can be done in $O(m)$. In the remainer of this section, we will therefore focus on the matrix completion setup exclusively. We note that many of the following properties are similar, for example, for Fourier-type rank-one measurements as present in many variations of the phase retrieval problem [CESV13, TLRW14, XSH+18].

First, we note the if $\Phi : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^m$ is such that there exists a set of distinct locations $\Omega = (i_\ell, j_\ell)_{l=1}^m \subset [d_1] \times [d_2]$ with

$$\Phi(\mathbf{X})_\ell = \langle e_{i_\ell}, e_{j_\ell}^*, \mathbf{X} \rangle$$

for all $\ell \in [m]$, it holds that

$$\Phi\Phi^* = \mathbf{I} \in \mathbb{R}^{m \times m},$$

which simplifies some terms above.

Therefore, we can summarize the following simplified outline of an implementation of step (2.90) of `MatrixIRLS` for matrix completion:

1. Calculate $P_{T_k}^* \Phi^*(\mathbf{y}) \in T_k$.

2. Solve
$$\left(\frac{\epsilon_k^{2-p} \mathbf{I}}{\mathscr{D}_{H^{(k)}}^{-1} - \epsilon_k^{2-p} \mathbf{I}} + P_{T_k}^* \Phi^* \Phi P_{T_k} \right) \gamma_k = P_{T_k}^* \Phi^*(\mathbf{y}) \tag{2.167}$$
for $\gamma_k \in T_k$ by a conjugate gradient method [HS52], [QSS10, Chapter 4].

3. Calculate residual $\mathbf{r}_k := \mathbf{y} - \Phi P_{T_k}(\gamma_k) \in \mathbb{R}^m$.

4. Calculate $\widetilde{\gamma_k} = \left(\frac{\mathscr{D}_{H^{(k)}}^{-1}}{\mathscr{D}_{H^{(k)}}^{-1} - \epsilon_k^{2-p} \mathbf{I}} \right) \gamma_k - P_{T_k}^* \Phi^*(\mathbf{r}_k) \in T_k$.

By the calculations above, we obtain an implicit representation of the new iterate $\mathbf{X}^{(k+1)} \in \mathbb{R}^{d_1 \times d_2}$ such that

$$\mathbf{X}^{(k+1)} = \Phi^*(\mathbf{r}_k) + P_{T_k}(\widetilde{\gamma_\mathbf{k}}). \tag{2.168}$$

For elements of T_k such as γ_k, we use the representation of (2.157) such that we keep the $r(d_1 + d_2 + r)$ parameters of the matrices $(\gamma_k)_1 \in \mathbb{R}^{r \times r}$, $(\gamma_k)_2 \in \mathbb{R}^{r \times d_2}$ and $(\gamma_k)_3 \in \mathbb{R}^{d_1 \times r}$ such that

$$P_{T_k}(\gamma_k) = \mathbf{U}^{(k)}(\gamma_k)_1 \mathbf{V}^{(k)*} + \mathbf{U}^{(k)}(\gamma_k)_2(\mathbf{I} - \mathbf{V}^{(k)}\mathbf{V}^{(k)*}) + (\mathbf{I} - \mathbf{U}^{(k)}\mathbf{U}^{(k)*})(\gamma_k)_3 \mathbf{V}^{(k)*}. \tag{2.169}$$

By Lemma 2.6.2, applications of powers of the operator $\mathscr{D}_{H^{(k)}}$ then have a time complexity of $O(rD)$, and the same is true for the very similar operators

$$\frac{\epsilon_k^{2-p} \mathbf{I}}{\mathscr{D}_{H^{(k)}}^{-1} - \epsilon_k^{2-p} \mathbf{I}} \quad \text{and} \quad \frac{\mathscr{D}_{H^{(k)}}^{-1}}{\mathscr{D}_{H^{(k)}}^{-1} - \epsilon_k^{2-p} \mathbf{I}}.$$

What remains to be shown is that the linear operator $\Phi P_{T_k} : T_k \longrightarrow \mathbb{R}^m$ and its adjoint $(\Phi P_{T_k})^* = P_{T_k}^* \Phi^* : \mathbb{R}^m \longrightarrow T_k$ are similarly affordable.

Indeed, this is the case: To calculate, for example,

$$\Phi P_{T_k}(\gamma_k) \in \mathbb{R}^m$$

if $(\gamma_k) \in T_k$ is given as in (2.169), we note that

$$\begin{aligned} P_{T_k}(\gamma_k) &= \mathbf{U}^{(k)}(\gamma_k)_1\mathbf{V}^{(k)*} + \mathbf{U}^{(k)}(\gamma_k)_2(\mathbf{I} - \mathbf{V}^{(k)}\mathbf{V}^{(k)*}) + (\mathbf{I} - \mathbf{U}^{(k)}\mathbf{U}^{(k)*})(\gamma_k)_3\mathbf{V}^{(k)*} \\ &= \mathbf{U}^{(k)}\left[(\gamma_k)_1\mathbf{V}^{(k)*} + (\gamma_k)_2 - ((\gamma_k)_2\mathbf{V}^{(k)})\mathbf{V}^{(k)*}\right] + \left[(\gamma_k)_3 - \mathbf{U}^{(k)}(\mathbf{U}^{(k)*}(\gamma_k)_3)\right]\mathbf{V}^{(k)*} \\ &=: \mathbf{U}^{(k)}\mathbf{M}_1 + \mathbf{M}_2\mathbf{V}^{(k)*}, \end{aligned}$$

i.e., $P_{T_k}(\gamma_k)$ can be written as a sum of two products of $(d_1 \times r)$ and $(r \times d_2)$ matrices, whose entries can then be calculated in

$$O(m \cdot r),$$

as for, e.g., for all $\ell \in [m]$,

$$\Phi(\mathbf{U}^{(k)}\mathbf{M}_1)_\ell = \left(\mathbf{U}^{(k)}\mathbf{M}_1\right)_{i_\ell,j_\ell} = \sum_{k=1}^{r}(\mathbf{U}^{(k)})_{i_\ell,k}(\mathbf{M}_1)_{k,j_\ell}.$$

Furthermore, the calculation of the matrices $\mathbf{M}_1 \in \mathbb{R}^{r\times d_2}$ and $\mathbf{M}_2 \in \mathbb{R}^{d_1\times r}$ can be done in

$$O(r^2 D)$$

arithmetic operators as, for example $\mathbf{M}_1$ can be obtained by multiplying $(r \times r)$- with $(r \times d_2)$-matrices and $(r \times d_2)$- with $(d_2 \times r)$-matrices.

By considering the adjoint, it is possible to calculate also the image of $P_{T_k}^* \Phi^*$ in

$$O(mr + r^2 D)$$

basic operations.

We summarize the discussion of this section in the following theorem.

Theorem 2.8 (Sub-Quadratic Complexity of MatrixIRLS for Matrix Completion)**.** *Let $k, r \in \mathbb{N}$ and $\epsilon_k > 0$, let $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1\times d_2}$ be the k-th iterate of* `MatrixIRLS` *(Algorithm 1), assume we are given its first r singular vectors $\mathbf{U}^{(k)} \in \mathbb{R}^{d_1\times r}$, $\mathbf{V}^{(k)} \in \mathbb{R}^{d_2\times r}$ and singular vectors $\sigma_1(\mathbf{X}^{(k)}), \ldots, \sigma_r(\mathbf{X}^{(k)})$. Assume that $\sigma_\ell(\mathbf{X}^{(k)}) \leq \epsilon_k$ for all $\ell > r$.*

If Φ is as in the matrix completion *problem, i.e., the input operator $\Phi : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$ is as in Section 2.5.3, then the new iterate $\mathbf{X}^{(k+1)} \in \mathbb{R}^{d_1\times d_2}$ can be calculated implicity in a* time complexity *of*

$$O\left((mr + r^2 D) \cdot N_{CG\_inner}\right),$$

where N_{CG_inner} is the number of inner iterations used in the conjugate gradient method to solve (2.167).

In particular, if the sample complexity m is as the lower bound in (2.137), then this can be written as

$$O\left(\left(\mu_0 r^2 D\log(D) + r^2 D\right) \cdot N_{CG\_inner}\right),$$

where μ_0 is the incoherence factor of Definition 3.3.1.

Furthermore, the space complexity *of the representation of $\mathbf{X}^{(k+1)}$ is*

$$O(m + rD)$$

or

$$O(\mu_0 r D\log(D)),$$

respectively.

The result of Theorem 2.8 comes from the fact that the per iteration cost of an iteration of the conjugate gradient (CG) method is dominated by the cost of matrix-vector multiplication with the system matrix

$$\left(\frac{\epsilon_k^{2-p}\mathbf{I}}{\mathcal{D}_{H^{(k)}}^{-1} - \epsilon_k^{2-p}\mathbf{I}} + P_{T_k}^* \Phi^* \Phi P_{T_k}\right),$$

and by the cost of calculating scalar products in the domain [QSS10, Chapter 4]. As the domain of the system is T_k and due to the use of the representation (2.169), one scalar product costs only $O(r^2 D)$, which is dominated by the cost of matrix-vector multiplication $(mr + r^2 D)$.

We note that the number of CG iterations $N_{\text{CG\_inner}}$ plays a critical role in the bound of Theorem 2.8. In exact arithmetic, it is known that the CG method terminates with the exact solution $\mathbf{X}^{(k+1)}$ after at most $N_{\text{CG\_inner}} = \dim(T_k) = O(rD)$ iterations [QSS10, Theorem 4.11]. As this would re-introduce a quadratic dependence on the dimension D, solving the system exactly is computationally prohibitive for the large-scale problems we are interested in.

However, for practical purposes, the conjugate gradient method can be stopped after a fixed number of iterations, or according to a fixed stopping rule. This returns an inexact solution $\mathbf{X}^{(k+1)}$ of (2.90), which might be enough to ensure overall progress of the IRLS method.

A discussion of stopping criteria is beyond the scope of this work, but we refer to [FPRW16] for a such a discussion for an IRLS algorithm for the sparse recovery problem.

We conclude this section with a short discussion of the other main computational step of Algorithm 1, the calculation of the best rank-$\widetilde{r}$ approximation

$$\mathcal{T}_{\widetilde{r}}(\mathbf{X}^{(k+1)}) = \mathbf{U}^{(k+1)} \operatorname{diag}(\sigma_i^{(k+1)})_{i=1}^{\widetilde{r}} \mathbf{V}^{(k)*} \tag{2.170}$$

and the $(\widetilde{r}+1)$-st singular value of $\mathbf{X}^{(k+1)}$.

Many modern methods as Lanczos bidiagonalization [SZ00, LXQ15], randomized block Krylov methods and others [ZL16b, MM15] can be used for this step, see [HMT11] for an overview.

Most of these methods are designed such that matrix-vector multiplications (or matrix-matrix multiplications with tall matrices) are their main computational steps, resulting in very accurate estimates of the first $\widetilde{r}$ singular values and singular vector pairs after $O(\widetilde{r})$ (up to logarithmic factors in the dimension D) matrix-vector multiplications.

In the following lemma, we show that the implicit representation (2.168) of $\mathbf{X}^{(k+1)}$ leads, similarly as in Theorem 2.8, to a *sub-quadratic* cost of matrix-vector multiplication.

Lemma 2.6.3 (Matrix-vector multiplication with iterates in implicit representation)**.** *Let $\mathbf{X}^{(k+1)}$ be the updated iterate of* `MatrixIRLS` *as in (2.168), let $\mathbf{v} \in \mathbb{R}^{d_2}$. Then the time complexity for the calculation of*

$$\mathbf{X}^{(k+1)}\mathbf{v} \in \mathbb{R}^{d_1}$$

is

$$O(m + rD).$$

Proof. Let $\mathbf{r}_k \in \mathbb{R}^m$, let $\widetilde{\gamma} \in T_k$ be such that

$$P_{T_k}(\widetilde{\gamma}) = \mathbf{U}^{(k)}(\widetilde{\gamma}_k)_1 \mathbf{V}^{(k)*} + \mathbf{U}^{(k)}(\widetilde{\gamma}_k)_2(\mathbf{I} - \mathbf{V}^{(k)}\mathbf{V}^{(k)*}) + (\mathbf{I} - \mathbf{U}^{(k)}\mathbf{U}^{(k)*})(\widetilde{\gamma}_k)_3 \mathbf{V}^{(k)*}$$

for matrices $(\widetilde{\gamma}_k)_1 \in \mathbb{R}^{r\times r}$, $(\widetilde{\gamma}_k)_2 \in \mathbb{R}^{r\times d_2}$ and $(\widetilde{\gamma}_k)_3 \in \mathbb{R}^{d_1\times r}$.

Then

$$
\begin{aligned}
\mathbf{X}^{(k+1)}\mathbf{v} &= \left(\Phi^*(\mathbf{r}_k) + P_{T_k}(\widetilde{\gamma_{\mathbf{k}}})\right)\mathbf{v} \\
&= \sum_{l=1}^{m} e_{i_\ell}(\mathbf{r}_k)_\ell e_{j_\ell}^* \mathbf{v} + \left[\mathbf{U}^{(k)}(\widetilde{\gamma_k})_1\mathbf{V}^{(k)*} + \mathbf{U}^{(k)}(\widetilde{\gamma_k})_2(\mathbf{I} - \mathbf{V}^{(k)}\mathbf{V}^{(k)*}) + (\mathbf{I} - \mathbf{U}^{(k)}\mathbf{U}^{(k)*})(\widetilde{\widetilde{\gamma_k}})_3\mathbf{V}^{(k)*}\right]\mathbf{v} \\
&= \sum_{l=1}^{m} e_{i_\ell}(\mathbf{r}_k)_\ell e_{j_\ell}^* \mathbf{v} + \mathbf{U}^{(k)}\underbrace{(\widetilde{\gamma_k})_1\mathbf{V}^{(k)*} + (\widetilde{\gamma_k})_2 - ((\widetilde{\gamma_k})_2\mathbf{V}^{(k)})\mathbf{V}^{(k)*}}_{=:\mathbf{M}_1}\mathbf{v} + \underbrace{(\widetilde{\gamma_k})_3 - \mathbf{U}^{(k)}(\mathbf{U}^{(k)*}(\widetilde{\gamma_k})_3)}_{=:\mathbf{M}_2}\mathbf{V}^{(k)*}\mathbf{v} \\
&= \sum_{l=1}^{m} e_{i_\ell}(\mathbf{r}_k)_\ell \mathbf{v}_{j_\ell} + \mathbf{U}^{(k)}(\mathbf{M}_1\mathbf{v}) + \mathbf{M}_2(\mathbf{V}^{(k)*}\mathbf{v})
\end{aligned}
$$

for any $\mathbf{v} \in \mathbb{R}^{d_2}$.

The result follows because $\sum_{l=1}^{m} e_{i_\ell}(\mathbf{r}_k)_\ell \mathbf{v}_{j_\ell}$ is a sum of m elements, because $\mathbf{M}_1 v$ and $\mathbf{V}^{(k)*}v$ can be done in $O(rd_2)$ operations, and because the multiplication of these with $\mathbf{U}^{(k)}, \mathbf{M}_2 \in \mathbb{R}^{d_1 \times r}$ then costs another $O(rd_1)$. □

The methods for the partial singular value decompositions mentioned above also need matrix-vector multiplications with the transpose $\mathbf{X}^{(k+1)*} \in \mathbb{R}^{d_2 \times d_1}$ of $\mathbf{X}^{(k+1)}$, but, as can be easily verified, the same bound as in Lemma 2.6.3 also holds for these.

2.6.3 Conditioning of System Matrix

In this section, we provide a result about the *spectrum* of the system matrix

$$
\mathbf{A} := \mathbf{D}_k + P_{T_k}^*\Phi^*\Phi P_{T_k} := \left(\frac{\epsilon_k^{2-p}\mathbf{I}}{\mathscr{D}_{H^{(k)}}^{-1} - \epsilon_k^{2-p}\mathbf{I}} + P_{T_k}^*\Phi^*\Phi P_{T_k}\right) \tag{2.171}
$$

of (2.167). It is well-known that the shape of the spectrum of the system matrix $\mathbf{A}$ plays an important role in the convergence of the conjugate gradient iterations. In particular, the CG method terminates after ℓ iterations (in exact arithmetic) if $\mathbf{A}$ has ℓ distinct eigenvalues [NW06, Theorem 5.4] and a bound on the error $\gamma_\ell - \gamma^*$ of the ℓ-th iterate γ_ℓ to the exact solution γ^* of the linear system [NW06, (5.36)] can be provided by

$$
\langle \gamma_\ell - \gamma^*, \mathbf{A}(\gamma_\ell - \gamma^*) \rangle \le 2\left(\frac{\sqrt{\kappa(\mathbf{A})} - 1}{\sqrt{\kappa(\mathbf{A})} + 1}\right)\langle \gamma_0 - \gamma^*, \mathbf{A}(\gamma_0 - \gamma^*) \rangle,
$$

where

$$
\kappa(\mathbf{A}) := \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})} \tag{2.172}
$$

is the condition number of $\mathbf{A}$.

It has been a common problem for IRLS methods that the linear systems to be solved become ill-conditioned close to the desired (low-rank or sparse, depending on the problem) solution [DDFG10, FRW11, FPRW16]. Close to the solution the smoothing parameter ϵ_k is typically very small, resulting in “very large weights” on large parts of the domain induced by the quadratic form

$$
\langle \mathbf{X}, W^{(k)}(\mathbf{X}) \rangle.
$$

For the sparse recovery problem, it has been observed [Vor12] that this blow-up can be a problem for an iterative approach, and in [FPRW16], an analysis was pursued for an IRLS algorithm for the sparse recovery problem about with which precision the linear system for each outer iteration k needs to be solved by a conjugate gradient method to ensure overall convergence.

However, the underlying issue of bad conditioning of the IRLS system matrices was not addressed or solved in [FPRW16] (see Section 5.2 of [FPRW16] for a discussion).

We now argue that by solving the linear systems (2.171), these issues are mitigated, and we show the following statement about its good conditioing in the case that $\mathbf{X}^{(k)}$ is close enough to a rank-r matrix $\mathbf{X}_0$.

Theorem 2.9. *Let $A : \mathbb{R}^{r(d_1+d_2-r)\times r(d_1+d_2-r)}$ be as in (2.171), but acting on a basis representation of T_k. Let $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1\times d_2}$ be the k-th output of* `MatrixIRLS` *for an input with operator $\Phi : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$ as in Section 2.5.3, observation $\mathbf{y} = \Phi(\mathbf{X}_0)$, where $\mathbf{X}_0 \in \mathbb{R}^{d_1\times d_2}$ is a μ_0-incoherent rank-r matrix, and rank estimate $\widetilde{r} = r$. Let $\epsilon_k = \sigma_{r+1}(\mathbf{X}^{(k)}) < \sigma_r(\mathbf{X}^{(k)})$ and C as in Corollary 2.5.3, let $C_1 > 0$ be a constant that is small enough. If*

$$\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \leq \min\left(C_1^{\frac{2}{2-p}} \left(\frac{\mu_0 r}{d}\right)^{\frac{1}{2-p}}, \frac{1}{4}\right) \sigma_r(\mathbf{X}_0),$$

and

$$m \geq C\mu_0 r(d_1+d_2)\log(d_1+d_2),$$

it holds that with high probability, the spectrum $\lambda(\mathbf{A})$ fulfills

$$\lambda(\mathbf{A}) \subset \frac{m}{d_1 d_2}\left[\frac{6}{10}; \frac{24}{10}\right].$$

In particular, it then holds that the condition number of $\mathbf{A}$ fulfills

$$\kappa(\mathbf{A}) \leq 4.$$

Remark 2.6.1. *As a basis representation of T_k, we can use for example γ_k represented by $\{(\widetilde{\gamma_k})_1, (\widetilde{\gamma_k})_2, (\widetilde{\gamma_k})_3\}$ such that $(\widetilde{\gamma_k})_1 \in \mathbb{R}^{r\times r}$, $(\widetilde{\gamma_k})_2 \in \mathbb{R}^{r\times(d_2-r)}$ and $(\widetilde{\gamma_k})_2 \in \mathbb{R}^{(d_1-r)\times r}$ such that*

$$\mathbf{U}^{(k)}(\widetilde{\gamma_k})_1\mathbf{V}^{(k)*} + \mathbf{U}^{(k)}(\widetilde{\gamma_k})_2\mathbf{V}_\perp^{(k)*} + \mathbf{U}_\perp^{(k)}(\widetilde{\gamma_k})_3\mathbf{V}^{(k)*}. \tag{2.173}$$

Proof. Recall the definitions $\mathbf{D}_k = \frac{\epsilon_k^{2-p}\mathbf{I}}{\mathcal{D}_{H^{(k)}}^{-1} - \epsilon_k^{2-p}\mathbf{I}}$ and $\mathbf{A} = \mathbf{D}_k + P_{T_k}^*\Phi^*\Phi P_{T_k}$. We know that $\mathbf{D}_k \succ 0$ since $\mathcal{D}_{H^{(k)}}^{-1} \succ \epsilon_k^{2-p}\mathbf{I}$, as all of its eigenvalues λ_i are such that $\lambda_i \geq \sigma_r^{2-p}(\mathbf{X}^{(k)}) \geq \epsilon_k^{2-p}$, see (2.159). Furthermore, the eigenvalues of $\mathbf{D}_k$ are given by

$$\begin{aligned}\lambda(\mathbf{D}_k) = &\left\{(\mathbf{H}_1^{(k)})_{ij}^{-1} : i \in [d_1], j \in [d_2], i \leq j, i \leq r \text{ or } j \leq r\right\} \\ &\cup \left\{(\mathbf{H}_2^{(k)})_{ij}^{-1} : i \in [d_1], j \in [d_1], i < j, i \leq r \text{ or } j \leq r\right\},\end{aligned}$$

which means that

$$\|\mathbf{D}_k\|_{S_\infty} \leq \frac{\epsilon_k^{2-p}}{\sigma_r(\mathbf{X})^{2-p} - \epsilon_k^{2-p}} \leq \frac{\epsilon_k^{2-p}}{(3/4\sigma_r(\mathbf{X}_0))^{2-p} - \epsilon_k^{2-p}}, \tag{2.174}$$

using that $\sigma_r(\mathbf{X}^{(k)}) \geq (1-1/4)\sigma_r(\mathbf{X}_0)$ since $\|\mathbf{X}^{(k)}-\mathbf{X}_0\|_{S_\infty} \leq \frac{1}{4}\sigma_r(\mathbf{X}_0)$, see also (2.114). Also, since $\epsilon_k = \sigma_{r+1}(\mathbf{X}^{(k)}) \leq \|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \leq \frac{1}{4}\sigma_r(\mathbf{X}_0)$ and further $\epsilon_k = \sigma_{r+1}(\mathbf{X}^{(k)}) \leq C_1^{\frac{2}{2-p}} \left(\frac{\mu_0 r}{d}\right)^{\frac{1}{2-p}} \sigma_r(\mathbf{X}_0)$, we have that

$$\frac{\epsilon_k^{2-p}}{(3/4\sigma_r(\mathbf{X}_0))^{2-p} - \epsilon_k^{2-p}} \le \frac{C_1^2 \frac{\mu_0 r}{d}}{(3/4 - 1/4)^{2-p}} \frac{\sigma_r(\mathbf{X}_0)^{2-p}}{\sigma_r(\mathbf{X}_0)^{2-p}} \le 4C_1^2 \frac{\mu_0 r}{d}.$$

This implies that

$$0 \le \lambda_{\min}(\mathbf{D}_k) \le \lambda_{\max}(\mathbf{D}_k) = \|\mathbf{D}_k\|_{S_\infty} \le 4C_1^2 \frac{\mu_0 r}{d} \le 4C_1^2 \frac{m}{CdD\log(D)} \le \frac{m}{d_1 d_2},$$

if the constant $C_1 > 0$ is small enough, using the lower bound on the sample complexity m.

The second summand in $\mathbf{A}$, the matrix $P_{T_k}^* \Phi^* \Phi P_{T_k}$, is positive semidefinite already due to its factorized form. We note that by following the proof of Corollary 2.5.3, we see that under the assumptions of Theorem 2.9, we have that for $\mathscr{P}_{T_k} : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}, \mathbf{Z} \mapsto P_{T_k} P_{T_k}^*(\mathbf{Z})$ and $\mathscr{P}_\Omega : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}, \mathbf{Z} \mapsto \Phi^*\Phi(\mathbf{Z})$,

$$\begin{aligned} \frac{d_1 d_2}{m} \left\| P_{T_k} \left[P_{T_k}^* \Phi^* \Phi P_{T_k} - \frac{m}{d_1 d_2} \mathbf{I} \right] P_{T_k}^* \right\|_{S_\infty} &= \left\| \frac{d_1 d_2}{m} P_{T_k} P_{T_k}^* \Phi^* \Phi P_{T_k} P_{T_k}^* - P_{T_k} P_{T_k}^* \right\|_{S_\infty} \\ = \left\| \frac{d_1 d_2}{m} \mathscr{P}_T \mathscr{P}_\Omega \mathscr{P}_T - \mathscr{P}_T \right\|_{S_\infty} \le \frac{4}{10} \end{aligned}$$

on an event E that holds with high probability.

As P_{T_k} is a matrix with orthonormal columns such that $P_{T_k}^* P_{T_k} = \mathbf{I}$, this implies that

$$\left\| P_{T_k}^* \Phi^* \Phi P_{T_k} - \frac{m}{d_1 d_2} \mathbf{I} \right\|_{S_\infty} \le \frac{4m}{10 d_1 d_2}$$

on the event E. Thus, the bound on the spectrum of $\mathbf{A}$ follows from this and (2.174) since

$$\begin{aligned} \left\| \mathbf{A} - \frac{3}{2} \frac{m}{d_1 d_2} \mathbf{I} \right\|_{S_\infty} &= \left\| \mathbf{D}_k + P_{T_k}^* \Phi^* \Phi P_{T_k} - \frac{3}{2} \frac{m}{d_1 d_2} \mathbf{I} \right\|_{S_\infty} \le \left\| \mathbf{D}_k - \frac{1}{2} \frac{m}{d_1 d_2} \mathbf{I} \right\|_{S_\infty} + \left\| P_{T_k}^* \Phi^* \Phi P_{T_k} - \frac{m}{d_1 d_2} \mathbf{I} \right\|_{S_\infty} \\ &\le \frac{1}{2} \frac{m}{d_1 d_2} + \frac{4}{10} \frac{m}{d_1 d_2} = \frac{9}{10} \frac{m}{d_1 d_2}. \end{aligned}$$

The condition number bound follows immediately since $\kappa(\mathbf{A}) = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})} = 4$. □

There is a caveat for the applicability of Theorem 2.9 for the analysis of a practical implementation of `MatrixIRLS`: It might be undesirable to use basis representations as (2.173) of elements of the linear space T_k for γ_k in (2.167), since these representations use the matrices $\mathbf{U}_\perp^{(k)} \in \mathbb{R}^{d_1 \times (d_1 - r)}$ and $\mathbf{V}_\perp^{(k)} \in \mathbb{R}^{d_2 \times (d_2 - r)}$ of last singular vectors of $\mathbf{X}^{(k)}$ explicitly, which we might not want to use as the resulting projections ΦP_{T_k} will not be as cheap as in Section 2.6.2.

However, Theorem 2.9 can be used to show at least that $\mathbf{A}$ acting on elements of T_k given by the redundant representation of (2.157) (which uses $r(d_1 + d_2 + r)$ instead of $\dim(T_k) = r(d_1 + d_2 - r)$ parameters) has $r(d_1 + d_2)$ concentrated eigenvalues and only $2r^2$ small ones.

As a summary, since Theorem 2.9 gives a bound on the condition number of the linear system matrix $\mathbf{A}$ that is a *small constant*, the theory of the conjugate gradient methods suggests that very good solutions can be found already after few, in particular, after

$$N_{\text{CG\_inner}} = \text{cst}.$$

CG iterations (where cst. is small), for each IRLS iteration, at least in the neighborhood of a low-rank matrix $\mathbf{X}_0$ that is compatible with the measurements.

Taking into account the statement of Theorem 2.8, this suggests that at least locally, a new iterate $\mathbf{X}^{(k+1)}$ can be calculated with a time complexity of

$$O\left((mr + r^2D) \cdot N_{\text{CG\_inner}}\right) = O\left(mr + r^2D\right).$$

Together with the superlinear convergence rate of the (outer) IRLS iterations, this suggests that `MatrixIRLS` is indeed an algorithm that can achieve high-accuracy low-rank solutions of underdetermined systems sub-linearly in the matrix dimension $(d_1 \times d_2)$, or sub-quadratically in $D = \max(d_1, d_2)$.

2.7 Numerical Experiments

In this section, we explore the numerical performance of IRLS algorithms for low-rank matrix recovery. We focus on a verification of *convergence rates* and on the comparison of various low-rank matrix recovery algorithms of different kind in terms of *data efficiency*.

As a representative of IRLS algorithms for low-rank matrix recovery using tight quadratic bounds as studied in this dissertation, we use the algorithm `HM-IRLS`, presenting the experiments of the paper [KS18] in Sections 2.7.1 to 2.7.3. `HM-IRLS` has similar properties as `MatrixIRLS`, especially in terms of convergence rates (see Theorem 2.7 and the discussion thereafter) and data efficiency, which is why the experiments can be regarded as representative for `MatrixIRLS` up to a certain degree.

In particular, we demonstrate first in Section 2.7.2 that the superlinear convergence rate that was proven theoretically for Algorithm 1/`MatrixIRLS` in Theorem 2.7 (and for `HM-IRLS` in [KS18]) can indeed be accurately verified in numerical experiments, which is in contrast to older variants of IRLS.

In Section 2.7.3, we then examine the recovery performance of `HM-IRLS` for the matrix completion setting with the performance of other state-of-the-art algorithms comparing the measurement complexities that are needed for successful recovery for many random instances.

The numerical experiments are conducted on Linux and Mac systems with MATLAB R2017b. An implementation of `HM-IRLS` for matrix completion including code reproducing many conducted experiments is available at `https://github.com/ckuemmerle/`.

2.7.1 Experimental Setup

In the experiments, we sample $(d_1 \times d_2)$ dimensional ground truth matrices $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ of rank r such that $\mathbf{X}_0 = \mathbf{U}_0 \mathbf{\Sigma}_0 \mathbf{V}^*$, where $\mathbf{U}_0 \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V}_0 \in \mathbb{R}^{d_2 \times r}$ are independent matrices with i.i.d. standard Gaussian entries and $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ is a diagonal matrix with i.i.d. standard Gaussian diagonal entries, independent from $\mathbf{U}_0$ and $\mathbf{V}_0$.

We recall that a rank-r matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ has $\deg_f = r(d_1 + d_2 - r)$ degrees of freedom, which is the theoretical lower bound on the number of measurements that are necessary for exact reconstruction [CP11]. The random measurement setting we use in the experiments can be described as follows: We take measurements of matrix completion type, sampling $m = \lfloor \rho \deg_f \rfloor$ locations $\Omega = (i_\ell, j_\ell)_{\ell=1}^m$ and corresponding entries of $\mathbf{X}_0$ uniformly over the $d_1 d_2$ possibles to obtain $\mathbf{y} = \Phi(\mathbf{X}_0)$, cf. Section 2.5.3. Here, ρ is such that $\frac{d_1 d_2}{\deg_f} \geq \rho \geq 1$ and parametrizes the difficulty of the reconstruction problem, from very hard problems for $\rho \approx 1$ to easier problems for larger ρ.

However, this uniform sampling of Φ could yield instances of measurement operators whose information content is not large enough to ensure well-posedness of the corresponding low-rank matrix recovery problem, even if $\rho > 1$. More precisely, it is impossible to recover a matrix exactly if the number of revealed entries in any row or column is smaller than its rank r, which is explained and shown in the context of the proof of [PABN16, Theorem 1].

Thus, in order to provide for a sensible measurement model for small ρ, we exclude operators Φ that sample fewer than r entries in any row or column. Therefore, we adapt slightly the sampling model covered by Theorem 2.7 such that operators Φ are discarded and sampled again until the requirement of at least r entries per column and row is met and recovery can be achieved from a theoretical point of view.

We note that the described phenomenon is very related to the fact that matrix completion recovery guarantees for the uniform sampling model require at least one additional log factor, i.e., they require at least $m \geq \log(\max(d_1, d_2)) \deg_f = \log(D) \deg_f$ sampled entries [DR16, Section V].

While we detail the experiments for the matrix completion measurement setting just described in the remaining section, we add that Gaussian measurement models also lead to very similar results in experiments.

2.7.2 Convergence Rate Comparison with Other IRLS Algorithms

In this subsection, we vary the Schatten-p parameter between 0 and 1 and compare the corresponding convergence behavior of `HM-IRLS` with the IRLS variant `IRLS-col`, which performs the reweighting just in the column space, and with a variant called `AM-IRLS`, corresponding to weight operators with are the *arithmetic mean* of the ones of `IRLS-col` and `IRLS-row` as they have been defined and discussed in Section 2.3.2.

In particular, all these IRLS variants optimize the *smoothed Schatten-p* objective

$$F_\epsilon(\mathbf{X}) = \sum_{i=1}^{d} \left(\sigma_i^2(\mathbf{X}) + \epsilon^2\right)^{\frac{p}{2}},$$

for $0 < p \leq 1$ by iteratively solving the weighted least squares problems (2.90), i.e.,

$$\mathbf{X}^{(k+1)} = \arg\min_{\Phi(\mathbf{X})=\mathbf{y}} \langle \mathbf{X}, W^{(k1)}(\mathbf{X}) \rangle,$$

for weight operators $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \longrightarrow \mathbb{R}^{d_1 \times d_2}$ defined such that

$$W^{(k)}(\mathbf{X}) = \mathbf{U}_k \left[\mathbf{H}^{(k)} \circ \left(\mathbf{U}_k^* \mathbf{X} \mathbf{V}_k \right) \right] \mathbf{V}_k^*,$$

where $\mathbf{X}^{(k)} = \mathbf{U}_k \operatorname{dg}(\sigma_i^{(k)}) \mathbf{V}_k^*$, $\mathbf{U}_k \in \mathbb{O}^{d_1}$, $\mathbf{V}_k \in \mathbb{O}^{d_2}$ is a singular value decomposition of the previous iterate $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$, and $\mathbf{H}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ such that

$$\left(\mathbf{H}^{(k)}\right)_{ij} = \begin{cases} 2\left[\left((\sigma_i^{(k)})^2 + \epsilon_k^2\right)^{\frac{2-p}{2}} + \left((\sigma_j^{(k)})^2 + \epsilon_k^2\right)\right)^{\frac{2-p}{2}}\right]^{-1} & \text{for } \texttt{HM-IRLS}, \\ \left((\sigma_i^{(k)})^2 + \epsilon_k^2\right)^{\frac{p-2}{2}} & \text{for } \texttt{IRLS-col}, \\ \left((\sigma_j^{(k)})^2 + \epsilon_k^2\right)^{\frac{p-2}{2}} & \text{for } \texttt{IRLS-row}, \text{ and} \\ 2.5.3\frac{1}{2} \cdot \left[\left((\sigma_i^{(k)})^2 + \epsilon_k^2\right)^{\frac{p-2}{2}} + \left((\sigma_i^{(k)})^2 + \epsilon_k^2\right)^{\frac{p-2}{2}}\right] & \text{for } \texttt{AM-IRLS}. \end{cases} \tag{2.175}$$

It can be checked that the corresponding weight operator $W^{(k)}$ for `IRLS-col` and `IRLS-row` is equal to the definitions of (2.34) and (2.31), respectively.

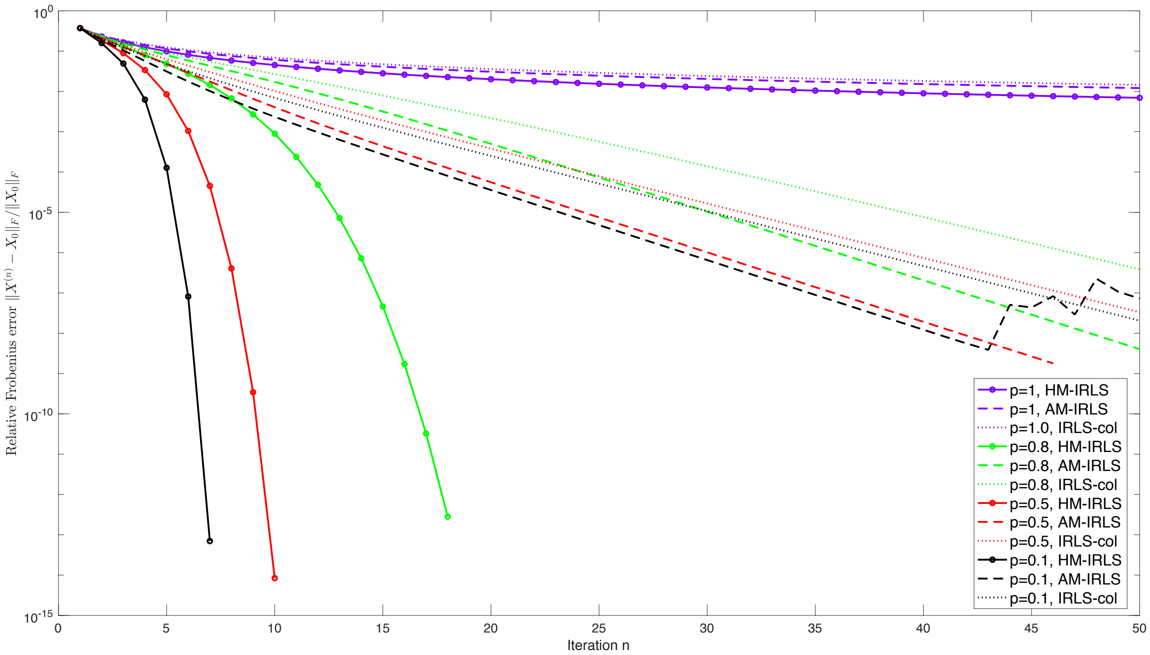

Figure 2.1 – Relative Frobenius errors as a function of the iteration n for oversampling factor $\rho = 2$ ("easy" problem).

The smoothing parameter ϵ_k is chosen according to the rule

$$\epsilon_{k+1} = \min\left(\epsilon_k, \sigma_{r+1}(\mathbf{X}^{(k+1)})\right),$$

which is the same rule (2.91) as in Algorithm 1 resp. `MatrixIRLS`. In Algorithm 1 this would correspond to choosing $\widetilde{r} = r$, i.e., the correct rank estimate.

We recall that `IRLS-col` is very similar to the IRLS algorithms of [FRW11] and [MF12] and differs from them basically just in the choice of the ϵ-smoothing. We present the experiments with `IRLS-col` to isolate the influence of the weight matrix type, but very similar results can be observed for the algorithms of [FRW11] and [MF12].[6]

In the matrix completion setup of Section 2.7.1, we choose $d_1 = d_2 = 40$, $r = 10$ and distinguish easy, hard and very hard problems corresponding to oversampling factors ρ of 2.0, 1.2 and 1.0, respectively. The algorithms are provided with the ground truth rank r and are stopped whenever the relative change of Frobenius norm $\|X^{(k)} - X^{(n-1)}\|_F/\|X^{(n-1)}\|_F$ drops below the threshold of 10^{-10} or a maximal iteration of iterations $n_{\max}$ is reached.

Convergence Rates

First, we study the behavior of the three IRLS algorithms for the "easy" setting of an oversampling factor of $\rho = 2$, which means that $\frac{2r(d_1+d_2-r)}{d_1 d_2} = 0.875$ of the entries are sampled, and parameters $p \in \{0.1, 0.5, 0.8, 1\}$.

In Figure 2.1, we observe that for $p = 1$, `HM-IRLS`, `AM-IRLS` and `IRLS-col` have a quite similar behavior, as the relative Frobenius errors $\|X^{(n)} - X_0\|_F/\|X_0\|_F$ decrease only slowly, i.e., even a linear rate is hardly identifiable. For choices $p < 1$ that correspond to non-convex objectives, we observe a very fast, superlinear convergence of `HM-IRLS`, as the iterates $X^{(n)}$ converge up to a relative error of less than 10^{-12} within fewer than 20 iterations for $p \in \{0.8, 0.5, 0.1\}$. Precise calculations verify that the rate of convergences are indeed of

[6] Implementations of the mentioned authors' algorithms were downloaded from `https://faculty.washington.edu/mfazel/` and `https://github.com/rward314/IRLSM`, respectively.

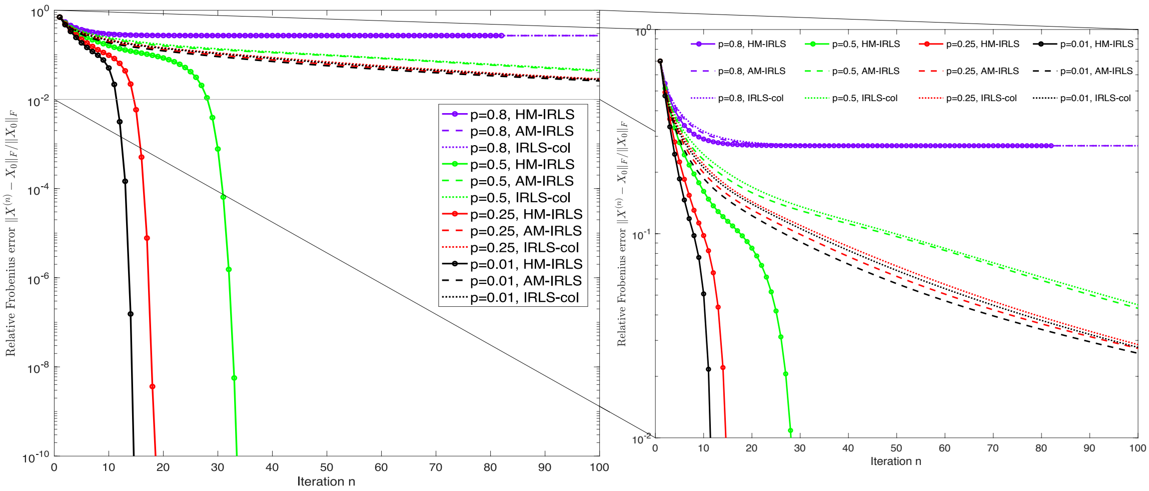

Figure 2.2 – Relative Frobenius errors as a function of the iteration n for oversampling factor $\rho = 1.2$ (hard problem). Left column: y-range $[10^{-10}; 10^0]$. Right column: Enlarged section of left column corresponding to y-range of $[10^{-2}; 10^0]$.

order $2 - p$, the order predicted by Theorem 2.7. We note that for these examples, the fast convergence rate not only kicks in locally, but starting from the very first iteration.

On the other hand, it is easy to see that AM-IRLS and IRLS-col converge *linearly, but not superlinearly* to the ground truth X_0 for $p \in \{0.8, 0.5, 0.1\}$. The linear rate of AM-IRLS is slightly better than the one of IRLS-col, but the numerical stability of AM-IRLS deteriorates for $p = 0.1$ close to the ground truth (after iteration 43). This is due to a bad conditioning of the quadratic problems as the $\mathbf{X}^{(n)}$ are close to rank-r matrices. In contrast, no numerical instability issues can be observed for HM-IRLS.

For the hard matrix completion problems with oversampling factor of $\rho = 1.2$, we observe that for $p = 0.8$, the three algorithms typically do not converge to ground truth. This can be seen in the example that is shown in Figure 2.2, where HM-IRLS, AM-IRLS and IRLS-col all exhibit a relative error of 0.27 after 100 iterations. We do not visualize the result for $p = 1$, as the iterates of the three algorithms do not converge to the ground truth either, which is to be expected: In some sense, they implement nuclear norm minimization, which is typically not able to recover a low-rank matrix from measurements with an oversampling factor as small as $\rho = 1.2$ [DGM13]. The dramatic difference in behavior between HM-IRLS and the other approaches becomes very apparent for more non-convex choices of $p \in \{0.01, 0.25, 0.5\}$, where the former converges up to a relative Frobenius error of less than 10^{-10} within 15 to 35 iterations, while the others do not reach a relative error of 10^{-2} even after 100 iterations. For HM-IRLS, the convergence of order $2 - p$ can be very well locally observed also here, it just takes some iterations until the superlinear convergence begins, which is due to the increased difficulty of the recovery problem.

Finally, we see in the example shown in Figure 2.3 that even for the very hard problems where $\rho = 1$, which means that the number of sampled entries corresponds exactly to the degrees of freedom $r(d_1 + d_2 - r)$, HM-IRLS can be successful to recover the rank-r matrix if the parameter p is chosen small enough (here: $p \leq 0.25$). This is not the case for the algorithms AM-IRLS and IRLS-col.

We summarize that among the three variants HM-IRLS, AM-IRLS and IRLS-col, only HM-IRLS is able to solve the low-rank matrix recovery problem for very low sample complexities corresponding to $\rho \approx 1$. Furthermore, among the four considered algorithms, it is the only algorithm that exhibits a superlinear rate of convergence at all.

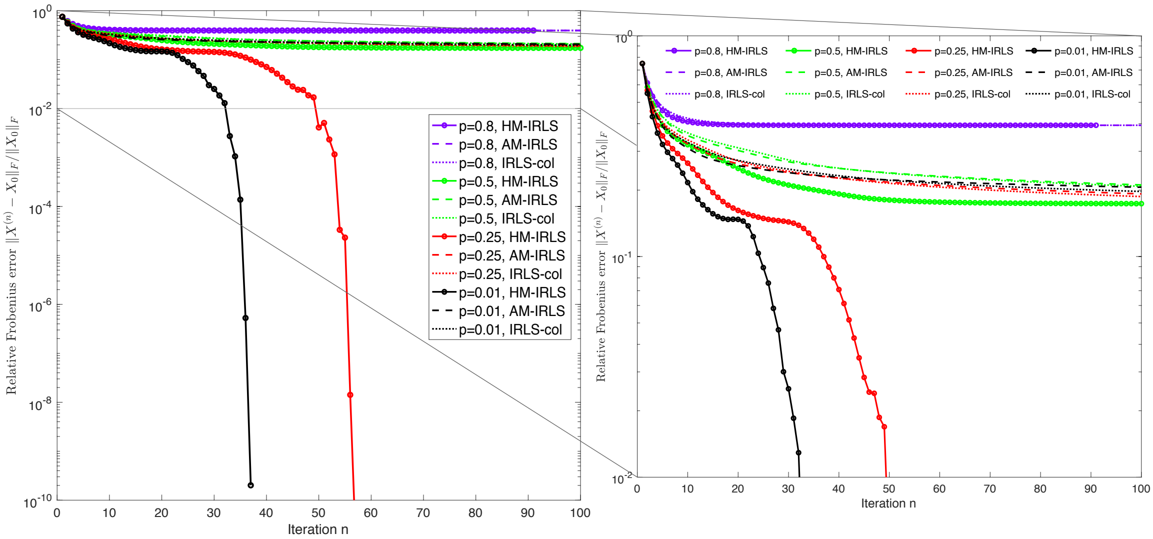

Figure 2.3 – Relative Frobenius errors as a function of the iteration n for oversampling factor $\rho = 1.0$ (very hard problem). Left column: y-range $[10^{-10}; 10^{0}]$. Right column: Enlarged section of left column corresponding to y-range of $[10^{-2}; 10^{0}]$.

2.7.3 Matrix Completion: Recovery Performance Compared to State-Of-The-Art Algorithms

After comparing the performance of HM-IRLS with other IRLS variants, we now conduct experiments to compare wit the empirical performance of of low-rank matrix recovery algorithms different from IRLS.

To obtain a comprehensive picture, we consider not only the IRLS variants AM-IRLS and IRLS-col, but a selection of other methods, which are chosen to be a representative set of state-of-the-art methods based on different strategies in our next experiments. In particular, we conduct experiments to compare to the Riemannian optimization algorithm Riemann_Opt [Van13], the alternating minimization approaches AltMin [HH09], ASD [TW16] and BFGD [PKCS18], and finally the algorithms Matrix ALPS II [KC14] and CGIHT_Matrix [BTW15], which are based on iterative hard thresholding. As the IRLS variants, all these algorithms use knowledge about the actual ground truth rank r.

In the experiments, we examine the empirical recovery probabilities of the different algorithms systematically for varying oversampling factors ρ, determining the difficulty of the low-rank recovery problem as the sample complexity fulfills $m = \lfloor \rho \deg_f \rfloor$. We recall that a large parameter ρ corresponds to an *easy* reconstruction problem, while a small ρ, e.g., $\rho \approx 1$, defines a very *hard* problem.

We choose $d_1 = d_2 = 100$ and the $r = 8$ as parameter of the experimental setting, conducting the experiments to recover rank-8 matrices $\mathbf{X}_0 \in \mathbb{R}^{100\times 100}$. We remain in the matrix completion measurement setting described in Section 2.7.1, but now sample 150 random instances of $\mathbf{X}_0$ and Φ for different numbers of measurements varying between $m_{\min} = 1500$ to $m_{\max} = 4000$. This means that the oversampling factor ρ increases from $\rho_{\min} = 0.975$ to $\rho_{\max} = 2.60$. For each algorithm, a successful recovery of $\mathbf{X}_0$ is defined as a relative Frobenius error $\|\mathbf{X}^{\text{out}} - \mathbf{X}_0\|_F / \|\mathbf{X}_0\|_F$ of the matrix $\mathbf{X}^{\text{out}}$ returned by the algorithm of smaller than 10^{-3}. The algorithms are run until stagnation of the iterates or until the maximal number of iterations $n_{\max} = 3000$ is reached. The number $n_{\max}$ is chosen large enough to ensure that a recovery failure is not due to a lack of iterations.

In the experiments, except for AltMin, for which we used our own implementation, we used implementations provided by the authors of the corresponding papers for the respective

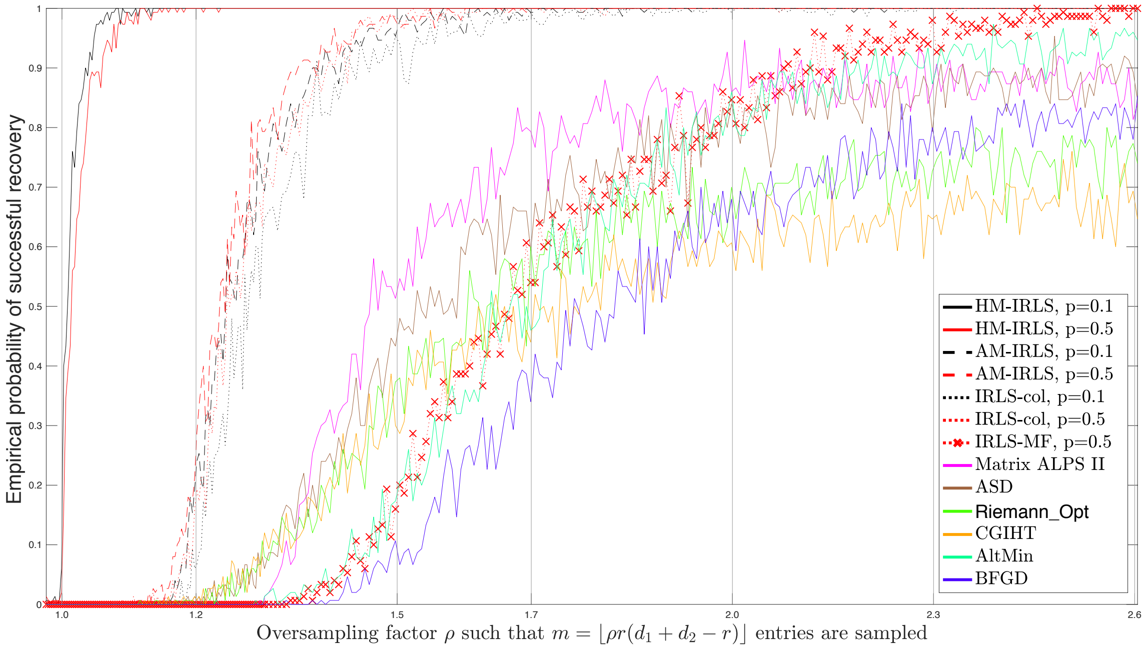

Figure 2.4 – Comparison of empirical success rates of state-of-the-art algorithms, as a function of the oversampling factor ρ

algorithms, using default input parameters provided by the authors. The respective code sources can be found in the references.

Beyond the state-of-the-art performance of proposed IRLS method

The results of the experiment can be seen in Figure 2.4. We observe that HM-IRLS exhibits a very high empirical recovery probability for $p = 0.1$ and $p = 0.5$ as soon as the sample complexity parameter ρ is slightly larger than 1.0, which means that $m = \lfloor \rho r(d_1 + d_2 - r) \rfloor$ measurements suffice to recover $(d_1 \times d_2)$-dimensional rank-r matrices with ρ close to 1. This is very close to the information theoretical lower bound of $\deg_f = r(d_1 + d_2 - r)$. Very interestingly, it can be observed that the empirical recovery probability reaches almost 1 already for an oversampling factor of $\rho \approx 1.1$, and remains at exactly 1 starting from $\rho \approx 1.2$.

Relatively good success rates can also be observed for the algorithms AM-IRLS and IRLS-col for non-convex parameter choices $p \in \{0.1, 0.5\}$, reaching an empirical success probability of almost 100% at around $\rho = 1.5$. AM-IRLS performs only marginally better than the classical IRLS strategy IRLS-col, which are both outperformed considerably by HM-IRLS. It is important to note that in accordance to what was observed in Section 2.7.2, in the successful instances, the error threshold that defines successful recovery is achieved already after a few dozen iterations for HM-IRLS, while typically only after several or many hundreds for AM-IRLS and IRLS-col. Furthermore, it is interesting to observe that the algorithm IRLS-MF, which corresponds to the variant studied and implemented by [MF12] and differs from IRLS-col mainly only in the choice of the ϵ-smoothing (2.91), has a considerably worse performance than the other IRLS methods. This is plausible since the smoothing severely influences the optimization landscape of the objective to be minimized.

The strong performance of HM-IRLS is in stark contrast to the behavior of all the algorithms that are based on different approaches than IRLS and that we considered in our experiments. They basically never recover any rank-r matrix if $\rho < 1.2$, and most of the algorithms need a sample complexity parameter of $\rho > 1.7$ to exceed a empirical recovery probability of a mere 0.5. A success rate of close to 0.8 is reached not before raising ρ above 2.0 in our experimental setting, and also only for a subset of the comparison algorithms, in particular for Matrix

`ALPS II`, `ASD`, `AltMin`. The empirical probability of 1 is only reached for some of the IRLS methods, and not for any competing method in our experimental setting, even for rather large oversampling factors such as $\rho = 2.5$. While we do not rule out that a possible parameter tuning could improve the performance of any of the algorithms slightly, we conclude that for hard matrix completion problems, the experimental evidence for the vast differences in the recovery performance of `HM-IRLS` compared to other methods is very apparent.

Thus, our observation is that the tight quadratic bound IRLS algorithms as `HM-IRLS` algorithm *recover low-rank matrices systematically with nearly the optimal number of measurements, and need fewer measurements than all the state-of-the-art algorithms we included in our experiments*, if the non-convexity parameter p is chosen such that $p \ll 1$.

We note that the very sharp phase transition between failure and success that can be observed in Figure 2.4 for `HM-IRLS` indicates that the sample complexity parameter ρ is indeed the major variable determining the success of `HM-IRLS`. In contrast, the wider phase transitions for the other algorithms suggest that they might depend more on other factors, as the realizations of the random sampling model and the interplay of measurement operator Φ and ground truth matrix X_0.

Another conclusion that can be drawn from the empirical recovery probability of 1 is that, despite the severe non-convexity of the underlying Schatten-p quasi-norm for, e.g., $p = 0.1$, `HM-IRLS` with the initialization of $X^{(1)}$ as the Frobenius norm minimizer does not get stuck in stationary points if the oversampling factor is large enough. The preceding discussion of this chapter gives also a meaningful interpretation of choosing the non-convexity parameter $p = 0$, resulting in a log-det type objective (2.27), which has not been studied by the theory of [KS18]. Experiments suggest that a choice of $p = 0$ admits usually one of the best performances, if the not the very best among all $p \in [0; 1]$.

Further experiments conducted with random initializations as well as severely adversary initializations, e.g., with starting points chosen in the orthogonal complement of the spaces spanned by the singular vectors of the ground truth matrix X_0, lead to comparable results. We leave the systematic study of this aspect to future work.

As a summary, we claim that, for the low-rank matrix completion problem, `HM-IRLS` exhibits a global convergence behavior for oversampling factors in a range for which competing non-convex low-rank matrix recovery algorithms fail to succeed.

This good global convergence behavior goes beyond our local analysis of Section 2.4.2. A detailed theoretical investigation of such behavior remains for future work.

2.8 Further Proofs

In this section, we provide proofs of some properties that have been shown already for previous IRLS algorithms [DDFG10, FRW11, MF12].

2.8.1 Proof of Theorem 2.6

Proof of Theorem 2.6. 1. To show the first statement, let $k \in \mathbb{N}$ and let $\epsilon_{k-1} > 0$, $\mathbf{X}^{(k-1)} \in \mathbb{R}^{d_1 \times d_2}$ be the output of `MatrixIRLS` at iteration $k-1$. As the quadratic $Q_{\epsilon_{k-1}}(\cdot|\mathbf{X}^{(k-1)}$ (see (2.61)) defined by the weight operator $W^{(k-1)}$ of `MatrixIRLS` at iteration $k-1$ is defined as

$$Q_{\epsilon_{k-1}}(\mathbf{X}|\mathbf{X}^{(k-1)}) = F_{\epsilon_{k-1}}(\mathbf{X}^{(k-1)}) + \langle \nabla F_{\epsilon_{k-1}}(\mathbf{X}^{(k-1)}), \mathbf{X}-\mathbf{X}^{(k-1)} \rangle + \frac{1}{2} \langle \mathbf{X}-\mathbf{X}^{(k-1)}, W^{(k-1)}(\mathbf{X}-\mathbf{X}^{(k-1)}) \rangle$$

and since

$$\nabla F_{\epsilon_{k-1}}(\mathbf{X}^{(k-1)}) = W^{(k-1)}(\mathbf{X}^{(k-1)}),$$

due to Theorem 2.4, it holds that

$$Q_{\epsilon_{k-1}}(\mathbf{X}|\mathbf{X}^{(k-1)}) = F_{\epsilon_{k-1}}(\mathbf{X}^{(k-1)}) + \frac{1}{2}\left(\langle \mathbf{X}, W^{(k-1)}(\mathbf{X})\rangle - \langle \mathbf{X}^{(k-1)}, W^{(k-1)}(\mathbf{X}^{(k-1)})\rangle\right). \quad (2.176)$$

In particular, as $2Q_{\epsilon_{k-1}}(\mathbf{X}|\mathbf{X}^{(k-1)})$ and $\langle \mathbf{X}, W^{(k-1)}(\mathbf{X})\rangle$ differ just by terms not depending on $\mathbf{X}$, it follows that

$$\mathbf{X}^{(k)} = \underset{\Phi(\mathbf{X})=\mathbf{Y}}{\arg\min} \langle \mathbf{X}, W^{(k-1)}(\mathbf{X})\rangle = \underset{\Phi(\mathbf{X})=\mathbf{Y}}{\arg\min}\, Q_{\epsilon_{k-1}}(\mathbf{X}|\mathbf{X}^{(k-1)}). \quad (2.177)$$

Thus, we can use this to see that

$$F_{\epsilon_{k-1}}(\mathbf{X}^{(k)}) \leq Q_{\epsilon_{k-1}}(\mathbf{X}^{(k)}|\mathbf{X}^{(k-1)}) \leq Q_{\epsilon_{k-1}}(\mathbf{X}^{(k-1)}|\mathbf{X}^{(k-1)}) = F_{\epsilon_{k-1}}(\mathbf{X}^{(k-1)}),$$

where we used the global majorization statement of Theorem 2.4 in the first inequality and (2.177) in the second inequality, and the last equality follows directly from (2.176).

Taking this into account, the statement of Theorem 2.6.1 is shown if

$$F_{\epsilon_k}(\mathbf{X}^{(k)}) \leq F_{\epsilon_{k-1}}(\mathbf{X}^{(k)}), \quad (2.178)$$

for ϵ_k chosen as in the smoothing update step (2.91) of `MatrixIRLS`. And indeed, that the latter is true can be inferred from the dependence on ϵ of the univariate functions $f_{p,\epsilon}$ that constitute the objective $F_\epsilon(\mathbf{X}^{(k)}) = F_{p,\epsilon}(\mathbf{X}^{(k)})$ of `MatrixIRLS` (see (2.48)) such that

$$F_\epsilon(\mathbf{X}^{(k)}) = \sum_{i=1}^{d} f_{p,\epsilon}(\sigma_i(\mathbf{X}^{(k)})),$$

where $f_{p,\epsilon}$ is such that

$$f_{p,\epsilon}(\sigma) = \begin{cases} \log\left(\max(\sigma,\epsilon)\right) + \frac{1}{2}\left(\frac{\min(\sigma,\epsilon)^2}{\epsilon^2} - 1\right), & \text{if } p = 0, \\ \frac{1}{p}\max(\sigma,\epsilon)^p + \frac{1}{2}\left(\frac{\min(\sigma,\epsilon)^2}{\epsilon^{2-p}} - \epsilon^p\right), & \text{if } 0 < p \leq 1 \end{cases}$$

for any $\sigma \geq 0$. To show the claim (2.178), we first note that

$$\epsilon_k \leq \epsilon_{k-1}$$

due to (2.91). Defining fixed σ a function $h_\sigma : \mathbb{R}_{\leq 0} \longrightarrow \mathbb{R}$ such that $h_\sigma(\epsilon) = f_{p,\epsilon}(\sigma)$, we obtain

$$h_\sigma'(\epsilon) = \begin{cases} \frac{2-p}{2}\epsilon^{p-1} + \frac{p-2}{2}\frac{\sigma^2}{\epsilon^{3-p}} = \frac{2-p}{2}\epsilon^{p-3}\left(\epsilon^2 - \sigma^2\right) & \text{if } \epsilon > \sigma, \\ 0 & \text{if } \epsilon < \sigma, \end{cases}$$

for all $0 \leq p \leq 1$. This means that h_σ is a piecewise differentiable function with almost everywhere non-negative derivative and, thus, h_σ is a non-decreasing function and in particular,

$$f_{p,\epsilon_k}(\sigma) = h_\sigma(\epsilon_k) \leq h_\sigma(\epsilon_{k-1}) = f_{p,\epsilon_{k-1}}(\sigma). \quad (2.179)$$

As (2.179) holds for all $\sigma = \sigma_i(\mathbf{X}^{(k)})$, the claim (2.178) is shown.

2. We note that

$$F_{\epsilon_k}(\mathbf{X}) = \sum_{i:\sigma_i(\mathbf{X})>\epsilon_k} \log(\sigma_i(\mathbf{X})) + \sum_{i:\sigma_i(\mathbf{X})\leq\epsilon_k} \left(\log(\epsilon_k) + \frac{1}{2}\left(\frac{\sigma_i(\mathbf{X})^2}{\epsilon_k^2} - 1\right)\right)$$

if $p = 0$. Since $\widetilde{\sigma} \mapsto \log(\widetilde{\sigma}^2)$ is concave, we have that

$$\left(\log(\epsilon_k) + \frac{1}{2}\left(\frac{\sigma_i(\mathbf{X})^2}{\epsilon_k^2} - 1\right)\right) \geq \log\left(\sigma_i(\mathbf{X})\right)$$

for any $i \in [d]$ with $\sigma_i(\mathbf{X}) \leq \epsilon_k$, as in this case

$$\log(\sigma) = \frac{1}{2}\log(\sigma^2) \leq \frac{1}{2}\log(\epsilon_k^2) + \frac{1}{2}(\epsilon_k^2)^{-1}\left(\sigma^2 - \epsilon_k^2\right) = \log(\epsilon_k) + \frac{1}{2}\epsilon_k^{-2}\left(\sigma^2 - \epsilon_k^2\right)$$

for any $\sigma > 0$. Similarly for $0 < p \leq 1$, the function he function $h : \mathbb{R}_{\leq 0} \longrightarrow \mathbb{R}$, $h(\widetilde{\sigma}) = \frac{1}{p}\widetilde{\sigma}^{p/2}$ is concave, and therefore

$$\frac{1}{p}\sigma^p = \frac{1}{p}(\sigma^2)^{p/2} \leq \frac{1}{p}(\epsilon_k^2)^{p/2} + \frac{1}{2}(\epsilon_k^2)^{p/2-1}\left(\sigma^2 - \epsilon_k^2\right) = \frac{1}{2}\frac{\sigma^2}{\epsilon_k^{2-p}} + \left(\frac{1}{p} - \frac{1}{2}\right)\epsilon_k^p.$$

With this, we argue that

$$f_{p,\epsilon}(\sigma) \geq \begin{cases} \log(\sigma), \\ \frac{1}{p}\sigma^p, \end{cases}$$

which finishes the proof of Theorem 2.6.2 by inserting $\sigma = \sigma_i(\mathbf{X})$ and summing over all $i \in [d]$.

3. By (2.177), and statement 1., it holds that

$$\begin{aligned} F_{\epsilon_{k-1}}(\mathbf{X}^{(k-1)}) - F_{\epsilon_k}(\mathbf{X}^{(k)}) &\geq F_{\epsilon_{k-1}}(\mathbf{X}^{(k-1)}) - Q_{\epsilon_{k-1}}(\mathbf{X}^{(k)}|\mathbf{X}^{(k-1)}) \\ &= -\frac{1}{2}\left(\langle\mathbf{X}^{(k)}, W^{(k-1)}(\mathbf{X}^{(k)})\rangle - \langle\mathbf{X}^{(k-1)}, W^{(k-1)}(\mathbf{X}^{(k-1)})\rangle\right). \end{aligned}$$

By Lemma 2.4.1, we know that $\mathbf{X}^{(k)}$ from (2.90) fulfills the optimality condition

$$\langle W^{(k-1)}(\mathbf{X}^{(k)}), \eta\rangle = 0 \text{ for all } \eta \in \ker\Phi \text{ and } \Phi(\mathbf{X}^{(k)}) = \mathbf{Y}.$$

Choosing $\eta = \mathbf{X}^{(k-1)} - \mathbf{X}^{(k)}$, we see that

$$\begin{aligned} &\langle\mathbf{X}^{(k)}, W^{(k-1)}(\mathbf{X}^{(k)})\rangle - \langle\mathbf{X}^{(k-1)}, W^{(k-1)}(\mathbf{X}^{(k-1)})\rangle = \\ &= \langle\mathbf{X}^{(k)}, W^{(k-1)}(\mathbf{X}^{(k)})\rangle - \langle\mathbf{X}^{(k-1)}, W^{(k-1)}(\mathbf{X}^{(k-1)})\rangle + 2\langle W^{(k-1)}(\mathbf{X}^{(k)}), \mathbf{X}^{(k-1)} - \mathbf{X}^{(k)}\rangle \\ &= -\left(\langle\mathbf{X}^{(k-1)}, W^{(k-1)}(\mathbf{X}^{(k-1)})\rangle + 2\langle W^{(k-1)}(\mathbf{X}^{(k)}), \mathbf{X}^{(k-1)}\rangle + \langle\mathbf{X}^{(k)}, W^{(k-1)}(\mathbf{X}^{(k)})\rangle\right) \\ &= -\langle(\mathbf{X}^{(k)} - \mathbf{X}^{(k-1)}), W^{(k-1)}(\mathbf{X}^{(k)} - \mathbf{X}^{(k-1)})\rangle \end{aligned}$$

and therefore

$$F_{\epsilon_{k-1}}(\mathbf{X}^{(k-1)}) - F_{\epsilon_k}(\mathbf{X}^{(k)}) \geq \sigma_{\min}(W^{(k-1)})\|\mathbf{X}^{(k)} - \mathbf{X}^{(k-1)}\|_F^2 \geq \max(\sigma_1(\mathbf{X}^{(k-1)}), \epsilon_{k-1})^{p-2}\|\mathbf{X}^{(k)} - \mathbf{X}^{(k-1)}\|_F^2. \tag{2.180}$$

Furthermore, if $p > 0$ or if $\bar{\epsilon} = \lim_{k\to\infty} \epsilon_k > 0$, we know that the sequence $(\mathbf{X}^{(k)})_{k\geq 1}$ is bounded. Indeed, if $p > 0$,

$$\frac{1}{p}\|\mathbf{X}^{(k)}\|_{S_\infty}^p = \frac{1}{p}\sigma_1(\mathbf{X}^{(k)})^p \leq F_{\epsilon_k}(\mathbf{X}^{(k)}) \leq F_{\epsilon_1}(\mathbf{X}^{(1)}) \leq d\frac{1}{p}\max(\sigma_1(\mathbf{X}^{(1)}), \epsilon_1)^p =: c_p$$

for all $k \in \mathbb{N}$ and therefore

$$\sigma_{\min}(W^{(k-1)}) \geq \max(\sigma_1(\mathbf{X}^{(k-1)}), \epsilon_{k-1})^{p-2} \geq \max(pc_p^{1/p}, \epsilon_{k-1})^{p-2} \geq \max(pc_p^{1/p}, \epsilon_1)^{p-2} =: C_p,$$

where we used that $\min_{1\leq i\leq d_1, 1\leq j\leq d_2} \min\left((\mathbf{H}_1^{(k-1)})_{ij}, (\mathbf{H}_2^{(k-1)})_{ij}\right) = \max(\sigma_1(\mathbf{X}^{(k-1)}), \epsilon_{k-1})^{p-2}$ due (2.58) and (2.59) of Definition 2.3.4.
Inserting this into (2.180), we obtain that

$$F_{\epsilon_{k-1}}(\mathbf{X}^{(k-1)}) - F_{\epsilon_k}(\mathbf{X}^{(k)}) \geq C_p\|\mathbf{X}^{(k)} - \mathbf{X}^{(k-1)}\|_F^2$$

and, summing over $2 \leq k \leq K$, this results in

$$F_{\epsilon_1}(\mathbf{X}^{(1)}) \geq F_{\epsilon_1}(\mathbf{X}^{(1)}) - F_{\epsilon_K}(\mathbf{X}^{(K)}) \geq C_p \sum_{k=1}^{K-1} \|\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}\|_F^2$$

and $\sum_{k=1}^{\infty} \|\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}\|_F^2 = 0$ for $K \to \infty$, and therefore

$$\lim_{k\to\infty} \|\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}\|_F = 0.$$

If, on the other hand, $p = 0$, suppose that $\bar{\epsilon} = \lim_{k\to\infty} \epsilon_k > 0$. In this case we can likewise argue that the sequence $(\mathbf{X}^{(k)})_{k\geq 1}$ is bounded as

$$F_{\epsilon_{k-1}}(\mathbf{X}^{(k-1)}) \leq F_{\epsilon_1}(\mathbf{X}^{(1)}) \leq \max(0, \log(\max(\sigma_1(\mathbf{X}^{(1)}), \epsilon_1))) =: c_0$$

and

$$\begin{aligned} F_{\epsilon_{k-1}}(\mathbf{X}^{(k-1)}) &\geq \log(\max(\sigma_1(\mathbf{X}^{(k-1)}), \epsilon_{k-1})) - d\log(\epsilon_{k-1}) - d/2 \\ &\geq \log(\max(\sigma_1(\mathbf{X}^{(k-1)}), \epsilon_{k-1})) - d\log(\bar{\epsilon}) - d/2. \end{aligned}$$

Rearranging the last two inequalities, we obtain that

$$\max(\sigma_1(\mathbf{X}^{(k-1)}), \epsilon_{k-1}) \geq \exp\left(c_0 + d\log(\bar{\epsilon}) + d/2\right)$$

and therefore

$$\sigma_{\min}(W^{(k-1)}) \geq \max(\sigma_1(\mathbf{X}^{(k-1)}), \epsilon_{k-1})^{-2} \geq \exp\left[(-2)\left(c_0 + d\log(\bar{\epsilon}) + d/2\right)\right] =: C_0.$$

As above for $p > 0$, this shows that $\lim_{k\to\infty} \|\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}\|_F = 0$.

4. In the proof of statement 3., we showed that the sequence $(\mathbf{X}^{(k)})_{k\geq 1}$ is bounded. In particular, this already shows that each subsequence of $(\mathbf{X}^{(k)})_{k\geq 1}$ has a convergent subsequence. Let $(\mathbf{X}^{(k_\ell)})_{\ell\geq 1}$ be such a sequence with $\lim_{\ell\to\infty} \mathbf{X}^{(k_\ell)} = \bar{\mathbf{X}}$. As the matrix operator $W^{(k_\ell)}$ of Definition 2.3.4 depends continuously on $\mathbf{X}^{(k_\ell)}$, there exists a weight operator $\bar{W} : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ such that $\bar{W} = \lim_{\ell\to\infty} W^{(k_\ell)}$. Using statement 3. of this theorem,

it also holds that $\mathbf{X}^{(k_\ell+1)} \longrightarrow \bar{\mathbf{Z}}$ and therefore

$$\langle \bar{W}\bar{\mathbf{X}}, \eta\rangle = \lim_{\ell\to\infty} \langle W^{(k_\ell)}(\mathbf{X}^{(k_\ell+1)}), \eta\rangle = 0$$

for all $\eta \in \ker \Phi$. Noting now that the first statement of Theorem 2.4 implies that

$$\nabla F_{\bar{\epsilon}}(\bar{\mathbf{X}}) = \bar{W}(\bar{\mathbf{X}}),$$

as $\bar{W}$ is the weight operator corresponding to $\bar{\mathbf{X}}$, we obtain that

$$\langle F_{\bar{\epsilon}}(\bar{\mathbf{X}}), \eta\rangle = 0 \text{ for all } \eta \in \ker\Phi \text{ and } \Phi(\bar{\mathbf{X}}) = \mathbf{Y},$$

which means that $\bar{\mathbf{X}}$ is a stationary point of $F_{\bar{\epsilon}}$ subject to the linear constraint $\Phi(\mathbf{X}) = \mathbf{Y}$. This finishes the proof of Theorem 2.6.

□

2.8.2 Proof of Lemma 2.4.1

Proof of Lemma 2.4.1. First, we observe that the function $\mathbf{X} \mapsto \langle \mathbf{X}, W^{(k-1)}(\mathbf{X})\rangle$ is strictly convex as the matrices $\mathbf{H}_1^{(k-1)}$ and $\mathbf{H}_2^{(k-1)}$ used in the definition of $W^{(k-1)}$ have all positive entries, cf. Definition 2.3.4. Therefore, we can define the fractional operator $(W^{(k-1)})^{1/2} : \mathbb{R}^{d_1\times d_2} \longrightarrow \mathbb{R}^{d_1\times d_2}$ such that

$$\langle \mathbf{X}, W^{(k-1)}(\mathbf{X})\rangle = \langle (W^{(k-1)})^{1/2}(\mathbf{X}), (W^{(k-1)})^{1/2}(\mathbf{X})\rangle = \langle \mathbf{Z}, \mathbf{Z}\rangle$$

if we define $\mathbf{Z} = (W^{(k-1)})^{1/2}(\mathbf{X})$. With this change of variables, the constraint $\Phi(\mathbf{X}) = \mathbf{Y}$ can be written such that

$$\widetilde{\Phi^{(k-1)}}(\mathbf{Z}) := \Phi((W^{(k-1)})^{-1/2}(\mathbf{Z})) = \mathbf{Y}$$

with the (underdetermined) linear operator $\widetilde{\Phi^{(k-1)}} : \mathbb{R}^{d_1\times d_2} \longrightarrow \mathbb{R}^m$. Then,

$$\mathbf{Z}^{(k)} = \underset{\mathbf{Z}\in\mathbb{R}^{d_1\times d_2}:\widetilde{\Phi^{(k-1)}}(\mathbf{Z})=\mathbf{Y}}{\arg\min} \langle \mathbf{Z}, \mathbf{Z}\rangle = \widetilde{\Phi^{(k-1)}}^{\dagger}(\mathbf{Y}) = \widetilde{\Phi^{(k-1)}}^{*}\left(\widetilde{\Phi^{(k-1)}}\widetilde{\Phi^{(k-1)}}^{*}\right)^{-1}(\mathbf{Y}),$$

where $\widetilde{\Phi^{(k-1)}}^{\dagger}$ is the *Moore-Penrose inverse* of $\widetilde{\Phi^{(k-1)}}$ [Ber09, Proposition 6.1.5] and, reversing the change of variables,

$$\begin{aligned}\mathbf{X}^{(k)} = (W^{(k-1)})^{-1/2}(\mathbf{Z}^{(k)}) &= (W^{(k-1)})^{-1/2} \circ (W^{(k-1)})^{-1/2}\left(\Phi^*(\Phi(W^{(k-1)})^{-1}\Phi^*)^{-1}(\mathbf{Y})\right)\\ &= (W^{(k-1)})^{-1}\left(\Phi^*\left(\Phi(W^{(k-1)})^{-1}\Phi^*\right)^{-1}(\mathbf{Y})\right),\end{aligned}$$

which shows the equality of (2.93).

To show the second statement, we note that (2.93) is equivalent to

$$\langle \mathbf{X}^{(k)} + \eta, W^{(k-1)}(\mathbf{X}^{(k)} + \eta)\rangle > \langle \mathbf{X}^{(k)}, W^{(k-1)}(\mathbf{X}^{(k)})\rangle \text{ for all } \eta \in \ker\Phi \setminus \{0\} \text{ and } \Phi(\mathbf{X}^{(k)}) = \mathbf{Y},$$

which in turn is equivalent to

$$2\langle W^{(k-1)}(\mathbf{X}^{(k)}), \eta\rangle + \langle \eta, W^{(k-1)}(\eta)\rangle > 0 \text{ for all } \eta \in \ker\Phi \setminus \{0\} \text{ and } \Phi(\mathbf{X}^{(k)}) = \mathbf{Y}. \qquad (2.181)$$

It is now easy to observe that holds if and only if (2.94) is true: Suppose that (2.94) holds. As

$W^{(k-1)}$ is positive definite, $\langle \eta, W^{(k-1)}(\eta)\rangle > 0$ for $\eta \in \ker \Phi \setminus \{0\}$ and therefore (2.181) is true.

Conversely, supposing that (2.181) is true would lead to a contradiction if there existed an $\eta \in \ker \setminus \{0\}$ such that $\langle W^{(k-1)}(\mathbf{X}^{(k)}), \eta\rangle > 0$, as then $t\eta \in \ker \Phi \setminus \{0\}$ would fulfill

$$2\langle W^{(k-1)}(\mathbf{X}^{(k)}), t\eta\rangle + \langle (t\eta), W^{(k-1)}(t\eta)\rangle < 0$$

for $t > 0$ chosen appropriately small. Similarly, $\langle W^{(k-1)}(\mathbf{X}^{(k)}), \eta\rangle < 0$ leads to a contradiction. □

2.9 Discussion and Summary of Contributions

We conclude this chapter with a summary of the contributions of our investigations.

In Section 2.2, we reviewed the theory of non-convex rank surrogate objectives that are non-convex *spectral functions*, arguing that they can be considered a promising tool to derive more tractable formulations of the rank minimization problem

$$\min_{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}} \operatorname{rank}(\mathbf{X}) \quad \text{subject to } \Phi(\mathbf{X}) = \mathbf{y},$$

while themselves resulting in optimization problems that are challenging in their own right due to their non-convexity.

We then explored a family of algorithms called *Iteratively Reweighted Least Squares* (IRLS), reviewing existing results and investigating the adaption of this framework to the optimization of rank surrogate objectives. While this has been attempted already in the works [MF12, FRW11, LXY13], we pinpoint why their derivation, which is related to the optimization of high-dimensional functions depending on *separable* variables, leads to multiple different, but sub-optimal IRLS formulations, which we called `IRLS-col` and IRLS-row.

In Sections 2.3.2 to 2.3.4, we proposed a new formulation of IRLS for low-rank matrix recovery named `MatrixIRLS`, which improves on `IRLS-col` and IRLS-row as the underlying *weight operator* corresponds to considerably tighter quadratic bounds on the spectral objectives. The tightness of the quadratic bounds is further related to the second-order derivative structure of spectral functions.

`MatrixIRLS` in comparison with other IRLS algorithms

In this context, we proved Theorem 2.4, which indicates that the quadratic functions induced by the weight operator $W^{(k)}$ (cf. Definition 2.3.4) of `MatrixIRLS` is indeed still a global majorizer of the function to be minimized, ensuring a monotonous decrease of the objective with subsequent iterations of `MatrixIRLS`. The monotonous decrease results in a statement about subsequence convergence of the algorithm's iterates to stationary points of the smoothed logdet/Schatten-p objective (Theorem 2.6)–a statement that had previously also been shown for the older IRLS variants of [MF12, FRW11]. Optimally, we would like, of course, to obtain a statement that ensures convergence to *global minimizers* of our objectives. A statement that strong is, unfortunately, difficult to achieve due to the non-convexity of the Schatten-p quasi norm/the smoothed objectives F_ϵ. From some point of view, a result that strong is beyond the scope of existing theory even for structurally simpler settings IRLS for the related *sparse recovery* problem [DDFG10].

However, for `MatrixIRLS`, it is possible to develop a *local convergence theory*, which we did with Theorem 2.7, our second main theorem. It suggests that under regularity assumptions on the linear operator Φ, `MatrixIRLS` converges with a locally *superlinear convergence rate* of order $2 - p$, where $p \in [0; 1)$ is the non-convexity parameter corresponding to a Schatten-p

quasi-norm objective for $p > 0$ and to a logdet-objective for $p = 0$, and a lower bound on the radius of fast convergence is provided.

This fast convergence rate has neither been proved nor observed in experiments for previous IRLS algorithms [MF12, FRW11, LXY13] of the type of `IRLS-col` and `IRLS-row` (see Remark 2.4.3 and Section 2.7.2), suggesting that our algorithms constitute a substantial improvement on them, see also the experiments of Section 2.7.1 for a numerical verification. Compared to the local convergence result [KS18, Theorem 11], which implies a local convergence rate of order $2 - p$ for `HM-IRLS` if $0 < p \leq< 1$, we note that Theorem 2.7 also covers the important case of $p = 0$, and furthermore [KS18, Theorem 11] admits only a smaller bound on the radius implying fast local convergence [KS18, Theorem 11] than Theorem 2.7.

It is worthwhile to compare the properties of `MatrixIRLS` with the behavior of the IRLS algorithm of [DDFG10] designed to solve the sparse vector recovery problem by mimicking ℓ_p-minimization for $0 < p \leq 1$. [DDFG10, Theorem 7.9, Figure 8.3] provided a comparable local convergence statement as Theorem 2.7. This means that our algorithms can be seen as an adequate generalization of the well-known IRLS for sparse vector recovery [GR97, DDFG10] the setting of the recovery of low-rank matrices.

Furthermore, it can be observed that the convergence of the algorithm of [DDFG10] to a sparse vector often breaks down if p is smaller than 0.5 [DDFG10, Section 8]. In contrast to that, we observed that IRLS algorithms with tight quadratic bounds do not suffer from this loss of global convergence for $p \ll 0.5$ in the low-rank case Section 2.7.3. We explain this difference by the choice of the ϵ-smoothing, which would, translated to the matrix, correspond to

$$\epsilon_k = \min\left(\epsilon_{k-1}, \frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{d}\right),$$

see [DDFG10, Eq. (1.9)]. This smoothing is more adapted to solving the problem for a convex objective with $p = 1$, but too small to be adequate for very non-convex objectives corresponding to $p \approx 0$. Elaborating on this is beyond the scope of this thesis, but we claim that our investigations also contribute how to improve IRLS for sparse vector recovery–we leave translating the proof of Theorem 2.7 to the sparse vector case to future work.

In the context of the large literature about low-rank matrix recovery algorithms, also the *convergence rate* result of Theorem 2.7 is remarkable, since there are only few low-rank recovery algorithms which exhibit either theoretically or practically verifiable superlinear convergence rates. In particular, although the algorithms of [MMBS13] and `NewtonSLRA` of [SS16] do show superlinear convergence rates, the first is not competitive to `MatrixIRLS` in terms of sample complexity and the second has neither applicable theoretical guarantees for most of the interesting problems nor the ability of solving medium size problems.

Optimal Sample Complexity for Local Guarantees of `MatrixIRLS`

We now briefly review the implications of our analysis on the *data efficiency* of `MatrixIRLS`.

In Section 2.5, we verified the assumptions ensuring fast local convergence of `MatrixIRLS` for three models for the measurement operator Φ: Sub-Gaussian operators, phase retrieval and matrix completion. Our analysis showed that local convergence occurs already for a number of measurements or samples m that is *optimal* (up to constants) from a information theoretic point of view, i.e., compared to the number of degrees of freedom of the model.

In this context, we single out the bound on the sample complexity for the problem of completing μ_0-incoherent rank-r matrices of dimension $(d \times d)$ from of Corollary 2.5.3, which guarantees local convergence of `MatrixIRLS` to the rank-r matrix with high probability for m

randomly sampled entries if m fulfills

$$m \geq C\mu_0 rd \log(d), \tag{2.182}$$

for some constant $C \geq 1$. First, we note that the statement of Corollary 2.5.3 has not yet been proven in the paper [KS18], as stronger assumptions than Assumption 2.1 were used to show the local convergence statement.

The bound (2.182) is *optimal* up to the constant C, as has been shown in [CT10, Theorem 1.7], [Che15, Proposition 1], and improves, to the best of our knowledge, on the best bounds for *any* algorithm for low-rank matrix completion. In particular, the state-of-the-art theoretical guarantee by Chen [Che15, Theorem 1], which built on previous work of [CT10, Gro11, Rec11], for the success of the nuclear norm minimization program

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \|\mathbf{X}\|_* \text{ subject to } \Phi(\mathbf{X}) = \mathbf{y}$$

requires

$$m \geq \mu_0 r \log^2(d)$$

sampled entries, which is a *factor of* $\log(d)$ *more samples* than (2.182)[7].

Most of the recent research on *non-convex approaches* for the low-rank matrix recovery or completion problem is considerably different from our approach, as they often start already from a *matrix factorization* approach that assumes iterates $\mathbf{X} = \mathbf{LR}^* \in \mathbb{R}^{d_1\times d_2}$ are already being represented as an (outer) *product* of two matrices $\mathbf{L} \in \mathbb{R}^{d\times r}$, $\mathbf{R} \in \mathbb{R}^{d\times r}$, and then optimize an objective like $F : \mathbb{R}^{d\times r} \times \mathbb{R}^{d\times r} \to \mathbb{R}$,

$$F(\mathbf{L},\mathbf{R}) = \|\Phi(\mathbf{LR}^*) - \mathbf{y}\|_F^2,$$

by regularized or non-regularized gradient descent, or by alternating minimization, see [CLC19] for a survey.

Also for these approaches, convergence guarantees have been obtained, but even the best guarantees require *more* samples than (2.182): For unregularized gradient descent, recent results [MWCC19, CLL19] showed that

$$m \geq Cc(\kappa)\mu_0^2 r^2 \log(d)$$

samples suffice for convergence (with $c(\kappa)$ being a constant depending on the condition number κ of $\mathbf{X}_0$), which is suboptimal with respect to (2.182) by a factor of $\mu_0 r$. Another algorithm with one of the smallest requirements on m is the Grassmann manifold based gradient descent algorithm of [KMO10], which provably converges if a good initialization is provided and if the sample complexity fulfills

$$m \geq \mu_0^2\kappa^2 rd \max(\kappa^4 r, \log(d)).$$

We would like to stress here that the provided convergence analysis of `MatrixIRLS` is of *local* nature, and does not guarantee that the algorithm converges from any initialization to the desired low-rank matrix.

[7] We also note that our assumption on the μ_0-incoherence of the matrix $\mathbf{X}_0$ μ_0-incoherence is slightly weaker than the one of [Che15], see Definition 3.3.1.

State-of-the-art Data Efficiency

However, the experimental results of Section 2.7.3 suggest that at least in the investigated setting of recovering random low-rank matrices from few, random entrywise samples (low-rank matrix completion), tight quadratic bound IRLS algorithms for non-convex rank surrogates perform very favorably in terms of data efficiency, i.e., they need very few data samples m to reconstruct low-rank matrices. The experiment was designed in a representative manner, comparing to previous IRLS algorithms as `IRLS-col`, but also to other algorithms representative of the state-of-the art, such as Riemannian optimization techniques [Van13], alternating minimization approaches [HH09, TW16], algorithms based on iterative hard thresholding [KC14, BTW15], and others [PKCS18]. We saw that `HM-IRLS` recovers low-rank matrices systematically with an optimal number of measurements that is very close to the theoretical lower bound on the number of measurements that is known to be necessary for recovery, with high empirical probability.

One possible conclusion of the experiments is that in fact, despite the severe non-convexity of the underlying objective for $p \ll 1$, convergence of `MatrixIRLS` to bad local minima or saddle points happens only rarely, as otherwise, the good empirical recovery rate observed in Section 2.7.3 could not be realized. This indicates that the performance of `MatrixIRLS` is robust to the choice of the initialization and that it can be used as competetive a standalone algorithm to recover low-rank matrices, and low requirements on the sample complexity indicated by the *local* statements of Section 2.5.1 might be extendable to a *global* convergence analysis.

Scalability: Breaking Quadratic Complexity for IRLS Algorithms

Finally, another important aspect of our contributions are the computational considerations for `MatrixIRLS` presented in Section 2.6. In that section, we presented how it is possible to implement `MatrixIRLS` in the case of matrix completion actually in a way where all basic operations require only a time complexity of

$$O(mr + r^2 D),$$

where r is the target rank, $D = \max(d_1, d_2)$ the maximum of the ambient dimensions of the matrix to be recovered and m the number of observed entries. This is precisely the order of the complexity of gradient descent approaches optimizing a data fit objective such as (2.3) based on matrix factorization [CC18, p. 23].

From a scalability point of view, this makes IRLS competitive with the currently most studied and most popular approaches for the low-rank matrix recovery problem. Compared to first-order methods based on matrix factorization, our method has the advantage that there is no need to choose step sizes of gradient steps, for which it might be difficult to find good general rules–our method's main free parameters are related to the accuracy of the solutions linear system to be solved by a conjugate gradient method, which has been studied in abundance in numerical linear algebra for a long time.

The papers [MF12] and [FRW11], which introduced the first IRLS variants for low-rank recovery, also elaborated on the implementation of their respective methods for matrix completion.

[MF12] uses two different projected gradient descent approaches in their implementation, working always with full matrices and full singular values decompositions, which scale in general cubically in the matrix dimension such that $O(d^2 D)$.

The paper [FRW11] covers an algorithm of the type of `IRLS-col` for $p = 1$, but not the generalization for $p < 1$. However, generalizing their implementation in the most natural way to $p < 1$ would lead to an objective similar to the objective (2.92) of MatrixIRLS, i.e., similar to

$$F_\epsilon(\mathbf{X}) = \sum_{i=1}^{d} \left[\frac{1}{p} \max(\sigma_i(\mathbf{X}), \epsilon)^p + \frac{1}{2} \left(\frac{\min(\sigma_i(\mathbf{X}), \epsilon)^2}{\epsilon^{2-p}} - \epsilon^p \right) \right].$$

Furthermore, a variant of the "Woodbury trick" of Section 2.6.2 is used [FRW11, Section 4.2.1] to reduce the problem size from solving the $(m \times m)$ linear systems

$$\left(\Phi (W^{(k)})^{-1} \Phi^* \right) \mathbf{z} = \mathbf{y}$$

to d_2 systems of size $r \times r$. The authors use crucially that a weight operator $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ such as (2.28) acts *column-wise* on $\mathbf{X}$, creating a certain separability of the problem in the case of matrix completion. This separability is not preserved if we use weight operators such as the harmonic mean operator of [KS18] or MatrixIRLS, which means that the simplification of [FRW11] is not applicable in our case. Overall, [FRW11] still works with full matrices stored across iterations in memory.

We note that due to the column- or row-wise action of the weight operators of [MF12, FRW11], it is not possible to obtain an algorithm for the *phase retrieval* setting of Section 2.5.2 with their weight operator, as the Hermiticity of the matrices is not preserved, unlike for MatrixIRLS (see Proposition 2.5.2).

Due to the usage of an objective F_ϵ that is slightly different from (2.92), i.e., by using

$$F_\epsilon(\mathbf{X}) = \sum_{i=1}^{d} \left(\sigma_i^2(\mathbf{X}) + \epsilon^2 \right)^{\frac{p}{2}}$$

for $0 < p \leq 1$ which necessitates the calculation of all singular values in each iteration, an implementation with the advantages of Section 2.6 had not been possible for the tight quadratic bound IRLS variant HM-IRLS described in [KS18].

2.10 Outlook

Beyond the results we covered in this chapter, there are many questions related to IRLS algorithms for low-rank matrix recovery that remain for future investigations.

First of all, the part of our convergence analysis that includes convergence rates is *inherently local* and does not imply a convergence of MatrixIRLS to the solution of the affine rank minimization problem (2.1) starting from *any* initialization point $\mathbf{X}^{(1)}$.

There are ways how to modify our algorithm to obtain a provably *globally convergent* algorithm with superlinear local convergence rates for interesting sample complexities. For example, it would be possible to devise a two-stage procedure, which in the first stage, for example, uses a projected gradient scheme as in [CW15, ZL16a] to bring an iterate $\mathbf{X}^{(k)}$ into the basin of fast convergence (2.138) of MatrixIRLS, and then use MatrixIRLS in a second phase to obtain fast convergence starting from $\mathbf{X}^{(k)}$. Using this first stage, it would only take $O(log(D))$ (sub-Gaussian operators) or $O(\mu_0 r \log(D))$ (matrix completion) iterations [CC18] to arrive at our basin of fast convergence in the case of $p = 0$, but the convergence of the first stage would require a larger sample complexity (which is still *near-optimal* in the sense that m would need to scale linearly in D up to a logarithmic factor, but with a dependence on higher powers of r).

However, we did not pursue this road due to the good empirical behavior of our framework, which does not seem to require a distinct first stage. We think that an explanation of this phenomenon is that while there are many saddle points in the landscape of F_ϵ if the number of observations m is close to the number of degrees of freedom $\deg_f$, `MatrixIRLS` is able to *escape* them quickly. Connecting the structure of the weight operator to the second-order structure of the spectral objectives (see Section 2.3.2) can be a key tool for establishing this rigorously, building on recent ideas about saddle-escaping modified Newton methods for non-convex optimiztion [PMR19]. In fact, our framework fits well into the one of [PMR19], as the entries of the matrices $\mathbf{H}_1^{(k)}$ and $\mathbf{H}_2^{(k)}$ of (2.58) and (2.59) in the definition of the weight operator $W^{(k)}$ are very similar to the *absolute value* of the eigenvalues of the Hessian of of F_ϵ (see Lemma 2.3.1).

Furthermore, it remains slightly unclear why using weight operators with the *harmonic mean* of $\widetilde{\sigma}_i^{(k)}$ and $\widetilde{\sigma}_j^{(k)}$ (in the notation of Definition 2.3.4) or even taking weight operators $W^{(k)}$ corresponding to the q-power mean with $q = -\infty$ and setting $\mathbf{H}_2^{(k)} = \mathbf{H}_1^{(k)}$, resulting in the formula $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ such that

$$W^{(k)}(\mathbf{Z}) = \mathbf{U}_k \left[\mathbf{H}^{(k)} \circ (\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k)\right] \mathbf{V}_k^*,$$

for $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$ where $\mathbf{H}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ with

$$(\mathbf{H}^{(k)})_{ij} = M_{-\infty}(\widetilde{\sigma}_i^{(k)}, \widetilde{\sigma}_j^{(k)}) = \min\left(\max(\sigma_i^{(k)}, \epsilon)^{p-2}, \max(\sigma_j^{(k)}, \epsilon)^{p-2}\right),$$

leads to a very good and even better empirical behavior than the precise choice of weight operators in `MatrixIRLS`. While for these “extremely tight” weight operators our proof of the global majorization result Theorem 2.4 does not go through and therefore leaves the interpretation of the corresponding IRLS algorithm as a monotonously minimizing scheme of the objective F_ϵ (see Theorem 2.6) still open, our result on locally superlinear convergence under very weak assumptions on Φ can be easily extended to these operators. This leads to the question whether the proof of Theorem 2.4 can be improved, ensuring the same property for even tighter weight operators.

In many, if not almost all practical applications of low-rank modelling, it is to be expected that the linear measurements $\mathbf{y} = \Phi(\mathbf{X}_0)$ are not exact, but perturbed by (unknown) noise, similar to the sparse recovery setting of Chapter 1. In these cases, the measurements would rather correspond to

$$\mathbf{y} = \Phi(\mathbf{X}_0) + \mathbf{w}$$

with $\mathbf{w} \in \mathbb{R}^m$ such that, e.g., $\|\mathbf{w}\|_2 \leq \eta$ for some small $\eta > 0$.

The IRLS framework, including an analysis similar to [DDFG10, FRW11], has already been extended to this case in [LXY13]. Without any conceptual changes, it is possible to translate `MatrixIRLS` to this setting by solving unconstrained problems

$$\mathbf{X}^{(k)} = \operatorname*{arg\,min}_{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}} \|\Phi(\mathbf{X}) - \mathbf{y}\|_2^2 + \lambda \langle \mathbf{X}, W^{(k-1)}(\mathbf{X}) \rangle.$$

instead of (2.90) for a suitable tuning parameter $\lambda > 0$, optimizing an objective $\mathcal{J}_{p,\epsilon}$ that is a sum of a smoothed Schatten-p or log-det objective and a data-fit term such that

$$\mathcal{J}_{p,\epsilon}(\mathbf{X}) = \|\Phi(\mathbf{X}) - \mathbf{y}\|_2^2 + \lambda F_{p,\epsilon}(\mathbf{X})$$

with $F_{p,\epsilon}$ as in (2.48).

We leave the adaption of the theoretical analysis to this unconstrained variant to future work, noting that on a computational level, the two algorithms will be very similar.

Finally, we think that an adaption of our framework to other, related low-rank matrix optimization problems is very well possible and promising. For example,an extension of these ideas to problems where both *sparsity* and *low-rank* structures are available, such as in *robust principical component analysis* [CLMW11], *sparse phase retrieval* [JOH13, IVW17] or *blind deconvolution* [ARR14, LLSW19] remain for future work. The first work in this direction is presented in Chapter 3.

Chapter 3

Recovery of Structured Low-Rank Matrices by Iteratively Reweighted Least Squares

In this chapter, we build on the optimization framework presented in Chapter 2 to develop and analyze an algorithm for a different problem.

Similarly to Chapter 2, consider a linear operator Φ mapping from complex $(d_1 \times d_2)$-matrices to a m-dimensional vector which is such that it is has non-trivial nullspace, i.e., with

$$m < d_1 d_2.$$

If $r \in \mathbb{N}$ is the target rank and working on a *complex* domain, taking a different point of view, the low-rank matrix recovery problem of Chapter 2 can be reformulated using the non-convex and non-smooth set

$$\mathcal{M}_{\leq r} := \{\mathbf{X} \in \mathbb{C}^{d_1 \times d_2} : \operatorname{rank}(\mathbf{X}) \leq r\}$$

of matrices with rank at most r as solving the problem

$$\text{Find } \mathbf{X} \in \mathbb{C}^{d_1 \times d_2} \quad \text{such that } \Phi(\mathbf{X}) = \mathbf{y} \text{ and } \mathbf{X} \in \mathcal{M}_{\leq r}, \tag{3.1}$$

where $\mathbf{y} \in \mathbb{C}^m$ is the vector with measurement data.

Now what if we knew, for some reason, that the matrix $\mathbf{X}$ we are looking for additionally *lies in a subspace* $\mathcal{S}$ of considerably smaller dimension $\dim(\mathcal{S}) \ll d_1 d_2$ than the ambient matrix space?

In this scenario, (3.1) would become the problem

$$\text{Find } \mathbf{X} \in \mathbb{C}^{d_1 \times d_2} \quad \text{such that } \Phi(\mathbf{X}) = \mathbf{y} \text{ and } \mathbf{X} \in \mathcal{M}_{\leq r} \cap \mathcal{S}. \tag{3.2}$$

As we are interested here in the recovery of low-rank matrices that possess an *an additional linear structure* indicated by $\mathcal{S}$, we call (3.2) a *structured low-rank matrix recovery* problem [MU13, CC14, SLO$^+$14].

From a mathematical point of view, the additional subspace constraint can actually be helpful as there is hope to identify a matrix $\mathbf{X}_0 \in \mathcal{M}_{\leq r} \cap \mathcal{S}$ such that $\mathbf{y} = \Phi(\mathbf{X}_0) \in \mathbb{C}^m$ from a number of measurements m that is *smaller* than what we needed in Chapter 2, as the number of degrees of freedom of the model might by smaller than $r(d_1 + d_2 - r)$. However, the additional constraint might pose computational difficulties.

The content of this chapter is partially based on the conference publications [KV18, KV19]. The author of this dissertation is the main author of these publications, which are based on joint work and co-authored with Claudio Verdun. Preliminary versions of this work have been presented at the 4th International Traveling Workshop on Interactions between low-complexity data models and Sensing Techniques (iTWIST) in November 2018 in Marseille, France, in July

2019 at the 13th International Conference on Sampling Theory and Applications in Bordeaux, France, and in September 2019 at the Workshop on Low-Rank Models and Applications (LRMA) in Mons, Belgium.

3.1 Introduction

We first motivate the setting described by (3.1) and define the problem precisely in Sections 3.1.1 and 3.1.2, before we state our contributions and related work in Sections 3.1.3 and 3.1.4.

3.1.1 Motivation

Already in some of the first works in which the nuclear norm relaxation of the rank mentioned in Chapter 2 was proposed [Faz02, FHB04], an important application of their framework was seen in *system identification* and *control theory*. In fact, if the discrete, noiseless trajectory of a linear time-invariant system is arranged into a *Hankel* matrix, it can be seen that the resulting matrix is of the rank of the model order of the system [Mar19]. This modelling can be used to identify the system from realizations with noisy or incomplete data, for example by using the nuclear norm of the Hankel matrix as a part of an objective function [LV10, FPST13].

A *Hankel* is simply a matrix with *constant anti-diagonals*. More precisely, for any $n, d_1 \in \mathbb{N}$ such that $d_2 = n - d_1 + 1$, we can define the (linear) *Hankel operator* $\mathscr{H}$ (associated to (n, d_1)) such that $\mathscr{H} : \mathbb{C}^n \longrightarrow \mathbb{C}^{d_1 \times d_2}$,

$$\mathbf{x} = (\mathbf{x}_0, \ldots, \mathbf{x}_{n-1}) \mapsto \mathscr{H}(\mathbf{x}) = \begin{bmatrix} \mathbf{x}_0 & \mathbf{x}_1 & \mathbf{x}_2 & \iddots & \mathbf{x}_{d_2-1} \\ \mathbf{x}_1 & \mathbf{x}_2 & \iddots & \iddots & \mathbf{x}_{d_2} \\ \mathbf{x}_2 & \iddots & \iddots & \iddots & \iddots \\ \iddots & \iddots & \iddots & \iddots & \iddots \\ \mathbf{x}_{d_1-1} & \iddots & \iddots & \iddots & \mathbf{x}_{n-2} \\ \mathbf{x}_{d_1} & \mathbf{x}_{d_1+1} & \iddots & \mathbf{x}_{n-2} & \mathbf{x}_{n-1} \end{bmatrix}, \tag{3.3}$$

and say that $\mathbf{X} \in \mathbb{C}^{d_1 \times d_2}$ is a *Hankel matrix* if there exists $\mathbf{x} \in \mathbb{C}^n$ such that

$$\mathbf{X} = \mathscr{H}(\mathbf{x}).$$

The parameter d_1 is often called *pencil parameter* or *window length* [HS90, YKJL17, GZ13, PTV16].

In the sense of (3.2), the Hankel operator $\mathscr{H}$ then defines a linear subspace $\mathcal{S}$ of $\mathbb{C}^{d_1 \times d_2}$ such that

$$\mathcal{S} = \left\{ \mathbf{X} \in \mathbb{C}^{d_1 \times d_2} : \exists \mathbf{x} \in \mathbb{C}^n \text{ such that } \mathscr{H}(\mathbf{x}) = \mathbf{X} \right\},$$

and $\mathcal{S}$ has the dimension $\dim(\mathcal{S}) = n$. Since $n = d_1 + d_2 - 1$ and cannot be larger than $D = \max(d_1, d_2)$, this is often a subspace of *considerably smaller dimension* than the ambient matrix space $\mathbb{C}^{d_1 \times d_2}$.

The occurrence of matrices with Hankel or related structures goes far beyond control theory. In signal processing, an important problem is the estimation or the retrieval of harmonic signals from just a finite number of continuous time samples [SM05]. Some of the most common methods for this problem used in applications are so-called *subspace methods* such as `ESPRIT` [RK89], `MUSIC` [Sch86] and the *matrix pencil method* [HS90, Hua92], whose main steps can be considered as partial singular value decompositions of the Hankel matrix related to the sample sequence [PT13, PTV16]. The underlying mathematical problem goes back to the year 1795 and G. de Prony [dP95][1] and can be stated as follows:

If $x : [0,1] \to \mathbb{C}$ is a periodic, continuous function ("signal") that is a superposition of $r \ll n$ exponentials, i.e., its value at the time $t \in [0,1]$ fulfills

$$x(t) = \sum_{k=1}^{r} \alpha_k \, \mathrm{e}^{2\pi i f_k t}, \tag{3.4}$$

where $\alpha_1, \ldots, \alpha_r \in \mathbb{C}$ are (unknown) complex amplitudes and $f_1, \ldots, f_r \in [0,1)$ are fixed (unknown) frequencies, we would like to reconstruct the amplitudes $(\alpha_k)_{k=1}^r$ and the frequencies $(f_k)_{k=1}^r$ just from a finite sequence of n samples

$$\mathbf{x} = (\mathbf{x}_0, \ldots, \mathbf{x}_{n_1})$$

with

$$\mathbf{x}_j = x\left(\frac{j}{n}\right) \tag{3.5}$$

for $j \in \{0, 1, \ldots, n-1\}$.

This problem had been solved in its simplest form by Prony already more than 200 years ago, but unfortunately, the classical method is quite sensitive to noise on the time samples [SP95]. For this reason, a variety of methods has been proposed to obtain more stable estimators. In this context, the above mentioned subspace methods are one important class, but also different approaches such as approximate Prony's methods [PT10], generalized Prony's methods [PP13], approaches based on continuous total variation or atomic norm regularization [BTR13, BTSR13, CFG14, YX16], on compressed sensing (by discretizing the frequency space) [MCW05, SB12], sparse Bayesian learning [HBFR14], and methods based on non-convex optimization [FWS$^+$16].

If we include variations and multi-dimensional generalizations of the model (3.4), there are very different names for the above modelling in different communities, including spectral compressed sensing [DB13], line spectral estimation [HS90, BTR13], harmonic retrieval [KAR83] and super-resolution [CFG14].

Techniques solving (3.4) are useful to solve a variety of engineering problems in direction of arrival estimation [YLSX17], fluorescence microscopy [SHL10], analog-to-digital conversion [TLD$^+$10] and radar and sonar imaging [PEPC10]. Finally, we refer to a generalization of the exponential sum model by the concept of *finite rate of innovation* (FRI) [DVB07, MV05, VMB02], which is well-known in signal processing.

Importantly, what relates (3.4) with structured low-rank matrices is the following observation.

Proposition 3.1.1 (Low-rankness of Hankel matrices from sum-of-exponential samples (see e.g. [Eld15, Proposition 15.2])). *Let $x : [0,1] \to \mathbb{C}$ be a continuous function fulfilling (3.4) for parameters $\alpha_1, \ldots, \alpha_r \in \mathbb{C}$ and $f_1, \ldots, f_r \in [0,1)$ such that none of the f_k coincide. If $\mathcal{H} : \mathbb{C}^n \to \mathbb{C}^{d_1 \times d_2}$ is the Hankel operator (3.3) associated to (n, d_1) and if*

$$r < \min(d_1, d_2),$$

then

$$\operatorname{rank}(\mathcal{H}(\mathbf{x})) = r,$$

[1] The reader can check the appearance of Hankel matrices in the original manuscript by de Prony: http://users.polytech.unice.fr/~leroux/PRONY.pdf

where $\mathbf{x} = ((\mathbf{x})_0, \ldots, (\mathbf{x})_{n_1})$ *is the vector of consecutive samples* $\mathbf{x}_j = x\left(\frac{j}{n}\right)$, $j \in \{0, 1, \ldots, n-1\}$, *of the function* x.

Proposition 3.1.1 is the foundation of subspace methods like ESPRIT. They use that in the case of noisy samples

$$\widetilde{\mathbf{x}}_j = \mathbf{x}_j + \mathbf{w}_j \tag{3.6}$$

with some noise vector $\mathbf{w} \in \mathbb{C}^n$ with not too large entries $\mathbf{w}_j$, the Hankel matrix $\mathscr{H}(\widetilde{\mathbf{x}})$ is not too far from a rank-r matrix.

This concept can be generalized in multiple ways, for example, to multi-dimensional signals x, where the corresponding low-rank matrix is then not Hankel, but *block-Hankel (with Hankel blocks)* or *multi-level Hankel* [Mar19]. In several applications, it is not necessary to estimate the frequencies $(f_k)_{k=1}^r$, but recovering the noiseless samples by using the low-rank property of their associated structured matrix implicitly is sufficient. This is for example the case in control theory [FPST13, Mar19], in *geophysics*, where there has been a recent interest in methods involving rank-reduction of Hankel-type matrices related to missing data problems [GSC13, OS11, JYLM16], or in time series analysis, where this technique is known as *singular spectrum analysis* [GNZ01]. The same idea applies to *inpainting* or *denoising* problems in imaging, where the spectral sparsity of *image patches* in the Fourier domain can be taken advantage of, and the *rank* can play the role of a regularizer [YKJL17, JY15, OJ16]. In particular in magnetic resonance imaging (MRI), these techniques have been applied successfully [Hal13, LJK+16, MJKM17, WM17].

All the mentioned ideas can be also transferred to low-rank (block-/multi-level) *Toeplitz matrices*, which are matrices with constant *diagonals*. Indeed, this can be easily seen from the fact that a Hankel matrix can be transformed into a Toeplitz matrix through a permutation of its rows and columns, and this operation, due to the orthonormality of permutation matrices, does not affect the rank.

Encompassing also other generalizations, we call a matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ *structured* if its elements can be determined from substantially fewer than $d_1 d_2$, i.e., if it has $O(D) = O(\max(d_1, d_2))$ degrees of freedom.

3.1.2 Problem statement

In the rest of this chapter, we will restrict ourselves to matrices with *Hankel* structure as in (3.3), noting that the setting can be easily adapted to other linear matrix structures.

As we saw in Proposition 3.1.1, low-rank Hankel matrices arise naturally in many applications. In analogy to the unstructured low-rank matrices discussed in Chapter 2, we now consider the problem of their *recovery* from underdetermined linear measurements.

In particular, we focus on the *completion* of low-rank Hankel matrices. This setting arises under the exponential sum model of (3.4) in the case of *missing samples*: Assuming that not a sequence of equidistant samples $\mathbf{x} = ((\mathbf{x})_0, \ldots, (\mathbf{x})_{n_1})$ fulfilling (3.5) are given, but, on the other hand, just a *subset* $\Omega \subset \{0, 1, \ldots, n-1\}$ of these, completing the sequence of samples by using the low-rank prior on the Hankel matrix is of interest as a means to compensate for the loss of data.

In particular, let $m < n$, $\Omega = \{\omega_1, \ldots, \omega_m\} \subset \{0, \ldots, n-1\}$ be an index set of cardinality $|\Omega| = m$, and $P_\Omega : \mathbb{C}^n \to \mathbb{C}^m, \mathbf{x} \mapsto P_\Omega(\mathbf{x}) = \sum_{\omega_j \in \Omega} e_{\omega_j}^* \mathbf{x}$ be the linear subsampling operator associated to the subset Ω. Let $\mathbf{y} := P_\Omega(\mathbf{x}_0) \in \mathbb{C}^m$ correspond to the partial samples of an unknown full sample vector $\mathbf{x}_0 \in \mathbb{C}^n$. Then we define the *low-rank Hankel matrix completion* problem such that

$$\widehat{\mathbf{X}} = \mathscr{H}(\widehat{\mathbf{x}}) \text{ s.t. } \widehat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{C}^n, P_\Omega(\mathbf{x})=\mathbf{y}}{\arg\min} \ \operatorname{rank}(\mathscr{H}(\mathbf{x})), \tag{3.7}$$

or in other words, the problem of finding the Hankel-structured matrix $\mathbf{X} \in \mathbb{C}^{d_1\times d_2}$ of lowest rank compatible with the samples $\mathbf{y} \in \mathbb{C}^m$.

While the problem (3.7) is the main focus of this chapter, we also consider the *structured low-rank approximation problem* (SLRA) associated to $\mathbf{x} \in \mathbb{C}^n$ and rank $r < \min(d_1, d_2)$ defined as

$$\widehat{\mathbf{x}} = \underset{\substack{\mathbf{z}\in\mathbb{C}^n s.t.\\ \operatorname{rank}(\mathscr{H}(\mathbf{z}))\leq r}}{\arg\min} \ \|\mathscr{H}(\mathbf{z}) - \mathscr{H}(\mathbf{x})\|_{F(w)}^2, \tag{3.8}$$

where $\|\cdot\|_{F(w)}$ is a suitable *weighted Frobenius norm* [CH15, Mar19] associated to a collection of weights $\mathbf{w} \in \mathbb{C}^{d_1\times d_2}$, $\mathbf{w}_{ij} > 0$ for all $i \in [d_1], j \in [d_2]$ induced by the inner product

$$\langle \mathbf{A}, \mathbf{B}\rangle_{F(\mathbf{w})} = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} \mathbf{w}_{ij}\overline{\mathbf{A}}_{ij}\mathbf{B}_{ij},$$

where $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{d_1\times d_2}$.

Due to the interplay of the linear Hankel subspace and the non-convex low-rank structure, it is non-trivial to find efficient computational approaches to solve (3.7) or (3.8).

3.1.3 Our Contribution

In this chapter, we develop a new Iteratively Reweighted Least Squares (IRLS) algorithm for the *structured low-rank completion* problem. It optimizes a log-det objective over a subspace of structured matrices, as for example, the space of Hankel matrices. These matrices are of relevance for the harmonic retrieval problem [CWW18] and for system identification [FPST13].

As IRLS algorithms for similar problems, our approach minimizes quadratic upper bounds on the smoothed surrogate objective with iteratively updated smoothing. Building on the framework introduced in Chapter 2, the proposed algorithm called `StrucIRLS` uses tight quadratic bounds motivated from the second-order derivative structure of the smoothed rank surrogate. Extending the general framework of Chapter 2 to the structured case, we incorporate the linear subspace structure into the IRLS framework, which is crucial to solve the original problem in a information theoretically and computationally attractive way.

We prove that locally, `StrucIRLS` converges to an incoherent rank-r Hankel matrix $\mathscr{H}(\mathbf{x}_0)$ with locally quadratic convergence rate, with high probability, under the assumption that the sample locations $\Omega = \{\omega_1, \ldots, \omega_m\}$ are uniformly distributed in $\{0, \ldots, n-1\}$, if the sample complexity or number of samples m fulfills

$$m \geq C\mu_0 c_s r \log(n), \tag{3.9}$$

where C is an absolute constant and

$$c_s = \max\left(\frac{n}{d_1}, \frac{n}{d_2}\right) \tag{3.10}$$

is a constant that depends on the window length d_1, see Theorem 3.2 for details. This result indicates a favorable *data-efficiency* of the algorithm, as the required sample complexity is near-optimal, being just a constant and a logarithmic factor in n larger than the number of degrees of freedom

$$\deg_f = 2r$$

of the rank-r Hankel matrix. The requirement on the sample complexity (3.9) is *weaker* than comparable conditions implying the convergence of other methods for the problem.

Apart from incorporating the subspace constraint without any need of additional projections, we show that our approach can be computationally attractive, as it only contains operations that scale *sub-linearly* in the dimension of the matrix space. This is of particular importance in real-world engineering instances of the problem, where n, d_1 and d_2 can be very large and a quadratic growth of dimensionalty from $\mathbf{x}$ to $\mathscr{H}(\mathbf{x})$ can be a computational issue. Furthermore, for multi-dimensional harmonic retrieval problems [CC18, JLY16, AC14], $\mathscr{H}$ can be a block Hankel matrix operator whose size is non-trivial except from for very small problem instances,

In experiments with artificial data arising from the problem of spectral super-resolution of frequencies with small separation, we observe that `StrucIRLS` exhibits an empirical recovery probability close to one from fewer samples than existing state-of-the-art approaches for the low-rank Hankel matrix completion task.

3.1.4 Related Work

The low-rank Hankel approximation problem (3.8) is older than the corresponding completion problem (3.7): A simple approach for the former was proposed by Cadzow [Cad88], alternating between projections onto the rank-r manifold (which correspond to partial singular value decompositions) and onto the Hankel subspace. However, the optimality of the method is not clear, and counterexamples for the convergence of the method can be constructed, cf. [Moo94, Section VI]. Beside recalling the above-mentioned connection to subspace methods in signal processing, we mention the non-convex optimization approaches that have been proposed in [CH15, IUM14].

For the low-rank Hankel completion problem, an alternating projection method adapting the denoising variant [Cad88] had been proposed in [OS11] in a geophysics journal. Building up on the success of nuclear norm minmization for unstructured matrix recovery problems, Chen and Chi [CC14] analyzed the *Hankel nuclear norm minimization* (`HNNM`) problem (that had been proposed before in control theory [FHB04, FPST13, LV10])

$$\widehat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{C}^n, \mathscr{P}_\Omega(z)=\mathbf{y}}{\arg\min} \; \|\mathscr{H}(\mathbf{x})\|_{S_1} \tag{3.11}$$

and showed guaranteed recovery for (3.11), with high probability, under the assumption that the ground truth $\mathscr{H}(\mathbf{x}_0)$ is μ_0-incoherent, and that $m = |\Omega|$ locations in Ω are sampled uniformly at random, given a number of samples that is near-optimal up to several factors logarithmic in n, or more precisely, that is such that

$$m \geq C\mu_1 c_s r \log^4(d_1 d_2) \tag{3.12}$$

with absolute constant C, c_s as in (3.10) and μ_1 being a suitable incoherence factor. This was improved by [JLY16] to two log factors in the dimension in the special case of *wrap-around Hankel matrices* [JLY16, Theorem 2]. As we saw in Chapter 2, nuclear norm-based modelling has their approaches computational limitations as it is equivalent to solving a semidefinite program (SDP). Even for fast solvers, this results in the very best case in optimization problems with $O(D^2) = O(n^2)$ variables. They also considered the multi-dimensional generalizations (multi-level/block).

Computationally more advantageous non-convex optimization strategies such as *(fast) iterative hard thresholding* (`FIHT`) [CWW19] and *projected gradient descent* (`PGD`) [CWW18] have also been recently developed for low-rank Hankel completion. In practical applications, these algorithms are of greater interest than nuclear norm minimization due to their higher empirical recovery rate and their more scalable implementation [CWW19]. If a two-step procedure with distinct intialization phase is used for FIHT, the sample complexity requirement of [CWW19] is

$$m \geq C\mu_1 c_s \kappa^6 r^2 \log(n) \log(\sqrt{n}\log(n)). \tag{3.13}$$

Furthermore, the projected gradient descent approach of [CWW18] possesses a sufficient condition for convergence from

$$m \geq C\mu_1^2 c_s \kappa^2 r^2 \log(n)). \tag{3.14}$$

In both formulas, $\kappa = \frac{\sigma_1(\mathscr{H}(\mathbf{x}_0))}{\sigma_r \mathscr{H}(\mathbf{x}_0))}$ corresponds to the condition number of the ground truth Hankel matrix $\mathscr{H}(\mathbf{x}_0)$.

We note that in comparison to (3.9), the assumptions on the sample complexity of (3.12)–(3.14) are *stronger*. However, we would like to stress that unlike the guarantees for HNNM, PGD or FIHT, our result is of *local* nature, i.e., Theorem 3.2 not even implies that StrucIRLS converges from the initialization we chose. In this sense, the situation corresponds to what we proved about MatrixIRLS for unstructured reccovery in Chapter 2.

As a summary, we observe that just from the current state of the theory, it is hard to quantify which algorithmic approach provides an optimal tradeoff between *data efficiency*, i.e., the ability to identify the model for a number of samples m as small as possible, *space complexity* and *time complexity*.

Finally, we note that in [OJ17], an IRLS approach has already been used for the Hankel or Toeplitz structured low-rank matrix recovery problem. The algorithm GIRAF (Generic Iterative Reweighted Annihilating Filter) of [OJ17] is a fast algorithm desinged for large-scale 2D MRI reconstruction problems. It corresponds to a low-rank IRLS algorithm based on IRLS-row (see Section 2.3.2 and [MF12, FRW11]), but applied to a *half-circulant extension* of the Toeplitz matrix arising from the data recovery problem. This half-circulant extension is not expected to be very low-rank even if $\mathscr{H}(\mathbf{x})$ is, which renders the connection to the original problem unclear. However, the authors of [OJ17] report good empirical results despite this fact. The authors of [OJ17] do not provide a convergence analysis for their algorithm.

3.1.5 Outline

This chapter is organized as follows. In Section 3.2, we introduce the proposed IRLS algorithm called StrucIRLS for the completion of structured low-rank matrices, outlining the main steps of the iterative procedure. In Section 3.3 we present our main theoretical results, interpreting the algorithm in terms of a non-convex rank objective and providing the local convergence guarantee with quadratic convergence rate. In Section 3.4 we detail computational aspects of StrucIRLS.

Moreover, the connection to harmonic retrieval or super-resolution problems is detailed in Section 3.5. Then numerical experiments and comparisons to state-of-the-art methods for structured low-rank matrix recovery are carried out in Section 3.6. In Section 3.7, we provide the proof of the locally quadratic convergence rate under appropriate assumptions on the coherence of such matrices before concluding

3.2 Formulation of StrucIRLS

We recall the framework of *Iteratively Reweighted Least Squares* algorithms whose main steps were already introduced in the beginning of Section 2.3: Given a family of functions $(F_\epsilon :$

$\mathbb{C}^{d_1\times d_2} \to \mathbb{R})_{\epsilon>0}$ that are *ϵ-smoothings* of the objective $F : \mathbb{C}^{d_1\times d_2} \to \mathbb{R}$ to be minimized in the first place, we define, for a given $\epsilon_k > 0$ and $\mathbf{X}^{(k)} \in \mathbb{C}^{d_1\times d_2}$, a *quadratic upper bound* $Q_{\epsilon_k}(\cdot|\mathbf{X}^{(k)})$, which is *tight*, but still a global majorizer of F_{ϵ_k}.

The global quadratic upper bound function $Q_{\epsilon_k}(\cdot|\mathbf{X}^{(k)})$ is then minimized under the data constraint, resulting in a weighted least squares problem, returning the update $\mathbf{X}^{\mathbf{(k+1)}} \in \mathbb{C}^{d_1\times d_2}$ of the iterate. As a last step of a given iteration, the smoothing parameter $\epsilon_k 0$ is updated to $\epsilon_{k+1} \leq \epsilon_k$, leading with $\mathbf{X}^{\mathbf{(k+1)}}$ to the definition of the updated quadratic bound function $Q_{\epsilon_{k+1}}(\cdot|\mathbf{X}^{(k+1)})$ of the next iteration.

Recalling the formulation of the low-rank Hankel matrix completion problem (3.7) we want to solve, i.e.,

$$\widehat{\mathbf{X}} = \mathscr{H}(\widehat{\mathbf{x}}) \text{ s.t. } \widehat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{C}^n, \mathscr{P}_\Omega(\mathbf{x})=\mathbf{y}}{\arg\min} \quad \operatorname{rank}(\mathscr{H}(\mathbf{x})),$$

where $\mathbf{y} \in \mathbb{C}^m$ is a vector of samples and $P_\Omega : \mathbb{C}^n \to \mathbb{C}^m$ a subsampling operator corresponding to a set $\Omega = \{\omega_1, \dots, \omega_m\} \subset \{0, \dots, n-1\}$, we see that the (intractable) *objective* to optimize is still the rank function rank$(\cdot)$, as in Chapter 2. However, the domain of the objective is now restricted to the set of Hankel matrices.

Motivated by the discussion of Section 2.2.1, where the log-det objective was identified as the one whose minimizers most closely resembles those of the rank objective, and by the fast predicted convergence rate due to Theorem 2.7, we define the objective $F : \mathbb{C}^{d_1\times d_2} \to \mathbb{R}$,

$$\mathbf{X} \to F(\mathbf{X}) = \sum_{i=1}^{d} f\left(\sigma_i(\mathbf{X})\right) := \sum_{i=1}^{d} \log\left(\sigma_i(\mathbf{X})\right)$$

recalling the convention $d = \min(d_1, d_2)$, and its ϵ-smoothing $F_\epsilon : \mathbb{C}^{d_1\times d_2} \to \mathbb{R}$,

$$\mathbf{X} \to F_\epsilon(\mathbf{X}) := \sum_{i=1}^{d} f_\epsilon\left(\sigma_i(\mathbf{X})\right) = \sum_{i=1}^{d} \left[\log\left(\max(\sigma_i(\mathbf{X}), \epsilon)\right) + \frac{1}{2}\left(\frac{\min(\sigma_i(\mathbf{X}), \epsilon)^2}{\epsilon^2} - 1\right)\right].$$

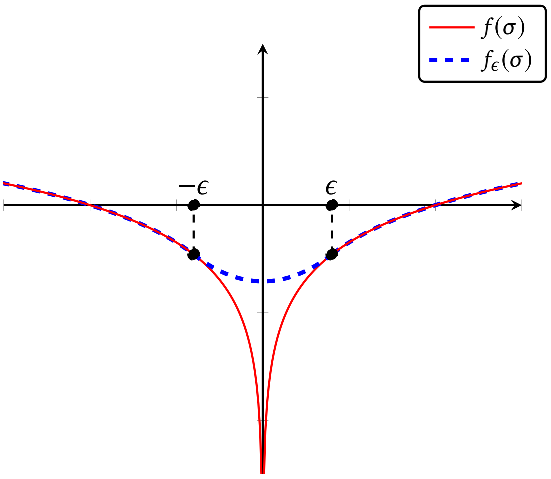

Figure 3.1 – Smooth approximation $f_\epsilon(\sigma)$ for $f(\sigma) = \log|\sigma|$.

For a visualization of the relationship between F and F_ϵ, we refer to Figure 3.1, which clarifies the geometry of their singular value-wise one-dimensional functions $f : \mathbb{R} \to \mathbb{R}$ and

$f_\epsilon : \mathbb{R} \to \mathbb{R}$ with $f(\sigma) = \log|\sigma|$ and

$$f_\epsilon(\sigma) = \begin{cases} \log|\sigma|, & \text{if } |\sigma| \geq \epsilon, \\ \log(\epsilon) + \frac{1}{2}\left(\frac{\sigma^2}{\epsilon^2} - 1\right), & \text{if } |\sigma| < \epsilon. \end{cases}$$

Now, we note that the Hankel structure of matrices induced by the Hankel operator $\mathscr{H} : \mathbb{C}^n \to \mathbb{C}^{d_1 \times d_2}$ implies, in particular, that all Hankel matrices $\mathbf{X} \in \mathcal{S}$ with

$$\mathcal{S} = \{\mathbf{X} \in \mathbb{C}^{d_1 \times d_2} : \exists \mathbf{x} \in \mathbb{C}^n \text{ such that } \mathscr{H}(\mathbf{x}) = \mathbf{X}\}$$

are *symmetric* in the sense that the symmetrization operator (definition without loss of generality for $d_2 \geq d_1$) $\widetilde{S} : \mathbb{C}^{d_1 \times d_2} \to \mathbb{C}^{d_1 \times d_2}$,

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 \end{bmatrix} \mapsto \widetilde{S}(\mathbf{Z}) = \begin{bmatrix} \frac{1}{2}(\mathbf{Z}_1 + \mathbf{Z}_1^*) & \mathbf{Z}_2 \end{bmatrix},$$

where $\mathbf{Z}_1 \in \mathbb{C}^{d_1 \times d_1}$, $\mathbf{Z}_2 \in \mathbb{R}^{d_1 \times (d_2 - d_1)}$, returns the Hankel matrix itself, i.e.,

$$\widetilde{S}(\mathbf{X}) = \mathbf{X} \tag{3.15}$$

for any $\mathbf{X} \in \mathcal{S}$.

Specifying the case $p = 0$, we can therefore leverage Definition 2.3.4 of the weight operator $W^{(k)} : \mathbb{C}^{d_1 \times d_2} \to \mathbb{C}^{d_1 \times d_2}$ of `MatrixIRLS` and its global majorization result, Theorem 2.4, to obtain, for any $\mathbf{x} \in \mathbb{C}^n$, that

$$\begin{aligned} \mathcal{J}_{\epsilon_k}(\mathbf{x}) &:= F_\epsilon(\mathscr{H}(\mathbf{x})) \leq Q_{\epsilon_k}\left(\mathscr{H}(\mathbf{x}) | \mathscr{H}(\mathbf{x}^{(k)})\right) \\ &= \mathcal{J}_{\epsilon_k}(\mathbf{x}^{(k)}) + \frac{1}{2}\left(\langle \mathscr{H}(\mathbf{x}), W^{(k)}(\mathscr{H}(\mathbf{x}))\rangle - \langle \mathscr{H}(\mathbf{x}^{(k)}), W^{(k)}(\mathscr{H}(\mathbf{x}^{(k)}))\rangle\right), \end{aligned} \tag{3.16}$$

for any Hankel matrix $\mathbf{X}^{(k)} = \mathscr{H}(\mathbf{x}^{(k)}) \in \mathbb{C}^{d_1 \times d_2}$ defined by the coordinates of a vector $\mathbf{x}^{(k)} \in \mathbb{C}^n$.

In (3.16), the weight operator $W^{(k)} : \mathbb{C}^{d_1 \times d_2} \to \mathbb{C}^{d_1 \times d_2}$ is defined such that

$$W^{(k)}(\mathbf{Z}) = \mathbf{U}_k \left[\mathbf{H}_1^{(k)} \circ (\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k)\right] \mathbf{V}_k^* \tag{3.17}$$

where $\mathbf{U}_k \in \mathbb{C}^{d_1 \times d_1}$, $\mathbf{V}_k \in \mathbb{C}^{d_2 \times d_2}$ are unitary matrices and $\sigma_i^{(k)} = \sigma_i(\mathscr{H}(\mathbf{x}^{(k)}))$ for $i \in [d]$, such that

$$\mathscr{H}(\mathbf{x}^{(k)}) = \mathbf{U}_k \operatorname{dg}(\sigma^{(k)}) \mathbf{V}_k^*$$

is a singular value decomposition of $\mathscr{H}(\mathbf{x}^{(k)})$, and $\mathbf{H}_1^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ is such that

$$(\mathbf{H}_1^{(k)})_{ij} = \frac{1}{\max(\sigma_i^{(k)}, \epsilon_k) \max(\sigma_j^{(k)}, \epsilon_k)} \tag{3.18}$$

for any $i \in [d_1]$, $j \in [d_2]$, cf. (2.58). We note that due to (3.15), only the matrix of weights $\mathbf{H}_1^{(k)}$ appears, but not $\mathbf{H}_2^{(k)}$ of (2.59).

In particular, the inequality of (3.16) is due to the following important fact: Since the global majorization statement of Theorem 2.4,

$$F_{\epsilon_k}(\mathbf{X}) \leq Q_{\epsilon_k}(\mathbf{X} | \mathscr{H}(\mathbf{x}^{(k)})),$$

holds for *any* $(d_1 \times d_2)$-matrix $\mathbf{X}$, it in particular also holds for any element $\mathcal{H}(\mathbf{x})$ of the subspace of Hankel matrices $\mathcal{S}$.

Incorporating the Subspace Constraint

As in the unstructured case studied in Chapter 2, we can now devise the first step of an IRLS strategy by minimizing the right hand side of (3.16) under the linear constraint described by the set

$$\{\mathbf{x} \in \mathbb{C}^n : P_\Omega(\mathbf{x}) = \mathbf{y}\}.$$

As the first and third summand of the right hand side of (3.16) are constant and do not depend on the unknown $\mathbf{x}$, we obtain the new iterate $\mathbf{x}^{(k+1)} \in \mathbb{C}^n$ as the solution of the *weighted least squares problem*

$$\mathbf{x}^{(k+1)} := \underset{P_\Omega(\mathbf{X})=\mathbf{y}}{\arg\min} \left\langle \mathcal{H}(\mathbf{x}), W^{(k)}(\mathcal{H}(\mathbf{x})) \right\rangle. \tag{3.19}$$

In fact, by writing down the optimization problem (3.19), we can already observe that *we have reduced the domain of our optimization from the $d_1 d_2$-dimensional matrix domain* to the *n-dimensional domain $\mathbb{C}^n$*.

Using Hankel matrices not explicitly, but the *subspace structure implicitly* through the Hankel operator $\mathcal{H} : \mathbb{C}^n \to \mathbb{C}^{d_1 \times d_2}$ and its *adjoint* $\mathcal{H}^* : \mathbb{C}^{d_1 \times d_2} \to \mathbb{C}^n$, we see that the objective function of (3.19) is indeed *quadratic* in $\mathbf{x} \in \mathbb{C}^n$: Since the adjoint $\mathcal{H}^*$ of the Hankel operator fulfills

$$\langle \mathcal{H}(\mathbf{x}), \mathbf{Z} \rangle = \langle \mathbf{x}, \mathcal{H}^*(\mathbf{Z}) \rangle$$

for any $\mathbf{Z} \in \mathbb{C}^{d_1 \times d_1}$ and $\mathbf{x} \in \mathbb{C}^n$, it follows that

$$\left\langle \mathcal{H}(\mathbf{x}), W^{(k)}(\mathcal{H}(\mathbf{x})) \right\rangle = \left\langle \mathbf{x}, \mathcal{H}^* W^{(k)} \mathcal{H} \mathbf{x} \right\rangle =: \langle \mathbf{x}, \widetilde{W}^{(k)} \mathbf{x} \rangle,$$

using the definition of the *structured weight operator*

$$\widetilde{W}^{(k)} := \mathcal{H}^* W^{(k)} \mathcal{H} \in \mathbb{C}^{n \times n}, \tag{3.20}$$

which is a positive definite $(n \times n)$ matrix.

This reasoning shows that the constraint of an Hankel subspace is not only compatible well with the quadratic nature of an IRLS approach, but is even able to reduce the problem size.

Based on this, we define Algorithm 2, our proposed *Structured Iteratively Reweighted Least Squares* algorithm (`StrucIRLS`) for the low-rank Hankel matrix completion problem.

We recall the following notation: We *best rank-r* approximation $\mathcal{T}_r(\mathbf{X})$ of a matrix $\mathbf{X} \in \mathbb{C}^{d_1 \times d_2}$ with singular value decomposition $\mathbf{X} = \begin{bmatrix} \mathbf{U}^{(k)} & \mathbf{U}_\perp^{(k)} \end{bmatrix} \begin{bmatrix} \operatorname{diag}(\sigma_i(\mathbf{X}))_{i=1}^r & 0 \\ 0 & \operatorname{dg}(\sigma_i(\mathbf{X}))_{r=+1}^d \end{bmatrix} \begin{bmatrix} \mathbf{V}^{(k)} \\ \mathbf{V}_\perp^{(k)*} \end{bmatrix}$, where $\mathbf{U}^{(k)} \in \mathbb{C}^{d_1 \times r}$ and $\mathbf{V}^{(k)} \in \mathbb{C}^{d_2 \times r}$ are matrices with orthonormal columns, is a matrix that fulfills

$$\mathcal{T}_r(\mathbf{X}) = \mathbf{U}^{(k)} \operatorname{diag}(\sigma_i(\mathbf{X}))_{i=1}^r \mathbf{V}^{(k)*} = \underset{\mathbf{Z}:\operatorname{rank}(\mathbf{Z}) \leq r}{\arg\min} \|\mathbf{Z} - \mathbf{X}\|, \tag{3.21}$$

where $\|\cdot\|$ can be any unitarily invariant norm.

Algorithm 2 `StrucIRLS` for structured low-rank matrix completion

```
Input: Set of observed coordinates $\Omega \subset \{0, 1, \dots n-1\}$ with $|\Omega| = m$, structure operator
       $\mathscr{H} : \mathbb{C}^n \to \mathbb{C}^{d_1 \times d_2}$, vector of samples $\mathbf{y} \in \mathbb{C}^m$, rank estimate $\widetilde{r}$.
Output: Sequences $(\mathbf{x}^{(k)})_{k \geq 1} \subset \mathbb{C}^n$ and $(\epsilon_k)_{k \geq 1}$.
Initialize $k = 0$, $\epsilon^{(0)} = \infty$ and $\widetilde{W}^{(0)} = \mathscr{H}^* \mathscr{H} : \mathbb{C}^n \to \mathbb{C}^n$.
for $k = 1, 2, \dots$ do
    1. Use a conjugate gradient method to solve linearly constrained quadratic program
           $\mathbf{x}^{(k)} = \underset{P_\Omega(\mathbf{x}) = \mathbf{y}}{\arg\min} \langle \mathbf{x}, \widetilde{W}^{(k-1)} \mathbf{x} \rangle.$    (3.22)
    2. Find best rank-$(\widetilde{r}+1)$ approximation of $\mathscr{H}(\mathbf{x}^{(k)})$ to obtain
           $\mathscr{T}_{\widetilde{r}}(\mathscr{H}(\mathbf{x}^{(k)})) = \mathbf{U}^{(k)} \operatorname{diag}(\sigma_i^{(k)})_{i=1}^{\widetilde{r}} \mathbf{V}^{(k)*}$    (3.23)
       and $\sigma_{\widetilde{r}+1}^{(k)}$, update smoothing:
           $\epsilon_k = \min\left(\epsilon_{k-1}, \sigma_{\widetilde{r}+1}(\mathscr{H}(\mathbf{x}^{(k)}))\right)$    (3.24)
    3. Update $\widetilde{W}^{(k)}$ as in (3.20) (see also (3.17) and (3.18)) , using the information $\mathbf{U}^{(k)}, \mathbf{V}^{(k)}$
       and $\sigma_1^{(k)}, \dots, \sigma_{\widetilde{r}+1}$ from item 2.
```

The algorithm `StrucIRLS` as described Algorithm 2 can be regarded as an algorithm for structured low-rank matrix completion for *any* kind of linear matrix structures induced by a linear operator $\mathscr{H} : \mathbb{C}^n \to \mathbb{C}^{d_1 \times d_2}$ with $n < d_1 d_2$. However, it is not true that the linear system (3.22) corresponding to step 1 can be solved efficiently for *any* operator $\mathscr{H}$, since the cost of performing a matrix-vector multiplication with the structured weight operator matrix

$$\widetilde{W}^{(k-1)} = \mathscr{H}^* W^{(k-1)} \mathscr{H} \in \mathbb{C}^{n \times n},$$

which, as we saw in Section 2.6, plays a crucial role in implementations of iterative solvers, depends on $\mathscr{H}$.

Furthermore, the cost of calculating the low-rank approximations (3.23) of $\mathscr{H}(\mathbf{x}^{(k)})$ will depend on $\mathscr{H}$ if methods like Lanczos bidiagonalization [SZ00, LXQ15], randomized power iteration [HMT11] or randomized Block Krylov [MM15] methods are used, not using structured matrices $\mathscr{H}(\mathbf{x}^{(k)})$ explicitly, but rather using its action on vectors.

We refer to Section 3.4 for a discussion of these aspects in the case of *Hankel* operators $\mathscr{H}$. In Section 3.3, we provide a convergence analysis for Algorithm 2 similar to the one obtained for `MatrixIRLS` in Section 2.4.

As in the last chapter, we note that the choice of the rule (3.24) for the update of the smoothing parameter ϵ plays an important role in the convergence analysis and also has major implications on the geometry of the optimization landscape.

Remark 3.2.1. *Similar to other IRLS algorithms,* `StrucIRLS` *can be interpreted within the* majorization-minimization *(MM) framework* [Lan16, SBP17]. *As described in* [SBP17], *in this framework, functions $Q(\cdot|\mathscr{H}(\mathbf{x}^{(k)}))$ with simple (not necessarily quadratic) structure are defined* *to fulfill properties (3.16) and*

$$\mathcal{J}(x^{(k)}) = Q\left(\mathcal{H}(\mathbf{x}^{(k)})|\mathcal{H}(\mathbf{x}^{(k)})\right)$$

with respect to a function $\mathcal{J}$, and then minimized to obtain an iterative optimization strategy to minimize $\mathcal{J}$.

An important additional ingredient of StrucIRLS *(and of* MatrixIRLS *for the unstructured case) which is usually not present in the MM framework, is the updated smoothing of (3.24), which changes the objective $\mathcal{J}$ itself across the iterations as well, and not only the surrogate* $Q(\cdot|\mathcal{H}(\mathbf{x}^{(k)}))$.

In this sense, the interplay between all the different steps are crucial for IRLS. As the survey [SBP17]*, which described MM procedures, states:* "On one hand, to achieve a fast convergence rate, a surrogate function that tries to follow the shape of the objective function is preferable. On the other hand, it should be simple to minimize so that the computational cost per iteration is low". *Despite this difference to MM procedures, IRLS can be interpreted as a method that tries to find a good compromise between these competing goals.*

3.3 Convergence Analysis of StrucIRLS

In this section, we provide a convergence analysis of StrucIRLS. In Section 3.3.1, we first state results about a basic interpretation of StrucIRLS in terms of log-det objective functions. These results hold for any subsampling set $\Omega \subset \{0, 1, \dots, n-1\}$ and any structure operator $\mathcal{H}$.

Subsequently, we provide a more refined local convergence analysis in Section 3.3.2 for the Hankel matrix completion case, assuming that a sufficient number of samples $m = |\Omega|$ are uniformly drawn at random.

3.3.1 Basic Properties and Minimization of Smoothed Objective

The statements of the following theorem guarantee that StrucIRLS is actually a valid strategy to minimize the auxiliary objectives $\mathcal{J}_\epsilon$, which are define in Theorem 3.1.

Theorem 3.1 (Monotonicity & Stationary Points Are Accumulation Points)**.** *Let $P_\Omega : \mathbb{C}^n \to \mathbb{C}^m$, $\mathbf{y} \in \mathbb{C}^m$ and $\mathcal{H} : \mathbb{C}^n \to \mathbb{C}^{d_1 \times d_2}$ be linear operators and $\widetilde{r} \in \mathbb{N}$. Let $(\mathbf{x}^{(k)}, \epsilon_k)_{k\geq 1}$ be a sequence of outputs of* StrucIRLS *corresponding to these parameters. Define, for any $\epsilon > 0$, the functions (smoothed) log-det objectives $\mathcal{J} : \mathbb{C}^n \to \mathbb{R}$,*

$$\mathbf{x} \to \mathcal{J}(\mathbf{x}) = \sum_{i=1}^{d} \log\left(\sigma_i(\mathcal{H}(\mathbf{x}))\right)$$

and $\mathcal{J}_\epsilon : \mathbb{C}^{d_1 \times d_2} \to \mathbb{R}$,

$$\mathbf{x} \to \mathcal{J}_\epsilon(\mathbf{x}) = \sum_{i=1}^{d} \left[\log\left(\max(\sigma_i(\mathcal{H}(\mathbf{x})), \epsilon)\right) + \frac{1}{2}\left(\frac{\min(\sigma_i(\mathcal{H}(\mathbf{x})), \epsilon)^2}{\epsilon^2} - 1\right)\right]. \tag{3.25}$$

Then the following holds.

1. *The sequence $\left(\mathcal{J}_{\epsilon_k}(\mathbf{x}^{(k)})\right)_{k\geq 1}$ is non-increasing, i.e., $\mathcal{J}_{\epsilon_k}(\mathbf{x}^{(k)}) \leq \mathcal{J}_{\epsilon_{k-1}}(\mathbf{x}^{(k-1)})$ for any $k \in \mathbb{N}$.*
2. *For any $k \in \mathbb{N}$,*

$$\mathcal{J}_{\epsilon_k}(\mathbf{x}^{(k)}) \geq \sum_{i=1}^{d} \log\left(\sigma_i(\mathcal{H}(\mathbf{x}^{(k)}))\right)$$

where $\mathscr{H}(\mathbf{x}^{(k)}) = \mathbf{U}_k \operatorname{dg}(\sigma_i(\mathscr{H}(\mathbf{x}^{(k)})))\mathbf{V}_k^*$ *is a singular value decomposition of* $\mathscr{H}(\mathbf{x}^{(k)})$.

3. *If* $\bar{\epsilon} := \lim_{k\to\infty} \epsilon_k > 0$, *then the iterates* $\mathbf{x}^{(k)}$ *and* $\mathbf{x}^{(k+1)}$ *fulfill* $\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}\|_2 = 0$.

4. *If* $\bar{\epsilon} := \lim_{k\to\infty} \epsilon_k > 0$, *then each subsequence of* $(\mathbf{x}^{(k)})_k$ *has a convergent subsequence, and each accumulation point* $\bar{\mathbf{x}}$ *of* $(\mathbf{x}^{(k)})_k$ *is a stationary point of* $\mathcal{J}_{\bar{\epsilon}}$ *subject to the linear constraint* $\{\mathbf{x} : P_\Omega(\mathbf{x}) = \mathbf{y}\}$.

Proof. This results follow directly from following the proof of Theorem 2.6, in particular using the observation (3.16). □

For a visualization of statement 2., we refer to Figure 3.1.

3.3.2 Local Convergence Theory: Quadratic Convergence Rate

In this section, we provide a convergence analysis for `StrucIRLS` that is related to the one provided in Section 2.4 for `MatrixIRLS`.

In particular, we provide a local convergence guarantee stating that `StrucIRLS` recovers a rank-r Hankel matrix $\mathscr{H}(\mathbf{x}_0)$ if we obtain an iterate $\mathbf{x}^{(k)}$ that is close enough to the ground truth $\mathbf{x}_0$ that parametrized the rank$-r$ matrix $\mathscr{H}(\mathbf{x}_0)$. This is the case, with high probability, already from a small number of noiseless samples $y = P_\Omega(\mathbf{x}_0)$ provided that the $\mathscr{H}(\mathbf{x}_0)$ has enough information spread across the singular values, a property known as *incoherence* which is analogue to Definition 3.3.1 in Chapter 2. Moreover, we show that the local convergence attains a locally *quadratic convergence rate*.

To define the incoherence notion that is adequate for our purposes, we recall the definition (2.99) of a *tangent space* T onto the manifold of rank-r matrices

$$\mathscr{M}_r = \{\mathbf{X} \in \mathbb{C}^{d_1 \times d_2} : \operatorname{rank}(\mathbf{X}) = r\}$$

at a matrix

$$\mathbf{Z} = \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix} \begin{bmatrix}\mathbf{\Sigma} & 0 \\ 0 & \mathbf{\Sigma}_\perp\end{bmatrix} \begin{bmatrix}\mathbf{V}^* \\ \mathbf{V}_\perp^*\end{bmatrix},$$

where $\mathbf{U} \in \mathbb{C}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{C}^{d_2 \times r}$ are matrices with orthonormal bases of the first r left resp. right singular vectors of $\mathbf{X}$ their columns, and $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ and $\mathbf{\Sigma}_\perp \in \mathbb{R}^{(d_1 - r)\times(d_2 - r)}$ being diagonal matrices with the (ordered) singular values of $\mathbf{X}$ on their diagonals: In this case, we define $T = T_\mathbf{Z}$ such that

$$T_\mathbf{Z} := \{\mathbf{U}\mathbf{M}^* + \widetilde{\mathbf{M}}\mathbf{V}^* : \mathbf{M} \in \mathbb{C}^{d_2 \times r}, \ \widetilde{\mathbf{M}} \in \mathbb{C}^{d_1 \times r}\}. \tag{3.26}$$

Intuitively, we can say that a rank-r matrix $\mathbf{Z}$ has low incoherence if the projections of all elements of the *standard Hankel basis* onto the tangent space T associated to $\mathbf{Z}$ are *small*. In particular, we define the *standard Hankel basis* as follows: If $\mathscr{H} : \mathbb{C}^n \to \mathbb{C}^{d_1 \times d_2}$ is a Hankel operator and if $e_k \in \mathbb{C}^n$ is the k-th standard basis vector of $\mathbb{C}^n$, we define the matrices

$$\mathscr{H}_k := \frac{\mathscr{H}(e_k)}{\|\mathscr{H}(e_k)\|_F} \in \mathbb{C}^{d_1 \times d_2} \tag{3.27}$$

for $k = 1, 2, \ldots, n$. In fact, it can be easily seen that they attain the value $\min(k, n+1-k)^{-1/2}$ on the indices corresponding to the k-th antidiagonal and the value 0 for all other indices. We can observe that the set

$$\{\mathscr{H}_k : k \in [n]\}$$

is an orthonormal basis of the subspace $\mathcal{S}$ of $(d_1 \times d_2)$ Hankel matrices.

The following definition is a *weaker* version of the incoherence notion of [CC14], see also Definition 3.3.1 and the following Remark 2.5.1.

Definition 3.3.1. *We say that a rank-r Hankel matrix* $\mathcal{H}(\mathbf{x}) \in \mathbb{C}^{d_1 \times d_2}$ *with singular value decomposition* $\mathcal{H}(\mathbf{x}) = \mathbf{U\Sigma V}^*$ *and corresponding tangent space* $T = \{\mathbf{UM}^* + \widetilde{\mathbf{M}}\mathbf{V}^* : \mathbf{M} \in \mathbb{C}^{d_2 \times r}, \widetilde{\mathbf{M}} \in \mathbb{C}^{d_1 \times r}\}$ *is* μ_0-incoherent *if there exists a constant* $\mu_0 \geq 1$ *such that*

$$\max_{1 \leq k \leq n} \|\mathcal{P}_T(\mathcal{H}_k)\|_F \leq \sqrt{\mu_0 r \frac{d_1 + d_2}{d_1 d_2}} \tag{3.28}$$

where $\mathcal{H}_k$ *denotes the k-th element* (3.27) *of the orthonormal standard Hankel basis.*

Remark 3.3.1. *We note that the assumption that a tangent space* T *is* μ_0*-incoherent according to Definition 3.3.1* is weaker *than similar assumptions* [CC14, Definition 1] *,*[YKJL17, (A.18)], [CWW19, Definition 1; CWW18]. *In particular, assuming that*

$$\max_{1 \leq k \leq n} \|\mathbf{U}^*\mathcal{H}_k\|_F \leq \sqrt{\frac{\mu_0 r}{d_1}} \quad \text{and} \quad \max_{1 \leq k \leq n} \|\mathbf{V}^*\mathcal{H}_k^*\|_F \leq \sqrt{\frac{\mu_0 r}{d_2}}$$

as in [YKJL17, (A.18)]*, we obtain*

$$\begin{aligned}
\|\mathcal{P}_T(\mathcal{H}_k)\|_F^2 &= \|\mathbf{UU}^*\mathcal{H}_k + \mathcal{H}_k\mathbf{VV}^* - \mathbf{UU}^*\mathcal{H}_k\mathbf{VV}^*\|_F^2 = \|\mathbf{UU}^*\mathcal{H}_k(\mathbf{I} - \mathbf{VV}^*) + \mathcal{H}_k\mathbf{VV}^*\|_F^2 \\
&= \|\mathbf{UU}^*\mathcal{H}_k(\mathbf{I} - \mathbf{VV}^*)\|_F^2 + \|\mathcal{H}_k\mathbf{VV}^*\|_F^2 \leq \|\mathbf{UU}^*\mathcal{H}_k\|_F^2\|\mathbf{I} - \mathbf{VV}^*\|^2 + \|\mathcal{H}_k\mathbf{VV}^*\|_F^2 \\
&\leq \|\mathbf{U}^*\mathcal{H}_k\|_F^2 + \|\mathcal{H}_k\mathbf{V}\|_F^2 \leq \frac{\mu_0 r}{d_1} + \frac{\mu_0 r}{d_2} = \frac{\mu_0 r(d_1 + d_2)}{d_1 d_2}.
\end{aligned}$$

We recall that in our completion setting stated in Section 3.1.2, it was assumed that we are given an index set

$$\Omega = \{\omega_1, \ldots, \omega_m\} \subset \{0, \ldots, n-1\}$$

of cardinality m to define the subsampling operator $P_\Omega : \mathbb{C}^n \to \mathbb{C}^m, \mathbf{x} \mapsto P_\Omega(\mathbf{x}) = \sum_{\omega_j \in \Omega} e_{\omega_j}^* \mathbf{x}$. To prevent arbitrarily adversary sampling sets Ω which would not convey the structure of $\mathbf{x}$ well, we resort to randomness in the distribution of the indices $\omega_1, \ldots, \omega_m$ in order to obtain a convergence theorem. In particular, we will assume that we are given m indices $\omega_1, \ldots, \omega_m$ that are sampled uniformly at random among the n possible ones $\{0, \ldots, n_1\}$ (without replacement).

We then obtain a local convergence statement implying fast local convergence given enough randomly distributed samples, formalized in Theorem 3.2.

Theorem 3.2 (Local Convergence with Quadratic Rate)**.** *Let* $\mathcal{H} : \mathbb{C}^n \to \mathbb{C}^{d_1 \times d_2}$ *be the Hankel operator associated to the parameters* (n, d_1)*. Let* $\mathbf{x}_0 \in \mathbb{C}^n$ *be vector such that* $\mathcal{H}(\mathbf{x}_0) \in \mathbb{C}^{d_1 \times d_2}$ *is a rank-r matrix that is* μ_0*-incoherent in the sense of Definition 3.3.1. Let* $\Omega \subset \{1, \ldots, n\}$ *be a collection of m indices* Ω *sampled uniformly at random among* $\{0, 1, \ldots, n-1\}$*, let* $c_s = \max\left(\frac{n}{d_1}, \frac{n}{d_2}\right)$*. There exist constants* $C \geq 1$*,* C_1 *and* C_2 *such that if*

$$m \geq C\mu_0 c_s r \log(n), \tag{3.29}$$

with probability at least $1 - 2n^{-2}$*, the following holds: If* $\epsilon_k = \sigma_{r+1}(\mathcal{H}(\mathbf{x}^{(k)}))$ *and if the output vector* $\mathbf{x}^{(k)} \in \mathbb{C}^n$ *of the k-th iteration of* `StrucIRLS` *with inputs* Ω*,* $\mathcal{H}$*,* $\mathbf{y} = P_\Omega(\mathbf{x}_0)$ *and* $\widetilde{r} = r$ *fulfills*

$$\|\mathcal{H}(\mathbf{x}^{(k)}) - \mathcal{H}(\mathbf{x}_0)\|_{S_\infty} \leq \min\left(C_1 \frac{\sqrt{\mu_0 c_s r}}{\sqrt{n \log(n)}}, C_2 \frac{1}{n^2 r \kappa}\right) \sigma_r(\mathcal{H}(\mathbf{x}_0)),$$

then

$$\mathbf{x}^{(k)} \xrightarrow{k\to\infty} \mathbf{x}_0$$

and furthermore, the local convergence rate is quadratic *in the sense that there is a constant μ such that,*

$$\left\|\mathcal{H}(\mathbf{x}^{(k+1)}) - \mathcal{H}(x_0)\right\|_{S_\infty} \leq \mu\left\|\mathcal{H}(\mathbf{x}^{(k)}) - \mathcal{H}(\mathbf{x}_0)\right\|_{S_\infty}^2,$$

which can be chosen such that

$$\mu = \frac{24(9n^2+1)r\kappa}{\sigma_r(\mathcal{H}(\mathbf{x}_0))},$$

where $\kappa = \frac{\sigma_1(\mathcal{H}(\mathbf{x}_0))}{\sigma_r(\mathcal{H}(\mathbf{x}_0))}$ is the condition number of $\mathcal{H}(\mathbf{x}_0)$.

It is instructive to compare the sufficient condition for the sample complexity (3.29) with corresponding conditions for local convergence results for `PGD` [CWW18] and `FIHT` [CWW19], which require an order of $\Omega(\mu_0^2\kappa^2 r^2 \log(n))$ and $\Omega(\mu_0 r \log(n))$ samples respectively. We note that a *global convergence* theory for `HNNM` [CC14], a convex optimization procedure, is available if $\Omega(\mu_0 r \log^4(n))$ samples are provided.

Furthermore, while being a sufficient condition, we note that (3.29) can be seen as an indication to choose the window length parameter $d_1 \approx \frac{n}{2}$ to make the Hankel matrices as square as possible, as in this case, the parameter

$$c_s = \max\left(\frac{n}{d_1}, \frac{n}{d_2}\right)$$

in the bound is minimized. However, the same conclusions can be already drawn from the bounds in [CC14, CWW18, CWW19].

3.4 Computational Considerations

While the presentation of `StrucIRLS` and its convergence analysis is somewhat similar to the algorithm for unstructured recovery discussed in Chapter 2, it is not possible to translate the implementation of Section 2.6 in a one-by-one manner.

As in Section 2.6, we are interested in using an algorithm that runs with a time *and* space complexity of

$$o(d_1 d_2).$$

In most applications, due to statistical reasons, we would choose the window length parameter d_1 such that the structured matrices are approximately square. In this case, this means that we are interested in a space and time complexity of

$$o(n^2)$$

for `StrucIRLS`.

In this section, we restrict ourselves to the case of *Hankel*-structured matrices defined by a Hankel operator $\mathcal{H} : \mathbb{C}^{d_1\times d_2} \to \mathbb{C}^n$ such as in (3.3) with $d_1 = \Theta(n)$ and $d_2 = \Theta(n)$.

This is due to the fact that Hankel matrices possess a *convolutional structure*, i.e., they can be seen as a restricted, rearranged circulant matrices. These circulant matrices can be diagonalized by a discrete Fourier transform, which means that, by using the Fast Fourier Transform (FFT), matrix-vector multiplications can be performed in $O(n\log(n))$ instead of $\Theta(n^2)$, as it would be the case for unstructured matrices [KS99, Pan01, BQW09, Kor10, PT15].

To review how to perform these fast matrix-vector multiplication with Hankel matrices, let $\mathbf{x} \in \mathbb{C}^n$ and $\mathcal{H}(\mathbf{x}) \in \mathbb{C}^{d_1\times d_2}$. $\mathcal{H}(x^{(k)})$ now can be embedded into a so-called *circulant matrix*

$\mathscr{C}(\mathbf{x})$, defined by the *circulant operator* $\mathscr{C} : \mathbb{C}^n \to \mathbb{C}^{n\times n}$ that maps a vector $\mathbf{x} \in \mathbb{C}^n$ to

$$\mathscr{C}(\mathbf{x}) = \left(\mathscr{C}(\mathbf{x})_{ij}\right)_{i=1,j=1}^{n,n} = \left((\mathbf{x}_{(i-j) \mod n}\right)_{i=1,j=1}^{n,n} \in \mathbb{C}^{n\times n}, \tag{3.30}$$

which then can be diagonalized this by a discrete Fourier transform. As a preparation of what follows below, we recall formulas for the actions of the circulant operator $\mathscr{C}$ and its adjoint operator $\mathscr{C}^* : \mathbb{C}^{n\times n} \mapsto \mathbb{C}^n$ of $\mathscr{C}$.

Lemma 3.4.1 (Cf. [Pan01]). *Let $\mathscr{C} : \mathbb{C}^n \to \mathbb{C}^{n\times n}$ be the circulant operator of (3.30) and $\mathscr{C}^* : \mathbb{C}^{n\times n} \mapsto \mathbb{C}^n$ its adjoint. Then the action of $\mathscr{C}$ on a vector $\mathbf{x} \in \mathbb{C}^n$ can be written such that*

$$\mathscr{C}(x) = \sqrt{n}\mathbf{F}\operatorname{diag}(\mathbf{F}\mathbf{x})\mathbf{F}^*, \tag{3.31}$$

and the action of $\mathscr{C}^*$ on a matrix $\mathbf{Y} \in \mathbb{C}^{n\times n}$ such that

$$\mathscr{C}^*(\mathbf{Y}) = \sqrt{n}\mathbf{F}^*\operatorname{diag}(\mathbf{F}^*\mathbf{Y}^*\mathbf{F}),$$

where $\mathbf{F} \in \mathbb{C}^{n\times n}$ is the unitary discrete Fourier transform (DFT) *matrix and* diag $: \mathbb{C}^{n\times n} \to \mathbb{C}^n$ *such that* $\operatorname{diag}(\mathbf{M}) = (\mathbf{M}_{ii})_{i=1}^n$ *for a matrix $\mathbf{M} \in \mathbb{C}^{n\times n}$.*

Proof. The formula (3.31) follows, e.g., from [Pan01, Theorem 2.6.4].

Let $\mathbf{Y} \in \mathbb{C}^{n\times n}$ and $\mathbf{x} \in \mathbb{C}^n$. Using that

$$\mathscr{C}(\mathbf{x}) = \sqrt{n}\mathbf{F}\operatorname{diag}(\mathbf{F}\mathbf{x})\mathbf{F}^*,$$

$$\begin{aligned}\langle \mathbf{Y}, \mathscr{C}(\mathbf{x})\rangle &= \sqrt{n}\langle \mathbf{Y}, \mathbf{F}\operatorname{diag}(\mathbf{F}\mathbf{x})\mathbf{F}^*\rangle \\ &= \langle \mathbf{F}^*\mathbf{Y}\mathbf{F}, \operatorname{diag}(\mathbf{F}\mathbf{x})\rangle = \sqrt{n}\operatorname{tr}(\mathbf{F}^*\mathbf{Y}^*\mathbf{F}\operatorname{diag}(\mathbf{F}\mathbf{x})) = \sqrt{n}\sum_{i=1}^n [\mathbf{F}^*\mathbf{Y}^*\mathbf{F}]_{ii}[\operatorname{diag}(\mathbf{F}\mathbf{x})]_i \\ &= \sqrt{n}\langle \operatorname{diag}(\mathbf{F}^*\mathbf{Y}^*\mathbf{F}), \mathbf{F}\mathbf{x}\rangle = \sqrt{n}\langle \mathbf{F}^*\operatorname{diag}(\mathbf{F}^*\mathbf{Y}^*\mathbf{F}), \mathbf{x}\rangle.\end{aligned}$$

□

As described in Section 2.6.2, fast matrix-vector multiplications are very helpful in the second step of an IRLS iteration, where information for the definition of the weight operator updated is calculated by obtaining a best rank-$r+1$ approximation. In `StrucIRLS`, this corresponds to (3.23), and by saving the interates $\mathbf{x}^{(k)} \in \mathbb{C}^n$ in memory, but not $\mathscr{H}(\mathbf{x}^{(k)}) \in \mathbb{C}^{d_1\times d_2}$, we can make used of the above considerations to obtain a sub-quadratic time complexity.

We will not detail here how to solve the *weighted least squares problem* (3.22) in the most efficient way, However, in a implementation that uses *conjugate gradients* [HS52] with an appropriate stopping criterion, to obtain a fast implementation, it is necessary that matrix-vector multiplications with the (positive definite) *structured weight operator matrices* $\widetilde{W}^{(k)}$ from (3.20), i.e., with

$$\widetilde{W}^{(k)} = \mathscr{H}^* W^{(k)}\mathscr{H} \in \mathbb{C}^{n\times n}$$

where $W^{(k)} : \mathbb{C}^{d_1\times d_2} \to \mathbb{C}^{d_1\times d_2}$ is as in (3.17), are computationally affordable.

To make sure that this is the case, we provide in the following theorem a time complexity bound on the matrix-vector multiplication with $\widetilde{W}^{(k)}$.

Theorem 3.3. *Let $\mathbf{x}^{(k)} \in \mathbb{C}^n$ and $\epsilon_k > 0$ such that $\sigma_{r+1}(\mathscr{H}(\mathbf{x}^{(k)})) \leq \epsilon_k$, let $\mathscr{H} : \mathbb{C}^n \to \mathbb{C}^{d_1\times d_2}$ be the Hankel operator. Then the multiplication of $\widetilde{W}^{(k)} \in \mathbb{C}^{n\times n}$ with a vector $\mathbf{x} \in \mathbb{C}^n$ can be computed in $O(nr^2 + nr\log n)$ operations.*

Proof. Let $\mathbf{x} \in \mathbb{C}^n$ be arbitrary. Recalling the equation (3.20), we have $\widetilde{W}^{(k)} = \mathscr{H}^* W^{(k)} \mathscr{H}$. First, we note that the Hankel matrix $\mathscr{H}(\mathbf{x}) = [\mathscr{H}(\mathbf{x})]_{i,j=1}^{d_1,d_2} = (\mathbf{x}_{i+j-1})_{i,j=1}^{d_1,d_2}$ can be embedded into a circulant matrix $[\mathscr{C}(\mathbf{x})]_{i,j=1}^{n,n} = \mathbf{x}_{((i-j) \mod n)+1}$ with the vector $\mathbf{x} \in \mathbb{C}^n$ as its first row such that

$$\mathscr{H}(\mathbf{x}) = P_{d_1} \mathbb{I}_{\text{anti}} \mathscr{C}(\mathbf{x}) J P_{d_2}^* = \mathscr{P}_{d_1} \mathscr{C}(\mathbf{x}) \mathscr{P}_{d_2}^*, \tag{3.32}$$

where $\mathbf{J}$ is the permutation matrix such that $[\mathbf{J}]_{ij} = 1$ if $j = i + 1$ or $i = 1$ and $j = n$ and zero elsewhere and $\mathbb{I}_{\text{anti}}$ the matrix with entries equal to 1 only on the main antidiagonal. Furthermore, $P_{d_1} : \mathbb{C}^n \longrightarrow \mathbb{C}^{d_1}, \mathbf{x} \mapsto (\mathbf{x}_1, \ldots, \mathbf{x}_{d_1})^T$ and $P_{d_2} : \mathbb{C}^n \longrightarrow \mathbb{C}^{d_2}, \mathbf{x} \mapsto (\mathbf{x}_1, \ldots, \mathbf{x}_{d_2})^T$ denote the projection into the first d_1 and d_2 coordinates, respectively, and $\mathscr{P}_{d_1} \in \mathbb{C}^{d_1 \times n}$ and $\mathscr{P}_{d_2} \in \mathbb{C}^{d_2 \times n}$ are given by

$$\mathscr{P}_{d_1} = P_{d_1} \mathbb{I}_{\text{anti}} \text{ and } \mathscr{P}_{d_2} = P_{d_2} \mathbf{J}^*.$$

By Lemma 3.4.1, it follows that that the circulant matrix $\mathscr{C}(\mathbf{x})$ can be diagonalized by the (unitary) discrete Fourier transform $\mathbf{F} = \left(\frac{1}{\sqrt{n}} \mathrm{e}^{-2\pi \mathrm{i}(j-1)(k-1)/n}\right)_{k,j=1}^{n} \in \mathbb{C}^{n \times n}$, such that

$$\mathscr{C}(\mathbf{x}) = \mathbf{F} \mathbf{D}_{\sqrt{n}\mathbf{F}\mathbf{x}} \mathbf{F}^*,$$

where $\mathbf{D}_{\sqrt{n}\mathbf{F}\mathbf{x}} = \operatorname{diag}(\sqrt{n}\mathbf{F}\mathbf{x}) \in \mathbb{C}^{n \times n}$ is as in Lemma 3.4.1. Combined with (3.32), we have

$$\mathscr{H}(\mathbf{x}) = P_{d_1} \mathbb{I}_{\text{anti}} \mathscr{C}(\mathbf{x}) \mathbf{J} P_{d_2}^* = \mathscr{P}_{d_1} \mathbf{F} D_{\sqrt{n}\mathbf{F}\mathbf{x}} \mathbf{F}^* \mathscr{P}_{d_2}^*.$$

Next, we use our assumption that there at most r singular values of $\mathscr{H}(x^{(k)})$ larger than ϵ_k to obtain a simplified version of (2.155) from Lemma 2.6.1: Thus, we can write the action of the weight operator $W^{(k)} : \mathbb{C}^{d_1 \times d_2} \longrightarrow \mathbb{C}^{d_1 \times d_2}$ on a matrix $\mathbf{Z} \in \mathbb{C}^{d_1 \times d_2}$ such that

$$\begin{aligned} W^{(k)}(\mathbf{Z}) = {} & \mathbf{U}[\mathbf{H}_{\mathbf{U},\mathbf{V}} \circ (\mathbf{U}^* \mathbf{Z} \mathbf{V})] \mathbf{V}^* + \\ & + \mathbf{U}\,(\mathbf{D})\,\mathbf{U}^* \mathbf{Z}(\mathbf{I} - \mathbf{V}\mathbf{V}^*) \\ & + (\mathbf{I} - \mathbf{U}\mathbf{U}^*) \mathbf{Z} \mathbf{V}\,(\mathbf{D})\,\mathbf{V}^* \\ & + \epsilon_k^{(-2)}\,(\mathbf{I} - \mathbf{U}\mathbf{U}^*)\,\mathbf{Z}\,(\mathbf{I} - \mathbf{V}\mathbf{V}^*)\,, \end{aligned}$$

where $\mathbf{H}_{\mathbf{U},\mathbf{V}} \in \mathbb{R}^{r \times r}$ such that $(\mathbf{H}_{\mathbf{U},\mathbf{V}})_{ij} = \frac{1}{\sigma_i^{(k)} \sigma_j^{(k)}}$ for $i, j \in [r]$, $\mathbf{U} = \mathbf{U}^{(k)}$, $\mathbf{V} = \mathbf{V}^{(k)}$ and $\mathbf{D} \in \mathbb{R}^{r \times r}$ diagonal such that $(\mathbf{D})_{ii} = \frac{1}{\sigma_i^{(k)} \epsilon_k}$ for $i \in [r]$.

Then, using the linearity of $\mathscr{H}^*$, we rearranged the matrix vector product $\widetilde{W}^{(k)} \mathbf{x}$ such that

$$\begin{aligned} \widetilde{W}^{(k)} \mathbf{x} = {} & \mathscr{H}^* W^{(k)}(\mathscr{H}(\mathbf{x})) = \mathscr{H}^* \mathbf{U}[H_{\mathbf{U},\mathbf{V}} \circ (\mathbf{U}^* \mathscr{H}(\mathbf{x}) \mathbf{V})] \mathbf{V}^* + \mathscr{H}^* \mathbf{U}\mathbf{D}(\mathbf{U}^* \mathscr{H}(\mathbf{x}) \mathbf{V}_\perp) \mathbf{V}_\perp^* \\ & + \mathscr{H}^* \mathbf{U}_\perp \mathbf{U}_\perp^* \mathscr{H}(\mathbf{x}) \mathbf{V}) \mathbf{D} \mathbf{V}^* + \epsilon_k^{-2} \mathscr{H}^* \mathbf{U}_\perp \mathbf{U}_\perp^* \mathscr{H}(\mathbf{x}) \mathbf{V}_\perp \mathbf{V}_\perp^* \\ & = \mathscr{H}^* \mathbf{U}[H_{\mathbf{U},\mathbf{V}} \circ (\mathbf{U}^* \mathscr{H}(\mathbf{x}) \mathbf{V})] \mathbf{V}^* + \mathscr{H}^* \mathbf{U}\mathbf{D}\mathbf{U}^* \mathscr{H}(\mathbf{x})(\mathbf{I} - \mathbf{V}\mathbf{V}^*) \\ & + \mathscr{H}^* (\mathbf{I} - \mathbf{U}\mathbf{U}^*) \mathscr{H}(\mathbf{x}) \mathbf{V}\mathbf{D}\mathbf{V}^* + \epsilon_k^{-2} \mathscr{H}^* (\mathbf{I} - \mathbf{U}\mathbf{U}^*) \mathscr{H}(\mathbf{x})(\mathbf{I} - \mathbf{V}\mathbf{V}^*) \\ & = \epsilon_k^{-2} \mathscr{H}^* \mathscr{H}(\mathbf{x}) + \mathscr{H}^* \left[[\mathbf{U}(\mathbf{D} - \epsilon_k^{-2} \mathbf{I}) \mathbf{U}^*] \mathscr{H}(\mathbf{x})\right] + \mathscr{H}^* \left[\mathscr{H}(\mathbf{x})[\mathbf{V}(\mathbf{D} - \epsilon_k^{-2} \mathbf{I}) \mathbf{V}^*]\right] \\ & + \mathscr{H}^* \left[\mathbf{U}(H_{\mathbf{U},\mathbf{V}} \circ (\mathbf{U}^* \mathscr{H}(\mathbf{x}) \mathbf{V}) - \mathbf{U}^* \mathscr{H}(\mathbf{x}) \mathbf{V}\mathbf{D} - \mathbf{D}\mathbf{U}^* \mathscr{H}(\mathbf{x}) \mathbf{V} + \epsilon_k^{-2} \mathbf{U}^* \mathscr{H}(\mathbf{x}) \mathbf{V}) \mathbf{V}^*\right] \end{aligned} \tag{3.33}$$

The four terms can be separately analyzed. Before continuing, we note that the action of the adjoint of the Hankel operator is given by $\mathscr{H}^*(Y) = \mathscr{C}^*(\mathscr{P}_{d_1}^* Y \mathscr{P}_{d_2}) = \sqrt{n} \mathbf{F}^* \operatorname{diag}(\mathbf{F}^* \mathscr{P}_{d_2}^* Y^* \mathscr{P}_{d_1} \mathbf{F})$, since indeed,

$$\begin{aligned}\langle Y, \mathscr{H}(x)\rangle &= \langle Y, \mathscr{P}_{d_1}\mathscr{C}(x)\mathscr{P}_{d_2}^*\rangle = \mathrm{tr}(Y^*\mathscr{P}_{d_1}\mathscr{C}(x)\mathscr{P}_{d_2}^*) = \mathrm{tr}(\mathscr{P}_{d_2}^* Y^*\mathscr{P}_{d_1}\mathscr{C}(x)) \\ &= \langle \mathscr{P}_{d_1}^* Y \mathscr{P}_{d_2}\ \mathscr{C}(x)\rangle = \langle \mathscr{C}^*\mathscr{P}_{d_1}^* Y \mathscr{P}_{d_2}, x\rangle = \langle \mathscr{H}^*(Y), x\rangle.\end{aligned} \tag{3.34}$$

Then, inserting the result from Lemma 3.4.1, leads us to the result. Defining $\bar{\mathbf{D}} = \mathbf{D} - \epsilon_k^{-2}\mathbf{I}$ and plugging (3.32) into the previous equations implies the following four observations:

1. Since $\mathscr{H}^*\mathscr{H} : \mathbb{C}^n \longrightarrow \mathbb{C}^n$ is a diagonal operator multiplying the i-th element with a fixed number ν_i, we obtain that $\epsilon^{-2}\mathscr{H}^*\mathscr{H}(\mathbf{x}) = \{\epsilon^{-2}\nu_i \mathbf{x}_i\}_{i=0}^{n-1}$ can be performed in $O(n)$ operations.

2. $\mathscr{H}^* \left[\mathbf{U}\bar{\mathbf{D}}\mathbf{U}^*\mathscr{H}(\mathbf{x})\right] = \sqrt{n}\mathbf{F}^* \,\mathrm{diag}(\mathbf{F}^*\mathscr{P}_{d_2}^*(\mathbf{U}\bar{\mathbf{D}}\mathbf{U}^*\mathscr{H}(x))^*\mathscr{P}_{d_1}\mathbf{F})$
   $= \sqrt{n}\mathbf{F}^* \,\mathrm{diag}\,(\mathscr{P}_{d_2}^*\mathscr{P}_{d_2}\mathbf{F}\mathrm{diag}\,(\sqrt{n}\mathbf{F}\mathbf{x})\mathbf{F}^*) = n\mathbf{F}^*\,\mathrm{diag}\,(\mathbf{F}^*\mathscr{P}_{d_2}^*\mathscr{P}_{d_2}\mathbf{F}\overline{\mathrm{diag}\,(\mathbf{F}\mathbf{x})}\mathbf{F}^*\mathscr{P}_{d_1}^*\mathbf{U}\bar{\mathbf{D}}\mathbf{U}^*\mathscr{P}_{d_1}\mathbf{F})$.
   The matrix $M_1 = \mathbf{F}^*\mathscr{P}_{d_2}^*\mathscr{P}_{d_2}\mathbf{F}\overline{\mathrm{diag}\,(\mathbf{F}x)}\mathbf{F}^*\mathscr{P}_{d_1}^*\mathbf{U} \in \mathbb{C}^{n\times r}$ can be computed in $O(Rn\log n)$ by using the FFT r times while, by counting the dimensions, the precomputed matrix $M_2 = \bar{\mathbf{D}}\mathbf{U}^*\mathscr{P}_{d_1}\mathbf{F} \in \mathbb{C}^{r\times n}$ can be performed in $O(r^2 n)$.

3. $\mathscr{H}^* \left[\mathscr{H}(x)\mathbf{V}\bar{\mathbf{D}}\mathbf{V}^*\right]$ is analogous to the previous case and can also be calculated in $O(r^2 n + Rn\log n)$ operations.

4. $\mathscr{H}^* \left[\mathbf{U}(H \circ (\mathbf{U}^*\mathscr{H}(x)\mathbf{V}) - \mathbf{U}^*\mathscr{H}(x)\mathbf{V}\bar{\mathbf{D}} - \bar{\mathbf{D}}\mathbf{U}^*\mathscr{H}(x)\mathbf{V} + \epsilon^{-2}\mathbf{U}^*\mathscr{H}(x)\mathbf{V})\mathbf{V}^*\right]$ is of the form $\mathscr{H}^* \left[\mathbf{U}(H \circ A - A\bar{\mathbf{D}} - \bar{\mathbf{D}}A - kA)\right]$ and the Schur product can be simplified leading to $n\mathbf{F}^*\,\mathrm{diag}\left(\mathbf{F}^*\mathscr{P}_{d_2}^*\mathbf{V}(H_{\mathrm{new}}^* \circ [\mathbf{V}^*\mathscr{P}_{d_2}\mathbf{F}\overline{\mathrm{diag}\,(\mathbf{F}x)}\mathbf{F}^*\mathscr{P}_{d_1}^*\mathbf{U}])\mathbf{U}^*\mathscr{P}_{d_1}\mathbf{F}\right)$, where $[H_{\mathrm{new}}]_{ij} = [H]_{ij} + \epsilon^{-2} - \bar{\mathbf{D}}_{ii} - \bar{\mathbf{D}}_{jj}$.
   Again, by counting dimensions, $M_3 = (H_{\mathrm{new}}^* \circ [\mathbf{V}^*\mathscr{P}_{d_2}\mathbf{F}\overline{\mathrm{diag}\,(\mathbf{F}x)}\mathbf{F}^*\mathscr{P}_{d_1}^*\mathbf{U}])\mathbf{U}^*\mathscr{P}_{d_1}\mathbf{F}$ can be performed in $O(r^2 n)$ operations, since $H_{\mathrm{new}} \in \mathbb{C}^{r\times r}$ and $\mathbf{U}^*\mathscr{P}_{d_1}\mathbf{F} \in \mathbb{C}^{r\times n}$. Then, the product of the precomputed matrix $M_4 = \mathbf{F}^*\mathscr{P}_{d_2}^*\mathbf{V} \in \mathbb{C}^{n\times r}$ by M_3 has also complexity $O(r^2 n)$. Lastly, the outer operation $\mathbf{F}^*\,\mathrm{diag}\,(M_3 M_4)$ needs $O(n\log n)$ by using FFT.

This shows an overall complexity of $O(nr\log n + r^2 n)$. □

Thus, we see that matrix-vector multiplication with $\widetilde{W}^{(k)}$ is, up to a logarithmic factor in n, almost linear in the dimension of $\mathbf{x} \in \mathbb{C}^n$ if $r^2 \ll n$.

3.5 Connection to Harmonic Retrieval

As mentioned in the Section 3.1.1, in many applications the parsimonious signals one encounters is specified by parameters in a continuous domain. They are often modelled as a superposition of point sources in the spectral or the time domain and the goal is to retrieve these highly localized patterns through the acquisition of a small number of samples, which means that a sum of exponentials model such as 3.4 is available. If in the application the samples are not consecutive, as in 3.5, but *irregular*, for example, corresponding to a *subsampled* equidistant grid, low-rank Hankel matrix completion techniques such as StrucIRLS can be used to complete the sequence to obtain an equidistant sample sequence.

In a second step, to retrieve the model parameters $\{\alpha_k\}_{k=1}^r$ and $\{f_k\}_{k=1}^r$ of (3.4), we could consequently use classical methods such as the original Prony's method [dP95], MUSIC [Sch86], ESPRIT [RK89], or the Matrix Pencil method [HS90]. This might be an interesting application, as these methods are not designed to deal with the case of non-consecutive samples.

Our methods can also be extended not to cover the case of *missing* samples, but of *noisy* samples such that

$$\widetilde{\mathbf{x}}_j = \mathbf{x}_j + \mathbf{w}_j$$

for $j \in \{0, 1, \ldots, n-1\}$ with $\mathbf{w}_j$, for example, independent Gaussian noise for all $j \in \{0, 1, \ldots, n-1\}$, cf. (3.6).

In this case, it is possible to adapt our ideas to solve first the Hankel structured low-rank approximation problem (3.8) with input $\mathbf{x} = \widetilde{\mathbf{x}}$, and *then* use a method for harmonic retrieval of choice such as Prony or ESPRIT or others (see Section 3.1.1) to retrieve the model parameters $\{\alpha_k\}_{k=1}^r$ and $\{f_k\}_{k=1}^r$ based on the *denoised* sample vector $\widehat{x} \in \mathbb{C}^n$.

In our framework, this would correspond to dropping the linear constraint $P_\Omega(\mathbf{x}) = \mathbf{y}$ and optimizing instead the unconstrained optimization problem

$$\min_{\mathbf{z} \in \mathbb{C}^n} \|\mathbf{z} - \mathbf{x}\|_2^2 + \lambda \mathcal{J}_\epsilon(\mathbf{z})$$

with $\mathcal{J}_\epsilon$ as in (3.25), for a suitable $\lambda > 0$. From the theory presented in this thesis, it is clear that this would involve solving the sequence of weighted least squares problems

$$\mathbf{x}^{(k)} = \arg\min_{\mathbf{z} \in \mathbb{C}^n} \left(\langle \mathbf{z} - \mathbf{x}, \mathbf{z} - \mathbf{x} \rangle + \lambda \left\langle \mathbf{z}, \widetilde{W}^{(k-1)} \mathbf{z} \right\rangle \right), \tag{3.35}$$

while updating the weight operator matrix $\widetilde{W}^{(k)}$ and the smoothing parameter ϵ_k as in StrucIRLS.

If the model order r of the sum of exponentials is known, it is possible to find good choices for the regularization parameter λ by a simple grid search.

3.6 Numerical Experiments

In this section, explore the performance of StrucIRLS empirically. We focus on the setting of completing low-rank Hankel matrices completion setting from random samples.

We conduct an experiment mimicking the setup of the experiments of [CWW18], considering the completion of Hankel matrices $\mathcal{H}(x)$ from m sample coordinates Ω that are drawn uniformly at random, of a vector $\mathbf{x} = ((\mathbf{x})_0, \ldots, (\mathbf{x})_{n-1})$ such that $(\mathbf{x})_j = x(j/n)$ for all $j \in \{0, \ldots, n-1\}$, $n = 127$, with sum-of-exponentials model $x : [0, 1] \to \mathbb{C}$,

$$x(t) = \sum_{k=1}^{r} \alpha_k e^{(2\pi i f_k)t},$$

where the f_k and c_k are sampled at random independently such that $f_k \sim \mathcal{U}([0, 1])$, $|\alpha_k| = 1 + 10^{c_k}$, $c_k \sim \mathcal{U}([0, 1])$. In Figure 3.2, the empirical recovery probabilities averaged across 50 simulations for each pair of m and r are documented in comparison with the atomic norm minimization algorithm (ANM) [BTSR13], projected gradient descent (PGD)[CWW18], fast iterative hard thresholding (FIHT) [CWW19], Hankel nuclear norm minimization (NNM) [CC14], and a non-convex optimization method operator on the frequency parameter space [FWS$^+$16].

In the experiment, we do not us the weight operator of StrucIRLS indicated by (3.17), but one that corresponds to an even tighter quadratic bound based on a harmonic mean weight

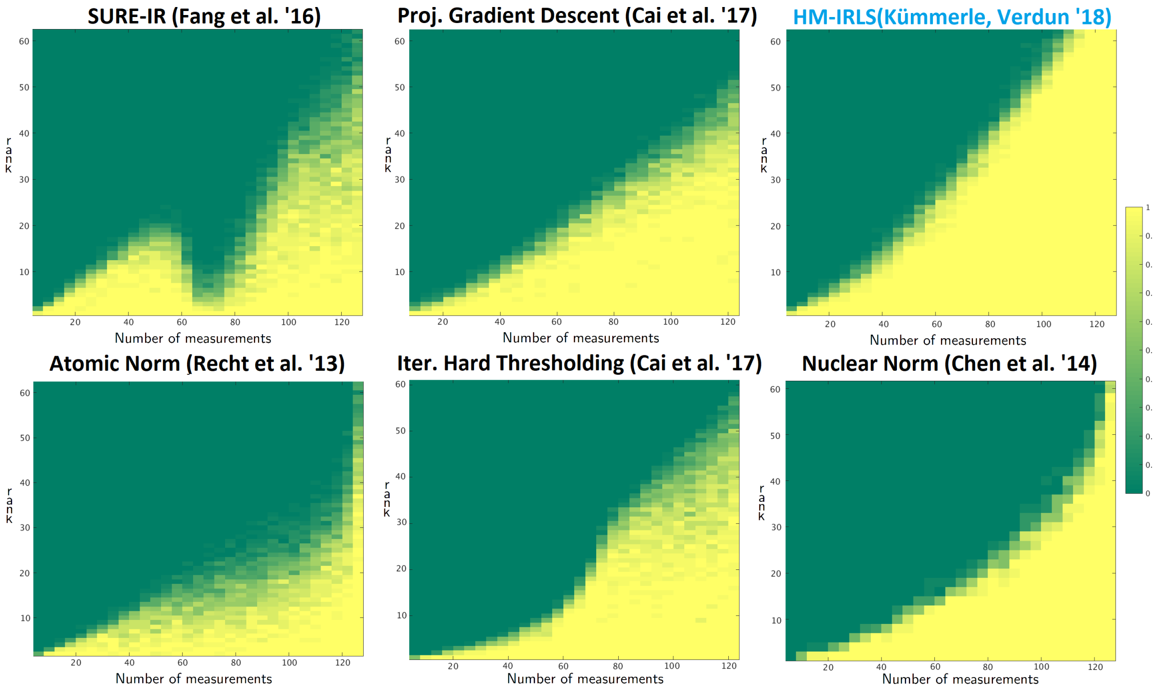

Figure 3.2 – Hankel matrix completion, m measurements of vector $\mathbf{x} \in \mathbb{C}^n$ with $n = 127$. x-axis: number of measurements m, y-axis: model order r.

operator. Furthermore, we use the objective

$$\mathcal{J}_\epsilon(\mathbf{x}) = \sum_{i=1}^{d} \log\left(\sigma_i^2(\mathcal{H}(\mathbf{x})) + \epsilon^2\right)$$

instead of the one in (3.25).

We observe our IRLS algorithm exhibits the best performance, with successful recovery already when $m \approx 2r$, despite the fact that some frequencies f_k will be very close if r is not too small, which is known to compromise the performance of methods such as ANM [BTR13]. The data corresponding to the same experiment is visualized also in Figure Figure 3.3.

Furthermore, we provide an experiment that investigates the potential of an IRLS approach for Hankel low-rank approximation (cf. Section 3.1.2), if used as a preprocessing for the harmonic retrieval problem as indicated in Section 3.5.

For this, we repeat an experiment of [HV18, Section VI]: Using equispaced samples from a signal that is a sum of two frequencies located at $f_1 = 0.35$ and $f_2 = 0.40$ (both with unitary amplitude) that are perturbed by additive Gaussian noise (with different *signal-to-noise ratios (SNR)*), we first use a variant of `StrucIRLS` for the Hankel low-rank approximation problem to obtained *denoised* samples. In the algorithm, we use an automated rule for the tuning parameter λ (see (3.35)).

After denoising, we use ESPRIT to obtain the frequencies, which is arguably one of best algorithms for frequency estimation for low noise levels [Fan16]. As a comparison, we use the algorithms [Cad88, CH15, HV18] (combined with ESPRIT for frequency retrieval, respectively), vanilla-ESPRIT [RK89] and Prony's method. For our method, we choose the regularization parameter λ from (3.35) according to an adaptive rule that uses the information of the model order $r = 2$.

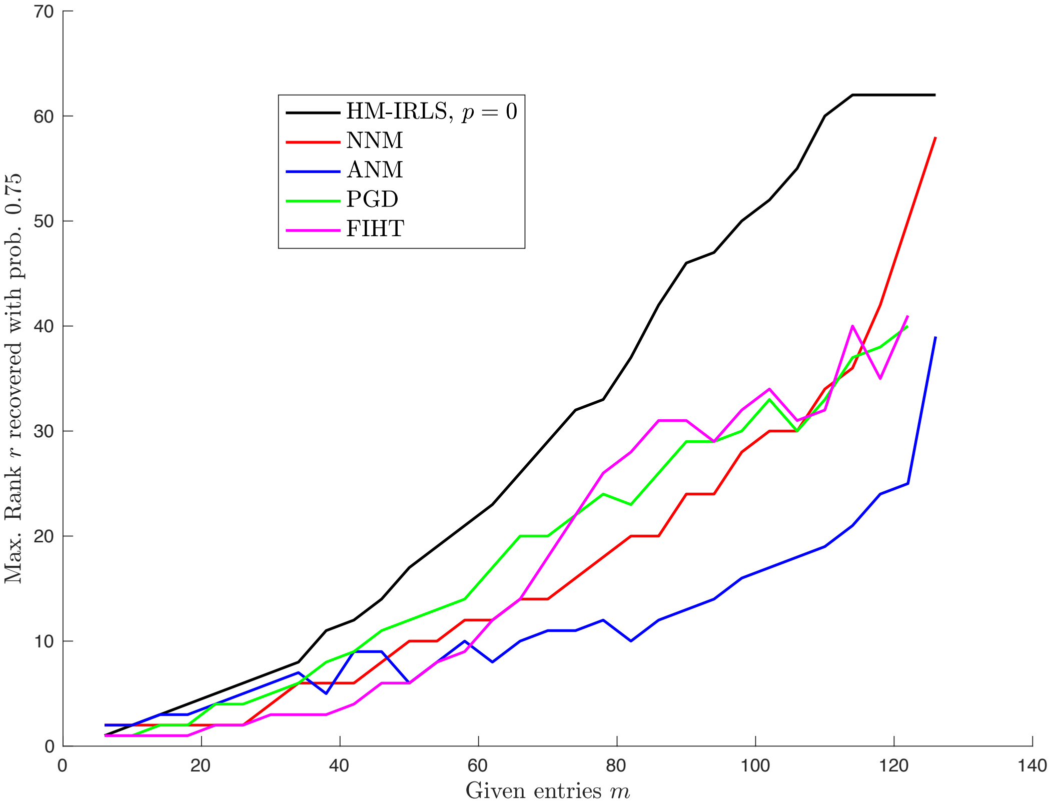

Figure 3.3 – Hankel matrix completion, Max. rank r with 75% empirical recovery probability for a number of m samples $\mathbf{x} \in \mathbb{C}^n$ with $n = 127$.

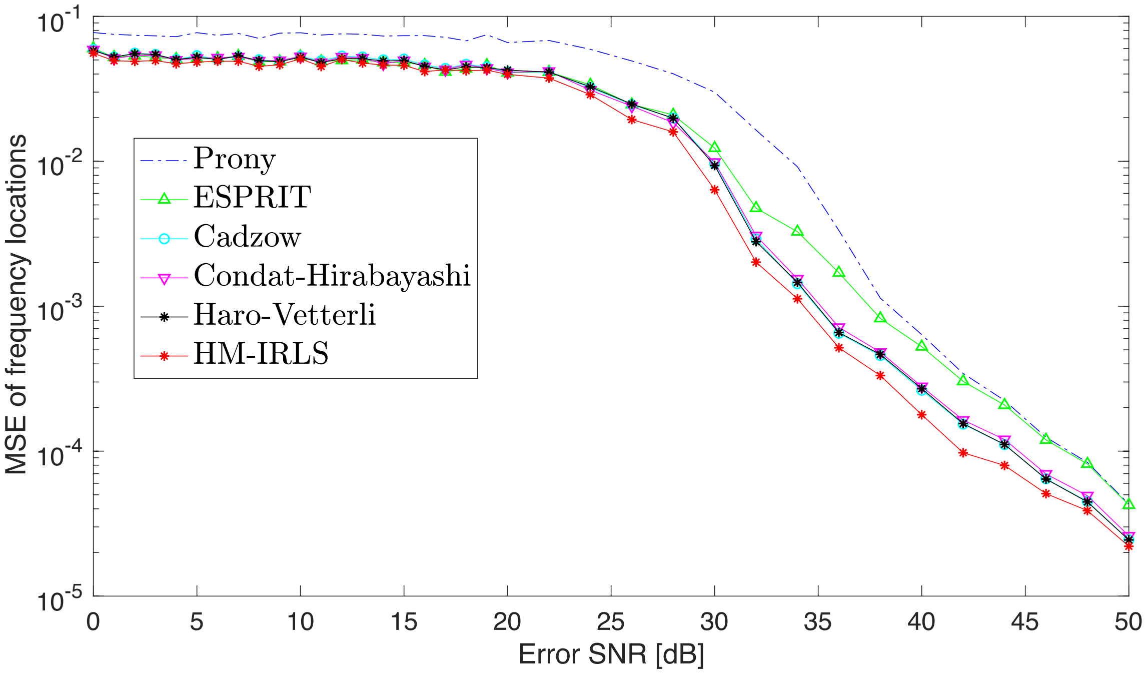

Figure 3.4 – Frequency estimation experiment, $r = 2$, $\alpha_1 = \alpha_2 = 1$ and $f_1 = 0.35$ and $f_2 = 0.40$.

The results corresponding to an average over 500 independent noise realizations for each SNR value can be seen in Figure 3.4. In this simple experiment, and our method obtains a lower MSE on the vector of frequencies $f = (f_1, f_2)$ than the competing methods across different noise SNRs.

3.7 Proof of Theorem 3.2

For the proof of Theorem 3.2, we can use the strategy that had been used in Chapter 2 for proving the respective convergence result for `MatrixIRLS`.

Similarly to the proof of Corollary 2.5.3, we start by adapting the sampling model of Theorem 3.2 as follows: We sample the collection Ω of indices $\omega_1, \ldots, \omega_m \in \{0, 1, \ldots, n-1\}$ from a random model of independent sampling with replacement (this means that Ω could contain the same index more than once).

Lemma 3.7.1 (Based on Proposition 5 from [Rec11]). *Let $\beta > 1$, let $\Omega = (\omega_\ell)_{\ell=1}^m$ be a multiset of indices in $\{0, 1, \ldots, n-1\}$ fulfilling $m < n$, that is sampled independently with replacement. Then with probability at least $1 - n^{1-2\beta}$, the maximum number of repetitions of any entry in Ω is less than $\frac{8}{3}\beta \log(n)$ for all $n \geq 9$.*

Defining the operator $\mathcal{R}_\Omega : \mathbb{C}^n \to \mathbb{C}^n$ such that

$$\mathcal{R}_\Omega(\mathbf{x}) := P_\Omega^* P_\Omega(\mathbf{x})) = \sum_{\ell=1}^m e_{\omega_\ell} e_{\omega_{ell}}^* \mathbf{x}, \tag{3.36}$$

we have consequently that with probability at least $1 - n^{1-2\beta}$

$$\|\mathcal{R}_\Omega\|_{S_\infty} \leq \frac{8}{3}\beta \log(n).$$

Proof. The proof follows exactly the same lines as [Rec11, Proposition 5], with the only difference that in the application of the Chernoff bound, the probability of a head is given by $\frac{1}{n}$ instead of $\frac{1}{d_1 d_2}$ because we are sampling a vector $\mathbf{x} \in \mathbb{C}^n$. This leads to a bound $n^{1-2\beta}$ instead of $\max(d_1, d_2)^{2-2\beta}$. □

As a preparation of the next result, which corresponds to a *local restricted isometry property* (see [CR09, Section 4.2] for some discussion) around matrices that are μ_0-incoherent to the standard Hankel basis, we define the operator $\mathcal{G} : \mathbb{C}^n \to \mathbb{C}^{d_1 \times d_2}$ such that

$$\mathcal{G}(\mathbf{x}) = \sum_{k=0}^{n-1} (\mathbf{x})_k \mathcal{H}_k,$$

where the $\mathcal{H}_k \in \mathbb{C}^{d_1 \times d_2}$ are the standard Hankel basis matrices (3.27).

Lemma 3.7.2 ([CC14, Lemma 3]). *Let $\mathcal{H}(\mathbf{x}) \in \mathbb{C}^{d_1 \times d_2}$ be a rank-r Hankel matrix with left and right singular vector matrices $\mathbf{U} \in \mathbb{C}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{C}^{d_2 \times r}$ and tangent space T (3.26). If $\mathcal{H}(\mathbf{x})$ is μ_0-incoherent with respect to Definition 3.3.1, then*

$$\left\| \mathcal{P}_T \mathcal{G} \mathcal{G}^* \mathcal{P}_T - \frac{n}{m} \mathcal{P}_T \mathcal{G} \mathcal{P}_\Omega \mathcal{G}^* \mathcal{P}_T \right\|_{S_\infty} \leq \varepsilon \tag{3.37}$$

. *holds with probability at least $1 - n^{-2}$ provided that $m \geq \frac{6}{7\varepsilon^2}\mu_0 c_s r \log(n)$, where $c_s =$* $\max\left\{\frac{n}{d_1}, \frac{n}{d_2}\right\}$.

Proof. This lemma is a special case of the result of [LLJY18, Lemma 23]. For completeness, we provide a proof. First we define the family of operators

$$\mathscr{Z}_\ell := \frac{n}{m}\mathscr{P}_T\mathscr{G}e_{\omega_\ell}e_{\omega_\ell}^*\mathscr{G}^*\mathscr{P}_T - \frac{1}{m}\mathscr{P}_T\mathscr{G}\mathscr{G}^*\mathscr{P}_T$$

for any $\ell \in [m]$. Then

$$\mathbb{E}[\mathscr{Z}_\ell] = \frac{1}{n}\sum_{i=1}^{n}\frac{n}{m}\mathscr{P}_T\mathscr{G}e_ie_i^*\mathscr{G}^*\mathscr{P}_T - \frac{1}{m}\mathscr{P}_T\mathscr{G}\mathscr{G}^*\mathscr{P}_T = \frac{1}{n}\frac{n}{m}\mathscr{P}_T\mathscr{G}\mathscr{G}^*\mathscr{P}_T - \frac{1}{m}\mathscr{P}_T\mathscr{G}\mathscr{G}^*\mathscr{P}_T = 0.$$

Since for $X \in \mathbb{C}^{d_1\times d_2}$

$$\mathscr{P}_T\mathscr{G}e_{\omega_\ell}e_{\omega_\ell}^*\mathscr{G}^*\mathscr{P}_T(X) = \langle\mathscr{G}(e_{\omega_\ell}),\mathscr{P}_T(X)\rangle_F\mathscr{P}_T\mathscr{G}(e_{\omega_\ell}) = \langle\mathscr{P}_T\mathscr{G}(e_{\omega_\ell}),X\rangle_F\mathscr{P}_T\mathscr{G}(e_{\omega_\ell}),$$

we obtain

$$\|\mathscr{P}_T\mathscr{G}e_{\omega_\ell}e_{\omega_\ell}^*\mathscr{G}^*\mathscr{P}_T(X)\|_F \leq \left|\langle\mathscr{P}_T\mathscr{G}(e_{\omega_\ell}),X\rangle_F\right|\|\mathscr{P}_T\mathscr{G}(e_{\omega_\ell})\|_F \leq \|\mathscr{P}_T\mathscr{G}(e_{\omega_\ell})\|_F^2\|X\|_F$$

by Cauchy-Schwartz, and thus the norm bound

$$\left\|\frac{n}{m}\mathscr{P}_T\mathscr{G}e_{\omega_\ell}e_{\omega_\ell}^*\mathscr{G}^*\mathscr{P}_T\right\|_2 \leq \frac{n}{m}\|\mathscr{P}_T\mathscr{G}(e_{\omega_\ell})\|_F^2 \leq \frac{n}{m}\max_{k\in[n]}\|\mathscr{P}_T(\mathscr{H}_k)\|_F^2 \leq \frac{n}{m}\frac{2\mu_0c_sr}{n} = \frac{2\mu_0c_sr}{m} \tag{3.38}$$

using the incoherence assumption on the subspace T in the last inequality. Similarly,

$$\left\|\frac{1}{m}\mathscr{P}_T\mathscr{G}\mathscr{G}^*\mathscr{P}_T\right\|_2 \leq \frac{1}{m}\sum_{k=1}^{n}\left\|\mathscr{P}_T\mathscr{G}e_ke_k^*\mathscr{G}^*\mathscr{P}_T\right\|_2 \leq \frac{2\mu_0c_sr}{m} \tag{3.39}$$

and therefore $\|\mathscr{Z}_\ell\|_2 \leq \frac{4\mu_0c_sr}{m}$ for all $\ell \in [m]$. For the expectation of the squares of $\mathscr{Z}_\ell$, we obtain

$$\mathbb{E}\mathscr{Z}_\ell\mathscr{Z}_\ell^* = \frac{n^2}{m^2}\mathbb{E}\left[(\mathscr{P}_T\mathscr{G}e_{\omega_\ell}e_{\omega_\ell}^*\mathscr{G}^*\mathscr{P}_T)^*(\mathscr{P}_T\mathscr{G}e_{\omega_\ell}e_{\omega_\ell}^*\mathscr{G}^*\mathscr{P}_T)\right] - \frac{1}{m^2}(\mathscr{P}_T\mathscr{G}\mathscr{G}^*\mathscr{P}_T)^*(\mathscr{P}_T\mathscr{G}\mathscr{G}^*\mathscr{P}_T)$$

and therefore

$$\begin{aligned}\left\|\sum_{\ell=1}^{m}\mathbb{E}\mathscr{Z}_\ell\mathscr{Z}_\ell^*\right\|_2 &\leq \sum_{\ell=1}^{m}\left\|\mathbb{E}\mathscr{Z}_\ell\mathscr{Z}_\ell^*\right\|_2 = \sum_{\ell=1}^{m}\left(\frac{n^2}{m^2}\left\|\mathbb{E}\left[(\mathscr{P}_T\mathscr{G}e_{\omega_\ell}e_{\omega_\ell}^*\mathscr{G}^*\mathscr{P}_T)^2\right]\right\|_2 + \frac{1}{m^2}\|\mathscr{P}_T\mathscr{G}\mathscr{G}^*\mathscr{P}_T\|_2^2\right)\\ &\leq \sum_{\ell=1}^{m}\left(\frac{n^2}{m^2}\|\mathscr{P}_T\mathscr{G}e_{\omega_\ell}e_{\omega_\ell}^*\mathscr{G}^*\mathscr{P}_T\|_2\left\|\mathbb{E}\left[\mathscr{P}_T\mathscr{G}e_{\omega_\ell}e_{\omega_\ell}^*\mathscr{G}^*\mathscr{P}_T\right]\right\|_2 + \frac{1}{m^2}2\mu_0c_sr\cdot 1\right)\\ &\leq \sum_{\ell=1}^{m}\left(\frac{n^2}{m^2}\frac{2\mu_0c_sr}{n}\left\|\mathbb{E}\left[\mathscr{P}_T\mathscr{G}e_{\omega_\ell}e_{\omega_\ell}^*\mathscr{G}^*\mathscr{P}_T\right]\right\|_2 + \frac{1}{m^2}2\mu_0c_sr\right)\\ &= \sum_{\ell=1}^{m}\left(\frac{n^2}{m^2}\frac{2\mu_0c_sr}{n}\left\|\frac{1}{n}\mathscr{P}_T\mathscr{G}\mathscr{G}^*\mathscr{P}_T\right\|_2 + \frac{1}{m^2}2\mu_0c_sr\right) = \frac{1}{m}4\mu_0c_sr,\end{aligned}$$

where we used that $\|\mathscr{P}_T\mathscr{G}\mathscr{G}^*\mathscr{P}_T\|_2 \leq 1$ since $\mathscr{P}_T\mathscr{G}\mathscr{G}^*\mathscr{P}_T$ is a projection and (3.39) in the second inequality, (3.38) in the third inequality and the fact that the $\{\omega_\ell\}$ are uniformly distributed at random among $[n]$ in the second equality. As the $\mathscr{X}_\ell$ are Hermitian, it follows by the matrix Bernstein inequality (see, e.g., [Ver18, Theorem 5.4.1]) that

$$\begin{aligned} &\mathbb{P}\left(\left\|\frac{n}{m}\mathscr{P}_T\mathscr{G}\mathscr{P}_\Omega\mathscr{G}^*\mathscr{P}_T - \mathscr{P}_T\mathscr{G}\mathscr{G}^*\mathscr{P}_T\right\|_2 \geq \epsilon\right) \leq 2d_1d_2\exp\left(-\frac{m\epsilon^2/2}{4\mu_0c_sr + 4\mu_0c_sr\epsilon/3}\right) \\ &\leq 2\frac{(n+1)^2}{4}\exp\left(-\frac{m\epsilon^2}{2\mu_0c_sr + \mu_0c_sr/3}\right) = \frac{1}{2}(n+1)^2\exp\left(-\frac{m\epsilon^2}{\frac{7}{3}\mu_0c_sr}\right) \leq n^{-2}, \end{aligned} \tag{3.40}$$

where the last inequality holds if $\frac{1}{2}(n+1)^2n^2 \leq \exp\left(\frac{3m\epsilon^2}{7\mu_0c_sr}\right)$, which is implied by the condition $m \geq \frac{6}{7\epsilon^2}\mu_0c_sr\log(n)$. □

As a consequence of it, it is possible to establish a crucial inequality that controls the projection onto the tangent space.

Lemma 3.7.3 ([YKJL17, Lemma 20]). *Let $\mathbf{x} \in \mathbb{C}^n$ be such that T is the tangent space of the embedded rank r matrix manifold at the point $\mathscr{H}(\mathbf{x})$. Assume that (3.37) holds for $\varepsilon \leq \frac{1}{2}$. Then, for any $\mathbf{z} \in \mathbb{C}^n$, it holds that*

$$\|\mathscr{P}_T\mathscr{H}(\mathbf{z})\|_F < 3n\,\|\mathscr{P}_{T^\perp}\mathscr{H}(\mathbf{z})\|_F \qquad \textit{for all } \mathbf{z} \in \ker P_\Omega.$$

Proof. The proof follows from [CC14, |Lemma 1]. □

The next lemma is the counterpart Lemma 2.5.3. We included the refinement of [WCCL16b, Lemma 4.2], obtaining weaker third condition.

Lemma 3.7.4 ([CWW19, Lemma 8]). *Let $\mathbf{x}_0, \mathbf{x} \in \mathbb{C}^n$ be such that $\mathscr{H}(\mathbf{x}_0)$ and $\mathscr{H}(\mathbf{x})$ are rank-r matrices with tangent spaces onto the rank-r manifold T_0 and T, respectively. Assume that $0 < \varepsilon < 1/10$ and that the following three conditions hold:*

a. $\|\mathscr{R}_\Omega\| \leq \frac{8}{3}\log(n)$ *for* $\mathscr{R}_\Omega$ *as in* (3.36),

b. $\|\mathscr{P}_{T_0}\mathscr{G}\mathscr{G}^*\mathscr{P}_{T_0} - \frac{n}{m}\mathscr{P}_{T_0}\mathscr{G}\mathscr{P}_\Omega\mathscr{G}^*\mathscr{P}_{T_0}\| \leq \varepsilon$

c. $\|\mathscr{H}(\mathbf{x}^{(k)}) - \mathscr{H}(\mathbf{x}_0)\|_{S_\infty} \leq \frac{\sqrt{m}\varepsilon}{16\sqrt{n}\log(n)(1+\varepsilon)}\sigma_r(\mathscr{H}(\mathbf{x}_0))$.

Then we have

$$\left\|\mathscr{P}_T\mathscr{G}\mathscr{G}^*\mathscr{P}_T - p^{-1}\mathscr{P}_T\mathscr{G}\mathscr{P}_\Omega\mathscr{G}^*\mathscr{P}_T\right\| \leq 4\varepsilon, \tag{3.41}$$

and, as a consequence of Lemma 3.7.3, we have

$$\|\mathscr{P}_T\mathscr{H}(\mathbf{z})\|_F < 3n\,\|\mathscr{P}_{T^\perp}\mathscr{H}(\mathbf{z})\|_F \qquad \forall\mathbf{z} \in \ker P_\Omega. \tag{3.42}$$

We can finally prove our main result, Theorem 3.2.

Proof of Theorem 3.2. We proceed as in the proof of Corollary 2.5.3.

If $m \geq \frac{600}{7}\mu_0c_sr\log(n)$, then on the event $E_\Omega \cap E_{\Omega,T_0}$, on which the statements of both Lemma 3.7.1 (with $\beta = 2$) and Lemma 3.7.2 hold, we have that

$$\|\mathscr{R}_\Omega\|_{S_\infty} \leq \frac{16}{3}\log(n)$$

and

$$\left\| \mathcal{P}_T \mathcal{G}\mathcal{G}^* \mathcal{P}_T - \frac{n}{m} \mathcal{P}_T \mathcal{G} \mathcal{P}_\Omega \mathcal{G}^* \mathcal{P}_T \right\|_{S_\infty} \leq \varepsilon$$

for $\varepsilon = 1/10$, whereas the probability that $E_\Omega \cap E_{\Omega,T_0}$ hold is at least $1 - 2n^{-2}$.

On $E_\Omega \cap E_{\Omega,T_0}$, it holds then that for any $\mathbf{x} \in \mathbb{C}^n$ such that

$$\|\mathcal{H}(\mathbf{x}) - \mathcal{H}(\mathbf{x}_0)\|_{S_\infty} \leq \widetilde{\xi}\sigma_r(\mathcal{H}(\mathbf{x}_0)),$$

with

$$\widetilde{\xi} = \frac{\sqrt{m}\varepsilon}{16\sqrt{n}\log(n)(1+\varepsilon)},$$

that

$$\|\mathcal{H}(\mathbf{z})\|_F^2 \leq \|\mathcal{P}_T(\mathcal{H}(\mathbf{z}))\|_F^2 + \|\mathcal{P}_{T^\perp}(\mathcal{H}(\mathbf{z}))\|_F^2 \leq \left(9n^2 + 1\right) \|\mathcal{P}_{T^\perp}(\mathcal{H}(\mathbf{z}))\|_F^2.$$

Due to the assumption on the sample complexity m, this implies that Assumption 2.1 is fulfilled with constant $c(d_1, d_2) = \sqrt{9n^2+1}$ for $\mathcal{H}(\mathbf{x}_0)$ with radius

$$\xi = C\frac{\sqrt{\mu_0 c_s r}}{\sqrt{n\log(n)}}.$$

Inserting this into the statement of Theorem 2.7, we obtain that for any $\mathcal{H}(\mathbf{x}^{(k)})$ such that

$$\|\mathcal{H}(\mathbf{x}^{(k)}) - \mathcal{H}(\mathbf{x}_0)\|_{S_\infty} \leq \min\left(C_1\frac{\sqrt{\mu_0 c_s r}}{\sqrt{n\log(n)}}, C_2\frac{1}{n^2 r\kappa}\right)\sigma_r(\mathcal{H}(\mathbf{x}_0)),$$

where C_1 and C_2 are appropriately chosen absolute constants, we have that

$$\|\mathcal{H}(\mathbf{x}^{(k+1)}) - \mathcal{H}(\mathbf{x}_0)\|_{S_\infty} \leq \mu\|\mathcal{H}(\mathbf{x}^{(k)}) - \mathcal{H}(\mathbf{x}_0)\|_{S_\infty}^2$$

with $\mu = 4(9n^2+1)r6\kappa/\sigma_r(\mathcal{H}(\mathbf{x}_0))$.

This finishes the proof of Theorem 3.2.

□

3.8 Conclusion and Future work

In this chapter, we formulated an IRLS algorithm designed for structured low-rank matrix recovery problems. We exhibited beneficial computational properties and theoretical guarantees for local convergence. In particular, we proved that the proposed algorithm StrucIRLS converges *quadratically* to the global minimizer of the low-rank completion problem of matrices Hankel structure if the iterates are in a neighborhood of this minimizer, under a weak assumption on the number of provided random samples. This number depends only linearly on the rank of the Hankel matrix and logarithmically on the ambient dimension, which is considerably weaker than the sufficient conditions that are available for other methods such as projected gradient descent [CWW18], iterative hard thresholding [CWW19] or those based on convex relaxations [FHB04, CC14, JLY16].

The numerical experiments we conducted show that the strong local convergence statements for our algorithm capture indeed a very competitive *data efficiency*, as the method is able to complete structured low-rank Hankel matrices *from fewer samples* than other state-of-the-art methods. An important open question that remains for future work is the establishment of a global convergence theory, as our convergence theory is inherently *local*, allowing for strong convergence statements just in a neighborhood of a ground truth solution. The promising numerical experiments suggest that convergence to the ground truth, however, is in fact *not* a mere local phenomenon, but can be observed with empirical certainty once enough measurements or known entry of the structured matrix are provided, starting from the generic initialization of Algorithm 2. We expect that analysing the saddle escaping behavior of the algorithm as outlined in Section 2.10 will lead to a more comprehensive understanding.

We also showed that an IRLS approach designed for the structured low-rank *approximation* problem has potentially interesting applications as a preprocessing step in harmonic retrieval problems. A generalization of our framework can also be used to cope with *noise* on the samples in the structured matrix completion problem, using an unconstrained formulation as in Section 3.5 with similarities to [LXY13]. We leave the establishment of a convergence theory that includes error bounds for this case to future investigations.

The wide applicability of structured low-rank modelling in signal processing and in control theory was briefly mentioned in the motivation for this chapter. We expect that methods such as ours that are able to cope well with missing or non-equidistantly sampled data can lead to new applications, which would not be possible to realize with conventional algorithms such as *subspace methods*, as these methods typically require very regular samples.

Bibliography

[AAGM15] S. Artstein-Avidan, A. Giannopoulos, and V. D. Milman, *Asymptotic Geometric Analysis, Part I*, Math. Surveys and Monogr., vol. 202, American Mathematical Soc., 2015.

[ABB19] B. Adcock, A. Bao, and S. Brugiapaglia, *Correcting for unknown errors in sparse high-dimensional function approximation*, Numer. Math. **142** (2019), no. 3, 667–711.

[AC14] F. Andersson and M. Carlsson, *ESPRIT for Multidimensional General Grids*, SIAM J. Matrix Anal. Appl. **39** (2014), no. 3, 1470–1488.

[ACP16] F. Andersson, M. Carlsson, and K.-M. Perfekt, *Operator-Lipschitz estimates for the singular value functional calculus*, Proc. Amer. Math. Soc. **144** (2016), no. 5, 1867–1875.

[ALPTJ11] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann, *Restricted Isometry Property of Matrices with Independent Columns and Neighborly Polytopes by Random Sampling*, Constr. Approx. **34** (2011), no. 1, 61–88.

[AR15] A. Ahmed and J. Romberg, *Compressive Multiplexing of Correlated Signals*, IEEE Trans. Inf. Theory **61** (2015), no. 1, 479–498.

[ARR14] A. Ahmed, B. Recht, and J. Romberg, *Blind Deconvolution Using Convex Programming*, IEEE Trans. Inf. Theory **60** (2014), no. 3, 1711–1732.

[Aud14] K. M. R. Audenaert, *A generalisation of Mirsky's singular value inequalities*, arXiv preprint (2014). preprint, arXiv:1410.4941 [math.FA].

[AV97] G. Aubert and L. Vese, *A variational method in image recovery*, SIAM J. Numer. Anal. **34** (1997), no. 5, 1948–1979.

[BA18] S. Brugiapaglia and B. Adcock, *Robustness to unknown error in sparse regularization*, IEEE Trans. Inf. Theory **64** (2018), no. 10, 6638–6661.

[BCB14] D. Bahdanau, K. Cho, and Y. Bengio, *Neural Machine Translation by Jointly Learning to Align and Translate*, arXiv preprints, arXiv:1409.0473 (2014).

[BCE06] R. Balan, P. Casazza, and D. Edidin, *On signal reconstruction without phase*, Appl. Comput. Harmon. Anal. **20** (2006), no. 3, 345 –356.

[BCW11] A. Belloni, V. Chernozhukov, and L. Wang, *Square-root lasso: pivotal recovery of sparse signals via conic programming*, Biometrika **98** (2011), no. 4, 791–806.

[BD09] T. Blumensath and M. E. Davies, *Iterative hard thresholding for compressed sensing*, Appl. Comput. Harmon. Anal. **27** (2009), no. 3, 265 –274.

[BDDW08] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A simple proof of the restricted isometry property for random matrices*, Constr. Approx. **28** (2008), no. 3, 253–263.

[BDMS09] N. Bissantz, L. Dümbgen, A. Munk, and B. Stratmann, *Convergence Analysis of Generalized Iteratively Reweighted Least Squares Algorithms on Convex Function Spaces*, SIAM J. Optim. **19** (2009), no. 4, 1828–1845.

[Bec15] A. Beck, *On the Convergence of Alternating Minimization for Convex Programming with Applications to Iteratively Reweighted Least Squares and Decomposition Schemes*, SIAM J. Optim. **25** (2015), no. 1, 185–209.

[Bec17] ______, *First-Order Methods in Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

[Ber09] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas (Second Edition)*, Princeton University Press, 2009.

[Ber99] D. P. Bertsekas, *Nonlinear programming (Second edition)*, Athena Scientific Optimization and Computation Series, Athena Scientific, Belmont, MA, 1999.

[BGMN05] F. Barthe, O. Guédon, S. Mendelson, and A. Naor, *A probabilistic approach to the geometry of the ℓ_p-ball*, Ann. Probab. **33** (2005), no. 2, 480–513.

[BGVV14] S. Brazitikos, A. Giannopoulos, P. Valettas, and B.-H. Vritsiou, *Geometry of isotropic convex bodies*, Mathematical Surveys and Monographs, vol. 196, American Mathematical Society, Providence, Rhode Island, 2014.

[BH09] T. Bühler and M. Hein, *Spectral clustering based on the graph p-Laplacian*, International Conference on Machine Learning (ICML), 2009, pp. 81–88.

[BK10] R. Blume-Kohout, *Optimal, reliable estimation of quantum states*, New J. Phys. **12** (2010), no. 4, 043034.

[BL07] J. Bennett and S. Lanning, *The Netflix Prize*, Proceedings of KDD cup and workshop, 2007, pp. 35.

[BL88] D. Böhning and B. G. Lindsay, *Monotonicity of quadratic-approximation algorithms*, Ann. Inst. Statist. Math. **40** (1988), no. 4, 641–663.

[BLMK12] S. Babacan, M. Luessi, R. Molina, and A. Katsaggelos, *Sparse Bayesian methods for low-rank matrix estimation*, IEEE Trans. Signal Process. **60** (2012), no. 8, 3964–3977.

[BML$^+$17] L. Brown, C. C. Mosher, C. Li, R. Olson, J. Doherty, T. C. Carey, L. Williams, J. Chang, and E. Staples, *Application of compressive seismic imaging at Lookout Field, Alaska*, The Leading Edge **36** (2017), no. 8, 670–676.

[BNS16] S. Bhojanapalli, B. Neyshabur, and N. Srebro, *Global Optimality of Local Search for Low Rank Matrix Recovery*, Advances in Neural Information Processing Systems (NIPS), 2016, pp. 3873–3881.

[BP16] J. Bolte and E. Pauwels, *Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs*, Math. Oper. Res. **41** (2016), no. 2, 442–465.

[BQW09] K. Browne, S. Qiao, and Y. Wei, *A Lanczos bidiagonalization algorithm for Hankel matrices*, Linear Algebra Appl. **430** (2009), no. 5, 1531–1543. Special Issue devoted to the 14th ILAS Conference.

[BS15] A. Beck and S. Sabach, *Weiszfeld's Method: Old and New Results*, J. Optim. Theory Appl. **164** (2015), no. 1, 1–40.

[BT74] A. E. Beaton and J. W. Tukey, *The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data*, Technometrics **16** (1974), no. 2, 147–185.

[BTR13] B. N. Bhaskar, G. Tang, and B. Recht, *Atomic Norm Denoising With Applications to Line Spectral Estimation*, IEEE Trans. Signal Process. **61** (2013), no. 23, 5987–5999.

[BTSR13] B. N. Bhaskar, G. Tang, P. Shah, and B. Recht, *Compressed Sensing Off the Grid*, IEEE Trans. Inf. Theory **59** (2013), no. 11, 7465–7490.

[BTW15] J. D. Blanchard, J. Tanner, and K. Wei, *CGIHT: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion*, Inf. Inference **4** (2015), no. 4, 289–327.

[Bul03] P. S. Bullen, *Handbook of means and their inequalities*, Mathematics and Its Applications, vol. 560, Springer Science & Business Media, 2003.

[BV04] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 2004.

[Cad88] J. A. Cadzow, *Signal enhancement - A composite property mapping algorithm*, IEEE Trans. Acoust., Speech, Signal Process. **36** (1988), no. 1, 49–62.

[Can08] E. J. Candès, *The restricted isometry property and its implications for compressed sensing*, C.R. Math. **346** (2008), no. 9, 589 –592.

[CBAB97] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, *Deterministic edge-preserving regularization in computed imaging*, IEEE Trans. Image Process. **6** (1997), no. 2, 298–311.

[CC14] Y. Chen and Y. Chi, *Robust Spectral Compressed Sensing via Structured Matrix Completion*, IEEE Trans. Inf. Theory **60** (2014), no. 10, 6576–6601.

[CC18] Y. Chen and Y. Chi, *Harnessing Structures in Big Data via Guaranteed Low-Rank Matrix Estimation: Recent Theory and Fast Algorithms via Convex and Nonconvex Optimization*, IEEE Signal Process. Mag. **35** (2018), no. 4, 14–31.

[CC90] R. Cominetti and R. Correa, *A generalized second-order derivative in nonsmooth optimization*, SIAM J. Control Optim. **28** (1990), no. 4, 789–809.

[CCS10] J.-F. Cai, E. J Candès, and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM J. Optim. **20** (2010), no. 4, 1956–1982.

[CDK15] J. A. Chavez-Dominguez and D. Kutzarova, *Stability of low-rank matrix recovery and its connections to Banach space geometry*, J. Math. Anal. Appl. **427** (2015), no. 1, 320–335.

[CESV13] E. J. Candès, Y. Eldar, T. Strohmer, and V. Voroninski, *Phase Retrieval via Matrix Completion*, SIAM J. Imag. Sci. **6** (2013), no. 1, 199–225.

[CFG14] E. J. Candès and C. Fernandez-Granda, *Towards a Mathematical Theory of Super-resolution*, Comm. Pure Appl. Math. **67** (2014), no. 6, 906–956.

[CH15] L. Condat and A. Hirabayashi, *Cadzow denoising upgraded: A new projection method for the recovery of dirac pulses from noisy linear measurements*, Sampl. Theory Signal Image Process. **14** (2015), no. 1, 17–47.

[Cha07] R. Chartrand, *Exact reconstructions of sparse signals via nonconvex minimization*, IEEE Signal Process. Lett. **14** (2007), no. 10, 707–710.

[Che15] Y. Chen, *Incoherence-Optimal Matrix Completion*, IEEE Trans. Inf. Theory **61** (2015), no. 5, 2909–2923.

[CL14] E. J. Candès and X. Li, *Solving Quadratic Equations via PhaseLift When There Are About as Many Equations as Unknowns*, Found. Comput. Math. **14** (2014), no. 5, 1017–1026.

[Cla75] F. H. Clarke, *Generalized gradients and applications*, Trans. Amer. Math. Soc. **205** (1975), 247–262.

[Cla90] F. Clarke, *Optimization and Nonsmooth Analysis*, Society for Industrial and Applied Mathematics, 1990.

[CLC19] Y. Chi, Y. M. Lu, and Y. Chen, *Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview*, IEEE Trans. Signal Process. **67** (2019), no. 20, 5239–5269.

[CLL19] J. Chen, D. Liu, and X. Li, *Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$-regularization*, arXiv preprint arXiv:1901.06116 (2019).

[CLMW11] E. J. Candès, X. Li, Y. Ma, and J. Wright, *Robust principal component analysis?*, J. ACM **58** (2011), no. 3, 11:1–11:37.

[CLS15] E. J. Candès, X. Li, and M. Soltanolkotabi, *Phase Retrieval via Wirtinger Flow: Theory and Algorithms*, IEEE Trans. Inf. Theory **61** (2015), no. 4, 1985–2007.

[CP11] E. J. Candès and Y. Plan, *Tight Oracle Inequalities for Low-Rank Matrix Recovery From a Minimal Number of Noisy Random Measurements*, IEEE Trans. Inf. Theory **57** (2011), no. 4, 2342–2359.

[CR09] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math. **9** (2009), no. 6, 717–772.

[CRPW12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, *The Convex Geometry of Linear Inverse Problems*, Found. Comput. Math. **12** (2012), no. 6, 805–849.

[CRT06a] E. J. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inf. Theory **52** (2006), no. 2, 489–509.

[CRT06b] E. J. Candès, J. K. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Commun. Pure Appl. Math. **59** (2006), no. 8, 1207–1223.

[CRWX18] J.-F. Cai, Y. Rong, Y. Wang, and Z. Xu, *Data recovery on a manifold from linear samples: theory and computation*, Annals of Mathematical Sciences and Applications **3** (2018), no. 1, 337–365.

[CSV13] E. J. Candès, T. Strohmer, and V. Voroninski, *PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming*, Commun. Pure Appl. Math. **66** (2013), no. 8, 1241–1274.

[CT05] E. J. Candès and T. Tao, *Decoding by Linear Programming*, IEEE Trans. Inf. Theory **51** (December 2005), no. 12, 4203–4215.

[CT06] E. J. Candes and T. Tao, *Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?*, IEEE Trans. Inf. Theory **52** (2006), no. 12, 5406–5425.

[CT10] E. J. Candès and T. Tao, *The Power of Convex Relaxation: Near-Optimal Matrix Completion*, IEEE Trans. Inf. Theory **56** (2010), no. 5, 2053–2080.

[CW15] Y. Chen and M. Wainwright, *Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees*, arXiv preprint arXiv:1509.03025 (2015).

[CW18] J.-F. Cai and K. Wei, *Exploiting the structure effectively and efficiently in low-rank matrix recovery*, Processing, Analyzing and Learning of Images, Shapes, and Forms **19** (2018), 21 pp.

[CWB08] E. J. Candès, M. B. Wakin, and S. P. Boyd, *Enhancing sparsity by reweighted ℓ_1-minimization*, The Journal of Fourier Analysis and Applications **14** (2008), 877–905.

[CWW18] J.-F. Cai, T. Wang, and K. Wei, *Spectral compressed sensing via projected gradient descent*, SIAM J. Optim **28** (2018), no. 3, 2625–2653.

[CWW19] ———, *Fast and provable algorithms for spectrally sparse signal reconstruction via low-rank hankel matrix completion*, Appl. Comput. Harmon. Anal. **46** (2019), no. 1, 94–121.

[CY08] R. Chartrand and W. T. Yin, *Iteratively reweighted algorithms for compressive sensing*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008, pp. 3869–3872.

[Dau88] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Commun. Pure Appl. Math. **41** (1988), no. 7, 909–996.

[Dau92] I. Daubechies, *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, 1992.

[Dax10] A. Dax, *On extremum properties of orthogonal quotients matrices*, Linear Algebra Appl. **432** (2010), no. 5, 1234–1257.

[DB13] M. F. Duarte and R. G. Baraniuk, *Spectral compressive sensing*, Appl. Comput. Harmon. Anal. **35** (2013), no. 1, 111–129.

[DDFG10] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, *Iteratively reweighted least squares minimization for sparse recovery*, Commun. Pure Appl. Math. **63** (2010), 1–38.

[DGM13] D. L. Donoho, M. Gavish, and A. Montanari, *The phase transition of matrix recovery from Gaussian measurements matches the minimax MSE of matrix denoising*, Proc. Nat. Acad. Sci. U.S.A. **110** (2013), no. 21, 8405–8410.

[DGT09] N. Dafnis, A. Giannopoulos, and A. Tsolomitis, *Asymptotic shape of a random polytope in a convex body*, J. Funct. Anal. **257** (2009), no. 9, 2820 –2839.

[DK15] D. Drusvyatskiy and C. Kempton, *Variational analysis of spectral functions simplified*, arXiv preprint arXiv:1506.05170 (2015).

[DlPG12] V. De la Peña and E. Giné, *Decoupling. From dependence to independence. Randomly stopped processes, U-statistics and processes, martingales and beyond*, Springer Science & Business Media, 2012.

[DLR18] S. Dirksen, G. Lecué, and H. Rauhut, *On the gap between restricted isometry properties and sparse recovery conditions*, IEEE Trans. Inf. Theory **64** (2018), no. 8, 5478–5487.

[DM87] G. R. Ducharme and P. Milasevic, *Spatial median and directional data*, Biometrika **74** (1987), no. 1, 212–215.

[Don05] D. L. Donoho, *Neighborly polytopes and sparse solutions of underdetermined linear equations*, preprint, Stanford Univ., 2005.

[Don06] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inf. Theory **52** (2006), no. 4, 1289 –1306.

[Don18] D. Donoho, *Blackboard to bedside: How high-dimensional geometry is transforming the MRI industry*, Notices of the AMS **65** (2018), no. 1.

[dP95] G. de Prony, *Essai experimental et analytique: sur les lois de la dilatabilité des fluids elastiques et sur celles la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, à différents temperatures*, Journal de l'Ecole Polytechnique **1** (1795), no. 22, 24–76.

[DPW09] R. DeVore, G. Petrova, and P. Wojtaszczyk, *Instance-optimality in probability with an ℓ_1-minimization decoder*, Appl. Comput. Harmon. Anal. **27** (2009), no. 3, 275 –288.

[DR16] M. A. Davenport and J. Romberg, *An Overview of Low-Rank Matrix Recovery From Incomplete Observations*, IEEE J. Sel. Topics Signal Process. **10** (2016), 608–622.

[DSST18] C. Ding, D. Sun, J. Sun, and K.-C. Toh, *Spectral operators of matrices*, Math. Program. **168** (2018), no. 1, 509–531.

[DST09] D. Donoho, V. Stodden, and Y. Tsaig, *Reproducible research in computational harmonic analysis*, Computing in Science Engineering **11** (2009), no. 1, 8 –18.

[DT09] D. Donoho and J. Tanner, *Counting faces of randomly projected polytopes when the projection radically lowers dimension*, J. Amer. Math. Soc **22** (2009), no. 1, 1–53.

[DVB07] P. L. Dragotti, M. Vetterli, and T. Blu, *Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets strang–fix*, IEEE Trans. Signal Process. **55** (2007), no. 5, 1741–1757.

[EACR$^+$16] A. El Alaoui, X. Cheng, A. Ramdas, M. J. Wainwright, and M. I. Jordan, *Asymptotic behavior of ℓ_p-based Laplacian regularization in semi-supervised learning*, Conference on Learning Theory (COLT), 2016, pp. 879–906.

[Eld15] Y. C. Eldar, *Sampling theory: Beyond bandlimited systems*, 1st ed., Cambridge University Press, New York, NY, USA, 2015.

[EM14] Y. C. Eldar and S. Mendelson, *Phase retrieval: Stability and recovery guarantees*, Appl. Comput. Harmon. Anal. **36** (2014), no. 3, 473 –494.

[ENP12] Y. C. Eldar, D. Needell, and Y. Plan, *Uniqueness conditions for low-rank matrix recovery*, Appl. Comput. Harmon. Anal. **33** (2012), no. 2, 309–314.

[EY36] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika **1** (1936), no. 3, 211–218.

[Fac95] F. Facchinei, *Minimization of SC1 functions and the Maratos effect*, Operations Research Letters **17** (1995), no. 3, 131 –137.

[Fan16] A. Fannjiang, *Compressive spectral estimation with single-snapshot esprit: Stability and resolution*, arXiv preprint arXiv:1607.01827 **abs/1607.01827** (2016).

[Faz02] M. Fazel, *Matrix rank minimization with applications*, Ph.D. Thesis, Electrical Engineering Department, Stanford University, 2002.

[FD87] C. Fienup and J. Dainty, *Phase retrieval and image reconstruction for astronomy*, Image recovery: theory and application, 1987, pp. 231–275.

[FGM17] R. M Freund, P. Grigas, and R. Mazumder, *An Extended Frank–Wolfe Method with "In-Face" Directions, and Its Application to Low-Rank Matrix Completion*, SIAM J. Optim. **27** (2017), no. 1, 319–346.

[FHB03] M. Fazel, H. Hindi, and S. P. Boyd, *Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices*, Proceedings of the American Control Conference, 2003, pp. 2156–2162.

[FHB04] M. Fazel, H. Hindi, and S. Boyd, *Rank minimization and applications in system theory*, Proceedings of the american control conference, 2004, pp. 3273–3278.

[Fie82] J. R. Fienup, *Phase retrieval algorithms: a comparison*, Appl. Opt. **21** (1982), no. 15, 2758–2769.

[FL17] S. Foucart and G. Lecué, *An IHT Algorithm for Sparse Recovery From Subexponential Measurements*, IEEE Signal Process. Lett. **24** (2017), no. 9, 1280–1283.

[Fou14] S. Foucart, *Stability and robustness of ℓ_1-minimizations with weibull matrices and redundant dictionaries*, Linear Algebra Appl. **441** (2014), 4–21. special issue on Sparse Approximate Solution of Linear Systems.

[Fou18] S. Foucart, *Concave Mirsky Inequality and Low-Rank Recovery*, SIAM J. Matrix Anal. Appl. **39** (2018), no. 1, 99–103.

[FPRW16] M. Fornasier, S. Peter, H. Rauhut, and S. Worm, *Conjugate gradient acceleration of iteratively re-weighted least squares methods*, Comput. Optim. Appl. **65** (2016), no. 1, 205–259.

[FPST13] M. Fazel, T. Pong, D. Sun, and P. Tseng, *Hankel Matrix Rank Minimization with Applications to System Identification and Realization*, SIAM J. Matrix Anal. Appl. **34** (2013), no. 3, 946–977.

[FR13] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Applied and Numerical Harmonic Analysis, Springer New York, 2013.

[Fri81] S. Friedland, *Convex spectral functions*, Linear and Multilinear Algebra **9** (1981), no. 4, 299–316.

[FRW11] M. Fornasier, H. Rauhut, and R. Ward, *Low-rank matrix recovery via iteratively reweighted least squares minimization*, SIAM J. Optim. **21** (2011), no. 4, 1614–1640.

[Fun06] S. Funk, *Netflix update: Try this at home*, 2006. Blog entry, available at http://sifter.org/~simon/journal/20061211.html.

[FWHS16] C. Forman, J. Wetzl, C. Hayes, and M. Schmidt, *Compressed Sensing: a Paradigm Shift in MRI*, MAGNETOM Flash (2016), 19.

[FWS$^+$16] J. Fang, F. Wang, Y. Shen, H. Li, and R. S. Blum, *Super-Resolution Compressed Sensing for Line Spectral Estimation: An Iterative Reweighted Approach*, IEEE Trans. Signal Process. **64** (2016), no. 18, 4649–4662.

[GB14] M. Grant and S. Boyd, *CVX: Matlab software for disciplined convex programming, version 2.1*, 2014.

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT press, 2016.

[GGK00] I. Gohberg, S. Goldberg, and N. Krupnik, *Traces and determinants of linear operators*, Operator Theory: Advances and Applications, vol. 116, Springer, Basel, 2000.

[GKK15] D. Gross, F. Krahmer, and R. Kueng, *A Partial Derandomization of PhaseLift Using Spherical Designs*, J. Fourier Anal. Appl. **21** (2015), no. 2, 229–266.

[GKK$^+$19] O. Guédon, F. Krahmer, C. Kümmerle, S. Mendelson, and H. Rauhut, *On the geometry of polytopes generated by heavy-tailed random vectors*, arXiv preprint arXiv:1907.07258 (2019).

[GLF$^+$10] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, *Quantum state tomography via compressed sensing*, Phys. Rev. Lett. **105** (2010), 150401.

[GLM16] R. Ge, J. D. Lee, and T. Ma, *Matrix Completion has No Spurious Local Minimum*, Advances in Neural Information Processing Systems (NIPS), 2016, pp. 2973–2981.

[GLT19] O Guédon, A. Litvak, and K Tatarko, *Random polytopes obtained by matrices with heavy-tailed entries*, Communications in Contemporary Mathematics (2019), 1950027.

[Glu89] E. D. Gluskin, *Extremal Properties of Orthogonal Parallelepipeds and Their Applications to the Geometry of Banach Spaces*, Sb. Math. **64** (1989), no. 1, 85.

[GNOT92] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, *Using Collaborative Filtering to Weave an Information Tapestry*, Commun. ACM **35** (1992), no. 12, 61–70.

[GNZ01] N. Golyandina, V. Nekrutkin, and A. A. Zhigljavsky, *Analysis of time series structure: SSA and related techniques*, Chapman and Hall/CRC, 2001.

[Grü03] B. Grünbaum, *Convex polytopes*, 2nd ed. (P. McMullen, G. C. Shephard, J. E. Reeve, and A. Ball, eds.), Springer New York, 2003.

[GR92] D. Geman and G. Reynolds, *Constrained Restoration and the Recovery of Discontinuities*, IEEE Trans. Pattern Anal. Mach. Intell. **14** (1992), no. 03, 367–383.

[GR97] I. F. Gorodnitsky and B. D. Rao, *Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm*, IEEE Trans. Signal Process. (1997), 600–616.

[Gro11] D. Gross, *Recovering Low-Rank Matrices From Few Coefficients in Any Basis*, IEEE Trans. Inf. Theory **57** (2011), no. 3, 1548–1566.

[GS72] R. W. Gerchberg and W. O. Saxton, *A practical algorithm for the determination of the phase from image and diffraction plane pictures*, Optik (Jena) **35** (1972), 227–246.

[GSC13] J. Gao, M. D. Sacchi, and X. Chen, *A fast reduced-rank interpolation method for prestack seismic volumes that depend on four spatial dimensions*, Geophysics **78** (2013), no. 1, V21–V30.

[GXM$^+$17] S. Gu, Q. Xie, D. Meng, W. Zuo, X. Feng, and L. Zhang, *Weighted nuclear norm minimization and its applications to low level vision*, International journal of computer vision **121** (2017), no. 2, 183–208.

[GY95] D. Geman and C. Yang, *Nonlinear image recovery with half-quadratic regularization*, IEEE Trans. Image Process. **4** (1995), no. 7, 932–946.

[GZ13] N. Golyandina and A. Zhigljavsky, *Singular spectrum analysis for time series*, Springer Science & Business Media, 2013.

[GZZF14] S. Gu, L. Zhang, W. Zuo, and X. Feng, *Weighted Nuclear Norm Minimization with Application to Image Denoising*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2862–2869.

[Hal13] J. P. Haldar, *Low-rank modeling of local k-space neighborhoods (LORAKS) for constrained MRI*, IEEE Trans. Med. Imag. **33** (2013), no. 3, 668–681.

[HBFR14] T. L. Hansen, M. A. Badiu, B. H. Fleury, and B. D. Rao, *A sparse Bayesian learning algorithm with dictionary parameter estimation*, IEEE 8th Sensor Array and Multichannel Signal Processing Workshop (SAM), 2014, pp. 385–388.

[HBI73] J. B. Hawkins and A. Ben-Israel, *On generalized matrix functions*, Linear and Multilinear Algebra **1** (1973), no. 2, 163–171.

[Her10] F. J. Herrmann, *Randomized sampling and sparsity: Getting more information from fewer samples*, Geophysics **75** (2010), no. 6, WB173–WB187.

[HH09] J. P. Haldar and D. Hernando, *Rank-Constrained Solutions to Linear Matrix Equations Using PowerFactorization*, IEEE Signal Process. Lett. **16** (2009), no. 7, 584–587.

[HIB$^+$18] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, *Deep reinforcement learning that matters*, Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[HJ12] R. A. Horn and C. R. Johnson, *Matrix analysis*, 2nd ed., Cambridge University Press, 2012.

[HJ91] R. Horn and C. Johnson, *Topics in matrix analysis*, 1st Edition, Cambridge University Press, 1991.

[HLSJ18] C. Herrmann, Y. Lu, C. Scheunert, and P. Jung, *Improving robustness for anisotropic sparse recovery using matrix extensions*, Wsa 2018; 22nd international itg workshop on smart antennas, 2018, pp. 1–7.

[HMT11] N. Halko, P.-G. Martinsson, and J. A. Tropp, *Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions*, SIAM Rev. **53** (2011), no. 2, 217–288.

[HR17] I. Haviv and O. Regev, *The Restricted Isometry Property of Subsampled Fourier Matrices*, Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2014–2016, 2017, pp. 163–179.

[HS52] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, Vol. 49, NBS Washington, DC, 1952.

[HS90] Y. Hua and T. K. Sarkar, *Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise*, IEEE Trans. Signal Process. **38** (1990), no. 5, 814–824.

[HT10] M. Hardt and K. Talwar, *On the geometry of differential privacy*, Proceedings of the forty-second ACM symposium on Theory of computing, 2010, pp. 705–714.

[HU77] J.-B. Hiriart-Urruty, *Contributions à la programmation mathématique: cas déterministe et stochastique*, Ph.D. Thesis, Universiteé de Clermont-Ferrand II, 1977.

[Hua92] Y. Hua, *Estimating two-dimensional frequencies by matrix enhancement and matrix pencil*, IEEE Trans. Signal Process. **40** (1992), no. 9, 2267 –2280.

[HUSN84] J.-B. Hiriart-Urruty, J.-J. Strodiot, and V. H. Nguyen, *Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data*, Appl. Math. Optim. **11** (1984), no. 1, 43–56.

[HV18] B. B. Haro and M. Vetterli, *Sampling continuous-time sparse signals: A frequency-domain perspective*, IEEE Trans. Signal Process. **63** (2018), no. 2, 777–801.

[HW77] P. W. Holland and R. E. Welsch, *Robust regression using iteratively reweighted least-squares*, Communications in Statistics - Theory and Methods **6** (1977), no. 9, 813–827.

[IUM14] M. Ishteva, K. Usevich, and I. Markovsky, *Factorization approach to structured low-rank approximation with applications*, SIAM J. Matrix Anal. Appl. **35** (2014), no. 3, 1180–1204.

[IVW17] M. Iwen, A. Viswanathan, and Y. Wang, *Robust sparse phase retrieval made easy*, Appl. Comput. Harmon. Anal. **42** (2017), no. 1, 135–142.

[JLY16] K. H. Jin, D. Lee, and J. C. Ye, *A general framework for compressed sensing and parallel MRI using annihilating filter based low-rank Hankel matrix*, IEEE Trans. Comput. Imag. **2** (2016), no. 4, 480–495.

[JMD10] P. Jain, R. Meka, and I. S. Dhillon, *Guaranteed Rank Minimization via Singular Value Projection*, Advances in Neural Information Processing Systems (NIPS), 2010, pp. 937–945.

[JNS13] P. Jain, P. Netrapalli, and S. Sanghavi, *Low-rank Matrix Completion Using Alternating Minimization*, Proc. ACM Symp. Theory Comput. (STOC), June 2013, pp. 665–674.

[JOH13] K. Jaganathan, S. Oymak, and B. Hassibi, *Sparse phase retrieval: Convex algorithms and limitations*, IEEE International Symposium on Information Theory, 2013, pp. 1022–1026.

[JY15] K. H. Jin and J. C. Ye, *Annihilating Filter-Based Low-Rank Hankel Matrix Approach for Image Inpainting*, IEEE Trans. Image Process. **24** (2015), no. 11, 3498–3511.

[JYLM16] Y. Jia, S. Yu, L. Liu, and J. Ma, *A fast rank-reduction algorithm for three-dimensional seismic data interpolation*, Journal of Applied Geophysics **132** (2016), 137 –145.

[KAR83] S.-Y. Kung, K. S. Arun, and D. V. B. Rao, *State-space and singular-value decomposition-based approximation methods for the harmonic retrieval problem*, J. Opt. Soc. Am. **73** (1983), no. 12, 1799 –1811.

[Kas83] B. S. Kashin, *On properties of random sections of an n-dimensional cube*, Vestnik Moskov. Univ. Ser. I Mat. Mekh. **3** (1983), 8–11. English transl. in Moscow Univ. Math. Bull. 38 (1983).

[KBV09] Y. Koren, R. Bell, and C. Volinsky, *Matrix Factorization Techniques for Recommender Systems*, Computer **42** (2009), no. 8, 30–37.

[KC14] A. Kyrillidis and V. Cevher, *Matrix recipes for hard thresholding methods*, J. Math. Imaging Vision **48** (2014), no. 2, 235–265. [using `Matrix ALPS II` ('Matrix ALgrebraic PursuitS II') algorithm, code from http://akyrillidis.github.io/projects/].

[KKM$^+$16] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata, and L. Zdeborová, *Phase Transitions and Sample Complexity in Bayes-Optimal Matrix Factorization*, IEEE Trans. Inf. Theory **62** (2016), no. 7, 4228–4265.

[KKR18] F. Krahmer, C. Kümmerle, and H. Rauhut, *A quotient property for matrices with heavy-tailed entries and its application to noise-blind compressed sensing*, arXiv e-prints arXiv:1806.04261 (2018).

[KKRT16] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege, *Stable low-rank matrix recovery via null space properties*, Inf. Inference **5** (2016), no. 4, 405–441.

[KM15] V. Koltchinskii and S. Mendelson, *Bounding the smallest singular value of a random matrix without concentration*, Int. Math. Res. Notices **2015** (2015), no. 23, 12991–13008.

[KMO10] R. H. Keshavan, A. Montanari, and S. Oh, *Matrix Completion From a Few Entries*, IEEE Trans. Inf. Theory **56** (2010), no. 6, 2980–2998.

[KMR14] F. Krahmer, S. Mendelson, and H. Rauhut, *Suprema of chaos processes and the restricted isometry property*, Commun. Pure Appl. Math. **67** (2014), no. 11, 1877–1904.

[Kol11] V. Koltchinskii, *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, Springer, Heidelberg, 2011.

[Kor09] Y. Koren, *The BellKor Solution to the Netflix Grand Prize*, Netflix prize documentation **81** (2009), no. 2009, 1–10.

[Kor10] A. Korobeynikov, *Computation-and space-efficient implementation of SSA*, Statistics and Its Interface **3** (2010), no. 3, 357–368.

[KS17] C. Kümmerle and J. Sigl, *Harmonic Mean Iteratively Reweighted Least Squares for low-rank matrix recovery*, 12th International Conference on Sampling Theory and Applications (SampTA), Tallinn, Estonia, 2017, pp. 489–493.

[KS18] ———, *Harmonic Mean Iteratively Reweighted Least Squares for Low-Rank Matrix Recovery*, Journal of Machine Learning Research **19** (2018), no. 47, 1–49.

[KS99] T. Kailath and A. Sayed, *Fast reliable algorithms for matrices with structure* (T. Kailath and A. H. Sayed, eds.), Society for Industrial and Applied Mathematics, 1999.

[KSH12] A. Krizhevsky, I. Sutskever, and G. E Hinton, *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems (nips), 2012, pp. 1097–1105.

[KT88] D. Klatte and K. Tammek, *On second-order sufficient optimality conditions for $C^{1,1}$-optimization problems*, Optimization **19** (1988), no. 2, 169–179.

[KV18] C. Kümmerle and C. M. Verdun, *Denoising and completion of structured low-rank matrices via iteratively reweighted least squares*, International Traveling Workshop on Interactions between Low-Complexity Data Models and Sensing Techniques (iTWIST), Marseille, France, 2018.

[KV19] C. Kümmerle and C. M. Verdun, *Completion of structured low-rank matrices via iteratively reweighted least squares*, 13th International Conference on Sampling Theory and Applications (SampTA), Bordeaux, France, 2019, pp. 1–4.

[KW92] S. Kwapień and W. A. Woyczyński, *Random series and stochastic integrals: single and multiple*, Probability and its Applications, Birkhäuser Boston, Inc., Boston, MA, 1992. MR1167198

[KY03] H. Kushner and G G. Yin, *Stochastic approximation and recursive algorithms and applications*, Vol. 35, Springer Science & Business Media, 2003.

[Löw34] K. Löwner, *Über monotone Matrixfunktionen*, Mathematische Zeitschrift **38** (1934Dec), no. 1, 177–216.

[Lan16] K. Lange, *MM Optimization Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016.

[Law61] C. L. Lawson, *Contributions to the theory of linear least maximum approximation*, Ph. D. Thesis, Univ. of Calif., Los Angeles, 1961.

[LBH15] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature **521** (2015), no. 7553, 436.

[LDSP08] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, *Compressed Sensing MRI*, IEEE Signal Processing Magazine **25** (2008), no. 2, 72–82.

[Lee13] J. M. Lee, *Introduction to Smooth Manifolds*, 2nd ed., Springer, 2013.

[Lew95] A. S. Lewis, *The Convex Analysis of Unitarily Invariant Matrix Functions*, J. Convex Anal. **2** (1995), no. 1–2, 173–183.

[Lew96] ———, *Convex Analysis on the Hermitian Matrices*, SIAM J. Optim. **6** (1996), no. 1, 164–177.

[LHV13] Z. Liu, A. Hansson, and L. Vandenberghe, *Nuclear norm system identification with missing inputs and outputs*, Systems Control Lett. **62** (2013), no. 8, 605–612.

[Li93] Y. Li, *A Globally Convergent Method for ℓ_p Problems*, SIAM J. Optim. **3** (1993), no. 3, 609–629.

[LJK^{+}16] D. Lee, K. H. Jin, E. Y. Kim, S.-H. Park, and J. C. Ye, *Acceleration of MR parameter mapping using annihilating filter-based low rank hankel matrix (ALOHA)*, Magnetic resonance in medicine **76** (2016), no. 6, 1848–1864.

[LLJY18] K. Lee, Y. Li, K. H. Jin, and J. C. Ye, *Unified Theory for Recovery of Sparse Signals in a General Transform Domain*, IEEE Trans. Inf. Theory **64** (2018), no. 8, 5457–5477.

[LLSW19] X. Li, S. Ling, T. Strohmer, and K. Wei, *Rapid, robust, and reliable blind deconvolution via nonconvex optimization*, Appl. Comput. Harmon. Anal. **47** (2019), no. 3, 893–934.

[Log65] B. F. Logan, *Properties of high-pass signals*, Ph.D. Thesis, Columbia University, 1965.

[LPRTJ05] A. E. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann, *Smallest singular value of random matrices and geometry of random polytopes*, Adv. Math. **195** (2005), no. 2, 491 –523.

[LS01] A. S. Lewis and H. S. Sendov, *Twice Differentiable Spectral Functions*, SIAM J. Matrix Anal. Appl. **23** (2001), no. 2, 368–386.

[LS02] A. S. Lewis and H. S. Sendov, *Quadratic expansions of spectral functions*, Linear Algebra Appl. **340** (2002), no. 1, 97 –121.

[LS05a] ______, *Nonsmooth Analysis of Singular Values. Part I: Theory*, Set-Valued Analysis **13** (2005), no. 3, 213–241.

[LS05b] ______, *Nonsmooth Analysis of Singular Values. Part II: Applications*, Set-Valued Analysis **13** (2005), no. 3, 243–264.

[LTYL15] C. Lu, J. Tang, S. Yan, and Z. Lin, *Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm*, IEEE Trans. Image Process. **25** (2015), no. 2, 829–839.

[LV10] Z. Liu and L. Vandenberghe, *Interior-point method for nuclear norm approximation with application to system identification*, SIAM J. Matrix Anal. Appl. **31** (2010), no. 3, 1235–1256.

[LXQ15] L. Lu, W. Xu, and S. Qiao, *A fast SVD for multilevel block Hankel matrices with minimal memory storage*, Numer. Algor. **69** (2015), no. 4, 875–891.

[LXY13] M. Lai, Y. Xu, and W. Yin, *Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization*, SIAM J. Numer. Anal (2013), 2013.

[LZ97] E. Lutwak and G. Zhang, *Blaschke-santaló inequalities*, Journal of Differential Geometry **47** (1997), no. 1, 1–16.

[Mar19] I. Markovsky, *Structured low-rank approximation: Algorithms, implementation, applications*, 2nd ed., Springer International Publishing, 2019.

[MBS14] M. Malek-Mohammadi, M. Babaie-Zadeh, and M. Skoglund, *Iterative Concave Rank Approximation for Recovering Low-Rank Matrices*, IEEE Trans. Signal Process. **62** (2014), no. 20, 5213–5226.

[MCW05] D. Malioutov, M. Cetin, and A. S. Willsky, *A sparse signal reconstruction perspective for source localization with sensor arrays*, IEEE Trans. Signal Process. **53** (2005), no. 8, 3010–3022.

[Men15] S. Mendelson, *Learning without concentration*, J. ACM **62** (2015), no. 3, 21:1–21:25.

[Men19] ______, *On the geometry of random polytopes*, arXiv preprint arXiv:1902.01664 (2019).

[MF10a] K. Mohan and M. Fazel, *Iterative reweighted least squares for matrix rank minimization*, 2010 48th annual allerton conference on communication, control, and computing (allerton), 2010Sep., pp. 653–661.

[MF10b] K. Mohan and M. Fazel, *Reweighted nuclear norm minimization with application to system identification*, Proceedings of the American Control Conference, 2010, pp. 2953–2959.

[MF12] ______, *Iterative reweighted algorithms for matrix rank minimization*, J. Mach. Learn. Res. **13** (2012), no. 1, 3441–3473.

[MHT10] R. Mazumder, T. Hastie, and R. Tibshirani, *Spectral regularization algorithms for learning large incomplete matrices*, Journal of Machine Learning Research **11** (2010), no. Aug, 2287–2322.

[Mil90] R. P. Millane, *Phase retrieval in crystallography and optics*, J. Opt. Soc. Am. A **7** (1990), no. 3, 394–411.

[Mir60] L. Mirsky, *Symmetric Gauge Functions And Unitarily Invariant Norms*, The Quarterly Journal of Mathematics **11** (1960), no. 1, 50–59.

[MJKM17] M. Mani, M. Jacob, D. Kelley, and V. Magnotta, *Multi-shot sensitivity-encoded diffusion data recovery using structured low-rank matrix completion (MUSSELS)*, Magnetic resonance in medicine **78** (2017), no. 2, 494–507.

[ML17] S. Mendelson and G. Lecué, *Sparse recovery under weak moment assumptions*, J. Eur. Math. Soc. **19** (2017), no. 3, 881–904.

[MLJ$^+$17] C. C. Mosher, C. Li, F. D. Janiszewski, L. S. Williams, T. C. Carey, and Y. Ji, *Operational deployment of compressive sensing systems for seismic data acquisition*, The Leading Edge **36** (2017), no. 8, 661–669.

[MM15] C. Musco and C. Musco, *Randomized Block Krylov Methods for Stronger and Faster Approximate Singular Value Decomposition*, Advances in Neural Information Processing Systems (NIPS), 2015, pp. 1396–1404.

[MMBS13] B. Mishra, G. Meyer, F. Bach, and R. Sepulchre, *Low-Rank Optimization with Trace Norm Penalty*, SIAM J. Optim. **23** (2013), no. 4, 2124–2149.

[Moo94] B. D. Moor, *Total least squares for affinely structured matrices and the noisy realization problem*, IEEE Trans. Signal Proc. **42** (1994), 3104–3113.

[MPTJ08] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, *Uniform uncertainty principle for bernoulli and subgaussian ensembles*, Constructive Approximation **28** (2008), no. 3, 277–289.

[MS18] M. Mousavi and H. S. Sendov, *A Unified Approach to Spectral and Isotropic Functions*, SIAM J. Matrix Anal. Appl. **39** (2018), no. 2, 632–663.

[MS74] G. Merle and H. Späth, *Computational experiences with discrete lp-approximation*, Computing **12** (1974), no. 4, 315–321.

[MU13] I. Markovsky and K. Usevich, *Structured low-rank approximation with missing data*, SIAM J. Matrix Anal. Appl. **34** (2013), no. 2, 814–830.

[MV05] I. Maravić and M. Vetterli, *Sampling and reconstruction of signals with finite rate of innovation in the presence of noise*, IEEE Trans. Signal Process. **53** (2005), no. 8, 2788–2805.

[MWCC19] C. Ma, K. Wang, Y. Chi, and Y. Chen, *Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution*, Foundations of Computational Mathematics (2019).

[Nat95] B. K. Natarajan, *Sparse approximate solutions to linear systems*, SIAM J. Comput. **24** (1995), no. 2, 227–234.

[NN05] M. Nikolova and M. K. Ng, *Analysis of half-quadratic minimization methods for signal and image recovery*, SIAM J. Sci. Comput. **27** (2005), no. 3, 937–966.

[Nof17] V. Noferini, *A Formula for the Fréchet Derivative of a Generalized Matrix Function*, SIAM J. Matrix Anal. Appl. **38** (2017), no. 2, 434–457.

[NT09] D. Needell and J. A. Tropp, *CoSaMP: Iterative signal recovery from incomplete and inaccurate samples*, Appl. Comput. Harmon. Anal. **26** (2009), no. 3, 301 –321.

[NW06] J. Nocedal and S. Wright, *Numerical optimization*, Springer Science & Business Media, 2006.

[Nyq83] H. Nyquist, *The optimal lp norm estimator in linear regression models*, Communications in Statistics - Theory and Methods **12** (1983), no. 21, 2511–2524.

[ODBP15] P. Ochs, A. Dosovitskiy, T. Brox, and T. Pock, *On Iteratively Reweighted Algorithms for Nonsmooth Nonconvex Optimization in Computer Vision*, SIAM J. Imaging Sci. **8** (2015), no. 1, 331–372.

[OJ16] G. Ongie and M. Jacob, *Off-the-grid recovery of piecewise constant images from few fourier samples*, SIAM J. Imaging Sci. **9** (2016), no. 3, 1004–1041.

[OJ17] G. Ongie and M. Jacob, *A Fast Algorithm for Convolutional Structured Low-Rank Matrix Recovery*, IEEE Trans. Comput. Imag. **3** (2017), no. 4, 535 –550.

[OMFH11] S. Oymak, K. Mohan, M. Fazel, and B. Hassibi, *A simplified approach to recovery conditions for low rank matrices*, Proceedings of the IEEE International Symposium on Information Theory (ISIT), 2011, pp. 2318–2322.

[OS11] V. E. Oropeza and M. D. Sacchi, *Simultaneous seismic data de-noising and reconstruction via multichannel singular spectrum analysis*, Geophysics **76** (2011), no. 3, V25–V32.

[Osb85] M. R. Osborne, *Finite algorithms in optimization and data analysis*, John Wiley & Sons, Inc., New York, NY, USA, 1985.

[PABN16] D. L. Pimentel-Alarcón, N. Boston, and R. D. Nowak, *A Characterization of Deterministic Sampling Patterns for Low-Rank Matrix Completion*, 2016. arXiv preprint arXiv:1503.02596v3 [stat.ML].

[Pan01] V. Y. Pan, *Structured Matrices and Polynomials: Unified Superfast Algorithms*, Springer-Verlag, Berlin, Heidelberg, 2001.

[Pao06] G. Paouris, *Concentration of mass on convex bodies*, Geometric & Functional Analysis GAFA **16** (2006), no. 5, 1021–1049.

[PEPC10] L. C. Potter, E. Ertin, J. T. Parker, and M. Cetin, *Sparsity and Compressed Sensing in Radar Imaging*, Proc. IEEE **98** (2010), no. 6, 1006–1020.

[PKCS18] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, *Finding Low-Rank Solutions via Nonconvex Matrix Factorization, Efficiently and Provably*, SIAM J. Imaging Sci. **11** (2018), no. 4, 2165–2204. [using BFGD ('Bi-Factored Gradient Descent') algorithm, code from http://akyrillidis.github.io/projects/].

[PMR19] S. Paternain, A. Mokhtari, and A. Ribeiro, *A Newton-Based Method for Nonconvex Optimization with Fast Evasion of Saddle Points*, SIAM J. Optim. **29** (2019), no. 1, 343–368.

[PP13] T. Peter and G. Plonka, *A generalized Prony method for reconstruction of sparse sums of eigenfunctions of linear operators*, Inverse Problems **29** (2013), no. 2, 025001.

[PRK93] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, *Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition*, Proceedings of 27th Asilomar Conference on Signals, Systems and Computers, 1993, pp. 40–44.

[PRT13] G. E. Pfander, H. Rauhut, and J. A. Tropp, *The restricted isometry property for time–frequency structured random matrices*, Probability Theory and Related Fields **156** (2013), no. 3, 707–737.

[PSC14a] J. T. Parker, P. Schniter, and V. Cevher, *Bilinear Generalized Approximate Message Passing-Part I: Derivation*, IEEE Trans. Signal Process. **62** (2014), no. 22, 5839–5853.

[PSC14b] ______, *Bilinear Generalized Approximate Message Passing-Part II: Applications*, IEEE Trans. Signal Process. **62** (2014), no. 22, 5854–5867.

[PT10] D. Potts and M. Tasche, *Parameter estimation for exponential sums by approximate Prony method*, Signal Processing **90** (2010), no. 5, 1631–1642.

[PT13] ______, *Parameter estimation for nonincreasing exponential sums by Prony-like methods*, Linear Algebra Appl. **439** (2013), no. 4, 1024–1039.

[PT15] ______, *Fast ESPRIT algorithms based on partial singular value decompositions*, Appl. Numer. Math. **88** (2015), 31–45.

[PTV16] D. Potts, M. Tasche, and T. Volkmer, *Efficient Spectral Estimation by MUSIC and ESPRIT with Application to Sparse FFT*, Frontiers in Applied Mathematics and Statistics **2** (2016), 1.

[Qi94] L. Qi, *Superlinearly convergent approximate Newton methods for LC1 optimization problems*, Math. Program. **64** (1994), no. 1, 277–294.

[QSS10] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical mathematics*, Vol. 37, Springer Science & Business Media, 2010.

[QY03] H. Qi and X. Yang, *Semismoothness of Spectral Functions*, SIAM J. Matrix Anal. Appl. **25** (2003), no. 3, 766–783.

[Rec11] B. Recht, *A Simpler Approach to Matrix Completion*, Journal of Machine Learning Research **12** (2011), 3413–3430.

[RFP10] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Rev. **52** (2010), no. 3, 471–501.

[RK89] R. Roy and T. Kailath, *ESPRIT-estimation of signal parameters via rotational invariance techniques*, IEEE Trans. Acoust., Speech, Signal Process. **37** (1989), no. 7, 984–995.

[RK99] B. D. Rao and K. Kreutz-Delgado, *An affine scaling methodology for best basis selection*, IEEE Trans. Signal Process. **47** (1999), no. 1, 187–200.

[RM51] H. Robbins and S. Monro, *A stochastic approximation method*, The Annals of Mathematical Statistics (1951), 400–407.

[ROF92] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D **60** (1992), no. 1-4, 259–268.

[Rom09] J. Romberg, *Compressive sensing by random convolution*, SIAM J. Imaging Sci. **2** (2009), no. 4, 1098–1128.

[RR13] B. Recht and C. Ré, *Parallel stochastic gradient algorithms for large-scale matrix completion*, Math. Program. Comput. **5** (2013), no. 2, 201–226.

[RRT12] H. Rauhut, J. Romberg, and J. A. Tropp, *Restricted isometries for partial random circulant matrices*, Appl. Comput. Harmon. Anal. **32** (2012), no. 2, 242 –254.

[RSW15] N. Rao, P. Shah, and S. Wright, *Forward–backward greedy algorithms for atomic norm regularization*, IEEE Trans. Signal Process. **63** (2015), no. 21, 5798–5811.

[RU68] J. R. Rice and K. H. Usow, *The Lawson Algorithm and Extensions*, Math. Comput. **22** (1968), no. 101, 118–127.

[RV08] M. Rudelson and R. Vershynin, *On sparse reconstruction from Fourier and Gaussian measurements*, Comm. Pure Appl. Math. **61** (2008), no. 8, 1025–1045.

[RXH11] B. Recht, W. Xu, and B. Hassibi, *Null space conditions and thresholds for rank minimization*, Math. Program. **127** (2011), no. 1, 175–202.

[RYZ18] M. C. Robini, F. Yang, and Y. Zhu, *Inexact half-quadratic optimization for linear inverse problems*, SIAM J. Imaging Sci. **11** (2018), no. 2, 1078–1133.

[RZ15] M. C. Robini and Y. Zhu, *Generic Half-Quadratic Optimization for Image Reconstruction*, SIAM J. Imaging Sci. **8** (2015), no. 3, 1752–1797.

[SB12] P. Stoica and P. Babu, *Spice and likes: Two hyperparameter-free methods for sparse-parameter estimation*, Signal Processing **92** (2012), no. 7, 1580–1590.

[SBP17] Y. Sun, P. Babu, and D. P. Palomar, *Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning*, IEEE Trans. Signal Process. **65** (2017), no. 3, 794–816.

[SCD02] J.-L. Starck, E. J Candés, and D. L Donoho, *The curvelet transform for image denoising*, IEEE Trans. Image Process. **11** (2002), no. 6, 670–684.

[Sch73] E. J. Schlossmacher, *An iterative technique for absolute deviations curve fitting*, J. Am. Stat. Assoc. **68** (1973), no. 344, 857–859.

[Sch86] R. O. Schmidt, *Multiple Emitter Location and Signal Parameter Estimation*, IEEE Trans. Antennas Propag. **34** (1986), no. 3, 276–280.

[SHL10] L. Schermelleh, R. Heintzmann, and H. Leonhardt, *A guide to super-resolution fluorescence microscopy*, J. Cell Biol. **190** (2010), no. 2, 165–175.

[Sig18] J. Sigl, *Iteratively Reweighted Least Squares-Nonlinear Regression and Low-Dimensional Structure Learning for Big Data*, Ph.D. Thesis, Technische Universität München, 2018.

[Sil16] D. Silver, *Mastering the game of Go with deep neural networks and tree search*, Nature **529** (2016), no. 7587, 484.

[SL16] R. Sun and Z. Q. Luo, *Guaranteed Matrix Completion via Non-Convex Factorization*, IEEE Trans. Inf. Theory **62** (2016), no. 11, 6535–6579.

[SLO$^+$14] P. Shin, P. Larson, M. Ohliger, M. Elad, J. Pauly, D. Vigneron, and M. Lustig, *Calibrationless parallel imaging reconstruction based on structured low-rank matrix completion*, Magnetic resonance in medicine **72** (2014), no. 4, 959–970.

[SM05] P. Stoica and R. L. Moses, *Spectral analysis of signals*, Pearson Prentice Hall Upper Saddle River, NJ, 2005.

[SP95] T. K. Sarkar and O. Pereira, *Using the matrix pencil method to estimate the parameters of a sum of complex exponentials*, IEEE Antennas and Propagation Magazine **37** (1995), no. 1, 48–55.

[SRJ05] N. Srebro, J. Rennie, and T. S. Jaakkola, *Maximum-Margin Matrix Factorization*, Advances in Neural Information Processing Systems (NIPS), 2005, pp. 1329–1336.

[SS08] D. Sun and J. Sun, *Löwner's Operator and Spectral Functions in Euclidean Jordan Algebras*, Math. Oper. Res. **33** (2008), no. 2, 421–445.

[SS16] É. Schost and P.-J. Spaenlehauer, *A quadratically convergent algorithm for structured low-rank approximation*, Found. Comput. Math. **16** (2016), no. 2, 457–492.

[SS92] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, 1992.

[Ste06] M. Stewart, *Perturbation of the SVD in the presence of small singular values*, Linear Algebra Appl. **419** (2006), no. 1, 53–77.

[SY10] R. Saab and Ö. Yılmaz, *Sparse recovery by non-convex optimization – instance optimality*, Appl. Comput. Harmon. Anal. **29** (2010), no. 1, 30–48.

[SZ00] H. D. Simon and H. Zha, *Low rank matrix approximation using the Lanczos bidiagonalization process with applications*, SIAM J. Sci. Comput. **21** (2000), no. 6, 2257–2274.

[TBM79] H. L. Taylor, S. C. Banks, and J. F. McCoy, *Deconvolution with the ℓ_1-norm*, Geophysics **44** (1979), no. 1, 39–52.

[TBS⁺16] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, *Low-rank Solutions of Linear Matrix Equations via Procrustes Flow*, International Conference on Machine Learning (ICML), 2016, pp. 964–973.

[Tib96] R. Tibshirani, *Regression Shrinkage and Selection Via the Lasso*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **58** (1996), no. 1, 267–288.

[TLD⁺10] J. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, *Beyond Nyquist: Efficient sampling of sparse bandlimited signals*, IEEE Trans. Inf. Theory **56** (2010), no. 1, 520–544.

[TLRW14] L. Tian, X. Li, K. Ramchandran, and L. Waller, *Multiplexed coded illumination for fourier ptychography with an led array microscope*, Biomed. Opt. Express **5** (2014), no. 7, 2376–2389.

[TLT18] X. Tian, J. R Loftus, and J. E Taylor, *Selective inference with unknown variance via the square-root lasso*, Biometrika **105** (2018), no. 4, 755–768.

[TTT99] K.-C. Toh, M. J Todd, and R. H Tütüncü, *SDPT3-a MATLAB software package for semidefinite programming, version 1.3*, Optim. Method. Softw. **11** (1999), no. 1-4, 545–581.

[TW13] J. Tanner and K. Wei, *Normalized Iterative Hard Thresholding for Matrix Completion*, SIAM J. Sci. Comput. **35** (2013), no. 5, S104–S125.

[TW16] ______, *Low rank matrix completion by alternating steepest descent methods*, Appl. Comput. Harmon. Anal. **40** (2016), no. 2, 417–429. [using `ASD` ('Alternating Steepest Descent') algorithm, code from https://www.math.ucdavis.edu/~kewei/publications.html].

[TWH15] R. Tibshirani, M. Wainwright, and T. Hastie, *Statistical learning with sparsity: the lasso and generalizations*, Chapman and Hall/CRC, 2015.

[Van13] B. Vandereycken, *Low-Rank Matrix Completion by Riemannian Optimization*, SIAM J. Optim. **23** (2013), no. 2, 1214–1236. [using `Riemann_Opt` ('Riemannian Optimization') algorithm, code from http://www.unige.ch/math/vandereycken/matrix_completion.html].

[VD17] S. Voronin and I. Daubechies, *An iteratively reweighted least squares algorithm for sparse regularization*, Functional analysis, harmonic analysis, and image processing, 2017, pp. 391–411.

[Ver12] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, Compressed sensing, 2012, pp. 210–268.

[Ver18] ______, *High-dimensional probability: An introduction with applications in data science*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2018.

[VMB02] M. Vetterli, P. Marziliano, and T. Blu, *Sampling signals with finite rate of innovation*, IEEE Trans. Signal Process. **50** (2002), no. 6, 1417–1428.

[Vor12] S. Voronin, *Regularization of linear systems with sparsity constraints with applications to large scale inverse problems*, Ph.D. Thesis, Princeton University, 2012.

[VZ00] Y. Vardi and C.-H. Zhang, *The multivariate l1-median and associated data depth*, Proceedings of the National Academy of Sciences **97** (2000), no. 4, 1423–1426.

[Wai19] M. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, Vol. 48, Cambridge University Press, 2019.

[Wat01] G. A. Watson, *Approximation in normed linear spaces*, Numerical analysis: Historical developments in the 20th century, elsevier, 2001, pp. 41–76.

[Wat77] ______, *On two methods for discrete Lp-approximation*, Computing **18** (1977), no. 3, 263–266.

[WCCL16a] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung, *Guarantees of Riemannian Optimization for Low Rank Matrix Recovery*, SIAM J. Matrix Anal. Appl. **37** (2016), no. 3, 1198–1222.

[WCCL16b] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung, *Guarantees of Riemannian optimization for low rank matrix completion*, arXiv preprint arXiv:1603.06610 (2016).

[Wed72] P.-Å. Wedin, *Perturbation bounds in connection with singular value decomposition*, BIT **12** (1972), no. 1, 99–111.

[Wei37] E. Weiszfeld, *Sur le point pour lequel la somme des distances de n points donnés est minimum*, Tohoku Mathematical Journal, First Series **43** (1937), 355–386.

[WM17] W. Wu and K. L. Miller, *Image formation in diffusion MRI: a review of recent technical developments*, J. Magn. Reson. Imaging **46** (2017), no. 3, 646–662.

[WN10] D. Wipf and S. Nagarajan, *Iterative Reweighted Methods for Finding Sparse Solutions*, IEEE J. Sel. Topics Signal Process. **4** (2010), no. 2, 317–329.

[Woj10] P. Wojtaszczyk, *Stability and instance optimality for Gaussian measurements in compressed sensing*, Found. Comput. Math. **10** (2010), no. 1, 1–13.

[Woj12] ______, *ℓ_1 Minimization with Noisy Data*, SIAM J. Numer. Anal. **50** (2012), no. 2, 458–467.

[Woo50] M. A. Woodbury, *Inverting modified matrices*, Memorandum report **42** (1950), no. 106, 336.

[WP09] E. Weiszfeld and F. Plastria, *On the point for which the sum of the distances to n given points is minimum*, Ann. Oper. Res. **167** (2009Mar), no. 1, 7–41.

[WYZ12] Z. Wen, W. Yin, and Y. Zhang, *Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm*, Math. Program. Comput. **4** (2012), no. 4, 333–361.

[XSH$^+$18] R. Xu, M. Soltanolkotabi, J. P Haldar, W. Unglaub, J. Zusman, A. F. Levi, and R. M Leahy, *Accelerated wirtinger flow: A fast algorithm for ptychography*, arXiv preprint arXiv:1806.05546 (2018).

[Xu18] Z. Xu, *The minimal measurement number for low-rank matrix recovery*, Appl. Comput. Harmon. Anal. **44** (2018), no. 2, 497 –508.

[XW15] B. Xin and D. Wipf, *Pushing the limits of affine rank minimization by adapting probabilistic PCA*, International Conference on Machine Learning (ICML), 2015, pp. 419–427.

[Yan09] Z. Yang, *A study on nonsymmetric matrix-valued functions*, 2009. Master's thesis, Department of Mathematics, National University of Singapore, https://www.polyu.edu.hk/ama/profile/dfsun/Main_YZ.pdf.

[YJ92] X. Q. Yang and V. Jeyakumar, *Generalized second-order directional derivatives and optimization with C1,1 functions*, Optimization **26** (1992), no. 3-4, 165–185.

[YK15] Q. Yao and J. Kwok, *inexact soft-impute for fast large-scale matrix completion*, Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI), 2015.

[YKJL17] J. C. Ye, J. M. Kim, K. H. Jin, and K. Lee, *Compressive Sampling Using Annihilating Filter-Based Low-Rank Interpolation*, IEEE Trans. Inf. Theory **63** (2017), no. 2, 777–801.

[YLSX17] Z. Yang, J. Li, P. Stoica, and L. Xie, *Sparse Methods for Direction-of-Arrival Estimation*, Academic Press Library in Signal Processing, 2017, pp. 509–581.

[YM18] N. Yair and T. Michaeli, *Multi-scale weighted nuclear norm image restoration*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3165–3174.

[YX16] Z. Yang and L. Xie, *Enhancing Sparsity and Resolution via Reweighted Atomic Norm Minimization*, IEEE Trans. Inf. Theory **64** (2016), no. 4, 995–1006.

[Zha11] T. Zhang, *Sparse Recovery With Orthogonal Matching Pursuit Under RIP*, IEEE Trans. Inf. Theory **57** (2011), no. 9, 6215–6221.

[Zie95] G. M. Ziegler, *Lectures on polytopes*, Vol. 152, Springer-Verlag, New York, 1995. seventh updated printing 2007.

[ZL12] Y.-B. Zhao and D. Li, *Reweighted ℓ_1-minimization for sparse solutions to underdetermined linear systems*, SIAM J. Optim. **22** (2012), no. 3, 1065–1088.

[ZL16a] Q. Zheng and J. Lafferty, *Convergence Analysis for Rectangular Matrix Completion Using Burer-Monteiro Factorization and Gradient Descent*, arXiv preprint arXiv:1605.07051 (2016).

[ZL16b] Z. A. Zhu and Y. Li, *Even faster SVD decomposition yet without agonizing pain*, Advances in Neural Information Processing Systems (NIPS), 2016, pp. 974–982.

[ZLTW18] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, *Global Optimality in Low-Rank Matrix Optimization*, IEEE Trans. Signal Process. **66** (2018), no. 13, 3614–3628.

[ZMW$^+$17] L. Zheng, A. Maleki, H. Weng, X. Wang, and T. Long, *Does ℓ_p -Minimization Outperform ℓ_1 -Minimization?*, IEEE Trans. Inf. Theory **63** (2017), no. 11, 6896–6935.

[ZWSP08] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, *Large-Scale Parallel Collaborative Filtering for the Netflix Prize*, Algorithmic Aspects in Information and Management, 2008, pp. 337–348.