Technische Universität München

Fakultät für Mathematik

Lehrstuhl für Angewandte Numerische Analysis und Optimierung und Data Analysis

Iteratively Reweighted Least Squares -
Nonlinear Regression and Low-Dimensional Structure Learning for Big Data

Juliane Sigl

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.

Vorsitzende: Prof. Dr. Nina Gantert

Prüfer der Dissertation:

1. Prof. Dr. Massimo Fornasier
2. Prof. Dr. Sergei Pereverzyev
   Johann Radon Institute for Computational and Applied Mathematics (RICAM), Linz (Österreich)
3. Prof. Rachel Ward, PhD
   University of Texas at Austin, Austin (USA)
   (schriftliche Beurteilung)

Die Dissertation wurde am 28.11.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 10.03.2018 angenommen.

Thanks and Acknowledgements

In a first place, I would like to thank my supervisor Prof. Dr. Massimo Fornasier for his strong confidence in my abilities and his constant encouragement of my work and ideas. His guidance and generous support during this years influenced my development not only on a scientific but also on a personal level.

My research was supported by the DFG-project no. 254193214 “Multi-parameter regularization in high-dimensional learning”, which I gratefully acknowledge. Additionally, I am thankful for the opportunity to participate in international research training group IGDK Munich-Graz as an associate member.

Furthermore, I want to thank Patrick Lenz and Martin Ehler for the successful scientific collaboration, as well as Felix Krahmer, Valeriya Naumova, Johannes Maly, Philipp Walk and Francesco Solombrino for the helpful discussions. In this regard, I am also grateful to Henrik Ohlsson, Gerlinde Plonka-Hoch and Sergei Pereverzyev for their invitations to research visits, respectively, their welcoming hospitality and for sharing with me their expertise. Thanks also to Ricardo Acevedo Cabra for our fruitful collaboration in the context of the TUM Data Innovation Lab.

Sincerely, I would like to express my gratitude to Christian Kümmerle for the continuous exchange of ideas, the technical discussions and feedback, which turned out to be very profitable for our joint work.

Moreover, I would like to thank the old M15 Italian/German-connection Marco, Francesco, Giacomo and Steffen bringing Italian vitality (including coffee) to Munich and celebrating traditional Bavarian culture. Special thanks go to Mattia for being the best office mate one could wish for.

Great thanks also to Frank for his support and encouragement making him an inofficial mentor for me and for our vivid exchange of cookies and Tupperware boxes over all

the years.

I also owe thanks to the new M15-generation and the coffee crew across M-units for the funny moments and their good company in breaks for lunch, coffee or inbetween.

Ich möchte mich bei allen lieben Menschen bedanken, die mich in den letzten Jahren begleitet und unterstützt haben, die mich in schwierigen Phasen motiviert haben und die guten Momente mit mir gefeiert haben.

Diesen Punkt, an dem ich stehe, hätte ich niemals erreicht ohne meine guten Freunde Michi, Katha, Baschi, Nati, Joe, Sandi, Rüdy, Stefan, Till, Tobi, Barbara und viele mehr. Ich danke euch für die unvergessliche Zeit, die wir zusammen hatten und dafür, dass ihr München zu meinem Zuhause gemacht habt.

Ein großes Danke verdient auch Eva für die außergewöhnlichen Momente, die vielen Gespräche und einfach dafür, dass sie seit vielen Jahren eine so gute Freundin für mich ist.

Danke auch an mein Minga, mit allen meinen Lieblingsecken. München ist für mich keine Stadt sondern ein wichtiges Kapitel meines Lebens.

Meiner Familie möchte ich ein besonderes Dankeschön sagen: meiner Oma Fanne für vieles und Apfelstrudel, meinem ersten besten Freund und Opa Hans, meinen Eltern Franz und Rosi dafür, dass ich immer uneingeschränkt auf sie vertrauen kann und sie mir immer vertraut haben, dass ich den für mich richtigen Weg gehe, meiner Schwester Conny, die immer für mich da ist und ich für sie.

Danke an Christoph, dass er mit mir durchgehalten, mich ausgehalten und immer zu mir gehalten hat. Er ermutigt mich, Ziele zu setzen, die Zeit zu nutzen, Grenzen zu verschieben und gemeinsam mehr zu sehen und zu erleben.

Abstract - Zusammenfassung

Abstract

In this thesis, we develop and investigate new computational methods for data analytical approximation problems belonging to the family of iteratively reweighted least squares (IRLS) algorithms. First, we discuss the applicability of IRLS-methods in regression problems with nonlinear measurement settings entailing nonconvex or even nonsmooth optimization problems. Next, we introduce an IRLS-variant with a novel reweighting strategy for learning low-rank matrices from few random measurements that substantially enhances performance with respect to state-of-the-art methods. Finally, we present an IRLS-algorithm with a very general formulation allowing for learning signals with multiple or composed low-dimensional structures from a minimal number of measurements.

Zusammenfassung

Diese Arbeit befasst sich mit der Entwicklung und Untersuchung von Berechnungsverfahren für Approximationsprobleme im Bereich der Datenanalyse, die zur Familie der Iteratively Reweighted Least Squares (IRLS)-Methoden gehören. Als Erstes wird die Anwendung von IRLS-Methoden auf Regressionsprobleme diskutiert, bei denen der zugrunde liegende Messprozess nichtlinear ist, was zu nichtkonvexen oder sogar nichtglatten Optimierungsproblemen führt. Anschließend wird eine Variante eines IRLS-Algorithmus mit einer neuartigen Gewichtungsstrategie für das Lernen von Niedrigrangmatrizen aus wenigen Zufallsmessungen vorgestellt, die im Vergleich zu state-of-the-art-Methoden wesentliche Performanceverbesserungen aufweist. Abschließend präsentieren wir eine IRLS-Methode mit sehr allgemeiner Formulierung, die das Lernen von Signalen mit mehreren oder zusammengesetzten Strukturen aus einer minimalen Anzahl von Messungen erlaubt.

Contents

Notation **1**

1 Introduction **5**

2 En route to iteratively reweighted least squares: From data sets to optimization problems **15**

2.1 Data analysis and approximation problems 15

2.1.1 About data sets and approximation problems 16

2.1.2 Least squares: examples for error and penalization functions . . 18

2.1.3 Generalizing the example: weighted least squares 21

2.2 Nonlinear optimization problems and surrogate functionals 24

2.2.1 Prerequisites from optimization 24

2.2.2 Optimality conditions and a detour to nonlinear least squares problems 29

2.2.3 Optimization of surrogate functionals 37

2.3 High-dimensional data: fighting the curse of dimensionality 40

2.3.1 Low-dimensional subspace structures 41

2.3.2 An introduction: sparse vectors and related concepts for matrix valued data 43

2.3.3 Theoretical foundations of vector and matrix valued compressed sensing 53

2.4 Iteratively reweighted least squares methods in data analysis 61

2.4.1 General concept of iteratively reweighted least squares 62

2.4.2 `IRLS` for sparse vector recovery and matrix valued signals 65

3 Nonlinear residual minimization via IRLS 71

3.1 Auxiliary functional and nonlinear residual iteratively reweighted least squares algorithm 74

3.2 Theoretical analysis and convergence results for `NR-IRLS` 76

3.2.1 Preliminary results 79

3.2.2 Convergence and error decay rates 82

3.3 Local convexification of the auxiliary functional 87

3.4 Convergence analysis for the convexified algorithm 90

3.4.1 Preliminary results 90

3.4.2 Convergence 94

3.5 Numerical experiments 96

3.5.1 Visually accessible example 96

3.5.2 High dimensional examples in a nonlinear compressed sensing application context 98

3.5.3 Recovery from data with impulsive noise perturbation 106

4 Harmonic mean IRLS for low-rank matrix recovery 109

4.1 Towards a harmonic mean weight matrix 113

4.1.1 Kronecker and Hadamard products 113

4.1.2 Averaging of weight matrices 115

4.2 Harmonic mean iteratively reweighted least squares algorithm 118

4.3 Theoretical analysis 121

4.3.1 Optimization of the auxiliary functional with respect to W . . . 122

4.3.2 Optimization of the auxiliary functional with respect to X . . . 128

4.3.3 Basic properties of the algorithm 130

4.3.4 Strong Schatten-p-null space property 134

4.3.5 Convergence results 136

4.3.6 Locally superlinear convergence 139

4.3.7 Discussion and comparison with existing IRLS algorithms 150

4.4 Numerical experiments 152

4.4.1 Measurement setting 152

4.4.2 Convergence rate comparison with other IRLS type algorithms . 153

4.4.3 Recovery performance comparison with state-of-the-art algorithms158

4.4.4 Computational complexity 161

5 Generalized IRLS for recovery of matrices with multiple structures 163

5.1 Auxiliary functional and generalized `IRLS` algorithm for structured matrices 167

5.2 Algorithm formulation for the constrained case 169

5.3 Theoretical analysis and convergence results for the constrained case . 171

5.3.1 Unified presentation of properties for different sparsity structures 171

5.3.2 Preliminary results 172

5.3.3 Convergence results 178

5.4 Algorithm formulation for the unconstrained case 184

5.5 Theoretical analysis and convergence results for the unconstrained case 186

5.5.1 Preliminary results 186

5.5.2 Convergence results 188

Literature 204

Notation

Sets and operations on sets:

$\emptyset$	empty set
Λ, Λ^c	set and its complement
$[n]$	set of natural numbers $\{1, \ldots, n\}$
$\mathbb{N}$	set of natural numbers
$\mathbb{N}_0$	set of natural numbers including 0
$\mathbb{R}$	set of real numbers
$\mathbb{R}_+$	set of positive real numbers
$\mathbb{C}$	set of complex numbers
$B_\varepsilon(x)$	ball with radius $\varepsilon > 0$ around x w.r.t. the Euclidean norm
$B_{\lvert\lvert\lvert\cdot\rvert\rvert\rvert}(x, \varepsilon)$	ball with radius $\varepsilon > 0$ around x w.r.t. the norm $\lvert\lvert\lvert\cdot\rvert\rvert\rvert$

Functions and operations on functions:

In the following, let $F : \mathbb{R}^d \to \mathbb{R}$, and $\varphi : \mathbb{R}^d \to \mathbb{R}^N$ be arbitrary functions.

dom F, domφ	domain of the function F, domain of the function ϕ
$\nabla F(x)$, $\nabla^2 F(x)$	the gradient and the Hessian of F at x
$\partial F(x)$	subdifferential of F at x
$\ell_{F,\mathcal{C}}(c)$	level set of F on $\mathcal{C}$ corresponding to the value c, i.e., $$\ell_{F,\mathcal{C}}(c) = \{x \in \mathcal{C} : F(x) \leq c\}\,.$$

In the following, consider the vectors $x \in \mathbb{R}^d, y \in \mathbb{R}^m$ and the matrices $X \in \mathbb{R}^{d_1 \times d_2}$ and an operator $\Phi : \mathbb{R}^d \to \mathbb{R}^m$

Vectors, matrices and operators:

$M_{d_1 \times d_2}$	set of matrices in $\mathbb{R}^{d_1 \times d_2}$
$X_{[\mathcal{I},\mathcal{J}]}$	submatrix of X w.r.t. the index sets $\mathcal{I} \subset \{1, ..., m\}$, $\mathcal{J} \subset \{1, ..., m\}$
$X_{[i,\cdot]}$, $X_{[\cdot,j]}$	i-th row and j-th column of X
$(F(x_i))_{i=1}^d$	$(F(x_i))_{i=1}^d = [F(x_1), \ldots, F(x_d)]^T$
$\mathbf{I}_d$	identity matrix in dimension $d \times d$
$\mathbf{0}_{d_1 \times d_2}$	$d_1 \times d_2$-matrix with only 0-entries
$\mathbf{1}_{d_1 \times d_2}$	$d_1 \times d_2$-matrix with only 1-entries
$\mathrm{diag}(x)$	diagonal matrix with $\mathrm{diag}(x)_{[ij]} = \begin{cases} x_i, i \in [d] & i = j \\ 0 & i \neq j \end{cases}$
$\mathbb{S}^d$	set of symmetric, real $d \times d$ matrices
$\mathbb{S}_+^d$	set of symmetric, real, positive semidefinite $d \times d$-matrices
$\mathbb{S}_{++}^d$	set of symmetric, real, and positive definite $d \times d$-matrices
$\mathbb{H}^d$	set of Hermitian $d \times d$-matrices
$\mathbb{H}_+^d$	set of Hermitian and positive semidefinite $d \times d$-matrices
$\mathbb{H}_{++}^d$	set of Hermitian and positive definite $d \times d$-matrices
O_d	set of $d \times d$-orthonormal matrices, i.e., $\{O \in M_{d \times d} \vert OO^T = \mathbf{I}_d\}$
$O_{d_1 \times d_2}$	set of $d_1 \times d_2$-matrices with orthonormal columns, i.e., $\{O \in M_{d_1 \times d_2} \vert O^T O = \mathbf{I}_{d_2}\}$
U_d	set of $d \times d$-unitary matrices, i.e., $\{U \in \mathbb{C}^{d \times d} \vert UU^* = \mathbf{I}_d\}$
$U_{d_1 \times d_2}$	set of $d_1 \times d_2$-matrices with unitary columns, i.e., $\{U \in \mathbb{C}^{d_1 \times d_2} \vert U^* U = \mathbf{I}_{d_2}\}$
$\mathcal{D}(\cdot)$	domain of an operator
$\mathrm{Ran}(\cdot)$	range of an operator
$\mathcal{N}(\cdot)$	nullspace of an operator
$\mathcal{F}(\cdot, \cdot)$	set of elements in the domain that are mapped on y by Φ, i.e., $\mathcal{F}(y, \Phi) = \left\{z \in \mathbb{R}^d \vert \Phi(z) = y\right\}$

In the following, let the vectors $x, y \in \mathbb{R}^d, z \in \mathbb{R}^{d_1 d_2}$ and matrices $A \in \mathbb{R}^{d \times d_1}$, $B \in \mathbb{R}^{d \times d_2}$, $C, D \in \mathbb{R}^{d \times d}$, $E \in \mathbb{R}^{d_1 \times d_1}$ and $F \in \mathbb{R}^{d_2 \times d_2}$

Operations on vectors and matrices and operators:

$(\cdot)^T$	transpose of a vector or matrix
$(\cdot)^*$	conjugate transpose of a vector or matrix, adjoint operator
$(\cdot)^{-1}$, $(\cdot)^\dagger$	inverse of a square matrix, Moore-Penrose pseudo inverse
$\mathbf{r}(\cdot)$	non-increasing rearrangement of the vector entries w.r.t. their absolute values, i.e., $\mathbf{r}_1(x) \geq \mathbf{r}_2(x) \geq \cdots > 0$
$\langle \cdot, \cdot \rangle_{\ell_2}$, $\| \cdot \|_{\ell_2}$	Euclidean inner product and the Euclidean norm, i.e., $\langle x, y \rangle_{\ell_2} = \sum_i x_i y_i$, $\|x\|_{\ell_2} = (\sum_i |x_i|^2)^{\frac{1}{2}}$
$\langle \cdot, \cdot \rangle_W$, $\| \cdot \|_{\ell_2(W)}$	weighted Euclidean inner product, i.e., $\langle x, y \rangle_W = \langle x, Wy \rangle$ for a weight matrix $W \in \mathbb{S}_+^{d \times d}$ and the induced norm
$\| \cdot \|_{\ell_p}$	ℓ_p-vector (quasi)norm, $0 < p \leq \infty$, i.e., $\|x\|_{\ell_p} = (\sum_i |x_i|^p)^{\frac{1}{p}}$
$\mathrm{tr}(\cdot)$	trace of a matrix, i.e., $\mathrm{tr}(A) = \sum_i A_{i,i}$
$\sigma_i(\cdot)$	the i-th singular values of a matrix
$\mathrm{rank}(\cdot)$	rank of a matrix
$\langle \cdot, \cdot \rangle_F$, $\| \cdot \|_F$,	Frobenius inner product $\langle X, Y \rangle_F = \mathrm{tr}\,[X^* Y]$ and induced Frobenius-norm of a matrix
$\langle \cdot, \cdot \rangle_W$, $\|\cdot\|_{F(W)}$	weighted Frobenius inner product, i.e., $\langle A, B \rangle_W = \langle A, WB \rangle$ for a weight matrix $W \in \mathbb{S}_+^{d \times d}$ and the induced norm
$\| \cdot \|_{S_p}$	Schatten-p-matrix (quasi)norm, $0 < p \leq \infty$, i.e., $\|A\|_{S_p} = (\sum_i \sigma_i(A)^p)^{\frac{1}{p}} = \left(\mathrm{tr}(XX^T)\right)^{\frac{p}{2}}$
$\| \cdot \|_{\ell_{p,q}}$	$\ell_{p,q}$-matrix (quasi) norm, $0 < p \leq \infty$, $1 \leq q$, i.e., $\|A\|_{\ell_{p,q}} = \left(\sum_i \|A_{[i,\cdot]}\|_{\ell_p}^q \right) \frac{1}{q}$
$\succ$, $\succeq$	partial ordering on the space $\mathbb{S}^{d \times d}$, i.e., $D \succ C \;:\Leftrightarrow\; D - C \in \mathbb{S}_{++}^{d \times d}, \quad D \succeq C \;:\Leftrightarrow\; D - C \in \mathbb{S}_+^{d \times d}$
$\otimes$	Kronecker product, i.e., the tensor product of matrices w.r.t. the standard bases
$\oplus$	Kronecker sum, i.e., $E \oplus F = E \otimes \mathbf{I}_{d_2} + \mathbf{I}_{d_1} \otimes F$
$\odot$	Hadamard product, i.e., $(C \odot D)_{(i,j)} = C_{(i,j)} \cdot D_{(i,j)}$
$(\cdot)_{\mathrm{vec}}$	vectorization of a matrix by stacking its columns in a vector
$(\cdot)_{\mathrm{mat}(d_1,d_2)}$,$(\cdot)_{\mathrm{mat}}$	reshaping a vector of length $d_1 d_2$ into a $d_1 \times d_2$-matrix, i.e., $z_{\mathrm{mat}(d_1,d_2)} = Z$, where $Z_{[i,j]} = z_{(j-1)\cdot d_1 + i}$; $Z = (Z_{\mathrm{vec}})_{\mathrm{mat}}$

Chapter 1
Introduction

The immersion into new environmental conditions confronts us with the interpretation of so far not experienced, maybe initially incomprehensible information obtained from sensory impressions. In finding our way in the world around us - capturing new concepts, learning to speak a language or deducing causalities - in a quite quick and efficient manner, our brain is able to find structure and meaning in these streams of incoming signals. Our minds have the impressive capability to make inferences which apparently reach far beyond the available data helping us to predict and prepare ourselves for future actions [147, 157]. How do we do it?

Just consider the situation of a young child trying to decipher the meanings of new words. Parents experience and scientists confirm [10, 167] that average 2-year-old infants can acquire the proper usage of an unknown word such as “dog” or “chair” from facing few examples only. It is possible for them to capture the meaning, not only the phonetic pronunciation by generalization and appropriately utilize the new word in unfamiliar situations. This is indeed a remarkable feat considering this as a computational result from very limited sensory input data. Imagining the infinite space of all possible items, there exist still infinite but substantially constrained subsets of objects belonging to the categories “dog” or “chair” [147]. This rises the question, how a child is able to capture the boundaries of these subsets from the observation of a handful examples.

It is basic statistics knowledge that correlation does not imply causation. However, on a regular basis, little children infer causal relationships from only a small number of samples [67] that is by far too low to even reliably establish a correlation. Consequently, there is no chance to draw exact inference for arbitrary models neither for the brain nor any type of computer. Nevertheless, humans are capable to derive complex causalities, formulate strong generalizations, and establish powerful abstractions from data that is insufficient, noisy and corrupted with outliers or ambiguous - in many aspects very limited leading to a significant discrepancy between the level of available sensory information and the level of insights and cognition gained from it [147].

This conundrum became known as "the problem of induction", that concerned great philosophers for ages, from ancient Greek Plato and Aristotle, over Hume to Carnap, Quine, Goodman and later in the 20th century [61]. Solutions for this problem proposed by philosophy did not change much in the course of time since Plato: The brain can defeat this "curse of dimensionality" [6] by invoking the "blessing of abstraction" [66, 120]. This means that, if the brain can reach beyond the data provided, this gap has to be closed by another source of information. Our mind places assumptions on the world around us constituting limitations what can be meaningfully represented, manipulated or learned in general - the central statement of the "no free lunch theorem" [120, 165]. In other words, sensible generalization is not feasible without further abstract background knowledge generating and restricting the brain's hypotheses [132]. Different scientific disciplines in mathematics have come up with specific terms for this additional information that is invoked: optimization experts speak of "constraints", machine learning researchers call it "inductive bias" and statisticians name it "priors"[147].

It is a common hypothesis that our brain builds up a simplified model of the world by a-priori assuming that not every input and output variable and all relations between them are relevant. Instead, only a small number of important interactions exists. From a computational perspective, such an extraction of these relevant variables and their relationship can be regarded as learning of these underlying structures with reduced complexity. More precisely, lower-dimensional structures represent constraints on the input-output map which are appropriate for the natural world. Exploiting these representations lets our brain reveal the hidden structure in our experienced phenomenons and allows drawing inferences and conclusions - forming the basis for human intelligence[13, 147].

These days researchers and scientists want to take a step further by creating artificial intelligence (AI), enabling machines to learn from data reaching human or even superhuman level. Already today AI is changing our lives and advancing rapidly in many areas from business over healthcare to sciences. Actually, AI and machine learning are omnipresent and facilitating our day-to-day life, often in a rather obvious way, by web search with all-knowing Google, offering assistance via Siri or Alexa, using our smartphone that can recogonize your face and soon autonomously driving our cars and operating our smart home devices by remote control. Beyond that, less directly visible, machine learning is now behind any major service, detecting credit card fraud or spam mails, making purchase suggestions on Amazon, breaking down language barriers with translation tools and smart non-player elements in games.

Similar to the exploitation of inherent structure with reduced complexity in sensory information in our brains, also machine learning methods take advantage of low-dimensional representations of incoming signals or the output data.

In this thesis, we are concerned with the design and analysis of algorithms tailored for data analytical problems involving input signals or output variables with a specific type of inherent structures. These particular structures allow for so-called sparse representations with reduced dimensionality that we exploit for the computational efficient treatment of the considered data analysis problems.
Let us discuss a few interesting application examples for illustration.

(i) **Detection of faulty sensors in wireless sensor networks**
We consider a wireless sensor network as in Figure 1.1 that contains several but few faulty, compromised, or jammed sensors either sending no meaningful or no signal at all to a signal receiver collecting the incoming measurements. Hence, in such cases, sparse measurement outliers occur and it would be advantageous to smartly detect and ignore the erroneous information transmitted by the damaged sensors. The exploitation of this structural a-priori knowledge on the measurement error is essential for the robust regression of wireless sensor data contaminated by faulty sensors and allowing their detection and neglection[107].

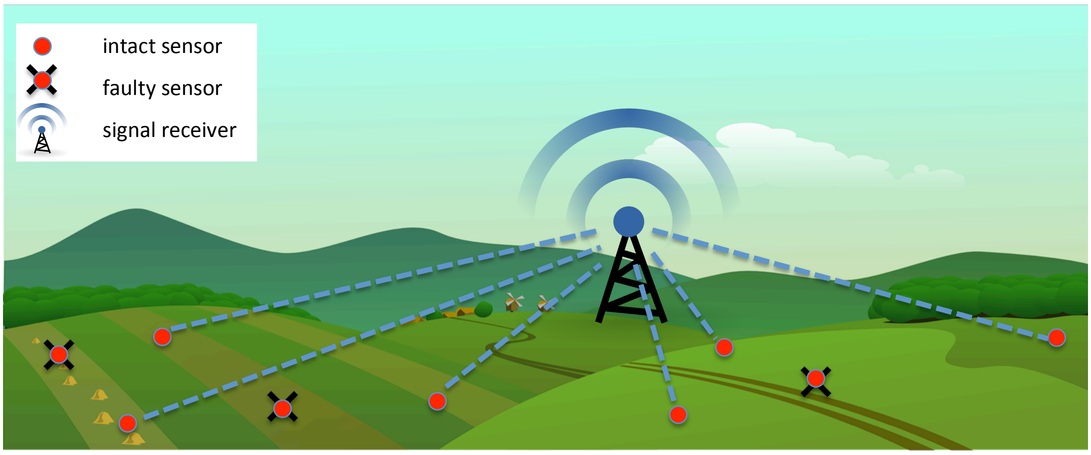


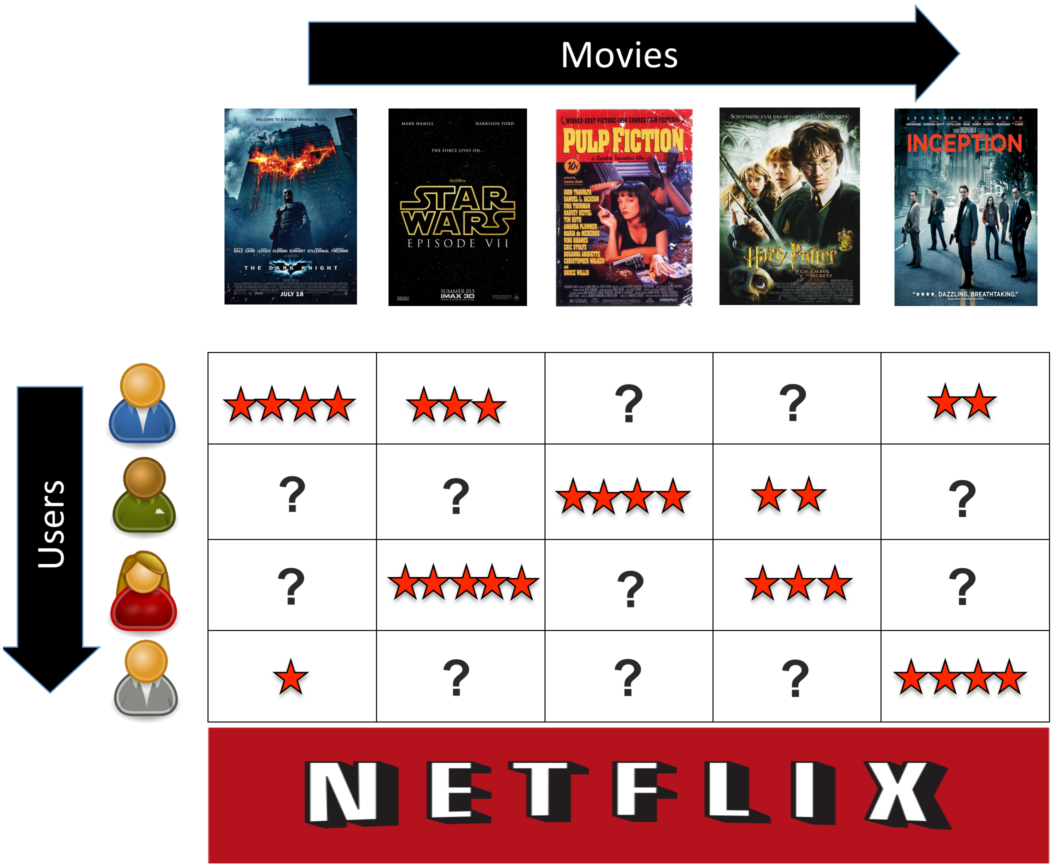
Figure 1.1: Visualization of a wireless sensor network with faulty sensors (graphic by Juliane Sigl, using pictures under creative commons licensing allowing modifications [1])

[1] $https://pixabay.com/en/landscape - countryside - fields - nature - 409551/$, $https://commons.wikimedia.org/wiki/File\%3AWifi.svg$, By RRZEicons (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

(ii) **Netflix movie recommendations**

Recommendation engines are very powerful personalization tools that online shopping retailers, for instance Amazon, and online movie services like Netflix use to provide relevant content for the customer from a potentially overwhelming set of choices. In the words of Steve Jobs - "a lot of times, people don't know what they want until you show it to them" [62, 127].

In 2009, Netflix selected the winner of its famous contest, the so-called Netflix prize, for the best algorithm able to accurately predict customers ratings of movies based on those of previously seen movies [112] as illustrated in Figure 1.2.



★★★★	★★★	?	?	★★
?	?	★★★★	★★	?
?	★★★★★	?	★★★	?
★	?	?	?	★★★★

Figure 1.2: Schematic representation of Netflix's data base with user's movie ratings (graphic by Juliane Sigl, using pictures under creative commons licensing allowing modifications [2])

[2] https://commons.wikimedia.org/wiki/File:User_icon_1.svg, https://commons.wikimedia.org/wiki/File:User_icon_2.svg, https://commons.wikimedia.org/wiki/File:User_icon_3.svg, https://commons.wikimedia.org/wiki/File:Emblem-person-grey.svg, https://commons.wikimedia.org/wiki/File:Netflix_logo.svg, https://www.flickr.com/photos/22284790@N06/2443224485, https://www.flickr.com/photos/jonatasciccone/8180401101, https://www.flickr.com/photos/krisolin/6859796173, https://www.flickr.com/photos/methodshop/5245100040, https://www.flickr.com/photos/colinzhu/49684002

In fact, movies in the Netflix data base can be clustered according to genre, e.g., Comedy, Sci-Fi etc. or they star the same actor or actress and particular user groups may identify their movie taste with certain of those categorizable features. The basic assumption of Netflix is that, if the previous movie ratings of user A and B have been similar in the past, they can be assigned to a subset of users with matching movie taste. Based on the assumptions that the number of categories of movies is way smaller than the total number of available films and the overall count of customers is far larger than the number of groups of users with matching taste, Netflix builds recommendations using similarity scores.

This representation of the mass of customers and movies in preference groups with reduced dimension is another instance of sparsity structures: the Netflix data base corresponds to a user-movie rating matrix with a large number of missing entries and, due to the correlations between the inherent user and movie groups, this matrix has low rank, i.e., the vector of its singular values is sparse.

This problem is called *(low-rank) matrix completion* and has received much attention in recent years as there is a high demand for efficient solvers also for large scale data sets in practical contexts [7]. Further examples of relevant, related application scenarios, in which the recovery of low rank matrices occurs are system identification [153] or global positioning in sensor networks [21].

(iii) **Moving object detection and background separation in videos**
One of the classical computational tasks in video surveillance is the detection of objects in front of detailed background as exemplary shown in Figure 1.3. Given a sequence of video frames, we want to identify objects and activities that stand out from the background, which is difficult in busy video scenes, where every frame may include perturbations.

If we stack all video frames as the columns of a large matrix X, due to the correlation between frames, the stationary background will correspond to highly similar columns, i.e., the matrix showcases approximate low-rankness. In general, moving foreground objects, such as vehicles or pedestrians, constitute only a small fraction of the pixels in the image frames and, therefore, can be treated as additive sparse perturbation errors. Nevertheless, each single image frame has thousands of pixels, and a video fragment consists of hundreds or thousands of frames. Consequently, without efficient computational solution methods, the decomposition of X in a low-rank and a sparse perturbation component is infeasible[19].

Another interesting application problem, where a similiar solution strategy can be used, is face recognition, e.g., considering varying illumination situations.

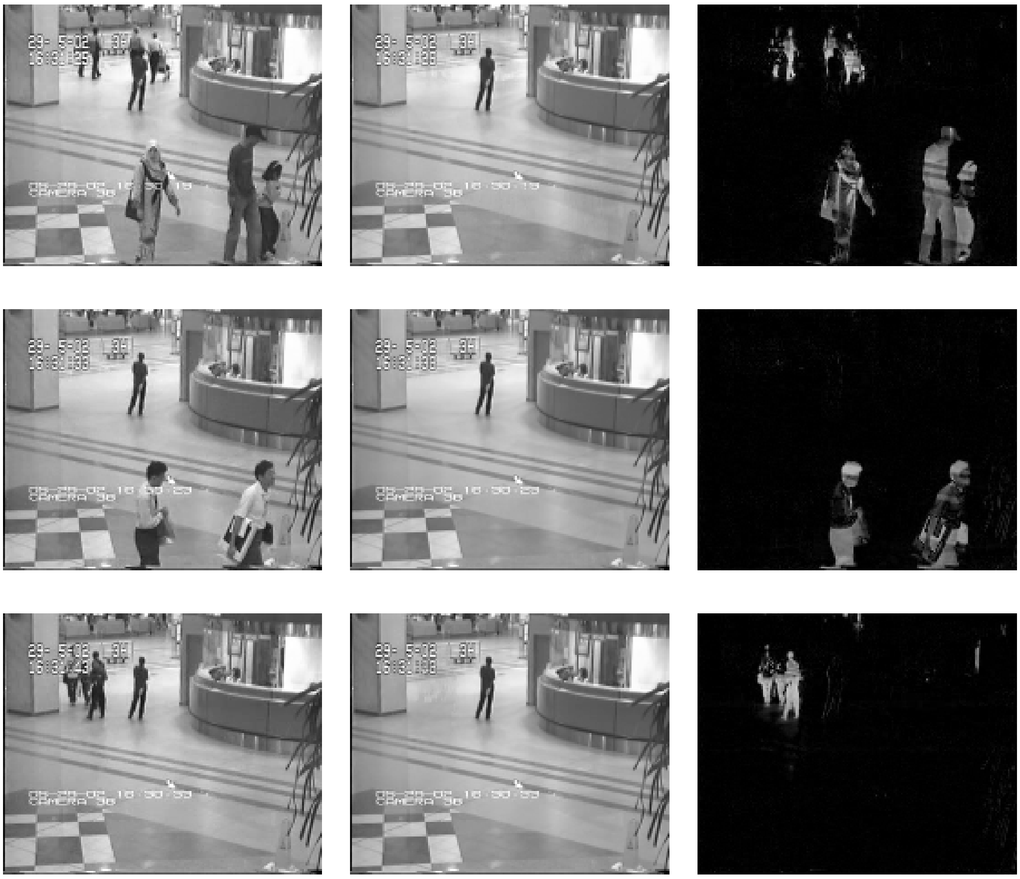


Figure 1.3: Moving object detection in video surveillance [19, Figure 2]; Left: video surveillance frames, middle: separated background, right: moving foreground objects

In these applications mentioned above, the given data sets exhibit intrinsic low-dimensional structures representing different kinds of sparse data features, that need to be reflected in an appropriate model of the data analysis problems at hand. The formulation of the corresponding approximation problems that have to be approached for their solution typically involve ℓ_p-quasinorms for vectors or appropriate related matrix-valued quasinorms.

Despite their, at first sight, very different nature, the three data analytical problems can be translated into optimization problems that are efficiently solvable with algorithms belonging to the class of Iteratively Reweighted Least Squares (IRLS) methods.

IRLS is an algorithmic strategy which classically imitates ℓ_p-minimization for vectors in residual minimization or signal recovery. Its formulation can also be adapted to the minimization of related matrix quasinorms. The algorithm is performing a successive approximation of the ℓ_p-minimization problem by solving a weighted least squares problem with an iteratively adapted weight matrix in each iteration. This procedure is hopefully leading to convergence to the actual ℓ_p-minimizer.

IRLS approximation constitutes a powerful and adaptable algorithmic scheme for a wide range of problems in engineering and applied sciences. It can be employed as a fast and robust computational tool in a large number of application scenarios, in particular in statistics for robust regression, general nonlinear parameter estimation, maximum or quasi-likelihood estimation, and in the expectation-maximization context. Beyond that, IRLS methods were able to deliver remarkable results in signal processing in sparse vector [35] and low-rank matrix [51] recovery as well as for the solution of minimization problems involving bounded variation functions [26]. A superlinear convergence rate can be reached even for nonsmooth and nonconvex optimization problems [36].

This thesis is a self-contained compendium of our research work, collected in the research papers [86, 87, 135], and the so far unpublished research work in Section 5.1-5.2, Subsection 5.3.2-5.3.3, Section 5.4-5.5. In this work, we contribute with novel iteratively reweighted least squares algorithms both extending the applicability spectrum of this algorithm class and enhancing performance for classical application settings with respect to the state-of-the-art. The results in this thesis for this class of algorithms can be conceptually summarized in the following two directions:

(i) **IRLS for minimization of sparse nonlinear residuals [Chapter 3]**
In the context of sparse measurement outliers, as in the example of faulty sensor detection in a wireless sensor network, robust methods for regression also for nonlinear measurement settings are desired.

In Chapter 3, that presents results first appearing in the paper [135] by the author of this thesis, we discuss an IRLS method for sparsity-promoting ℓ_p-norm-residual minimization involving nonlinear operators which allows the efficient numerical treatment of such problems. In particular, the investigation of the algorithmic behaviour and performance of IRLS as well as its applicability conditions and limitations are of high interest for statistical analysis in relevant application cases.

We present a rigorous theoretical analysis of the convergence behaviour for this IRLS-type algorithm called NR-IRLS under certain applicability conditions on the nonlinear measurement operators. Actually, this work includes the cases, where the measurement map is nonlinear and mildly smooth with parameter $1 \leq p < 2$ and, hence, we face not only a nonconvex but even a nonsmooth optimization problem. More precisely, the novelty is its ability to deal with severe nonsmoothness resulting from the cases, where $p \approx 1$.

(ii) **IRLS for learning of signals with low-dimensional intrinsic structures**

(a) *Low-rank matrices [Chapter 4]*

The occurrence of the matrix completion problem in recommender systems, exemplary described above for the case of the Netflix movie streaming service, and further highly relevant applications like system identification [153] or global positioning in sensor networks [21] made the low-rank matrix recovery problem a widely studied problem in the machine learning community.

The strategy of using an `IRLS`-type method for the approximation of the low-rank matrix recovery problem via a Schatten-p-quasinorm minimization problem appeared already in the papers [51, 106] published several years ago. Still, both of the `IRLS` approaches presented in those publications are not able to fully generalize the properties of the algorithm for sparse vector recovery in [35]. In particular, for the algorithms defined in [51, 106] there was no way to establish a provable rate of convergence.

However, the algorithm under discussion in this chapter, the harmonic mean iteratively reweighted least squares (`HM-IRLS`), which was first presented in [86] and [87] by the author of this thesis in collaboration with Christian Kümmerle, introduces an important novelty. We use a new kind of weight matrices, the so-called *harmonic mean weight matrices*, which are *more symmetrical* than the weight matrices previously used [51, 106]. This empowers `HM-IRLS` to overcome the disadvantages of the weight matrices in [51, 106].

Similar to existing work, we introduce an auxiliary functional as a tool to extend the previous convergence results to `HM-IRLS`, partly under null space property assumptions. Moreover, as a main theoretical result, we show that, in contrast to other `IRLS` variants, `HM-IRLS` can exhibits a locally superlinear rate of convergence. This rate can be accurately verified in our numerical tests and also practically reaches rates arbitrarily close to quadratic for $p \to 0$. Even more surprisingly, `HM-IRLS` demonstrates global convergence and superior performance in terms of sample complexity in comparison with various state-of-the art methods in the literature, in particular in the strongly nonconvex regime of $p \ll 1$.

(b) *Signals with multiple structures or combination of structures [Chapter 5]*

In this chapter, that contains so far unpublished work, we want to pass over from the recovery of low-rank matrices to more general high dimensional signals with multiple underlying structures from a minimal amount of linear measurements. In practical applications like video surveillance or face recognition, often the signal to be recovered is either a matrix with multiple sparsity-type structures occurring simultaneously or is the linear combination of several signals with different sparsity structures.

Recently, the negative results of Oymak e.a.[116] revealed that the intuitive attempt of combining convex norms usually minimized to promote each of the single structures will require just as many measurements as exploiting only one (dominating) structure. Only the combination of the nonconvex penalization functionals that are promoting a certain structural property will be beneficial for a reduction of the number of measurements. Motivated by this recently discovered fact, we investigate and analyse an `IRLS`-type method as this algorithm family has proven to be an efficient tool even in the nonconvex regime.

We present a very general formulation of an `IRLS` algorithm, named `GIRLS`, fusing different reweighting strategies into one unified weight matrix. We again utilize auxiliary functionals combining terms corresponding to the single sparsity structures and allow constrained as well as unconstrained formulations to incorporate the measurement information. Using this tool, we show convergence results and error bounds, some of which are based on null space properties of the measurement operator.

Additionally, we provide in Chapter 2 a synthetic overview on the fundamentals and most important aspects in data analysis and optimization that, finally, pave the way towards the formulation of iteratively reweighted least squares algorithms. We access the topic from a data analysis point of view presenting the types of approximation goals that will be of interest in this work and formulate corresponding optimization problems for their solution. Of course, we provide a collection of useful tools from optimization and explain the special aspects to consider for high dimensional data. Furthermore, we introduce the concept of sparsity for vector and matrix valued data and its theoretical foundations. With this chapter, we provide for the reader useful results from other literature sources, that are required for the elaboration of the topics in the subsequent chapters.

In accordance with the latest regulations concerning the authorship of the results obtained in Ph.D. theses, I hereby declare that, if not clearly stated otherwise, the content of the following chapters, sections and subsections are original and were obtained thanks to my contribution

Chapter 3, Subsection 4.1.2, Section 4.2, Subsections 4.3.1-4.3.3, Subsections 4.3.5-4.3.7, Section 4.4, Section 5.1-5.2, Subsection 5.3.2-5.3.3, Section 5.4-5.5.

Moreover, I hereby declare that, Christian Kümmerle and myself equally contributed to all results in the above listed sections and subsections belonging to Chapter 4 and 5. The remaining chapters, sections and subsections are reelaborated versions of results that can be found in the literature and that are necessary to make the thesis self-contained.

Some pictures shown in this introduction were produced by myself using images licensed under a Creative Commons license that allows modifications as can be verified by following the related web addresses. Additionally, this introduction contains pictures found in the literature and we give a reference to their source in the caption. The pictures in the rest of the work were either autonomously produced in the context of [135] or the outcome of the scientific partnership with the co-author of the paper [87].

Chapter 2

En route to iteratively reweighted least squares: From data sets to optimization problems

This chapter paves the way for the reader towards the investigation of the powerful class of Iteratively Reweighted Least Squares (IRLS) algorithms.

Starting from its origin in data analysis as a tool for the solution of approximation problems, we present the classic least squares as an introductory example.

Thereafter, we continue with the modeling of data with low dimensional structures and the translation into an optimization problem with an appropriate objective. Of course we also provide the essential background in optimization. Subsequently, we introduce Iteratively Reweighted Least Squares (IRLS) as the sequence of minimizations of surrogate functionals to the original objective.

2.1 Data analysis and approximation problems

The acquisition of data in the form of output values of measurements with unidentified influence parameters or an unknown input signal is ubiquitous, e.g., in scientific experiments, engineering tests, and financial market observations.

In many applications, it is necessary to relate these measured or observed values, that are possibly afflicted with measurement inaccuracies, with the unknown input variables by formulating a mathematical model describing the measurement process.

This de facto means, we employ modeling as an approximation to reality that gives us further insights on the influence factors of the observed process.

The upcoming subsection takes inspiration from parts of Chapter 6 on approximation and fitting of the classic book [12] .

2.1.1 About data sets and approximation problems

Formulating this mathematically more precisely, we assume we are given a data set $y \in \mathbb{R}^m$ with a number of m measurements $y_j, j \in [m]$ resulting from a measurement process that can be described by the map $\Phi : \mathbb{R}^d \to \mathbb{R}^m$, often called measurement or sampling operator. In some cases, we assume that an error $e \in \mathbb{R}^m$ is involved in the measurement results as well , also referred to as measurement noise. The input variable $x \in \mathbb{R}^d$ to the map Φ represents d possible influence parameters $x_i, i \in [d]$.

This relationship can be cast into a system of equations with m equations and d unknowns as follows

$$\Phi(x) + e = y. \tag{2.1}$$

Let us for simplicity assume now that $\Phi(x)$ is a linear map with matrix representation $\Phi(x) = \Phi x$ for $\Phi \in M_{m\times d}$ and that the error $e = 0$ resulting in the system

$$\Phi x = y. \tag{2.2}$$

In the case that $m = d$ and Φ is full rank, Φ is invertible and there exists a unique solution to (2.2) that can be calculated explicitly

$$x = \Phi^{-1} y.$$

However, if e is nonzero, the result above might not be accurate enough, or if the sampling operator Φ in (2.1) is nonlinear, not even in the case $d = m$ a (unique) solution of the equation system might exist, e.g., if $y \notin \operatorname{ran}(\Phi)$. As a consequence, an appropriate approximation is desired.

This comes into effect in particular in the cases where $m \neq d$:

(i) $m > d$: equation (2.1) is an *overdetermined* equation system, for which it is impossible to fulfill all equations exactly.
We need an approximation to the solution with minimal error with respect to a certain error function in dependence of the residual $e = \Phi(x) - y$, which we denote with $f(e)$.

(ii) $m < d$: equation (2.1) is an *underdetermined* equation system with infinitely many solutions, and further assumptions on the signal x are necessary to ensure uniqueness. This can be realized by introducing a penalization or regularization function $g(x)$ which puts at disadvantage unfavored solutions, and then consider minimization of $g(x)$ under the constraint $\Phi(x) = y$.

Our strategy is to identify an optimal solution $x = x_{\text{opt}}$ to the approximation problem above as the minimizer of an objective functional $\mathcal{J}(x)$

$$x_{\text{opt}} = \arg\min_{x} \mathcal{J}(x), \tag{2.3}$$

where $\mathcal{J}(x)$ reflects the requirements of the particular problem setting. We design $\mathcal{J}(x)$ by the incorporation of function terms enforcing certain characteristics for the optimal solution to (2.1) as follows:

(i) achieving a certain accuracy in the data fit, i.e., minimizing the approximation error $f(e)$

$$x_{\text{opt}} = \arg\min_{x} \mathcal{J}(x) = \arg\min_{x} f(e) = \arg\min_{x} f(\Phi(x) - y) \tag{2.4}$$

(ii) generalizing best to future observations by identifying the relevant parameters and the underlying data structure, i.e., minimizing the regularization term $g(x)$ and thereby penalize undesired features

$$x_{\text{opt}} = \arg\min_{x} \mathcal{J}(x) = \arg\min_{\Phi(x)=y} g(x) \tag{2.5}$$

If both aspects are of relevance, we suggest solving a minimization problem combining the two components $f(e)$ and $g(x)$ weighted by a factor $\lambda > 0$ according to their importance in the problem context:

$$x_{\text{opt}} = \arg\min_{x} \mathcal{J}(x) = \arg\min_{x} f(e) + \lambda g(x) = \arg\min_{x} f(\Phi(x) - y) + \lambda g(x) \tag{2.6}$$

This optimization problem always has a unique solution depending on the regularization parameter λ.

2.1.2 Least squares: examples for error and penalization functions

Let us now consider the problem (2.2) for the overdetermined case with $\Phi \in M_{m \times d}$ for $m > d$ and measurement data that might be corrupted by nonzero noise e

$$\Phi x + e = y. \tag{2.7}$$

A standard approach in statistics for developing estimates of the model parameters x in (2.7) is the so-called *linear least squares fitting*, i.e., minimizing the sum of squares of the differences between the values predicted by a linear measurement model and the actually observed data or measured results.

This corresponds to the approximation solution to (2.2) via the minimization problem as in (2.4) with choice of the Euclidean norm as an error function

$$f(e) = \|e\|_{\ell_2} \tag{2.8}$$

leading to

$$x_{LS} = \arg\min_x \mathcal{J}(x) = \arg\min_x \|e\|_{\ell_2} = \arg\min_x \|\Phi x - y\|_{\ell_2}. \tag{2.9}$$

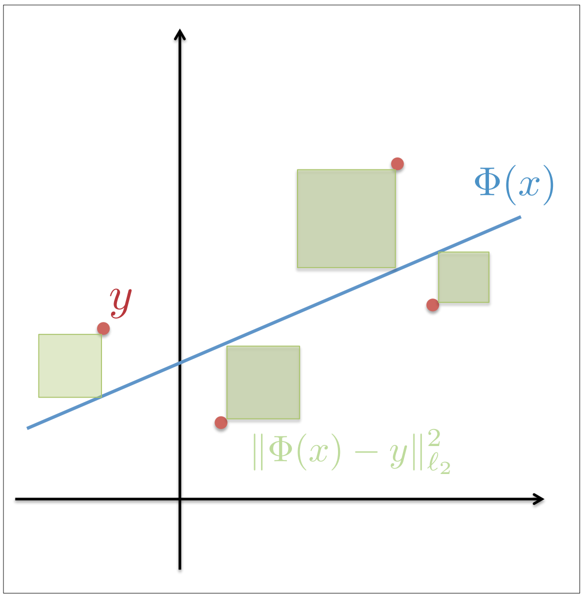


Figure 2.1: Visualization of least squares fitting

For Φ with full column rank the solution x_{LS} in (2.9) has the following closed-form representation

$$x_{LS} = (\Phi^T\Phi)^{-1}\Phi^T y = \Phi^\dagger y, \tag{2.10}$$

where $\Phi^\dagger \in \mathbb{R}^{d\times m}$ is the Moore-Penrose pseudo inverse of $\Phi \in M_{m\times d}$ with $m > d$.

The least squares fitting constitutes the simplest and most common form of linear regression with a wide field of practical applications.

It is possible to solve the for x_{LS} in (2.10) by employing one of the approaches listed below with increasing order of computational complexity and stability [59, 92]:

(i) Cholesky factorization of $\Phi^T\Phi$,

(ii) QR-factorization of Φ, and

(iii) Singular Value Decomposition of Φ.

In applications, where least squares approximations to the solution of (2.1) are employed, often a Gaussian distribution of the error e is assumed, i.e., $e_j \sim \mathcal{N}(0,\sigma)$ are i.i.d. zero-mean Gaussian random variables with standard deviation $\sigma > 0$ for $j \in [m]$ [65, 134]. Even in the case, where the measurement model map Φ is very accurately approximating reality, the measurement noise e can disturb the quality of the approximation result x. Therefore, the characteristic nature of the error occurring in the measurements and the choice of the error function influence the goodness of the solution for the particular problem.

As an example, for modeling different kinds of noise in statistical and data analytical applications, from Gauss or Laplace distributed to impulsive noise, a widely used type of error function is the more general ℓ_p-norm [134], i.e.,

$$f(e) = \|e\|_{\ell_p} = \left(\sum_{j=1}^{m} |e_j|^p\right)^{\frac{1}{p}} \tag{2.11}$$

leading to the solution of the optimization problem

$$x_{LP} = \arg\min_x \mathcal{J}(x) = \arg\min_x \|e\|_{\ell_p} = \arg\min_x \|\Phi(x) - y\|_{\ell_p}. \tag{2.12}$$

The parameter $0 \leq p \leq 2$ is adjusted in dependence on the type of residual error [12], e.g., in the case of Laplacian noise the Euclidean norm is replaced by the ℓ_1-norm

$$f(e) = \|e\|_{\ell_1} = \sum_{j=1}^{m} |e_j| \tag{2.13}$$

resulting in so-called least absolute deviation fitting [104]

$$x_{L1} = \arg\min_x \mathcal{J}(x) = \arg\min_x \|e\|_{\ell_1} = \arg\min_x \|\Phi(x) - y\|_{\ell_1}. \tag{2.14}$$

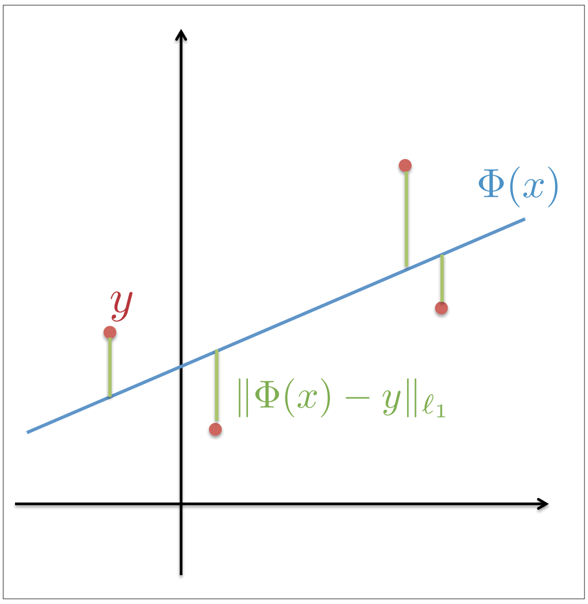


Figure 2.2: Visualization of least absolute deviation fitting

In general, if the noise error is assumed to follow a certain design pattern, often corresponding to a known type of probability distribution, one wants to enforce the choice of an x that results in an error $e = \Phi(x) - y$ with the expected properties by employing a suitable error function.

In contrast, considering the case of an underdetermined linear equation system, i.e., $m < d$, with error $e = 0$

$$\Phi x = y, \tag{2.15}$$

the map Φ is not injective and we are confronted with infinitely many solutions. Indeed, there is a $(d-m)$-dimensional affine space of vectors $x = x_p + \mathcal{N}(\Phi)$ solving this system, where x_p is any particular solution and $\mathcal{N}(\Phi)$ denotes the null space of Φ [48].

The task here is to select the most suitable solution x_{opt} for the specific problem out of this set and to pose further assumptions on the solution vector, in the best case also allowing unique identifiability of x_{opt}.

In a like manner as described for assumptions on the noise above, to determine the solution x_{opt} matching best the desired characteristics, a-priori information on the signal x needs to be integrated. This can be realized by the minimization of an appropriate penalization function under the linear constraint $\Phi x = y$ in (2.5).

As an example, in the case that it is known to be close to the origin in Euclidean distance, one can choose as a penalization function

$$g(x) = \|x\|_{\ell_2} \tag{2.16}$$

giving rise to the least Euclidean norm solution

$$x_{LN} = \arg\min_{x} \mathcal{J}(x) = \arg\min_{\Phi x = y} \|x\|_{\ell_2}. \tag{2.17}$$

Again, if Φ has full row rank, the solution x_{LN} in (2.17) can be represented in closed-form as

$$x_{LN} = \Phi^T(\Phi\Phi^T)^{-1}y = \Phi^{\dagger}y, \tag{2.18}$$

where $\Phi^{\dagger} \in M_{d\times m}$ is the Moore-Penrose pseudo inverse of $\Phi \in M_{m\times d}$ with $m < d$.

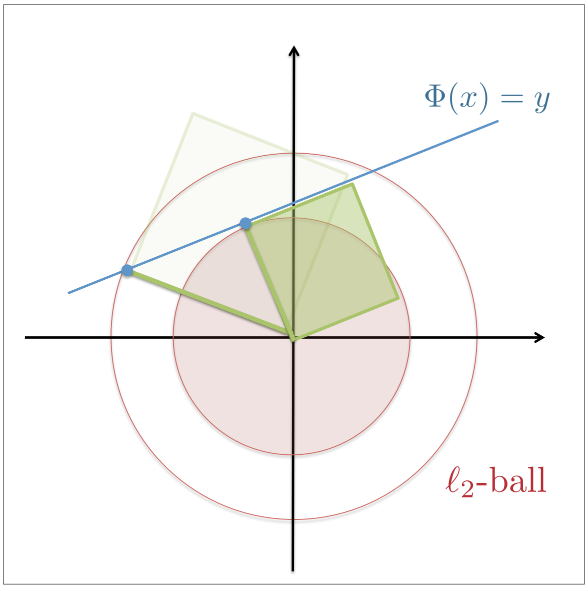


Figure 2.3: Visualization of least norm or ℓ_2-norm penalty minimization

For additional examples and discussions on possible a priori assumption on the signal, introducing the concept of sparsity and the deduction of appropriate penalization functions, we refer to Section 2.3.

2.1.3 Generalizing the example: weighted least squares

As seen in the examples above, for the choice of the Euclidean norm as an error or penalization function, practical explicit calculation formulas for the solutions to the optimization problems (2.4) and (2.5) come in handy. We now want to consider a generalization of this approach by introducing a *weighting matrix* to be multiplied to the argument of the Euclidean norm function. Thereby we can emphasize or deemphasize certain components of the argument vector, i.e., of the error e for (2.4) or the signal x for (2.5) respectively. [141]

Definition 2.1 (Weighted ℓ_2-norm)**.** Let $z, \bar{z} \in \mathbb{R}^N$ and $W \in \mathbf{S}_+^N$ be a symmetric, positive semidefinite weight matrix. We define the weighted ℓ_2-scalar product

$$\langle z, \bar{z} \rangle_{\ell_2(W)} = \langle Wz, \bar{z} \rangle = \langle z, W\bar{z} \rangle$$

and the induced reweighted ℓ_2-norm with weight matrix W as

$$\|z\|_{\ell_2(W)} = \langle z, z \rangle_{\ell_2(W)}. \tag{2.19}$$

As a special case, we consider W to be a diagonal matrix with positive entries, i.e., $W = \operatorname{diag}(w)$, where $w \in \mathbb{R}^N$ with $w_i > 0$ for $i \in [N]$. In this case, we can define the ℓ_2-norm with a vector valued weight w as

$$\|z\|_{\ell_2(w)} = \langle z, z \rangle_{\ell_2(W)} = \sum_{i=1}^{N} w_i z^2. \tag{2.20}$$

The optimization problems in (2.4) can be formulated with $f(e) = \|e\|_{\ell_2(W)}$ with an appropriate choice for $W \in \mathbb{S}_+^m$ and (2.5) with $g(x) = \|x\|_{\ell_2(W)}$ with $W \in \mathbb{S}_+^d$ respectively.

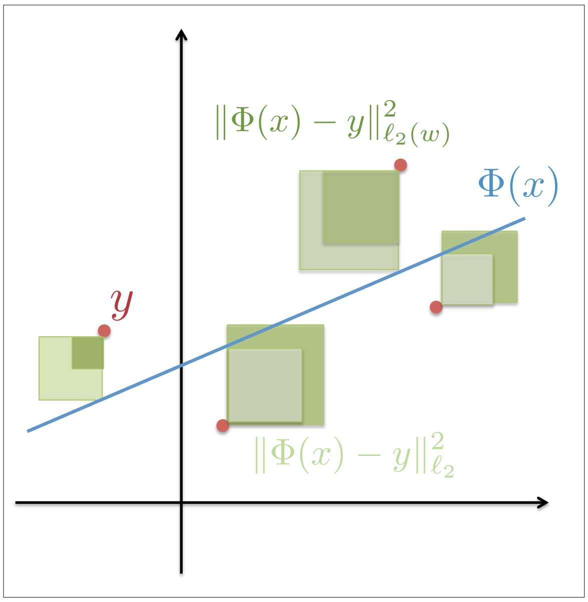


Figure 2.4: Visualization of weighted least squares fitting

The formulas for the solutions in (2.10) and (2.18) generalize for the case of a reweighted ℓ_2-norm to the weighted least squares solution

$$x_{LS_W} = (\Phi^T W \Phi)^{-1} \Phi^T W y, \tag{2.21}$$

and the minimum weighted norm solution

$$x_{LN_W} = W^{-1} \Phi^T (\Phi W^{-1} \Phi^T)^{-1} y, \tag{2.22}$$

respectively.

These closed-form calculation rules for the solutions to the corresponding optimization problems in (2.4) and (2.5) can be used as a standalone approximation, but they will, moreover, be one of the fundamental building blocks for the development of an iteratively reweighted least squares algorithm in Section 2.4.

2.2 Nonlinear optimization problems and surrogate functionals

In this section, we want to pave the ground for the solution of the optimization tasks arising from data analysis problems as mentioned above. Our goal is to prepare the reader for the generalization of least squares problems to nonlinear measurement operators in Chapter 3. The solution of the nonconvex minimization problems resulting from the introduction of the nonlinearity can be challenging and we will provide some helpful tools for handling these difficulties.

The content of the next two subsections in general follows loosely the presentation of Chapter 2 and 10 in [113] and the introduction to convex analysis and the subdifferential in [54].

2.2.1 Prerequisites from optimization

Finding optimal solutions to application problems, e.g., in engineering, science, economics or within more complex types of mathematical problem settings, is often formulated as the minimization (or maximation) of functions in one or several variables, possibly involving constraints on the variables as well. The field of optimization has its mathematical foundations in linear algebra and multivariate calculus and utilizes many tools from these branches of mathematics.

Let $F : \mathbb{R}^d \to \mathbb{R}$ be a real-valued function with d input variables in form of a vector in $\mathbb{R}^d$, where its function value at the point $x \in \mathbb{R}^d$ is denoted by $F(x)$. In the literature, the function F that we aim to minimize (or maximize) is often referred to as the objective function. In the following, we will restrict our considerations without loss of generality to minimization problems.

First let us consider unconstrained minimization problems of the form

$$\min_{x \in \mathbb{R}^d} F(x). \tag{2.23}$$

In this case, the vector $x^* \in \mathbb{R}^d$ is called a *local minimizer* of F if there exists an $\epsilon > 0$ such that

$$F(x^*) < F(x) \text{ for all } x \in B_\epsilon(x^*).$$

We introduce the stronger notion that x^* is a *global minimizer* of F if

$$F(x^*) < F(x) \text{ for all } x \in \mathbb{R}^d.$$

We now want to transfer the notion of minimizer also to constrained problems: we aim at the minimization of the objective function $F(x)$ subject to $x \in \mathcal{C}$. This means, we minimize $F(x)$ over all x lying in a predefined set $\mathcal{C} \subset \mathbb{R}^d$, corresponding to the constraint.

$$\min_{x \in \mathcal{C}} F(x). \tag{2.24}$$

We obtain a local minimizer x^* of F subject to the set $\mathcal{C}$ if

$$F(x^*) < F(x) \text{ for all } x \in B_\epsilon(x^*) \cap \mathcal{C}.$$

Similarly, we define of the global minimizer of F over the set $\mathcal{C}$

$$F(x^*) < F(x) \text{ for all } x \in \mathcal{C}.$$

As a first step, we introduce further terminology that will be useful for the upcoming discussion on the existence of optimal solutions.

Definition 2.2. Let $\mathcal{C} \subset \mathbb{R}^d$ be a subset of $\mathbb{R}^d$, $F : \mathcal{C} \to \mathbb{R}$ be a real valued function and $c \in \mathbb{R}$ be a constant. Then the *level set* of F on $\mathcal{C}$ corresponding to the value c is a set of the form

$$\ell_{F,\mathcal{C}}(c) = \{x \in \mathcal{C} : F(x) \leq c\}.$$

Next we the proceed with the notion of lower semicontinuity, which is a useful generalization of the continuity concept.

Definition 2.3. A function $F : \mathbb{R}^d \to \mathbb{R}$ is called *lower semicontinuous* if, for every $x \in \mathbb{R}^d$ and every sequence $(x_j)_{j \geq 1}$ converging to x,

$$\lim \inf_{j \to \infty} F(x_j) \geq F(x).$$

Of course, a continuous function $F : \mathbb{R}^d \to \mathbb{R}$ is lower semicontinuous. Let us mention that lower semicontinuity of F is equivalent to the closedness of all its level sets $\ell_{F,\mathbb{R}^d}(c)$.

To present also a nontrivial example of a lower semicontinuous function, we consider the characteristic function $\mathcal{X}_\mathcal{C}$ of a proper subset $\mathcal{C}$, which is not continuous, but lower semicontinuous if and only if $\mathcal{C}$ is closed.

Definition 2.4. A function $F : \mathbb{R}^d \to \mathbb{R}$ is called *coercive* with respect to the set $\mathcal{C}$ if either $\mathcal{C}$ is bounded or it holds that

$$\lim_{\substack{\|x\|_{\ell_2} \to \infty \\ x \in \mathcal{C}}} F(x) \to \infty.$$

The coercivity of $F : \mathcal{C} \to \mathbb{R}$ is equivalent to the property that F has bounded level sets $\ell_{F,\mathcal{C}}(c)$.

Using the definitions above, we are able to present the following statement on the *existence of minimizers* to the problems (2.23) and (2.24), derived from Weierstrass' theorem in different variants [34, 58].

Theorem 2.5. *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a nonempty and closed set, and $F : \mathcal{C} \to \mathbb{R}$ be a lower semicontinuous function on $\mathcal{C}$. If any of the below conditions holds, then there exists a nonempty, compact set of minimizers of F subject to $\mathcal{C} \subseteq \mathbb{R}^d$*

(i) $\mathcal{C}$ is bounded.

(ii) F bounded from below by a value α, i.e., $F(x) \geq \alpha$ for all $x \in \mathcal{C}$ and, moreover, F is coercive with respect to the set $\mathcal{C} \subseteq \mathbb{R}^d$.

This result is also known as the "direct method of calculus of variations" [34]. The proof of this statement uses the compactness of the level sets following from the lower semicontinuity and the conditions (i) and (ii). Note that the second condition allows $\mathcal{C} = \mathbb{R}^d$.

The question arising from these results is how to identify local or even global minimizers of a given function $F : \mathbb{R}^d \to \mathbb{R}$ possibly subject to a constraint set $\mathcal{C}$ in a general setting.

In the very general, difficult case that we aim at the minimization of a nondifferentiable function F, it is possible to apply algorithms based only on function evaluation values such as the Nelder-Mead algorithm [111] or pattern search [148]. However, these algorithms are computationally demanding and smoothness of the objective function is in general a very desirable property.

In a next step, we introduce the important concept of *convexity* discussing convex sets and convex functions later on. Moreover, we give reasons why the convexity property for constraint sets and objective functions is advanteguous as well. Convexity originates from the field of mathematical analysis dealing with convex sets and convex functions and constitutes a fundamental concept for optimization.

Definition 2.6. We call a subset $\mathcal{C} \in \mathbb{R}^d$ *convex*, if for all $x, \bar{x} \in \mathcal{C}$,

$$tx + (1-t)\bar{x} \in \mathcal{C} \text{ for all } 0 \leq t \leq 1,$$

meaning that the line segment connecting the points x and $\bar{x}$ is contained in $\mathcal{C}$. Moreover, it holds that a set $\mathcal{C}$ is convex if and only if, for all $x_1, \dots, x_d \in \mathcal{C}$ and $t_1, \dots, t_d \geq 0$ such that $\sum_{i=1}^d t_i = 1$, the convex combination $\sum_{i=1}^d t_i x_i = 1$ is also contained in $\mathcal{C}$.

As typical examples for convex sets can be listed subspaces, affine spaces or norm balls $B_\epsilon(x)$, and intersections of convex sets are convex sets again.

Now we go on with the introduction of different types of convexity of functions with the following definition

Definition 2.7. A function $F : \mathbb{R}^d \to \mathbb{R}$ is called

(i) *convex* if, for all $x, \bar{x} \in \mathbb{R}^d$ and $0 \leq t \leq 1$,

$$F(tx + (1-t)\bar{x}) \leq tF(x) + (1-t)F(\bar{x}),$$

(ii) *strictly convex* if, for all $x \neq \bar{x} \in \mathbb{R}^d$ and $0 < t < 1$,

$$F(tx + (1-t)\bar{x}) < tF(x) + (1-t)F(\bar{x}),$$

(iii) *strongly convex* with parameter $\gamma > 0$ if, for all $x, \bar{x} \in \mathbb{R}^d$ and $0 \leq t \leq 1$,

$$F(tx + (1-t)\bar{x}) \leq tF(x) + (1-t)F(\bar{x}) - \frac{\gamma}{2} t(1-t) \|x - \bar{x}\|_{\ell_2}^2.$$

Examples of convex functions include linear and affine functions as well as all kinds of norms. Moreover, note that a strongly convex function is strictly convex as well.

Finally, a fact pointing out an interesting connection between convex sets and convex functions is the following

Proposition 2.8. *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set, $c \in \mathbb{R}$ and consider a convex function $F : \mathcal{C} \to \mathbb{R}$. Then the level sets of F on $\mathcal{C}$ are convex.*

Similar as for convex sets, convexity for functions can be defined using convex combinations: a function $F : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if for all $x_1, \dots, x_d \in \mathbb{R}^d$ and

$t_1, \dots, t_d \geq 0$ such that $\sum_{i=1}^{d} t_i = 1$,

$$F\left(\sum_{i=1}^{d} t_i x_i\right) \leq \sum_{i=1}^{d} t_i F(x_i).$$

We continue with a short note about continuity of convex functions.

Proposition 2.9. *A convex function $F : \mathbb{R}^d \to \mathbb{R}$ is continuous on $\mathbb{R}^d$.*

A further characterization for different types of convexity of differentiable functions can be formulated as follows:

Proposition 2.10. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function.*

(i) F is convex if and only if, for all $x, \bar{x} \in \mathbb{R}^d$,

$$F(x) \geq F(\bar{x}) + \langle \nabla F(\bar{x}), x - \bar{x} \rangle,$$

where $\nabla F(\bar{x})$ is the gradient of F at $\bar{x}$.

(ii) F is strongly convex with parameter $\gamma > 0$ if and only if, for all $x, \bar{x} \in \mathbb{R}^d$,

$$F(x) \geq F(\bar{x}) + \langle \nabla F(\bar{x}), x - \bar{x} \rangle + \frac{\gamma}{2} \|x - \bar{x}\|_{\ell_2}^2.$$

(iii) If F is twice differentiable, then it is convex if and only if, for all $x \in \mathbb{R}^d$,

$$\nabla^2 F(x) \succeq 0,$$

where $\nabla^2 F(x)$ is the Hessian of F at x.

Having collected important properties of convex sets and functions, we point out the properties that make this class of sets and functions especially useful in the optimization context both in theoretical and practical aspects.

We call an optimization problem a convex optimization problem if we aim at the minimization of a convex function F over a convex set $\mathcal{C} \subseteq \mathbb{R}^d$.

Theorem 2.11. *Let $F : \mathcal{C} \to \mathbb{R}$ be a convex function defined on a convex set $\mathcal{C} \subseteq \mathbb{R}^d$.*

(i) Then every local minimizer of F over $\mathcal{C}$ is also a global minimizer.

(ii) If F is continuous and $\mathcal{C}$ is closed, then the set of local (and therefore global) minimum points of F over $\mathcal{C}$ is a closed convex set.

(iii) If $F : \mathcal{C} \to \mathbb{R}$ is strongly convex, the minimizer of F over $\mathcal{C}$ is unique.

From this theorem, it becomes clear why these types of problems are particularly interesting, as property (i) implies that convex optimization problems allow for an efficient algorithmic treatment.

The question arising from these results is, how to identify local or even global minimizers of a given function $F : \mathbb{R}^d \to \mathbb{R}$ possibly subject to a constraint set $\mathcal{C}$, also in a general setting not restricted to the convex case.

To provide an adequate answer, we give conditions on the optimality of certain function input values, that can also be the core idea for the construction of efficient algorithms.

2.2.2 Optimality conditions and a detour to nonlinear least squares problems

In this next part, we focus on differentiable, nonlinear functions F and at least partly give corresponding results for the nondifferentiable but convex case as well. However, a detailed discussion of nonsmooth optimization will not be part of this thesis and we refer to [4] for further results in this direction.

We start our considerations with unconstrained problems and present powerful optimality conditions that can be derived from Taylor's theorem.

Theorem 2.12. *Let $x^* \in \mathbb{R}^d$ and assume that $F : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable. We consider the*

(i) first-order necessary and sufficient condition: $\nabla F(x^*) = 0$,

(ii) second-order necessary condition: $\nabla^2 F(x^*) \succeq 0$, *i.e.*, $\nabla^2 F(x^*)$ *pos. semidef.*,

(iii) second-order sufficient condition: $\nabla^2 F(x^*) \succ 0$, *i.e.*, $\nabla^2 F(x^*)$ *pos. def.*

Points x^* fulfilling condition (i) are called critical (or stationary) points *of F.*
On the one hand, if x^* is a local minimizer of F, necessarily conditions (i) and (ii) have to hold. On the other hand, if condition (i) and (iii) hold, x^* is a local minimizer of F.

The first-order condition stated in (i) implies that the gradient is zero at x^*. Furthermore, the second-order condition can be interpreted as a kind of local convexity condition on F in the neighbourhood of x^* if we compare conditions (ii) and (iii) to Proposition 2.10 (iii).

Condition (i) gives us an instruction, how to find at least critical points of a given function: by finding solutions x^* to the equation system $\nabla F(x) = 0$.

It is clear that if F is a quadratic function of the variable x, $\nabla F(x) = 0$ is a linear system of equations, where direct solvers and effective computational methods are readily available.

In the general case, however, $\nabla F(x) = 0$ is a $d \times d$-dimensional nonlinear system of equations, which can be very demanding to solve. In the case that F is nonconvex, even determining whether a solution exists and whether it is unique is sometimes already a hard problem. As in general no direct solution methods are applicable, we will have to rely on heuristic algorithms that are not guaranteed to succeed but often work well in practice.

One can try to solve this equation with iterative methods, for example Newton's method, and in this way obtain a critical point as a candidate for a local minimizer. Nevertheless, in the general case, it is only possible to find stationary points and check whether they are local minimizers. However, it is not clear whether a global or local minimizer was identified by employing this approach. In practise, such strategies only succeed in finding a global minimizer if the initialization of the iterative method is chosen smartly and sufficiently close to it already.

To illustrate the approach of an iterative algorithm for the solution of nonlinear optimization problems, we sketch the basic idea and the typical algorithmic procedure that these strategies share:

Starting from a given *initialization point* $x^{(0)}$, the algorithm outputs a series of vectors $x^{(1)}, x^{(2)}, \ldots$ called iterates, for which we hope that it converges to a point $\lim\limits_{n \to \infty} x^{(n)} = \bar{x}$, coinciding with a local minimizer x^* for the objective function F.

Most of these methods implement criteria which enforce a descent condition on the iterates. This can be realized by performing the n-th iteration step starting from the current iterate $x^{(n)}$ as follows

- Determine a *descent direction* d (most popular choice: the gradient of F), and
- Choose a *step length* $\alpha > 0$ giving a good decrease in the objective function value.
- Set $x^{(n+1)} = x^{(n)} + \alpha d$.

Therefore, a descent method pursues the target

$$F(x^{(n+1)}) < F(x^{(n)})$$

in each step and often the algorithm terminates if the function values of the iterates does not change significantly anymore. This means that if stopping criterion $F(x^{(n)}) - F(x^{(n-1)}) < \delta$ for $\delta > 0$ is reached, for $n = \bar{n}$ the final output result of the algorithm is $\bar{x} = x^{(\bar{n})}$.

From this procedure it is clear that the output vector of the algorithm and therefore, also the local minimizer of the objective that is approached will depend on the initialization value $x^{(0)}$.

With the decrease of the objective function in each step we aim at the decrease of the distance of the iterates $x^{(n)}$ to the local minimizer x^* as well, at least from a certain iteration N on. Reformulating this goal in terms of the error $E^{(n)} = x^{(n)} - x^*$ gives

$$|||E^{(n+1)}||| < |||E^{(n)}||| \text{ for } n > N$$

for an appropriate norm $|||\cdot|||$. Theoretically, convergence of the algorithm to the local minimizer occurs if $\lim\limits_{n\to\infty} e^{(n)} = 0$ and we practically hope for an output $\bar{x} = x^{(\bar{n})}$ of the algorithm with $E^{(\bar{n})} \approx 0$.

Please note that for different algorithmic strategies we can have different convergence behaviour for the same problem. Some may exhibit convergence while others don't. Moreover, this decrease of the error per iteration and, therefore, the number of iterations necessary until convergence can be different for different algorithmic strategies.

Algorithms can also exhibit different so-called *convergence speed* or *convergence rates.* Depending on the relationship between $|||E^{(n+1)}|||$ and $|||E^{(n)}|||$ for $n > N$, i.e., when $|||E^{(n)}|||$ is already small, we can distinguish between

- *Linear convergence rate*: $|||E^{(n+1)}||| \leq \mu |||E^{(n)}|||$ with $0 < \mu < 1$,
- *Quadratic convergence rate*: $|||E^{(n+1)}||| \leq \mu |||E^{(n)}|||^2$ with $0 < \mu$,
- *Superlinear convergence rate*: $\frac{|||E^{(n+1)}|||}{|||E^{(n)}|||} \to 0$ for $n \to \infty$.

The fact that convergence occurs or not is independent of the choice of $|||\cdot|||$ but not the rate of convergence, that can vary for different choices of $|||\cdot|||$.

At this point, we want to come back to the generalization of the sum of squares minimization problem to *nonlinear* measurement operators $\Phi : \mathbb{R}^d \to \mathbb{R}^m$.

In the linear case, the concatenation of the error function $f(x) = \|x\|_{\ell_2}$ and the linear residual $e(x) = \Phi x - y$ leads to a quadratic objective $F(x) = f(e(x))$. For the minimization of this quadratic function one has to solve a linear equation resulting from the first-order optimality condition, where the efficient solvers mentioned above are available.

In contrast, the introduction of the nonlinear operator Φ leads to the concatenation of $f(x) = \|x\|_{\ell_2}$ with the nonlinear residual $e(x) = \Phi(x) - y$ and in the end leads to the minimization of a in general nonconvex objective function

$$\min_x F(x) = \min_x \|e(x)\|_{\ell_2} = \min_x \|\Phi(x) - y\|_{\ell_2}. \tag{2.25}$$

Even though nonlinear least squares residual minimization is a widespread problem, appearing for example in regression curve fitting or parameter determination from scientific experiments, this field is not an extensively discussed subject within analysis.

We present the application of different iterative descent methods as introduced above for the minimization of the nonlinear objective $F(x)$ as in (2.25) exemplary for this problem.

Let in the following $J(x)$ be the Jacobian of F at the point x. If $F(x)$ is twice continuously differentiable, then we solve the nonlinear equation

$$\nabla F(x) = J(x)^* e(x) = 0,$$

which provides local stationary points for $F(x)$.

The first algorithm we will investigate is the well-known *Newton's method.* Its derivation from Taylor's theorem indicates a descent direction d_N that is determined via the solution of the equation

$$\nabla^2 F(x^{(n)}) d_N = -\nabla F(x^{(n)}).$$

Using the result with step length $\alpha = 1$, we can formulate one iteration of this method, the so-called Newton step, as

$$x^{(n+1)} = x^{(n)} - (\nabla^2 F(x^{(n)}))^{-1} \nabla F(x^{(n)}) = x^{(n)} - (J(x^{(n)})^* J(x^{(n)}) + S(x^{(n)}))^{-1} J(x^{(n)})^* e(x^{(n)}), \tag{2.26}$$

where $S(x^{(n)})$ denotes the matrix $S(x^{(n)}) = \sum_{j=1}^{m} e_j(x^{(n)}) \nabla^2 e_j(x^{(n)})$.

Newton's method can reach *quadratic* convergence rate, but still it can be computationally expensive as it requires the calculation of md^2 derivatives for the evaluation of $S(x^{(n)})$ at each step.

A key observation towards computationally more efficient solvers related to Newton's method is the approximation of the Hessian

$$\nabla^2 F(x) = J(x)^T J(x) + S(x) \approx J(x)^T J(x). \tag{2.27}$$

resulting in the Gauß-Newton method with the subproblem:

$$\left[J(x^{(n)})^T J(x^{(n)})\right] d_{GN}^{(n)} = -\nabla F(x^{(n)}). \tag{2.28}$$

This approximation can be justified with the assumption that, especially in the case of mild nonlinearity of Φ, the residuals is small in the neighborhood of the solution and, therefore, the first term is significantly more important.

However, performing a sequence of such iterations with updates $x^{(n+1)} = x^{(n)} + d^{(n)}$, with the step size choice $\alpha = 1$ as chosen above in (2.26) fails to reach convergence. Therefore, for this kind of algorithms a more careful step size control is necessary. Depending on the strategy used to determine the step size $0 \leq \alpha \leq 1$, one distinguishes two main algorithm categories: line-search algorithms and trust-region algorithms.

(i) *Line Search:* As a first task within the iteration step of a line search method, a descent direction d is determined, along which the value of the objective function shall be decreased. Second we find a step size α that decides how far to move along the direction to obtain a minimization of the function value. The value of α can be optimized exactly but often an approximation reaching a decrease in the objective function value is sufficient. In fact, most line search algorithms also solve approximate models to obtain search directions, e.g., Gauß-Newton.

(ii) *Trust Region:* A trust region method uses a (in most cases quadratic) surrogate model function to approximate the true objective function. It is crucial for this kind of approaches that this approximation is only "trusted" to be appropriate over a subset of the search space centered around the current iterate, the so-called trust region. In the next step, the trust region is either expanded in case that the surrogate function proves to be a sufficiently good approximation and successfully minimizes the actual objective function or otherwise it is contracted if the approximation via the model function fails to decrease the objective function. Then the surrogate optimization problem is adjusted and repeatedly solved.

We notice the duality between trust region methods and line search methods from the perspective that they execute the two main steps of descent methods in the reversed order: line search methods determine a step direction and then set an appropriate step size for the chosen direction while trust region methods first find an acceptable step size determined by the size of the trust region and then decide for a step direction.

As an example, we will point out the most popular algorithm for solving non-linear least squares problems referred to as the *Levenberg-Marquardt* algorithm [94, 108]. It actually was also the very first trust region algorithm to be developed and can be considered as a trust-region modification of the Gauss-Newton algorithm with the subproblem

$$d_{LM}^{(n)} = \arg\min_{d\in\mathbb{R}^d} \frac{1}{2}\|J(x^{(n)})d + e^{(n)}\|_{\ell_2}^2, \text{ subject to } \|d\| \leq \Delta^{(n)}. \tag{2.29}$$

The Levenberg-Marquardt algorithm exhibits the same local convergence behaviour as the Gauß-Newton method and is at least locally equivalent to a linear least squares problem, but it has the major advantage that it is also globally convergent.

Now we end our detour to the solution of the nonlinear least squares problem and make a step towards optimality results for functions that are nondifferentiable but convex and, therefore, still have favourable optimization properties.

We first generalize the gradient to nondifferentiable functions with the definition of the subdifferential.

Definition 2.13. The subdifferential of a function $F : \mathbb{R}^d \to (-\infty, \infty]$ at a point $x \in \mathbb{R}^d$ is defined by

$$\partial F(x) = \left\{ y \in \mathbb{R}^d : f(\bar{x}) \geq F(x) + \langle y, \bar{x} - x\rangle \text{ for all } \bar{x} \in \mathbb{R}^d \right\}.$$

The elements of $\partial F(x)$ are called subgradients of F at x.

In the case that F is differentiable at a point x, it holds that $\partial F(x) = \{\nabla f(x)\}$, which means that the subdifferential $\partial F(x)$ consists of one single element, the gradient of F at x. The subdifferential ∂F can also be the empty set but for a convex function $F : \mathbb{R}^d \to \mathbb{R}$ it is always nonempty.

To give an illustrative example of a function with a nontrivial subdifferential, we consider $F(x) = |x|$, where $\partial F = \begin{cases} \{\mathrm{sign}(x)\} & x \neq 0 \\ [-1, 1] & x = 0 \end{cases}$.

Now, interestingly, the subdifferential allows a simple characterization of minimizers of convex functions via a first-order type necessary and sufficient condition as follows:

Theorem 2.14. *A vector x is a minimizer of a convex function F if and only if $0 \in \partial F(x)$.*

For the rest of the subsection, we will turn to constrained optimization problems. We will concentrate on specific constraint sets $\mathcal{C}$, that can be described by equalities or inequalities.

We define the nonlinear optimization problem with equality/inequality constraints

$$\begin{aligned} &\text{minimize } F(x) \\ &\text{subject to } h_k(x) = 0, k \in [m_1] \\ &\qquad\qquad g_l(x) \leq 0, l \in [m_2] \end{aligned} \tag{2.30}$$

where F, $h_1, h_2, \ldots, h_{m_1}$ and $g_1, g_2, \ldots, g_{m_2}$ are continuously differentiable functions from $\mathbb{R}^d$ into $\mathbb{R}$.

Moreover, we introduce the operators $H : \mathbb{R}^d \to \mathbb{R}^{m_1}$ and $H(x) = (h_1(x), h_2(x), ..., h_{m_1}(x))$ and analogously $G : \mathbb{R}^d \to \mathbb{R}^{m_2}$ and $G(x) = (g_1(x), g_2(x), ..., g_{m_2}(x))$.

In problems with inequality constraints it is often difficult to determine which inequalities are active in an optimal solution, meaning that they are fulfilled with equality. If we knew the active inequalities, we would essentially have a problem with only equality constraints, $H(x) = 0$ plus the active equalities. The set of indices of the active inequalities at x is denoted by $\Lambda(x)$, so $\Lambda(x) = \{l \leq m_2 : g_l(x) = 0\}$.

In the following, a point x is called regular if $\{\nabla h_1(x), \ldots, \nabla h_{m_1}(x)\} \cup \{\nabla g_l(x) : l \in \Lambda(x)\}$ is linearly independent.

We now present a main result in nonlinear optimization which gives the necessary and sufficient first-order optimality conditions to the problem (2.30), the so-called Karush-Kuhn-Tucker conditions, or simply the KKT conditions. Additionally, we will give second-order conditions to obtain a full presentation of necessary and sufficient conditions on optimality.

In order to present these conditions, we introduce the Lagrangian function $L : \mathbb{R}^d \times \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \to \mathbb{R}$ given by

$$L(x, \lambda, \mu) = F(x) + \sum_{k=1}^{m_1} \lambda_k h_k(x) + \sum_{l=1}^{m_2} \mu_l g_l(x) = f(x). \tag{2.31}$$

with Lagrange multiplier vectors $\lambda = (\lambda_1, \lambda_2, ..., \lambda_{m_1})$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_{m_2})$.

The gradient of L with respect to x is denoted by

$$\nabla_x L(x, \lambda, \mu) = \nabla F(x) + \sum_{k=1}^{m_1} \lambda_k \nabla h_k(x) + \sum_{l=1}^{m_2} \mu_l \nabla g_l(x). \tag{2.32}$$

and the Hessian matrix of L at (x, λ, μ) by $\nabla_{xx} L(x, \lambda, \mu)$ containing the corresponding second order partial derivatives of the Langrangian with respect to x.

The upcoming theorem first presents the first-order conditions, known as the KKT conditions, as well as second order conditions that are subsequently put in context and explained.

Theorem 2.15. *Let $x^* \in \mathbb{R}^d$ and assume that f, $h_1, h_2, \dots, h_{m_1}$ and $g_1, g_2, \dots, g_{m_2}$ are twice continuously differentiable functions from $\mathbb{R}^d$ into $\mathbb{R}$. We consider the*

(i) first-order necessary and sufficient conditions (KKT conditions):

$$\begin{aligned} \nabla_x L(x^*, \lambda^*, \mu^*) &= 0 & & \\ \mu_l^* &\geq 0 & l &\in [m_2] \\ \mu_l^* &= 0 & l &\notin \Lambda(x^*). \end{aligned} \tag{2.33}$$

(ii) second-order necessary condition:

$$y^T \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) y \geq 0 \tag{2.34}$$

for all y with $\nabla h_k(x^*)^T y = 0$ for $k \in [m_1]$ and $\nabla g_l(x^*)^T y = 0$ for $l \in \Lambda(x^*)$.

(iii) second-order sufficient condition:

$$y^T \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) y > 0 \tag{2.35}$$

for all y with $\nabla h_k(x^*)^T y = 0$ for $k \in [m_1]$ and $\nabla g_l(x^*)^T y = 0$ for $l \in \Lambda(x^*)$.

If x^* is a local minimizer of (2.30) *and x^* is a regular point, then there are unique Lagrange multiplier vectors $\lambda^* = (\lambda_1^*, \lambda_2^*, \dots, \lambda_{m_1}^*)$ and $\mu^* = (\mu_1^*, \mu_2^*, \dots, \lambda_{m_2})$ such that necessarily conditions (i) and (ii) hold.*

In case that x^*, λ^* and μ^* are such that x^* is a feasible point and (i) and (iii) hold, then x^* is a local minimizer of (2.30)*.*

2.2.3 Optimization of surrogate functionals

Surrogate-based optimization [82, 122] refers to a branch of optimization frameworks for the development of problem-driven algorithms that use *surrogate model problems* to approximate the solution of an otherwise computationally intractable problem. More concretely, surrogate methods iteratively solve a sequence of optimization problems at a low computational cost, that typically consist of two main steps: First a model function $G(x)$ with favourable optimization properties is determined to substitute the original objective function $F(x)$. Then computationally efficient optimization techniques are applied for this surrogate function, hoping for a good approximation of the optimizer of the true objective function.

Of course, the design of the surrogate function is crucial for the success of the surrogate model algorithms and one needs to find an appropriate trade-off between the following two ambitions: On the one hand, the tighter the approximation of $G(x)$ to the objective $F(x)$, the faster the convergence and therefore, the more efficient the derived algorithm. On the other hand, a closed-form solution to the optimization problem resulting from the surrogate model is preferable.

Finding the right balance between the two often contrary goals mentioned above requires the smart application of inequalities bounding the objective function for the construction of surrogate functions.

Some general principles and methods for the construction of surrogate functions, just mentioning here three important representatives, can be

(i) *Separation in variables*: allowing parallel computing implementations in high dimensions.

(ii) *Convexity and smoothness*: favourable optimization properties, e.g., via linearization of the concave part of the objective or Taylor expansion.

(iii) *Special inequalities*, e.g., Arithmetic-Geometric Mean Inequality, Cauchy-Schwartz Inequality, Jensen's Inequality.

For discussion of these techniques and further examples we refer to Lange et al. [90] In the following, we want to give an example for the class of so-called *maximation-minimization* algorithms (MM-algorithms). For this special type of methods the surrogate function is "maximizing" the objective, meaning that it is a local upperbound approximation of the objective function, where the difference is minimal in the point of the current iterate [89].

Example [Convexification via quadratic perturbation]:
In the case we want to optimize a nonconvex objective function $F(x)$, one can try to approach this problem iteratively solving a sequence of locally convex problems with a maximation-minimization strategy. More concretely, we consider, starting from an initial point $x^{(0)}$, to perform a majorizing local convexification of the objective $F(x)$ around the current iterate $x^{(n)}$ at each iteration $n > 0$.

We propose an appropriate convexification $F_{\mu,u}$ by quadratic perturbations

$$F_{\mu,u}(x) := F(x) + \mu\|x - u\|_{\ell_2}^2 \tag{2.36}$$

for $\mu > 0$ and a reference point $u \in \mathbb{R}^d$, which is rather standard and well-known in the nonlinear optimization literature, for instance in sequential quadratic programming [3]. Notice that $F_{\mu,u}(x)$ is coercive whenever F is bounded from below.

Related to this type of convexification we define μ-convexity of a function

Definition 2.16. Let $F : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function with piecewise continuous and bounded second derivatives. We say F is μ-convex if there exists $\mu > 0$ such that $F(\cdot) + \mu\|\cdot\|_{\ell_2}^2$ is convex.

We observe that, if F is μ-convex we can always assume that μ is chosen in such a way that also $F_{\mu,u}(x)$ is ν-strongly convex with ν depending on F and μ, but not on u. [2]

The function $F_{\frac{1}{2},z}$ serves as a foundation for the formulation of the iterative proximal point algorithm [89]

$$x^{(n+1)} = \operatorname{prox}_f^{\mu}(x^{(n)}),$$

where the so-called proximal map is defined by

$$\operatorname{prox}_F^{\mu}(z) = \arg\min_x F_{\frac{1}{2},z} = \arg\min_x \left[F(x) - \frac{1}{2}\|x - z\|_{\ell_2}^2\right].$$

Its scaled version for $\mu > 0$ is often referred to as the the Moreau-Yosida regularization

$$F_{MY}^{\mu}(z) = \arg\min_x F_{\mu,z} = \arg\min_x \left[F(x) - \mu\|x - z\|_{\ell_2}^2\right].$$

Now we want to continue with the concepts of relaxation and convex envelopes, which also belongs to the class of surrogate optimization methods, but one considers a minorization of the actual objective function [16].

Definition 2.17. Let $\mathrm{dom}(F) \subset \mathbb{R}^d$ be the domain of the function $F : \mathbb{R}^d \to \mathbb{R}$. A relaxation of the minimization problem

$$\min_{x \in \mathrm{dom}(F)} F(x) \tag{2.37}$$

is a surrogate minimization problem

$$\min_{x \in X_R} F_R(x) \tag{2.38}$$

where $X_R \supseteq \mathrm{dom}(f)$ and $F_R(x) \leq F(x)$ for all $x \in X$. Let x^* denote an optimal solution of the original minimization problem and $x_R^* \in X_R$ the optimal solution of the surrogate problem, then $x^* \in X \subseteq X_R$ and $F(x^*) \geq F_R(x^*) \geq F_R(x_R^*)$.

A special, quite important instance of a relaxation of an optimization problem is its convex relaxation, which chooses the convex envelope of a function as the relaxation function.

Definition 2.18. The *convex envelope* $\bar{F}$ of a function F is the convex approximation of F from below

$$\bar{F}(x) := \sup \{ g(x) \leq F(x) : g \text{ is a convex function} \}.$$

The convex minimization problem for $\bar{F}$ on a compact set $\mathcal{C}$ has a unique global solution, which will coincide with the a global minimizer on $\mathcal{C}$ of the relaxed function F, although F might have several local minimizers.

2.3 High-dimensional data: fighting the curse of dimensionality

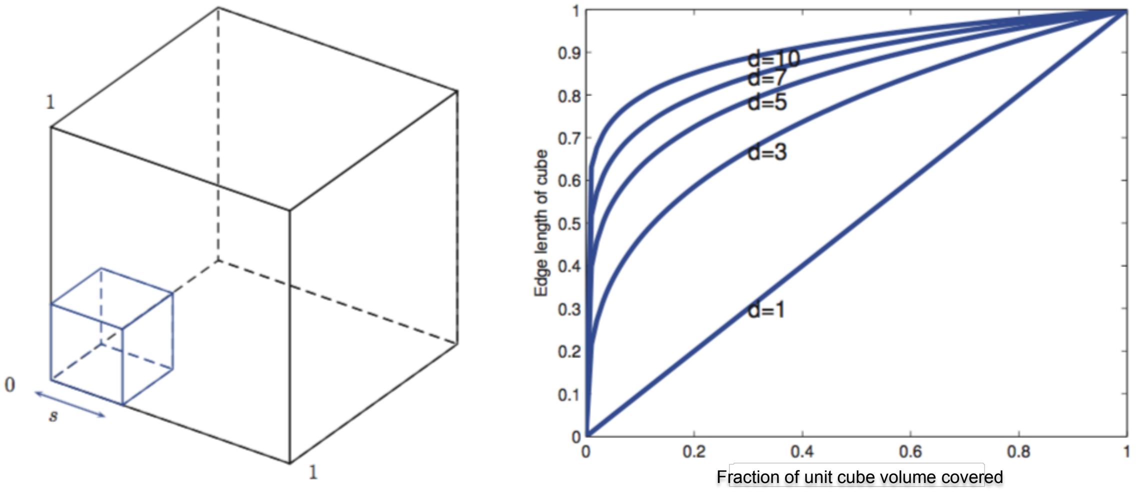
Many real-life problems naturally involve high-dimensional signal vectors, and we face the case that the number of measurements m is way less than the number of variables d we need to estimate, i.e., we consider the (linear) equation system

$$\Phi(x) = y \tag{2.39}$$

with $x \in \mathbb{R}^d$, $y \in \mathbb{R}^m$ and $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ where $m \ll d$ as typical, for example in image processing [60, 139].

Theoretical considerations motivate the study of high-dimensional spaces, which evince unexpected properties that completely differ from our intuition for low-dimensional spaces as 2D or 3D, often entitled as the *"curse of dimensionality"*.

The methaphorical expression "curse of dimensionality" appeared first in the work of Bellman [6]. He used this term in connection with the difficulties arising from the overwhelming number of function evaluations necessary for the optimization of a continuous function to a certain accuracy by searching on a discrete grid with growing number of function parameters. If for example one searches the optimum of a function on the unit cube in dimension d using a Cartesian grid of spacing $\frac{1}{s}$, one has to carry out s^d function evaluations.



(a) (b)

Figure 2.5: Visual explanation of the exponential growth of the space volume with the dimension causing the "curse of dimensionality" (a) Embedding of a cube of side length $s \in [0, 1]$ into the unit cube of dimension d (b) Fraction of the volume of the unit cube covered by the embedded cube in dependence on the edge lengths. Picture source: [109]

Hence in this scenario, without further simplifying assumptions, the number of data samples required grows exponentially with the dimension. As usually only a limited amount of data is available, high-dimensional spaces exhibit an inherent sparsity. This fact, causing the curse of dimensionality in this case, is often referred to as the *empty space phenomenon* [93].

Critical quantities related to the problem dimension or the number of measurements, e.g., computational complexity are growing exponentially as well making the application of most algorithmic approaches impractical.

In order to avoid or dampen the confrontation with the "curse of dimensionality" when working with high-dimensional data, one assumes that the significant influence parameters are in a interdependent relationship inducing a certain data structure. The geometry of this data structure in the ambient space provides a concise description of the data information content[100].

The goal of dimensionality reduction is to make use of these dependencies to find a lower complexity representation of the data but not loosing the structural information.

2.3.1 Low-dimensional subspace structures

In the case of (2.39), where $m < d$, from a traditional linear algebra perspective the knowledge of Φ and y only does not permit the calculation of x. Nevertheless, such underdetermined linear equation systems appear in the modeling of many practical application problems where *additional knowledge* about x is available.

In a lot of relevant cases, the signal $x \in \mathbb{R}^d$ is assumed to be concentrated around a subset Ω of the space representing an underlying structure, that has intrinsically a lower dimension K in the high-dimensional ambient space [151]. The intersection of this subset Ω with the affine solution space $\mathcal{F}(y, \Phi) = \{x | \Phi(x) = y\}$ of dimension $d-m$ then contains the set of qualified solutions.

The motivation justifying such a restriction of the space of eligible signals x to a lower-dimensional subset is the following: in many cases of interest, the process that is underlying the signal generation essentially only has few degrees of freedom compared to the dimension of the signal [100].

In such cases, there exists a representation of the signal $x \in \mathbb{R}^d$ with a reduced number of degrees of freedom via a parameter vector $z \in \mathbb{R}^K$ in the intrinsic dimension K, while still maintaining the full information content of the original signal x [151].

In a lot of interesting application cases, the subset Ω is supposed to be a manifold that can be smoothly parameterized by the K-dimensional parameter vector $z \in \mathbb{R}^K$, i.e., $\Omega = \left\{x | x = \Psi(z), \Psi : \mathbb{R}^K \to \mathbb{R}^d\right\}$ [45].

For example, manifolds have also been proposed as approximate models for signal classes such as images of handwritten digits [76].

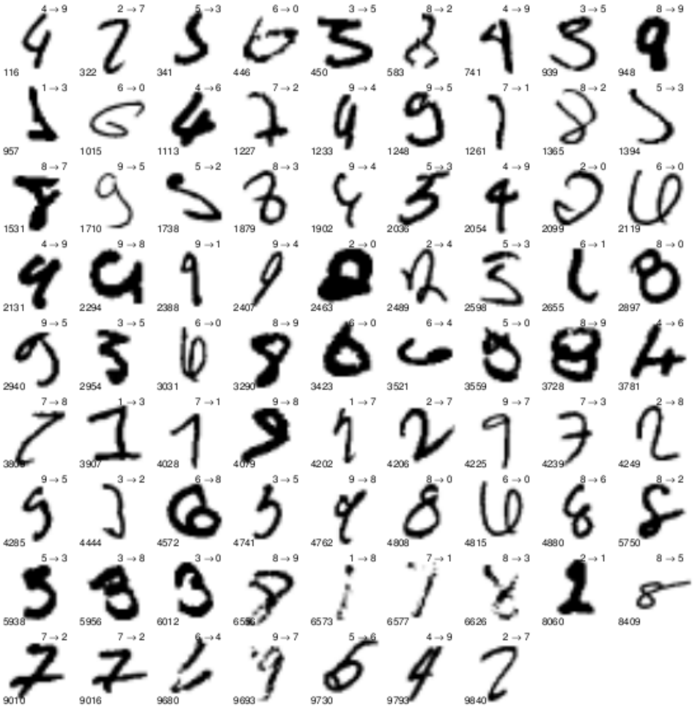
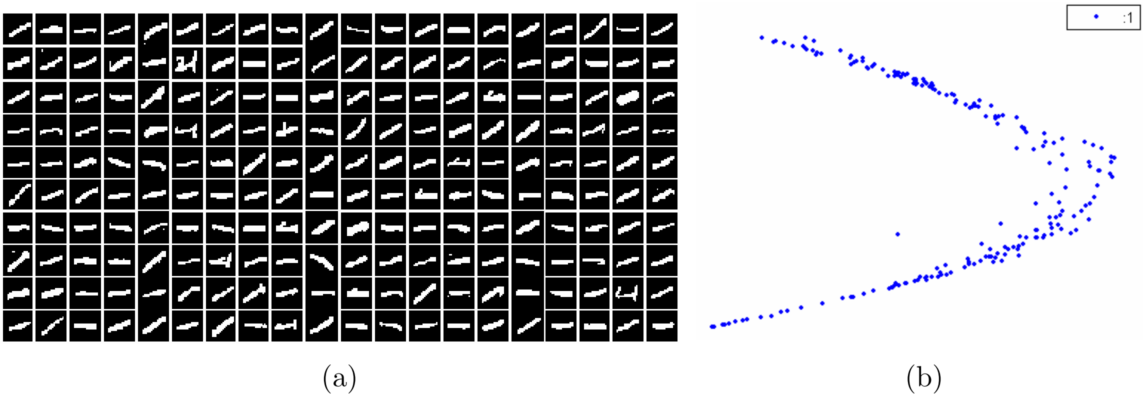


Figure 2.6: Images of handwritten digits: By making assumptions on the vague shape of handwritten signs that represent a certain digit, we reduce the degrees of freedom for the shapes that fall in the class of these digits; this allows for the classification of handwritten digits. Picture source: [102]



(a)

(b)

Figure 2.7: The manifold of the handwritten digit '1': (a) 200 samples of handwritten digit '1' (b) Visualization of the sample points in the shape space of handwritten signs and their accumulation around a lower-dimensional manifold. Picture source: [95]

We make the observation that the so-called manifold hypothesis is a way of avoiding the "curse of dimensionality" [100]: by exploiting the structure with reduced degrees of freedom K of the signal x, it is possible to perform a reduction of dimensionality. Thereby, we obtain a more compact signal representation facilitating the identification of structure specific characteristics of qualified solutions x as well as the formulation of an appropriate penalization function $g(x)$.

Another assumption which makes it possible to circumvent the curse of dimensionality is sparsity[100].

In the following, we focus on this specific type of signal structures which allow for a lower-dimensional representation of vectors $x \in \mathbb{R}^d$ via a k-dimensional coordinate vector in an appropriate basis, so-called k-sparse vectors. Related structural concepts for matrix valued data will be introduced as well.

The general outline and content of the next subsection is inspired by the books [48, 54].

2.3.2 An introduction: sparse vectors and related concepts for matrix valued data

In a large number of application contexts only a part of the components of a signal vector $x \in \mathbb{R}^d$ under consideration is of interest while the rest is negligible. More precisely, for a signal x represented in an appropriate basis, only few indexes $i \in \Lambda \subset [d]$ with $|\Lambda| = k \ll d$ with significantly large absolute values $|x_i|$ correspond to relevant influence parameters of, e.g., a physical phenomenon under consideration in the measurement process. The other indices $i \in \Lambda^c$ with $x_i \approx 0$ can essentially be dropped from the signal model [11, 121].

Such vectors with mainly nearly vanishing entries and only a small number of significant components for an advantageous choice of the basis are called *sparse* and the related described phenomenon is referred to as *sparsity.*

It is an important observation that the degrees of freedom for a signal vector $x \in \mathbb{R}^d$ with a fixed number of only $k \ll d$ nonzero components is only k and, therefore, the dimensionality of the signal model can be reduced.

As a consequence, by introducing the additional assumption of sparsity on the solution vector $x \in \mathbb{R}^d$ and taking advantage of the low-dimensional signal structure, reconstruction of x from the equation system

$$\Phi x = y \tag{2.40}$$

given the measurement result $y \in \mathbb{R}^m$ and a measurement matrix $\Phi \in \mathbb{R}^{m \times d}$ with $m \ll d$, becomes a feasible problem. This problem of finding a sparse solution to an underdetermined linear system became known as the *sparse recovery* or *compressed sensing problem.*

In the 2000's, the work of Candès, Romberg and Tao [25] and Donoho [40] built the ground for a rigorous mathematical theory of sparse recovery. This publication marks the start of the compressed sensing boom. A wide range of efficient algorithmic solvers for the compressed sensing problem is available these days and compressed sensing techniques are successfully applied to numerous relevant real-life problems, e.g. image processing. For an overview of the fast growing field of compressed sensing and further references, cf. [11, 54, 121].

Besides of this very classical type of sparsity for high-dimensional vectors with very few nonzero elements, sparsity can have diverse manifestations that vary from one application domain to another. We access the topic from sparse vector recovery and introduce appropriate convenient notation and problem formulations in the context of sparse vectors. These are applicable with modifications to the matrix case as well.

The concept of sparsity can be applied to the noise vector $e \in \mathbb{R}^m$, but for the rest of the section we focus on the sparse signal vectors $x \in \mathbb{R}^d$ and come back to noise or residual vectors later in Chapter 3.

In the following, let $k, d \in \mathbb{N}$ and $k \leq d$.

For an index set $\mathcal{I}$, we consider subsets $\Lambda \subset \mathcal{I}$ with its cardinality $|\Lambda|$ and its complement $\Lambda^c = \mathcal{I} \setminus \Lambda$. The submatrix of a matrix $Z \in \mathbb{R}^{d_1 \times d_2}$ only containing the columns with indices in $\Lambda = \{i_1, \dots, i_l\}$ is expressed by $Z_\Lambda = [Z_{[:,i_1]}, \dots, Z_{[:,i_l]}] \in \mathbb{R}^{d_1 \times d_2}$. Correspondingly, for the vector z, the restriction $z_\Lambda = [z_{i_1}, \dots, z_{i_l}]$ coincides with the entries of z for the indices contained in the set Λ.

The support, i.e., the set of nonzero coordinates of a vector $x \in \mathbb{R}^d$ is referred to as Λ or $\operatorname{supp}(x)$, i.e., $\Lambda = \operatorname{supp}(x) = \{i \in [d] | x_i \neq 0\}$.
Its cardinality is used to define the so called ℓ_0-*norm*, which actually is the counting measure, by

$$\|x\|_0 = \|x\|_{\ell_0} := |\operatorname{supp}(x)| = \sum_{i=1}^{d} |x_i|_0 \text{, with } |\xi|_0 := \begin{cases} 0 & \xi = 0 \\ 1 & \xi \neq 0. \end{cases}$$

So the ℓ_0-norm counts the number of nonzero entries in a vector.
Using this notion, we are able to give a quite natural, concrete formulation of sparsity.

The set of *k-sparse vectors*, i.e., vectors with at most k nonzero entries, is defined by

$$\Sigma_k = \left\{ x \in \mathbb{R}^d : \|x\|_{\ell_0} := |\mathrm{supp}(x)| \leq k \right\}. \tag{2.41}$$

We note that the set Σ_k constitutes a union of k-dimensional linear subspaces in the space $\mathbb{R}^d$ with ambient dimension d.

The ℓ_0-norm is not only nonconvex, nonsmooth but also discontinuous. It is in many cases approximated by the ℓ_p-(quasi-)norm for a parameter $p > 0$

$$\|x\|_p = \|x\|_{\ell_p} := \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}}. \tag{2.42}$$

For $0 < p < 1$, we gain continuity but still have a nonconvex and nonsmooth quasi-norm, that does not fulfill all norm axioms, as the triangle inequality only holds with factor $C = 2^{1/p-1}$, while for $p \geq 1$ we obtain a norm, which is convex.

We note that

$$\lim_{p \to 0} \|x\|_{\ell_p}^p = |\mathrm{supp}(x)| = \|x\|_{\ell_0}. \tag{2.43}$$

We consider the case that signal information is obtained from *linear, nonadaptive* measurements using a measurement matrix $\Phi \in \mathbb{R}^{m \times d}$, where we expect a sparse solution but the signal x and its support $\mathrm{supp}(x)$ are unknown a priori. The most immediate approach is to search for the vector x with smallest support compatible with the measured data $y = \Phi x$.

As a support of low cardinality is desired, it comes natural to use an ℓ_0-quasinorm penalty function $g(x)$ for the formulation of an appropriate objective functional. This leads us to the so called ℓ_0-*minimization* problem

$$\min_{x \in \mathbb{R}^d} \|x\|_0 \text{ subject to } \Phi x = y \tag{2.44}$$

and hopefully its solution coincides with the vector in demand.

The theoretical setting that has to be complied to ensure the recovery of the correct solution is discussed in more detail in Section 2.3.3.

It should also be mentioned that, in contrast to the ℓ_2-norm, that is unitarily invariant, other norms usually are not, and also the ℓ_0-quasinorm related definition of "sparsity" is highly dependent on the underlying basis. The definition in (2.41) is assuming that the vector under consideration is sparse or nearly sparse in the canonical basis and therefore is formulated in terms of the cardinality of the support set, that is modified under basis change.

In numerous real-life measurement settings, the vectors $\tilde{x}$ of interest living in an Euclidean space $\mathbb{R}^n$ with $n \ll d$ have a *sparse representation* with respect to a suitable basis or frame $\{\varphi_j \in \mathbb{R}^n, j = 1, \cdots, d\}$. This means that $z \in \mathbb{R}^n$ can be represented such that $z = \sum_{j=1}^{d} x_j \varphi_j$, where x has *a small relevant set of indices with nonzero entries while the rest of the components are (nearly) zero* providing a sufficiently good approximation to the expansion of z.

The sparse representability of real-life signals explains the increasing popularity and successful practical applications of compressed sensing approaches, including very classical examples as image processing, where images are known to be sparsely representable with respect to Wavelets, Curvelets or Shearlets for instance [138].

Therefore, in practical problems, often a basis transformation $\Psi = \{\varphi_1, \ldots, \varphi_d\} \in \mathbb{R}^{d \times d}$ is required that allows for an appropriate sparse representation x of z by considering $z = \Psi x$.

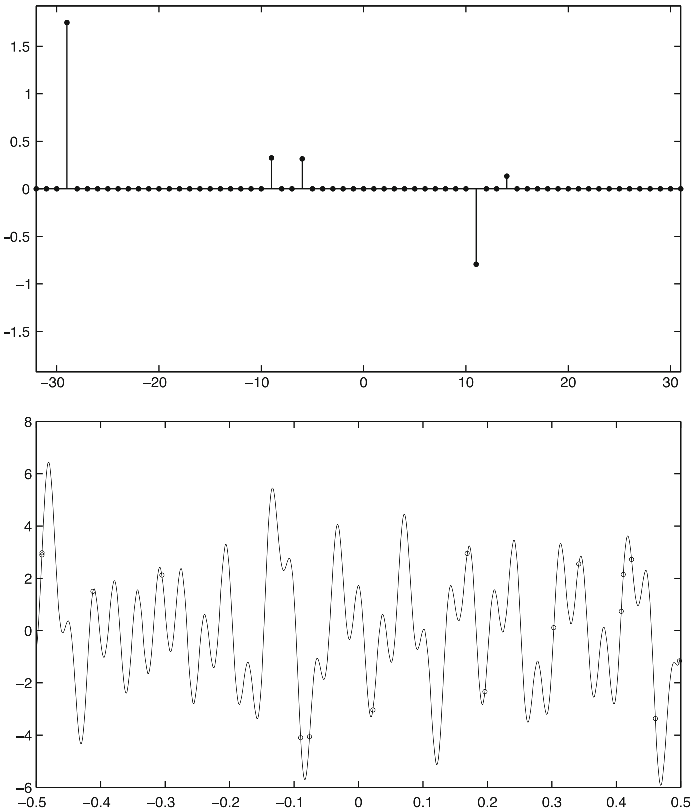


Figure 2.8: Top: 5-sparse vector of Fourier coefficients of length 64. Bottom: real part of the time-domain signal with 16 samples. Picture source: [54, Fig. 1.2]

In these cases instead of the linear system $\Phi z = y$ one considers the substitute linear system $\Phi z = \Phi\Psi x =: \widetilde{\Phi} x = y$ and the corresponding optimization problem

$$\min_{x\in\mathbb{R}^d} \|x\|_0 \text{ subject to } \widetilde{\Phi} x = y. \tag{2.45}$$

Such aspects require a careful modeling of the underlying problem for the correct application of sparse recovery methods, which is not subject of this work. In the following, we assume without loss of generality sparsity in the canonical basis of the solution vectors under consideration.

Unfortunately, computing the sparsest solution from (2.44) or (2.45) directly is a combinatorial optimization problem and, therefore, NP-hard [54, Theorem 2.1]. This means that it in general requires prohibitive computations of exponential growing complexity with respect to k, m, d. As a consequence, the solution of (2.44) quickly becomes computationally intractable with growing dimensions, especially for big data problems.

In the following, we propose to find tractable algorithms by noting that the function $|\cdot|_p$ is a continuous relaxation of $|\cdot|_0$ for $p > 0$ and relaxing (2.44) [133],[110].

In this spirit, also the observation (2.43) suggests to replace the problem (2.44) by the approximation of the ℓ_0-objective functional by ℓ_p-quasinorms for parameters $p > 0$. More precisely, we consider

$$\min_{x\in\mathbb{R}^d} \|x\|_p \text{ subject to } \Phi x = y. \tag{2.46}$$

Our hope is that by solving this relaxed problem, its solution is close to the solution of (2.44) as well.

In the case where $p > 1$, this optimization problem is convex, but it is not guaranteed to find sparse vectors and therefore, is suitable only with some reservations. Recovery of sparse vectors becomes possible for $p \leq 1$ if the solution is sparse enough and Φ fulfills certain spectral properties as we will explain in Section 2.3.3.

For the particular case of $p = 1$, probably the most studied case, this problem becomes the well-known ℓ_1-*norm minimization* (ℓ_1-*minimization*) problem. The ℓ_1-norm constitutes the convex relaxation of the ℓ_0-quasinorm, which makes it relatively easy to solve with standard linear programming techniques, e.g., interior point methods. The combination of these two properties makes the solution of the ℓ_1-minimization problem a very attractive choice also for practical applications [11, 121].

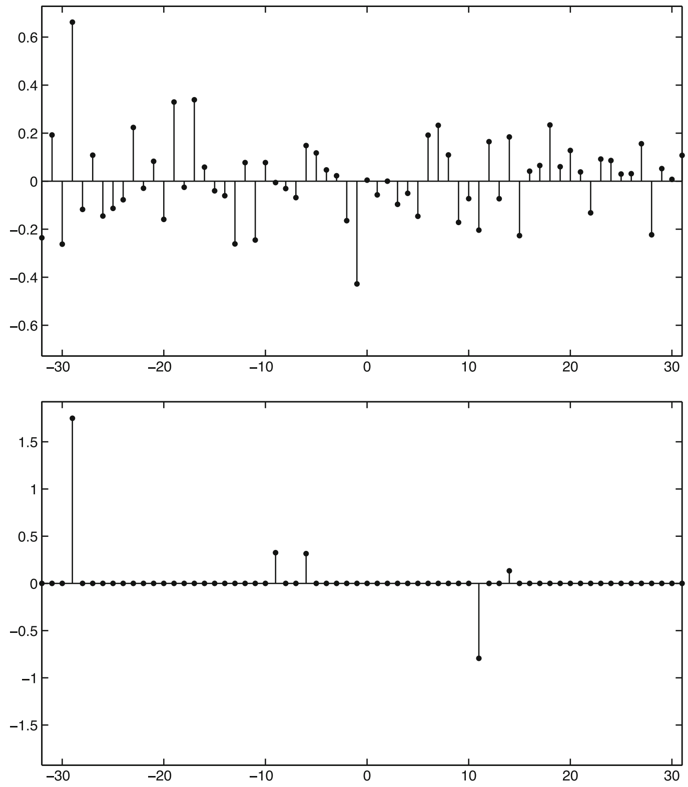


Figure 2.9: Top: poor reconstruction via ℓ_2-minimization. Bottom: exact reconstruction via ℓ_1 -minimization [54, Fig. 1.3]

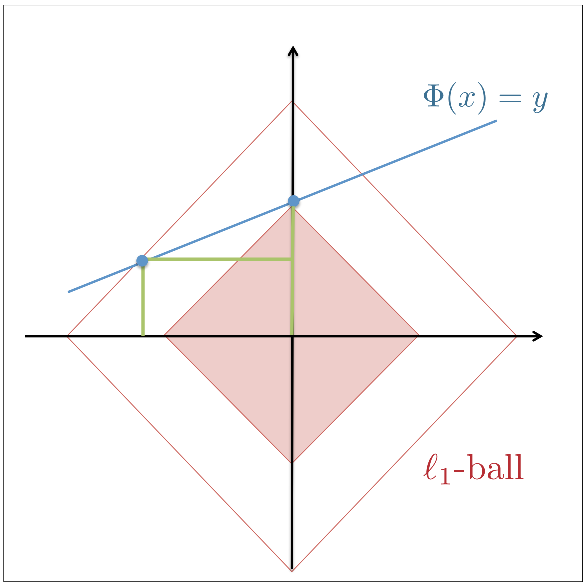


Figure 2.10: Visualization of the sparsity enhancing property of the ℓ_1-norm penalty, compare with Figure 2.3 for a visualization of ℓ_2-norm penalty minimization

For the parameter values $0 < p < 1$, the optimization problem is nonconvex and local minimizers can occur. Finding the global minimizer is again NP-hard as well as the

ℓ_0-norm minimization problem. Nonetheless, the properties of the ℓ_p -minimization for $0 < p < 1$ can prove useful from a theoretical point of view as the approximation to the ℓ_0-problem is closer, which can also bring practical advantages in the end.

We would like to illustrate in a simplified example in $\mathbb{R}^2$ now, why the ℓ_p-minimization is able to induce sparsity for $0 < p \leq 1$.
In the case for the signal dimension $d = 2$ and number of measurements $m = 1$, as a representation for the solution space $\mathcal{F}(y, \Phi) = \{z : \Phi z = y\}$ one can simply take an affine line in $\mathbb{R}^2$. Moreover, the sparsest solution in this case only has one single nonzero component. Visually explained, the nonconvex ℓ_p-norm-balls can be expanded from small size until one of its spikes meets the affine space $\mathcal{F}(y, \Phi)$ in a sparse solution, while for $p > 1$ the first occurrence of tangency of the affine space to the norm-ball will yield a non-sparse solution.

Our intuition tells us that the ℓ_p-minimization problem does not provide a worse approximation of the sparse solution than the solution of the original ℓ_0-minimization problem for small p. However, we still have to justify this conjecture as it is not yet clear when a global minimizer of (2.46) really coincides with a solution to (2.44). For this purpose, we will introduce the so-called null space property and important related matrix properties in the upcoming Section 2.3.3.

We note that many ideas from compressed sensing were recently applied to the recovery of matrices with certain sparsity-type structures from incomplete linear measurements [21, 125]. We now introduce some popular instances of structured matrices and generalize the concepts and notations presented above for sparse vectors accordingly. A comparable collection and presentation of sparsity-type structures and their modeling can be found in [17] and [84]. In Chapter 4 and Chapter 5 we will discuss compressed sensing algorithms and applications related to these sparsity structures in detail.

Let $X \in M_{d_1 \times d_2}$ be an arbitrary matrix and let $X_1, \ldots, X_{d_2} \in \mathbb{R}^{d_1}$ denote its columns and $X_1^T, \ldots, X_{d_1}^T \in \mathbb{R}^{d_2}$ its rows respectively.

First we assume the columns of $X \in M_{d_1 \times d_2}$ to have a common sparsity pattern, a common support $\Lambda = \operatorname{supp}(X_1) = \cdots = \operatorname{supp}(X_{d_2})$ of cardinality $K < d_1$ in the sense of vector sparsity. We call this common support the *row-support* (or row-sparsity pattern) of sparsity K, i.e., there exists a set of coordinates $\Lambda_{row} := \operatorname{supp}_{row}(X) \subset [d]$ with $|\Lambda| = K$ such that

$$\Lambda_{row}^c := [d_1] \setminus \Lambda_{row} = \{i \in [d_1] \mid X_{[i,j]} = 0 \text{ for all } j \in [d_2]\}, \tag{2.47}$$

or equivalently, $\Lambda_{row} = \{i \in [d_1] \mid \|X_i\|_{\ell_2} > 0\}$.

The set of *K-row-sparse matrices*, i.e., matrices with row-support size at most K, is defined by

$$\Sigma_K^{row} = \{X \in M_{d_1 \times d_2} : |\text{supp}_{row}(X)| \leq K\}. \tag{2.48}$$

Moreover, the counterpart of the ℓ_0-quasinorm is the function $\|\cdot\|_{\ell_{2,0}} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}_{\geq 0}$ with $\|X\|_{2,0} := \|X\|_{\ell_{2,0}} := |\Lambda| = |\{i \in [d_1] \mid \|X_i\|_{\ell_2} > 0\}|$, which we call the $\ell_{2,0}$-*norm.*

Similarly, we generalize the ℓ_p-quasinorm, for $0 < p < \infty$ and $q \geq 1$, and define the *(mixed)* $\ell_{q,p}$*-norm of* X as the non-negative number $\|X\|_{q,p} := \|X\|_{\ell_{q,p}} := \left(\sum_{i=1}^{n_1} \|X_i\|_q^p\right)^{1/p}$. We note that for the parameter choice $p = q = 2$ we obtain the Frobenius norm, meaning $\|X\|_{2,2} = \|X\|_F$.

At this point, we remember the discussion in the context of sparse vectors using the low-dimensional signal structure to find solutions to underdetermined linear systems. We see that for the setting of row-sparse matrices also the reconstruction of a matrix X from the equation system

$$\Phi(X) = Y \tag{2.49}$$

with $Y \in \mathbb{R}^m$ and $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m$ with $m \ll d_1 d_2$, becomes possible under the structural assumption on the solution matrix and under appropriate conditions on the map Φ, cf. Section 2.3.3 This results in solving the optimization [98] problem

$$\min \|X\|_{2,0} \text{ subject to } \Phi(X) = Y. \tag{2.50}$$

The relaxed formulation of the problem above

$$\min \|X\|_{2,p} \text{ subject to } \Phi(X) = Y \tag{2.51}$$

corresponding to the ℓ_p-norm minimization problem for sparse vectors can be considered as a proxy for (2.50) as well.

Completely analogously, we can define the concept of *column-sparsity* along with the $\|\cdot\|_{\ell_{0,2}}$-quasinorm as $\|\cdot\|_{\ell_{0,2}} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}_{\geq 0}$ with $\|X\|_{\ell_{0,2}} := |\Lambda| = |\{i \in [d_1] \mid \|X_i^T\|_{\ell_2} > 0\}|$. For a relaxed version of $\|\cdot\|_{\ell_{0,2}}$ we employ the (mixed) $\ell_{p,q}$-norm with appropriately chosen parameters $0 < p < \infty$ and $q \geq 1$, where p and q take the reversed roles in the definition of the quasinorm above.

In the course of this thesis we will focus on the use of the $\|\cdot\|_{\ell_{p,2}}$-norm and $\|\cdot\|_{\ell_{2,p}}$-norm only.

Another very useful matrix structure results from the application of the sparsity concept to the singular values of a matrix, better known as *low-rankness*, that we will explain

now in detail.
Let $X \in M_{d_1 \times d_2}$ again and for simplicity consider the case $d_1 \geq d_2$. We apply the singular value decomposition (SVD) to X

$$X = USV^*, \tag{2.52}$$

where $U \in U_{d_1}$ and $V \in U_{d_2}$ are unitary matrices and

$$S = \begin{pmatrix} \operatorname{diag}\left[\sigma_1(X), \ldots, \sigma_{d_2}(X)\right] \\ 0_{(d_1-d_2)\times m_X} \end{pmatrix} \in \mathbb{R}^{d_1 \times d_2}$$

the diagonal matrix containing the d_2 singular values $\sigma_1(X), \ldots, \sigma_{d_2}(X)$ of X, where $\sigma_1(X) \geq \sigma_2(X) \geq \ldots \geq \sigma_{d_2}(X) \geq 0$. Using the result of the SVD, we define the singular value vector $\bar{\sigma}(X) = [\sigma_1(X), \ldots, \sigma_{d_2}(X)]$.

Note that in the case that $\bar{\sigma}$ is r- sparse for some small $r \in \mathbb{N}$ with $r \ll d_2$, it holds that $\sigma_{r+1}(X) = \ldots = \sigma_{d_2}(X) = 0$ and as a consequence, it follows that X has low rank, i.e., $\operatorname{rank}(X) = r$. Therefore, it is possible to express the low-rank assumption in terms of the sparsity of the singular value vector $\bar{\sigma}(X)$. There exists a set $\Lambda_{rank} = \{i | \sigma_i(X) > 0\}$, where the sparsity measured by the ℓ_0-norm of the vector $\bar{\sigma}(X)$ corresponds to $\operatorname{rank}(X)$. We define the set of matrices in dimension $d_1 \times d_2$ of fixed rank $r \leq \min(d_1, d_2)$ as

$$M^r_{d_1 \times d_2} := \{X \in \mathbb{R}^{d_1 \times d_2} \mid \operatorname{rank}(X) = r\}. \tag{2.53}$$

Moreover, we define as the equivalent to the ℓ_p-quasinorm for low-rankness the so-called *Schatten-p (quasi-)norm* of $X \in \mathbb{R}^{d_1 \times d_2}$ for $0 \leq p \leq \infty$ as

$$\|X\|_{S_p} := \begin{cases} \operatorname{rank}(X), & \text{for } p = 0, \\ \left[\sum_{j=1}^{\min(d_1,d_2)} \sigma_j^p(X)\right]^{1/p}, & \text{for } 0 < p < \infty, \\ \sigma_{\max}(X), & \text{for } p = \infty. \end{cases} \tag{2.54}$$

It is useful to notice that the p-th power of the Schatten-p norm for $0 < p < \infty$ can be expressed as $\|X\|^p_{S_p} = \operatorname{tr}\left[(X^T X)^{p/2}\right]$, where $\operatorname{tr}[X]$ denotes the trace of X defined by the sum of its diagonal elements, $\operatorname{tr}[X] = \sum_{j=1}^{\min(d_1,d_2)} X_{jj}$.

Let us mention that for $p = 1$, the Schatten-p norm is also called *nuclear norm* sometimes denoted by $\|X\|_* = \|X\|_{S_1}$. The Schatten-2 norm corresponds to the *Frobenius norm*, i.e., $\|X\|_F = \|X\|_{S_2} = \sqrt{\langle X, X\rangle_F}$.

Having in mind the discussion above on the solutions to underdetermined linear systems, the equivalent to (2.44) for the recovery of a solution matrix under the additional assumption of low-rank structure is the solution of the affine rank minimization problem [54]

$$\min \operatorname{rank}(X) \text{ subject to } \Phi(X) = Y. \tag{2.55}$$

The relaxed version approximating (2.55) is formulated replacing the rank by the Schatten-p quasinorm. This results in the problem

$$\min \|X\|_{S_p} \text{ subject to } \Phi(X) = Y, \tag{2.56}$$

which corresponds to an ℓ_p-minimization of the singular values of the matrix X.

At this point, we want to provide the following result originally stated by Wedin [160], which corresponds to a bound on perturbations of the singular value decomposition and will be useful in the context of low- rank matrices later on. It gives a bound on the alignment of the subspaces spanned by the singular vectors of two matrices by their norm distance under the requirement of a sufficiently pronounced gap between the first singular values of the one matrix and the last singular values of the other matrix.

Lemma 2.19 (Wedin's bound [140]). *Let Z and $\bar{Z}$ be two matrices of the same size and their singular value decompositions*

$$Z = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} V_1^* \\ V_2^* \end{pmatrix} \quad \text{and} \quad \bar{Z} = \begin{pmatrix} \bar{U}_1 & \bar{U}_2 \end{pmatrix} \begin{pmatrix} \bar{\Sigma}_1 & 0 \\ 0 & \bar{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \bar{V}_1^* \\ \bar{V}_2^* \end{pmatrix},$$

where the submatrices have the sizes of corresponding dimensions. Suppose that δ, α satisfying $0 < \delta \leq \alpha$ are such that $\alpha \leq \sigma_{\min}(\Sigma_1)$ and $\sigma_{\max}(\bar{\Sigma}_2) < \alpha - \delta$. Then

$$\|\bar{U}_2^* U_1\|_{S_\infty} \leq \sqrt{2}\frac{\|Z - \bar{Z}\|_{S_\infty}}{\delta} \text{ and } \|\bar{V}_2^* V_1\|_{S_\infty} \leq \sqrt{2}\frac{\|Z - \bar{Z}\|_{S_\infty}}{\delta}. \tag{2.57}$$

We stress again, that the $\ell_{2,p}$-norm and $\ell_{p,2}$-norm as well as the Schatten-p norm are *norms* in the strict sense for $p \geq 1$ and *quasi-norms* for $0 < p < 1$, i.e. they fulfill the norm axioms except for the triangle inequality. Their relations to the Frobenius norm will become of major importance in the next Section 2.4.

By introducing the mentioned structural assumptions on matrices in $M_{d_1 \times d_2}$, we reduce the $d_1 \cdot d_2$ degrees of freedom of a general $d_1 \times d_2$-matrix considerably.

In the case of row or column sparsity, it is straightforward that the number of degrees of freedom is equivalent to the row support size times the number of columns $K \cdot d_2$ or the column support size times the number of rows $K \cdot d_1$, respectively. The set of row sparse matrices as well as the set of column sparse matrices correspond to a union of linear subspaces of these dimensions in the $d_1 \cdot d_2$-dimensional ambient space.

For the low-rank case, the set of $d_1 \times d_2$-matrices of fixed rank r corresponds to a *submanifold* of the $d_1 \cdot d_2$-dimensional space of dimension $r(d_1+d_2-r)$. The singular value decomposition of X can serve as a tool to calculate the dimensionality as demonstrated in [84, Lemma 3.1].

Lemma 2.20. *The number of degrees of freedom of a real matrix $X \in M^r_{d_1 \times d_2}$ of size $d_1 \times d_2$ with rank r is $r(d_1 + d_2 - r)$.*

2.3.3 Theoretical foundations of vector and matrix valued compressed sensing

We noted in the last chapter, that under certain conditions on the measurement matrix Φ and on the sparsity of the original signal vector x, the vector recovered by (2.46) coincides with the sparsest solution x to the equation system (2.40) and, therefore, also with the solution of (2.44).

In the following subsection, we investigate a necessary and sufficient condition for the exact reconstruction of every sparse vector x as a solution of the ℓ_p-minimization problem (2.46) called the null space property. Moreover, we introduce a popular related near isometry matrix property known as the *restricted isometry property.* The subsection is oriented at the presentation in [54] and [48].

We again start from the case of a linear measurement operator Φ and a vector valued signal x to give a first intuition of the most important tools in the field. Thereafter, we extend the results to sparsity structures for matrices as mentioned above.

As a first step, we introduce further notation for the case of nearly sparse vectors and the approximation of sparse solutions that will be useful later on.

In practice, in many measurement settings, the occurence of noise is typical and exactly sparse vectors are often not realistic. Therefore, we aim at vectors x, which are not necessarily exactly sparse but very close to an element of Σ_k in the sense of a suitable (quasi-)norm $\|\cdot\|_{\ell_p}$, which we call *compressible.*

This implies that a compressible vector should have a fast decaying *k-term best approximation error*, defined by

$$\beta_k(x)_{\ell_p} = \inf_{z \in \Sigma_k} \|x - z\|_{\ell_p} \qquad 0 < p < \infty.$$

The related *best k-term approximation* $x_{[k]}$ of a vector x constitutes the minimal distance of x to a k-sparse vector as follows

$$x_{[k]} = \arg\min_{z \in \Sigma_k} \|x - z\|_{\ell_p}, \quad 0 < p < \infty.$$

The best k-term approximation error basically results from setting to zero its $d-k$ smallest coefficients. Moreover, we define the *ϵ-smoothed ℓ_p-norm* as

$$\|x\|_{\ell_p,\epsilon} = \left(\sum_{i=1}^{d} (x_i^2 + \epsilon^2)^{\frac{p}{2}} \right)^{\frac{1}{p}} \tag{2.58}$$

In some cases, it is useful to describe sparse vectors by its *nonincreasing rearrangement* $r(x)$, for which

$$r(x)_1 \geq r(x)_2 \geq \ldots \geq r(x)_d \geq 0, \tag{2.59}$$

and there is a permutation $\pi : [d] \rightarrow [d]$ with $r(x)_i = |x_{\pi(i)}|$ for all $i \in [d]$.

This corresponds to the arrangement of the vector components according to their magnitude and, therefore, in the ranking of all entries and their corresponding indices according to their significant contribution to the signal. It can be used to express

$$\beta_k(x)_{\ell_p} = \left(\sum_{i=k+1}^{d} r_i(x)^p \right)^{1/p}, \quad 0 < p < \infty. \tag{2.60}$$

Refer to [48],[123] for more details.

As already proclaimed above, ℓ_p-norm minimization is known to perform stable recovery of the sparse solution vector x, in the sense that for the recovered vector x_p it holds that

$$\|x - x_p\|_{\ell_p} \leq c\beta_k(x)_{\ell_p}. \tag{2.61}$$

for a constant $c > 0$.

To verify this statement, we introduce a necessary and sufficient condition for exact recovery of sparse vectors called the null space property for parameter p (p-NSP).

Definition 2.21. A matrix $\Phi \in \mathbb{R}^{m \times d}$ has the *p-Null Space Property (p-NSP) of order* k for $0 < \gamma_k < 1$ if

$$\|\eta_\Lambda\|_{\ell_p}^p \leq \gamma_k \|\eta_{\Lambda^c}\|_{\ell_p}^p$$

for all sets $\Lambda \subset \{1, \ldots, d\}, |\Lambda| \leq k$ and for all $\eta \in \mathcal{N} = \mathrm{Ker}(\Phi)$.

The NSP essentially prohibits the existence of sparse or highly compressible vectors in the null space of Φ. This is a natural requirement, since otherwise no decoder would be able to robustly distinguish a (nearly) sparse vector from zero. This makes the NSP equivalent to stable recovery in the sense of (2.61) [33], [128].

Lemma 2.22. *If $\Phi \in \mathbb{R}^{m\times d}$ fulfills the p-null space property for some $0 < p \leq 1$ of order k with constant γ_k, then Φ also fulfills the q-null space property of same order and constants for all $0 < q < p$. [38, 70]*

Moreover, the p-NSP has the following stability result as consequence.

Lemma 2.23 ([35, Lemma 7.6], [54, Theorem 4.14])**.** *Assume that $\Phi \in \mathbb{R}^{m\times d}$ satisfies the p-NSP of order k with constant γ_k for $0 < p \leq 1$. Then for any vectors $x, x' \in \mathcal{F}(y, \Phi) = \{z : \Phi z = y\}$ it holds*

$$\|x' - x\|_{\ell_p}^p \leq \frac{1+\gamma_k}{1-\gamma_k}\left(\|x'\|_{\ell_p}^p - \|x\|_{\ell_p}^p + 2\beta_k(x)_{\ell_p}^p\right).$$

Next we familiarize the reader with a further interesting near isometry matrix property that is widely used in the context of sparse recovery for the analysis of ℓ_p-minimization.

Definition 2.24. A matrix $\Phi \in \mathbb{R}^{m\times d}$ has the *Restricted Isometry Property (RIP) of order k* if there exists $0 < \delta_k < 1$ such that

$$(1-\delta_k)\left\|x\right\|_{\ell_2} \leq \left\|\Phi x\right\|_{\ell_2} \leq (1+\delta_k)\left\|x\right\|_{\ell_2}$$

for all $x \in \Sigma_k$.

If for a matrix Φ the restricted isometry property is fulfilled, for every index set Λ with $|\Lambda| \leq k$ the submatrix Φ_Λ is well-conditioned or reformulated more explanatory, all columns of Φ with index contained in Λ are nearly orthonormal.

The NSP is an important tool for the analysis of convergence and stability but often difficult to show directly. Therefore, we establish its relationship to the RIP, which can be addressed easier in practise.

We first state that the RIP implies the 1-NSP for simplicity and as a consequence from Lemma 2.22 also the p-NSP for $0 < p < 1$:

Lemma 2.25. *Assume that $\Phi \in \mathbb{R}^{m\times d}$ has the RIP of order $K = k+h$ with $0 < \delta_K < 1$. Then Φ has the 1-NSP of order k and constant $\gamma_k = \sqrt{\frac{k}{h}\frac{1+\delta_K}{1-\delta_K}}$.*

The proof of this lemma can be found, for instance, in [50].

A more precise estimation result of the upper bound on the RIP constant for $0 < p < 1$ is given in Theorem 2.3. in [143]

Lemma 2.26 (Theorem 2.3. in [143]). *Let $0 < p \leq 1$, and m, d and k integers that satisfy $2k \leq m \leq d$, $\Phi \in \mathbb{R}^{m \times d}$ be a matrix fulfilling the restricted isometry property (RIP) with constant $0 < \delta_{2k} < 1$. Then Φ has the p-null space property (p-NSP) of order k with constant γ_k satisfying $\gamma_{2k} \leq b(p, \sqrt{\frac{1-\delta_{2k}}{1+\delta_{2k}}})$, where*

$$b(p,\delta) = \delta^{-1} \inf_{0<r_0<1} \max \left\{ \frac{1 + r_0\delta}{(1 + r_0^q \delta^q)^{1/q}}, \sup_{\sqrt{2}(1-r_0)\delta/2 \leq y \leq 1} \frac{2y}{(1 + 2^{-q/2} y^{2+q})^{1/q}}, \right.$$
$$\left. \sup_{\sqrt{2}(1-r_0)\delta/2 \leq y \leq 1} \frac{3y}{(1+y)^{1/q}, \sup_{1 \leq y} \frac{2y}{(1+y)^{1/q}}} \right\}$$

Further discussions on the relationship between RIP-type assumptions and exact recovery via ℓ_p-minimization can be found in [53].

The RIP does imply the NSP, but the converse is not true. Actually, the RIP is significantly more restrictive.

Summarizing, for matrices satisfying the RIP and, therefore, also the p-NSP, ℓ_p-minimization can provide exact recovery results under stable performance with error bounds as stated above.

Consequently, it is an important question to ask for which classes of matrices the RIP can be shown to hold with optimal constants, i.e.,

$$k \asymp \frac{m}{\log d/m + 1}.$$

Up to now, it was not possible to specify deterministic matrices for which optimal performance can be guaranteed. In contrast, as the RIP is a spectral concentration property, different classes of random matrices can fulfill optimal RIP, at least with high probability. Therefore, mainly random matrices are used in the context of compressed sensing, e.g., the most popular ones among them being Gaussian matrices.

Theorem 2.27. *(Theorem 2.14 in [48]) Suppose that m, d and $0 < \delta < 1$ are fixed. If Φ is a Gaussian random matrix of size $m \times d$, then there exist constants $c_1, c_2 > 0$ depending on δ such that the RIP holds for Φ with constant δ and $k < c_1 \frac{m}{\log d/m+1}$ with probability exceeding $1 - e^{-c_2 m}$.*

Moreover, for structured random matrices, for example, partial random circulant matrices or random partial Fourier matrices, the RIP can be shown to hold with high probability as soon as $m \geq Ck \log^4(N)$ [24, 48, 54, 83, 130]. As the verification of the RIP property for a special types of random matrices can be very demanding, this topic is not covered here in detail. An extensive review on RIP properties also for structured random matrices of other types can be found in [123].

As presented above, many ideas from sparse vector recovery can be applied to the recovery of matrices with sparsity structures from incomplete linear measurements [21, 125].

We want to pick up the matrix structures discussed in the previous subsection and transfer the theoretical tools corresponding to the classical NSP and RIP for sparse vectors also to the concepts of row- and column-sparsity and low-rankness.

Before we start, let us consider the vectorized form $X_{\text{vec}} = \left[X_1^T, \dots, X_j^T, \dots, X_{d_2}^T\right]^T \in \mathbb{R}^{d_1 d_2}$ of a matrix $X \in M_{d_1 \times d_2}$ with columns X_j, $j \in \{1, \dots, d_2\}$. The reverse recast of a vector $x \in \mathbb{R}^{d_1 d_2}$ into a matrix of dimension $d_1 \times d_2$ is denoted by $x_{\text{mat}(d_1,d_2)} = [X_1, \dots, X_j, \dots, X_{d_2}]$, where $X_j = [x_{(d_1-1)\cdot j+1}, \dots, x_{(d_1-1)\cdot j+d_1}]^T$, $j = 1, \dots, d_2$ are column vectors, or X_{mat} if the dimensions are clear from the context. Obviously, it holds that $X = (X_{\text{vec}})_{\text{mat}}$.
A useful connection is the equality of the Frobenius norm of a matrix X and the ℓ_2 norm of its vectorization X_{vec}, i.e., $\|X\|_F = \|X_{\text{vec}}\|_{\ell_2}$.

To give a unified presentation of the properties mentioned above for the matrix case, whose advantages will become clear in the upcoming Chapter 4 and Chapter 5, we use in the following the vectorized version of matrices and corresponding adaption of the dimensions of the measurement operator Φ. This means, we consider linear maps $\Phi : \mathbb{R}^{m \times d_1 d_2}$ which allow more general types of linear measurements than the ones presented in the linear equation systems (2.39) above.

All properties presented below related to row sparsity and the $\ell_{p,2}$-norm can be formulated analogously for the concept of column sparsity and the corresponding $\ell_{2,p}$-norm.

We begin with the presentation of the appropriate formulation of the *null space properties* for these structures

Definition 2.28. Let $0 < p \leq 1$. We say that a matrix $\Phi \in \mathbb{R}^{m \times d_1 d_2}$ fulfills the

(i) *$\ell_{2,p}$-null space property ($\ell_{2,p}$-NSP)* of order K with constant $0 < \gamma_K < 1$ if for all elements $\eta \in \mathcal{N}(\Phi) := \{\eta \in M_{d_1 \times d_2} \mid \Phi(\eta_{\text{vec}}) = 0\}$ in the null space $\mathcal{N}(\Phi)$ of Φ it holds that

$$\|\eta_\Lambda\|_{\ell_{2,p}}^p \leq \gamma_K \|\eta_{\Lambda^c}\|_{\ell_{2,p}}^p, \tag{2.62}$$

for all row support sets Λ of cardinality $|\Lambda| \leq K$.

(ii) *Schatten-p-null space property (Schatten-p-NSP)* of order r with constant $0 < \gamma_r < 1$ if for all elements $\eta \in \mathcal{N}(\Phi) := \{\eta \in M_{d_1 \times d_2} \mid \Phi(\eta_{\text{vec}}) = 0\}$ in the null space $\mathcal{N}(\Phi)$ of Φ it holds that

$$\sum_{i=1}^{r} \sigma_i^p(\eta) \leq \gamma_r \sum_{i=r+1}^{\min(d_1,d_2)} \sigma_i^p(\eta). \tag{2.63}$$

Moreover, we provide following *restricted isometry properties*:

Definition 2.29 (Row sparse RIP [44] and rank-RIP [125])**.** We say that a linear map $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m$ fulfills

(i) the *row sparse restricted isometry property (row sparse RIP)* of order $K \in \mathbb{N}$ with constant $\delta_K \in (0;1)$ if, for every K-joint sparse matrix $X \in M_{d_1 \times d_2}$, it holds that

$$(1 - \delta_K)\|X\|_F^2 \leq \|\Phi(X_{\text{vec}})\|_{\ell_2}^2 \leq (1 + \delta_K)\|X\|_F^2. \tag{2.64}$$

(ii) the *rank restricted isometry property (rank-RIP)* of order $r \in \mathbb{N}$ with constant $\delta_r \in (0;1)$ if, for every rank-r matrix $X \in M_{d_1 \times d_2}$, it holds that

$$(1 - \delta_r)\|X\|_F^2 \leq \|\Phi(X_{\text{vec}})\|_{\ell_2}^2 \leq (1 + \delta_r)\|X\|_F^2. \tag{2.65}$$

The constants δ_K resp. δ_r are called *restricted isometry constants* of the corresponding order.

Next we present a result that the RIP in its respective variants is fullfilled by Gaussian matrices $\Phi \in \mathbb{R}^{m \times d_1 d_2}$ if the number of measurements m is chosen large enough.

Theorem 2.30 ([21, 44])**.** *Let $\Phi \in \mathbb{R}^{m \times d_1 d_2}$ be a matrix with i.i.d. centered Gaussian entries with variance $1/m$. Then*

(i) for all $0 < \delta < 1$ and $0 < \epsilon < 1$, there exists a constant $C_\delta > 0$ such that the row sparse-RIP of order K with constant $\delta_K \leq \delta$ is fulfilled with probability at least $1 - \epsilon$ provided that

$$m \geq C_\delta\big(Kd_2 + K\log(d_1/K) + \log(2\epsilon^{-1})\big),$$

(ii) for all $0 < \delta < 1$ and $0 < \epsilon < 1$, there exists a constant $C_\delta > 0$ such that the rank-RIP of order r with constant $\delta_r \leq \delta$ is fulfilled with probability at least $1 - \epsilon$ provided that

$$m \geq C_\delta\big(r(d_1 + d_2) + \log(2\epsilon^{-1})\big).$$

Proof. (i) This follows from [44, Proposition 4], (ii) Follows from [21, Theorem 2.3]. □

For the sake of completeness, we present some important conclusions from these properties for the reader. Again it is possible to show that the null space properties as defined in Definition 2.28 hold for appropriately chosen constants if the above RIPs are satisfied and this result is established in the following theorem. The proofs of these results can be found in [84].

Theorem 2.31 (Connection of RIP with NSP)**.** *The following holds true:*

(i) Let $\Phi \in \mathbb{R}^{m \times d_1 d_2}$ be a matrix that satisfies the row sparse RIP (2.64) *with constant $\delta_{2K} < 1/2$. Then there exists a number $0 < p_0(\delta_{2K}) \leq 1$ such that for any $p < p_0(\delta_{2K})$, Φ fulfills the $\ell_{2,p}$-block-NSP* (2.62) *of order K with constant*

$$\gamma_K < \left[\frac{\delta_{2K}\frac{2-p}{2-\delta_{2K}} + p(\frac{1-p/2}{2-\delta_{2K}})^{2/p}}{\frac{2-p}{2-\delta_{2K}}(1-\delta_{2K})}\right]^{p/2}.$$

(ii) Let $\Phi \in \mathbb{R}^{m \times d_1 d_2}$ be a matrix that satisfies the rank-RIP (2.65) *with constant $0 < \delta_{2r} < 1$. Then there exists a number $0 < p_0(\delta_{2r}) \leq 1$ such that for any $p < p_0(\delta_{2r})$, Φ satisfies the Schatten-p-NSP* (2.63) *of order r with constant*

$$\gamma_r \leq b(p, \sqrt{(1-\delta_{2r})/(1+\delta_{2r})})^p < 1$$

with the function $b(q, \delta)$ as in Lemma 2.26

The following reverse triangle inequalities can be derived from the NSPs above and will serve as useful tools for the analysis of algorithms in situations where the NSP holds true [35, 51, 55]. For further details we again refer to [84].

Lemma 2.32. *Let $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m$ be a linear map.*

(i) If Φ fulfills the $\ell_{2,p}$-NSP of order K with constant γ_K from Definition 2.28, then

$$\|X - Z\|_{\ell_{2,p}}^p \leq \frac{1+\gamma_K}{1-\gamma_K}\Big(\|Z\|_{\ell_{2,p}}^p - \|X\|_{\ell_{2,p}}^p + 2\beta_K(X)_{\ell_{2,p}}^p\Big) \tag{2.66}$$

for all $X, Z \in M_{d_1 \times d_2}$ such that $\Phi X_{vec} = \Phi Z_{vec}$, where

$$\beta_K(X)_{\ell_{2,p}} := \inf\big\{\|X - X'\|_{\ell_{2,p}},\ X' \in M_{d_1 \times d_2} \textit{ is } K\textit{-row sparse}\big\}$$

is the best K-row approximation of $X \in M_{d_1 \times d_2}$ in the $\ell_{2,p}$-quasinorm.

(ii) If Φ fulfills the Schatten-p-NSP of order $2r$ with constant γ_{2r} from Definition 2.28, then

$$\|X - Z\|_{S_p}^p \leq \frac{1+\gamma_{2r}}{1-\gamma_{2r}}\Big(\|Z\|_{S_p}^p - \|X\|_{S_p}^p + 2\beta_r(X)_{S_p}^p\Big) \tag{2.67}$$

for all $X, Z \in M_{d_1 \times d_2}$ such that $\Phi X_{vec} = \Phi Z_{vec}$, where

$$\beta_r(X)_{S_p} := \inf\big\{\|X - X'\|_{S_p},\ X' \in M_{d_1 \times d_2} \textit{ has rank } r\big\}$$

is the best rank-r approximation of $X \in M_{d_1 \times d_2}$ in the Schatten-p-quasinorm.

2.4 Iteratively reweighted least squares methods in data analysis

In this section, we want to introduce the *iteratively reweighted least squares* (IRLS) method, a powerful optimization algorithm, which will be the central topic under discussion in this thesis. In the following chapters, several variants of the core algorithmic design of IRLS for different relevant application scenarios will be developed and analyzed.

Iteratively reweighted least squares is an algorithmic strategy which classically imitates ℓ_p-minimization for vectors. Moreover, it can also be extended to the minimization of related matrix quasinorms as already appearing in the problems (2.12) for residual minimization or (2.46) for signal recovery. The algorithm is performing a successive approximation of the ℓ_p-minimization problem. A weighted least squares problem with an iteratively adapted weight matrix is carried out in each iteration, hopefully leading to convergence to the actual ℓ_p-minimizer.

IRLS approximation constitutes a powerful and adaptable method for a vast number of problems in engineering and applied sciences. In a wide range of applications, it is employed as a fast and robust approximation tool, in particular in statistics for robust regression, maximum or quasi-likelihood estimation, general nonlinear parameter estimation as well as in the expectation-maximization framework. Besides that, IRLS has lead to very impressive results in signal processing in sparse vector [35] and low-rank matrix [51] recovery as it can exhibit superlinear convergence rate even for nonsmooth and nonconvex optimization problems [36]. Another area, where IRLS- algorithms are successfully applied is the solution of minimization problems involving bounded variation functions [26] and, in particular, for the approximation the weak solution of the p-Poisson problem [39].

First we will put forward the fundamental observations to clarify the relation between the algorithmic concepts of weighted least squares and those for ℓ_p-(quasi-)norm-minimization inspired by [48] and [35]. A generalization to matrix quasi-norms will be provided towards the end of this section.

Thereafter, we consider the variational nature of the problem. This will lead to the formulation of an iterative algorithm which characterizes an approximate solution as the minimizer of an energy functional.

2.4.1 General concept of iteratively reweighted least squares

As a start, we note that from an optimization perspective direct ℓ_p-minimization is somewhat inconvenient as it is a nonsmooth optimization. The crucial idea now is to substitute the occurring absolute value in the following simple way:

$$|t|^p = \frac{t^2}{|t|^{p-2}} \text{ for } t \neq 0.$$

Using this, we hope that we can recast the ℓ_p-quasinorm into a *weighted ℓ_2-norm*, which is smooth and quadratic and, hence, much more practical from an optimization point of view. We obtain for $z \in \mathbb{R}^N$ with $z_i \neq 0$ for $i \in [N]$

$$\|z\|_{\ell_p}^p = \sum_{i=1}^N |z_i|^p = \sum_{i=1}^N z_i^2 \, |z_i|^{p-2} = \sum_{i=1}^N z_i^2 w_i^{p-2} = \|z\|_{\ell_2(w)}^2 = \|z\|_{\ell_2(W)}^2$$

where $W = \operatorname{diag}(w) = \operatorname{diag}(w_1, \ldots, w_N)$ with $w_i = |z_i|^{p-2}$.

Here one has to take into account that the weights will approach infinity for $z_i \to 0$ in the case that z is tending to be sparse!
To avoid this issues, we introduce a smoothing parameter $\epsilon > 0$ and hope to obtain a good approximation by using

$$w_i = \left|(z_i^*)^2 + \epsilon^2\right|^{\frac{p-2}{2}} \approx |z_i^*|^{p-2} \,. \tag{2.68}$$

This enables us to use our observation to formulate the following iterative strategy for the ℓ_p-minimization in an optimization problem involving a function $\varphi : \mathbb{R}^d \to \mathbb{R}^N$ possibly with respect to a constraint set $\mathcal{C} \subseteq \mathbb{R}^d$

$$\min_{x \in \mathcal{C}} \|\varphi(x)\|_{\ell_p}. \tag{2.69}$$

- Suppose we are given a start weight matrix $W^0 = \operatorname{diag}(w^0) \in \mathbb{R}^{N \times N}$ with $w_i^0 > 0$ for $i \in [N]$.

- We then iterate for $n \geq 0$
  - Define
  $$x^{n+1} = \arg\min_{x \in \mathcal{C}} \|\varphi(x)\|_{\ell_2(w^n)}^2$$
  - Update ϵ^{n+1} such that $\epsilon^{n+1} \leq \epsilon^n$

- Construct a new weight matrix using the updated variables such that

$$w_i^{n+1} = \left|\varphi(x^{n+1})_i^2 + (\epsilon^{n+1})^2\right|^{\frac{p-2}{2}}$$

for all $i \in [N]$ to obtain $W^{n+1} = \operatorname{diag}(w^{n+1}) \in \mathbb{R}^{N \times N}$.

Hopefully, for a decreasing sequence of appropriately chosen $\epsilon_n \to 0$, the iteration of this procedure realizes a contraction principle, which allows for the convergence of the iterates $x^n \to x_0 = \min\limits_{x \in \mathcal{C}} \|\varphi(x)\|_{\ell_p}$ for $n \to \infty$.

Remark 2.33. The concept of iterative reweighting can not only be applied to (quasi-) norms as demonstrated above, but it can also be applied to other concave nondecreasing objective functions as for example indicated in Wipf e.a. [164], Malek-Mohammadi e.a.[103]. Nevertheless, we will not further discuss this extension to other more general functions in this thesis.

Before we come closer to the concrete establishment of the iteratively reweighted least squares algorithm, we want to give a variational interpretation of our developed concept and hereby show a key tool for its analysis.

The first step towards this makes clever use of the fact that $|t|^p$ for $t \in \mathbb{R}$ can be expressed as the minimum of a function of the weight $w > 0$

$$|t|^p = \min_{w>0} \frac{p}{2}\left(wt^2 + \frac{2-p}{p} w^{\frac{p}{p-2}}\right).$$

Its unique minimizer is $w = |t|^{p-2}$.

Combining this with our considerations above, we construct the following *surrogate energy functional*

$$\mathcal{J}(x, w, \epsilon) := \frac{p}{2}\left[\|\varphi(x)\|_{\ell_2(w)}^2 + \epsilon^2 \|\mathbf{1}_{N\times 1}\|_{\ell_2(w)}^2 + \|w\|_{\ell_{\frac{p}{p-2}}}^{\frac{p}{p-2}}\right] \tag{2.70}$$

majorizing the first weighted least squares term in the w-component. Now we can formulate our iteration process as the alternating minimization of this energy functional with respect to its different variables in Algorithm 1.

The advantages of simplicity, adaptability, and straightforward implementation of `IRLS` explain its popularity for quick and efficient numerical testing also for beginners and the long history of its application in statistics and engineering contexts.

The first appearance of iteratively reweighted least squares algorithms can be reported already in the 1960s. The doctoral thesis of Lawson in 1961 [91] introduces an `IRLS`-type

Algorithm 1 Typical structure iteratively reweighted least squares algorithm (IRLS)

```
Input: Map $\varphi : \mathbb{R}^d \to \mathbb{R}^N$, constraint set $\mathcal{C} \in \mathbb{R}^d$, non-convexity parameter $0 < p \leq 1$.
Output: Sequence $(x^{(n)})_{n=1}^{n_0} \subset \mathbb{R}^d$.
Initialize $n = 0$, $\epsilon^{(0)} = 1$ and $w^0 = \mathbf{1}_{N \times 1} \in \mathbb{R}^N$.
 repeat
    $x^{n+1} = \arg\min_{x \in \mathcal{C}} \mathcal{J}(x, w^n, \epsilon^n) = \arg\min_{x \in \mathcal{C}} \|\varphi(x)\|_{\ell_2(w^n)}$,    (2.71)
    $\epsilon^{n+1} < \epsilon^n$    (2.72)
    $w^{n+1} = \arg\min_{w > 0} \mathcal{J}(x^{n+1}, w, \epsilon^{n+1})$,    (2.73)
    $n = n + 1$.
until stopping criterion is met.
Set $n_0 = n$.
```

method in the form of an algorithm for achieving solutions to approximation problems, in particular involving Chebyshev polynomials, via limits of weighted ℓ_p-norm solutions. For this algorithm, often referred to as Lawson's algorithm, a linear convergence rate was shown in [32] and extensions of Lawson's algorithm for ℓ_p-minimization were proposed by Rice and Usow in 1968 [129]. In the context of robust regression, IRLS-type algorithms first appeared in the work of Beaton and Tukey in the mid 70s [5] followed by a discussion on local as well as global convergence properties by Dutter in 1975 [42]. Slightly later, Holland and Welsch came up with a variant of the IRLS algorithm using alternative explicitly defined weights instead of the standard weights.

Moreover, IRLS algorithms were suggested in inference related topics by Wedderburn in 1974 based on the concept of quasi-likelihood. Thereby he established the connection between the IRLS algorithm for maximum likelihood estimation and the Gauss-Newton method for least-squares fitting in nonlinear regression. These results could be generalized for the multivariate case by McCullagh a decade later [105, 159]. Around the same time, Green (1984 ,[69]) investigated the relation of IRLS to Newton-Raphson and Fisher scoring as appearing in generalized linear models, linear and nonlinear regression.

A thorough discussion of IRLS-type methods can be found in the works Huber [77] and the possibly most far-reaching mathematical performance analysis for IRLS with ℓ_p-minimization for the parameter range $1 < p < 3$ is provided by Osborne [115]. Further details on the history of IRLS methods in regression can be found in [14].

No substantial innovations in the field can be documented for some time, before total variation minimization in image processing as proposed by [131] attracted the attention of the community in the early 1990s and IRLS came to the fore again.

Its straightforward applicability and implementation for total variation regularized functionals is demonstrated in [26]. Moreover, also the availability of computationally very efficient preconditioning methods [156] make IRLS methods an attractive choice in this context, that outperforms universal optimization techniques such as interior point methods.

Shortly before the millenial, the publications [68] and [137] suggested the application of IRLS for the reconstruction of sparse vectors already before Candes, Romberg, Tao and Donoho [25, 40] lay the foundation for the literal boom in compressed sensing.

A comprehensive theoretical analysis of the convergence properties of IRLS for the ℓ_p-norm minimization problem under linear measurement constraints was developed in the papers [28, 29, 35, 124], where we will give an insight on the results of the later below.

In 2010, a further extension of IRLS to the problem of low-rank matrix recovery from a minimal number of linear measurements was pursued more or less in parallel by Fornasier e.a. [51] and Fazel e.a. [106]. Moreover, building upon the results in [26] for solving total variation minimization problems, IRLS is employed also for the solution of quasi-linear elliptic equations in [56] as so-called Kačanov iteration.

In the last ten years, a growing interest on topics related to IRLS, in particular in statistics and signal processing, resulted in an ongoing rapid research development. Beyond the breakthroughs mentioned above, it is hard to provide a complete survey of the most recent state-of-the-art results in the field. For a further collection of references to the quite recent literature in this direction we refer to [114].

2.4.2 IRLS for sparse vector recovery and matrix valued signals

Let us now examplary consider in more detail the application of an IRLS-strategy to the solution of a minimization problem as in (2.46)

$$\min_{\Phi x = y} \|x\|_{\ell_p} \tag{2.74}$$

involving a *linear* measurement map $\Phi \in \mathbb{R}^{m \times d}$ as analyzed in the context of sparse signal recovery in [35]. This corresponds to the realization of the algorithm sketched above for $\varphi(x) = x$ and $\mathcal{C} = \mathcal{F}(y, \Phi) = \big\{z \in \mathbb{R}^d | \Phi(z) = y\big\}$.

In this case, as presented in (2.18), the weighted ℓ_2-minimization step can be solved directly by calculating

$$x^{(n+1)} = (W^{(n)})^{-1}\Phi^T \left(\Phi(W^{(n)})^{-1}\Phi^T\right)^{-1} y,$$

where W^n is the diagonal weight matrix $\text{diag}(w^{(n)})$ with $w_i^{(n)} = \left|(x_i^{(n)})^2 + (\epsilon^{(n)})^2\right|^{\frac{p-2}{2}}$.

The update rule for the smoothing parameter sequence $\epsilon^{(n)}$ and its convergence limit $\lim_{n\to\infty} \epsilon^{(n)} = \bar{\epsilon}$ will play an important role in the algorithm's theoretical analysis. It is carried out as follows

$$\epsilon^{(n+1)} := \min\left(\epsilon^{(n)}, \frac{r(x^{(n+1)})_{K+1}}{d}\right), \tag{2.75}$$

where $r(X)$ is the non-increasing rearrangement as introduced in (2.59) and K is an approximate guess of the sparsity-level of the solution of (2.74).

A detailed pseudo code version of the classical IRLS algorithm for sparse recovery is presented in Section 7.2 in [35].

We observe, that the iterates x^n of this IRLS-algorithm are in general no sparse vectors and even their limit $\lim_{n\to\infty} x^{(n)} = \bar{x}$ is not necessarily exactly sparse.

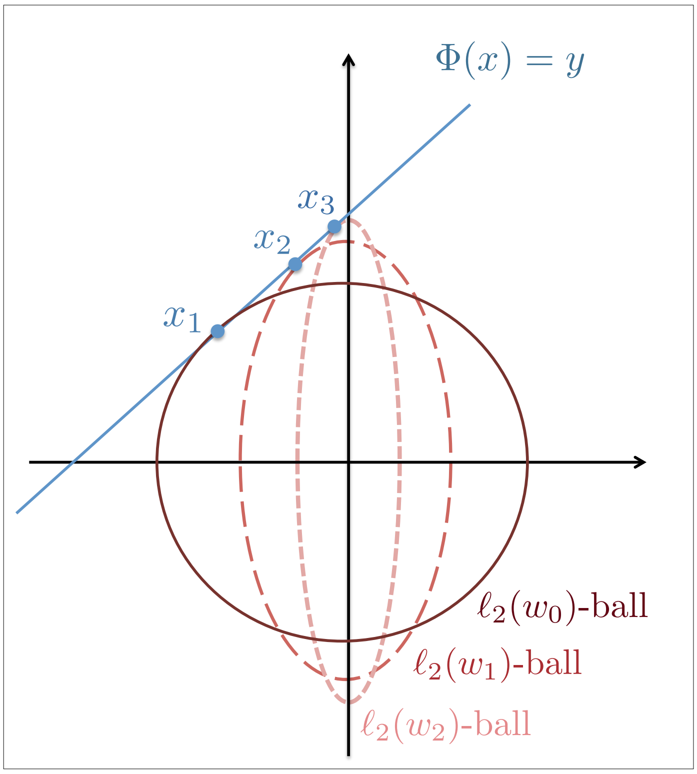


Figure 2.11: Visualization of the principle of the IRLS-algorithm

As a first step towards a discussion of theoretical analysis results for this algorithm in [35], we remember the sparse vector recovery guarantees as presented in (2.61) for ℓ_p-minimization. These hold true under the condition that Φ satisfies the corresponding p-NSP. Since we are mimicking the solution of (2.74) by the `IRLS` algorithm as explained above, these NSP-based recovery results stay applicable also for the analysis of the `IRLS` algorithm.

The central elements necessary to conduct the proof of convergence are some straightforward derivable variational properties of the corresponding auxiliary functional, the NSP, and the nonincreasing rearrangement. We will restrict our presentation to a sketch of the final result and refer to [48, p.115ff] and [35, p.12ff] for further details.

Theorem 2.34. *Let K (the same index as used in the update rule* (2.75) *) be chosen such that $\Phi \in \mathbb{R}^{m\times d}$ satisfies the p-NSP of order K, with $\gamma < 1 - \frac{2}{K+2}$. Then, for each $y \in \mathbb{R}^m$, the output of the `IRLS` algorithm in Section 7.2 in [35] converges to a vector $\bar{x}$, with $r(\bar{x})_{K+1} = d \lim_{n\to\infty} \epsilon^{(n)}$ and the following hold*

(i) If $\epsilon = \lim_{j\to\infty} \epsilon^{(n)} = 0$ then $\bar{x}$ is K-sparse; therefore, in this case, there is a unique ℓ_p-minimizer x_0 and $\bar{x} = x_0$;

(ii) If $\epsilon = \lim_{j\to\infty} \epsilon^{(n)} > 0$, then $\bar{x} = x^\epsilon = \arg\min_{x\in\mathcal{F}(y,\Phi)} \|x\|_{\ell_p,\epsilon}$, where $\|x\|_{\ell_p,\epsilon}$ is the ϵ-smoothed ℓ_p-norm as in (2.58).

Surprisingly, in the case that the approximation by the iterates is already close enough and the iterates enter a certain region around the actual minimizer, the convergence rate speeds up to superlinear.

To establish this superlinear rate of convergence, we define for the sequence of output vectors x^n the according error vector sequence $\eta^{(n)} := x^{(n)} - x_0 \in \mathcal{N}$ and

$$E^{(n)} := \left\|\eta^{(n)}\right\|_{\ell_p}^p = \left\|x^{(n)} - x_0\right\|_{\ell_p}^p .$$

We note that from Theorem 2.34 follows in the case that there exists a k-sparse vector $x_0 \in \mathcal{F}(y,\Phi)$ with $k < K - \frac{2\gamma}{1-\gamma}$ such that $x^{(n)} \to x_0$ and therefore $E^{(n)} \to 0$ and give the superlinear convergence rate for the `IRLS` algorithm in the following statement:

Theorem 2.35. *Assume Φ satisfies p-NSP of order K with constant $\gamma \in (0,1)$ and that $\mathcal{F}(y,\Phi)$ contains a k-sparse vector x_0 with $k \leq K$. Suppose that for a given $0 < \rho < 1$ and an iteration $n_0 \in \mathbb{N}$ we have*

$$E^{(n_0)} \leq (\rho r(x_0)_k)^p = R^*.$$

If ρ and γ are sufficiently small, there exists $\mu(\rho, K, \gamma, p, d) > 0$ such that for all $n \geq n_0$ we have

$$E^{(n+1)} < \mu(E^{(n)})^{2-p}.$$

For illustration of this result, one can think of superlinear convergence happening as soon as the iterates $x^{(n)}$ enter a ball centered at x_0 with radius R^*, i.e., $x^{(n)} \in \mathcal{B}_{R^*}^{\ell_p}(x_0)$. Moreover, note that the smaller the nonconvexity parameter p, the faster the convergence rate, even approaching quadratic rate for $p \to 0$.

The proof of this statement uses again the NSP as well as variational properties of the auxiliary functional as detailed in [48, p.122ff] and [35, p.16ff].

The interested reader is encouraged to get further information about `IRLS` for sparse vector recovery from [35, p.22ff].

After the discussion of theoretical aspects on `IRLS`-type methods we want to close the section with some comments on more practical issues:

Remark 2.36. (i) If we assume to be in the case where $\epsilon^{(n)} \to 0$, it follows that $|(x^{(n)})_{K+1}| \to 0$ for $n \to \infty$ from the definition of $\epsilon^{(n)}$. In this situation, the weights $(w^{(n)})_i$ will become extremely large until reaching the limits of machine representability for the indices $i > K$. On the one hand, this issue enforces the practical need for a lower bound $\hat{\epsilon} > 0$ to avoid computational instabilities. On the other hand, with the incorporation of a factor $\hat{\epsilon}$ we introduce an intrinsic limitation for the `IRLS` algorithm: it is only capable of finding an approximation of the exact solution and no longer coincides with the solution predicted by the theoretical analysis. One has to find an appropriate tradeoff: to reach sufficient recovery accuracy one has to choose $\hat{\epsilon}$ small enough, but this unfortunately can cause numerical complications, e.g., setting $\hat{\epsilon} = 1e^{-8}$ results in weight factors $(w^{(n)})_i$ of the order of $1e^{+8}$. As a consequence the execution of numerical operations can result in significant numerical errors strongly affecting the calculation results. As a conclusion, most likely the application of `IRLS` will not allow recovery errors in the regime of machine precision.

(ii) Moreover, if iterative methods are employed for the solution of the internal optimization problems in each step, we, additionally, encounter an approximation error depending on the particular termination tolerance of the chosen approach. This further deteriorates the expected accuracy of an `IRLS` method.

We now want to generalize the concept of `IRLS` also to the matrix valued case and draw the connection from the matrix (quasi-)norms corresponding to the underlying structures as considered above to weighted least squares problems.

Note at this point that in the vector case we considered mostly vector valued weights that could be used to obtain a formulation of the weighted least squares problem with a diagonal weight matrix with the weights on the diagonal. In the matrix valued case, one can in general use more flexible weighting concepts also allowing full matrices.

Definition 2.37. We now point out that each of the (quasi-)norms above can be expressed as a classical reweighted ℓ_2-norm $\|\cdot\|_{\ell_2(W)}$ if considered in a vectorized formulation.

(i) $$\|Z\|_{\ell_p}^p = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} |Z_{ij}|^p = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} |Z_{ij}|^{p-2} Z_{ij}^2 = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} \widetilde{W}_{ij} Z_{ij}^2 = \sum_{l}^{d_1\cdot d_2} \widetilde{W}_l Z_l^2$$
$$= \|W^{1/2} Z_{\text{vec}}\|_{\ell_2}^2 = \|Z_{\text{vec}}\|_{\ell_2(W)}^2$$
where $l = (i-1)\cdot d_1 + j$ and W is the diagonal weight matrix in $\mathbb{R}^{d_1\cdot d_2 \times d_1 \cdot d_2}$ with entries $W_{ll} = |Z_{ij}|^{p-2}$,

(ii) $$\|Z\|_{\ell_{2,p}}^p = \sum_{i=1}^{d_1}\left(\sum_{j=1}^{d_2} Z_{ij}^2\right)^{p/2} = \sum_{i=1}^{d_1}\left(\sum_{j=1}^{d_2} Z_{ij}^2\right)^{(p-2)/2}\left(\sum_{j=1}^{d_2} Z_{ij}^2\right) = \sum_{i=1}^{d_1} \tilde{W}_i \left(\sum_{j=1}^{d_2} Z_{ij}^2\right)$$
$$= \sum_{l=1}^{d_1\cdot d_2} \widetilde{W}_l Z_l^2 = \|W^{1/2} Z_{\text{vec}}\|_{\ell_2}^2 = \|Z_{\text{vec}}\|_{\ell_2(W)}^2,$$

where $l = (i-1)\cdot d_1 + j$ and W is the diagonal weight matrix in $\mathbb{R}^{d_1\cdot d_2\times d_1\cdot d_2}$ entries $W_{ll} = (\sum_{j=1}^{d_2} Z_{ij}^2)^{(p-2)/2}$.

(iii) $$\|Z\|_{\ell_{S_p}}^p = \text{tr}(ZZ^T)^{p/2} = \text{tr}[(ZZ^T)^{(p-2)/2}(ZZ^T)] = \text{tr}(\widetilde{W}ZZ^T) = \|\widetilde{W}^{1/2} Z\|_F^2$$
$$= \|W^{1/2} Z_{\text{vec}}\|_{\ell_2}^2 = \|Z_{\text{vec}}\|_{\ell_2(W)}^2,$$

where $\widetilde{W}$ is the symmetric weight matrix $(ZZ^T)^{p-2}$ in $\mathbb{R}^{d_1\times d_2}$ and W is the block diagonal weight matrix in $\mathbb{R}^{d_1\cdot d_2\times d_1\cdot d_2}$ with $\widetilde{W}$ repeating on the diagonal.

Similar to (2.68) we define the smoothed weight functions $W_s(Z,\epsilon)$ for the weight matrices introduced in Definition 2.37 by perturbations by a smoothing parameter $\epsilon > 0$ to avoid singularity and instability problems.

$$W_\epsilon(Z) = \begin{cases} diag\left(\left(((|Z_l|)^2+\epsilon^2\right)^{\frac{p_s-2}{2}}\right)_{l=1}^{d_1\cdot d_2}\right) & \text{for } (i) \\ diag\left(\left(\left(\sum_{j=1}^{d_2}(Z_{ij})^2+\epsilon^2\right)^{\frac{p-2}{2}}\right)_{l=1}^{d_1\cdot d_2}\right) & \text{for } (ii) \\ diag\left(\left(\left(ZZ^T+\epsilon^2\cdot I_{d_1}\right)^{\frac{p-2}{2}}\right)_{i=1}^{d_2}\right) & \text{for } (iii) \end{cases} . \qquad (2.76)$$

Remark 2.38. For all cases of the quasi-norms we obtain symmetric and positive definite weight matrices $W_\epsilon(Z)$.

The next chapter of this thesis discusses an `IRLS` algorithm for solving the nonlinear least squares problem as already introduced in (2.12), i.e., for a nonlinear measurement map $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ we attempt to solve the optimization problem

$$\min_x \|e(x)\|_{\ell_p} = \min_x \|\Phi(x) - y\|_{\ell_p}. \tag{2.77}$$

by the application of an `IRLS`-strategy as presented above, where $\varphi(x) = \Phi(x) - y$ and $\mathcal{C} = \mathbb{R}^d$.

Chapter 3

Nonlinear residual minimization via IRLS

In practice, measurement models in the applied sciences and engineering, that we represented by the operator $\Phi : \mathbb{R}^d \to \mathbb{R}^m, x \mapsto \Phi(x)$, are in most cases not linear but nonlinear, resulting in the task to find solutions to an overdetermined nonlinear equation system

$$\Phi(x) = y$$

for measurement results $y \in \mathbb{R}^m$ with $d \ll m$. Nevertheless, simplified linear models of the processes under consideration are used, or often nonlinearities are neglected. Unfortunately linearization, i.e., the assumption $\Phi(x) \approx \tilde{\Phi}x$ for $\tilde{\Phi} \in \mathbb{R}^{m \times d}$ is not reasonable in a wide range of applications with strong nonlinear behaviour of Φ, where linear model is not reflecting reality adequately enough.

When aiming at the fitting of the measurement results $y \in \mathbb{R}^m$ to the model output $\Phi(x) \in \mathbb{R}^m$ by minimization of the residual in the ℓ_p-norm, i.e., considering the problem formulation

$$\min_{x \in \mathbb{R}^d} \|\Phi(x) - y\|_{\ell_p}. \tag{3.1}$$

for $1 \leq p \leq 2$, satisfactory solutions can not be expected from an oversimplified approach.

For a nonlinear but smooth operator Φ and the parameter-range $1 \ll p \leq 2$, the overall objective function in (3.1) is smooth and, for example, also allows the application of the standard Newton method. However, such general types of optimization algorithms are not applicable for the case of nonsmooth operators Φ or the practically very relevant case of $p \approx 1$ (or even $p = 1$). In these cases, less efficient, adapted versions of the Newton methods, as for instance the semi-smooth Newton method [150] can be considered.

These methods, however, do not take advantage of the particular structure of the objective function involving a ℓ_p-norm term. It is reasonable to prefer more specialized approaches as an IRLS method, that more directly exploits the problem's peculiarities, over general optimization tools.

An instance of a practical application, where nonlinear regression problems of this type appear, was illustrated in the introductory example (i) "Detection of faulty sensors in wireless sensor networks" (see also Figure 1.1) and is discussed in more detail in [107].

Another motivation for the examination of this problem setting for the author of this thesis was that the model (3.1) intuitively occurs as an intermediate step in the application of greedy-type strategies for the solution of nonlinear equations with sparse solutions. These methods and applications were discussed in previous work of the author of this thesis together with Fornasier and Ehler in [43]. A popular instance of such a case is for the modeling of the phase retrieval problem, which we will discuss later in the context of our numerical test experiments in this chapter.

In the previous section, we demonstrated the formulation of an iteratively reweighted least squares strategy for general optimization problems as in (2.69). This chapter discusses its application for nonlinear residual minimization as already indicated above. More concretely, in the context of (2.69), we choose the constraint set $\mathcal{C} = \mathbb{R}^d$ and the objective function $F : \mathbb{R}^d \to \mathbb{R}^m, F(x) = \Phi(x) - y$ and obtain the setting of (3.1).

Moreover, from a practical point of view, the implementation of the `IRLS` algorithm for the nonlinear residual minimization problem (3.1) is easy and straightforward following the structure of Algorithm 1 as introduced above.

As a matter of fact, unfortunately, literature on algorithmic performance results such as convergence for the `IRLS` algorithm for residual minimization is very much limited to the case where Φ is a *linear* map.

Therefore, the investigation of the algorithmic behaviour and performance of `IRLS` for the ℓ_p-norm-minimization problem (3.1), involving nonlinear operators Φ, in particular its applicability conditions and limitations, is of high interest of the statistical as well as the applied sciences community.

In this chapter, we present a rigorous theoretical analysis of the convergence behaviour for `IRLS` for nonlinear residual minimizations under certain applicability conditions on the measurement setting for *nonlinear* operators Φ as in (3.1). These results have been introduced in the paper [135] by the author of this thesis.

The presentation in [135] includes the cases, where Φ is allowed to be nonlinear and mildly smooth, and $1 \leq p < 2$ and, hence, the objective functions can be not only nonconvex but nonsmooth. More precisely, the novelty in [135] is its ability to deal with severe nonsmoothness resulting from the cases, where $p \approx 1$, as appearing in, e.g., [2] that constitute difficult instances of optimization problems.

As already mentioned, [35, 51] provide analysis results for IRLS in the context of sparse vector and low-rank matrix recovery respectively, using an auxiliary functional similar to the one in (2.70). Following this variational methodology, in Section 3.1, an appropriate functional $\mathcal{J}_{NR}(x, \epsilon, w)$ (see Definition 3.1) is also formulated for the IRLS method in the nonlinear residual minimization context. The algorithm is deduced as an alternating minimization of its variables as demonstrated for Algorithm 1. In the publications [35, 51], as an additional coercivity requirement, the appropriate formulations of the Restricted Isometry Property (RIP) as introduced above in (2.24) and Definition 2.29 were assumed. Inspired by these previous approaches, in Section 3.2 a relaxed version of the RIP is employed as well, that was already introduced in the author's paper together with Fornasier and Ehler [43]. Using this coercivity assumption the convergence error decay rates results for the IRLS algorithm can be shown in case that the auxiliary function $\mathcal{J}_{NR}(x, \epsilon, w)$ is convex. Otherwise, it is shown in Section 3.3 that convexification by quadratic perturbations of the objective functional is a viable option. Convergence of the modified approach to a good approximation of stationary points of the original problem can be guaranteed in an analogous fashion as presented in Section 3.4. Within this theoretical analysis part, let us in particular point out the technically demanding calculations and results in Lemma 3.14 and Lemma 3.19 and Remark 3.20.

For illustration of the theoretical results, the chapter is closed by the presentation of several numerical experiment results in Section 3.5. First of all, the output vectors of standard built-in Matlab methods applied to the original ℓ_p-minimization problem are compared to the output vectors of our IRLS algorithm for a visually approachable toy example. The experiment demonstrates that in several instances the local minimizers found by IRLS and the ones output by standard methods are different, also when giving the same initialization point as an input for the algorithms. Moreover, the theoretical findings from previous sections are verified in the more complex context of the recovery of sparse solutions to phase retrieval problems, where IRLS is showing significantly superior performance with respect to standard MATLAB methods. The last numerical experiment explores the recovery capability of IRLS on the task of the recovery of measurement data that was corrupted by impulsive noise, aiming for a sparse residual vector.

3.1 Auxiliary functional and nonlinear residual iteratively reweighted least squares algorithm

In the setting of (3.1), for the realization of the iteratively reweighted least squares, instead of the original ℓ_p-minimization problem, we solve a sequence of weighted quadratic problems

$$x^{(n+1)} = \arg\min_{x \in \mathbb{R}^m} \|\Phi(x) - y\|^2_{\ell^2(w^{(n)})}, \tag{3.2}$$

with a smoothed weight sequence $w_i^n = |(\Phi(x^{(n)}) - y)_i^2 + (\epsilon^{(n)})^2|^{\frac{p-2}{2}}$, $i = 1, \ldots, m$, hoping for the convergence of the iterates to the ground truth vector x_0, i.e., $x^n \to x_0$ for $n \to \infty$.

It is important to note that if Φ is smooth enough, the sequence of problems (3.2) can be addressed by efficient and standard methods as detailed in Section 2.2.2 (including possible preconditioning and fine tuning etc.).

Similar to the convergence analysis in [26, 35, 51] for `IRLS`, we build a variational formulation for our algorithmic scheme denoted above. We employ an surrogate energy functional as presented in (2.70) for the deduction of the different steps of the `IRLS` algorithm. More precisely, the auxiliary functional for the problem (3.1) takes the form:

Definition 3.1. Given $\epsilon > 0$, $x \in \mathbb{R}^d$, and a weight vector $w \in \mathbb{R}^m$, with positive entries $w_i > 0, i \in [m]$ and $1 \leq p < 2$ we define

$$\mathcal{J}_{NR}(x, w, \epsilon) := \left[\sum_{i=1}^{m} w_i (\Phi_i(x) - y_i)^2 + \sum_{i=1}^{m} \left(\epsilon^2 w_i + \frac{2-p}{p} w_i^{p/(p-2)} \right) \right], \ x \in \mathbb{R}^d. \tag{3.3}$$

We now use $\mathcal{J}_{NR}$ to interpret an `IRLS` algorithm resulting from the scheme in (3.2) as an alternating minimization over its different variables.

Algorithm 2 Nonlinear residual iteratively reweighted least squares (`NR-IRLS`)

```
Input: A map $\Phi : \mathbb{R}^d \to \mathbb{R}^m$, image $y = \Phi(x_0) \in \mathbb{R}^m$ of ground truth vector $x_0$,
        parameter $1 < p \leq 2$.
Output: Sequence $(x^{(n)})_{n=1}^{n_0} \subset \mathbb{R}^d$.
Initialize $n = 0$, $\epsilon^{(0)} = 1$ and $w^0 = \mathbf{1}_{m\times 1} \in \mathbb{R}^m$.
repeat
    $x^{(n+1)} = \arg\min_{x\in\mathbb{R}^d} \mathcal{J}_{NR}(x, w^{(n)}, \epsilon^{(n)}) = \arg\min_{x\in\mathbb{R}^d} \|\Phi(x) - y\|_{\ell_2(w^{(n)})}^2$    (3.4)
    $\mathcal{N}^{(n+1)} = \min_i(|\Phi_i(x^{(n+1)}) - y_i)|)$ and $\mathcal{M}^{(n+1)} = \max_i(|\Phi_i(x^{(n+1)}) - y_i|)$,
    $\epsilon^{(n+1)} = \min\left(\max(\mathcal{N}^{(n+1)}, \tilde{\epsilon}), \epsilon^{(n)}, \mathcal{M}^{(n+1)}\right)$ with $\tilde{\epsilon} > 0$    (3.5)
    $w^{(n+1)} = \arg\min_{w\in\mathbb{R}_+^m} \mathcal{J}_{NR}(x^{(n+1)}, w, \epsilon^{(n+1)}) = \left(\left(\Phi(x^{(n+1)})_i - y_i\right)^2 + (\epsilon^{(n+1)})^2\right)^{\frac{p-2}{2}}\right)_{i=1}^m.$    (3.6)
    $n = n + 1.$
until stopping criterion is met.
Set $n_0 = n$.
```

In theory, the algorithm stops when $\epsilon^{(n)} = 0$ and one then defines $x^{(N)} := x^{(n)}$ for $N > n$. Nevertheless, in this way the output sequence $(x^n)_{n\in\mathbb{N}}$ of the algorithm will be a set with an infinite number of distinct iteration vectors, and a more practical criterion is to stop as soon as $\epsilon^{(n)}$ falls below a threshold δ fixed a-priori.

Remark 3.2. (i) In general, due to the nonlinearity of the map Φ, nonconvexity of the objective $\mathcal{J}(\cdot, w^{(0)}, \epsilon^{(0)})$ can not be excluded and, as a consequence, more than one critical point can occur. Therefore, starting from different points x_{start} with an iterative method for the solution of the nonlinear least squares problem in this first step might have an influence on the convergence behaviour and, hence, the resulting output of the algorithm.

(ii) At each step of the algorithm, we encounter a d-dimensional nonlinear weighted least squares problem, where the standard methods mentioned above are available for its numerical solution. We stress again, that these methods in general only converge to critical points, which constitutes an intrinsic limitation of the `NR-IRLS` algorithm!

3.2 Theoretical analysis and convergence results for `NR-IRLS`

As a first step towards an analysis of Algorithm 2, we will point out several properties as boundedness of the iterates $(x^{(n)})_{n\in\mathbb{N}}$ and their closeness for $n \to \infty$. They will serve as tools for the proof of convergence and the convergence rate of `NR-IRLS` assuming adapted versions of common condition in the context of the analysis of `IRLS`-type methods.

In the following, we make a suitable choice for the relevant search domain $\mathcal{C} \in \mathbb{R}^d$ for the optimimum of (3.1), containing the ground truth vector x_0 and the origin 0. Moreover, as a requirement we have that $\mathcal{C}$ contains the first iterate $x^{(1)}$.

Further requirements on the measurement map Φ will be pointed out in the following. Let Φ be continuous and bounded on $\mathcal{C}$ and, moreover, consider the following property that is an appropriate generalization of the RIP in Definition 2.24, which we entitle the *boundedness and coercivity condition (BCC)*:

Definition 3.3. Let $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ be a nonlinear, continuous map. We say that Φ fulfills the *boundedness and coercivity condition* (BCC) at $x \in \mathcal{C}$ if there exist $\alpha, \beta > 0$ such that

$$\alpha\|x - z\|_{\ell_2} \leq \|\Phi(x) - \Phi(z)\|_{\ell_p} \leq \beta\|x - z\|_{\ell_2}$$

for all $z \in \mathcal{C}$.

Remark 3.4. The lower bound in the BCC implies that the level set $\ell_{\Phi,\mathcal{C}}(\Phi(x))$ is a singleton only containing x itself. Therefore, the BCC at the ground truth solution x_0 is a necessary condition to guarantee identifiability from the nonlinear measurements $y = \Phi(x_0)$ without making further assumptions on x_0. The upper bound, however, is the requirement of Lipschitz continuity at x.

Next we comment on some straightforward observation for the functional in Definition 3.1. First we note that, after the n-th step, it holds that

$$\mathcal{J}_{NR}(x^{(n+1)}, w^{(n+1)}, \epsilon^{(n+1)}) = \sum_{i=1}^{m}[(\Phi(x^{(n+1)})_i - y_i)^2 + (\epsilon^{(n+1)})^2]^{p/2}.$$

From the minimization steps in Algorithm 2, we can use the optimality of the variable updates to deduce the following monotonicity property:

Lemma 3.5. *The inequalities*

$$\begin{aligned}\mathcal{J}_{NR}(x^{(n+1)}, w^{(n+1)}, \epsilon_{(n+1)}) \le \mathcal{J}_{NR}(x^{(n+1)}, w^{(n)}, \epsilon^{(n+1)}) \le \mathcal{J}_{NR}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)}) \\ \le \mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)})\end{aligned}$$

hold for all $n \geq 0$.

Proof: The first inequality is a result from the optimality property of $w^{(n+1)}$, while the second inequality is a simple consequence of $\epsilon^{(n+1)} \le \epsilon^{(n)}$. Furthermore, the last inequality follows from the minimization property of the update $x^{(n+1)}$.■

From Lemma 5.3 we deduce that $\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)})$ can be bounded by the value of $\mathcal{J}_{NR}$ at the first step, which is a constant, i.e., $\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) \le \mathcal{J}_{NR}(x^{(1)}, w^{(0)}, \epsilon^{(0)})$. This fact will be helpful to show the boundedness of the iterates $(x^{(n)})_{n\in\mathbb{N}}$:

Lemma 3.6. *Let $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ be a nonlinear, continuous map that fulfills the boundedness and coercivity condition (BCC) at $x_0 \in \mathcal{C}$ and $y = \Phi(x_0) \in \mathbb{R}^m$. Then the sequence of iterates $\left(x^{(n)}\right)_n$ defined by Algorithm 2 is bounded and, hence, lies in the ball $\mathcal{B}(0, R^*)$, where $R^* = \frac{1}{\alpha}\mathcal{J}_{NR}(x^{(0)}, w^{(0)}, \epsilon^{(0)})^{1/p} + \frac{1}{\alpha}\|\Phi(x_0) - y\|_{\ell_p} + \|x_0\|_{\ell_2}$.*

Proof. For all $n \in \mathbb{N}$

$$\begin{aligned}\|x^{(n)}\|_{\ell_2} &\le \|x^{(n)} - x_0\|_{\ell_2} + \|x_0\|_{\ell_2} \le \frac{1}{\alpha}\|\Phi(x^{(n)}) - \Phi(x_0)\|_{\ell_p} + \|x_0\|_{\ell_2} \\ &\le \frac{1}{\alpha}\left(\sum_{i=1}^{m}[(\Phi(x^{(n)})_i - y_i)^2 + (\epsilon^{(n)})^2]^{p/2}\right)^{1/p} + \frac{1}{\alpha}\|\Phi(x_0) - y\|_{\ell_p} + \|x_0\|_{\ell_2} \\ &\le \frac{1}{\alpha}\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)})^{1/p} + \frac{1}{\alpha}\|\Phi(x_0) - y\|_{\ell_p} + \|x_0\|_{\ell_2}.\end{aligned}$$

By the monotonicity property in Lemma 5.3, we obtain

$$\|x^{(n)}\|_{\ell_2} \le \frac{1}{\alpha}\mathcal{J}_{NR}(x^{(1)}, w^{(0)}, \epsilon^{(0)})^{1/p} + \frac{1}{\alpha}\|\Phi(x_0) - y\|_{\ell_p} + \|x_0\|_{\ell_2} = R^*,$$

where all terms composing on the right hand side are bounded. □

Remark 3.7. The ball $\mathcal{B}(0, R^*)$, that bounds the iterates, might be large, in particular in the case that the BCC constant α is small.

The ball's radius $R^* > 0$ depends on the unknown ground truth vector x_0, but it is possible to derive a sovereign estimate from above only depending on parameters that are given or fixed a-priori. First, we note that from the optimality of the solution x_0 it follows that

$$\|\Phi(x_0) - y\|_{\ell_p} \leq \|\Phi(0) - y\|_{\ell_p}$$

and we calculate

$$\begin{aligned}\|x_0\|_{\ell_2} = \|x_0 - 0\|_{\ell_2} &\leq \frac{1}{\alpha}\|\Phi(x_0) - \Phi(0)\|_{\ell_p} \\ &\leq \frac{1}{\alpha}\left(\|\Phi(x_0) - y\|_{\ell_p} + \|\Phi(0) - y\|_{\ell_p}\right) \leq \frac{2}{\alpha}\|\Phi(0) - y\|_{\ell_p}\end{aligned}$$

Therefore, we can give an upper bound $\hat{R}$ to R^*

$$R^* \leq \hat{R} := \frac{1}{\alpha}\left(\mathcal{J}_{NR}(x^{(1)}, w^{(0)}, \epsilon^{(0)})^{1/p} + 3\|\Phi(0) - y\|_{\ell_p}\right).$$

As mentioned earlier in Remark 3.2, nonconvexity of the functional $\mathcal{J}_{NR}(\cdot, w, \epsilon)$ might occur, which can cause serious difficulties in the optimization task that has to be carried out in the first step of Algorithm 2. Additionally, nonconvexity poses are more difficult theoretical problem for the analysis of the convergence behaviour of `NR-IRLS`.

For the convergence analysis of `NR-IRLS`, we will start with the easier case assuming local convexity of the functional $\mathcal{J}_{NR}(\cdot, w, \epsilon)$ and later extend and generalize these results by appropriate modification of the algorithmic scheme for the case, where local convexity can not be assumed.

For the rest of the section, we will place the assumption of strong convexity on the functional $\mathcal{J}_{NR}(\cdot, w^{(n)}, \epsilon^{(n)})$ locally at $x^{(n+1)}$ as in Definition 2.7 for all $n \geq 0$ and formulate this uniform property in the subsequent definition:

Definition 3.8. Let $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ be a nonlinear, continuous map and the functional $\mathcal{J}_{NR}(\cdot, w^{(n)}, \epsilon^{(n)})$ be defined for the variables $w^{(n)}, \epsilon^{(n)}$ as generated by Algorithm 2 for all $n \geq 0$ with minimizer $x^{(n+1)}$. We say that the *first uniform strong convexity condition (USCC-1)* is fulfilled if there exists a uniform constant $C > 0$ such that, for all $n \geq 0$, the following condition holds

$$\begin{aligned}&\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) - \mathcal{J}_{NR}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)}) \\ &= \|\Phi(x^{(n)}) - y\|^2_{\ell_2(w^{(n)})} - \|\Phi(x^{(n+1)}) - y\|^2_{\ell_2(w^{(n)})} \\ &\geq C\|x^{(n)} - x^{(n+1)}\|^2_{\ell_2}\end{aligned} \tag{3.7}$$

Remark 3.9. It is clear that the USCC-1 holds if the functional $\mathcal{J}_{NR}(\cdot, w^{(n)}, \epsilon^{(n)})$ with fixed variables $w^{(n)}, \epsilon^{(n)}$, for each step n, is strongly convex at $x^{(n+1)}$ with a constant $C > 0$ independent of n within the ball $\mathcal{B}(0, R^*)$.

Additionally, the strong convexity of the map $x \to \|\Phi(x) - y\|_{\ell_p}^2$ at the desired solution x_0 would be another advantageous property for the map Φ. This is expressed via the following property definition:

Definition 3.10. Let $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ be a nonlinear, continuous map and x_0 a minimizer of $\|\Phi(x) - y\|_{\ell_p}^2$. We say that the *second uniform strong convexity condition (USCC-2)* is fulfilled if there exists a uniform constant $\hat{C} > 0$ such that for all $n \geq 0$ the following condition holds

$$\|\Phi(x^{(n)}) - y\|_{\ell_p}^2 - \|\Phi(x_0) - y\|_{\ell_p}^2 \geq \hat{C}\|x^{(n)} - x_0\|_{\ell_2}^2. \tag{3.8}$$

Remark 3.11. (i) We observe that, in the case that $\|\Phi(\cdot) - y\|_{\ell_p}^2$ is totally convex at x_0 according to the definition of total convexity in [15], also (3.8) holds true. Another conclusion from the results in [15] is that the function $\|\Phi(\cdot) - y\|_{\ell_p}^2$ is also strictly convex in the set $\mathcal{B}(0, 2R^*)$.

(ii) In the case that $\Phi(x_0) = y$, note the equivalence of the condition (3.8) with the lower bound of the BCC with $\hat{C} = \alpha^2$.

3.2.1 Preliminary results

Next, we will provide further useful tools for the convergence proof for Algorithm 2 in the form of several Lemmata.The first important observation we note is that we can conclude from the convergence of the sequence $\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)})$ that the iterates $x^{(1)}, \cdots, x^{(n)}, x^{(n+1)}, \cdots \in \mathbb{R}^d$ generated by the `NR-IRLS` algorithm come arbitrarily close to each other for $n \to \infty$ under the assumption of the USCC-1.

Lemma 3.12. *Let $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ be a nonlinear, continuous map and given $y \in \mathbb{R}^m$. In the case that the USCC-1 property as in Definition 3.8 holds true with constant C, it follows for the iterates of Algorithms 2 thaty*

$$\lim_{n\to\infty} \|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2 = 0.$$

Proof. For each $n = 1, 2, ...$ it holds

$$\begin{aligned}
&\left[\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) - \mathcal{J}_{NR}(x^{(n+1)}, w^{(n+1)}, \epsilon^{(n+1)})\right] \\
&\geq \left[\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) - \mathcal{J}_{NR}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)})\right] \\
&\geq C \left\|x^{(n+1)} - x^{(n)}\right\|_{\ell_2}^2
\end{aligned}$$

We conclude from monotonicity as in Lemma 5.3 and boundedness of the sequence of functional values $\left(\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)})\right)_{n\in\mathbb{N}}$ that

$$\lim_{n\to\infty} (\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) - \mathcal{J}_{NR}(x^{(n+1)}, w^{(n+1)}, \epsilon^{(n+1)})) = 0.$$

Therefore, it follows

$$\lim_{n\to\infty} \|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2 = 0.$$

□

As the sequence $(\epsilon^{(n)})_{n\in\mathbb{N}}$ is monotone, it holds that the limit $\epsilon := \lim_{n\to\infty} \epsilon^{(n)}$ exists and also is non-negative. Next we introduce a functional that will be important for the formulation of the proof of convergence, in particular in the case $\epsilon > 0$.

Definition 3.13. (ϵ-perturbed ℓ_p-norm residual) Let $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ be a nonlinear map and $y \in \mathbb{R}^m$. We define the ϵ-perturbed ℓ_p-norm residual as the functional

$$f_\epsilon(x) := \sum_{i=1}^{m} ((\Phi(x)_i - y_i)^2 + \epsilon^2)^{p/2}.$$

If we assume for a moment that $x^{(n)}$ converges to a vector $\bar{x}$ and using

$$\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon_{(n)}) = \sum_{i=1}^{m} ((\Phi(x^{(n)}) - y)_i^2 + (\epsilon^{(n)})^2)^{p/2}, \tag{3.9}$$

we notice that the limit of the sequence $\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)})$ for $n \to \infty$ coincides with the ϵ-perturbed ℓ_p-norm residual in $\bar{x}$, $f_\epsilon(\bar{x})$. The corresponding minimizer depending on ϵ is denoted as

$$x^\epsilon \in \arg\min_x f_\epsilon(x). \tag{3.10}$$

The following lemma gives a characterization of these minimizers, that will be helpful for the convergence proof later on.

Lemma 3.14. *Let $\epsilon > 0$ and define the ϵ-smoothed weight vector $w(z, \epsilon) = ((\Phi(z) - y)_i^2 + \epsilon^2)^{(p-2)/2})_{i=1}^m$. If*

$$\|\Phi(z) - y\|_{\ell_2(w(z,\epsilon))}^2 \leq \|\Phi(\tilde{z}) - y\|_{\ell_2(w(z,\epsilon))}^2 \text{ for all } \tilde{z},$$

we have $z = x^\epsilon \in \arg\min_x f_\epsilon(x)$.

Proof. Our goal is to show that, if

$$\|\Phi(z)-y\|^2_{\ell_2(w(z,\epsilon))} \leq \|\Phi(\tilde{z})-y\|^2_{\ell_2(w(z,\epsilon))} \text{ for all } \tilde{z},$$

it holds $f_\epsilon(z) \leq f_\epsilon(\tilde{z})$ for all $\tilde{z}$.
As a start, we consider the inequality

$$\begin{aligned}\|\Phi(z)-y\|^2_{\ell_2(w(z,\epsilon))} &= \sum_i \frac{(\Phi(z)_i-y_i)^2}{[(\Phi(z)_i-y_i)^2+\epsilon^2]^{(2-p)/2}} \leq \sum_i \frac{(\Phi(\tilde{z})_i-y_i)^2}{[(\Phi(z)_i-y_i)^2+\epsilon^2]^{(2-p)/2}} \\ &= \|\Phi(\tilde{z})-y\|^2_{\ell_2(w(z,\epsilon))}\end{aligned}$$

and add ϵ^2 to each summand's numerator. Next we take the square root of both sides of the inequality, which is a monotone operation on the expressions that gives

$$\left(\sum_i \frac{[(\Phi(z)_i-y_i)^2+\epsilon^2]}{[(\Phi(z)_i-y_i)^2+\epsilon^2]^{(2-p)/2}}\right)^{1/2} \leq \left(\sum_i \frac{[(\Phi(\tilde{z})_i-y_i)^2+\epsilon^2]}{[(\Phi(z)_i-y_i)^2+\epsilon^2]^{(2-p)/2}}\right)^{1/2}.$$

We observe that the left side relates to $f_\epsilon(z)$ and employ the $\frac{1}{2}$- triangle inequality for the square root to obtain

$$(f_\epsilon(z))^{1/2} \leq \left(\sum_i \frac{[(\Phi(\tilde{z})_i-y_i)^2+\epsilon^2]}{[(\Phi(z)_i-y_i)^2+\epsilon^2]^{(2-p)/2}}\right)^{1/2} \leq \sum_i \frac{[(\Phi(\tilde{z})_i-y_i)^2+\epsilon^2]^{1/2}}{[(\Phi(z)_i-y_i)^2+\epsilon^2]^{(2-p)/4}}.$$

Using Hölder's inequality gives

$$\begin{aligned}(f_\epsilon(z))^{1/2} &\leq \left(\sum_i((\Phi(\tilde{z})_i-y_i)^2+\epsilon^2)^{p/2}\right)^{1/p} \cdot \left(\sum_i((\Phi(z)_i-y_i)^2+\epsilon^2)^{\frac{p-2}{4}\cdot\frac{p}{p-1}}\right)^{\frac{2(p-1)}{2p}} \\ &= (f_\epsilon(\tilde{z}))^{\frac{1}{p}} \cdot \left[\left(\sum_i((\Phi(z)_i-y_i)^2+\epsilon^2)^{\frac{p-2}{4}\cdot\frac{p}{p-1}}\right)^{2(p-1)/(p-2)}\right]^{\frac{p-2}{2p}}.\end{aligned}$$

Having in mind $\frac{1}{(a+b)^\tau} \leq \frac{1}{a^\tau} + \frac{1}{b^\tau}$ for $a, b, \tau > 0$, and noting that $\frac{2(p-1)}{p-2}$ is negative, we can use this estimate on each summand to see

$$(f_\epsilon(z))^{1/2} \leq (f_\epsilon(\tilde{z}))^{\frac{1}{p}} \cdot \left[\left(\sum_i((\Phi(z)_i-y_i)^2+\epsilon^2)^{\frac{p}{2}}\right)\right]^{\frac{p-2}{2p}} = (f_\epsilon(\tilde{z}))^{\frac{1}{p}} \cdot (f_\epsilon(z))^{\frac{p-2}{2p}}.$$

Rearranging the terms

$$(f_\epsilon(z))^{\frac{1}{2}-\frac{p-2}{2p}} = (f_\epsilon(z))^{\frac{1}{p}} \leq (f_\epsilon(\tilde{z}))^{\frac{1}{p}}$$

and using the monotonicity of the p-th square root gives the desired result

$$f_\epsilon(z) \leq f_\epsilon(\tilde{z}).$$

□

3.2.2 Convergence and error decay rates

At this point, we have established the foundations to formulate our convergence results for Algorithm 2, namely Definition 3.3 and Definition 3.8. In the course of the subsection, under the assumption of the strong convexity of $\|\Phi(\cdot) - y\|_{\ell_p}^2$ at x_0 with constant $\hat{C}$ and certain conditions on this constant, we will be able to establish a linear error decay rate if $y \in \mathrm{Ran}(\Phi)$, or otherwise, only adding an error term scaling in the ineliminable factor $\|\Phi(x_0) - y\|_{\ell_p}$.

Theorem 3.15. *Let $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ be a nonlinear, continuous map and given $y \in \mathbb{R}^m$. Consider the functionals $\mathcal{J}_{NR}(x, w^{(n)}, \epsilon^{(n)})$ for $w^{(n)}, \epsilon^{(n)}$ as generated by Algorithm 2 for all $n \geq 0$ and the following conditions shall hold*

(a) the boundedness and coercivity condition (BCC), i.e., there exist $\alpha, \beta > 0$ such that, for all $z \in \mathcal{B}(0, R^*)$:

$$\alpha\|x_0 - z\|_{\ell_2} \leq \|\Phi(x_0) - \Phi(z)\|_{\ell_p} \leq \beta\|x_0 - z\|_{\ell_2};$$

(b) and the first uniform strong convexity condition (USCC-1), i.e., there exists a uniform constant $C > 0$ such that for all $n \geq 0$ the following conditions holds

$$\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) - \mathcal{J}_{NR}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)}) \tag{3.11}$$

$$= \|\Phi(x^{(n)}) - y\|_{\ell_2(w^{(n)})}^2 - \|\Phi(x^{(n+1)}) - y\|_{\ell_2(w^{(n)})}^2 \geq C\|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2. \tag{3.12}$$

Then the sequence $(x^{(n)})_{n\in\mathbb{N}}$ generated by Algorithm 2 converges to a vector $\bar{x}$.

(i) if $\epsilon = \lim_{n\to\infty} \epsilon^{(n)} = 0$, and condition (a) holds, then $\bar{x} = x_0$ is the solution to the ℓ_p-minimization problem (3.1)*. Moreover, $y \in \mathrm{Ran}(\Phi)$ and $y = \Phi(x_0)$.*

(ii) if $\epsilon = \lim_{n\to\infty} \epsilon^{(n)} > 0$, and both conditions (a) and (b) hold, then $\bar{x} = x^\epsilon$ as defined in (5.39) *and $x^\epsilon \in \mathcal{B}(0, R^*)$. Here we assume that x^ϵ is indeed the unique global minimizer of f_ϵ.*

(c) *Denote the error at the n-th step as $E^{(n)}$ and the unavoidable error as $E_0 = \|\Phi(x_0) - y\|_{\ell_p}^2$. If condition (a) is fulfilled as well as the the second uniform convexity condition(USCC-2), i.e, there exists a uniform constant $\hat{C} > 0$ such that for all $n \geq 0$, the following conditions hold*

$$\|\Phi(x^{(n)}) - y\|_{\ell_p}^2 - \|\Phi(x_0) - y\|_{\ell_p}^2 \geq \hat{C}\|x^{(n)} - x_0\|_{\ell_2}^2$$

for all $n \geq 0$, where $\hat{C} > 0$ is such that $\mu := \frac{2^{1+2/p}(m^2+1)\beta^2}{\hat{C}} < 1$ and $\nu = \frac{2^{1+2/p}(m^2+1-2^{-2/p})}{\hat{C}}$, we can furthermore infer the property:

(iii) *the error decay rate can be characterized in terms of the errors $E^{(n)}$ and E_0 as follows:*

$$E^{(n+1)} \leq \mu E^{(n)} + \nu E_0 \tag{3.13}$$

or

$$E^{(n+1)} \leq \mu^n E_0 + \sum_{r=1}^{n} \mu^r \nu E_0. \tag{3.14}$$

Taking the limits for $n \to \infty$ gives an asymptotic error of the order of E_0

$$\bar{E} := \|\Phi(\bar{x}) - y\|_{\ell_p}^2 \leq \frac{\nu}{1-\mu} E_0. \tag{3.15}$$

Proof. (i) Our goal is to show the convergence of the sequence $x^{(n)}$ and that its limit coincides with the minimizer of problem (3.1). Let us first consider the case that it occurs $\epsilon^{(n_0)} = 0$ for some n_0 and, therefore, the stopping criterion is fulfilled and the algorithm sets $n = n_0$ and $x^{(n)} = x^{(n_0)}, n \geq n_0$. This implies that the output is $\bar{x} = x^{(n_0)}$. Then we can conclude from the definition of ϵ_n that also $\max_i((\Phi(x^{(n+1)})_i - y_i))^2 = 0$ and, hence, $\|\Phi(\bar{x}) - y\|_{\ell_p}^p = 0$. Having in mind (a), it follows that $\bar{x} = x_0$.

Next we consider the case, where $\epsilon^{(n)} > 0$ for all n. As we assumed $\epsilon^{(n)} \to 0$, there is an increasing sequence of indices n_l for which holds $\epsilon^{(n_l)} < \epsilon^{(n_l - 1)}$ for all l.
We observed in Lemma 5.4 that the sequence $x^{(n)}$ is bounded and hence there exists a convergent subsequence $(t_s)_{s\in\mathbb{N}}$ of $(n_l)_{l\in\mathbb{N}}$ yielding $(x^{(t_s)})_{s\in\mathbb{N}}$ whose limit point we denote by $\tilde{x}$. Using the definition of ϵ_{t_s}, we can conclude

$$\sum_i((\Phi(x^{(t_s)})_i - y_i)_i^2 + (\epsilon^{(t_s)})^2)^{p/2} < \sum_i 2^{p/2} \max_j |\Phi(x^{(t_s)})_j - y_j|^p.$$

In the case that $\epsilon^{(t_s)}$ falls below the small constant $\tilde{\epsilon}$, we infer from the definition of $(\epsilon_n)_{n\in\mathbb{N}}$ that $\epsilon_{t_s} = \max_j |\Phi(x^{(t_s)})_j - y_j| < \tilde{\epsilon}$. As a consequence, from $\epsilon^{(t_s)} \to 0$ follows that also $\max_j |\Phi(x^{(t_s)})_j - y_j| \to 0$. We make use of the fact that Φ is continuous, it follows that

$$0 \leq \sum_i |\Phi(\tilde{x})_i - y_i|^p = \lim_{s\to\infty} \sum_i ((\Phi(x^{(t_s)})_i - y_i)^2 + (\epsilon^{(t_s)})^2)^{p/2} \leq \lim_{s\to\infty} 2^{p/2} m (\epsilon^{(t_s)})^p = 0.$$

It remains to verify that $x^{(n)} \to x_0$. From $x^{(t_s)} \to x_0$ and $\epsilon^{(t_s)} \to 0$, we conclude $\mathcal{J}_{NR}(x^{(t_s)}, w^{(t_s)}, \epsilon^{(t_s)}) \to 0 = \sum_i |\Phi(x_0)_i - y_i|^p$ and by the monotonicity property of $\mathcal{J}_{NR}$, moreover, we have $\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) \to 0 = \sum_i |\Phi(x_0)_i - y_i|^p$.

We continue by using (3.9) to infer

$$\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) - m(\epsilon^{(n)})^p \leq \sum_i |\Phi(x^{(n)})_i - y_i|^p \leq \mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)}),$$

and combining this with the results above, we get

$$\lim_{n\to\infty} \sum_i |\Phi(x^{(n)})_i - y_i|^p = \sum_i |\Phi(x_0)_i - y_i|^p = 0.$$

Exploiting the BCC, we can deduce the statement that $x^{(n)} \to x_0$ in this case:

$$\begin{aligned} 0 &\leq \limsup_{n\to\infty} \left\| x^{(n)} - x_0 \right\|_2 \\ &< \limsup_{n\to\infty} \left(\frac{1}{\alpha} \left(\sum_i |\Phi(x^{(n)})_i - y_i|^p \right)^{1/p} + \frac{1}{\alpha} \left(\sum_i |\Phi(x_0)_i - y_i|^p \right)^{1/p} \right) \\ &= \frac{2}{\alpha} \lim_{n\to\infty} \left(\sum_i |\Phi(x^{(n)})_i - y_i|^p \right)^{1/p} = 0. \end{aligned}$$

(ii) As a first step, we aim to show that $x^{(n)} \to x^\epsilon$, $n \to \infty$ where $x^\epsilon \in \arg\min_x f_\epsilon(x)$. We already established the result that the sequence $(x^{(n)})_{n\in\mathbb{N}_0}$ is bounded and lies within the ball $\mathcal{B}(0, R^*)$ and therefore, accumulation points of this sequence exist. We denote any convergent subsequence of $(x^{(n)})_{n\in\mathbb{N}_0}$ with with $(x^{(n_l)})_{l\in\mathbb{N}_0}$ and its limit with $\bar{x}$. Our goal is now to prove $\bar{x} = x^\epsilon$.

Using that Φ is continuous, we get $\lim_{l\to\infty} w_i^{(n_l)} = [(\Phi(\bar{x})_i - y_i)^2 + \epsilon^2]^{(p-2)/2} = w(\bar{x}, \epsilon)_i := \bar{w}_i, i \in [m]$. Moreover, employing the result Lemma 5.5, it follows $x^{(n_l+1)} \to \bar{x}, i \to \infty$. From the definition of $x^{(n_l+1)}$ via the minimization step in the algorithm, we have that it holds

$$\left\| \Phi(x^{(n_l+1)}) - y \right\|_{\ell_2(w^{(n_l)})} \leq \left\| \Phi(z) - y \right\|_{\ell_2(w^{(n_l)})}, \text{for all } z \in \mathbb{R}^d. \qquad (3.16)$$

For a fixed value of z, we can infer for $n_l \to \infty$

$$\|\Phi(\bar{x}) - y\|_{\ell_2(\bar{w})} \leq \|\Phi(z) - y\|_{\ell_2(\bar{w})} .$$

From Lemma 3.14, we conclude that $\bar{x} = x^\epsilon$ as we made the assumption that x^ϵ constitutes the unique minimizer of f_ϵ. Hence, it has to be the unique accumulation point of $(x^{(n)})_{n\in\mathbb{N}}$ and also its limit which establishes the result.

(iii) Our goal is to show an error bound for the $(n+1)$-th iteration, starting from the error bound for the step before as follows, using the BCC,

$$\|x^{(n)} - x_0\|_{\ell_2}^2 \geq \frac{1}{\beta^2}\|\Phi(x^{(n)}) - \Phi(x_0)\|_{\ell_p}^2 \geq \frac{1}{\beta^2}\left(\frac{1}{2}\|\Phi(x^{(n)}) - y\|_{\ell_p}^2 - \|\Phi(x_0) - y\|_{\ell_p}^2\right). \tag{3.17}$$

We get back to our functional $\mathcal{J}_{NR}$ for exploiting its monotonicity along the iterations as derived in Lemma 5.3. First we define the term $\|\epsilon^{(n)}\|_{\ell_2(w^{(n)})} := \|\epsilon^{(n)}\cdot(1,\ldots,1)^T\|_{\ell_2(w^{(n)})}$ and observe that

$$\begin{aligned}\|\Phi(x^{(n)}) - y\|_{\ell_p}^2 = \left(\sum_{i=1}^m |\Phi(x^{(n)})_i - y_i|^p\right)^{\frac{2}{p}} &\geq \left(\sum_{i=1}^m \frac{(\Phi(x^{(n)})_i - y_i)^2 + (\epsilon^{(n)})^2 - (\epsilon^{(n)})^2}{((\Phi(x^{(n)})_i - y_i)^2 + (\epsilon^{(n)})^2)^{(2-p)/2}}\right)^{\frac{2}{p}} \\ &\geq 2^{1-2/p}\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)})^{\frac{2}{p}} - \|\epsilon^{(n)}\|_{\ell_2(w^{(n)})}^{\frac{4}{p}} \\ &\geq 2^{1-2/p}\mathcal{J}_{NR}(x^{(n+1)}, w^{(n+1)}, \epsilon^{(n+1)})^{\frac{2}{p}} - \|\epsilon^{(n)}\|_{\ell_2(w^{(n)})}^{\frac{4}{p}} \\ &\geq 2^{1-2/p}\|\Phi(x^{(n+1)}) - y\|_{\ell_p}^2 - \|\epsilon^{(n)}\|_{\ell_2(w^{(n)})}^{\frac{4}{p}}.\end{aligned}$$

From this result in combination with (3.17), we see that

$$\|x^{(n)} - x_0\|_{\ell_2}^2 \geq \frac{1}{2\beta^2}\left[2^{1-2/p}\|\Phi(x^{(n+1)}) - y\|_{\ell_p}^2 - \|\epsilon^{(n)}\|_{\ell_2(w^{(n)})}^{\frac{4}{p}} - 2\|\Phi(x_0) - y\|_{\ell_p}^2\right].$$

Adding and subtracting $\frac{1-2^{-2/p}}{\beta^2}\|\Phi(x_0) - y\|_{\ell_p}^2$ on both sides of the inequality and rearranging the terms gives

$$\begin{aligned}&\|x^{(n)} - x_0\|_{\ell_2}^2 + \frac{1-2^{-2/p}}{\beta^2}\|\Phi(x_0) - y\|_{\ell_p}^2 + \frac{1}{2\beta^2}\|\epsilon^{(n)}\|_{\ell_2(w^{(n)})}^{\frac{4}{p}} \\ &\geq \frac{1}{2^{2/p}\beta^2}\left(\|\Phi(x^{(n+1)}) - y\|_{\ell_p}^2 - \|\Phi(x_0) - y\|_{\ell_p}^2\right).\end{aligned}$$

Using Definition 3.10 and further rearrangement lead to

$$\|x^{(n)} - x_0\|_{\ell_2}^2 + \frac{1-2^{-2/p}}{\beta^2}\|\Phi(x_0) - y\|_{\ell_p}^2 + \frac{1}{2\beta^2}\|\epsilon^{(n)}\|_{\ell_2(w^{(n)})}^{\frac{4}{p}} \geq \frac{\hat{C}}{2^{1+2/p}\beta^2}\|x^{(n+1)} - x_0\|_{\ell_2}^2.$$

Next we estimate the expression $\|\epsilon^{(n)}\|_{\ell_2(w^{(n)})}^{\frac{4}{p}}$ from above, using the definition of $\epsilon^{(n)}$ and the simple observation that the norm of a vector is exceeding the maximum absolute value of a single vector entry by value:

$$\begin{aligned}
\|\epsilon^{(n)}\|_{\ell_2(w^{(n)})}^{\frac{4}{p}} &= \left(\sum_{i=1}^{m} \frac{(\epsilon^{(n)})^2}{[(\Phi(x^{(n)})_i - y_i)^2 + (\epsilon^{(n)})^2]^{(2-p)/2}}\right)^{\frac{2}{p}} \leq \left(\sum_{i=1}^{m} (\epsilon^{(n)})^p\right)^{\frac{2}{p}} \\
&\leq m^2\|\Phi(x^{(n)}) - y\|_{\ell_p}^2 \leq 2m^2\|\Phi(x^{(n)}) - \Phi(x_0)\|_{\ell_p}^2 + 2m^2\|\Phi(x_0) - y\|_{\ell_p}^2 \\
&\leq 2\beta^2 m^2\|x^{(n)} - x_0\|_{\ell_2}^2 + 2m^2\|\Phi(x_0) - y\|_{\ell_p}^2,
\end{aligned}$$

We summerize the results obtained above and achieve (3.13)

$$\begin{aligned}
E^{(n)} &= \|x^{(n+1)} - x_0\|_{\ell_2}^2 \\
&\leq \frac{2^{1+2/p}(m^2+1)\beta^2}{\hat{C}}\|x^{(n)} - x_0\|_{\ell_2}^2 + \frac{2^{1+2/p}(m^2+1-2^{-2/p})}{\hat{C}}\|\Phi(x_0) - y\|_{\ell_p}^2 \\
&= \mu E^{(n+1)} + \nu E_0.
\end{aligned}$$

The recurrent substitution of $E^{(n)}$ by its predecessors gives (3.14)

$$E^{(n+1)} \leq \mu^n E^{(0)} + \sum_{r=1}^{n} \mu^r \nu E_0.$$

From passing to the limit $n \to \infty$, we get (3.15). □

Remark 3.16. (i) We note that the values of μ and ν are worst upper bounds up to the point, where $\epsilon_n = \mathcal{M}^{(n)}$ in `NR-IRLS`. When $\epsilon^{(n)} = \mathcal{N}^{(n)}$, it is possible to define the constants $\tilde{\mu} = \frac{2^{2+2/p}\beta^2}{\hat{C}}, \tilde{\nu} = \frac{2^{2+2/p}-2}{\hat{C}}$ replacing μ, ν in these particular steps giving better constant values.

(ii) Due to the global minimization of $\mathcal{J}_{NR}(x, w^{(n_l)}, \epsilon^{(n_l)})$ w.r.t. x it is necessary that the inequality in Lemma 3.14 holds for all $\tilde{z}$ and not only for $z \in \mathcal{B}(0, 2R^*)$. From the corresponding minimization property, we obtain $x^{(n_l+1)}$, which constitutes the global minimizer in comparison to all other vectors z in (3.16) in step (ii).

(iii) We observe that in the case $\bar{x} = x_0$ the result (3.15) is trivial. On the other hand, for $\bar{x} = x^\epsilon$, we obtain from (3.15) further information on the vector x^ϵ as a quasi-minimizer.

3.3 Local convexification of the auxiliary functional

In this section, we want to turn to the more general case, where the uniform strong convexity condition of Definition 3.8 does not hold. This corresponds to a situation, where we cannot even assume locally convexity for the optimization problem we are confronted with. Therefore, we can not give theoretical guarantees as formulated above for the version of the `NR-IRLS` Algorithm 2. Instead, we follow the strategy of adaptive modification of Algorithm 2 by introducing local convexification around the current iterate, using the techniques presented in Section 2.2.3. This will enable us to show convergence of this adapted version of the algorithm to at least a critical point of the ϵ-perturbed ℓ_p-norm residual f_ϵ under appropriate assumptions.

The first iteration of Algorithm 2 as stated above, that is performed using $w^{(0)} = (1, \dots, 1)^T$ for the minimization of $\mathcal{J}_{NR}(x, w^{(0)}, \epsilon^{(0)})$, corresponds to a standard nonlinear ℓ_2-least squares step. At this early stage already, the local nonconvexity of the functional in the x-component can lead to the occurence of several local minimizers. Depending on the initialization vector $x^{(0)}$, that is provided as an input to the iterative solver for this nonconvex optimization problem, one of these local minimizers will be set as the next iterate $x^{(1)}$.

We need to keep this dependence on $x^{(0)}$ in mind for the local convexification centered around the current iterate that we aim at now. We have to be aware that the choice of the initialization vector influences the overall behaviour of the algorithm and the output results can strongly differ even for close starting points!

Assuming that $\Phi(x)$ is an analytic function, it follows from classical complex analysis that there only exists a finite set of isolated zeros of $\nabla\|\Phi(x) - y\|_{\ell_p}^p$ for $p > 1$ on any compact set. Our hope is that critical points of the functional $\nabla\|\Phi(x) - y\|_{\ell_p}^p$ would not change too strongly with p and that the global minimizer for $1 < p < 2$ will be in a neighborhood of a local minimizer of the least squares problem solved in the first step. Therefore, we propose to invest the computational effort to explore more than one or even as much as possible critical points of the nonconvex problem appearing in the first step using the methods described in Section 2.2.2, e.g., the Levenberg-Marquardt algorithm with several, possibly random initial points. The identified critical points will be listed as $x_{\ell_2}^{*1}, x_{\ell_2}^{*2}, \dots, x_{\ell_2}^{*L}$ and will serve as initialization points for the convexified `NR-IRLS` algorithm, that will be derived in this section. After having executed this adjusted version of `NR-IRLS` for all $x_2^{*1}, x_2^{*2}, \dots, x_2^{*L}$ and having obtained L possible solutions $x^{*_1}, x^{*_2}, \dots, x^{*_L}$ for the ℓ_p-minimization problem, we chose the x^{*_s} giving the lowest value of $\|\Phi(x^{*_s}) - y\|_{\ell_p}$ as our preferred approximation to the ℓ_p-minimizer.

We want to surmount the drawbacks of the failure of the condition in Definition 3.8 by establishing an approach for a locally convexifying adaption of Algorithm 2 and, thereby, acquire convergence guarantees also under these circumstances. We fix w, ϵ and now present the convexified version of the previously introduced functional $\mathcal{J}_{NR}$ by constructing its Moreau envelope (see Section 2.2.3) as a surrogate:

$$\mathcal{J}_{NR}^{\omega,u}(x, w, \epsilon) = \mathcal{J}_{NR}(x, w, \epsilon) + \omega\|x - u\|_{\ell_2}^2, \tag{3.18}$$

for a parameter $\omega > 0$ and $u \in \mathbb{R}^d$.

We will include this straightforward convexified formulation (3.18) into the first step of `NR-IRLS` to obtain a regularized minimization problem, resulting in the corresponding sequence of iterates $x^{(n)}$. This requires the appropriate choice of the parameters u and ω for the additional regularization term: we decide to fix $\omega > 0$ generously large and constant over all iterations and, moreover, $u = x^{(n)}$ for the n-th step, which leads to a iterative scheme as follows

$$x^{(n+1)} = \arg\min_x \mathcal{J}_{NR}^{\omega,x^{(n)}}(x, w^{(n)}, \epsilon^{(n)}). \tag{3.19}$$

An adapted version of the `NR-IRLS` algorithm can now be formulated as follows:

Algorithm 3 Convexified nonlinear residual IRLS (`NR-IRLS 2`)

Input: A map $\Phi : \mathbb{R}^d \to \mathbb{R}^m$, image $y = \Phi(x_0) \in \mathbb{R}^m$ of ground truth vector x_0, convexification parameter, ω, parameter $1 < p \leq 2$.

Output: Sequence $(x^{(n)})_{n=1}^{n_0} \subset \mathbb{R}^d$.

Initialize $n = 0$, $\epsilon^{(0)} = 1$ and $w^{(0)} = \mathbf{1}_{m\times 1} \in \mathbb{R}^m$.

repeat

$$x^{(n+1)} = \arg\min_{x\in\mathbb{R}^d} \mathcal{J}_{NR}^{\omega,x^{(n)}}(x, w^{(n)}, \epsilon^{(n)}) = \arg\min_{x\in\mathbb{R}^d} \|\Phi(x) - y\|_{\ell_2(w^{(n)})}^2 + \omega\|x - x^{(n)}\|_{\ell_2}^2 \tag{3.20}$$

$$\mathcal{N}^{(n+1)} = \min_i(|\Phi(x^{(n+1)})_i - y_i)|) \text{ and } \mathcal{M}^{(n+1)} = \max_i(|\Phi(x^{(n+1)})_i - y_i|),$$

$$\epsilon^{(n+1)} = \min\left(\max(\mathcal{N}^{(n+1)}, \tilde{\epsilon}), \epsilon^{(n)}, \mathcal{M}^{(n+1)}\right) \text{ with } \tilde{\epsilon} > 0 \tag{3.21}$$

$$w^{(n+1)} = \arg\min_{w\in\mathbb{R}_+^m} \mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n+1)}, w, \epsilon^{(n+1)}) = \left(\left((\Phi(x^{(n+1)})_i - y_i)^2 + (\epsilon^{(n+1)})^2)\right)^{\frac{p-2}{2}}\right)_{i=1}^m. \tag{3.22}$$

$n = n + 1.$

until *stopping criterion is met.*

Set $n_0 = n$.

Remark 3.17. (a) This particular choice for the modification of the objective functional and the first step of the algorithm that will introduce the desired local convexity will be justified later and presented alongside with recommendations for the concrete choice of the parameter ω in the theory section.

(b) We already mentioned a range of viable techniques for the solution of the convex minimization problem in (3.19), where a wide range of other methods exists beyond that.

3.4 Convergence analysis for the convexified algorithm

The study of the convergence behaviour of Algorithm 3 will be carried out in an analogous fashion to the analysis of Algorithm 2 in this section.

3.4.1 Preliminary results

Again we start with the monotonicity property of the modified functional:

Lemma 3.18. *The inequalities*

$$\begin{aligned}\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) &= \mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) \geq \mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)}) && (3.23)\\ &\geq \mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n+1)}, w^{(n)}, \epsilon^{(n+1)}) \geq \mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n+1)}, w^{(n+1)}, \epsilon^{(n+1)}) && (3.24)\\ &\geq \mathcal{J}_{NR}^{\omega,x^{(n+1)}}(x^{(n+1)}, w^{(n+1)}, \epsilon^{(n+1)}) = \mathcal{J}_{NR}(x^{(n+1)}, w^{(n+1)}, \epsilon^{(n+1)}) && (3.25)\end{aligned}$$

hold for all $n \geq 0$.

Proof. The first inequality is a consequence of the minimization property that defines $x^{(n+1)}$ in NR-IRLS2. Moreover, the second one results from the fact that $\epsilon^{(n+1)} \leq \epsilon^{(n)}$ and the third inequality from the minimization property of $w^{(n+1)}$. The last inequality follows from the non-negativity of the norm of a difference of vectors. □

We can conclude from this property and the boundedness of the sequence $(\mathcal{J}(x^{(n)}, w^{(n)}, \epsilon^{(n)}))_{n\in\mathbb{N}}$ that it also has to be convergent.

In an analogous manner to Lemma 5.4, we can conclude also for the convexified case that the sequence $(x^{(n)})_{n\in\mathbb{N}}$ bounded and the iterates lie in a ball of radius R^*, i.e., $(x^{(n)})_{n\in\mathbb{N}} \in \mathcal{B}(0, R^*)$.

Now we want to justify in more detail the choice of the formulation in (3.19) for the modification of the functional $\mathcal{J}$. Here the introduction of the Moreau envelope as a regularization is crucial for winning back the USCC-1 property for the modified version of the functional (3.18). Our aim is to establish the existence of a positive USCC-1 constant $\tilde{C}$, that will depend on ω and therefore can be influenced with its choice appropriately.

Lemma 3.19. *Let $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ be a nonlinear, continuous map and $\mathcal{J}(x, w, \epsilon)$ as defined in Definition 4.4 and $\mathcal{J}_{NR}^{\omega,u}(x, w, \epsilon)$ as defined in* (3.18)*. Moreover, we assume that*

$$\begin{aligned}&\Big|t[\|\Phi(tx^n + (1-t)x^{n+1}) - y\|^2_{\ell_2(w^{(n)})} - \|\Phi(x^{(n)}) - y\|^2_{\ell_2(w^{(n)})}] \qquad (3.26)\\ &+(1-t)[\|\Phi(tx^{(n)} + (1-t)x^{(n+1)}) - y\|^2_{\ell_2(w^{(n)})} - \|\Phi(x^{(n+1)}) - y\|^2_{\ell_2(w^{(n)})}]\Big|\\ &\leq Lt(t-1)\|x^{(n)} - x^{(n+1)}\|^2_{\ell_2}.\end{aligned}$$

for some $L > 0$ independent of $n \in \mathbb{N}$ and for all $t \in [0, 1]$. Let $(x^{(n)})_{n\in\mathbb{N}}$ be the output sequence of minimizers of Algorithm 3. Then for $\omega > 0$ large enough the USCC-1 is fulfilled for the adapted functional in (3.18)*, i.e., there exists a uniform constant $\tilde{C} > 0$ such that for all $n \geq 0$ holds*

$$\mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) - \mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)}) \geq \tilde{C}\|x^{(n+1)} - x^{(n)}\|^2_{\ell_2}$$

Remark 3.20. We want to explain the validity of (3.26) and assume for the moment that Φ is twice continuously differentiable and $\epsilon^{(n)} \geq \epsilon$ for all $n \in \mathbb{N}$ and that the Hessian of the map

$$x \to F_{w^{(n)}}(x) = \|\Phi(x) - y\|^2_{\ell_2(w^{(n)})},$$

which can be expressed as

$$\nabla^2 F_{w^{(n)}}(x) = \sum_{i=1}^m w_i^{(n)} \left[\nabla\Phi(x)_i \nabla\Phi(x)_i^* + (\Phi(x)_i - y_i)\nabla^2\Phi(x)_i\right],$$

is uniformly bounded on $B(0, R^*)$ by a constant $L' > 0$. We consider the Taylor expansion of the function $F_{w^{(n)}}(x) = \|\Phi(x) - y\|^2_{\ell_2(w^{(n)})}$ around the point $x = tx^{(n)} + (1-t)x^{(n+1)}$, to achieve a uniform estimate of the type (3.26):

$$\begin{aligned}&\Big|t[\|\Phi(tx^{(n)} + (1-t)x^{(n+1)}) - y\|^2_{\ell_2(w^{(n)})} - \|\Phi(x^{(n)}) - y\|^2_{\ell_2(w^{(n)})}]\\ &+(1-t)[\|A(tx^{(n)} + (1-t)x^{(n+1)}) - y\|^2_{\ell_2(w^{(n)})} - \|A(x^{(n+1)}) - y\|^2_{\ell_2(w^{(n)})}]\Big|\\ =\ &\Big|-t\nabla F_{w^{(n)}}(tx^{(n)} + (1-t)x^{(n+1)})^T(x^{(n)} - tx^{(n)} + (1-t)x^{(n+1)})\\ &-t(x^{(n)} - tx^{(n)} + (1-t)x^{(n+1)})^T\nabla^2 F_{w^{(n)}}(\xi_t^{(n)})(x^{(n)} - tx^{(n)} + (1-t)x^{(n+1)})\\ &-(1-t)\nabla F_{w^{(n)}}(tx^{(n)} + (1-t)x^{(n+1)})^T(x^{(n+1)} - tx^{(n)} + (1-t)x^{(n+1)}) +\\ &-(1-t)(x^{(n+1)} - tx^{(n)} + (1-t)x^{(n+1)})^T\nabla^2 F_{w^{(n)}}(\eta_t^{(n)})(x^{(n+1)} - tx^{(n)} + (1-t)x^{(n+1)})\Big|.\end{aligned}$$

Now, we have that

$$\begin{aligned}&- t\nabla F_{w^{(n)}}(tx^{(n)} + (1-t)x^{(n+1)})^T(x^{(n)} - tx^{(n)} + (1-t)x^{(n+1)})\\&= -t(1-t)\nabla F_{w^{(n)}}(tx^{(n)} + (1-t)x^{(n+1)})^T(x^{(n)} - x^{(n-1)})\end{aligned}$$

and

$$\begin{aligned}&- (1-t)\nabla F_{w^{(n)}}(tx^{(n)} + (1-t)x^{(n+1)})^T(x^{(n+1)} - tx^{(n)} + (1-t)x^{(n+1)})\\&= t(1-t)\nabla F_{w^{(n)}}(tx^{(n)} + (1-t)x^{(n+1)})^T(x^{(n)} - x^{(n-1)}).\end{aligned}$$

This means that the first order terms in the sum cancel each other and only the second order terms are remaining. We continue with the observation

$$\begin{aligned}&\|(x^{(n)} - tx^{(n)} + (1-t)x^{(n+1)}\|_{\ell_2}^2 = (1-t)^2\|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2, \text{ and}\\&\|(x^{(n+1)} - tx^{(n)} + (1-t)x^{(n+1)}\|_{\ell_2}^2 = t^2\|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2\end{aligned}$$

using the boundedness of the Hessians and we see that

$$\begin{aligned}&\Big|t[\|\Phi(tx^{(n)} + (1-t)x^{(n+1)}) - y\|_{\ell_2(w^{(n)})}^2 - \|\Phi(x^{(n)}) - y\|_{\ell_2(w^{(n)})}^2]\\&+(1-t)[\|\Phi(tx^{(n)} + (1-t)x^{(n+1)}) - y\|_{\ell_2(w^{(n)})}^2 - \|(x^{n+1}) - y\|_{\ell_2(w^{n})}^2]\Big|\\&\leq L't(1-t)^2\|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2 + L't^2(1-t)\|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2\\&\leq Lt(t-1)\|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2,\end{aligned}$$

using that $t \in [0,1]$ and $L = 2L'$. Consequently we assert that (3.26) is a reasonable assumption, even if the map Φ is not as smooth. A key point here is the fact that $\epsilon^{(n)} \geq \epsilon$ for all $n \in \mathbb{N}$, which is used in the proof of Theorem 3.22.

Proof. Having in mind (3.26), we carry out the estimates for $t \in [0,1]$

$$\begin{aligned}&\Big|\mathcal{J}_{NR}(tx^{(n)} + (1-t)x^{(n+1)}, w^{(n)}, \epsilon^{(n)}) - [t\mathcal{J}_{NR}^(x^{(n)}, w^{(n)}, \epsilon^{(n)}) + (1-t)\mathcal{J}_{NR}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)})]\Big|\\&\leq \Big|t[\|\Phi(tx^{(n)} + (1-t)x^{(n+1)}) - y\|_{\ell_2(w^{(n)})}^2 - \|\Phi(x^{(n)}) - y\|_{\ell_2(w^{(n)})}^2]\\&+(1-t)[\|\Phi(tx^{(n)} + (1-t)x^{(n+1)}) - y\|_{\ell_2(w^{(n)})}^2 - \|\Phi(x^{(n+1)}) - y\|_{\ell_2(w^{(n)})}^2]\Big|\\&\leq Lt(t-1)\|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2.\end{aligned}$$

Hence, we get

$$\begin{aligned}&\mathcal{J}_{NR}(tx^{(n)} + (1-t)x^{(n+1)}, w^{(n)}, \epsilon^{(n)})\\&\leq t\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) + (1-t)\mathcal{J}_{NR}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)}) - Ct(1-t)\|x^{(n)} - x^{(n-1)}\|_{\ell_2}^2,\end{aligned}$$

with a not necessarily positive, uniform constant $C = -L$, as there is not yet an assumption placed on the strong convexity for the functional $\mathcal{J}_{NR}(\cdot, w^{(n)}, \epsilon^{(n)})$ at this point but certainly $C > -\infty$.
Next we add the term $\omega\|tx^{(n)} + (1-t)x^{(n+1)} - x^{(n)}\|_{\ell_2}^2$ to both sides of the inequality

$$\begin{aligned}
&\mathcal{J}_{NR}(tx^{(n)} + (1-t)x^{(n+1)}, w^{(n)}, \epsilon_{(n)}) + \omega\|tx^{(n)} + (1-t)x^{(n+1)} - x^{(n)}\|_{\ell_2}^2 \\
&\leq t\mathcal{J}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) + (1-t)\mathcal{J}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)}) - Ct(1-t)\|x^{(n)} - x^{(n-1)}\|_{\ell_2}^2 \\
&+ \omega\|tx^{(n)} + (1-t)x^{(n+1)} - x^{(n)}\|_{\ell_2}^2
\end{aligned}$$

and rearrange

$$\begin{aligned}
&\mathcal{J}_{NR}^{\omega,x^{(n)}}(tx^{(n)} + (1-t)x^{(n+1)}, w^{(n)}, \epsilon^{(n)}) \leq t\mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) \\
&+ (1-t)\mathcal{J}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)}) - Ct(1-t)\|x^{(n)} - x^{(n-1)}\|_{\ell_2}^2 + (1-t)^2\omega\|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2.
\end{aligned}$$

Furthermore, by adding and subtracting the expression $(1-t)\omega\|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2$, we obtain

$$\begin{aligned}
&\mathcal{J}_{NR}^{\omega,x^{n}}(tx^{(n)} + (1-t)x^{(n+1)}, w^{(n)}, \epsilon^{(n)}) \\
&\leq t\mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) \\
&+ (1-t)\mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)}) - (C+\omega)t(1-t)\|x^{(n)} - x^{(n-1)}\|_{\ell_2}^2.
\end{aligned}$$

We note that, actually, the last inequality leads to the establishment of the strong convexity condition for the functional $\mathcal{J}_{NR}^{\omega,x^{(n)}}(\cdot, w^{(n)}, \epsilon^{(n)})$ at $x^{(n+1)}$ at $x^{(n)}$. Analogous calculations to those presented in proof of Lemma 5.5 executed also in this case lead to

$$\mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) - \mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)}) \geq \tilde{C}\|x^{(n+1)} - x^{(n)}\|_{\ell_2}^2,$$

with constant $\tilde{C} = C + \omega$. Here $\tilde{C}$ is positive for ω large enough. □

As a next step, we want to show that the iterates $\left(x^{(n)}\right)_{n\in N}$ come arbitrarily close for $n \to \infty$.

Lemma 3.21. *Let $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ be a nonlinear, continuous map with Φ and $(x^{(n)})_{n\in\mathbb{N}}$ and $(w^{(n)})_{n\in\mathbb{N}}$ be the sequences generated by Algorithm 3, so that condition* (3.26) *holds. Then, for $\omega > 0$ large enough*

$$\left\|x^{(n)} - x^{(n+1)}\right\|_{\ell_2}^2 \to 0 \text{ as } n \to \infty.$$

Proof. Using the monotonicity property, we see that

$$\begin{aligned}&\|\mathcal{J}_{NR}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) - \mathcal{J}_{NR}(x^{(n+1)}, w^{(n+1)}, \epsilon_{(n+1)})\|_{\ell_2}^2 \\ &\geq \|\mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) - \mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)})\|_{\ell_2}^2.\end{aligned}$$

Moreover, employing Lemma 3.19 it follows that

$$\|\mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n)}, w^{(n)}, \epsilon^{(n)}) - \mathcal{J}_{NR}^{\omega,x^{(n)}}(x^{(n+1)}, w^{(n)}, \epsilon^{(n)})\|_{\ell_2}^2 \geq \tilde{C}\|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2$$

Using the fact that $\|\mathcal{J}_{NR}(x^{(n)}w^{(n)}, \epsilon^{(n)}) - \mathcal{J}(x^{(n+1)}, w^{(n+1)}, \epsilon^{(n+1)})\|_{\ell_2}^2 \to 0$ as $n \to \infty$, we obtain

$$\|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2 \to 0 \text{ as } n \to \infty.$$

□

3.4.2 Convergence

Now we have all the necessary tools at hand to present the convergence results for Algorithm 3:

Theorem 3.22. *Fix $y \in \mathbb{R}^m$, $x_0 \in \mathbb{R}^d$. Let $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ be a nonlinear, countinuously differentiable map with Φ for which the boundedness and coercivity condition (BCC) holds, i.e., there exist $\alpha, \beta > 0$ such that, for all $z \in \mathcal{B}(0, R^*)$:*

$$\alpha\|x_0 - z\|_{\ell_2} \leq \|\Phi(x_0) - \Phi(z)\|_{\ell_p} \leq \beta\|x_0 - z\|_{\ell_2}.$$

Additionally, we require that, for the sequences $(x^n)_{n\in\mathbb{N}}$ and $(w^n)_{n\in\mathbb{N}}$ generated by Algorithm 3,

$$\begin{aligned}&\Big|t[\|\Phi(tx^{(n)} + (1-t)x^{(n+1)}) - y\|_{\ell_2(w^{(n)})}^2 - \|\Phi(x^{(n)}) - y\|_{\ell_2(w^{(n)})}^2] \\ &+(1-t)[\|\Phi(tx^{(n)} + (1-t)x^{(n+1)}) - y\|_{\ell_2(w^{(n)})}^2 - \|\Phi(x^{(n+1)}) - y\|_{\ell_2(w^{(n)})}^2]\Big| \\ &\leq Lt(t-1)\|x^{(n)} - x^{(n+1)}\|_{\ell_2}^2\end{aligned} \tag{3.27}$$

for some $L > 0$ independent of $n \in \mathbb{N}$ and for all $t \in [0, 1]$. For $\omega > 0$ large enough (determined according to Lemma 3.19), we get the following properties of Algorithm 3:

(i) If $\epsilon = \lim_{n\to\infty} \epsilon^{(n)} = 0$, then the sequence $(x^{(n)})_{n\in\mathbb{N}}$ converges to a vector $\bar{x}$, which is the solution to the ℓ_p-minimization problem (3.1). Moreover, if $y \in \mathrm{Ran}(\Phi)$ and $y = \Phi(x_0)$, then x_0 is the unique minimizer, thus $\bar{x}$ coincides with x_0.

(ii) if $\epsilon = \lim_{n\to\infty} \epsilon_n > 0$, then all accumulation points of $(x^{(n)})_{n\in\mathbb{N}}$ are critical points of the ϵ-perturbed ℓ_p-norm residual f_ϵ defined in (5.39)*, all lying in $\mathcal{B}(0, R^*)$.*

Proof. (i) the proof can be deduced in an analogous fashion to Theorem 5.8.

(ii) We know that the sequence is bounded $(x^{(n)})_{n\in\mathbb{N}}$ and, therefore, accumulation points exists. Denote with $(x^{(n_\ell)})_{\ell\in\mathbb{N}}$ any convergent subsequence of $(x^{(n)})_{n\in\mathbb{N}_0}$ and its limit with $\bar{x}$, for which we want to establish that it is a critical point of (5.39). From $w_i^{(n)} = [(\Phi(x^{(n)})_i - y_i)^2 + (\epsilon^{(n)})^2]^{(p-2)/2} \leq (\epsilon^{(n)})^{p-2} \leq \epsilon^{p-2}$ we deduce that up to the extraction of an additional subsequence, it holds $\lim_{\ell\to\infty} w_i^{(n_\ell)} = [(\Phi(\bar{x})_i - y_i)^2 + \epsilon^2]^{(p-2)/2} = w(\bar{x}, \epsilon)_i := \bar{w}_i, i = 1, \ldots, m$. Moreover, using Lemma 3.21, we see that $x^{(n_\ell+1)} \to \bar{x}, \ell \to \infty$. We observe that here $\epsilon^{(n)} \geq \epsilon > 0$ and the discussion in Remark 3.20 can be used to justify the assumption (3.27).) In a similar fashion, it follows $w^{(n_\ell+1)} \to \bar{w}$ for $\ell \to \infty$. From the assumption that Φ is continuously differentiable, it follows that the map $x \to \mathcal{J}_{NR}(\cdot, w^{(n)}, \epsilon^{(n)})$ is differentiable as well and we conclude using (3.19),

$$0 = \nabla_x \mathcal{J}_{NR}^{\omega, x^{(n_\ell)}}(x^{(n_\ell+1)}, w^{(n_\ell)}, \epsilon^{(n_\ell)}) = \nabla_x \mathcal{J}(x^{(n_\ell+1)}, w^{(n_\ell)}, \epsilon_{n_\ell}) + 2\omega(x^{n_\ell+1} - x^{n_\ell})$$

or

$$-2\omega(x^{(n_\ell+1)} - x^{(n_\ell)}) = \nabla_x \mathcal{J}(x^{(n_\ell+1)}, w^{(n_\ell)}, \epsilon_{(n_\ell)}).$$

Employing Lemma 3.21, we can conclude that by passing to the limit $\ell \to \infty$

$$0 \in \nabla_x \mathcal{J}(\bar{x}, \bar{w}, \epsilon) = \nabla f^\epsilon(\bar{x}).$$

□

Remark 3.23. (a) Instead of assuming that Φ is continuously differentiable as we did above, it is also possible to consider to lower smoothness, i.e., Φ continuous and require additional properties for subdifferentials. Nevertheless, generalizing our results to nonsmooth maps does not give us significantly new insight and is not considered here in detail.

(b) It is interesting to note that the error decay rate shown in (iii) in Theorem 5.8 can be validated also for Algorithm 3 in the case that condition (c) in Theorem 5.8 holds true.

We can summarize our results as follows: Either we reach the exact minimizer of the functional $\|\Phi(x) - y\|_{\ell_p}^p$ or otherwise, we have that every accumulation point is a critical point of the ϵ-perturbed ℓ_p-norm residual $f_\epsilon(x)$.

3.5 Numerical experiments

In this section, we illustrate and validate our theoretical findings by presenting several numerical experiments. Our first tests are carried out for a simple experiment framework to get a certain intuition for the algorithmic behaviour. Thereafter, the performance of `NR-IRLS` is evaluated in the context of higher dimensional ℓ_p-minimization problems, whose optimal solution is often difficult to investigate.

In the first example, we will examine the behaviour of the iterates of the `NR-IRLS` algorithm in each step and make comparisons of the algorithm output with the results of standard MATLAB optimization methods. Subsequent experiments consider nonlinear compressed sensing problems as examined in [43]. More concretely, we employ `NR-IRLS` in the intermediate step of of a greedy-type algorithm for the reconstruction of sparse vectors from quasilinear measurements. We claim that, if the overall recovery results obtained by the described algorithmic scheme is correct, the intermediate results must have lead to correct solutions as well. In the last experiment in this section, we will examine a measurement setting, where the measurement data is corrupted by so-called impulsive noise, which corresponds to sparsity structure appearing in the residual. We study the influence of the noise level on the recovery success of `NR-IRLS`.

All numerical experiments in this section were performed on a MacBook Pro 9.1. with a 2.6 GHz Intel Core i7 quad-core-processor and 8GB memory. Computations were run in MATLAB R2012b version 8.0.0.

3.5.1 Visually accessible example

In a simple test example case, we examine the algorithmic behaviour of `NR-IRLS` for a map

$$\Phi : \mathbb{R} \to \mathbb{R}^2, x \mapsto \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

and a measurement vector $y \in [0,1]^2$, and, consequently, the ℓ_p-minimizer $x_0 := \arg\min_x \|\Phi(x) - y\|_{\ell_p}^p$ will lie in $[0,1]$, too.
First we verify that the BCC in Definition 3.3 is fulfilled for $1 < p < 2$ in this particular setting with the lower BCC-bound $\alpha = 1$:

$$\|\Phi(x) - \Phi(x_0)\|_{\ell_p} = (|x - x_0|^p + |x^2 - (x_0)^2|^p)^{1/p} \geq |x - x_0| \geq \alpha \|x - x_0\|_{\ell_2}.$$

For the upper bound β we obtain $(1+2^p)^{1/p}$:

$$\begin{aligned}\|\Phi(x)-\Phi(x_0)\|_{\ell_p} &= (|x-x_0|^p+|x^2-(x_0)^2|^p)^{1/p} = (|x-x_0|^p+|x-x_0|^p\cdot|x+x_0|^p)^{1/p}\\ &\leq (|x-x_0|^p+|x-x_0|^p\cdot 2^p)^{1/p} = (1+2^p)^{1/p}|x-x_0| = \beta\|x-x_0\|_{\ell_2}\end{aligned}$$

In this situation, we usually encounter a nonconvex problem with possibly more than one local minimizer and aim to study the convergence behaviour of the of `NR-IRLS` in the version of Algorithm 2 in dependence of the nonconvexity paramter p. The variation of p also corresponds to the alteration of the underlying optimization problem and different minimizers or even a changing number of minimizers is possible. Here we investigate the differences in the optimization results for varying values of p and different choices of the initialization vector $x^{(0)}$. We compare the behaviour of `NR-IRLS` and MATLAB's built-in *lsqnonlin*-function, which is a realization of a trust-region-reflective or Levenberg-Marquardt strategy.

We provide a more detailed description of the parameter setting for our numerical tests in the following. For measurements $y=(0,0.9)^T$ and varying values of p in the range between 1 and 2, more precisely for the values $p\in\{1.1,1.3,1.7,1.9\}$, we study the algorithms recovery results.
For the specific setting of the algorithm parameters, we allow a maximum number of 50 iterations for `NR-IRLS`. Furthermore, we employ the MATLAB built-in function `fminunc` for the solution of the locally convex minimization problem in each internal step with default settings and use the last iterate for initialization. Also for running the `lsqnonlin`-function, which we directly use to solve the ℓ_p-minimization problem, we use MATLAB's default settings, too. For the experiments with both algorithms, we start from different points $x^{(0)}$ in the interval $[0,1]$, more precisely $x^{(0)}\in\{0,0.25,0.5,0.75,1\}$, and study their convergence behaviour resulting in different local minimizers as their outputs.

Via the graphical assessment of the algorithms' behaviour, we report the following experimental results: From Figure 3.1 and 3.2 it becomes clear that `NR-IRLS` converges to the critical point of the objective function with least distance to the ℓ_2-local minimizer resulting in the algorithms' first step, regardless of the value of p but in dependence on the starting point $x^{(0)}$ in the first step. The Figures 3.3 and 3.4 further underline the influence of the initialization point for the solution of the first nonlinear least squares problem. Also here `NR-IRLS` converges to the critical point that is closest to the minimizer of the ℓ_2-norm problem, while the standard MATLAB method converges to the local minimizer with least distance to the initialization point.

We draw as a conclusion that `NR-IRLS` has the potential to identify different local

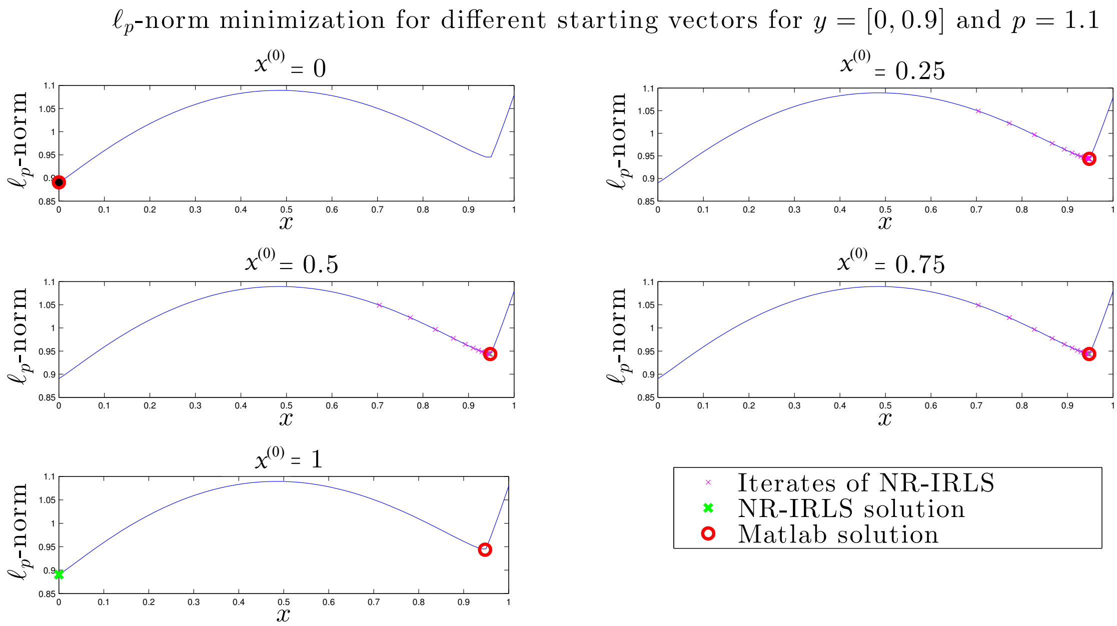


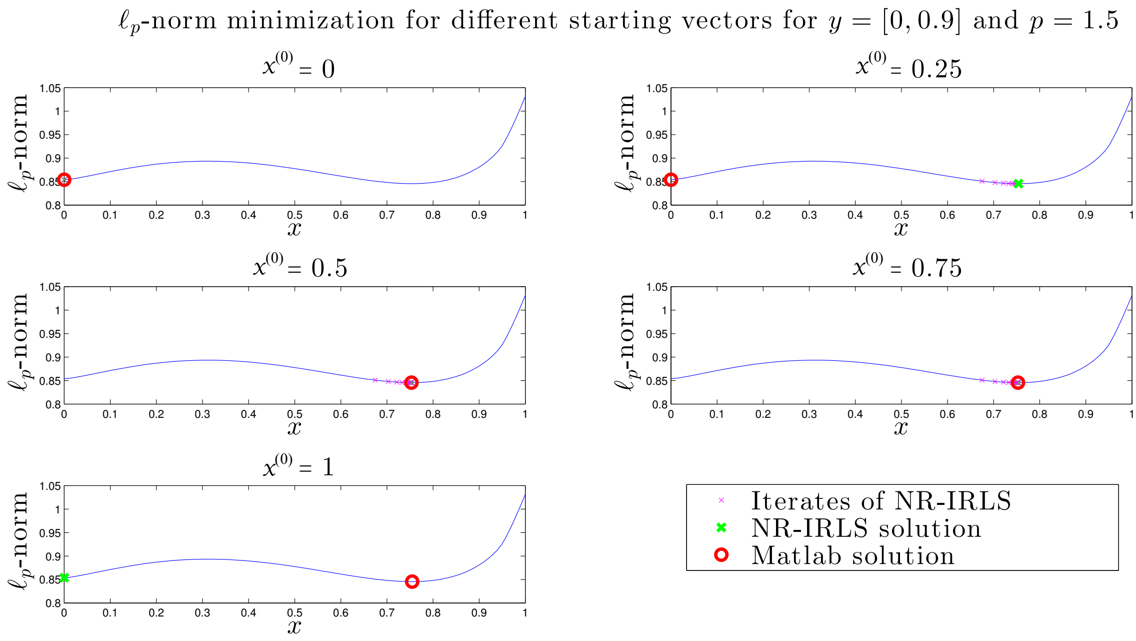
Figure 3.1

minimizers compared to standard gradient based methods also if the same starting point is provided to the methods.

3.5.2 High dimensional examples in a nonlinear compressed sensing application context

The paper [43] by the author of the thesis together with Fornasier and Ehler suggests a greedy algorithm (Algorithm 1) for the recovery of sparse vectors from a minimal amount of nonlinear measurements. This type of reconstruction problems will be called *nonlinear compressed sensing* problems, in particular for the case that the measurements generation involves randomness. We note that as a key operation of this algorithmic strategy is the solution of a nonlinear ℓ_p-minimization problem (3.1) in dimension d: at the d-th iteration of the algorithm it is necessary to identify the vector with at most d nonzero entries with best data fit, i.e., finding the solution to a minimal norm nonlinear residual problem as defined in (3.1).

In [43], the authors also considered $p \in [1, 2]$ as a norm parameter in the so-called Restricted Isometry Property (RIP), that is closely related to the BCC (see formula (3.1) in [43]). In the following experiments, we use nonlinear maps $\Phi : \mathbb{R}^k \to \mathbb{R}^m$, which are restrictions to k-dimensional index subspaces of two different types of measurement maps studied in [43]. One the one hand, we consider nonlinear maps that are constructed as the Lipschitz perturbations of matrices fulfilling the RIP. On the other



ℓ_p-norm minimization for different starting vectors for $y = [0, 0.9]$ and $p = 1.5$

$x^{(0)} = 0$

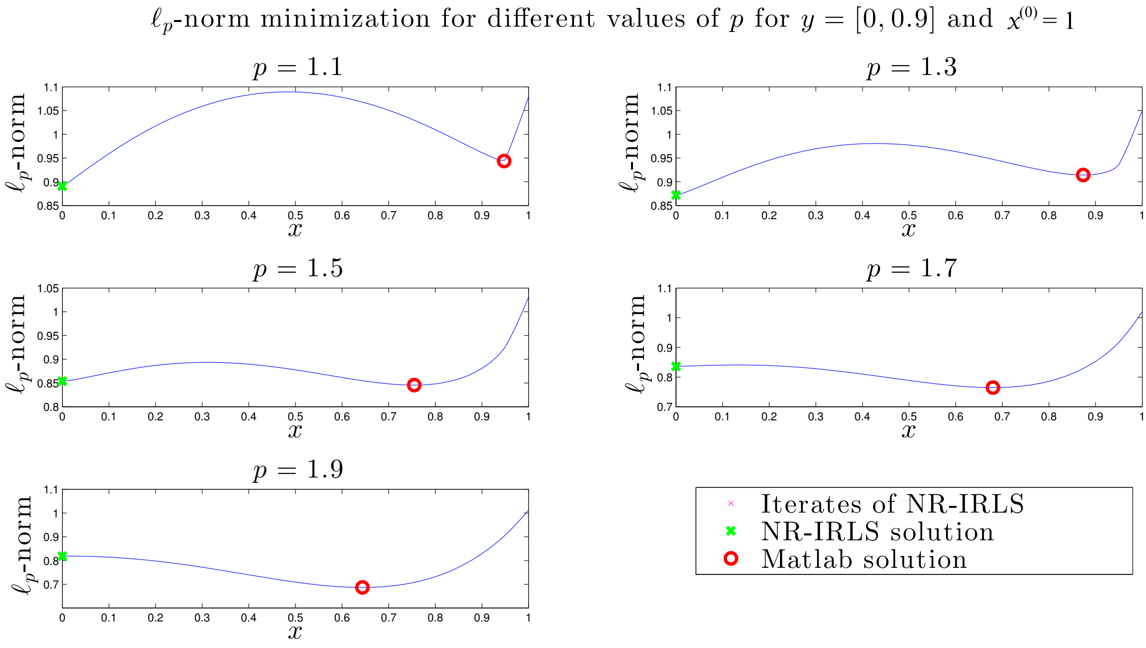
$x^{(0)} = 0.25$

$x^{(0)} = 0.5$

$x^{(0)} = 0.75$

$x^{(0)} = 1$

Figure 3.2



ℓ_p-norm minimization for different values of p for $y = [0, 0.9]$ and $x^{(0)} = 1$

$p = 1.1$

$p = 1.3$

$p = 1.5$

$p = 1.7$

$p = 1.9$

Figure 3.3

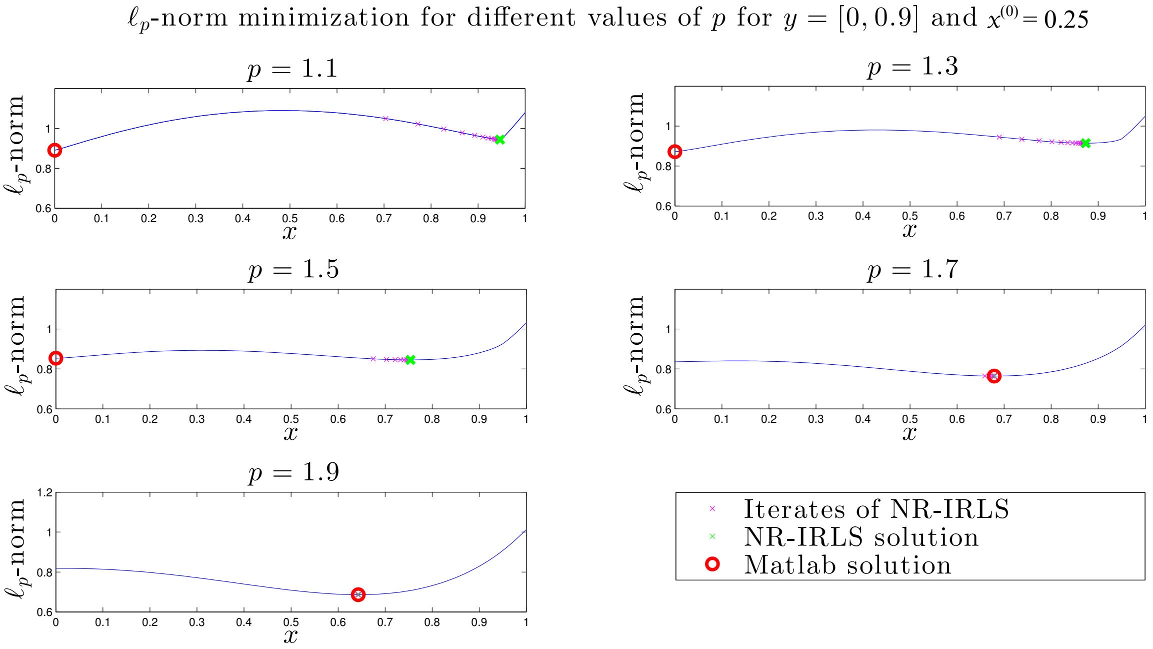


Figure 3.4

hand, the second setting in [43] involves the quadratic map $\Phi(x) = (|\langle x, \phi_i\rangle|^2)_{i=1,\ldots,m}$, which encodes the amplitudes of the scalar products of a vector x using a given collection of measurement vectors $\{\phi_1, \ldots, \phi_m\}$. We note that here for the second type of map solving the equation $\Phi(x) = y$ reduces to the recovery of the unknown signs of the scalar products, which is also the fundamental challenge in the solution of the more complex phase retrieval problem with applications, e.g., in X-ray crystallography [41, 47, 57].

For both of these measurement settings, we conduct numerical experiments testing `NR-IRLS` as in Algorithm 2 and the locally convexified version `NR-IRLS2` as in Algorithm 3 in comparison with standard MATLAB optimization methods. MATLAB source code implementing the greedy algorithm in the context of nonlinear compressed sensing is available at `http://www-m15.ma.tum.de/Allgemeines/SoftwareSite`.

3.5.2.1 Locally convex case: Nonlinear perturbation of linear RIP-matrices

We want to familiarize the reader first with a result presented in [43, Section 3.2.1].

Proposition 3.24. *Assume $k \leq m \leq N$ and $\Phi_1 \in \mathbb{R}^{m\times N}$ satisfies the δ-RIP of order $2k$, i.e.,*

$$(1-\delta)\|z\|_{\ell_2^N} \leq \|\Phi_1 z\|_{\ell_2^m} \leq (1+\delta)\|z\|_{\ell_2^N},$$

for all $z \in \mathbb{R}^N$ with at most $2k$ nonzero entries. If $\Phi_\rho : \mathbb{R}^N \to \mathbb{R}^m$ is chosen as

$$\Phi_\rho(z) := \Phi_1 z + \rho f(\|z - z^\circ\|_{\ell_2}^2)\Phi_2 z, \tag{3.28}$$

where $z^\circ \in \mathbb{R}^N$ is some reference vector in $\mathbb{R}^N$, $f : [0,\infty) \to \mathbb{R}$ is a bounded Lipschitz continuous function with $f(0) = 0$, $\rho > 0$ is a sufficiently small scaling factor, and $\Phi_2 \in \mathbb{R}^{m\times N}$ arbitrarily fixed, then there are constants $\alpha, \beta > 0$, such that for $p = 2$

$$\alpha\|z - \bar{z}\|_{\ell_2^N} \leq \|\Phi_\rho(z) - \Phi_\rho(\bar{z})\|_{\ell_p^m} \leq \beta\|z - \bar{z}\|_{\ell_2^N}$$

for all z with at most k nonzero entries and $\bar{z}$ is another fixed vector of at most k nonzero entries. For other $p \in [1, 2)$, these inequalities hold again with different constants α, β, derived, for instance, by equivalence of norms: for $0 < r < q$ we have $\|z\|_{\ell_q} \leq \|z\|_{\ell_r} \leq N^{1/r-1/q}\|z\|_{\ell_q}$.

One can infer from the proposition above that any restriction of Φ_ρ to vectors that are supported on a certain fixed index set $\Lambda \subset \{1, \dots, N\}$ with cardinality $\#\Lambda = k$ satisfies the BCC condition. Therefore, in the following, without loss of generality we place the assumption $\Lambda = \{1, \dots, k\}$ and let

$$\Phi : \mathbb{R}^k \times \mathbb{R}_+ \to \mathbb{R}^m, (x, \rho) \mapsto \Phi(x, \rho) = \Phi_\rho(x^\Lambda),$$

where $z = x^\Lambda$ represents the zero padding extension of x to a vector in higher dimension $\mathbb{R}^N$.

As shown in [135], in the linear case of $\Phi(\cdot, 0)$, i.e., where $\rho = 0$ and $\Phi(x, 0) = (\Phi_1)_{|\Lambda}$ boils down to a matrix in $\mathbb{R}^{m\times k}$, the first USCC holds true. Assuming that the parameter $\rho > 0$ is small, the map $\Phi(\cdot, \rho)$ is only a slight nonlinear perturbation of $\Phi(\cdot, 0)$. Moreover, we introduce the additional condition that f is twice continuously differentiable on $\mathbb{R}_+$ as used for the definition of Φ_ρ to extend the first USCC to $\Phi(\cdot, \rho)$ on a small ball around x_0 but do not present details of the rather clear elaboration of the argument.

We want to describe the setting for the upcoming numerical experiments where we perform the recovery of a sparse vector $z_0 \in \mathbb{R}^N$ with maximal k nonzero entries, where $k \in [1, 10] \cap \mathbb{N}$ from measurement results $y = \Phi_\rho(z_0)$ applying the method [43, Algorithm 1]. At each step of this algorithm, the minimization or a norm nonlinear residual has to be performed, where we employ Algorithm 2 of the present chapter. The ambient dimension $N = 80$ as well as the number of measurements $m = 30$ are fixed and we sample at random RIP matrices Φ_1 with i.i.d. Gaussian entries. Next we set Φ_2 to be as the matrix with all ones and the perturbation function f is the squared Euclidean distance from the given solution vector z_0, i.e., $f(\|z - z_0\|_{\ell_2}^2) = \|z - z_0\|_{\ell_2^N}^2$. We already mentioned earlier via the parameter $\rho > 0$, that steers the nonlinearity of the measurement operator, it is possible to regulate the validity of the BCC and the USCC property. In our tests, we explored the dependence of success rate on the nonlinearity by observing results for the parameter range $\rho \in \{0, 0.5, 1, 3, 5, 10, 20\}$ in 100 randomly generate synthetic problems for each of these choices of ρ. As we use synthetic data and the true sparse minimizer z_0 is known, we can use z_0 to measure the recovery success and categorize a reconstruction as successful as soon as the error is within a 1% of the norm of the solution vector z_0. Additional measurement noise was not included in these experiments.

We are quite generous and allow the execution of $3k$ steps of the greedy algorithm [43], which corresponds to a number ob iterations that is notably exceeding the intrinsic dimension of the sparse solution. This gives the algorithm the chance to correct wrongly chosen indices, that were added to the support set in previous iterations. The maximum number of iterations for the nonlinear ℓ_p-residual minimization performed with `NR-IRLS` is set to 50. Again, we employ the MATLAB built-in function `fminunc` with its default settings inititalization in the origin for solving the locally convex minimization problem appearing in each inner step.

First we present the empirical probability of successful recovery of sparse vectors for [43, Algorithm 1] implementing Algorithm 2 for the execution of the ℓ_p-minimization for varying values of p in the following Figure 3.5.

Our expectations on a decreasing recovery performance with growing sparsity level k and, consequently, increasing dimension of the ℓ_p-optimization problem were met by the experiment results. Also an increasing perturbation factor $\rho > 0$ for the construction of Φ_ρ, which corresponds to the severity of the nonlinearity, was supposed and verified in our experiments. The better recovery performance for the parameters p closer to 2 can be explained by the fact that the BCC condition has tighter bounds α and β in these cases.

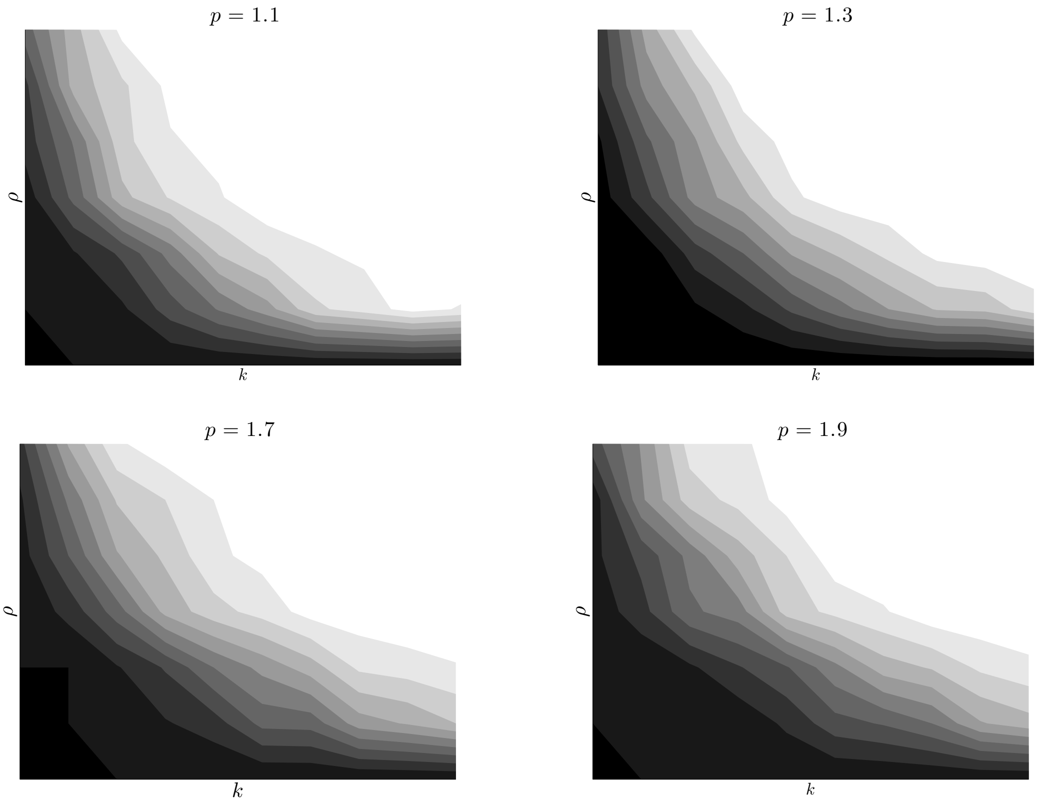


Figure 3.5: Recovery rates for the greedy strategy developed in [43] used for the measurement setting with perturbed RIP matrix as defined above in dimension $N = 80, m = 30$, where Φ_1 has i.i.d. Gaussian entries, Φ_2 being the matrix with all ones , $f(x) = \|x - x_0\|_{\ell_2}^2$, and solution vector x_0 with $\|x_0\|_{\ell_2} = 0.015$. Reconstruction is executed 50 times for each signal and sparsity level k to obtain stable recovery rates.

3.5.2.2 Phase retrieval problem

As introduced in [43] for the setting of phase retrieval, we consider a sequence of Gaussian random vectors $\phi_i \in \mathbb{R}^N$, $i = 1, \dots, m$ and construct the nonlinear measurement map as follows

$$\Phi(x) = (|\langle \phi_1, x \rangle|^2, \dots, |\langle \phi_m, x \rangle|^2)^\top. \tag{3.29}$$

With slight modifications to its original formulation, a BCC-type property holds for $p = 1$ replacing the ℓ_2-norm on both sides of the inequality by a Hilbert-Schmidt norm, which does not have disturb the validity of the results above. The existence of the corresponding BCC-constants $\alpha, \beta > 0$ can be assured by [43, Theorem 3.12] and according to [43, Formula (3.14)].

Unfortunately, in the general case, we can not assume that the USCC-property holds true in this setting and, therefore, the convexified `NR-IRLS` as in Algorithm 3 is applied.

As in the experiments above, the signal dimension is chosen to be $N = 80$ and the number of measurements and, thereby, the number of sampled i.i.d. Gaussian random vectors ϕ_i, $i = 1, \dots, m$ is set to $m = 30$. We generate synthetic sparse solutions z_0 with $\|z_0\|_{\ell_2} = 1$ with respective sparsity levels in the range $k \in \{1, 3, 6, 9, 12, 15, 18, 21\}$ and a nonincreasing rearrangement of their entries with decay rate $\kappa \in \{1, 0.8, 0.6, 0.4\}$, where a precise definition of the vector class $\mathcal{D}_\kappa$ can be found in [43]. The convergence results in [43] demand such a decay property and this theoretical requirement was justified by numerical tests involving MATLAB optimization routines in the intermediate steps.

For each of the mentioned parameter combinations, we create 50 noise free problem instances. Again, we exploit the knowledge of the solution vector z_0 for the classification of the recovery success which is defined to occur, when the error does not exceed 1% of the solution's norm.

Similar to the procedural settings in the experiments above, the greedy algorithm [43, Algorithm 1] performs maximally $3k$ steps and the maximum number iterations of the `NR-IRLS` method is bounded from above by 100. As we use the convexified version of `NR-IRLS`, we chose the regularization parameter $\omega = 100 > 0$, which is large enough in our context. Again, we use the MATLAB function $fminunc$ with its default settings now starting from randomly chosen points within a ball with radius of the solutions norm for performing the convex minimization in the internal step.

The graphics in Figures 3.6-3.7 below show the success rates of [43, Algorithm 1] implementing Algorithm 3 for the execution of the ℓ_p-minimization for varying values of p for the recovery of sparse vectors from measurements of the type Proposition 3.24. To our great surprise, the influence of the decay rate of the nonincreasing rearrangement of the solution vector z_0 on the recovery success is becoming less prominent when employing `NR-IRLS` for the solution of the internal ℓ_p-minimization problem. This observation is in stark contrast to the experiment results presented in [43], in which we applied the built-in MATLAB functions `fminunc`, `fminsearch` or `lsqnonlin`. For illustration of the performance differences, we compare the recovery results of the greedy algorithm incorporating `NR-IRLS` and the corresponding results obtained from an implementation using `lsqnonlin`. We observe that the later are significantly outperformed by the implementation using `NR-IRLS` in cases, where the decay rate of the nonincreasing rearrangement of the vector z_0 is not sufficiently pronounced.

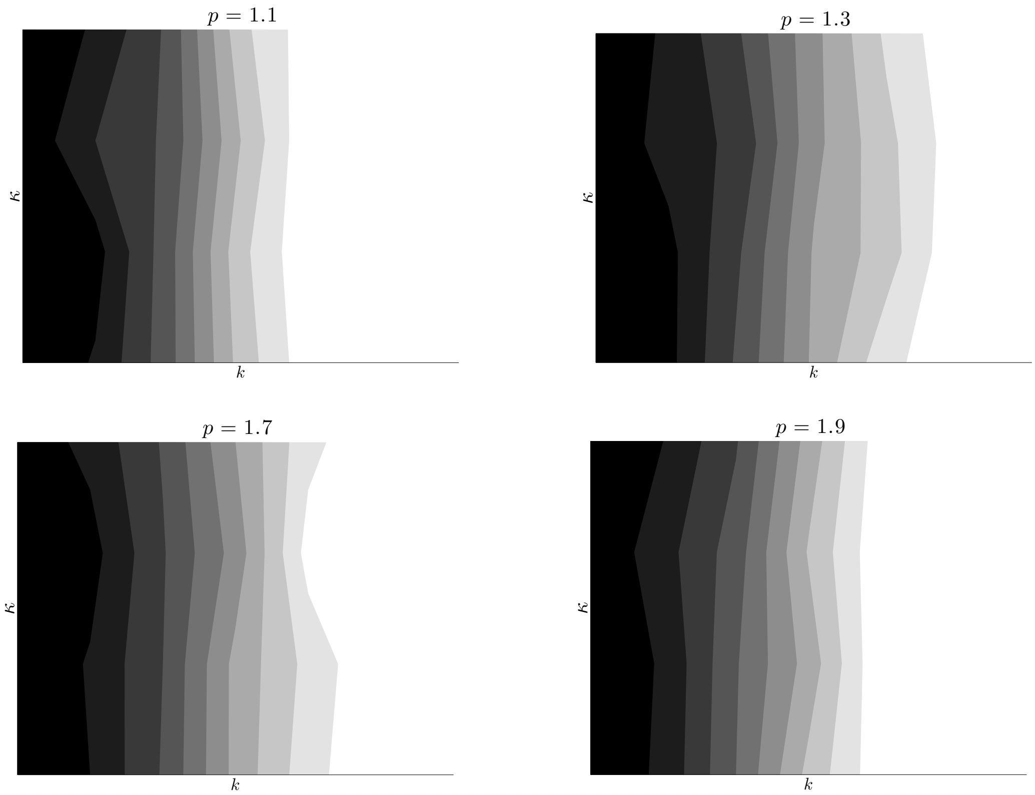


Figure 3.6: Recovery rates for the greedy strategy developed in [43] implemented with `NR-IRLS` used on the phase retrieval problem with Gaussian measurement vectors as above with $N = 80, m = 30$, and we use solutions x_0 with $\|x_0\| = 1$. Reconstruction is repeated 50 times for each signal and k, κ.

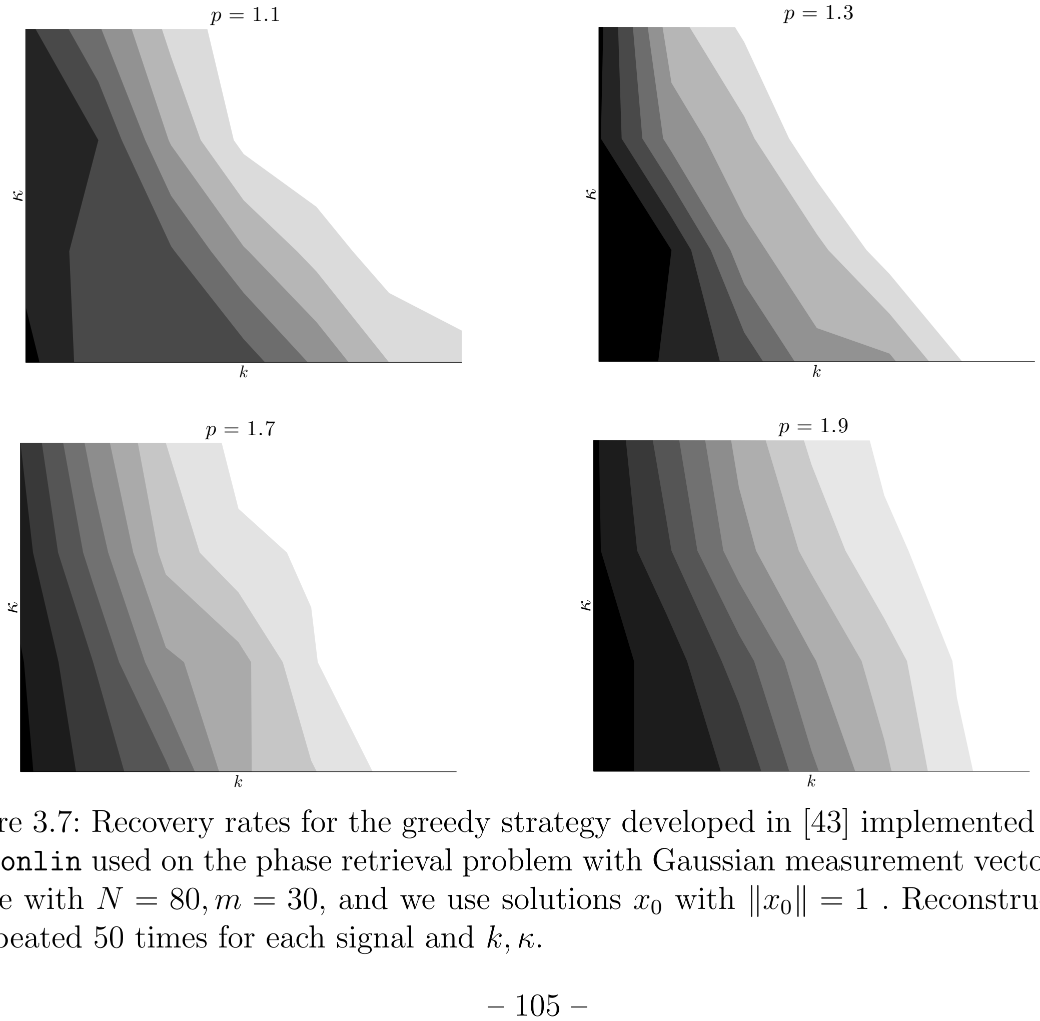


Figure 3.7: Recovery rates for the greedy strategy developed in [43] implemented with `lsqnonlin` used on the phase retrieval problem with Gaussian measurement vectors as above with $N = 80, m = 30$, and we use solutions x_0 with $\|x_0\| = 1$. Reconstruction is repeated 50 times for each signal and k, κ.

3.5.3 Recovery from data with impulsive noise perturbation

In this last subsection, we consider the case of observed measurement results y that are additionally corrupted by noise. As already explained in Section 2.1, the choice of the error function for the nonlinear residual minimization depends on the particular kind of noise. In the following, we will examine the particular case of measurement corruption by impulsive noise. This type of noise is characterized by the random appearences of instantaneous perturbations of the residual components which appear in the form of spikes or pulses with random amplitude. This corresponds the occurrence of sparse residual distortions and, therefore, sparsity enhancing error functions such as the ℓ_1-norm are a reasonable choice.

We again consider the phase retrieval problem as described above introducing impulsive noise on the measurements and adopt the experimental setting used in the prior numerical tests up to slight modifications mentioned below. With the experiments in this subsection, we aim at to explore the influence of the choice of the parameter $1 \leq p \leq 2$ on the recovery success rates for impulsive noise corrupted measurement results.

First we want to give a clear description of our statistical model for the impulsive noise: for this purpose we combine a binary-valued random sequence model for modeling the time of occurrence of the noise pulse with a continuous-valued random process model defining the pulse amplitude. An important instance of a statistical process for impulsive noise modeling by an amplitude modulated binary sequence is the so-called Bernoulli-Gaussian process [155]. In this specific model of an impulsive noise process, the random time of the impulse occurrence is modeled by a binary Bernoulli process B_{α_p} with a probability of success α_p and as an amplitude model a Gaussian process $\mathcal{N}_{(0,1)}$ with mean 0 and standard deviation 1 is used.

Now having a proper model for impulsive noise at hand, we give a detailed description of the measurement setting for the noisy phase retrieval problem.

As in previous experiments, we set $N = 80$, $m = 30$ and sample i.i.d. Gaussian random vectors ϕ_i, $i = 1, \dots, m$. Again we generate synthetic unit norm solutions z_0 with respective sparsity $k \in \{1, 2, 3, 5, 7, 9\}$ and decay rate $\kappa = 0.5$ for the nonincreasing rearrangement of the absolute value of their entries, where we refer to [43]for the definition of the vector class $\mathcal{D}_\kappa$. Next we create impulsive Bernoulli-Gaussian noise vectors with parameters $\alpha_p \in \{0.5, 0.4, 0.3, 0.2, 0.1, 0.0\}$, respectively adjusting the scaling to the norm of the measurement vector y, and add the result to the exact measurement vector. We generate 100 problem instances for each parameter combinations, and again use the solution z_0 for the determination of the recovery success, which we claim to be reached when the error is within a 5%-margin of the solution's norm. We use a similar algorithm set up as above executing $3k$ steps of the greedy algorithm [43, Algorithm 1] and for `NR-IRLS` itself the limit for the number of iterations is set to 50. We chose the parameter $\omega = 100$ and use the MATLAB built-in function `fminunc` with default settings and random starting points within the ball with radius of the solutions norm for the application to the convex minimization problem in each inner step.

The plots in Figure 3.8 illustrate the success rates of [43, Algorithm 1] for sparse vector recovery from measurements of the type Proposition 3.24 perturbed by impulsive noise, which implemens Algorithm 3 for the execution of the ℓ_p-minimization for variying values of p.

The visual assessment of the phase transition diagrams can be summarized as follows: minimization for small values of p and, therefore, the stronger the nonsmoothness of the error function, are preferable to standard ℓ_2-least squares minimization which matches our expectations. Additionally, we observe that, if p is getting close to 1 and low sparsity level k, recovery with `NR-IRLS` is very robust even for strong impulsive noise perturbations.

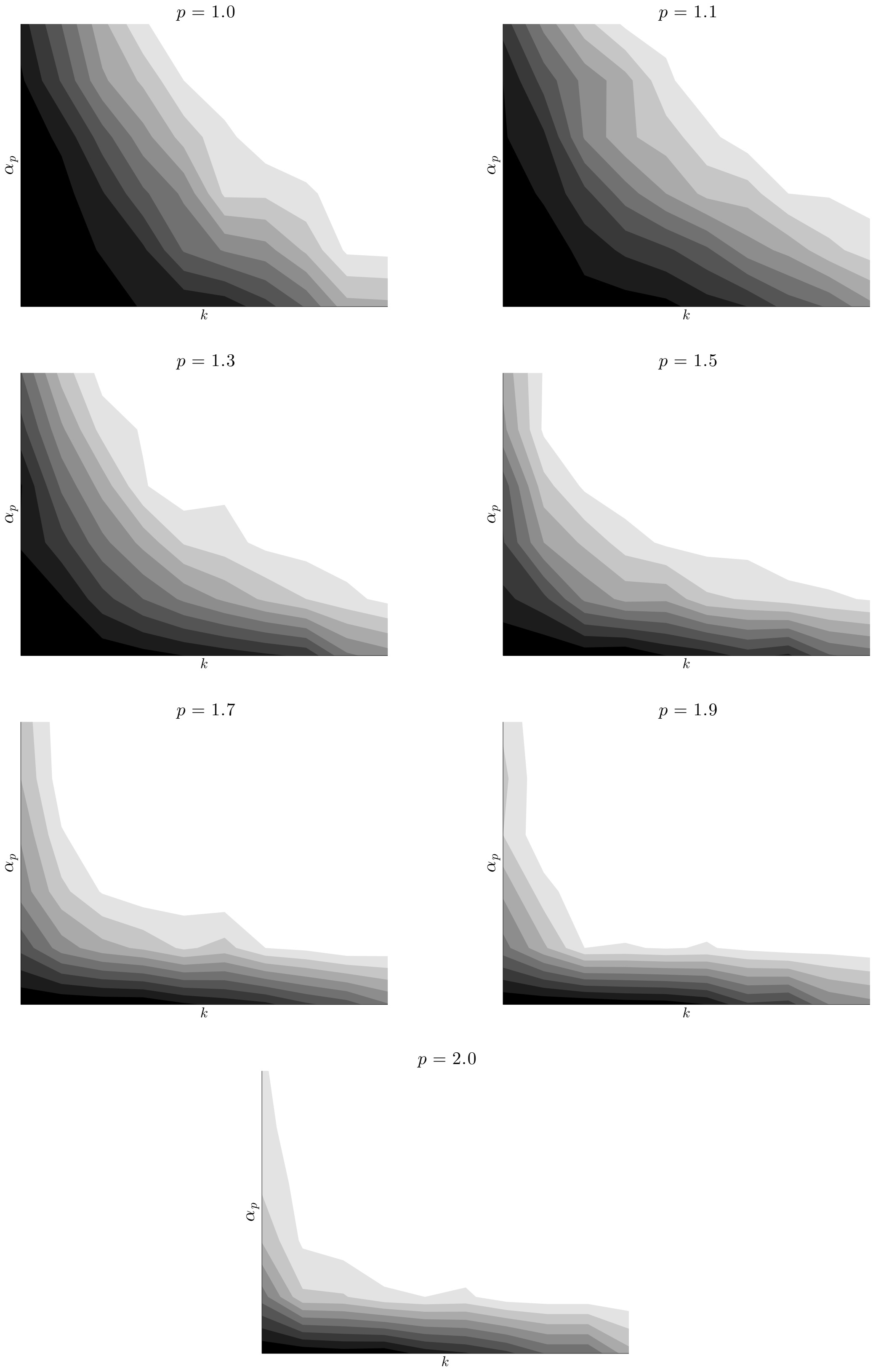


Figure 3.8: Recovery rates for the greedy strategy developed in [43] implemented with NR-IRLS used on the phase retrieval problem with Gaussian measurement vectors as above with $N = 80, m = 30$, and we use solutions x_0 with $\|x_0\| = 1$. Reconstruction is repeated 50 times for each signal with sparsity k and the particular noise perturbation as given above.

Chapter 4

Harmonic mean IRLS for low-rank matrix recovery

In the last years, the problem of recovering large scale matrix-valued signals with an inherent low-rank structure from incomplete linear measurements as discussed in Section 2.3 and especially in Section 2.3.2, attracted the attention of the signal processing and machine learning community. That means we aim at the unique identification of an unknown matrix $X_0 \in M_{d_1 \times d_2}$ from a linear equation system

$$\Phi(X) = Y \tag{4.1}$$

with a linear operator $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m$ and measurement vector $Y \in \mathbb{R}^m$ for $m \ll d_1 d_2$ under the additional assumption that X_0 has rank $r < \min(d_1, d_2)$.

Having in mind the discussion on intrinsic structures, in particular sparsity type structures in Section 2.3.1 and Section 2.3.2, we recognized that with the additional assumption of the low-rank structure of the solution, the recovery of X_0 from (4.1) becomes feasible by solving the affine rank minimization problem (2.55)

$$\min \operatorname{rank}(X) \text{ subject to } \Phi(X) = Y. \tag{4.2}$$

Low-rank matrix recovery problems of this type appear in application frameworks such as system identification [96, 97], signal processing [1], quantum tomography [71, 73], recommender systems [22, 64, 136] and phase retrieval [18, 23, 72].

A widely studied instance of the low-rank matrix recovery problem with great relevance in recommender systems is the problem of the identification and recovery of a large scale low-rank matrix from a subset of revealed entries, the so-called *matrix completion* *problem*, i.e., the choice of the measurement operator $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m$, where we are given m sample entries

$$\Phi(X)_\ell = X_{i_\ell, j_\ell}, \tag{4.3}$$

for $\ell \in [m]$ and some $i_\ell \in [d_1]$ and $j_\ell \in [d_2]$ depending on ℓ. A well-known example is the *Netflix problem* aiming at the completion of a matrix with user ratings for movies with more than 10^8 entries [7].

Despite the fact that the low-rank matrix recovery problem is NP-hard just like in the sparse vector case, a number of tractable methods that can provably achieve recovery in many relevant scenarios have been developed. For instance, the *nuclear norm minimization* (`NNM`) approach is particularly well-understood [22, 46], considering the convex relaxation of (4.2) as a proxy and solving the problem

$$\min \|X\|_{S_1} \text{ subject to } \Phi(X) = Y, \tag{4.4}$$

called *nuclear norm minimization* [21, 22, 126]. For `NMM`, the number of measurements necessary for successful recovery scales with *optimal order*, i.e., $m \geq \rho r(d_1 + d_2 - r)$. Nevertheless, the oversampling factor ρ is significantly larger than 1 . Therefore, the number of measurements is not *optimal* in the sense of the information theoretical lower bound $r(d_1+d_2-r)$ corresponding to the degrees of freedom presented in Lemma 2.20. For the interesting case of matrix completion the required number of measurements for reconstruction is even higher involving additional log-factors of the matrix dimension [31]. Although computation time for the solution of `NNM` scales polynomially, it becomes computationally challenging for growing dimensionality of the problem and intractable for many potential application settings.

With regard to these limitations of techniques based on convexification, the investigation of nonconvex optimization methods for low-rank matrix recovery [74, 78, 79, 145, 146, 149, 154, 161–163, 169] is proceeding rapidly already, where for several methods theoretical recovery guarantees comparable to those of `NNM` have been developed [20, 142, 144, 149, 169]. They have practical advantages such as a high empirical recovery rate and an efficient algorithm implementation, but, nevertheless, their successful convergence to the low-rank minimizer often heavily depends on a distinct, computationally demanding initialization step.

Following this path, in this chapter we will discuss a new iteratively reweighted least squares algorithm for the low-rank matrix recovery problem based on the minimization of the Schatten-p quasi-norm with non-convexity parameter $0 < p < 1$

$$\min \|X\|_{S_p}^p \text{ subject to } \Phi(X) = Y. \tag{4.5}$$

The abovementioned IRLS-algorithm was first introduced in the conference paper [86] and the respective journal article preprint [87] by Christian Kümmerle and the author of the thesis, where both contributed equally to all parts of the publication. The statement of the results in this chapter closely follows the presentation in [87].

The strategy of using an IRLS-type method for the approximation of (4.5) is not new and the corresponding algorithms appeared already in the papers[51, 106] published several years ago. Still, both of the IRLS-approaches presented in those publications are not able to fully generalize the properties of the algorithm for sparse vector recovery in [35]. The most important point is that neither the theoretical nor the numerical results in [51, 106] indicate the occurrence of the superlinear convergence rate for non-convex parameters $p < 1$ that is significantly pronounced in the vector case [35].

However, the algorithm under discussion in this chapter, introduces a new kind of weight matrices, so-called *harmonic mean weight matrices*, that will lead to several important improvements and innovations. The construction method for the weight matrices can be interpreted as the averaging of left- and right-sided weight matrices introduced in [51, 106] by taking their harmonic mean. This interpretation led to the choice of the name harmonic mean iteratively reweighted least squares (HM-IRLS) for the algorithm.

This new design of the weight matrices is *more symmetrical* than the weight matrices previously used [51, 106], and this empowers HM-IRLS to exploit the information in both the column and the row space of the iterates. More precisely, the specific structure of the harmonic mean weight matrices allows a better alignment of the left-singular and right-singular vectors of the iterates with those of the low-rank matrix to be recovered.

In the course of this chapter, we will demonstrate that by the employment of the harmonic mean weight matrices in HM-IRLS it is not only possible to overcome the disadvantages of the weight matrices used in [51, 106] but to also outperform them.

In the first section, we will introduce as a calculation tool the Kronecker product, which helps us to introduce the construction concept for the harmonic mean weight matrix.

Thereafter, in Section 4.3, we take inspiration from the theoretical analysis of existing IRLS-methods, where some of the findings are based on *null space properties* of the map Φ, to derive the corresponding analysis results also for HM-IRLS. To be more precise, using a similar auxiliary functional $\mathcal{J}_{HM}$, we are able to prove convergence of the sequence of iterates of HM-IRLS to stationary points of an ϵ-smoothed Schatten-p functional (analogous to (2.58)). In the case $\epsilon = 0$, recovery of a low-rank matrix is proven and in contrast to the IRLS-methods in [51, 106], we are able to establish a *local superlinear convergence rate* (of order $2 - p$) for `HM-IRLS` in Section 4.3.6, which implies that for the parameters $p \to 0$ the convergence rate is approaching quadratic.

Within the theoretical analysis section we want to draw the reader's attention in particular to the high technical sophistication of the proofs of the results in Lemma 4.6 and Theorem 4.16.

Our theoretical guarantees are validated by numerical experiments presented in Section 4.4 comparing the recovery ability and convergence speed of `HM-IRLS` with related IRLS-algorithms for low rank recovery. Moreover, we conduct extensive numerical tests comparing the recovery performance of `HM-IRLS` with the exisiting IRLS variants [51, 106], Riemannian optimization techniques [154], alternating minimization approaches [74, 146], algorithms based on iterative hard thresholding [9, 85], and others [118], with respect to sample complexity. Although our theoretical findings are not directly applicable to the matrix completion measurement model, we focused on this setting in our experiments due to its popularity in the machine learning community facilitating the comparison with other algorithms.

4.1 Towards a harmonic mean weight matrix

4.1.1 Kronecker and Hadamard products

Considering the matrices $A = (a_{ij})_{i\in[d_1],j\in[d_3]} \in \mathbb{R}^{d_1\times d_3}$ and $B \in \mathbb{R}^{d_2\times d_4}$, the representation in matrix form of their tensor product with respect to the standard bases is referred to as the *Kronecker product* $A\otimes B \in \mathbb{R}^{d_1\cdot d_2\times d_3\cdot d_4}$. The matrix resulting from this operation, $A\otimes B$ is a block matrix with $d_2 \times d_4$ blocks whose block of index $(i,j) \in [d_1]\times[d_3]$ is the matrix $a_{ij}B \in \mathbb{R}^{d_2\times d_4}$. Illustrating this exemplary for $A \in \mathbb{R}^{d_1\times d_3}$ with $d_1 = 2$ and $d_3 = 3$ we obtain

$$A \otimes B = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \otimes B = \left[\begin{array}{c|c|c} a_{11}B & a_{12}B & a_{13}B \\ \hline a_{21}B & a_{22}B & a_{23}B \end{array}\right].$$

A collection of basic properties of the Kronecker product can be found in [8, Chapter 7], [152] and we present some of them that will be particularly useful in calculations later on:

(i) $(A\otimes B)^* = A^* \otimes B^*$,

(ii) $(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$ (whenever A and B are invertible),

(iii) $(A\otimes B)(C\otimes D) = (AC\otimes BD)$,

(iv) $A\otimes(B\otimes C) = (A\otimes B)\otimes C$.

Next we introduce the *Hadamard product* $A\circ B \in \mathbb{R}^{d_1\times d_2}$ of two matrices $A\in\mathbb{R}^{d_1\times d_2}$ and $B \in \mathbb{R}^{d_1\times d_2}$, that corresponds to their entry-wise product

$$(A\circ B)_{i,j} = A_{i,j}B_{i,j}$$

with $i\in[d_1]$ and $j \in [d_2]$. In the literature, the Hadamard product is sometimes also referred to as *Schur product*, and for the reader's convenience we provide some of its basic calculation rules.

(i) $A\circ B = B\circ A$,

(ii) $(A\circ B)^* = A^*\circ B^*$,

(iii) $A\circ(B+C) = A\circ B + A\circ C$,

(iv) $A\circ(B\circ C) = (A\circ B)\circ C$.

Further properties are listed in [8, Chapter 7].

The Kronecker product often appears in the context of matrix equations which involve multiplications of matrices from the left and right side to the variable X as follows

$$AXB^* = Y \qquad \text{if and only if} \qquad (B \otimes A)X_{\text{vec}} = Y_{\text{vec}}.$$

Moreover, for matrices A and B with $d_1 = d_3$ and $d_2 = d_4$, we give the definition of the *Kronecker sum* $A \oplus B \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ of two matrices $A \in \mathbb{R}^{d_1 \times d_1}$ and $B \in \mathbb{R}^{d_2 \times d_2}$ as the following matrix in $\mathbb{R}^{d_1 d_2 \times d_1 d_2}$

$$A \oplus B = (\mathbf{I}_{d_2} \otimes A) + (B \otimes \mathbf{I}_{d_1}).$$

The Kronecker sums are a useful tool for the reformulation of the Sylvester matrix equation problem, where one wants to find $X \in M_{d_1 \times d_2}$ solving the equation system

$$AX + XB^* = Y \tag{4.6}$$

for fixed $A \in \mathbb{R}^{d_1 \times d_1}, B \in \mathbb{R}^{d_2 \times d_2}$ and $Y \in \mathbb{R}^{d_1 \times d_2}$ given. We can reformulate the equation above

$$(A \oplus B)X_{\text{vec}} = Y_{\text{vec}},$$

where, again, the vectorizations of X and Y is used. In this framework, we can exploit the explicit formula of the inverse $(A \oplus B)^{-1}$ of the Kronecker sum $A \oplus B$ that expresses this matrix in terms of singular vectors and singular values of A and B.

Lemma 4.1 ([80]). *Let $A \in \mathbb{H}^{d_1}$ and $B \in \mathbb{H}^{d_2}$, where one of the matrices is positive definite and the other positive semidefinite. Denote the singular vectors of A by $u_i \in \mathbb{R}^{d_1}$, $i \in [d_1]$, its singular values by σ_i, $i \in [d_1]$ and the singular vectors resp. values of B by $v_j \in \mathbb{R}^{d_2}$ resp. μ_j, $j \in [d_2]$, then*

$$(A \oplus B)^{-1} = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{v_j v_j^* \otimes u_i u_i^*}{\sigma_i + \mu_j} = (V \otimes U) D (V \otimes U)^*, \tag{4.7}$$

where $D \in M_{d_1 d_2 \times d_1 d_2}$ is a diagonal matrix with entries $d_l = (\sigma_i + \mu_j)^{-1} > 0$ for $l = (i-1)d_2 + j$, $U = [u_1, \ldots, u_{d_1}]$, and $V = [v_1, \ldots, v_{d_2}]$.

Furthermore, the action of $(A \oplus B)^{-1}$ on the matrix space $M_{d_1 \times d_2}$ can be written as

$$\big[(A \oplus B)^{-1} Z_{\text{vec}}\big]_{\text{mat}} = U\big(H \circ (U^* Z V)\big)V^*. \tag{4.8}$$

for $Z \in M_{d_1 \times d_2}$ and the matrix $H \in M_{d_1 \times d_2}$ with the entries $H_{i,j} = (\sigma_i + \mu_j)^{-1}$, $i \in [d_1]$, $j \in [d_2]$.

For proving Lemma 4.1, one can employ the Kronecker product calculation rules presented above.

4.1.2 Averaging of weight matrices

Let us for the moment suppose that $Z \in \mathbb{R}^{d_1 \times d_2}$ is a square matrix with $d_1 = d_2$ of full rank. Having in mind Definition 2.37, we now introduce two different reformulations of the p-th power of its Schatten-p quasi-norm as weighted ℓ_2-norms for the vectorized notation X_{vec} involving the Kronecker product,

(i) $\|Z\|_{S_p}^p = \operatorname{tr}[(ZZ^*)^{\frac{p}{2}}] = \operatorname{tr}[(ZZ^*)^{\frac{p-2}{2}}(ZZ^*)] = \operatorname{tr}(W_L ZZ^*) = \|W_L^{\frac{1}{2}} Z\|_F^2 = \|Z\|_{F(W_L)}^2$
$= \|(\mathbf{I}_{d_2} \otimes W_L)^{\frac{1}{2}} Z_{\text{vec}}\|_{\ell_2}^2 = \|Z_{\text{vec}}\|_{\ell_2(\mathbf{I}_{d_2} \otimes W_L)}^2$,
where W_L is the symmetric weight matrix $(ZZ^*)^{\frac{p-2}{2}}$ in $M_{d_1 \times d_1}$ and $\mathbf{I}_{d_2} \otimes W_L$ is the block diagonal weight matrix in $M_{d_1 \cdot d_2 \times d_1 \cdot d_2}$ with d_2 instances of W_L on the diagonal blocks,

(ii) $\|Z\|_{S_p}^p = \operatorname{tr}[(Z^*Z)^{\frac{p}{2}}] = \operatorname{tr}[(Z^*Z)(Z^*Z)^{\frac{p-2}{2}}] = \operatorname{tr}(Z^*ZW_R) = \|ZW_R^{\frac{1}{2}}\|_F^2 = \|Z^*\|_{F(W_R)}^2$
$= \|(W_R \otimes \mathbf{I}_{d_1})^{\frac{1}{2}} Z_{\text{vec}}\|_{\ell_2}^2 = \|Z_{\text{vec}}\|_{\ell_2(W_R \otimes \mathbf{I}_{d_1})}^2$,
where W_R is the symmetric weight matrix $(Z^*Z)^{\frac{p-2}{2}}$ in $M_{d_2 \times d_2}$. The weight matrix $W_R \otimes \mathbf{I}_{d_1} \in M_{d_1 d_2 \times d_1 d_2}$ is, as can be calculated from the definition of the Kronecker product, a block matrix of diagonal blocks of the type $\operatorname{diag}((W_R)_{ij}, \ldots, (W_R)_{ij}) \in \mathbb{R}^{d_1 \times d_1}$, $i, j \in [d_2]$.

Note that in the case that the matrix Z is not of full rank or also if $d_1 \neq d_2$, the calculations carried out above can not be well-defined: at least one of the matrices $ZZ^* \in \mathbb{R}^{d_1 \times d_1}$ or $Z^*Z \in \mathbb{R}^{d_2 \times d_2}$ is singular, which does not allow for the inversion appearing in the definition of the matrices $W_R = (Z^*Z)^{\frac{p-2}{2}}$ or $W_L = (ZZ^*)^{\frac{p-2}{2}}$ for $p < 2$. As already suggested in Section 2.4, we can avoid these problems by introducing a smoothing parameter $\epsilon > 0$ and defining the smoothed weight matrices $W_L(Z, \epsilon) \in \mathbb{H}_{++}^{d_1}$ and $W_R(Z, \epsilon) \in \mathbb{H}_{++}^{d_2}$ as follows

$$W_L(Z, \epsilon) := (ZZ^* + \epsilon^2 \mathbf{I}_{d_1})^{\frac{p-2}{2}}, \tag{4.9}$$

$$W_R(Z, \epsilon) := (Z^*Z + \epsilon^2 \mathbf{I}_{d_2})^{\frac{p-2}{2}}, \tag{4.10}$$

In the papers [51, 106] the strategy of the reformulation of the Schatten-p norm via the left-sided reweighted Frobenius-norm $\|\cdot\|^2_{F(W_L)}$ was already exploited for their versions of `IRLS` algorithms for low-rank matrix recovery [51, 106]. Let us stress again that none of the two papers pointed out the idea of a reformulation of the Schatten-p norm using a weighted Frobenius-norm which involves the multiplication of the weight matrix W_R from the right.

In this chapter, our goal is to take advantage of the low rank information in the column *and* the row space by the combination of both reweighting strategies into one weight matrix reflecting this symmetry.

We first want to present the possibly most intuitive or naive way to a more symmetric exploitation of the low rank structure in the following lemma.

Lemma 4.2. *Let $0 < p \leq 2$ and $Z \in M_{d_1 \times d_2}$ with $d = d_1 = d_2$ be a full rank matrix. Then*

$$\|Z\|^p_{S_p} = \frac{1}{2}\left(\|W_L^{\frac{1}{2}} Z\|_F^2 + \|Z W_R^{\frac{1}{2}}\|_F^2\right) = \left\|\left(\frac{W_L \oplus W_R}{2}\right)^{\frac{1}{2}} Z_{\text{vec}}\right\|^2_{\ell_2} = \|Z_{\text{vec}}\|^2_{\ell_2(W_{(arith)})},$$

$$\textit{where } \frac{1}{2}\left(\mathbf{I}_{d_2} \otimes W_L + W_R \otimes \mathbf{I}_{d_1}\right) = \frac{W_L \oplus W_R}{2} =: W_{(arith)}$$

is the arithmetic mean matrix of the symmetric and positive definite weight matrices $\mathbf{I}_{d_2} \otimes W_L$ and $W_R \otimes \mathbf{I}_{d_1}$, $W_L := (ZZ^*)^{\frac{p-2}{2}}$, and $W_R := (Z^*Z)^{\frac{p-2}{2}}$.

Proof. Using the computations above as well as the cyclicity of the trace, we calculate

$$\begin{aligned}
\|Z\|^p_{S_p} &= \frac{1}{2}\left(\|W_L^{\frac{1}{2}} Z\|_F^2 + \|Z W_R^{\frac{1}{2}}\|_F^2\right) = \frac{1}{2}\left(\|(\mathbf{I}_{d_2} \otimes W_L)^{\frac{1}{2}} Z_{\text{vec}}\|^2_{\ell_2} + \|(W_R \otimes \mathbf{I}_{d_1})^{\frac{1}{2}} Z_{\text{vec}}\|^2_{\ell_2}\right) \\
&= \frac{1}{2}\left[\operatorname{tr}\left((\mathbf{I}_{d_2} \otimes W_L)^{\frac{1}{2}} Z_{\text{vec}} Z_{\text{vec}}^* (\mathbf{I}_{d_2} \otimes W_L)^{\frac{1}{2}}\right) + \operatorname{tr}\left((W_R \otimes \mathbf{I}_{d_1})^{\frac{1}{2}} Z_{\text{vec}} Z_{\text{vec}}^* (W_R \otimes \mathbf{I}_{d_1})^{\frac{1}{2}}\right)\right] \\
&= \frac{1}{2}\left[\operatorname{tr}\left((\mathbf{I}_{d_2} \otimes W_L) Z_{\text{vec}} Z_{\text{vec}}^*\right) + \operatorname{tr}\left((W_R \otimes \mathbf{I}_{d_1}) Z_{\text{vec}} Z_{\text{vec}}^*\right)\right] \\
&= \frac{1}{2}\operatorname{tr}\left(\left[(\mathbf{I}_{d_2} \otimes W_L) + (W_R \otimes \mathbf{I}_{d_1})\right] Z_{\text{vec}} Z_{\text{vec}}^*\right) = \frac{1}{2}\left\|(\mathbf{I}_{d_2} \otimes W_L + W_R \otimes \mathbf{I}_{d_1})^{\frac{1}{2}} Z_{\text{vec}}\right\|^2_{\ell_2} \\
&= \left\|\left(\frac{W_L \oplus W_R}{2}\right)^{\frac{1}{2}} Z_{\text{vec}}\right\|^2_{\ell_2}.
\end{aligned}$$

□

Unfortunately, we make the observation that the introduction of the arithmetic mean weight matrix does not lead to convincing improvements with respect to the one-sided reweighting strategies used in [51, 106]. No particularly notable advantages neither in numerical experiments nor in the theoretical examination of the convergence rate of a corresponding IRLS-method for low-rank matrix recovery can be reported, cf. also subsection 4.4.2 and Remark 4.19.

On the contrary, using as a combination approach the harmonic mean of the weight matrices $\mathbf{I}_{d_2} \otimes W_L$ and $W_R \otimes \mathbf{I}_{d_1}$, i.e., weight matrices constructed as follows $2\left(W_R^{-1} \otimes \mathbf{I}_{d_1} + \mathbf{I}_{d_2} \otimes W_L^{-1}\right)^{-1} = 2\left(W_L^{-1} \oplus W_R^{-1}\right)^{-1} =: W_{(\text{harm})}$ significantly outperforms other weight matrix composition variants in both theoretical as well as practical aspects. These surprising results will be presented in detail in the subsequent sections of this chapter.

With the upcoming lemma we show that the harmonic mean of the weight matrices $\mathbf{I}_{d_2} \otimes W_L$ and $W_R \otimes \mathbf{I}_{d_1}$ can be used for a legitimate reformulation of the p-th power of the Schatten-p quasi-norm.

Lemma 4.3. *Let $0 < p \leq 2$ and $Z \in \mathbb{R}^{d_1 \times d_2}$ with $d = d_1 = d_2$ be a full rank matrix. Then*

$$\|Z\|_{S_p}^p = 2\left\|\left(W_L^{-1} \oplus W_R^{-1}\right)^{-\frac{1}{2}} Z_{\text{vec}}\right\|_{\ell_2}^2 = \|Z_{\text{vec}}\|_{\ell_2(W_{(harm)})}^2,$$

$$\textit{where } 2\left(W_R^{-1} \otimes \mathbf{I}_{d_1} + \mathbf{I}_{d_2} \otimes W_L^{-1}\right)^{-1} = 2\left(W_L^{-1} \oplus W_R^{-1}\right)^{-1} =: W_{(harm)}$$

is the harmonic mean matrix of the symmetric and positive definite weight matrices $\mathbf{I}_{d_2} \otimes W_L$ and $W_R \otimes \mathbf{I}_{d_2}$, $W_L := (ZZ^*)^{\frac{p-2}{2}}$ and $W_R := (Z^*Z)^{\frac{p-2}{2}}$.

Proof. Let $Z = USV^* \in M_{d \times d}$ be the singular value decomposition of Z. Hence for the vectorized version $Z_{\text{vec}} = (V \otimes U)S_{\text{vec}}$ holds true. Using the definitions of W_L and W_R, we express $W_L^{-1} = US^{2-p}U^*$ and $W_R^{-1} = VS^{2-p}V^*$. By the Kronecker sum inversion formula as stated in (4.7), we get $\left(W_L^{-1} \oplus W_R^{-1}\right)^{-1} = (V \otimes U)D(V \otimes U)^*$, where $D \in M_{d_1 d_2 \times d_1 d_2}$ is a diagonal matrix with entries $d_l = (s_i^{2-p} + s_j^{2-p})^{-1} > 0$ for $l = (i-1)d_2 + j$, if $i \in [d_1]$ and $j \in [d_2]$.

Using these facts, we obtain from the orthonormality of the columns of U and V and the particular structure of the diagonal matrix D

$$\begin{aligned}
\|Z_{\text{vec}}\|_{\ell_2(W_{(\text{harm})})}^2 &= \|W_{(\text{harm})}^{\frac{1}{2}} Z_{\text{vec}}\|_{\ell_2}^2 = 2\left\|\left(W_L^{-1} \oplus W_R^{-1}\right)^{-\frac{1}{2}} Z_{\text{vec}}\right\|_{\ell_2}^2 \\
&= 2\text{tr}\left(\left(\left(W_L^{-1} \oplus W_R^{-1}\right)^{-1} Z_{\text{vec}}\right)_{\text{mat}}^* Z\right) \\
&= 2\text{tr}\left([(V \otimes U)D(V \otimes U)^*(V \otimes U)S_{\text{vec}}]^* (V \otimes U)S_{\text{vec}}\right) \\
&= 2\text{tr}\left(S_{\text{vec}}^* D S_{\text{vec}}\right) \\
&= 2\left(\sum_{i=1}^{d} \frac{s_i^2}{2s_i^{2-p}}\right) = \|Z\|_{S_p}^p.
\end{aligned}$$

□

4.2 Harmonic mean iteratively reweighted least squares algorithm

At this point, we are ready to give a formulation of a new type of iteratively reweighted least squares algorithm for the low-rank matrix recovery problem, the so-called *harmonic mean iteratively reweighted least squares algorithm* (HM-IRLS). Similar as in existing variants, we perform the solution of a sequence of weighted least squares problems for the recovery of a low-rank matrix $X_0 \in M_{d_1 \times d_2}$ from only few linear measurements $Y = \Phi(X_0) \in \mathbb{R}^m$. The weight matrices that will be involved in the weighted least squares steps can be interpreted as the harmonic mean of the weight matrices in (4.9) and (4.10).

Let $0 < p \leq 1$ and denote $d = \min(d_1, d_2), D = \max(d_1, d_2)$. Let us now describe our suggested approach as follows: Given a non-increasing sequence of numbers $(\epsilon^{(n)})_{n=1}^{\infty}$, with $\epsilon^{(n)} \geq 0$ for $n \in \mathbb{N}$, we chose an initialization for a symmetric and positive definite weight matrix $\widetilde{W}^{(0)} \in \mathbb{H}_{++}^{d_1 d_2}$. Define recursively for $n = 1, 2, \dots,$ the iterate

$$X^{(n)} = \operatorname*{arg\,min}_{\Phi(X)=Y} \|X_{\text{vec}}\|^2_{\ell_2(\widetilde{W}^{(n-1)})} \tag{4.11}$$

and the weight matrix

$$\widetilde{W}^{(n)} = 2\left[U^{(n)}(\bar{\Sigma}^{(n)}_{d_1})^{2-p}U^{(n)*} \oplus V^{(n)}(\bar{\Sigma}^{(n)}_{d_2})^{2-p}V^{(n)*}\right]^{-1} \tag{4.12}$$

with the diagonal matrices $\bar{\Sigma}^{(n)}_{d_t} \in M_{d_t \times d_t}$ for $d_t = \{d_1, d_2\}$ such that

$$(\bar{\Sigma}^{(n)}_{d_t})_{ii} = \begin{cases} (\sigma_i(X^{(n)})^2 + \epsilon^{(n)2})^{\frac{1}{2}} & \text{if } i \leq d, \\ 0 & \text{if } d < i \leq D \end{cases} \tag{4.13}$$

and the matrices $U^{(n)} \in M_{d_1 \times d_1}$ and $V^{(n)} \in M_{d_2 \times d_2}$, containing the left and right singular vectors of $X^{(n)}$ in its columns, respectively.

Note that the update rule for $\widetilde{W}^{(n)}$ given above can be interpreted as an ϵ-stabilized version of the harmonic mean weight matrix $W_{(\text{harm})}$ in Lemma 4.3. The stabilization factor ϵ is introduced to avoid ill-conditioned instances of $\widetilde{W}^{(n)}$ as soon as some of the singular values of $X^{(n)}$ are getting close to zero and, even beyond that, for the matrices $(X^{(n)}X^{(n)*})^{\frac{2-p}{2}} \oplus (X^{(n)*}X^{(n)})^{\frac{2-p}{2}}$ we already face singularity when $X^{(n)}$ is not of full rank.

To enable the concise formulation of the first step (4.11) of HM-IRLS defining the next iterate, we introduce for $n \in \mathbb{N}$ the linear operator $\widetilde{\mathcal{W}}^{(n)-1} : M_{d_1 \times d_2} \to M_{d_1 \times d_2}$ as

$$(\widetilde{\mathcal{W}}^{(n)-1})(Z) := \frac{1}{2}\left[U^{(n)}(\bar{\Sigma}_{d_1}^{(n)})^{2-p}U^{(n)*}Z + ZV^{(n)}(\bar{\Sigma}_{d_2}^{(n)})^{2-p}V^{(n)*}\right], \tag{4.14}$$

which corresponds to the operation of the inverse of $\widetilde{W}^{(n)}$ on $M_{d_1 \times d_2}$.

We give a pseudo code summary of our formulation of `HM-IRLS` as follows.

Algorithm 4 Harmonic Mean IRLS for low-rank matrix recovery (`HM-IRLS`)

```
Input: A linear map $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m$, image $Y = \Phi(X_0)$ of the ground truth matrix
       $X_0 \in M_{d_1 \times d_2}$, rank parameter $\widetilde{r}$, non-convexity parameter $0 < p \leq 1$.
Output: Sequence $(X^{(n)})_{n=1}^{n_0} \subset M_{d_1 \times d_2}$.
Initialize $n = 0$, $\epsilon^{(0)} = 1$ and $\widetilde{W}^{(0)} = \mathbf{I}_{d_1 d_2} \in M_{d_1 d_2 \times d_1 d_2}$.
repeat
    $X^{(n+1)} = \underset{\Phi(X)=Y}{\arg\min} \|X_{\text{vec}}\|_{\ell_2(\widetilde{W}^{(n)})}^2 = (\widetilde{\mathcal{W}}^{(n)-1} \circ \Phi^* \circ (\Phi \circ \widetilde{\mathcal{W}}^{(n)-1} \circ \Phi^*)^{-1})(Y),$   (4.15)
    $\epsilon^{(n+1)} = \min\left(\epsilon^{(n)}, \sigma_{\widetilde{r}+1}(X^{(n+1)})\right),$   (4.16)
    $\widetilde{W}^{(n+1)} = 2\left[U^{(n+1)}(\bar{\Sigma}_{d_1}^{(n+1)})^{2-p}U^{(n+1)*} \oplus V^{(n+1)}(\bar{\Sigma}_{d_2}^{(n+1)})^{2-p}V^{(n+1)*}\right]^{-1},$   (4.17)
    where $U^{(n+1)} \in M_{d_1 \times d_1} V^{(n+1)} \in M_{d_2 \times d_2}$ are matrices containing the left and right
    singular vectors of $X^{(n+1)}$ in its columns, and $\bar{\Sigma}^{(n+1)}$ is defined as in (4.13).
        $n = n + 1$.
until stopping criterion is met.
Set $n_0 = n$.
```

We note that in practise the explicit calculation of the large weight matrices $\widetilde{W}^{(n+1)} \in \mathbb{H}_{++}^{d_1 d_2}$ (cf. (4.17)) does not have to be performed in an implementation of Algorithm 4. Fortunately, the formulas (4.14) and (4.15) indicate that only the operation of *its inverse* $(\widetilde{W}^{(n+1)})^{-1}$ *resp.* $(\widetilde{W}^{(n)})^{-1}$ has to be executed, which allows the implementation by matrix-matrix multiplications on the space $M_{d_1 \times d_2}$: For matrices $X, Z \in M_{d_1 \times d_2}$, it holds that $\widetilde{W}^{(n)} X_{\text{vec}} = Z_{\text{vec}}$ if and only if $X_{\text{vec}} = (\widetilde{W}^{(n)})^{-1} Z_{\text{vec}}$, which can be reformulated in matrix-matrix operations as follows

$$X = \frac{1}{2}\left[U^{(n)}(\bar{\Sigma}_{d_1}^{(n)})^{2-p}U^{(n)*}Z + ZV^{(n)}(\bar{\Sigma}_{d_2}^{(n)})^{2-p}V^{(n)*}\right].$$

Here the definition of $\widetilde{W}^{(n)}$ (cf. (4.17)) in combination with the Kronecker sum property implies the last equivalence.

In section 4.4.4, we will provide a more extensive discussion on the implementation details.

4.3 Theoretical analysis

Similar to previously discussed analysis approaches for `IRLS`, we again introduce an appropriate auxiliary functional $\mathcal{J}_{HM}$ to obtain a variational interpretation of the algorithmic procedure. For the rest of the section, let $d = \min(d_1, d_2)$ and $D = \max(d_1, d_2)$.

Definition 4.4. Let $0 < p \leq 1$. Given a full-rank matrix $W \in M_{d_1 \times d_2}$, let

$$\widetilde{W}(W) := 2\big[\mathbf{I}_{d_2} \otimes (WW^*)^{\frac{1}{2}}\big] \Big[(WW^*)^{\frac{1}{2}} \oplus (W^*W)^{\frac{1}{2}}\Big]^{-1} \big[(W^*W)^{\frac{1}{2}} \otimes \mathbf{I}_{d_1}\big] \in \mathbb{H}^{d_1 d_2 \times d_1 d_2} \tag{4.18}$$

be the *harmonic mean matrix* $\widetilde{W}$ *associated to* W.

We define the *auxiliary functional* $\mathcal{J}_{HM} : M_{d_1 \times d_2} \times \mathbb{R}_{\geq 0} \times M_{d_1 \times d_2} \to \mathbb{R}_{\geq 0}$ as

$$\begin{aligned} &\mathcal{J}_{HM}(X, \epsilon, W) \\ &:= \begin{cases} \frac{p}{2} \|X_{\mathrm{vec}}\|^2_{\ell_2(\widetilde{W}(W))} + \frac{\epsilon^2 p}{2} \sum\limits_{i=1}^{d} \sigma_i(W) + \frac{2-p}{2} \sum\limits_{i=1}^{d} \sigma_i(W)^{\frac{p}{(p-2)}} & \text{if } \operatorname{rank}(W) = d, \\ +\infty & \text{if } \operatorname{rank}(W) < d. \end{cases} \end{aligned} \tag{4.19}$$

Let us point out that the matrix $\widetilde{W}$ of (4.18) corresponds to forming the harmonic mean of the matrices $\widetilde{W_1} := \mathbf{I}_{d_2} \otimes (WW^*)^{\frac{1}{2}}$ and $\widetilde{W_2} = (W^*W)^{\frac{1}{2}} \otimes \mathbf{I}_{d_1}$, as explained in section 4.1.2, if $(WW^*)^{\frac{1}{2}}$ *and* $(W^*W)^{\frac{1}{2}}$ are positive definite. Please note that in this case, $(WW^*)^{\frac{1}{2}} \oplus (W^*W)^{\frac{1}{2}} = \widetilde{W_1} + \widetilde{W_2}$ is indeed an invertible matrix as $(A^{-1} + B^{-1})^{-1} = A(A+B)^{-1}B$ for any positive definite matrices A, B of the same dimensions,

$$\widetilde{W}(W) = 2\widetilde{W_1}\big(\widetilde{W_1} + \widetilde{W_2}\big)^{-1}\widetilde{W_2} = 2(\widetilde{W_1}^{-1} + \widetilde{W_2}^{-1})^{-1}. \tag{4.20}$$

In the following, we use the more general definition formulated in (4.18) as it is well-defined for any full-rank $W \in M_{d_1 \times d_2}$ and allows handling the case of non-square matrices, i.e., matrices W with $d_1 \neq d_2$, where either $(WW^*)^{\frac{1}{2}}$ or $(W^*W)^{\frac{1}{2}}$ has to be singular. Additionally, by involving the Moore-Penrose pseudo inverse $\widetilde{W_1}^{+}$ and $\widetilde{W_2}^{+}$ of both matrices $\widetilde{W_1}$ and $\widetilde{W_2}$, it is possible to reformulate (4.18) as follows

$$\widetilde{W}(W) = 2\widetilde{W_1}\big(\widetilde{W_1} + \widetilde{W_2}\big)^{-1}\widetilde{W_2} = 2(\widetilde{W_1}^{+} + \widetilde{W_2}^{+})^{-1}.$$

As a next step, we interpret Algorithm 4 as an alternating minimization of the auxiliary functional $\mathcal{J}_{HM}(X, \epsilon, W)$ with respect to its three arguments X, ϵ and W.

In order to do so, we need to justify the update formula (4.17) for the weight matrix

$\widetilde{W}^{(n+1)}$ as the evaluation result of the expression $\widetilde{W}^{(n+1)} = \widetilde{W}\big(W^{(n+1)}\big)$ of $\widetilde{W}$ from Definition 4.4 at the unique minimizer

$$W^{(n+1)} = \underset{W \in M_{d_1 \times d_2}}{\arg\min}\ \mathcal{J}_{HM}(X^{(n+1)}, \epsilon^{(n+1)}, W). \tag{4.21}$$

Furthermore, we need to show that the formula (4.15) can be interpreted as

$$X^{(n+1)} = \underset{\substack{X \in M_{d_1 \times d_2} \\ \Phi(X)=Y}}{\arg\min}\ \|X_{\text{vec}}\|^2_{\ell_2(\widetilde{W}(W^{(n)}))} = \underset{\substack{X \in M_{d_1 \times d_2} \\ \Phi(X)=Y}}{\arg\min}\ \mathcal{J}_{HM}(X, \epsilon^{(n)}, W^{(n)}). \tag{4.22}$$

In the following subsections we will verify these statements.

Remark 4.5. It is important to realize that $\|X_{\text{vec}}\|^2_{\ell_2(\widetilde{W}(W))} \neq \|X\|^2_{F((W^*W)^{1/2})} \neq \|X^*\|^2_{F((WW^*)^{1/2})}$. While the weighted norm on the left hand side involves a symmetrized weight acting in both the column *and* the row space, the norms on the right hand side only use one-sided reweighting in the column *or* row space respectively.

To close this subsection we introduce the ϵ-perturbed Schatten-p-norm of a matrix $X \in M_{d_1 \times d_2}$ as

$$f_\epsilon(X) = \sum_{i=1}^{d} (s_i(X)^2 + \epsilon^2)^{\frac{p}{2}}. \tag{4.23}$$

4.3.1 Optimization of the auxiliary functional with respect to W

Let us fix the matrix $X \in M_{d_1 \times d_2}$ with the corresponding singular value decomposition $X = \sum_{i=1}^{d} s_i u_i v_i^*$, where $u_i \in \mathbb{R}^{d_1}$ $v_i \in \mathbb{R}^{d_2}$ are the left and right singular vectors respectively and $s_i = s_i(X)$ the singular values for $i \in [d]$.

In this subsection, we aim at the justification of formula (4.17) via the building blocks that are used to construct the matrix $\widetilde{W}^{(n+1)}$. More precisely, we consider the minimization problem

$$\underset{W \in M_{d_1 \times d_2}}{\arg\min}\ \mathcal{J}_{HM}(X, \epsilon, W) \tag{4.24}$$

for $\epsilon > 0$.

Lemma 4.6. *The unique minimizer of* (4.24) *is given by*

$$W_{\text{opt}} = \sum_{i=1}^{d} (s_i(X)^2 + \epsilon^2)^{\frac{p-2}{2}} u_i v_i^*.$$

Furthermore, the value of $\mathcal{J}_{HM}$ *at the minimizer* W_{opt} *is*

$$\mathcal{J}_{HM}(X,\epsilon,W_{\text{opt}}) = \sum_{i=1}^{d}(s_i(X)^2+\epsilon^2)^{\frac{p}{2}} = f_\epsilon(X) \tag{4.25}$$

for $p>0$.

Proof. As a first step we introduce the function

$$f_{X,\epsilon}(W) = \mathcal{J}_{HM}(X,\epsilon,W) \\ = \begin{cases} \frac{p}{2}\|X_{\text{vec}}\|^2_{\ell_2(\widetilde{W}(W))} + \frac{\epsilon^2 p}{2}\sum\limits_{i=1}^{d}\sigma_i(W) + \frac{2-p}{2}\sum\limits_{i=1}^{d}\sigma_i(W)^{\frac{p}{(p-2)}} & \text{if } \operatorname{rank}(W)=d, \\ +\infty & \text{if } \operatorname{rank}(W)<d, \end{cases}$$

for $X \in M_{d_1\times d_2}$, $\epsilon>0$ fixed and with $W \in M_{d_1\times d_2}$ as its only argument. Please note that the set of minimizers of the function $f_{X,\epsilon}(W)$ does not contain an instance W with rank smaller than d as at such points $f_{X,\epsilon}(W)$ takes an infinite value. Consequently, we can limit our search for minimizers on the set of rank-d matrices $\Omega = \{Z \in M_{d_1\times d_2} | \operatorname{rank}(Z)=d\}$. We point out that the set Ω is an open set and the following properties of the function $f_{X,\epsilon}(W)$ hold true

(a) $f_{X,\epsilon}(W)$ is lower semicontinuous, which means that any sequence $(W^k)_{k\in\mathbb{N}}$ with $W^k \overset{k\to\infty}{\longrightarrow} W$ fulfills $\liminf\limits_{k\to\infty} f^p_{X,\epsilon}(W^k) \geq f_{X,\epsilon}(W)$,

(b) $f_{X,\epsilon}(W) \geq \alpha$ for all $W \in M_{d_1\times d_2}$ for some constant α,

(c) $f_{X,\epsilon}(W)$ is coercive, i.e., for any sequence $(W^k)_{k\in\mathbb{N}}$ with $\|W^k\|_F \overset{k\to\infty}{\longrightarrow} \infty$, we have $f^p_{X,\epsilon}(W^k) \overset{k\to\infty}{\longrightarrow} \infty$.

We verify the statements above: The function $f_{X,\epsilon}(W)|_\Omega$ is a concatenation of an indicator function of an open set, which is lower-semicontinuous and a sum of continuous functions on Ω and hence property (a) is true. Obviously property (b) is true for the parameter choice $\alpha=0$.

As a justification for (c), we point out that $f_{X,\epsilon}(W) > \frac{\epsilon^2 p}{2}\sum\limits_{i=1}^{d}\sigma_i(W) = \frac{\epsilon^2 p}{2}\|W\|_{S_1} \geq \frac{\epsilon^2 p}{2}\|W\|_F$ implies coercivity as directly following from its definition. We can conclude from (a) and (c), that the level sets as introduced in (2.2) $\ell_{f_{X,\epsilon}(W),\Omega}(c) = \{W \in M_{d_1\times d_2} | f_{X,\epsilon}(W) \leq c\}$ are closed and bounded and, hence, compact.

Via the direct method of calculus of variations as stated in Theorem 2.5, we derive from the validity of the properties (a) - (c) that $f_{X,\epsilon}(W)$ has at least one global minimizer,

which is contained in the set of critical points of $f_{X,\epsilon}(W)$ [34, Theorem 1].
As a next step, we want to find a characterization of the set of critical points of $f_{X,\epsilon}(W)$, by explicitly calculating its derivative with respect to W and equating the result with zero.

Let us without loss of generality consider the case $d = d_1 = d_2$ and define the set

$$\Omega = \{W \in M_{d\times d} \text{ s.t. } \operatorname{rank}(W) = d\}\,.$$

We already mentioned in (4.20), we can rewrite the harmonic mean matrix $\widetilde{W}(W)$ in the form

$$\widetilde{W}(W) = 2\widetilde{W}_1\big(\widetilde{W}_1 + \widetilde{W}_2\big)^{-1}\widetilde{W}_2 = 2(\widetilde{W}_1^{-1} + \widetilde{W}_2^{-1})^{-1}$$

for $W \in \Omega$ with the definitions $\widetilde{W}_1 := \mathbf{I}_d \otimes (WW^*)^{\frac{1}{2}}$ and $\widetilde{W}_2 = (W^*W)^{\frac{1}{2}} \otimes \mathbf{I}_d$. For $W \in \Omega$, we reformulate the auxiliary functional such that

$$\begin{aligned} f_{X,\epsilon}(W) = \mathcal{J}_{HM}(X,\epsilon,W) &= \frac{p}{2}\|X_{\text{vec}}\|^2_{\ell_2(\widetilde{W}(W))} + \frac{\epsilon^2 p}{2}\sum_{i=1}^{d}\sigma_i(W) + \frac{2-p}{2}\sum_{i=1}^{d}\sigma_i(W)^{\frac{p}{(p-2)}} \\ &= \frac{p}{2}\|X_{\text{vec}}\|^2_{\ell_2(\widetilde{W}(W))} + \frac{\epsilon^2 p}{2}\|(W^*W)^{1/2}\|_F^2 + \frac{2-p}{2}\|(W^*W)^{\frac{p}{2(p-2)}}\|_F^2. \end{aligned}$$

Now we aim at the identification of the set of critical points of $f_{X,\epsilon}(W)$ located in Ω and compute its derivative with respect to W using the derivative rules (7), (12), (13), (15), (16), (18), (20) in Chapter 8.2 and Theorem 3 in Chapter 8.4 of [101]. Using the notation of [101], we calculate

$$\begin{aligned} \frac{\partial f_{X,\epsilon}(W)}{\partial W} =& -\frac{p}{2}\operatorname{tr}\left(X_{\text{vec}}^*\widetilde{W}\frac{\partial \widetilde{W}^{-1}}{\partial W}\widetilde{W}X_{\text{vec}}\right) \\ &+ \frac{p\epsilon^2}{4}\left(\operatorname{tr}\left(W(W^*W)^{-\frac{1}{2}}\partial W^*\right) + \operatorname{tr}((W^*W)^{-\frac{1}{2}}W^*\partial W)\right) \\ &- \frac{p}{4}\left(\operatorname{tr}\left(W(W^*W)^{\frac{4-p}{2(p-2)}}\partial W^*\right) + \operatorname{tr}((W^*W)^{\frac{4-p}{2(p-2)}}W^*\partial W)\right) \end{aligned}$$

where

$$\begin{aligned} \frac{\partial\widetilde{W}^{-1}}{\partial W} &= \frac{1}{2}\frac{\partial\left[(WW^*)^{-\frac{1}{2}} \oplus (W^*W)^{-\frac{1}{2}}\right]}{\partial W} \\ &= -\frac{1}{4}\left[\left((W^*W)^{-\frac{3}{2}}W^*\partial W + \partial W^* W(W^*W)^{-\frac{3}{2}}\right)\otimes \mathbf{I}_{d_1}\right] \\ &- \frac{1}{4}\left[\mathbf{I}_{d_2}\otimes\left(\partial W(WW^*)^{-\frac{3}{2}}W^* + (WW^*)^{-\frac{3}{2}}W\partial W^*\right)\right]. \end{aligned} \tag{4.26}$$

Reformulating the first term as follows using the cyclicity of the trace, gives

$$
\begin{aligned}
-\frac{p}{2}\mathrm{tr}\left(X_{\mathrm{vec}}^{*}\widetilde{W}\frac{\partial\widetilde{W}^{-1}}{\partial W}\widetilde{W}X_{\mathrm{vec}}\right) = \frac{p}{8}\Big[&\mathrm{tr}\left((\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}^{*}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}(W^{*}W)^{-\frac{3}{2}}W^{*}\partial W\right)\\
&+\mathrm{tr}\left(W(W^{*}W)^{-\frac{3}{2}}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}^{*}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}\partial W^{*}\right)\\
&+\mathrm{tr}\left(W^{*}(WW^{*})^{-\frac{3}{2}}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}^{*}\partial W\right)\\
&+\mathrm{tr}\left((\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}^{*}(WW^{*})^{-\frac{3}{2}}W\partial W^{*}\right)\Big].
\end{aligned}
$$

Summarizing the calculations above, we obtain

$$
\begin{aligned}
\frac{\partial f_{X,\epsilon}(W)}{\partial W} = &\frac{p}{8}\Big[\mathrm{tr}\left((\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}^{*}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}(W^{*}W)^{-\frac{3}{2}}W^{*}\partial W\right)\\
&+\mathrm{tr}\left(W(W^{*}W)^{-\frac{3}{2}}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}^{*}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}\partial W^{*}\right)\\
&+\mathrm{tr}\left(W^{*}(WW^{*})^{-\frac{3}{2}}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}^{*}\partial W\right)\\
&+\mathrm{tr}\left((\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}^{*}(WW^{*})^{-\frac{3}{2}}W\partial W^{*}\right)\Big]\\
&+\frac{p\epsilon^{2}}{4}\left(\mathrm{tr}\left(W(W^{*}W)^{-\frac{1}{2}}\partial W^{*}\right)+\mathrm{tr}((W^{*}W)^{-\frac{1}{2}}W^{*}\partial W)\right)\\
&-\frac{p}{4}\left(\mathrm{tr}\left(W(W^{*}W)^{\frac{4-p}{2(p-2)}}\partial W^{*}\right)+\mathrm{tr}((W^{*}W)^{\frac{4-p}{2(p-2)}}W^{*}\partial W)\right).
\end{aligned}
$$

In order to find the critical points of $f_{X,\epsilon}(W)$, the terms above are rearranged, and we equate the derivative with zero, which yields

$$
\begin{aligned}
\frac{\partial f_{X,\epsilon}(W)}{\partial W} = &\frac{p}{8}\mathrm{tr}\Big(\Big[(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}^{*}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}(W^{*}W)^{-\frac{3}{2}}W^{*}\\
&\quad+W^{*}(WW^{*})^{-\frac{3}{2}}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}^{*}\\
&\quad+2\epsilon^{2}(W^{*}W)^{-\frac{1}{2}}W^{*}-2(W^{*}W)^{\frac{4-p}{2(p-2)}}W^{*}\Big]\partial W\Big)\\
&\frac{p}{8}\mathrm{tr}\Big(\Big[W(W^{*}W)^{-\frac{3}{2}}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}^{*}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}\\
&\quad+(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}(\widetilde{W}X_{\mathrm{vec}})_{\mathrm{mat}}^{*}(WW^{*})^{-\frac{3}{2}}W\\
&\quad+2\epsilon^{2}W(W^{*}W)^{-\frac{1}{2}}-2W(W^{*}W)^{\frac{4-p}{2(p-2)}}\Big]\partial W^{*}\Big)\\
:= &\frac{p}{8}\mathrm{tr}\left(A\partial W\right)+\frac{p}{8}\mathrm{tr}\left(A^{*}\partial W^{*}\right)\\
= &\frac{p}{8}\mathrm{tr}\left((A\oplus A)\partial W\right)\\
= &\,0,
\end{aligned}
$$

where

$$A = \Big[(\widetilde{W X}_{\text{vec}})_{\text{mat}}^{*}(\widetilde{W X}_{\text{vec}})_{\text{mat}}(W^*W)^{-\frac{3}{2}}W^* + W^*(WW^*)^{-\frac{3}{2}}(\widetilde{W X}_{\text{vec}})_{\text{mat}}(\widetilde{W X}_{\text{vec}})_{\text{mat}}^{*}$$
$$+2\epsilon^2(W^*W)^{-\frac{1}{2}}W^* - 2(W^*W)^{\frac{4-p}{2(p-2)}}W^*\Big]. \tag{4.27}$$

As a next step, we need to find W such that $A \oplus A = 0$. This has the implication that all eigenvalues of $A\oplus A = A\otimes \mathbf{I}_d + \mathbf{I}_d \otimes A$ have to be zero. We note at this point that the eigenvalues of the Kronecker sum matrix of two matrices A_1 and A_2 with eigenvalues λ_s and μ_t with $s, t \in [d]$ correspond to the sum of the eigenvalues $\lambda_s + \mu_t$. As we consider the case $A = A_1 = A_2$, it follows that all eigenvalues of A itself have to be zero, which is only possible if A is identical to the zero matrix.

Let $W = U\Sigma V^* \in M_{d\times d}$ with $U \in U_d, V \in U_d, \Sigma \in M_{d\times d}$, where $\Sigma = \operatorname{diag}(\sigma)$ is a diagonal matrix with *ascending* entries. Define the matrix $H = H_{i,j} = \frac{2}{\sigma_i^{-1}+\sigma_j^{-1}}$ for $i = 1, \dots, d, j = 1, \dots, d$, which corresponds to the result of reshaping the diagonal of the $d^2 \times d^2$-matrix $2(\Sigma\oplus\Sigma)$ into a $d\times d$-matrix. Note that we can express $(\widetilde{W X}_{\text{vec}})_{\text{mat}} = U\big(H \circ (U^*XV)\big)V^*$ using (4.8) and denote $B := H \circ (U^*XV)$.

Next we plug the decomposition $W = U\Sigma V^*$ into the equation (4.27) and calculate

$$\begin{aligned}
\text{A} = 0 \Leftrightarrow\ & (UBV^*)^*(UBV^*)(V\Sigma^2V^*)^{-3/2}(U\Sigma V^*)^* \\
& + (U\Sigma V^*)^*(U\Sigma^2U^*)^*)^{-3/2}(UBV^*)(UBV^*)^* \\
& + 2\epsilon^2(V\Sigma^2V^*)^{-1/2}(U\Sigma V^*)^* - 2(V\Sigma^2V^*)^{\frac{4-p}{2(p-2)}}(U\Sigma V^*)^* = 0 \qquad (4.28)\\
\Leftrightarrow\ & VB^*B\Sigma^{-2}U^* + V\Sigma^{-2}BB^*U^* + 2\epsilon^2V\mathbf{I}_dU^* - 2V\Sigma^{\frac{2}{p-2}}U^* = 0 \\
\Leftrightarrow\ & B^*B\Sigma^{-2} + \Sigma^{-2}BB^* + 2\epsilon^2\mathbf{I}_d - 2\Sigma^{\frac{2}{p-2}} = 0.
\end{aligned}$$

Noting that $2\epsilon^2\mathbf{I}_d - 2\Sigma^{\frac{2}{p-2}}$ is diagonal, it follows that also $B^*B\Sigma^{-2} + \Sigma^{-2}BB^*$ is diagonal. Moreover, observe that also the sum of matrices $B^*B + \Sigma^{-2}BB^*\Sigma^2$ is diagonal matrix as well, with a symmetric first summand B^*B. As the sum or difference of symmetric matrices is again symmetric, it follows that also the second summand $\Sigma^{-2}BB^*\Sigma^2$ is symmetric, i.e., $\Sigma^{-2}BB^*\Sigma^2 = (\Sigma^{-2}BB^*\Sigma^2)^* = \Sigma^2BB^*\Sigma^{-2}$. This implies that it holds that $BB^*\Sigma^4 = \Sigma^4BB^*$ and, as a consequence, Σ^4 and BB^* commute.

This is only the case if either Σ is a multiple of the identity or if the matrix BB^* is diagonal. Assuming the first case, we conclude from (4.28) that also BB^* and B^*B are a multiple of the identity. Hence, the first case, where Σ is assumed to be a multiple of the identity, is only a special case of the second possible scenario, where BB^* is a diagonal matrix. Therefore, it is sufficient to limit further considerations to the more general second case. (Considerations for B^*B can be carried out analogously.)

Diagonality of BB^* only occurs if B is either orthonormal or diagonal. The first possibility, orthonormality, leads to contradictions with the equations in (4.28). Thus the matrix $B = H \circ (U^*XV)$ has to be diagonal.

Let now be $X = \bar{U}\bar{S}\bar{V}^*$ the singular value decomposition of X. From the fact that H has no zero entries due to the full rank of W, we conclude the diagonality of $U^*\bar{U}\bar{S}\bar{V}^*V$. Consequently, U and V have to be chosen such that $P = [U^*\bar{U}]_{d\times d}$ and $P^* = [\bar{V}^*V]_{d\times d}$ for a permutation matrix $P \in U_d$. We denote the reshuffled indexing corresponding to the permutation P by $p(i) \in [d]$ for $i \in [d]$. Remembering that $H_{ii} = \sigma_i$ for $i \in [d]$, we get

$$\begin{aligned}
&(H\circ(P\bar{S}P^*))^*(H\circ(P\bar{S}P^*))\Sigma^{-2} + \Sigma^{-2}(H\circ(P\bar{S}P^*))(H\circ(P\bar{S}P^*))^* + 2\epsilon^2\mathbf{I}_d - 2\Sigma^{\frac{2}{p-2}} = 0\\
&\Leftrightarrow \quad 2\bar{s}_{p(i)}^2 + 2\epsilon^2 = 2\sigma_i^{\frac{2}{p-2}} \text{ for all } i \in [d]\\
&\Leftrightarrow \quad \sigma_i = (\bar{s}_{p(i)}^2 + \epsilon^2)^{\frac{p-2}{2}} \text{ for all } i \in [d].
\end{aligned}$$

Using the assumption that the diagonal of Σ has ascending entries and the diagonal of $\bar{S}$ has descending entries, we can conclude that the permutation matrix P coincides with the identity matrix. We infer from $P = \mathbf{I}_d$, that $U = \bar{U}$ and $V = \bar{V}$ and hence also $\Sigma = (\bar{S}^2 + \epsilon^2\mathbf{I}_d)^{\frac{p-2}{2}}$.

We can now summarize our detailed calculations above with the statement that

$$W_{\text{opt}} = \bar{U}\Sigma\bar{V}^* = \bar{U}(\bar{S}^2 + \epsilon^2\mathbf{I}_d)^{\frac{p-2}{2}}\bar{V}^*$$

is the only critical point of $f_{X,\epsilon}^p$ on the domain Ω.

We point out that the deduced results extend for the case $d_1 \neq d_2$, where the definition of $\widetilde{W}(W)$ is modified by the introduction of the Moore-Penrose pseudo inverse of $(WW^*)^{1/2}$

$$\widetilde{W}(W) = 2\widetilde{W}_1\big(\widetilde{W}_1 + \widetilde{W}_2\big)^{-1}\widetilde{W}_2 = 2(\widetilde{W}_1^{+} + \widetilde{W}_2^{-1})^{-1}.$$

In Theorem 5 in Chapter 8.4 of [101], one can find the corresponding derivative rule for the calculation in (4.26).

We close this part of the proof with stating that the only critical point and consequently the unique global minimizer of $f_{X,\epsilon}^p(W)$ is

$$W_{\text{opt}} = \sum_{i=1}^{d}(s_i^2 + \epsilon^2)^{\frac{p-2}{2}}u_iv_i^* =: \sum_{i=1}^{d}\sigma_iu_iv_i^*.$$

In a next step, we define the matrices $W_{\text{opt}}^L := \sum_{i=1}^d \sigma_i u_i u_i^*$ and $W_{\text{opt}}^R := \sum_{i=1}^d \sigma_i v_i v_i^*$, and note that

$$\widetilde{W}(W_{\text{opt}}) = 2(W_{\text{opt}}^{R^{-1}} \oplus W_{\text{opt}}^{L^{-1}})^{-1}$$

with Definition 4.4.

For the verification of the second part of the theorem, the optimal solution W_{opt} is plugged into the functional $\mathcal{J}_{HM}$ and we calculate using (4.7)

$$\begin{aligned}
\mathcal{J}_{HM}(X,\epsilon,W_{\text{opt}}) &= \frac{p}{2}\|X_{\text{vec}}\|^2_{\ell_2(\widetilde{W}(W_{\text{opt}}))} + \frac{\epsilon^2 p}{2}\sum_{i=1}^d \sigma_i(W_{\text{opt}}) + \frac{2-p}{2}\sum_{i=1}^d \sigma_i(W_{\text{opt}})^{\frac{p}{p-2}} \\
&= \frac{p}{2}\sum_{i=1}^d \left[s_i^2 (u_i^* \otimes v_i^*) 2 \left(\sum_{k=1}^{d_2}\sum_{j=1}^{d_1} \frac{u_k u_k^* \otimes v_j v_j^*}{\sigma_k^{-1} + \sigma_j^{-1}} \right) (u_i \otimes v_i) \right]_{ii} \\
&+ \frac{\epsilon^2 p}{2}\sum_{i=1}^d \sigma_i + \frac{2-p}{2}\sum_{i=1}^d \sigma_i^{\frac{p}{p-2}} \\
&= \frac{p}{2}\sum_{i=1}^d \left[2 s_i^2 \left(\sum_{k=1}^{d_2}\sum_{j=1}^{d_1} \frac{u_i^* u_k u_k^* u_i \otimes v_i^* v_j v_j^* v_i}{\sigma_k^{-1} + \sigma_j^{-1}} \right) \right]_{ii} \\
&+ \frac{\epsilon^2 p}{2}\sum_{i=1}^d \sigma_i + \frac{2-p}{2}\sum_{i=1}^d \sigma_i^{\frac{p}{p-2}} \\
&= \frac{p}{2}\sum_{i=1}^d (s_i^2 + \epsilon^2)\sigma_i + \frac{2-p}{2}\sum_{i=1}^d \sigma_i^{\frac{p}{p-2}} \\
&= \frac{p}{2}\sum_{i=1}^d (s_i^2 + \epsilon^2)(s_i^2 + \epsilon^2)^{\frac{p-2}{2}} + \frac{2-p}{2}\sum_{i=1}^d (s_i^2 + \epsilon^2)^{\frac{p}{2}} \\
&= \sum_{i=1}^d (s_i^2 + \epsilon^2)^{\frac{p}{2}}.
\end{aligned}$$

□

4.3.2 Optimization of the auxiliary functional with respect to X

Now we continue with the proof of the fact that the definition rule (4.15) of $X^{(n+1)}$ as used the first step of Algorithm 4 can be interpreted as the minimization of the auxiliary functional $\mathcal{J}_{HM}$ with respect to the variable X. Moreover, we provide arguments that this minimization step can be executed via the solution of a weighted least squares problem with weight matrix $\widetilde{W}^{(n)}$.

Lemma 4.7. *Let $0 < p \leq 1$. Given a full-rank matrix $W \in M_{d_1 \times d_2}$, let $\widetilde{W}(W) := 2([(WW^*)^{\frac{1}{2}}]^+ \oplus [(W^*W)^{\frac{1}{2}}]^+)^{-1} \in H_{d_1 d_2 \times d_1 d_2}$ be the matrix from Definition 4.4 and*

$\mathcal{W}^{-1}: M_{d_1\times d_2} \to M_{d_1\times d_2}$ the linear operator of its inverse

$$\mathcal{W}^{-1}(Z) := \frac{1}{2}\left[[(WW^*)^{\frac{1}{2}}]^+ Z + Z[(W^*W)^{\frac{1}{2}}]^+\right].$$

Then the matrix

$$X_{\text{opt}} = \big(\mathcal{W}^{-1} \circ \Phi^* \circ (\Phi \circ \mathcal{W}^{-1} \circ \Phi^*)^{-1}\big)\big(Y\big) \in M_{d_1\times d_2}$$

is the unique minimizer of the optimization problems

$$\underset{\Phi(X)=Y}{\arg\min}\ \mathcal{J}_{HM}(X,\epsilon,W) = \underset{\Phi(X)=Y}{\arg\min}\ \|X_{\text{vec}}\|^2_{\ell_2(\widetilde{W})}. \tag{4.29}$$

Moreover, a matrix $X_{\text{opt}} \in M_{d_1\times d_2}$ is a minimizer of the minimization problem (4.29) *if and only if it fulfills the property*

$$\langle \widetilde{W}(W)(X_{\text{opt}})_{\text{vec}}, H_{\text{vec}}\rangle_{\ell_2} = 0 \ \textit{ for all } \ H \in \mathcal{N}(\Phi) \quad \textit{and} \quad \Phi(X_{\text{opt}}) = Y. \tag{4.30}$$

Proof. We start with noting that the equality of the optimization problems (4.29) follows from the fact that only the first summand of the functional $\mathcal{J}_{HM}(X,\epsilon,W)$ depends on X.

Next we see that $\widetilde{W}(W) = 2([(W^*W)^{\frac{1}{2}}]^+ \oplus [(WW^*)^{\frac{1}{2}}]^+)^{-1}$ is positive definite: Let $W = \sum_{i=1}^d \sigma_i u_i v_i^*$, where u_i, v_i for $i \in [d]$ are the left and right singular vector respectively and σ_i for $i \in [d]$ are the singular values of W. Since $W^*W = \sum_{i=1}^d \sigma_i^2 v_i v_i^* \succeq 0$, also for the generalized inverse root holds that $[(WW^*)^{\frac{1}{2}}]^+ \succeq 0$ and for $WW^* = \sum_{i=1}^d \sigma_i^2 u_i u_i^* \succeq 0$, it follows that $[(WW^*)^{\frac{1}{2}}]^+ \succeq 0$. As already mentioned above, at least one of the matrices $(WW^*)^{\frac{1}{2}}$ and $(W^*W)^{\frac{1}{2}}$ is positive definite and consequently, $\frac{1}{2}[(W^*W)^{\frac{1}{2}}]^+ \oplus [(WW^*)^{\frac{1}{2}}]^+ = \frac{1}{2}[(W^*W)^{\frac{1}{2}}]^+ \otimes \mathbf{I}_{d_1} + \mathbf{I}_{d_2} \otimes [(WW^*)^{\frac{1}{2}}]^+ \succ 0$. We conclude that also for its inverse holds $\widetilde{W}(W) \succ 0$.

Using the fact that $\widetilde{W}(W) \succ 0$, we can show analogously to the standard equivalence of [54, Proposition A.23] that a matrix $\bar{X} \in M_{d_1\times d_2}$ is a minimizer of $\|X_{\text{vec}}\|^2_{\ell_2(\widetilde{W}(W))} = \langle \widetilde{W}(W)X_{\text{vec}}, X_{\text{vec}}\rangle_{\ell_2}$ under the linear constraint $\Phi(X) = Y$ if and only if

$$\langle \widetilde{W}(W)\bar{X}_{\text{vec}}, \psi_{\text{vec}}\rangle_{\ell_2} = 0 \ \text{ for all } \ \psi \in \mathcal{N}(\Phi) \quad \text{and} \quad \Phi(\bar{X}) = Y.$$

Moreover, the latter condition holds if and only if $X_{\text{opt}} = \big(\mathcal{W}^{-1} \circ \Phi^* \circ (\Phi \circ \mathcal{W}^{-1} \circ \Phi^*)^{-1}\big)\big(Y\big) \in M_{d_1\times d_2}$. Indeed, the property (4.30) is equivalent to the existence of vector $\lambda \in \mathbb{R}^m$, for which holds $\Phi^*(\lambda)_{\text{vec}} = \widetilde{W}(W)(X_{\text{opt}})_{\text{vec}}$. Using the definition of $\widetilde{W}(W)^{-1}$, we observe that $\widetilde{W}(W)^{-1} = \frac{1}{2}([(WW^*)^{\frac{1}{2}}]^+ \oplus [(W^*W)^{\frac{1}{2}}]^+)$ and we conclude that

$$
\begin{aligned}
X_{\mathrm{opt}} &= \left[\widetilde{W}(W)^{-1}(\Phi^*(\lambda)_{\mathrm{vec}})\right]_{\mathrm{mat}} = \frac{1}{2}\left([(WW^*)^{1/2}]^+\Phi^*(\lambda) + \Phi^*(\lambda)[(W^*W)^{1/2}]^+\right) \\
&= (\mathcal{W}^{-1} \circ \Phi^*)(\lambda).
\end{aligned}
\tag{4.31}
$$

As $Y = \Phi(X_{\mathrm{opt}}) = (\Phi \circ \mathcal{W}^{-1} \circ \Phi^*)(\lambda)$ in this case, we compute $\lambda = (\Phi \circ \mathcal{W}^{-1} \circ \Phi^*)^{-1}(Y)$ and we yield from (4.31) that $X_{\mathrm{opt}} = (\mathcal{W}^{-1} \circ \Phi^* \circ (\Phi \circ \mathcal{W}^{-1} \circ \Phi^*)^{-1})(Y)$. On the other hand, any matrix defined as $X_{\mathrm{opt}} = (\mathcal{W}^{-1} \circ \Phi^* \circ (\Phi \circ \mathcal{W}^{-1} \circ \Phi^*)^{-1})(Y)$ fulfills (4.30) by construction. Hence, we can conclude this proof by pointing out that X_{opt} fulfills the condition (4.30).

□

4.3.3 Basic properties of the algorithm

In the course of this subsection, we have a closer look at Algorithm 4 and examine some of its properties, that will be useful for developing the proof of convergence and to determine the rate of convergence later on. In particular, we show the boundedness of the sequence of iterates $(X^{(n)})_{n\in\mathbb{N}}$ and the fact that as $n \to \infty$ two successive iterates get arbitrarily close.

We start with a collection of properties of the functional $\mathcal{J}_{HM}$ that appear in a similar fashion already in the existing `IRLS`-literature.

Lemma 4.8. *Let $(X^{(n)}, \epsilon^{(n)})_{n\in\mathbb{N}}$ be the sequence of iterates and smoothing parameters of Algorithm 4. Let $X^{(n)} = \sum_{i=1}^{d} \sigma_i^{(n)} u_i^{(n)} v_i^{(n)*}$ be the SVD of the n-th iterate $X^{(n)}$. Let $(W^{(n)})_{n\in\mathbb{N}}$ be a corresponding sequence such that*

$$
W^{(n)} = \sum_{i=1}^{d} (\sigma_i^{(n)2} + \epsilon^{(n)2})^{\frac{p-2}{2}} u_i^{(n)} v_i^{(n)*}
$$

for $n \in \mathbb{N}$. Then the following properties hold:

(a) $\mathcal{J}_{HM}(X^{(n)}, \epsilon^{(n)}, W^{(n)}) \geq \mathcal{J}_{HM}(X^{(n+1)}, \epsilon^{(n+1)}, W^{(n+1)})$ for all $n \geq 1$,

(b) $\|X^{(n)}\|_{S_p}^p \leq \mathcal{J}_{HM}(X^{(1)}, \epsilon^{(0)}, W^{(0)}) =: \mathcal{J}_{p,0}$ for all $n \geq 1$,

(c) The iterates $X^{(n)}, X^{(n+1)}$ come arbitrarily close as $n \to \infty$, i.e., $\lim_{n\to\infty} \|(X^{(n)} - X^{(n+1)})_{\mathrm{vec}}\|_{\ell_2}^2 = 0$.

Proof. (a) Using the minimization property that defines $X^{(n+1)}$ in (4.22) together with the inequality $\epsilon^{(n+1)} \leq \epsilon^{(n)}$, we obtain

$$\mathcal{J}_{HM}(X^{(n)}, \epsilon^{(n)}, W^{(n)}) \geq \mathcal{J}_{HM}(X^{(n+1)}, \epsilon^{(n)}, W^{(n)}) \geq \mathcal{J}_{HM}(X^{(n+1)}, \epsilon^{(n+1)}, W^{(n)})$$

Combining this with the minimization property that defines $W^{(n+1)}$ in (4.21) and the results in Lemma 4.6, we have

$$\mathcal{J}_{HM}(X^{(n+1)}, \epsilon^{(n+1)}, W^{(n)}) \geq \mathcal{J}_{HM}(X^{(n+1)}, \epsilon^{(n+1)}, W^{(n+1)}),$$

which finishes the proof of (a).

(b) For all $n \in \mathbb{N}$, it holds that

$$\|X^{(n)}\|_{S_p}^p \leq g_{\epsilon^{(n)}}^p(X^{(n)}) = \mathcal{J}_{HM}(X^{(n)}, \epsilon^{(n)}, W^{(n)}) \leq \mathcal{J}_{HM}(X^{(1)}, \epsilon^{(0)}, W^{(0)}),$$

where we used Lemma 4.6 as well as the monotonicity property shown in (a).

(c) With property (a) and Definition 4.4 we obtain for each $n \in \mathbb{N}$

$$\begin{aligned}
&\frac{2}{p}\left[\mathcal{J}_{HM}(X^{(n)}, \epsilon^{(n)}, W^{(n)}) - \mathcal{J}_{HM}(X^{(n+1)}, \epsilon^{(n+1)}, W^{(n+1)})\right] \\
&\geq \frac{2}{p}\left[\mathcal{J}_{HM}(X^{(n)}, \epsilon^{(n)}, W^{(n)}) - \mathcal{J}_{HM}(X^{(n+1)}, \epsilon^{(n)}, W^{(n)}))\right] \\
&= \|X_{\text{vec}}^{(n)}\|_{\ell_2\left(\widetilde{W}(W^{(n)})\right)}^2 - \|X_{\text{vec}}^{(n+1)}\|_{\ell_2\left(\widetilde{W}(W^{(n)})\right)}^2 \\
&= \langle (X^{(n)} + X^{(n+1)})_{\text{vec}}, (X^{(n)} - X^{(n+1)})_{\text{vec}} \rangle_{\ell_2(\widetilde{W}^{(n)})},
\end{aligned}$$

with the notation $\widetilde{W}^{(n)} := \widetilde{W}(W^{(n)})$ in the last equality. Using the facts that $X^{(n+1)}$ is the minimizer of $\|X_{\text{vec}}\|_{\ell_2(\widetilde{W}^{(n)})}^2$ under the linear constraint and that $X^{(n)} - X^{(n+1)} \in \mathcal{N}(\Phi)$, we conclude together with (4.30) that

$$\langle \widetilde{W}^{(n)} X_{\text{vec}}^{(n+1)}, (X^{(n)} - X^{(n+1)})_{\text{vec}} \rangle = 0.$$

Therefore, we get

$$\begin{aligned}
&\langle (X^{(n)} + X^{(n+1)})_{\text{vec}}, (X^{(n)} - X^{(n+1)})_{\text{vec}} \rangle_{\ell_2(\widetilde{W}^{(n)})} \\
&= \langle (X^{(n)} - X^{(n+1)})_{\text{vec}}, (X^{(n)} - X^{(n+1)})_{\text{vec}} \rangle_{\ell_2(\widetilde{W}^{(n)})} \\
&= \|(X^{(n)} - X^{(n+1)})_{\text{vec}}\|_{\ell_2(\widetilde{W}^{(n)})}^2.
\end{aligned}$$

As a next step, we want to estimate $\sigma_{d_1 d_2}(\widetilde{W}^{(n)})$ to derive a bound on the difference of iterates independent of the involved weighting matrix in the expressions above. We remind the reader that $1 = \sigma_{d_1d_2}(Z)\sigma_1(Z^{-1})$ for any invertible matrix $Z \in M_{d_1d_2\times d_1d_2}$ and, hence, it remains to compute $\sigma_1\big((\widetilde{W}^{(n)})^{-1}\big)$ to gain sufficient information on $\sigma_{d_1d_2}\big(\widetilde{W}^{(n)}\big)$.

As discussed in [8, Proposition 7.2.3], the spectrum of a Kronecker sum matrix $A \oplus B$ consists of the pairwise sum of the spectra of the individual matrices A and B. We use this to compute,

$$\begin{aligned}\sigma_1^p\big((\widetilde{W}^{(n)})^{-1}\big) &= \left[\frac{1}{2}\Big((\sigma_1^{(n)2}+\epsilon^{(n)2})^{\frac{2-p}{2}} + (\sigma_1^{(n)2}+\epsilon^{(n)2})^{\frac{2-p}{2}}\Big)\right]^p = \big(\sigma_1^{(n)2}+\epsilon^{(n)2}\big)^{\frac{p}{2}(2-p)} \\ &\leq \big(f^p_{\epsilon^{(n)}}(X^{(n)})\big)^{2-p} = \big(\mathcal{J}_{HM}(X^{(n)},\epsilon^{(n)},W^{(n)})\big)^{2-p} \leq \big(\mathcal{J}_{HM}(X^{(1)},\epsilon^{(0)},W^{(0)})\big)^{2-p},\end{aligned}$$

employing Lemma 4.6 and the monotonicity of $\mathcal{J}_{HM}$.

Therefore, it follows that

$$\sigma_{\min}(\widetilde{W}^{(n)}) = \sigma_{d_1d_2}(\widetilde{W}^{(n)}) \geq \big(\mathcal{J}_{HM}(X^{(1)},\epsilon^{(0)},W^{(0)})\big)^{1-\frac{2}{p}} = \mathcal{J}_{p,0}^{1-\frac{2}{p}}$$

and we combine this result with the previous calculations yielding

$$\begin{aligned}&\frac{2}{p}\left[\mathcal{J}_{HM}(X^{(n)},\epsilon^{(n)},W^{(n)}) - \mathcal{J}_{HM}(X^{(n+1)},\epsilon^{(n+1)},W^{(n+1)})\right] = \frac{2}{p}\|(X^{(n)}-X^{(n+1)})_{\text{vec}}\|^2_{\ell_2(\widetilde{W}^{(n)})} \\ &\geq \frac{2}{p}\sigma_{\min}(\widetilde{W}^{(n)})\|(X^{(n)}-X^{(n+1)})_{\text{vec}}\|^2_{\ell_2} \geq \frac{2}{p}\mathcal{J}_{p,0}^{1-\frac{2}{p}}\|(X^{(n)}-X^{(n+1)})_{\text{vec}}\|^2_{\ell_2} \\ &= C_p\|(X^{(n)}-X^{(n+1)})_{\text{vec}}\|^2_{\ell_2}\end{aligned}$$

with the constant $C_p := \frac{2}{p}\mathcal{J}_{p,0}^{1-\frac{2}{p}} > 0$. Using the monotonicity as in (a) and the boundedness of the sequence $\big(\mathcal{J}_{HM}(X^{(n)},\epsilon^{(n)},W^{(n)})\big)_{n\in\mathbb{N}}$, we infer that

$$\lim_{n\to\infty}\left[\mathcal{J}_{HM}(X^{(n)},\epsilon^{(n)},W^{(n)}) - \mathcal{J}_{HM}(X^{(n+1)},\epsilon^{(n+1)},W^{(n+1)})\right] = 0,$$

and hence also

$$\lim_{n\to\infty}\|(X^{(n)}-X^{(n+1)})_{\text{vec}}\|^2_{\ell_2} = 0.$$

□

Now we note that with the assumption $X^{(n)} \to \bar{X}$ and $\epsilon^{(n)} \to \bar{\epsilon}$ for $n \to \infty$ with limit point $(\bar{X},\bar{\epsilon}) \in M_{d_1\times d_2} \times \mathbb{R}_{\geq 0}$, one can deduce that

$$\mathcal{J}_{HM}(X^{(n)},\epsilon^{(n)},W^{(n)}) \to f_{\bar{\epsilon}}(\bar{X})$$

for $n \to \infty$ by equation (4.25).

Given $\epsilon > 0$, a measurement vector $Y \in \mathbb{R}^m$ and the linear map Φ we consider the optimization problem

$$\min_{\Phi(X)=Y} f_\epsilon(X) \tag{4.32}$$

with $f_\epsilon(X) = \sum_{i=1}^d (\sigma_i(X)^2 + \epsilon^2)^{\frac{p}{2}}$ and $\sigma_i(X)$ being the i-th singular value of X, cf. (4.25). In the case that $f_\epsilon(X)$ is non-convex, i.e., for $p < 1$, one might practically only be capable of reaching critical points of the function.

Lemma 4.9. *Let $X \in M_{d_1 \times d_2}$ be a matrix with the SVD such that $X = \sum_{i=1}^d \sigma_i u_i v_i^*$, let $\epsilon > 0$. If we define*

$$\widetilde{W}(X,\epsilon) = 2\Bigg[\Big(\sum_{i=1}^d (\sigma_i^2 + \epsilon^2)^{\frac{2-p}{2}} u_i u_i^*\Big) \oplus \Big(\sum_{i=1}^d (\sigma_i^2 + \epsilon^2)^{\frac{2-p}{2}} v_i v_i^*\Big)\Bigg]^{-1} \in H_{d_1 d_2 \times d_1 d_2},$$

then $\widetilde{W}(X^{(n)}, \epsilon^{(n)}) = \widetilde{W}^{(n)}$, with $\widetilde{W}^{(n)}$ defined as in Algorithm 4, cf. (4.12)*.*

Furthermore, X is a critical point of the optimization problem (4.32) *if and only if*

$$\langle \widetilde{W}(X,\epsilon) X_{\text{vec}}, H_{\text{vec}} \rangle_{\ell_2} = 0 \text{ for all } H \in \mathcal{N}(\Phi) \quad \text{and} \quad \Phi(X) = Y. \tag{4.33}$$

In the case that f_ϵ is convex, i.e., if $p = 1$, (4.33) *implies that X is the unique minimizer of* (4.32)*.*

Proof. The first statement $\widetilde{W}(X^{(n)}, \epsilon^{(n)}) = \widetilde{W}^{(n)}$ follows straightforward from the definition of $\widetilde{W}(X,\epsilon)$ and (4.12).

Now we aim to show the necessity of (4.33). Let $X \in M_{d_1 \times d_2}$ be a critical point of (4.32) and without loss of generality we assume $d_1 \leq d_2$. We see from an easy calcualtion that in this case $f_\epsilon(X) = \operatorname{tr}\big[(XX^* + \epsilon^2 \mathbf{I}_{d_1})^{p/2}\big]$. We use the matrix derivative rules of [101, (7),(15),(18),(20) of Chapter 8.2] to derive that

$$\nabla g_\epsilon^p(X) = p(XX^* + \epsilon^2 \mathbf{I}_{d_1})^{\frac{p-2}{2}} X = p \sum_{i=1}^d (\sigma_i^2 + \epsilon^2)^{\frac{p-2}{2}} \sigma_i u_i v_i^*,$$

where the singular value decomposition $X = \sum_{i=1}^d \sigma_i u_i v_i^*$ is used in the last equality.

We employ the Kronecker sum inversion formula (4.7), to obtain

$$\begin{aligned}\Big[\widetilde{W}(X,\epsilon) X_{\text{vec}}\Big]_{\text{mat}} &= \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{2\sigma_k}{(\sigma_i^2 + \epsilon^2)^{\frac{2-p}{2}} + (\sigma_j^2 + \epsilon^2)^{\frac{2-p}{2}}} \sum_{k=1}^d u_i u_i^* u_k v_k^* v_j v_j^* \\ &= \sum_{i=1}^d (\sigma_i^2 + \epsilon^2)^{\frac{p-2}{2}} \sigma_i u_i v_i^*\end{aligned}$$

and thus $\nabla f_\epsilon(X) = p\big[\widetilde{W}(X,\epsilon)X_{\text{vec}}\big]_{\text{mat}}$.

Consider an arbitrary $H \in \mathcal{N}(\Phi)$ and the function $h_\epsilon(t) = f_\epsilon(X + tH) - f_\epsilon(X)$. We see that $(h_\epsilon)'(0) = \langle \nabla f_\epsilon(X), H\rangle_F$ and, hence, if X is a critical point of (4.32), then 0 is a critical point of h_ϵ as well, i.e., $(h_\epsilon)'(0) = 0$. Consequently,

$$0 = (h_\epsilon)'(0) = \langle \nabla f_\epsilon(X), H\rangle_F = p\big\langle \widetilde{W}(X,\epsilon)X_{\text{vec}}, H_{\text{vec}}\big\rangle_{\ell_2},$$

which implies (4.33).

To show the sufficiency of (4.33), let $X \in M_{d_1 \times d_2}$ such that $\Phi(X) = Y$ and also

$$\langle \widetilde{W}(X,\epsilon)X_{\text{vec}}, H_{\text{vec}}\rangle = 0$$

for all $H \in \mathcal{N}(\Phi)$. Using the calculation results above, it follows

$$0 = \langle \widetilde{W}(X,\epsilon)X_{\text{vec}}, H_{\text{vec}}\rangle = \frac{1}{p}\langle \nabla g_\epsilon^p(X), H\rangle_F.$$

This means that the gradient $\nabla f_\epsilon(X)$ is perpendicular to the null space $\mathcal{N}(\Phi)$ of Φ and as a consequence

$$\nabla f_\epsilon(X) \in \text{Ran}(\Phi^*) \text{ and } \Phi(X) = Y.$$

This corresponds exactly to first order optimality conditions of (4.32) and hence we can deduce that X is a critical point of f_ϵ under the linear constraint.

In the case $p = 1$, f_ϵ is a strictly convex function as $\epsilon > 0$ and therefore, the problem (4.32) has a unique minimizer. If we assume that $X \in M_{d_1 \times d_2}$ fulfills (4.33), this implies that this minimizer just conincides with X, as for any $X' \in M_{d_1 \times d_2}$ such that $\Phi(X') = Y$, it follows that $X - X' \in \mathcal{N}(\Phi)$ and thus by convexity of f_ϵ,

$$f_\epsilon(X') \geq f_\epsilon(X) + \langle \nabla f_\epsilon(X), X' - X\rangle_F = f_\epsilon(X) + \langle \widetilde{W}(X,\epsilon)X_{\text{vec}}, (X' - X)_{\text{vec}}\rangle_{\ell_2} = f_\epsilon(X).$$

□

4.3.4 Strong Schatten-p-null space property

For the analysis of `HM-IRLS` algorithm, we define the strong Schatten-p null space property which is closely related to the version we introduced already in Section 2.3.3 [51, 54, 117].

Definition 4.10 (Strong Schatten-p null space property)**.** Let $0 < p \leq 1$. We say that a linear map $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m$ matrix fulfills the strong Schatten-p null space property (Schatten-p NSP) of order r with constant $0 < \gamma_r \leq 1$ if

$$\left(\sum_{i=1}^{r} \sigma_i^2(Z) \right)^{p/2} < \frac{\gamma_r}{r^{1-\frac{p}{2}}} \left(\sum_{i=r+1}^{d} \sigma_i^p(Z) \right) \tag{4.34}$$

for all $Z \in \mathcal{N}(\Phi) \setminus \{0\}$.

We note that this version of the NSP implies the one in Definition 2.28 with constant $\gamma_r = 1$, which we call the weak Schatten-p-NSP, and constitutes a necessary and sufficient condition for solutions to the problem (4.2) to be rank-r matrices.

Theorem 4.11 ([52])**.** *Let $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m$ be a linear map, let $0 < p \leq 1$ and $r \in \mathbb{N}$. Then every matrix $X_0 \in M_{d_1 \times d_2}$ with* $\operatorname{rank}(X_0) \leq r$ *and $\Phi(X_0) = Y \in \mathbb{R}^m$ can be exactly recovered by Schatten-p minimization* (4.5) *if and only if Φ fulfills Definition 2.28 with constant $\gamma_r = 1$.*

Remark 4.12. The sufficiency of the weak Schatten-p NSP in Theorem 4.11 was already shown in [117]. However, to show the necessity as stated in the theorem, one also needs the recent generalization of Mirsky's singular value inequalities to concave functions as appeared in[3, 52].

We note that the (weak) Schatten-p NSP of Theorem 4.11 becomes a *stronger* requirement for growing p, which means that the Schatten-p property implies the Schatten-p' property if $0 < p' \leq p \leq 1$. As already presented in Theorem 2.31 for the weak NSP, also the strong Schatten-p null space property for any $0 < p \leq 1$ is implied by the rank restricted isometry property (rank-RIP) for a sufficiently small *rank restricted isometry constant* δ_r. This classical tool for the analysis of low-rank matrix recovery algorithms [21, 126] was already introduced in Definition 2.29. In the proof of [30, Theorem 4.1] it is shown that a restricted isometry constant of order $2r$ fulfilling $\delta_{2r} < \frac{2}{\sqrt{2}+3} \approx 0.4531$ indeed implies the strong Schatten-p NSP of order r with a constant $\gamma_r < 1$ for any $0 < p \leq 1$. To be more precise, one obtains that a constant $\delta_{2r} < \frac{2}{\sqrt{2}+3}$ implies that the validity of the strong Schatten-p NSP (5.17) of order r with $\gamma_r = \frac{(\sqrt{2}+1)^p}{2^p} \frac{\delta_{2r}^p}{(1-\delta_{2r})^p}$.

As already discussed in Section 2.3.3, in particular Theorem 2.30, the rank-RIP is fulfilled for a large number of random measurement models, e.g., for Gaussian, in the optimal measurement regime with overwhelming probability [21].

A useful tool that we will use for the convergence analysis of Algorithm 4, is the following version of the *reverse triangle inequalities* similar to Lemma 2.32 (ii), which is a consequence the strong Schatten-p NSP:

Lemma 4.13. *Let $0 < p \leq 1$. Assume that the linear map $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m$ fulfills the strong Schatten-p-NSP (5.17) of order r with constant $\gamma_r \in (0,1)$. Recall the definition of the best rank-r Schatten-p approximation error*

$$\beta_r(Z)_{S_p} := \inf\left\{\|Z - \widetilde{Z}\|_{S_p}^p,\ \widetilde{Z} \in M_{d_1\times d_2} \text{ has rank } r\right\} = \sum_{i=r+1}^{d} \sigma_i(Z)^p \tag{4.35}$$

of a matrix $Z \in M_{d_1 \times d_2}$.

Let $Z, Z' \in M_{d_1\times d_2}$ such that $\Phi(Z - Z') = 0$. Then

$$\|Z' - Z\|_F^p \leq \frac{2^p \gamma_r^{1-p/2}}{r^{1-p/2}} \frac{1}{1-\gamma_r} \left(\|Z'\|_{S_p}^p - \|Z\|_{S_p}^p + 2\beta_r(Z)_{S_p}\right). \tag{4.36}$$

Proof. The proof can be carried out by the modification of the proof of the corresponding result for ℓ_p-minimization [55, Theorem 13] by involving the generalization of Mirksy's singular value inequality to concave functions [3, 52]. Moreover, the proof of the statement [81, Threorem 12] can serve as the basis to show (4.36). □

Remark 4.14. It is important to note that, if $m < d_1 d_2$, null space property-type assumptions as (5.17) or the weak Schatten-p NSP are unfortunately not valid for the relevant case of matrix completion measurements [22], where $\Phi(X)$ is given as in (4.3).

4.3.5 Convergence results

Having provided some basic properties of HM-IRLS, we present now the convergence guarantees for the algorithm to at least critical points of a smoothed Schatten-p functional f_ϵ as defined in (4.25) without placing any additional assumptions. Beyond that, under the assumption of the strong Schatten-p null space property for the measurement operator Φ, we prove the a-posteriori exact recovery statement that HM-IRLS indeed converges to the low-rank minimizer X_0 in the case that $\lim\limits_{n\to\infty} \epsilon_n = 0$. Additionally, we provide a local convergence guarantee, which states that HM-IRLS recovers the low-rank matrix X_0 if we obtain an iterate $X^{(\bar{n})}$ that is in a close enough neighborhood to X_0.

Theorem 4.15. *Let $\Phi : M_{d_1\times d_2} \to \mathbb{R}^m$ be a linear operator, $Y \in \mathrm{Ran}(\Phi)$ a vector in its range. Let $(X^{(n)})_{n\geq 1}$ and $(\epsilon^{(n)})_{n\geq 1}$ be the sequences produced by Algorithm 4 for input parameters Φ, Y, r and $0 < p \leq 1$, let $\epsilon = \lim_{n\to\infty} \epsilon^{(n)}$.*

(i) If $\epsilon = 0$ and if Φ fulfills the strong Schatten-p NSP (5.17) of order r with constant $0 < \gamma_r < 1$, then the sequence $(X^{(n)})_{n\geq 1}$ converges to a matrix $\bar{X} \in M_{d_1\times d_2}$ of rank

at most r that is the unique minimizer of the Schatten-p minimization problem (4.5)*. Moreover, there exists an absolute constant $\hat{C} > 0$ such that for any Z with $\Phi(Z) = Y$ and any $\widetilde{r} \leq r$, it holds that*

$$\|Z - \bar{X}\|_F^p \leq \frac{\hat{C}}{r^{1-p/2}} \beta_{\widetilde{r}}(Z)_{S_p}, \tag{4.37}$$

where $\hat{C} = \frac{2^{p+1}\gamma_r^{1-p/2}}{1-\gamma_r}$ and $\beta_{\widetilde{r}}(Z)_{S_p}$ is the best rank-$\widetilde{r}$ Schatten-p approximation error, cf. (4.35)*.*

(ii) If $\epsilon > 0$, then each accumulation point $\bar{X}$ of $(X^{(n)})_{n\geq 1}$ is a stationary point of the ϵ-perturbed Schatten-p functional f_ϵ as in (4.23) *under the linear constraint $\Phi(X) = Y$. If, additionally, $p = 1$, then $\bar{X}$ is the unique global minimizer of f_ϵ.*

(iii) If Φ fulfills the strong Schatten-p NSP of order $2r$ with constant $\gamma_{2r} < 1$, assume that there exists a matrix $X_0 \in M_{d_1 \times d_2}$ with $\Phi(X_0) = Y$ of rank $\widetilde{r} \leq r \leq \frac{\min(d_1,d_2)}{2}$, a constant $0 < \rho < 1$ and an iteration $\bar{n} \in \mathbb{N}$ such that

$$\|X^{(\bar{n})} - X_0\|_{S_\infty} \leq \rho\sigma_{\widetilde{r}}(X_0)$$

and $\epsilon^{\bar{n}} = \sigma_{r+1}(X^{\bar{n}})$.
If the condition number $\kappa = \frac{\sigma_1(X_0)}{\sigma_{\widetilde{r}}(X_0)}$ of X_0 and ρ are sufficiently small (see (4.40) *and* (4.41)*), then*

$$X^{(n)} \to X_0 \quad \text{for } n \to \infty.$$

Proof. (i) Let us first assume that there exists an iteration $\bar{n} \in \mathbb{N}$ such that $\epsilon^{(\bar{n})} = 0$. Define $\bar{X} := X^{(\bar{n})}$, which fulfills by construction that $\Phi(\bar{X}) = Y$ and $\sigma_{r+1}(\bar{X}) = 0$, i.e., $\operatorname{rank}(\bar{X}) \leq r$.

In the other case, where $\epsilon^{(n)} > 0$ for all $n \in \mathbb{N}$, there exists a subsequence $(n_\ell)_{l\in\mathbb{N}}$ of $(n)_{n\geq n_0}$ such that $\epsilon^{(n_\ell)} < \epsilon^{(n_\ell - 1)}$ for all $\ell \in \mathbb{N}$. As shown in Lemma 5.4(b), $(X^{(n)})_n$ is bounded and one can extract a further subsequence, which we denote again by $(X^{(n_\ell)})_l$ converging to a limit $\bar{X} := \lim_{l\to\infty} X^{(n_\ell)}$. As $\lim_{l\to\infty} \epsilon^{(n_\ell)} = 0$, we conclude as well that $\lim_{l\to\infty} \sigma_{r+1}(X^{(n_\ell)}) \leq \lim_{l\to\infty} \epsilon^{(n_\ell - 1)} = 0$. Furthermore, we obtain by Weyl's stability estimate for the $(r+1)$-th singular value [51, Theorem 7.1] that $\sigma_{r+1}(\bar{X}) = 0$. As a consequence, also for this case, $\bar{X}$ fulfills $\Phi(\bar{X}) = Y$ and $\operatorname{rank}(\bar{X}) \leq r$.

As a next step, we aim to show that the whole sequence $(X^{(n)})_n$ converges to $\bar{X}$. According to (4.25), we have that

$$\mathcal{J}_{HM}(X^{(n)}, \epsilon^{(n)}, W^{(n)}) = f_{\epsilon^{(n)}}(X^{(n)})$$

for any $n \in \mathbb{N}$. Since $(X^{(n_\ell)}) \xrightarrow{l\to\infty} \bar{X}$ and $\epsilon^{(n_\ell)} \xrightarrow{l\to\infty} 0$, we see consequently that

$$\mathcal{J}_{HM}(X^{(n_\ell)}, \epsilon^{(n_\ell)}, W^{(n_\ell)}) \xrightarrow{l\to\infty} \|\bar{X}\|_{S_p}^p.$$

Using the non-increasing monotonicity from Lemma 5.3(a) it already follows that the same is valid for the whole sequence $(X^{(n)})_{n\geq 1}$, i.e., $\mathcal{J}_{HM}(X^{(n)}, \epsilon^{(n)}, W^{(n)}) \xrightarrow{n\to\infty} \|\bar{X}\|_{S_p}^p$. By the application of the triangle inequality for the p-power of the Schatten-p quasi-norms $\|\cdot\|_{S_p}^p$, it follows

$$\mathcal{J}_{HM}(X^{(n)}, \epsilon^{(n)}, W^{(n)}) - d(\epsilon^{(n)})^p \leq \|X^{(n)}\|_{S_p}^p \leq \mathcal{J}_{HM}(X^{(n)}, \epsilon^{(n)}, W^{(n)}).$$

Since $\lim_{n\to\infty} d(\epsilon^{(n)})^p = 0$, we get that

$$\|X^{(n)}\|_{S_p}^p \xrightarrow{n\to\infty} \|\bar{X}\|_{S_p}^p.$$

It still remains to prove that $X^{(n)} \to \bar{X}$. We note that from Lemma 4.13, it follows that

$$\|\bar{X} - X^{(n)}\|_F^p \leq \frac{2^p \gamma_r^{1-p/2}}{r^{1-p/2}} \frac{1}{1-\gamma_r} \left(\|\bar{X}\|_{S_p}^p - \|X^{(n)}\|_{S_p}^p + 2\beta_{r+1}(\bar{X})_{S_p} \right)$$

and therefore, using $\beta_{r+1}(\bar{X})_{S_p} = 0$, we get

$$\|\bar{X} - X^{(n)}\|_F^p \leq \frac{2^p \gamma_r^{1-p/2}}{r^{1-p/2}} \frac{1}{1-\gamma_r} \left(\|\bar{X}\|_{S_p}^p - \|X^{(n)}\|_{S_p}^p \right).$$

Now passing to the limit $n \to \infty$ gives that $\lim\limits_{n\to\infty} \|\bar{X} - X^{(n)}\|_F = 0$, and thus $X^{(n)} \to \bar{X}$ for $n \to \infty$.

Using the fact that $\bar{X}$ is of rank at most r and also fulfills $\Phi(X) = Y$, the strong Schatten-p null space property implies via Lemma 4.13 that $\bar{X}$ is indeed the unique solution to (4.5) coinciding with X_0.

For deriving the error bound (5.23), we observe that any matrix Z with $\Phi(Z) = Y$ fulfills

$$\|Z - \bar{X}\|_F^p \leq \frac{2^p \gamma_r^{1-p/2}}{r^{1-p/2}} \frac{1}{1-\gamma_r} \left(\|\bar{X}\|_{S_p}^p - \|Z\|_{S_p}^p + 2\beta_r(Z)_{S_p} \right) \leq \frac{2^{p+1} \gamma_r^{1-p/2}}{r^{1-p/2}} \frac{1}{1-\gamma_r} \beta_r(Z)_{S_p}, \tag{4.38}$$

where we used Lemma 4.13 in the first inequality and the minimization property of $\bar{X}$ in the second inequality. We conclude the proof of (5.23) with the observation that $\beta_r(Z)_{S_p} \leq \beta_{\widetilde{r}}(Z)_{S_p}$ if $\widetilde{r} \leq r$.

(ii) As a first step, we want to prove $X^{(n)} \to X^{\epsilon}$, $n \to \infty$, where $X^{\epsilon} \in M_{d_1 \times d_2}$ is a critical point of f_{ϵ} under the linear constraint. We already made the observation that the sequence $(X^{(n)})_{n\geq 1}$ is bounded and, hence, has accumulation points. Denote with $(X^{(n_\ell)})_{\ell\geq 1}$ any convergent subsequence of $(X^{(n)})_{n\geq 1}$ and with limit $\bar{X}$.

As the weight matrix $\widetilde{W}(X,\epsilon)$ as defined in Lemma 5.7 depends continuously on the variables X and ϵ, we can conclude

$$\lim_{\ell\to\infty} \widetilde{W}(X^{(n_\ell)}, \epsilon^{(n_\ell)}) = \widetilde{W}(\bar{X}, \epsilon) =: \overline{W}.$$

On the other hand, using Lemma 5.5(c), it also follows that $X^{(n_\ell+1)} \to \bar{X}$ for $\ell \to \infty$ and as a consequence we have $\lim\limits_{\ell\to\infty} \widetilde{W}^{(n_\ell+1)} = \overline{W}$ as well. Note that from interpreting $X^{(n_\ell+1)}$ as a minimizer of the functional $\mathcal{J}_{HM}$ in (4.29) and the proof of Lemma 4.7, we can conclude that

$$\widetilde{W}^{(n_\ell)} X^{(n_\ell+1)} \in \operatorname{Ran}(\Phi^*) \text{ and } \Phi(X^{(n_\ell+1)}) = Y.$$

This implies the existence of $\lambda \in \mathbb{R}^m$ such that $\widetilde{W}^{(n_\ell+1)} X^{(n_\ell)} = \Phi^*(\lambda)$. Next we note that for all $H \in \mathcal{N}(\Phi)$ and all $\ell \in \mathbb{N}$ holds true that,

$$\langle \widetilde{W}^{(n_\ell)} X_{\text{vec}}^{(n_\ell+1)}, H_{\text{vec}} \rangle = \langle \Phi^*(\lambda), H_{\text{vec}} \rangle = \langle \lambda, \Phi(\eta) \rangle = 0.$$

Consequently, $\langle \overline{W} \bar{X}_{\text{vec}}, \eta_{\text{vec}} \rangle = \lim\limits_{\ell\to\infty} \langle \widetilde{W}^{(n_\ell)} X_{\text{vec}}^{(n_\ell+1)}, H_{\text{vec}} \rangle = 0$ for all $H \in \mathcal{N}(\Phi)$. As shown in Lemma 5.7, this implies that $\bar{X}$ is a stationary point of f_{ϵ} and in the convex case, i.e., if $p = 1$, we even showed coincidence with the unique minimizer X^{ϵ}.

(iii) This statement follows from Theorem 4.16, which is proven below. □

4.3.6 Locally superlinear convergence

The goal of the next subsection is to introduce our locally superlinear convergence rate result for `HM-IRLS` in Theorem 4.16 under the assumption that the operator Φ fulfills an appropriate Schatten-p null space property.

Theorem 4.16 (Locally Superlinear Convergence Rate). *Assume that the linear map $\Phi : M_{d_1\times d_2} \to \mathbb{R}^m$ fulfills the strong Schatten-p NSP of order $2r$ with constant $\gamma_{2r} < 1$ and that there exists a matrix $X_0 \in M_{d_1\times d_2}$ with $\operatorname{rank}(X_0) = r \leq \frac{\min(d_1,d_2)}{2}$ such that $\Phi(X_0) = Y$. Let Φ, Y, r and $0 < p \leq 1$ be the input parameters of Algorithm 4. Moreover, let $\kappa = \frac{\sigma_1(X_0)}{\sigma_r(X_0)}$ be the condition number of X_0 and $\eta^{(n)} := X^{(n)} - X_0$ be the residuals of the n-th output matrix of Algorithm 4 for $n \in \mathbb{N}$.*

Assume that there exists an iteration $\bar{n} \in \mathbb{N}$ and a constant $0 < \rho < 1$ such that

$$\|\eta^{(\bar{n})}\|_{S_\infty} \leq \rho \sigma_r(X_0) \tag{4.39}$$

and $\epsilon^{(\bar{n})} = \sigma_{r+1}(X^{(\bar{n})})$.

If, additionally, the condition number κ and ρ are small enough such that

$$\mu \|\eta^{(\bar{n})}\|_{S_\infty}^{p(1-p)} < 1 \tag{4.40}$$

with the constant

$$\mu := 2^{5p}(1+\gamma_{2r})^p \Big(\frac{\gamma_{2r}(3+\gamma_{2r})(1+\gamma_{2r})}{(1-\gamma_{2r})}\Big)^{2-p} \Big(\frac{d-r}{r}\Big)^{2-\frac{p}{2}} r^p \frac{\sigma_r(X_0)^{p(p-1)}}{(1-\rho)^{2p}} \kappa^p, \tag{4.41}$$

then

$$\|\eta^{(n+1)}\|_{S_\infty} \leq \mu^{1/p} \left(\|\eta^{(n)}\|_{S_\infty}\right)^{2-p}$$

and

$$\|\eta^{(n+1)}\|_{S_p} \leq \mu^{1/p} \left(\|\eta^{(n)}\|_{S_p}\right)^{2-p}$$

for all $n \geq \bar{n}$.

As a first step towards the proof of Theorem 4.15, we show the following lemma.

Lemma 4.17. *Let $(X^{(n)})_n$ be the output sequence of Algorithm 4 for parameters Φ, Y, r and $0 < p \leq 1$, and $X_0 \in M_{d_1 \times d_2}$ be a matrix such that $\Phi(X_0) = Y$.*

(i) Let $\eta_{2r}^{(n+1)}$ be the best rank-2r approximation of $\eta^{(n+1)} = X^{(n+1)} - X_0$. Then

$$\|\eta^{(n+1)} - \eta_{2r}^{(n+1)}\|_{S_p}^{2p} \leq 2^{2-p} \Bigg(\sum_{i=r+1}^{d} \left(\sigma_i^2(X^{(n)}) + \epsilon^{(n)2}\right)^{\frac{p}{2}} \Bigg)^{2-p} \|\eta_{\text{vec}}^{(n+1)}\|_{\ell_2(\widetilde{W}^{(n)})}^{2p},$$

where $\widetilde{W}^{(n)}$ denotes the harmonic mean weight matrix from (4.12).

(ii) Assume that the linear map $\Phi : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ fulfills the strong Schatten-p NSP of order 2r with constant $\gamma_{2r} < 1$. Then

$$\|\eta^{(n+1)}\|_{S_2}^{2p} \leq 2^p \frac{\gamma_{2r}^{2-p}}{r^{2-p}} \Bigg(\sum_{i=r+1}^{d} \left(\sigma_i^2(X^{(n)}) + \epsilon^{(n)2}\right)^{\frac{p}{2}} \Bigg)^{2-p} \|\eta_{\text{vec}}^{(n+1)}\|_{\ell_2(\widetilde{W}^{(n)})}^{2p} \tag{4.42}$$

(i) Under the same assumption as for (ii), it holds that

$$\|\eta^{(n+1)}\|_{S_p}^{2p} \leq (1+\gamma_{2r})^2 2^{2-p} \bigg(\sum_{i=r+1}^{d} \big(\sigma_i^2(X^{(n)}) + \epsilon^{(n)2}\big)^{\frac{p}{2}} \bigg)^{2-p} \|\eta_{\text{vec}}^{(n+1)}\|_{\ell_2(\widetilde{W}^{(n)})}^{2p}.$$

Proof of Lemma 4.17. (i) Let the $X^{(n)} = \widetilde{U}^{(n)}\Sigma^{(n)}\widetilde{V}^{(n)*}$ be the (full) singular value decomposition of $X^{(n)}$, i.e., $\widetilde{U}^{(n)} \in U_{d_1}$ and $\widetilde{V}^{(n)} \in U_{d_2}$ are unitary matrices and $\Sigma^{(n)} = \operatorname{diag}(\sigma_1(X^{(n)}), \ldots, \sigma_r(X^{(n)})) \in M_{d_1 \times d_2}$. We define $U_T^{(n)} \in U_{d_1 \times r}$ as the matrix of the first r columns of $\widetilde{U}^{(n)}$ and $U_{T_c}^{(n)} \in U_{d_1 \times (d_1 - r)}$ as the matrix of its last $d_1 - r$ columns, so that $\widetilde{U}^{(n)} = \begin{pmatrix} U_T^{(n)} & U_{T_c}^{(n)} \end{pmatrix}$, and similarly $V_T^{(n)}$ and $V_{T_c}^{(n)}$.

As $\mathbf{I}_{d_1} = U_T^{(n)}U_T^{(n)*} + U_{T_c}^{(n)}U_{T_c}^{(n)*}$ and $\mathbf{I}_{d_2} = V_T^{(n)}V_T^{(n)*} + V_{T_c}^{(n)}V_{T_c}^{(n)*}$, we note that

$$U_{T_c}^{(n)}U_{T_c}^{(n)*}\eta^{(n+1)}V_{T_c}^{(n)}V_{T_c}^{(n)*} = \eta^{(n+1)} - U_T^{(n)}U_T^{(n)*}\eta^{(n+1)} + U_{T_c}^{(n)}U_{T_c}^{(n)*}\eta^{(n+1)}V_T^{(n)}V_T^{(n)*},$$

while $U_T^{(n)}U_T^{(n)*}\eta^{(n+1)} + U_{T_c}^{(n)}U_{T_c}^{(n)*}\eta^{(n+1)}V_T^{(n)}V_T^{(n)*}$ has a rank of at most $2r$. This implies that

$$\|\eta^{(n+1)} - \eta_{2r}^{(n+1)}\|_{S_p} \leq \|U_{T_c}^{(n)}U_{T_c}^{(n)*}\eta^{(n+1)}V_{T_c}^{(n)}V_{T_c}^{(n)*}\|_{S_p} = \|U_{T_c}^{(n)*}\eta^{(n+1)}V_{T_c}^{(n)}\|_{S_p}. \tag{4.43}$$

Using the definitions of $\widetilde{U}^{(n)}$ and $\widetilde{V}^{(n)}$, we write the harmonic mean weight matrices of the n-th iteration (4.12) as

$$\widetilde{W}^{(n)} = 2(\widetilde{V}^{(n)} \otimes \widetilde{U}^{(n)})\big(\bar{\Sigma}_{d_1}^{(n)2-p} \oplus \bar{\Sigma}_{d_2}^{(n)2-p}\big)^{-1}(\widetilde{V}^{(n)} \otimes \widetilde{U}^{(n)})^*, \tag{4.44}$$

where $\bar{\Sigma}_{d_1}^{(n)} \in M_{d_1 \times d_1}$ and $\bar{\Sigma}_{d_2}^{(n)} \in M_{d_2 \times d_2}$ are the diagonal matrices with the smoothed singular values of $X^{(n)}$ from (4.13), but filled up with zeros if necessary. Using the abbreviation

$$\Omega := (\widetilde{V}^{(n)} \otimes \widetilde{U}^{(n)})^* \widetilde{W}^{(n)\frac{1}{2}} \eta_{\text{vec}}^{(n+1)} \in \mathbb{R}^{d_1 d_2}, \tag{4.45}$$

we rewrite

$$\begin{aligned} \eta_{\text{vec}}^{(n+1)} = \widetilde{W}^{(n)-\frac{1}{2}}\widetilde{W}^{(n)\frac{1}{2}}\eta_{\text{vec}}^{(n+1)} &= 2^{-1/2}(\widetilde{V}^{(n)} \otimes \widetilde{U}^{(n)})\big(\bar{\Sigma}_{d_1}^{(n)2-p} \oplus \bar{\Sigma}_{d_2}^{(n)2-p}\big)^{1/2}\Omega \\ &= 2^{-1/2}(\widetilde{V}^{(n)} \otimes \widetilde{U}^{(n)}) \Big[(\mathbf{I}_{d_2} \otimes \bar{\Sigma}_{d_1}^{(n)\frac{2-p}{2}})D_L + (\bar{\Sigma}_{d_2}^{(n)\frac{2-p}{2}} \otimes \mathbf{I}_{d_1})D_R\Big]\,\Omega \end{aligned} \tag{4.46}$$

with the diagonal matrices $D_L, D_R \in M_{d_1 d_2 \times d_1 d_2}$ such that

$$(D_L)_{i+(j-1)d_1, i+(j-1)d_1} = \Big(1 + \Big(\frac{\sigma_j^2(X^{(n)}) + \epsilon^{(n)2}}{\sigma_i^2(X^{(n)}) + \epsilon^{(n)2}}\Big)^{\frac{2-p}{2}}\Big)^{-1/2}$$

and

$$(D_R)_{i+(j-1)d_1,i+(j-1)d_1} = \Big(\Big(\frac{\sigma_i^2(X^{(n)}) + \epsilon^{(n)2}}{\sigma_j^2(X^{(n)}) + \epsilon^{(n)2}}\Big)^{\frac{2-p}{2}} + 1\Big)^{-1/2}$$

for $i \in [d_1]$ and $j \in [d_2]$. This can be seen from the definitions of the Kronecker product $\otimes$ and the Kronecker sum $\oplus$ (cf. section 4.1.1), as $\Big(\big(\bar{\Sigma}_{d_1}^{(n)2-p} \oplus \bar{\Sigma}_{d_2}^{(n)2-p}\big)^{1/2}\Big)_{i+(j-1)d_1,i+(j-1)d_1} = (s_i + s_j)^{1/2} = s_i(s_i + s_j)^{-1/2} + s_j(s_i + s_j)^{-1/2} = s_i^{1/2}(1 + \frac{s_j}{s_i})^{-1/2} + s_j^{1/2}(\frac{s_i}{s_j} + 1)^{-1/2}$, if s_ℓ denotes the ℓ-th diagonal entry of $\bar{\Sigma}_{d_2}^{(n)2-p}$ and $\bar{\Sigma}_{d_1}^{(n)2-p}$ for $\ell \in [\max(d_1, d_2)]$.

If we write $\bar{\Sigma}_{d_1,T_c}^{(n)\frac{2-p}{2}} \in M_{(d_1-r)\times(d_1-r)}$ for the diagonal matrix containing the $d_1 - r$ last diagonal elements of $\bar{\Sigma}_{d_1}^{(n)2-p}$ and $\bar{\Sigma}_{d_2,T_c}^{(n)\frac{2-p}{2}} \in M_{(d_1-r)\times(d_1-r)}$ for the diagonal matrix containing the $d_2 - r$ last diagonal elements of $\bar{\Sigma}_{d_2}^{(n)2-p}$, it follows from (4.46) that

$$\begin{aligned}\|U_{T_c}^{(n)*}\eta^{(n+1)}V_{T_c}^{(n)}\|_{S_p}^p &= 2^{-\frac{p}{2}}\Big\|U_{T_c}^{(n)*}\widetilde{U}^{(n)}\Big[\bar{\Sigma}_{d_1}^{(n)\frac{2-p}{2}}(D_L\Omega)_{\mathrm{mat}} + (D_R\Omega)_{\mathrm{mat}}\bar{\Sigma}_{d_2}^{(n)\frac{2-p}{2}}\Big]\widetilde{V}^{(n)*}V_{T_c}^{(n)}\Big\|_{S_p}^p \\ &= 2^{-\frac{p}{2}}\Big\|\bar{\Sigma}_{d_1,T_c}^{(n)\frac{2-p}{2}}\big[(D_L\Omega)_{\mathrm{mat}}\big]_{T_c,T_c} + \big[(D_R\Omega)_{\mathrm{mat}}\big]_{T_c,T_c}\bar{\Sigma}_{d_2,T_c}^{(n)\frac{2-p}{2}}\Big\|_{S_p}^p \\ &\leq 2^{-\frac{p}{2}}\Big\|\bar{\Sigma}_{d_1,T_c}^{(n)\frac{2-p}{2}}\big[(D_L\Omega)_{\mathrm{mat}}\big]_{T_c,T_c}\Big\|_{S_p}^p + \Big\|\big[(D_R\Omega)_{\mathrm{mat}}\big]_{T_c,T_c}\bar{\Sigma}_{d_2,T_c}^{(n)\frac{2-p}{2}}\Big\|_{S_p}^p\end{aligned} \tag{4.47}$$

with the notation that M_{T_c,T_c} denotes the submatrix of M which contains the intersection of the last $d_1 - r$ rows of M with its last $d_2 - r$ columns.

Now, Hölder's inequality for Schatten-p quasinorms (e.g. [63, Theorem 11.2]) can be used to see that

$$\Big\|\bar{\Sigma}_{d_1,T_c}^{(n)\frac{2-p}{2}}\big[(D_L\Omega)_{\mathrm{mat}}\big]_{T_c,T_c}\Big\|_{S_p}^p \leq \Big\|\bar{\Sigma}_{T_c}^{(n)\frac{2-p}{2}}\Big\|_{S_{\frac{2p}{2-p}}}^p\Big\|\big[(D_L\Omega)_{\mathrm{mat}}\big]_{T_c,T_c}\Big\|_{S_2}^p. \tag{4.48}$$

Inserting the definition

$$\Big\|\bar{\Sigma}_{T_c}^{(n)\frac{2-p}{2}}\Big\|_{S_{\frac{2p}{2-p}}}^p = \left(\sum_{i=r+1}^{d}\big(\sigma_i^2(X^{(n)}) + \epsilon^{(n)2}\big)^{\frac{2-p}{4}\frac{2p}{2-p}}\right)^{\frac{2-p}{2}} = \left(\sum_{i=r+1}^{d}\big(\sigma_i^2(X^{(n)}) + \epsilon^{(n)2}\big)^{\frac{p}{2}}\right)^{\frac{2-p}{2}}$$

allows us to rewrite the first factor, while the second factor can be bounded by

$$\begin{aligned}\big\|\big[(D_L\Omega)_{\mathrm{mat}}\big]_{T_c,T_c}\big\|_{S_2}^p \leq \big\|(D_L\Omega)_{\mathrm{mat}}\big\|_{S_2}^p \leq \|\Omega_{\mathrm{mat}}\|_{S_2}^p = \|(\widetilde{V}^{(n)} \otimes \widetilde{U}^{(n)})^*\widetilde{W}^{(n)\frac{1}{2}}\eta_{\mathrm{vec}}^{(n+1)}\|_{\ell_2}^p \\ = \|\widetilde{W}^{(n)\frac{1}{2}}\eta_{\mathrm{vec}}^{(n+1)}\|_{\ell_2}^p = \|\eta_{\mathrm{vec}}^{(n+1)}\|_{\ell_2(\widetilde{W}^{(n)})}^p,\end{aligned}$$

as the matrix $D_L \in M_{d_1d_2\times d_1d_2}$ from (4.46) fulfills $\|D_L\|_{S_\infty} \leq 1$ since its entries are bounded by 1; we also recall the definition (4.45) of Ω and that $\widetilde{V}^{(n)}$ and $\widetilde{U}^{(n)}$ are unitary.

The term $\left\| \left[(D_R\Omega)_{\text{mat}} \right]_{T_c,T_c} \bar{\Sigma}_{d_2,T_c}^{(n)\frac{2-p}{2}} \right\|_{S_p}^p$ in (4.47) can be estimated analogously. Combining this with (4.43), we obtain

$$\|\eta^{(n+1)} - \eta_{2r}^{(n+1)}\|_{S_p}^{2p} \leq 2^{-p}\Big(2\Big(\sum_{i=r+1}^{d} \big(\sigma_i^2(X^{(n)}) + \epsilon^{(n)2}\big)^{\frac{p}{2}}\Big)^{\frac{2-p}{2}}\Big)^2 \|\eta_{\text{vec}}^{(n+1)}\|_{\ell_2(\widetilde{W}^{(n)})}^{2p},$$

concluding the proof of statement (i).

(ii) Using the strong Schatten-p null space property (5.17) of order $2r$ and that $\eta^{(n+1)} \in \mathcal{N}(\Phi)$, we estimate

$$\begin{aligned} \|\eta^{(n+1)}\|_{S_2}^{2p} &= \big(\|\eta_{2r}^{(n+1)}\|_{S_2}^2 + \|\eta^{(n+1)} - \eta_{2r}^{(n+1)}\|_{S_2}^2\big)^p \leq \Big(\frac{\gamma_{2r}^{2/p} + \gamma_{2r}^{2/p-1}}{(2r)^{2/p-1}}\|\eta^{(n+1)} - \eta_{2r}^{(n+1)}\|_{S_p}^2\Big)^p \\ &\leq \frac{\gamma_{2r}^{2-p}(\gamma_{2r}+1)^p}{(2r)^{2-p}}\|\eta^{(n+1)} - \eta_{2r}^{(n+1)}\|_{S_p}^{2p} \leq 2^p\frac{\gamma_{2r}^{2-p}}{2^{2-p}r^{2-p}}\|\eta^{(n+1)} - \eta_{2r}^{(n+1)}\|_{S_p}^{2p}, \end{aligned} \tag{4.49}$$

where we use in the second inequality a version of Stechkin's lemma [81, Lemma 3.1], which leads to the estimate

$$\|\eta^{(n+1)} - \eta_{2r}^{(n+1)}\|_{S_2}^2 \leq \Big(\frac{\|\eta_{2r}^{(n+1)}\|_{S_2}}{2r}\Big)^{2-p}\|\eta^{(n+1)} - \eta_{2r}^{(n+1)}\|_{S_p}^p \leq \frac{\gamma_{2r}^{2/p-1}}{(2r)^{2/p-1}}\|\eta^{(n+1)} - \eta_{2r}^{(n+1)}\|_{S_p}^2.$$

Combining (4.49) with statement (i), this results in

$$\|\eta^{(n+1)}\|_{S_2}^{2p} \leq 2^p\frac{\gamma_{2r}^{2-p}}{r^{2-p}}\Big(\sum_{i=r+1}^{d} \big(\sigma_i^2(X^{(n)}) + \epsilon^{(n)2}\big)^{\frac{p}{2}}\Big)^{2-p} \|\eta_{\text{vec}}^{(n+1)}\|_{\ell_2(\widetilde{W}^{(n)})}^{2p},$$

which shows statement (ii).

(iii) For the third statement, we use the strong Schatten-p NSP (5.17) to see that

$$\|\eta^{(n+1)}\|_{S_p}^p = \|\eta_{2r}^{(n+1)}\|_{S_p}^p + \|\eta^{(n+1)} - \eta_{2r}^{(n+1)}\|_{S_p}^p \leq (1+\gamma_{2r})\|\eta^{(n+1)} - \eta_{2r}^{(n+1)}\|_{S_p}^p,$$

and combine this with statement (i). □

Lemma 4.18. *Let $(X^{(n)})_n$ be the output sequence of Algorithm 4 with parameters Φ, Y, r and $0 < p \leq 1$, and $\widetilde{W}^{(n)}$ be the harmonic mean weight matrix (4.12) for $n \in \mathbb{N}$. Let $X_0 \in M_{d_1 \times d_2}$ be a rank-r matrix such that $\Phi(X_0) = Y$ with condition number $\kappa := \frac{\sigma_1(X_0)}{\sigma_r(X_0)}$.*

(i) If (4.39) *is fulfilled for iteration n, then $\eta^{(n+1)} = X^{(n)} - X_0$ fulfills*

$$\left\|\eta_{\text{vec}}^{(n+1)}\right\|_{\ell_2(\widetilde{W}^{(n)})}^{2p} \leq \frac{4^p r^{p/2} (s_r^0)^{p(p-1)}}{(1-\rho)^{2p}} \kappa^p \frac{\|\eta^{(n)}\|_{S_\infty}^{2p-p^2}}{(\epsilon^{(n)})^{2p-p^2}} \|\eta^{(n+1)}\|_{S_2}^p.$$

(ii) Under the same assumption as for (i), it holds that

$$\left\|\eta_{\text{vec}}^{(n+1)}\right\|_{\ell_2(\widetilde{W}^{(n)})}^{2p} \leq \frac{7^p r^{p/2} \max(r, d-r)^{p/2} (s_r^0)^{p(p-1)}}{(1-\rho)^{2p}} \kappa^p \frac{\|\eta^{(n)}\|_{S_\infty}^{2p-p^2}}{(\epsilon^{(n)})^{2p-p^2}} \|\eta^{(n+1)}\|_{S_\infty}^p.$$

Proof of Lemma 4.18. (i) Recall that $X^{(n+1)} = \underset{\Phi(X)=Y}{\arg\min} \|X_{\text{vec}}\|_{\ell_2(\widetilde{W}^{(n)})}^2$ is the minimizer of the weighted least squares problem with weight matrix $\widetilde{W}^{(n)}$. As $\eta^{(n+1)} = X^{(n+1)} - X_0$ is in the null space of the measurement map Φ, it follows from Lemma 4.7 that

$$0 = \langle \widetilde{W}^{(n)} X_{\text{vec}}^{(n+1)}, \eta_{\text{vec}}^{(n+1)} \rangle = \langle \widetilde{W}^{(n)} (\eta^{(n+1)} + X_0)_{\text{vec}}, \eta_{\text{vec}}^{(n+1)} \rangle,$$

which is equivalent to

$$\left\|\eta_{\text{vec}}^{(n+1)}\right\|_{\ell_2(\widetilde{W}^{(n)})}^2 = \langle \widetilde{W}^{(n)} \eta_{\text{vec}}^{(n+1)}, \eta_{\text{vec}}^{(n+1)} \rangle = -\langle \widetilde{W}^{(n)} (X_0)_{\text{vec}}, \eta_{\text{vec}}^{(n+1)} \rangle. \tag{4.50}$$

Using Hölder's inequality, we can therefore estimate

$$\begin{aligned} \left\|\eta_{\text{vec}}^{(n+1)}\right\|_{\ell_2(\widetilde{W}^{(n)})}^2 &= -\langle \widetilde{W}^{(n)} (X_0)_{\text{vec}}, \eta_{\text{vec}}^{(n+1)} \rangle_{\ell_2} = -\langle [\widetilde{W}^{(n)} (X_0)_{\text{vec}}]_{\text{mat}}, \eta^{(n+1)} \rangle_F \\ &\leq \left\| \left[\widetilde{W}^{(n)} (X_0)_{\text{vec}}\right]_{\text{mat}} \right\|_{S_2} \|\eta^{(n+1)}\|_{S_2}. \end{aligned} \tag{4.51}$$

To bound the first factor, we first rewrite the action of $\widetilde{W}^{(n)}$ on X_0 in the matrix space as

$$\begin{aligned} \left[\widetilde{W}^{(n)} (X_0)_{\text{vec}}\right]_{\text{mat}} &= 2[(\widetilde{V}^{(n)} \otimes \widetilde{U}^{(n)}) \big(\bar{\Sigma}_{d_1}^{(n)2-p} \oplus \bar{\Sigma}_{d_2}^{(n)2-p}\big)^{-1} (\widetilde{V}^{(n)} \otimes \widetilde{U}^{(n)})^* (X_0)_{\text{vec}}]_{\text{mat}} = \\ &= \widetilde{U}^{(n)} \big(H^{(n)} \circ (\widetilde{U}^{(n)*} X_0 \widetilde{V}^{(n)})\big) \widetilde{V}^{(n)*}, \end{aligned}$$

using (4.44) and Lemma 4.17 about the action of inverses of Kronecker sums, with the notation that $H^{(n)} \in M_{d_1 \times d_2}$ such that

$$H_{ij}^{(n)} = 2\left[\mathbf{1}_{\{i \leq d\}} (\sigma_i^2(X^{(n)}) + \epsilon^{(n)2})^{\frac{2-p}{2}} + \mathbf{1}_{\{j \leq d\}} (\sigma_j^2(X^{(n)}) + \epsilon^{(n)2})^{\frac{2-p}{2}}\right]^{-1}$$

for $i \in [d_1]$, $j \in [d_2]$, where $\mathbf{1}_{\{i \leq d\}} = 1$ if $i \leq d$ and $\mathbf{1}_{\{i \leq d\}} = 0$ otherwise. This enables us to estimate

$$\begin{aligned}\left\|\left[\widetilde{W}^{(n)}(X_0)_{\mathrm{vec}}\right]_{\mathrm{mat}}\right\|_{S_2}^2 &= \left\|\widetilde{U}^{(n)}\big(H^{(n)}\circ(\widetilde{U}^{(n)*}X_0\widetilde{V}^{(n)})\big)\widetilde{V}^{(n)*}\right\|_{S_2}^2 = \left\|H^{(n)}\circ(\widetilde{U}^{(n)^*}X_0\widetilde{V}^{(n)})\right\|_{S_2}^2\\ &= \left\|H^{(n)}\circ\begin{pmatrix}U_T^{(n)*}X_0V_T^{(n)} & U_T^{(n)*}X_0V_{T_c}^{(n)}\\ U_{T_c}^{(n)^*}X_0V_T^{(n)} & U_{T_c}^{(n)*}X_0V_{T_c}^{(n)}\end{pmatrix}\right\|_{S_2}^2\\ &= \left\|H_{T,T}^{(n)}\circ(U_T^{(n)*}X_0V_T^{(n)})\right\|_{S_2}^2 + \left\|H_{T,T_c}^{(n)}\circ(U_T^{(n)*}X_0V_{T_c}^{(n)})\right\|_{S_2}^2\\ &+ \left\|H_{T_c,T}^{(n)}\circ(U_{T_c}^{(n)*}X_0V_T^{(n)})\right\|_{S_2}^2 + \left\|H_{T_c,T_c}^{(n)}\circ(U_{T_c}^{(n)*}X_0V_{T_c}^{(n)})\right\|_{S_2}^2,\end{aligned} \tag{4.52}$$

using the notation from the proof of Lemma 4.17. To bound the first summand, we calculate

$$\begin{aligned}&\left\|H_{T,T}^{(n)}\circ(U_T^{(n)*}X_0V_T^{(n)})\right\|_{S_2} \le \left\|H_{T,T}^{(n)}\circ(U_T^{(n)*}X^{(n)}V_T^{(n)})\right\|_{S_2} + \left\|H_{T,T}^{(n)}\circ(U_T^{(n)*}(-\eta^{(n)})V_T^{(n)})\right\|_{S_2}\\ &\le \left\|H_{T,T}^{(n)}\circ\Sigma_T^{(n)}\right\|_{S_2} + \left\|H_{T,T}^{(n)}\circ(U_T^{(n)*}\eta^{(n)}V_T^{(n)})\right\|_{S_2}\\ &\le \left(\sum_{i=1}^r \frac{\sigma_i^2(X^{(n)})}{\left(\sigma_i^2(X^{(n)})+\epsilon^{(n)2}\right)^{2-p}}\right)^{1/2} + \max_{i,j=1}^r |H_{i,j}^{(n)}|\|U_T^{(n)*}\eta^{(n)}V_T^{(n)}\|_{S_2}\\ &\le \sqrt{r}\sigma_r^{p-1}(X^{(n)}) + (\sigma_r^2(X^{(n)})+\epsilon^{(n)2})^{\frac{p-2}{2}}\|U_T^{(n)*}\eta^{(n)}V_T^{(n)}\|_{S_2}\\ &\le \sqrt{r}\sigma_r^{p-1}(X^{(n)}) + \sigma_r^{p-2}(X^{(n)})\sqrt{r}\|\eta^{(n)}\|_{S_\infty} = \sqrt{r}\sigma_r^{p-2}(X^{(n)})\big[\sigma_r(X^{(n)}) + \|\eta^{(n)}\|_{S_\infty}\big],\end{aligned} \tag{4.53}$$

denoting $\Sigma_T^{(n)} = \mathrm{diag}(\sigma_i(X^{(n)}))_{i=1}^r$ and that the matrices $U_T^{(n)}$ and $V_T^{(n)}$ contain the first r left resp. right singular vectors of $X^{(n)}$ in the second inequality, together with the estimates $\|X\|_{S_1} \le \sqrt{r}\|X\|_{S_2} \le r\|X\|_{S_\infty}$ for $(r\times r)$-matrices X.

We recall the notations $s_r^0 = \sigma_r(X_0)$ and $s_1^0 = \sigma_1(X_0)$ and note that

$$\sigma_r(X^{(n)}) \ge s_r^0(1-\rho),$$

as the assumption (4.39) implies that

$$s_r^0 = \sigma_r(X_0) = \sigma_r(X^{(n)} - \eta^{(n)}) \le \sigma_r(X^{(n)}) + \sigma_1(\eta^{(n)}) \le \sigma_r(X^{(n)}) + \rho s_r^0,$$

using [8, Proposition 9.6.8] in the first inequality.

Therefore, we can bound the term of (4.53) such that

$$\left\| H^{(n)}_{T,T} \circ (U^{(n)*}_{T} X_0 V^{(n)}_{T}) \right\|_{S_2} \leq \sqrt{r}(s^0_r(1-\rho))^{p-2}[s^0_r(1-\rho) + \rho s^0_r] = \sqrt{r}(s^0_r)^{p-1}(1-\rho)^{p-2}. \tag{4.54}$$

For the second summand in the estimate of $\left\| \left[\widetilde{W}^{(n)}(X_0)_{\text{vec}}\right]_{\text{mat}} \right\|^2_{S_2}$, similar arguments and again assumption (4.39) are used to compute

$$\begin{aligned}
&\left\| H^{(n)}_{T,T_c} \circ (U^{(n)*}_{T} X_0 V^{(n)}_{T_c}) \right\|_{S_2} \leq \left\| H^{(n)}_{T,T_c} \circ (U^{(n)*}_{T} X^{(n)} V^{(n)}_{T_c}) \right\|_{S_2} + \left\| H^{(n)}_{T,T_c} \circ (U^{(n)*}_{T} \eta^{(n)} V^{(n)}_{T_c}) \right\|_{S_2} \\
&= \left\| H^{(n)}_{T,T_c} \circ (U^{(n)*}_{T} U^{(n)}_{T} \Sigma^{(n)}_{T} \overbrace{V^{(n)*}_{T} V^{(n)}_{T_c}}^{=0} + \overbrace{U^{(n)*}_{T} U^{(n)}_{T_c}}^{=0} \Sigma^{(n)}_{T_c} V^{(n)*}_{T_c} V^{(n)}_{T_c}) \right\|_{S_2} \\
&+ \left\| H^{(n)}_{T,T_c} \circ (U^{(n)*}_{T} \eta^{(n)} V^{(n)}_{T_c}) \right\|_{S_2} \\
&\leq \max_{i \in [r], j \in \{r+1,\ldots,d_2\}} |H^{(n)}_{i,j}| \| U^{(n)*}_{T} \eta^{(n)} V^{(n)}_{T_c} \|_{S_2} \leq 2\big[(\sigma_r(X^{(n)})^2 + \epsilon^{(n)2})^{\frac{2-p}{2}}\big]^{-1} \| U^{(n)*}_{T} \eta^{(n)} V^{(n)}_{T_c} \|_F
\end{aligned} \tag{4.55}$$

$$\begin{aligned}
&\leq 2\sigma_r(X^{(n)})^{p-2} \| U^{(n)*}_{T} \eta^{(n)} V^{(n)}_{T_c} \|_{S_2} \leq 2\sqrt{r}(s^0_r(1-\rho))^{p-2} \| \eta^{(n)} \|_{S_\infty} \\
&\leq 2\rho\sqrt{r}(s^0_r)^{p-1}(1-\rho)^{p-2}.
\end{aligned} \tag{4.56}$$

From exactly the same arguments it follows that also

$$\left\| H^{(n)}_{T_c,T} \circ (U^{(n)*}_{T_c} X_0 V^{(n)}_{T}) \right\|_{S_2} \leq 2\rho\sqrt{r}(s^0_r)^{p-1}(1-\rho)^{p-2}. \tag{4.57}$$

It remains to bound the last summand $\left\| H^{(n)}_{T_c,T_c} \circ (U^{(n)*}_{T_c} X_0 V^{(n)}_{T_c}) \right\|^2_{S_2}$. We see that

$$\begin{aligned}
&\left\| H^{(n)}_{T_c,T_c} \circ (U^{(n)*}_{T_c} X_0 V^{(n)}_{T_c}) \right\|_{S_2} \leq \max_{\substack{i \in \{r+1,\ldots,d_1\} \\ j \in \{r+1,\ldots,d_2\}}} \left| H^{(n)}_{i,j} \right| \| U^{(n)*}_{T_c} X_0 V^{(n)}_{T_c} \|_{S_2} \\
&\leq (\epsilon^{(n)})^{p-2} \| U^{(n)*}_{T_c} X_0 V^{(n)}_{T_c} \|_{S_2} \leq (\epsilon^{(n)})^{p-2} \| U^{(n)*}_{T_c} U^0_T \|_{S_\infty} \| S^0 \|_{S_2} \| V^{0*}_{T} V^{(n)}_{T_c} \|_{S_\infty} \\
&\leq (\epsilon^{(n)})^{p-2} \frac{\sqrt{2}\|\eta^{(n)}\|_{S_\infty}}{(1-\rho)s^0_r} \sqrt{r} s^0_1 \frac{\sqrt{2}\|\eta^{(n)}\|_{S_\infty}}{(1-\rho)s^0_r} = 2\sqrt{r}\|\eta^{(n)}\|^2_{S_\infty} (\epsilon^{(n)})^{p-2}(1-\rho)^{-2}(s^0_r)^{-1}\frac{s^0_1}{s^0_r}
\end{aligned} \tag{4.58}$$

where Hölder's inequality for Schatten norms was used in the third inequality. In the fourth inequality, Wedin's singular value perturbation bound of Lemma 2.19 is used with the choice $Z = X_0$, $\bar{Z} = X^{(n)}$, $\alpha = s^0_r$ and $\delta = (1-\rho)s^0_r$, and finally $\epsilon^{(n)} \leq \rho s^0_r$ in the last inequality, which is implied by the rule (4.16) for $\epsilon^{(n)}$ together with assumption (4.39).

Summarizing the estimates (4.54), (4.55), (4.57) and (4.58), we conclude that

$$
\begin{aligned}
&\left\| \left[\widetilde{W}^{(n)}(X_0)_{\text{vec}}\right]_{\text{mat}} \right\|_{S_2}^2 \\
&\leq r(s_r^0)^{2p-2}(1-\rho)^{2p-4}\left[1+8\rho^2+4\frac{\|\eta^{(n)}\|_{S_\infty}^4}{(1-\rho)^{2p}}(\epsilon^{(n)})^{2p-4}(s_r^0)^{-2p}\left(\frac{s_1^0}{s_r^0}\right)^2\right] \\
&= \frac{r(s_r^0)^{2p-2}}{(1-\rho)^4}\left[(1+8\rho^2)(1-\rho)^{2p}+4\frac{\|\eta^{(n)}\|_{S_\infty}^{4-2p}}{(\epsilon^{(n)})^{4-2p}}\frac{\|\eta^{(n)}\|_{S_\infty}^{2p}}{(s_r^0)^{2p}}\left(\frac{s_1^0}{s_r^0}\right)^2\right] \\
&\leq \frac{r(s_r^0)^{2p-2}}{(1-\rho)^4}\left[9+4\frac{\|\eta^{(n)}\|_{S_\infty}^{4-2p}}{(\epsilon^{(n)})^{4-2p}}\rho^{2p}\kappa^2\right] \leq \frac{13r(s_r^0)^{2p-2}}{(1-\rho)^4}\left[\frac{\|\eta^{(n)}\|_{S_\infty}^{4-2p}}{(\epsilon^{(n)})^{4-2p}}\kappa^2\right],
\end{aligned}
$$

as $0 < \rho < 1$, $\epsilon^{(n)} \leq \sigma_{r+1}(X^{(n)}) = \|X_{T_c}^{(n)}\|_{S_\infty} \leq \|\eta^{(n)}\|_{S_\infty}$ and using the assumption (4.39) in the second inequality. This concludes the proof of Lemma 4.18(i) together with inequality (4.51) as $13^{p/2} \leq 16^{p/2} = 4^p$.

(ii) For the second statement of Lemma 4.18, we proceed similarly as before, but note that by Hölder's inequality, also

$$\left\|\eta_{\text{vec}}^{(n+1)}\right\|_{\ell_2(\widetilde{W}^{(n)})}^2 \leq \left\|\left[\widetilde{W}^{(n)}(X_0)_{\text{vec}}\right]_{\text{mat}}\right\|_{S_1}\|\eta^{(n+1)}\|_{S_\infty},$$

cf. (4.51). Furthermore

$$
\begin{aligned}
\left\|\left[\widetilde{W}^{(n)}(X_0)_{\text{vec}}\right]_{\text{mat}}\right\|_{S_1} \leq{}& \left\|H_{T,T}^{(n)} \circ (U_T^{(n)*}X_0V_T^{(n)})\right\|_{S_1} + \left\|H_{T,T_c}^{(n)} \circ (U_T^{(n)*}X_0V_{T_c}^{(n)})\right\|_{S_1} \\
&+ \left\|H_{T_c,T}^{(n)} \circ (U_{T_c}^{(n)*}X_0V_T^{(n)})\right\|_{S_1} + \left\|H_{T_c,T_c}^{(n)} \circ (U_{T_c}^{(n)*}X_0V_{T_c}^{(n)})\right\|_{S_1}.
\end{aligned}
$$

The four Schatten-1 norms can then be estimated by $\max(r,(d-r))^{1/2}$ times the corresponding Schatten-2 norms. Using then again inequalities (4.54) – (4.58), we conclude the proof of (ii). □

Proof of Theorem 4.16. First we note that

$$\left(\sum_{i=r+1}^{d}\left(\sigma_i^2(X^{(n)})+\epsilon^{(n)2}\right)^{\frac{p}{2}}\right)^{2-p} \leq 2^{p-\frac{p^2}{2}}(d-r)^{2-p}\sigma_{r+1}(X^{(n)})^{p(2-p)} \tag{4.59}$$

as $\epsilon^{(n)} \leq \sigma_{r+1}(X^{(n+1)})$ due to the choice of $\epsilon^{(n)}$ in (4.16). We proceed by induction over $n \geq \bar{n}$. Lemmas 4.17(ii) and 4.18(ii) imply together with (4.59) that for $n = \bar{n}$,

$$
\begin{aligned}
\|\eta^{(n+1)}\|_{S_\infty}^p \leq \frac{\|\eta^{(n+1)}\|_{S_2}^{2p}}{\|\eta^{(n+1)}\|_{S_\infty}^p} &\leq 2^p\gamma_{2r}^{2-p}2^{p-\frac{p^2}{2}}\left(\frac{d-r}{r}\right)^{2-p/2}\frac{7^pr^p(s_r^0)^{p(p-1)}}{(1-\rho)^{2p}}\kappa^p\|\eta^{(n)}\|_{S_\infty}^{2p-p^2} \\
&\leq 2^{5p}\gamma_{2r}^{2-p}\left(\frac{d-r}{r}\right)^{2-p/2}\frac{r^p(s_r^0)^{p(p-1)}}{(1-\rho)^{2p}}\kappa^p\|\eta^{(n)}\|_{S_\infty}^{p(2-p)}
\end{aligned}
\tag{4.60}
$$

as $\sigma_{r+1}(X^{(n)}) = \epsilon^{(n)}$ by assumption for $n = \bar{n}$.

Similarly, by Lemmas 4.17(iii) and 4.18(ii) and (4.59), the error in the Schatten-p quasinorm fulfills

$$\|\eta^{(n+1)}\|_{S_p}^{2p} \leq (1+\gamma_{2r})^2 2^{2+2p} \big(d-r\big)^{2-p} \frac{r^{p/2}(s_r^0)^{p(p-1)}}{(1-\rho)^{2p}} \kappa^p \|\eta^{(n)}\|_{S_\infty}^{p(2-p)} \|\eta^{(n+1)}\|_{S_2}^p \tag{4.61}$$

for $n = \bar{n}$. Using the strong Schatten-p null space property of order $2r$ for the operator Φ, we see from the arguments of (4.49) that

$$\|\eta^{(n)}\|_{S_\infty}^p \leq \|\eta^{(n)}\|_{S_2}^p \leq \frac{2^{p-1}\gamma_{2r}^{1-p/2}}{r^{1-p/2}} \|\eta^{(n)}\|_{S_p}^p$$

and also $\|\eta^{(n+1)}\|_{S_2}^p \leq \frac{2^{p-1}\gamma_{2r}^{1-p/2}}{r^{1-p/2}} \|\eta^{(n+1)}\|_{S_p}^p$. Inserting that in (4.61) and dividing by $\|\eta^{(n+1)}\|_{S_p}^p$, we obtain

$$\|\eta^{(n+1)}\|_{S_p}^p \leq 2^{4p}(1+\gamma_{2r})^2 \gamma_{2r}^{2-p} \Big(\frac{d-r}{r}\Big)^{2-p} \frac{r^{p/2}(s_r^0)^{p(p-1)}}{(1-\rho)^{2p}} \kappa^p \|\eta^{(n)}\|_{S_\infty}^{p(1-p)} \|\eta^{(n)}\|_{S_p}^p.$$

Under the assumption that (4.40) holds, it follows from this and (4.60) that

$$\|\eta^{(n+1)}\|_{S_\infty}^p \leq \|\eta^{(n)}\|_{S_\infty}^p \text{ and } \|\eta^{(n+1)}\|_{S_p}^p \leq \|\eta^{(n)}\|_{S_p}^p \tag{4.62}$$

for $n = \bar{n}$, which also entails the statement of Theorem 4.16 for this iteration.

Let now $n' > \bar{n}$ be such that (4.62) is true for all n with $n' > n \geq \bar{n}$. If $\sigma_{r+1}(X^{(n')}) \leq \epsilon^{(n'-1)}$, then $\epsilon^{(n')} = \sigma_{r+1}(X^{(n')})$ and the arguments from above show (4.62) also for $n = n'$.

Otherwise, it holds that $\sigma_{r+1}(X^{(n')}) > \epsilon^{(n'-1)}$ and there exists $n' > n'' \geq \bar{n}$ such that $\epsilon^{(n')} = \epsilon^{(n'')} = \sigma_{r+1}(X^{(n'')})$. Then

$$\begin{aligned} &\|\eta^{(n'+1)}\|_{S_\infty}^p \\ &\leq 2^p \frac{\gamma_{2r}^{2-p}}{r^{2-p}} \Bigg(\sum_{i=r+1}^{d} \Big(\frac{\sigma_i^2(X^{(n')})}{\epsilon^{(n'')2}} + 1 \Big)^{\frac{p}{2}} \Bigg)^{2-p} \frac{7^p r^{p/2} \max(r, d-r)^{p/2} (s_r^0)^{p(p-1)}}{(1-\rho)^{2p}} \kappa^p \|\eta^{(n')}\|_{S_\infty}^{p(2-p)} \end{aligned}$$

and we compute

$$\Big(\sum_{i=r+1}^{d}\Big(\frac{\sigma_i^2(X^{(n')})}{\epsilon^{(n'')2}}+1\Big)^{\frac{p}{2}}\Big)^{2-p} \leq \Big(\sum_{i=r+1}^{d}\frac{\sigma_i^p(X^{(n')})}{\epsilon^{(n'')p}}+(d-r)\Big)^{2-p}$$
$$\leq \Big(\frac{\|\eta^{(n')}\|_{S_p}^p}{\epsilon^{(n'')p}}+(d-r)\Big)^{2-p} \leq \Big(\frac{\|\eta^{(n'')}\|_{S_p}^p}{\epsilon^{(n'')p}}+(d-r)\Big)^{2-p}$$
$$\leq \Big(\frac{2(1+\gamma_{2r})\|X_{T_c}^{(n'')}\|_{S_p}^p}{(1-\gamma_{2r})\epsilon^{(n'')p}}+(d-r)\Big)^{2-p} \leq \Big(\frac{3+\gamma_{2r}}{1-\gamma_{2r}}\Big)^{2-p}(d-r)^{2-p},$$

using that X_0 is a matrix of rank at most r in the second inequality, the inductive hypothesis in the third and an analogue of Lemma 4.13 for a Schatten-p quasinorm on the left hand side (cf. [81, Lemma 3.2] for the corresponding result for $p = 1$). The latter argument uses the assumption on the null space property. This shows that

$$\|\eta^{(n'+1)}\|_{S_\infty}^p \leq \mu\|\eta^{(n')}\|_{S_\infty}^{p(2-p)}$$

for

$$\widetilde{\mu} := 2^{4p}\gamma_{2r}^{2-p}\Big(\frac{(3+\gamma_{2r})(d-r)}{(1-\gamma_{2r})r}\Big)^{2-p}\frac{r^{p/2}(s_r^0)^{p(p-1)}}{(1-\rho)^{2p}}\kappa^p\max\Big(2^p(d-r)^{\frac{p}{2}},(1+\gamma_{2r})^2\Big), \tag{4.63}$$

and $\|\eta^{(n'+1)}\|_{S_\infty}^p \leq \|\eta^{(n')}\|_{S_\infty}^p$ under the assumption (4.40) of Theorem 4.16, as $\widetilde{\mu} \leq \mu$ with μ as in (4.41). Indeed $\widetilde{\mu} \leq \mu$ since

$$\max\Big(2^p(d-r)^{\frac{p}{2}},(1+\gamma_{2r})^2\Big)\Big(\frac{d-r}{r}\Big)^{2-p}r^{p/2} \leq 2^p(1+\gamma_{2r})^2\Big(\frac{d-r}{r}\Big)^{2-p/2}r^p.$$

The same argument shows that $\|\eta^{(n'+1)}\|_{S_p}^p \leq \|\eta^{(n')}\|_{S_p}^p$, which finishes the proof. □

Proof of Theorem 4.15(iii). The statement follows from Theorem 4.16, since for $\widetilde{r} \leq r$, an operator Φ fulfilling the Schatten-p NSP of order $2r$ with constant $\gamma_{2r} < 1$ trivially fulfills the Schatten-p NSP of order $2\widetilde{r}$ with constant $\gamma_{2\widetilde{r}} \leq \gamma_{2r} < 1$. □

Remark 4.19. We note that the left- and right-sided weight matrices of previous IRLS approaches [51, 106] at iteration n could be expressed in our notation as

$$\mathbf{I}_{d_2}\otimes W_L^{(n)} := \mathbf{I}_{d_2}\otimes U^{(n)}(\bar{\Sigma}_{d_1}^{(n)})^{p-2}U^{(n)*}$$

and

$$W_R^{(n)}\otimes\mathbf{I}_{d_1} := V^{(n)}(\bar{\Sigma}_{d_2}^{(n)})^{p-2}V^{(n)*}\otimes\mathbf{I}_{d_1},$$

respectively.

Let

$$T^{(n)} := \operatorname{span}\left\{u_i^{(n)} y^*, x v_i^{(n)*} \mid x \in \mathbb{R}^{d_1}, y \in \mathbb{R}^{d_2}, i \in [r]\right\},$$

where $u_1^{(n)}, \ldots, u_r^{(n)}$ resp. $v_1^{(n)}, \ldots, v_r^{(n)}$ are the first r left and right singular vectors of $X^{(n)}$. $T^{(n)}$ is a space that can be considered as a *generalized support* of the best rank-r approximation of $X^{(n)}$.

With this remark, we want to give an explanation for the fact that left- or right-sided weight matrices do not lead to algorithms with superlinear convergence rates for $p < 1$. This argument will be based on the observation that there are always parts of the space $T^{(n)}$ that are equipped with too large weights if $X^{(n)} = U^{(n)}\Sigma^{(n)}V^{(n)*}$ is already approximately low-rank. In particular, proceeding as in (4.52), we obtain for $\mathbf{I}_{d_2} \otimes W_L^{(n)}$

$$\begin{aligned} \left\| \left[\left(\mathbf{I}_{d_2} \otimes W_L^{(n)}\right)(X_0)_{\text{vec}}\right]_{\text{mat}} \right\|_{S_2}^2 = {} & \left\| \left(\bar{\Sigma}_T^{(n)}\right)^{p-2} U_T^{(n)*} X_0 V_T^{(n)} \right\|_{S_2}^2 + \left\| \left(\bar{\Sigma}_T^{(n)}\right)^{p-2} U_T^{(n)*} X_0 V_{T_c}^{(n)} \right\|_{S_2}^2 \\ & + \left\| \left(\bar{\Sigma}_{T_c}^{(n)}\right)^{p-2} U_{T_c}^{(n)*} X_0 V_T^{(n)} \right\|_{S_2}^2 + \left\| \left(\bar{\Sigma}_{T_c}^{(n)}\right)^{p-2} U_{T_c}^{(n)*} X_0 V_{T_c}^{(n)} \right\|_{S_2}^2 \end{aligned}$$

if $\bar{\Sigma}_T^{(n)}$ denotes the diagonal matrix with the first r non-zero entries of $\bar{\Sigma}_{d_1}^{(n)}$ and $\bar{\Sigma}_{T_c}^{(n)}$ the one of the remaining entries.

Here, the third of the four summands would become too large for $p < 1$ to allow for a superlinear convergence when the last $d - r$ singular values of $X^{(n)}$ approach zero. An analogous argument can be used for the right-sided weight matrix $W_R^{(n)} \otimes \mathbf{I}_{d_1}$ and, notably, also for arithmetic mean weight matrices $W_{(\text{arith})}^{(n)} = \mathbf{I}_{d_2} \otimes W_L^{(n)} + W_R^{(n)} \otimes \mathbf{I}_{d_1}$, cf. section 4.1.2.

4.3.7 Discussion and comparison with existing IRLS algorithms

Optimally, one could ask for a statement in Theorem 4.15 about the accumulation points $\bar{X}$ being *global minimizers* of f_ϵ, instead of mere stationary points, cf. [51, Theorem 6.11], [35, Theorem 5.3]. As we will see in the next section, numerical experiments indicate that at least empirically the recovery of global minimizers for a large number of problem instances in the matrix completion context is achieved. Due to the strong nonconvexity of the Schatten-p quasinorm and of the ϵ-perturbed version f_ϵ for small ranges of p, such a strong theoretical statement is unfortunately difficult to prove.

Nevertheless, our results can be interpreted as analogues of the results in [35, Theorem 7.7], which discusses the convergence behaviour of an `IRLS` algorithm for the sparse vector case based on ℓ_p-minimization with $p < 1$.

As already mentioned, one can view the algorithm of [51] as an asymmetric variant of `HM-IRLS` with parameter choice $p = 1$ in our notation and under this point of view, our result Theorem 4.15 recovers the results of [51, Theorem 6.11(i-ii)] for $p = 1$ and provides a generalization, although with weaker conclusions due to the non-convexity, also to the cases $0 < p < 1$. Non-convex choices $0 < p < 1$ have been considered in [106] for the algorithm `IRLS-`p, that is very similar to the one in [51]. However, the convergence result [106, Theorem 5.1] in the non-convex case corresponds to Theorem 4.15(ii) but does not give statements similar to (i) and (iii) of Theorem 4.15.

To the best of our knowledge, the convergence rate result in Theorem 4.16 is new in the sense that so far,in the literature, there are no convergence rate proofs for IRLS algorithms for the low-rank matrix recovery problem. In fact, Remark 4.19 provides an argument why it is not possible for existing `IRLS`-variants of [51] and [106] to exhibit superlinear convergence rates, unlike `HM-IRLS`.

Finally, let us point out the close connection between the statements of Theorems 4.15 and 4.16 and results presented for `IRLS` in the context of the sparse vector recovery problem [35, Theorems 7.7 and 7.9].

4.4 Numerical experiments

In this section, we verify numerically the theoretically predicted superlinear convergence rate for Algorithm 4 (HM-IRLS) in Theorem 4.16 even for relevant measurement operators not fulfilling theoretical requirements, more precisely the strong null space property.

In particular, also for the important framework of matrix completion, HM-IRLS exhibits a superlinear convergence rate of order $2 - p$ that can be observed very clearly in the experiment results reported in subsection 4.4.2. Moreover, we report that for other IRLS-type algorithms as in [51, 106], and a variant implementing arithmetic mean weight matrices (see Lemma 4.2) instead of harmonic mean weight matrices such a superlinear rate of any form was *not* observed in our numerical tests.

Beyond that, we study in our experiments the recovery performance of HM-IRLS as well as of other algorithms for the matrix completion measurement setting. We compare them with a focus on the measurement complexities necessary for successful recovery for a large number of random problem instances in subsection 4.4.3. The methods covered in this comparison to HM-IRLS include not only variants of IRLS, but also other types of cutting-edge low-rank matrix recovery approaches. Interestingly, our numerical tests reveal that even for cases where the oversampling factor is very low ($\rho \approx 1$), HM-IRLS is able to recover the desired low-rank matrix, without requesting a special initialization, although the underlying recovery problem is severely non-convex. In particular, HM-IRLS *recovers low-rank matrices systematically with nearly the optimal number of measurements and needs fewer measurements than all the state-of-the-art algorithms, including previously existing IRLS methods, involved in our experiments.*

All numerical experiments discussed in this section are performed on a MacBook Pro 9.1 with a 2.6 GHz Intel Core i7 quad-core-processor and 8GB memory. Computations are run in MATLAB R2014a, version 8.3.0.532. An implementation of the HM-IRLS algorithm and a minimal test example are available at https://www-m15.ma.tum.de/Allgemeines/SoftwareSite.

4.4.1 Measurement setting

In our experiments, we consider $d_1 \times d_2$ low-rank matrices X_0 of $\operatorname{rank}(X_0) = r$, which we construct by the multiplication of matrices $U\Sigma V$, where $U \in \mathbb{R}^{d_1 \times r}$, $V \in \mathbb{R}^{r \times d_2}$ are matrices with i.i.d. standard Gaussian entries and $\Sigma \in \mathbb{R}^{r \times r}$ is diagonal with i.i.d. Gaussian entries as well.

As noted in Lemma 2.20, a low-rank matrix $X \in M_{d_1 \times d_2}$ of rank r has $d_f = r(d_1+d_2-r)$ degrees of freedom, which corresponds to the theoretical lower bound on the number of measurements necessary for exact recovery [21].

We give a detailed description of the random measurement setting for our experiments. We will consider the matrix completion framework, choosing $m = \lfloor \rho \cdot d_f \rfloor$ entries of X_0 uniformly over its $d_1 \cdot d_2$ indices to get our measurement result $Y = \Phi(X_0)$. The so-called oversampling factor $\frac{d_1 d_2}{d_f} \geq \rho \geq 1$ regulates the hardness of the recovery problem.

However, a sampling scheme as just described above can yield instances of measurement maps Φ with insufficient information content to guarantee the well-posedness of the corresponding low-rank matrix recovery problem, even for the cases where $\rho > 1$. To be more precise, if the number of sampled entries in any row or column is below its rank r it is impossible to recover a matrix exactly. A more detailed explanation and proof can be found in the context of [119, Theorem 1].

Therefore, the uniform sampling model is adapted in such a way that measurement operators Φ are excluded and generated again until the requirement of minimium r entries per column/row is met. Thereby, we ensure that reconstruction is possible from a theoretical point of view.

The phenomenon just described above is closely related to the fact that recovery guarantees for matrix completion for the uniform sampling model require at least one additional log factor, which means that at least $m \geq \log(\max(d_1, d_2))d_f$ sampled entries are required. [37, Section V].

Although we only present experiments for the matrix completion setting in this section, we point out that also in the case of Gaussian measurement models we obtain similar results in numerical tests.

4.4.2 Convergence rate comparison with other IRLS type algorithms

In the following, we compare the `HM-IRLS` algorithm to existing variants of `IRLS` for low-rank matrix recovery that only employ reweighting in the column space as presented in [51] (IRLS-M) called `IRLS-FRW`, and the strongly related version in[106] (IRLS-p) denoted by `IRLS-MF`. Additionally, we consider the performance comparison of `HM-IRLS` with an alternative method incorporating reweightings in both the row and column space: we add to our list of test algorithms an arithmetic mean iteratively reweighted least squares (`AM-IRLS`), which employs a weight matrix composed according to the arithmetic mean

$$\widetilde{W}^{(n)} = \frac{1}{2}\left[\left(U^{(n)}S^{(n)}S^{(n)*}U^{(n)*} + \epsilon^{(n)2}\mathbf{I}_{d_1}\right)^{\frac{p-2}{2}} \oplus \left(V^{(n)}S^{(n)*}S^{(n)}V^{(n)*} + \epsilon^{(n)2}\mathbf{I}_{d_2}\right)^{\frac{p-2}{2}}\right]$$

of left- and right-sided weight matrices, where $X^{(n)} = U^{(n)}S^{(n)}V^{(n)*}$ is the full SVD of the iterate $X^{(n)}$. We refer to subsection 4.1.2 for the introduction of the weight matrix used in `AM-IRLS`.

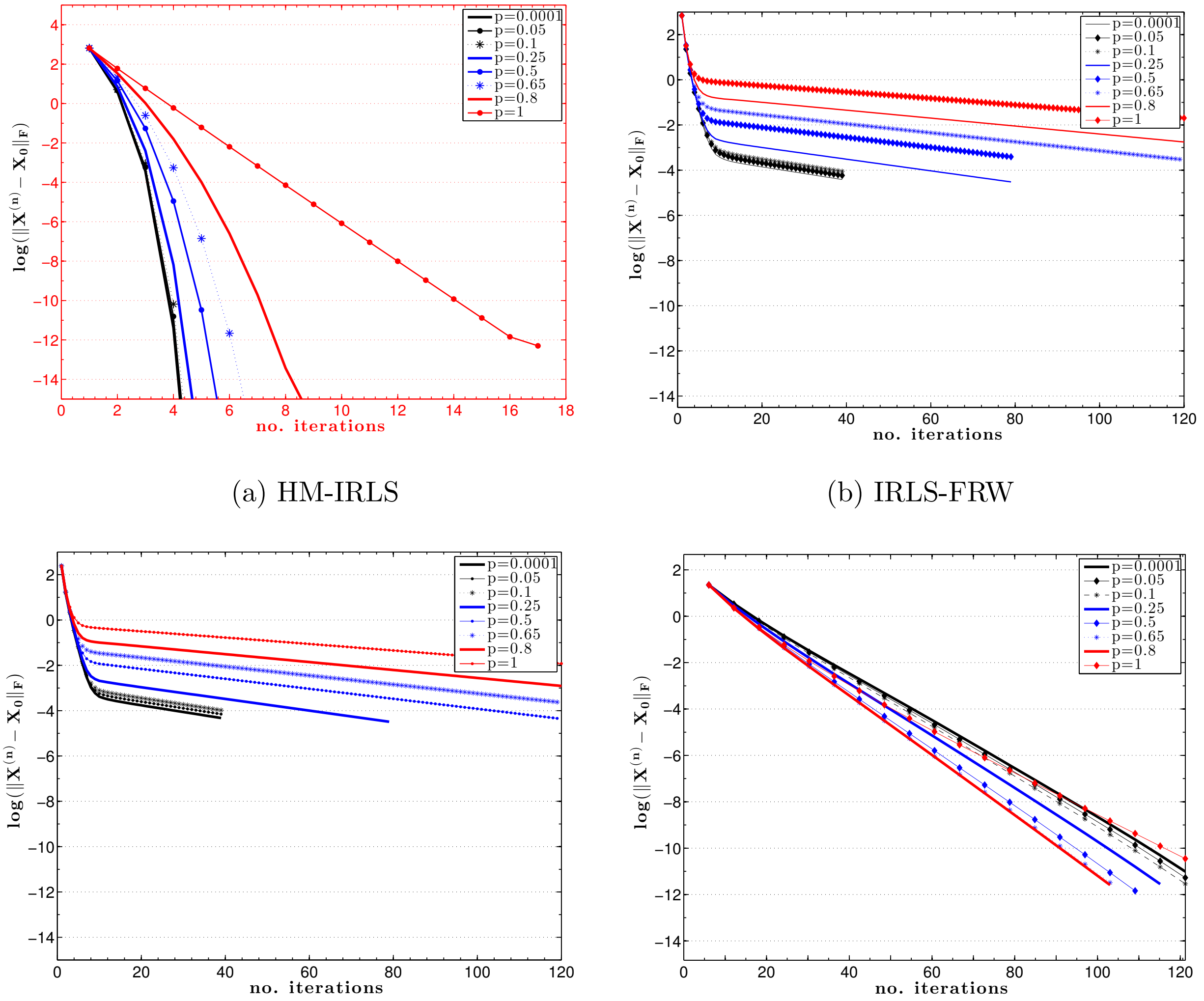
With our present experiments we aim at the examination of the convergence behaviour, in particular the convergence speeds of the four test algorithms for hard and easy matrix completion problems. In our numerical tests, we set the dimension of the solution matrix to $d_1 = d_2 = 40$, rank $r = 10$, oversampling factor $\rho = 2.0, 1.2, 1.0$ and consider the instances of the random model as explained in subsection 4.4.1. The experiments are performed for non-convexity parameters $p = \{0.0001, 0.05, 0.1, 0.25, 0.5, 0.65, 0.8, 1.0\}$ appearing in the Schatten-p minimization problems (4.5), which all variants of the `IRLS` algorithm strive to solve.

The plots in Figures 4.1 to 4.3 show the behaviour of the Frobenius error in logarithmic scale $\log(\|X^{(n)} - X_0\|_F)$ for the iterations $n = 1, \dots, \bar{n}$ of the listed algorithms, where we denote with $\bar{n}$ the first iteration at which the Frobenius error falls below a certain tolerance level or a maximum number of iterations $n_{\max}$ is exceeded.

4.4.2.1 Results for `HM-IRLS`

It can be verified from inspecting Figure 4.1a that, for parameters $p < 1$, `HM-IRLS` exhibits superlinear rates of convergence very accurately of the orders $2-p$ as theoretically predicted by Theorem 4.16. As a consequence, we observe a dramatic enhancement of the convergence rate *from linear to arbitrarily close to quadratic* for p tending from 1 to 0. When decreasing the oversampling factor ρ from $\rho = 2$ in Figures 4.1 to $\rho = 1.2$ in Figure 4.2 and eventually even to $\rho = 1$ in Figure 4.3, which corresponds to the increase of the hardness of the matrix completion problem, we observe the divergence of the `HM-IRLS` algorithm for larger values $p \gg 0$. This behaviour is very predictable, as it is known for nuclear norm minimization to fail at the recovery of low-rank matrix in cases where the oversampling factor ρ is getting close to 1 and, for the parameter choice $p = 1$, `HM-IRLS` just approximates `NNM`.

On the other hand, we get an interesting result illustrated by Figures 4.2a and 4.3: with the choice of p close to 0, even for the very difficult reconstruction problems with very low sample complexities $\rho = 1.2$ and $\rho = 1$, the `HM-IRLS` algorithm is able to successfully recover the low-rank matrix, still performing with a convergence rate of



(a) HM-IRLS

(b) IRLS-FRW

(c) IRLS-MF

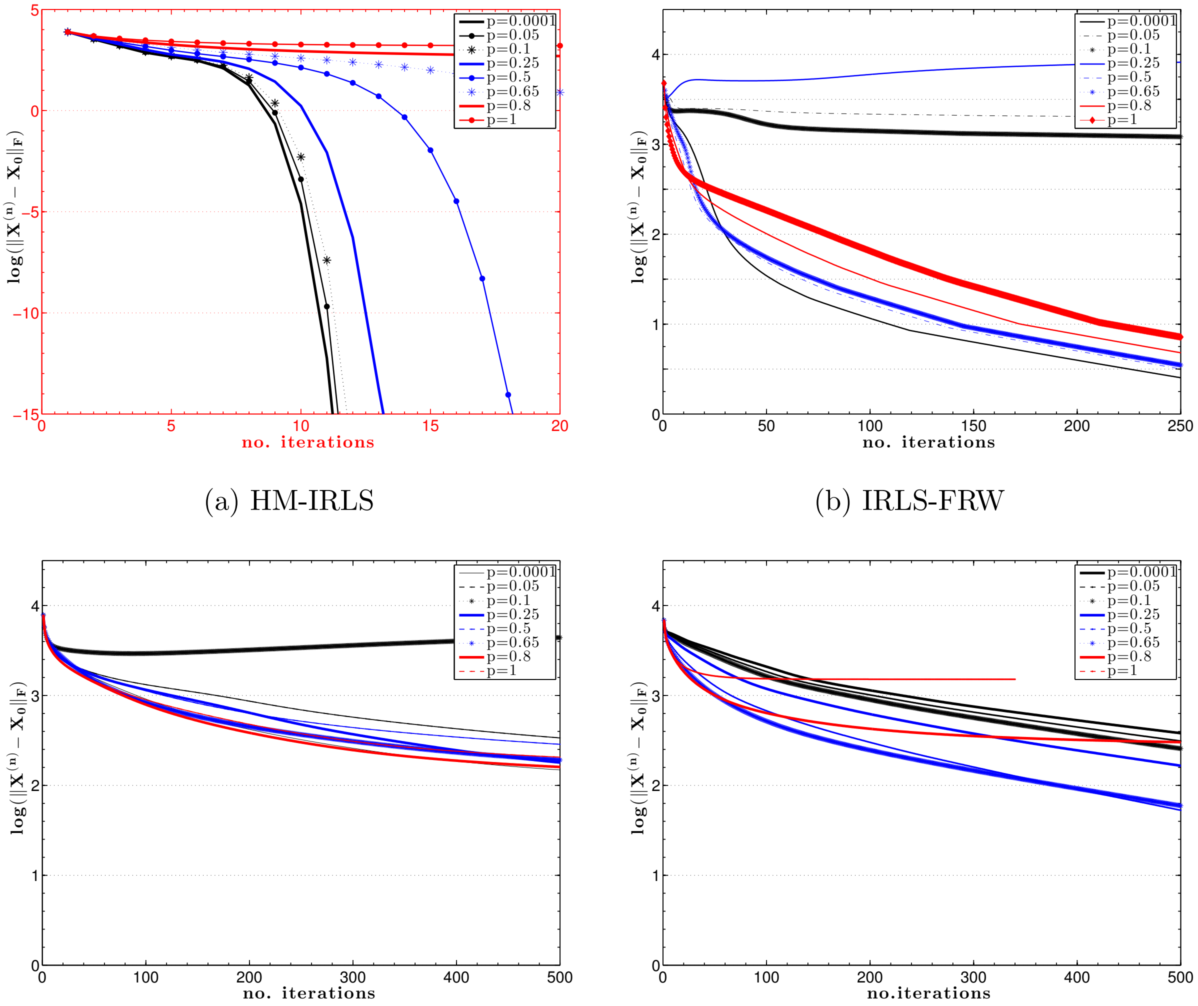
(d) AM-IRLS

Figure 4.1: Behaviour of the log error $\log(\|X^{(n)} - X_0\|_F)$ for successive iterations for different parameter values of $0 < p \leq 1$ and fixed measurement oversampling factor $\rho = 2.0$. Note that the x-axis in (a) has a different scaling, indicating a much faster convergence of `HM-IRLS`.

order $2 - p$.

4.4.2.2 Results for other `IRLS`-type algorithms for low rank matrix recovery

For the other variants `IRLS-FRW`, `IRLS-MF` and `AM-IRLS`, we observe a contrasting algorithmic behaviour. Figures 4.2b–4.2d show that these methods do not converge to the ground truth low-rank matrix X_0 for hard reconstructions problems with low sampling complexity rates $\rho = 1.2$ and $\rho = 1.0$, regardless of the choice of the parameter p. The corresponding plots for $\rho = 1.0$ are omitted as in these cases a lack of convergence is observed as well.



(a) HM-IRLS

(b) IRLS-FRW

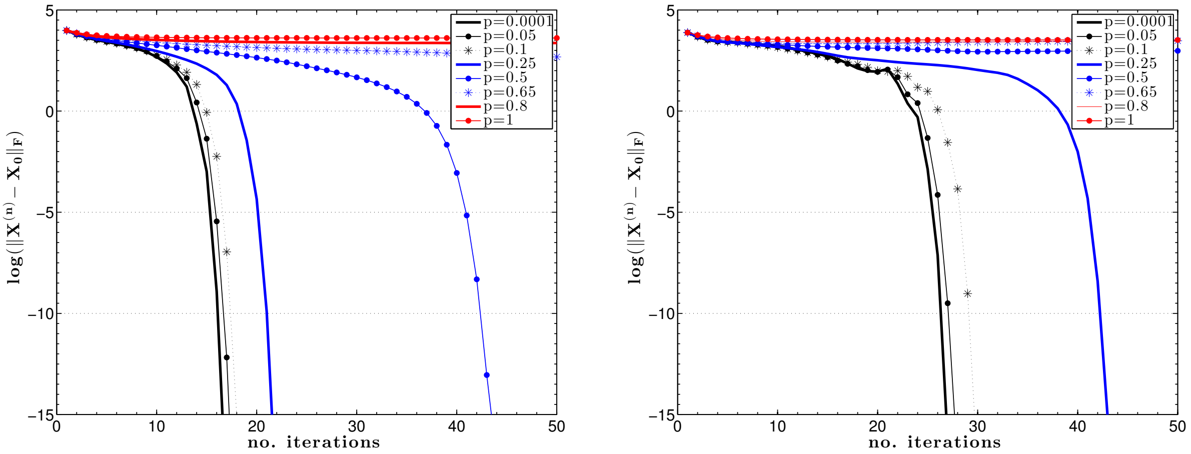
(c) IRLS-MF

(d) AM-IRLS

Figure 4.2: Behaviour of the log error $\log(\|X^{(n)} - X_0\|_F)$ for successive iterations for different parameter values of $0 < p \leq 1$ and fixed measurement oversampling factor $\rho = 1.2$. Note that the x-axis in (a) has a different scaling, indicating a much faster convergence of `HM-IRLS`.

We observe convergence of the mentioned methods for easier matrix completion problems corresponding to $\rho = 2.0$ as shown in Figure 4.1b–4.1d. Nevertheless in these cases at best they exhibit a *linear* rate of convergence only, also if the parameter p is chosen significantly smaller than 1. If we provide such a generous amount of measurements, we find that `IRLS-FRW` and `IRLS-MF` show slightly faster convergence for p approaching 0 and only for `AM-IRLS` larger values of p yield more promising results (cf. again Figure 4.1b–4.1d).

For the oversampling factor $\rho = 1.2$, which gives intermediate difficulty level for reconstruction, the methods `IRLS-FRW`, `IRLS-MF` and `AM-IRLS` become unstable very easily



(a) HM-IRLS

(b) HM-IRLS

Figure 4.3: Behaviour of the log error $\log(\|X^{(n)} - X_0\|_F)$ for successive iterations for different parameter values of $0 < p \leq 1$ for measurement oversampling factor $\rho = 1.0$, two different instances of the measurement model of section 4.4.1.

for small choices of p. Therefore, we pose the conjecture that there exists an optimal value $0 < p_{\text{opt}} < 1$ for each algorithm, that we do not investigate further.

4.4.2.3 `HM-IRLS` as the best extension of IRLS for sparse recovery

As a summary of the experiments above, we can state that among the four variants `HM-IRLS`, `IRLS-FRW`, `IRLS-MF` and `AM-IRLS`, only `HM-IRLS` is capable of solving the low-rank matrix recovery problem for very low oversampling factor $\rho \approx 1$. Additionally, `HM-IRLS` is the only `IRLS` algorithm for low-rank matrix recovery, which showcases a superlinear convergence rate at all.

We consider it just as interesting to compare the algorithmic behaviour of `HM-IRLS` with the properties of the `IRLS` algorithm of [35] for sparse vector recovery, which mimics the ℓ_p-minimization for $0 < p \leq 1$. The superlinear convergence rate of the algorithm in [35] as illustrated in Figure 8.3 in [35] could not be generalized to the low-rank matrix recovery problem by any of the versions `IRLS-FRW`, `IRLS-MF` or `AM-IRLS`, as obvious from Figures 4.1a, 4.2a and 4.3.

Taking the theoretical guarantees as well as the numerical evidence into account, we claim that *`HM-IRLS` is the best extension of IRLS for vector recovery [35] to the low-rank matrix recovery setting*, providing a substantial improvement over the reweighting strategies of [51, 106].

We even go a step further, by pointing out two observations which suggest that `HM-IRLS` in some sense even exhibits more favorable properties than the version of `IRLS` for the vector case in [35]:

(i) First, the superlinear convergence in the vector case is only observable locally after a considerable number of iterations with a linear error decay have been executed as discussed in [35, Section 8]. In contrast to such behaviour, for our algorithm `HM-IRLS`, superlinear error decay can be observed quite early (i.e., for example even as early as after 2 or 3 iterations), at least in cases where a large enough oversampling factor can be provided, cf. Figure 4.1a.

(ii) Second, in cases where $p < 0.5$, we observe a loss of global convergence of the algorithm in [35] [35, Section 8]. In contrast to that, the `HM-IRLS` algorithm does not suffer from this convergence breakdown for $p \ll 0.5$. Consequently, we suggest the choice of very small parameters $p \leq 0.1$ in order to achieve very fast convergence, cf. Figure 4.3.

4.4.3 Recovery performance comparison with state-of-the-art algorithms

After the performance comparison of `HM-IRLS` with other `IRLS`-type methods, we extend the list of our test algorithms in our experiments with strategies different from `IRLS`.

In order to provide a comprehensive picture in our numerical tests, we involve a broad variety of state-of-the-art algorithms in the experiments: from the already studied `IRLS` algorithms `IRLS-FRW`, `IRLS-MF` [51, 106], over Riemannian optimization techniques `Riemann_Opt` of [154], alternating minimization approaches `p_MC_AltMin`, `ASD` and `BFGD` of [74, 118, 146], finally to iterative thresholding-based methods such as `MatrixALPSII`, `CGIHT_Matrix` in [9, 85].

Our goal in the following experiments is to systematically study the empirical recovery probabilities of the different algorithms for varying sample complexities $m = \lfloor \rho \cdot d_f \rfloor$, parametrized by the oversampling factor ρ, which regulates the hardness of the low-rank recovery problem. Here, a large parameter ρ corresponds to an easy reconstruction problem, and a small value of ρ, e.g., $\rho \approx 1$ describes very hard problems.

Again, we adopt the matrix completion measurement setting as explained in subsection 4.4.1, setting the dimensions of the ground truth matrices X_0 to $d_1 = d_2 = 100$ and the rank to $r = 8$. We now randomly sample 150 instances of X_0 and Φ for different numbers of measurements increasing from $m_{\min} = 1500$ to $m_{\max} = 4000$, which corresponds to a growing oversampling factor ρ from $\rho_{\min} = 0.975$ to $\rho_{\max} = 2.60$.

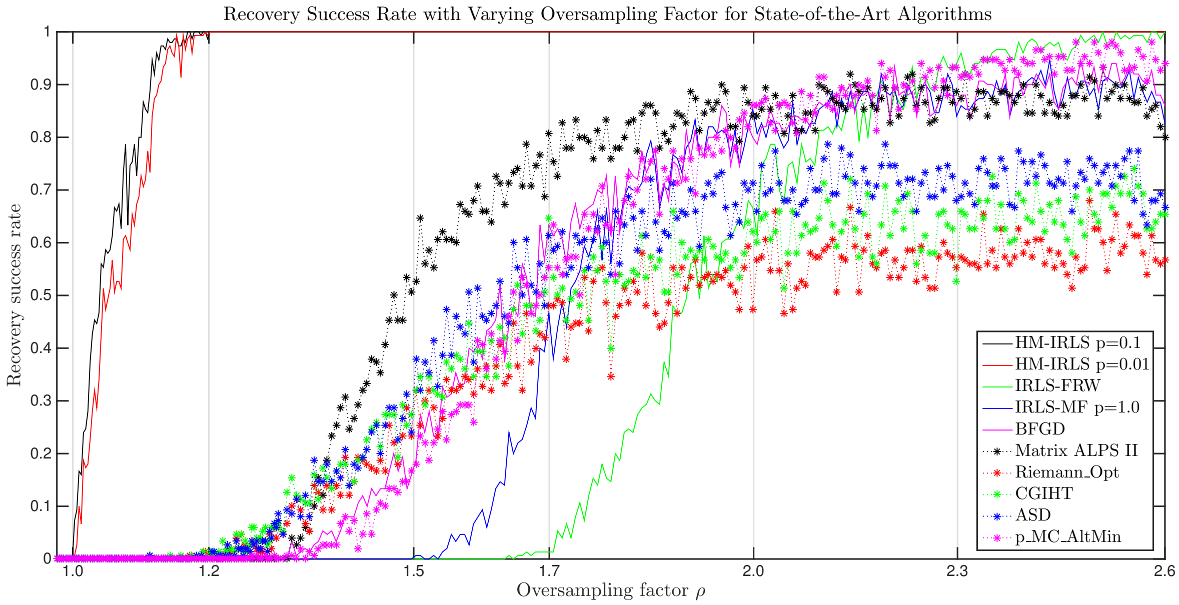


Figure 4.4: Recovery success rate with varying oversampling factor ρ for state-of-the-art-algorithms

We define an attempt for the recovery of X_0 as successful if the relative Frobenius error $\|X^{\text{out}} - X_0\|_F/\|X_0\|_F$ for the output matrix X^{out} is smaller than 10^{-3}. We run the different algorithms either until convergence or until a maximal number of iterations $n_{\max}$ (e.g. $n_{\max} = 10000$) is exceeded, where $n_{\max}$ is chosen generously large enough to guarantee that a recovery failure is not caused by a lack of iterations.

For the conduction of our experiments, we employ implementations that were made available by the authors of the corresponding research papers for our comparison algorithms, setting the default input parameters provided with the authors' software packages. We collected the respective code sources in the reference section.

4.4.3.1 Beyond the state-of-the-art performance of HM-IRLS

Surprisingly, the experiment results displayed in Figure 4.4 reveal that HM-IRLS reaches a very high empirical recovery probability for parameters $p = 0.1$ and $p = 0.01$ as soon as the oversampling factor ρ is larger than 1.0. This implies that for the recovery of $(d_1 \times d_2)$-dimensional rank-r matrices a number of $m = \lfloor \rho r(d_1 + d_2 - r) \rfloor$ measurements with $\rho \approx 1$ is already sufficient, which is extremely close to the information theoretical lower bound of $d_f = r(d_1 + d_2 - r)$. Moreover, we report the interesting fact, that already for a measurement complexity factor of $\rho \approx 1.15$ HM-IRLS achieves an empirical recovery percentage of 100%.

For all algorithms tested, we make the contrasting observation that they basically always fail to perform recovery for any rank-r matrix if $\rho < 1.2$, and in most cases need an oversampling factor of $\rho > 1.7$ to exceed an empirical recovery success rate of a mere 50%. Only after rising ρ above 2.0, a recovery percentage of nearly 80% is approached at least for a subset of the comparison algorithms, more precisely for `Matrix ALPS II, BFGD, p_MC_AltMin` and the existing `IRLS` approaches `IRLS-FRW` and `IRLS-MF`. Even for large oversampling factors up to $\rho = 2.5$, all other competing algorithms are incapable of systematically achieving the empirical probability of 100%. Although we do not rule out that possible parameter tuning can slightly enhance the recovery performance of any of the other algorithms, we report that, for hard matrix completion problems, the experimental evidence for the very significant gap in the recovery performance of `HM-IRLS` in comparison with all other methods is striking.

Hence, our observations can be summarized as follows: for the choice of the non-convexity parameter $p \ll 1$ the proposed `HM-IRLS` algorithm is able to *recover low-rank matrices systematically with nearly the optimal number of measurements and needs fewer measurements than all the state-of-the-art algorithms included in our experiments.*

In Figure 4.4 a very sharp phase transition between failure and success of recovery can be observed very clearly for `HM-IRLS`. This indicates that the oversampling factor ρ indeed plays a major role for the determination of the success of `HM-IRLS`. In contrast, for all other algorithms we have wider phase transitions suggesting the existence of further influence factors such as the realizations of the random sampling model or possible interactions between the measurement operator Φ and solution matrix X_0.

A last important conclusion that we draw from the very high empirical recovery probability of 100% in those cases where the sample complexity factor ρ is large enough, is that local minimizers are not an issue for `HM-IRLS`, but it always converges to the global minimizer, although the underlying Schatten-p quasinorm for, e.g., $p = 0.01$, is severely nonconvex.

Therefore, we conclude that the initialization of $X^{(1)}$ as the Frobenius norm minimizer that we choose can already ensure global convergence. In contrast, other non-convex low-rank recovery methods might show a heavy dependence on a smartly chosen starting point. In this point of view, our experimental results indicate that the non-convex low-rank matrix recovery algorithms included in our tests do not seem to be able to capture the desired basin of attraction of the global minimum in many cases if the sample complexity is low (i.e., if ρ is small). This entails the discovery of local minimzers only.

4.4.4 Computational complexity

Although the harmonic mean weight matrix $\widetilde{W}^{(n)}$ (cf. (4.17)) is the inverse of a $(d_1d_2 \times d_1d_2)$-matrix and consequently a dense $(d_1d_2 \times d_1d_2)$-matrix in the general case, this is not relevant for the practical implementation of the algorithm as it never has to be computed explicitly. Moreover, neither there is a practical need to compute its inverse $(\widetilde{W}^{(n)})^{-1} = \frac{1}{2}\left(U^{(n)}(\bar{\Sigma}^{(n)})^{2-p}U^{(n)*} \oplus V^{(n)}(\bar{\Sigma}^{(n)})^{2-p}V^{(n)*}\right)$ explicitly.

Indeed, the harmonic mean weight matrix is just involved in the form of the linear operator $(\mathcal{W}^{(n)})^{-1}$ (cf. (4.14)) acting on the space of matrices $M_{d_1 \times d_2}$, which can be represented as a left- and right-sided matrix multiplication as it can be derived from (4.11) and the definition of the Kronecker sum (4.6).

Consequently, the application of the operator $(\mathcal{W}^{(n)})^{-1}$ is feasible in $O(d_1d_2(d_1+d_2))$ and can be implemented via the naive matrix multiplication algorithm, and, hence, can be easily parallelized.

The computation costs for the expression $\Phi \circ \widetilde{\mathcal{W}}^{(n)-1} \circ \Phi^* \in M_{m \times m}$ also depend on the linear measurement operator Φ. We note that in particular, for the matrix completion setting (4.3), where Φ is a just an entry selection operator, we do not have to perform additional arithmetic operations.

We point out that the execution of the `HM-IRLS` algorithm involves of two major computational steps in each iteration: On the one hand, the computation of the SVD of the $d_1 \times d_2$-matrix $X^{(n)}$ with time complexity $O(d_1d_2 \min(d_1, d_2))$. On the other hand the solution of the least squares problem under the linear constraint in (4.15), whose time complexity depends on Φ. In the matrix completion case, the second step is dominated by the inversion of a symmetric, $m \times m$ sparse linear system, which has a time complexity of at most $O(\max(d_1, d_2)^3 r^3)$.

For the matrix completion setting, we are able to perform recovery of low-rank matrices up to, e.g., $d_1 = d_2 = 3000$ on a single machine from only very few given entries.

Acceleration possibilities and extensions

A key idea for also enabling the solution of higher dimensional problems in reasonable runtimes is to speed up the solution of the $m \times m$ linear system in (4.15), which constitutes the computational bottleneck of the algorithm, by employing iterative solvers. In the sparse vector case, a significant gain in computational speed of the corresponding `IRLS` algorithm [35] could be reached by the incorporation of conjugate gradient (CG) methods as discussed in [49]. In this work, a competitive solver for the sparse recovery problem is suggested, which introduces an effective preconditioning and couples the accuracy of the CG solutions to the outer IRLS iteration. A similar modification could be employed for an acceleration of `HM-IRLS`.

Furthermore, it could be interesting to explore whether additional computational speed up is possible by replacing the full SVDs of the iterates $X^{(n)}$, which are used to define the linear operator $(\mathcal{W}^{(n)})^{-1}$ in Algorithm 4, by an approximation via truncated and randomized SVDs [75].

Chapter 5

Generalized IRLS for recovery of matrices with multiple structures

In this chapter, we want to pass over from the recovery of low-rank matrices to more general high dimensional signals with multiple underlying structures from a minimal amount of linear measurements

$$\Phi(X) = Y \tag{5.1}$$

with a linear map $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m, Y \in \mathbb{R}^m$ and $m \ll d_1 \cdot d_2$. More precisely, we want to consider matrix recovery problems that involve the different sparsity-type structures introduced in Section 2.3.2. In practical applications, often signals with even more than one structural property or a combination of the above mentioned ones - sparse vectors or vectorized matrices, row and column-sparse matrices and low-rank matrices - emerge, e.g., row-sparse and low rank matrices or the sum of sparse and low-rank matrices. In the following, we will explore recovery problems, where the matrix to be recovered is either a matrix with multiple sparsity-type structures occurring simultaneously or is the linear combination of several matrices with different sparsity structures.

For a matrix X with sparsity structures $s_s, s \in [t]$, the equivalents to “support” S_s, ”support size or order” k_s in the vector case are summarized in the following table:

property s_s	support S_s	order $k_s = \lvert S_s \rvert$
sparsity	$\Lambda = \{ l \mid X_{i,j} \neq 0, l = (i-1) \cdot d_1 + j \ \}$	# entries $\neq 0$
row-sparsity	$\Lambda_{row} = \left\{ i \mid \sum_{j=1}^{d_2} X_{i,j} \neq 0 \right\}$	# rows $\neq 0$
column-sparsity	$\Lambda_{col} = \left\{ j \mid \sum_{i=1}^{d_2} X_{i,j} \neq 0 \right\}$	# columns $\neq 0$
low-rank	$\Lambda_{rank} = \{ i \mid$ singular value $\sigma_i(X) \neq 0 \}$	rank

Table 5.1: Structural properties, analog to support, and order in the vector case

For this problem setting of recovery of multi-structured matrices from an equation system (5.1), we want to consider the following nonconvex and nonsmooth model in the *constrained* formulation :

$$\min F_1^0(X) = \sum_{s=1}^{t} \lambda_s \|X\|_{N_s^0}, \qquad (5.2)$$
$$\text{s.t. } \Phi(X) = Y,$$

where $\|\cdot\|_{N_s^0}$ is one the following structure inducing quasi-norms

(i) $\|\cdot\|_{\ell_0}$ for sparsity,

(ii) $\|\cdot\|_{\ell_{2,0}}$ for row-sparsity,

(iii) $\|\cdot\|_{\ell_{0,2}}$ for column-sparsity,

(iv) $\|\cdot\|_{\ell_{S_0}}$ for low rank.

Moreover, it is possible that the measurement data is supposed to have structured perturbations and that the residual $\Phi(X) = Y$ shows sparsity type features as well. We can incorporate this information via an *unconstrained* model formulation:

$$\min F_2^0(X) = \sum_{s=1}^{t} \lambda_s \|X\|_{N_s^0} + \mu \|\Phi(X) - Y\|_{N_r^0}. \qquad (5.3)$$

In the case, a matrix valued representation of measurement results is possible, i.e., $Y = \widetilde{Y}_{\text{vec}}$ for $\widetilde{Y} \in \mathbb{R}^{m_1 \times m_2}$ with $m = m_1 m_2$, we can consider $\|\cdot\|_{N_r^0}$ to be again any of the above mentioned quasi-norms applied to $(\Phi(X) - Y)_{\text{mat}(m_1,m_2)}$. Otherwise, in the case of a vector valued residual $\Phi(X) - Y$, we only assume sparsity as a reasonable structure and, therefore, $\|\cdot\|_{N_r^0} = \|\cdot\|_{\ell_0}$.

As already mentioned, sparsity-type recovery problems in their formulation via the minimization of nonconvex functionals involving terms promoting the specific sparsity structure are NP-hard to solve. Therefore, it is useful to consider their relaxation and substitution by an appropriate convex norm minimization problem (see Table 5.2).

A straightforward approach towards recovery strategies for the case of simultaneously structured signals would be the linear combination of the convex norms usually minimized for each of the single structures. Recently, the negative results of Oymak e.a.[116] surprised the community. They revealed that this intuitive attempt of combining convex norms will require just as many measurements as exploiting only one (dominating)

property	nonconvex functional	convex relaxation
sparsity	$\|\cdot\|_{\ell_0}$	$\|\cdot\|_{\ell_1}$
row-sparsity	$\|\cdot\|_{\ell_{2,0}}$	$\|\cdot\|_{\ell_{2,1}}$
column-sparsity	$\|\cdot\|_{\ell_{0,2}}$	$\|\cdot\|_{\ell_{1,2}}$
low-rank	$\mathrm{rank}(\cdot)$	$\|\cdot\|_*$

Table 5.2: Structural properties and nonconvex and convex promotion functional

structure. Only the combination of the nonconvex functionals that are promoting a certain structural property will be beneficial for a reduction of the number of measurements.

In general, by combining different structural assumptions, we reduce the d_1d_2 degrees of freedom of a general $(d_1 \times d_2)$-matrix considerably. Nevertheless, the necessary number of measurements depends on the recovery strategy that is employed and Oymak e.a. [116] show that there is a significant difference for convex and nonconvex approaches for simultaneously structured matrix recovery.

A first approach towards a solution is to only mildly relax the formulation of the problem (5.2) to attenuate its nonconvexity but not progressing until reaching convexity. This leads, finally, to the model problems we want to work on in this chapter:

(1) *constrained problem formulation:* the objective functional $F_1(X)$ to be minimized is a combination of (quasi-)norms of the signal in matrix form X under the linear constraint

$$\min F_1(X) = \sum_{s=1}^{t} \lambda_s \|X\|_{N_s}, \tag{5.4}$$
$$\text{s.t. } \Phi X = Y$$

(2) *unconstrained problem formulation:* the objective functional $F_2(X)$ to be minimized is a combination of (quasi-)norms of the signal in matrix form X itself and some (quasi-)norm of its residuals

$$\min F_2(X) = \sum_{s=1}^{t} \lambda_s \|X\|_{N_s} + \mu \|\Phi(X) - Y\|_{N_r}, \tag{5.5}$$

where $\lambda_s \in \mathbb{R}_+$, $s \in [t]$ and $\|\cdot\|_{N_s}, \|\cdot\|_{N_r}$ are one the following (quasi-)norms

(i) $\|\cdot\|_{\ell_p}^p$, for $0 < p \leq 2$ promoting sparsity,

(ii) $\|\cdot\|_{\ell_{2,p'}}^{p'}$, for $0 < p' \leq 2$ promoting row-sparsity,

(ii) $\| \cdot \|_{\ell_{p'',2}}^{p''}$, for $0 < p'' \leq 2$ promoting column-sparsity,

(iii) $\| \cdot \|_{\ell_{S_p'''}}^{p'''}$, for $0 < p''' \leq 2$ promoting low rank.

We note that the popular problem of low rank and sparse matrix decomposition also known as *robust principal component analysis (RPCA)* as in [19, 27, 166, 168, 170] is a special case of our setting with $\Phi = I_{d_1}$.

The solution of the nonconvex minimization problems above is in general a hard problem and standard relaxation based compressed sensing approaches will fail to be applicable. In this chapter, we present a generalized Iteratively Reweighted Least Squares method inspired by the ability of IRLS-type algorithms to approximate the different nonconvex problems for the individual structures, for some of them even with a superlinear rate of convergence. This approach was already explored for the special case of row-sparse and low-rank matrices in the master's thesis of Christian Kümmerle [84], which was co-supervised by the author of this thesis. The concept is extended to general simultaneously occurring sparsity-type structures for matrices and convex combinations of those in this chapter of the thesis, based on unpublished results in joint work with Christian Kümmerle. The present version of IRLS, named Generalized Iteratively Reweighted Least Squares (GIRLS) will be able to handle any kind of combination of nonconvex (quasi-)norms (5.2) and (5.3) as described in full detail in the later section of either the signal in matrix form X itself or different kinds of linear measurement residuals.

This extension to a generalized framework contributes novel theoretical results while no further numerical experiments were conducted beyond Kümmerle's master's thesis. Therefore, we refer to his work for details on numerical tests in this context.

Moreover, the optimal choice of the parameters $\lambda_s \in \mathbb{R}_+, s \in [t]$ is still an interesting open problem, where it would be possible to explore methods used in the framework of multi-parameter regularization [99] and to apply them in our setting as well.

5.1 Auxiliary functional and generalized IRLS algorithm for structured matrices

As we observed in Definition 2.37, each of the (quasi-)norms $\|\cdot\|_{N_s}$ (or also $\|\cdot\|_{N_r}$) above can be expressed as a classical reweighted ℓ_2-norm $\|\cdot\|_{\ell_2(W_s)}$ if considered in a vectorized formulation. The weight matrices corresponding to the different sparsity-types have certain structural patterns and Table 5.3 summarizes these structures presented in Section 2.4. Note that the weight matrices W_s for $\ell_2(W_s)$- minimization (or analogously for W_r and $\ell_2(W_r)$) given in the table are applied to the vectorized version of Z, $Z_{\text{vec}} \in \mathbb{R}^{d_1 d_2}$ and we use the indices $l \in [d_1 d_2], l = (i-1) \cdot d_1 + j$, $i \in [d_1], j \in [d_2]$.

property	$\|\cdot\|_{N_s}$	weight matrix $W_s \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$
sparsity	$\|\cdot\|_{\ell_p}$	$W_s = \mathrm{diag}(W_l)_{l=1}^{d_1 d_2}$ with $W_l = \|Z_l\|^{p-2}$
row-sparsity	$\|\cdot\|_{\ell_{2,p}}$	$W_s = \mathrm{diag}(W_l)_{l=1}^{d_1 d_2}$ with $W_l = (\sum_{j=1}^{d_2} \|Z_{ij}\|^2)^{\frac{p-2}{2}}$
column-sparsity	$\|\cdot\|_{\ell_{p,2}}$	$W_s = \mathrm{diag}(W_l)_{l=1}^{d_1 d_2}$ with $W_l = (\sum_{i=1}^{d_1} \|Z_{ij}\|^2)^{\frac{p-2}{2}}$
low rankness	$\|\cdot\|_{S_p}$	$W_s = \mathbf{I}_{d_2} \otimes W_L$ with $W_L = ZZ^T$
low rankness	$\|\cdot\|_{S_p}$	$W_s = W_R \otimes \mathbf{I}_{d_1}$ with $W_R = Z^T Z$

Table 5.3: Structural properties, corresponding quasinorms $\|\cdot\|_{N_s}$ and weight matrices for $\ell_2(W_s)$-minimization

To prevent singularity and instability problems a smoothing factor ϵ can be incorporated into the weight matrices and we refer to (2.76) for the respective smoothed versions $W_{s,\epsilon}(X)$.

Having understood these facts, we can use the linearity of $\|\cdot\|^2_{\ell_2(W)}$ to unify the sum of reweighted ℓ_2-norms for a vector $z \in \mathbb{R}^d$ and weight matrices $W_n \in \mathbb{R}^d$ as follows:

$$\sum_{n=1}^{N} \|z\|^2_{\ell_2(W_n)} = \sum_{n=1}^{N} \sum_{i=1}^{d} (W_n)_i z_i^2 = \sum_{i=1}^{d} \left(\sum_{n=1}^{N} (W_n)_i \right) z_i^2 = \|z\|^2_{\ell_2(\sum_{n=1}^{N} W_n)}. \tag{5.6}$$

We can apply (5.6) to obtain the objective functional F_1 in (5.4) and the first term in the objective functional F_2 in (5.5)

$$\begin{aligned} \sum_{s=1}^{t} \lambda_s \|X_{\text{vec}}\|^2_{\ell_2(N_s)} &= \sum_{s=1}^{t} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \lambda_s (W_s)_{ij} X_{ij}^2 = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \left(\sum_{n=1}^{t} \lambda_s (W_s)_{ij} \right) X_{ij}^2 \\ &= \|X_{\text{vec}}\|^2_{\ell_2(\sum_{s=1}^{t} \lambda_s W_s)} = \|X_{\text{vec}}\|^2_{\ell_2(W_1)}, \end{aligned} \tag{5.7}$$

where the *weight matrix* $W_1 = \sum_{s=1}^{t} \lambda_s W_s$.

We define the *weight matrix* W_2 of the residual term $\Phi(X) - Y$ as

$$W_2 = \mu W_r. \tag{5.8}$$

As a conclusion from the calculations above, we can reformulate (5.4) and (5.5) in a unified simple structure as follows

(1°) *constrained problem formulation:* the objective functional $F_1(X)$ to be minimized will be a combination of (quasi-)norms of the signal in matrix form X under linear constraints

$$\min F_1^\circ(X) = \|X_{\text{vec}}\|_{\ell_2(W_1)}^2 \qquad \text{s.t. } \Phi(X) = Y, \tag{5.9}$$

where W_1 is defined as above.

(2°) *unconstrained problem formulation:* the objective functional $F_2(X)$ to be minimized will be a combination of (quasi-)norms of the signal in matrix form X itself and its residuals

$$\min F_2^\circ(X) = \|X_{\text{vec}}\|_{\ell_2(W_1)}^2 + \|\Phi(X) - Y\|_{\ell_2(W_2)}^2, \tag{5.10}$$

where W_1 and W_2 are defined as above.

Remark 5.1. In the even more general case, one can consider several measurement sets

$$\Phi_r(X) = Y_r, \quad r \in [R],$$

where $\Phi_r : M_{d_1 \times d_2} \to \mathbb{R}^{m_r}$, $Y_r \in \mathbb{R}^{m_r}$ and the corresponding residuals with different structures enforced by the different norms $\|\cdot\|_{N_r}$

$$\|\Phi_r(X) - Y_r\|_{N_r}, r \in [R].$$

The unconstrained problem formulation generalizes to the objective functional $F_2(X)$ as follows

$$\min F_2(X) = \sum_{s=1}^{t} \lambda_s \|X\|_{N_s} + \sum_{r=1}^{R} \mu_r \|\Phi_r(X) - Y_r\|_{N_r}. \tag{5.11}$$

To keep notations simple, we restrict the formulation of the problems, algorithms and analysis in the rest of the paper to a single linear constraint.

One key tool to treat the different terms to be minimized are the transformations of (5.2) and (5.3) to their unified reweighted ℓ_2-norm minimization of the form of (5.9) and (5.10).

5.2 Algorithm formulation for the constrained case

At this point, we introduce a useful tool for the formulation and theoretical analysis of an iteratively reweighted least squares algorithm for problems of type (5.2) in the form of the following functional:

Definition 5.2. Let $\Phi : M_{d_1\times d_2} \to \mathbb{R}^m, Y \in \mathbb{R}^m$ and $X \in M_{d_1\times d_2}$. Moreover, consider the quasi-norms $\|\cdot\|_{N_s}$ for $s \in [t]$ and the corresponding parameters $\epsilon_s > 0, s \in [t]$ as well as the weight matrices $W_s \in \mathbb{R}^{d_1\cdot d_2\times d_1\cdot d_2}$ corresponding to the quasi-norms in dependence of X and ϵ_s. Set $W_1 = \sum\limits_{s=1}^{t} \lambda_s p_s W_s$. We define the *auxiliary functional* for the constrained algorithm as

$$\mathcal{J}_{GIRLS}(X, (\epsilon_s)_{s=1,\ldots,t}, (W_s)_{s=1,\ldots,t}) :=$$
$$\frac{1}{2}\left[\|X_{\text{vec}}\|^2_{\ell_2(W_1)} + \sum_{s=1}^{t} \|\epsilon_s \cdot \mathbf{1}_{d_1\cdot d_2}\|^2_{\ell_2(\lambda_s p_s W_s)} + 2 - p_s\lambda_s \|W_s^{p_s/(p_s-2)}\|_F^2\right]. \tag{5.12}$$

Next, we define the auxiliary variable for a matrix $X \in M_{d_1\times d_2}$

$$\mathcal{N}_s(X) = \begin{cases} r_{K_s+1}(X)/(d_1\cdot d_2)^{1/p_s}, & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{\ell_{p_s}}, \\ r_{K_s+1}\left(\left(\sum\limits_{j=1}^{d_2}(X_{ij})^2\right)^{1/2}\right)/d_1^{1/p_s}, & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{\ell_{2,p_s}}, \\ r_{K_s+1}\left(\left(\sum\limits_{i=1}^{d_1}(X_{ij})^2\right)^{1/2}\right)/d_2^{1/p_s}, & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{\ell_{p_s,2}}, \\ \sigma_{K_s+1}(X)/(\min(d_1,d_2))^{1/p_s}, & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{S_{p_s}}. \end{cases}$$

Moreover, let

$$\widetilde{\mathcal{J}}_{GIRLS}(W)^{(n+1)}_{\bar{s}} = \mathcal{J}_{GIRLS}(X^{(n+1)}, (\epsilon_s^{(n+1)})_{s=1,\ldots,t}, (W_s^{(n+1)})_{s=1,\ldots,\bar{s}-1}, W, (W_s^{(n)})_{s=\bar{s}+1,\ldots,t}).$$

An iteratively reweighted least squares algorithm for the approximation of the solution of (5.2) can be formulated as an alternating minimization of the just defined auxiliary functional with respect to its arguments.

Algorithm 5 Generalized IRLS for structured matrices (GIRLS)

```
Input: $\Phi : M_{d_1 \cdot d_2} \to \mathbb{R}^m$, $Y = \Phi(X_0) \in \mathbb{R}^m$ for ground truth matrix $X_0 \in M_{d_1 \times d_2}$,
        nonconvexity parameters $p_s$ for $s \in [t]$.
Output: $X^{(1)}, X^{(2)}, \ldots \in M_{d_1 \times d_2}$
Initialize $\epsilon_s^{(0)} = 1$, set $W_s^{(0)} = \lambda_s p_s \cdot I_{d_1 \cdot d_2}$ for $s \in [t]$.
repeat
    $X^{(n+1)} = \arg\min_{\Phi(X)=Y} \mathcal{J}_{GIRLS}(X, (\epsilon_s^{(n)})_{s=1,\ldots,t}, (W_s^{(n)})_{s=1,\ldots,t})$        (5.13)
    $= \arg\min_{\Phi(X)=Y} \|X_{\text{vec}}\|^2_{\ell_2(W_1^{(n)})}$
    $= \big(W_1^{(n)-1} \circ \Phi^* \circ (\Phi \circ W_1^{(n)-1} \circ \Phi^*)^{-1}\big)\big(Y\big)$
    for $s = 1, \ldots, t$
        $\epsilon_s^{(n+1)} = \min\Big(\epsilon_s^{(n)}, \max(\mathcal{N}_s(X^{(n+1)}), \tilde{\epsilon}), \max_s(\mathcal{N}_s(X^{(n+1)}))\Big)$ with $\tilde{\epsilon} > 0$,        (5.14)
    for $s = 1, \ldots, t$
        $W_s^{(n+1)} = \begin{cases} \arg\min_{W>0, W \text{ diag}} \tilde{\mathcal{J}}_{GIRLS}(W)_s^{(n+1)}, & \text{for } \|\cdot\|_{N_s} = \|\cdot\|_{\ell_{p_s}}^{p_s}, \\ \arg\min_{W=\mathbf{I}_{d_2}\otimes\mathbf{W}, \mathbf{W}>0, \mathbf{W}\text{diag}} \tilde{\mathcal{J}}_{GIRLS}(W)_s^{(n+1)}, & \text{for } \|\cdot\|_{N_s} = \|\cdot\|_{\ell_{2,p_s}}^{p_s}, \\ \arg\min_{W=\mathbf{W}\otimes\mathbf{I}_{d_1}, \mathbf{W}>0, \mathbf{W}\text{diag}} \tilde{\mathcal{J}}_{GIRLS}(W)_s^{(n+1)}, & \text{for } \|\cdot\|_{N_s} = \|\cdot\|_{\ell_{2,p_s}}^{p_s}, \\ \arg\min_{W=\mathbf{I}_{d_2}\otimes\mathbf{W}, \mathbf{W}>0} \tilde{\mathcal{J}}_{GIRLS}(W)_s^{(n+1)}, & \text{for } \|\cdot\|_{N_s} = \|\cdot\|_{S_{p_s}}^{p_s}, \\ \arg\min_{W=\mathbf{W}\otimes\mathbf{I}_{d_1}, \mathbf{W}>0} \tilde{\mathcal{J}}_{GIRLS}(W)_s^{(n+1)}, & \text{for } \|\cdot\|_{N_s} = \|\cdot\|_{S_{p_s}}^{p_s} \end{cases}$
        $= W_{s,\epsilon_s^{(n+1)}}(X^{(n+1)})$ as defined in (2.76).        (5.15)
    $W_1^{(n+1)} = \sum_{s=1}^{t} \lambda_s p_s W_s^{(n+1)}$.
    $n = n + 1$.
until stopping criterion is met;
Set $n_0 = n$.
```

We stop the algorithm if $\epsilon_s^{(n)} = 0$ for $s \in [t]$ and set $X^{(j)} := X^{(n)}$ for $j > n$. However, in general, the algorithm will generate an infinite sequence $(X^{(n)})_{n \in \mathbb{N}}$ of distinct matrices and it is convenient to keep the variables ϵ_s, W_s fixed as soon as $\epsilon_s^{(n)}$ falls below an appropriately chosen threshold and only continue updating the other variables.

The details of the derivation of explicit expressions to calculate $X^{(n+1)}, W_s^{(n)}$ as defined in (2.76) is omitted here. They can be obtained by deducing the appropriate Langrangian from J_{GIRLS} and the corresponding constraints and minimizing the resulting functional, where each of the minimization steps carried out in the algorithm constitutes a convex optimization problem.

5.3 Theoretical analysis and convergence results for the constrained case

In the following section, we will have a closer look at Algorithm 5 and point out some of its properties. In particular, we show that the iterates $(X^{(n)})_{n\in\mathbb{N}}$ stay bounded and the fact that two consecutive iterates are getting arbitrarily close as $n \to \infty$. These results will be useful to develop finally the proof of convergence for Algorithm 5 under conditions determined along the way.

5.3.1 Unified presentation of properties for different sparsity structures

At this point, we want to summarize certain useful notations and matrix properties in their specific variants for the sparsity structures mentioned above and give their formulation for general structured $X \in M_{d_1\times d_2}$:

(i) **RIP**: A map $\Phi : M_{d_1\times d_2} \to \mathbb{R}^m$ fulfills the RIP for a structure s_s of order k_s with constant $\delta_s \in (0,1)$ if for every matrix X with sparsity structure s_s of order k_s holds

$$(1-\delta_s)\|X\|_F^2 \leq \|\Phi(X)\|_{\ell_2}^2 \leq (1+\delta_s)\|X\|_F^2. \tag{5.16}$$

Moreover, each of these versions of the restricted isometry properties implies the corresponding nullspace property (NSP) ([88, 117, 124, 158]):

(ii) **NSP**: A map $\Phi : M_{d_1\times d_2} \to \mathbb{R}^m$ fulfills the NSP for a structure s_s of order k_s with constant $\gamma_s \in (0,1)$ if for all elements $\eta \in M_{d_1\times d_2}$ of the nullspace of Φ, $\mathcal{N}(\Phi)$ holds

$$\|\eta_{S_s}\|_{N_s} \leq \gamma_k \|\eta_{S_s^c}\|_{N_s}. \tag{5.17}$$

(iii) **Best k_s-term approximation error**[54]: For $p_s > 0$, the *best k_s-term approximation error* to a matrix $X \in M_{d_1\times d_2}$ is defined by

$$\beta_{k_s}(X)_{N_s} := \inf\left\{\|X-Z\|_{N_s},\, Z \text{ has sparsity structure } s_s \text{ of order } k_s\right\}.$$

A quite straightforward consequence of the above NSP is the following corresponding inequality

(iv) **Inequality from NSP** [35, 55, 116]: If a map $\Phi : M_{d_1\times d_2} \to \mathbb{R}^m$ fulfills the NSP for a structure s_s of order k_s with constant $\gamma_s \in (0,1)$ and Z, Z' with $\Phi(Z) = Y$ and $\Phi(Z') = Y$ we have

$$\|Z' - Z\|_{N_s} \leq \frac{1+\gamma_s}{1-\gamma_s}\left(\|Z'\|_{N_s} - \|Z\|_{N_s} + 2\beta_{k_s}(Z)_{N_s}\right). \tag{5.18}$$

5.3.2 Preliminary results

In this section, we formulate several Lemmata that will be fundamental ingredients for the proof of convergence of Algorithm 5.

In the following, we want to assume that our desired solution matrix X_0 has sparsity structure $s_s, s \in [t]$ of order $k_s, s \in [t]$ and the map Φ fulfills the corresponding NSP of order $K_s, s \in [t]$ with $K_s > k_s$ respectively, where K_s is representing a generous guess of the sparsity level k_s.

Denote $\mathcal{J}^{(n)}_{GIRLS} = \mathcal{J}_{GIRLS}(X^{(n)}, (\epsilon_s^{(n)})_{s=1,\dots,t}, (W_s^{(n)})_{s=1,\dots,t})$. Our first quite straightforward observation is that at iteration n of Algorithm 5 the following holds

$$\begin{aligned}\mathcal{J}^{(n)}_{GIRLS} &= \frac{1}{2}\left(\|X^{(n)}_{\text{vec}}\|^2_{\ell_2(\bar{W}^{(n)})} + \sum_{s=1}^{t}\|\epsilon_s^{(n)}\cdot \mathbf{1}_{d_1\cdot d_2}\|^2_{\ell_2(\lambda_s W_s^{(n)})} + \sum_{s=1}^{t}(2-p_s)\|(W_s^{(n)})^{p_s/(p_s-2)}\|^2_{\ell_2}\right)\\ &= \sum_{s=1}^{t}\lambda_s\frac{p_s}{2}\sum_{l=1}^{d_1\cdot d_2} W_s^l (X_l^{(n)})^2 + \lambda_s\frac{p_s}{2}\sum_{l=1}^{d_1\cdot d_2} W_s^l(\epsilon_s^{(n)})^2 + \frac{2-p_s}{2}\lambda_s (W_s^l)^{\frac{p_s}{p_s-2}}\\ &= \sum_{s=1}^{t}\lambda_s f_{N_s}^{\epsilon_s^{(n)}}(X^{(n)}),\end{aligned} \tag{5.19}$$

$$\text{where } f_{N_s}^{\epsilon_s}(X) = \begin{cases} \sum_{l=1}^{d_1\cdot d_2}\left(|X_l|^2 + \epsilon_s^2\right)^{\frac{p_s}{2}} & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{\ell_{p_s}},\\ \sum_{i=1}^{d_1}\left(\sum_{j=1}^{d_2}|X_{ij}|^2 + \epsilon_s^2\right)^{\frac{p_s}{2}} & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{\ell_{2,p_s}},\\ \sum_{j=1}^{d_2}\left(\sum_{i=1}^{d_1}|X_{ij}|^2 + \epsilon_s^2\right)^{\frac{p_s}{2}} & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{\ell_{p_s,2}},\\ \operatorname{tr}\left(\left(XX^T + \epsilon_s^2\cdot I_{d_1}\right)^{\frac{p_s}{2}}\right) & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{S_{p_s}},\\ \operatorname{tr}\left(\left(X^TX + \epsilon_s^2\cdot I_{d_2}\right)^{\frac{p_s}{2}}\right) & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{S_{p_s}}.\end{cases} \tag{5.20}$$

We note that $f_{N_s}^{\epsilon_s}$ is a good approximations to $\|\cdot\|_{N_s}$, which will be useful later.

Furthermore, we observe that due to the minimization properties resulting from Algorithm 5, the following monotonicity property holds.

Lemma 5.3. *The inequalities*

$$\begin{aligned}\mathcal{J}_{GIRLS}^{(n)} &= \mathcal{J}_1(X^{(n)}, (\epsilon_s^{(n)})_{s=1,\ldots,t}, (W_s^{(n)})_{s=1,\ldots,t}) \\ &\geq \mathcal{J}_{GIRLS}(X^{(n+1)}, (\epsilon_s^{(n)})_{s=1,\ldots,t}, (W_s^{(n)})_{s=1,\ldots,t}) \\ &\geq \mathcal{J}_{GIRLS}(X^{(n+1)}, (\epsilon_s^{(n+1)})_{s=1,\ldots,t}, (W_s^{(n)})_{s=1,\ldots,t}) \\ &\geq \mathcal{J}_{GIRLS}(X^{(n+1)}, (\epsilon_s^{(n+1)})_{s=1,\ldots,t}, (W_s^{(n+1)})_{s=1,\ldots,t}) = \mathcal{J}_{GIRLS}^{(n+1)}\end{aligned}$$

hold for all $n \geq 0$.

Proof. Here the first inequality follows from the minimization property that defines $X^{(n+1)}$, the next inequalities from $\epsilon_s^{(n+1)} \leq \epsilon_s^{(n)}$, and the last inequality from the minimization properties that define $W_s^{(n+1)}$. □

Due to Lemma 5.3, we can state that

$$\mathcal{J}_{GIRLS}(X^{(n)}, (\epsilon_s^{(n)})_{s=1,\ldots,t}, (W_s^{(n)})_{s=1,\ldots,t}) \leq \mathcal{J}_{GIRLS}(X^{(1)}, (\epsilon_s^{(0)})_{s=1,\ldots,t}, (W_s^{(0)})_{s=1,\ldots,t}),$$

where the right hand side is a constant and this will help to obtain the boundedness of the iterates $(X^{(n)})_{n\in\mathbb{N}}$:

Lemma 5.4. *The sequence of iterates $\left(X^{(n)}\right)_{n\in\mathbb{N}}$ defined by Algorithm 5 fulfills*

$$\sum_{s=1}^{t} \lambda_s \|X^{(n)}\|_{N_s} \leq \mathcal{J}_{GIRLS}(X^1, (\epsilon_s^{(0)})_{s=1,\ldots,t}, (W_s^{(0)})_{s=1,\ldots,t}) := \mathcal{J}_{GIRLS}^{(0)}.$$

Proof. For all $n \in \mathbb{N}$

$$\begin{aligned}\sum_{s=1}^{t} \lambda_s \|X^{(n)}\|_{N_s} \leq \sum_{s=1}^{t} \lambda_s f_{N_i}^{\epsilon_s^{(n)}}(X^{(n)}) &= \mathcal{J}_{GIRLS}(X^{(n)}, (\epsilon_s^{(n)})_{s=1,\ldots,t}, (W_s^{(n)})_{s=1,\ldots,t}) \\ &\leq \mathcal{J}_{GIRLS}(X^{(1)}, (\epsilon_s^{(0)})_{s=1,\ldots,t}, (W_s^{(0)})_{s=1,\ldots,t}) = \mathcal{J}_{GIRLS}(0),\end{aligned}$$

where the last inequality is a consequence of the monotonicity property stated in Lemma 5.3. □

As a next result, we would like to state that from the sequence $\mathcal{J}_{GIRLS}^{(n)}$ being convergent it follows that the iterates $X^{(0)}, \ldots, X^{(n)}, X^{(n+1)}, \ldots$ of Algorithm 5 are getting arbitrarily close for $n \to \infty$.

Lemma 5.5. *For the iterates of Algorithms 5 it holds*

$$\lim_{n\to\infty} \|X^{(n)} - X^{(n+1)}\|_{\ell_2}^2 = 0.$$

Proof. For each $n = 1, 2, ...$ we have

$$\begin{aligned} 2\left[\mathcal{J}^{(n)}_{GIRLS} - \mathcal{J}^{(n+1)}_{GIRLS}\right] &\geq 2\left[\mathcal{J}^{(n)}_{GIRLS} - \mathcal{J}_{GIRLS}(X^{(n+1)}, (\epsilon_s^{(n)})_{s=1,...,t}, (W_s^{(n)})_{s=1,...,t})\right] \\ &= \|X^{(n)}_{\text{vec}}\|^2_{\ell_2(W_1^{(n)})} - \|X^{(n+1)}_{\text{vec}}\|^2_{\ell_2(W_1^{(n)})} \\ &= \langle (X^{(n)} + X^{(n+1)})_{\text{vec}}, (X^{(n)} - X^{(n+1)})_{\text{vec}}\rangle_{\ell_2(W_1^{(n)})} \end{aligned}$$

Analogously to [35], we notice that the $\ell_2(W)$-norm is strictly convex and, therefore, its minimizer, which we denote by X_W, is unique. It is possible to characterize this minimizer by

$$\langle (X_W)_{\text{vec}}, \eta_{\text{vec}}\rangle_{\ell_2(W)} = 0$$

for all $\eta \in \mathcal{N}(\Phi)$. Since $X^{(n+1)}$ is the minimizer of $\|X_{\text{vec}}\|^2_{\ell_2(W_1^{(n)})}$ and $X^{(n)} - X^{(n+1)} \in \mathcal{N}(\Phi)$, it also holds that

$$\langle X^{(n+1)}_{\text{vec}}, (X^{(n)} - X^{(n+1)})_{\text{vec}})\rangle_{\ell_2(W_1^{(n)})} = 0.$$

Moreover, we need an estimate on $\sigma_{\min}(W_1^{(n)})$ to obtain a bound on the difference of iterates independent of the reweighting matrix. Since $1 = \sigma_{\min}(X)\sigma_{\max}(X^{-1})$ for any invertible matrix X, it is sufficient to calculate $\sigma_{\max}((W_1^{(n)})^{-1})$ to gain information on $\sigma_{\min}(W_1^{(n)})$. Notice that $(W_s^{(n)})^{-1}$ can be bounded by direct calculation as follows

$$\sigma_1\big((W_s^{(n)})^{-1}\big) \leq \Big(\lambda_s f_s^{\epsilon_s^{(n)}}(X^{(n)})\Big)^{\frac{2-p_s}{p_s}} \leq (\mathcal{J}^{(0)}_{GIRLS})^{\frac{2-p_s}{p_s}}.$$

Hence, we conclude that $\sigma_{\min}(W_s^{(n)}) \geq (\mathcal{J}^{(0)}_{GIRLS})^{1-\frac{2}{p_s}}$.

We can then summarize the results above to obtain

$$\begin{aligned} 2\left[\mathcal{J}^{(n)}_{GIRLS} - \mathcal{J}^{(n+1)}_{GIRLS}\right] &= \langle (X^{(n)} + X^{(n+1)})_{\text{vec}}, (X^{(n)} - X^{(n+1)})_{\text{vec}}\rangle_{\ell_2(W_1^{(n)})} \\ &= \langle (X^{(n)} - X^{(n+1)})_{\text{vec}}, (X^{(n)} - X^{(n+1)})_{\text{vec}}\rangle_{\ell_2(W_1^{(n)})} \\ &= \|(X^{(n)} - X^{(n+1)})_{\text{vec}}\|^2_{\ell_2(\sum\limits_{s=1}^{t} \lambda_s p_s W_s)} \\ &= \sum_{s=1}^{t} \lambda_s p_s \|(X^{(n)} - X^{(n+1)})_{\text{vec}}\|^2_{\ell_2(W_s)} \\ &\geq \sum_{s=1}^{t} \lambda_s p_s \sigma_{\min}(W_s^{(n)})\|(X^{(n)} - X^{(n+1)})_{\text{vec}}\|^2_{\ell_2} \\ &= \sum_{s=1}^{t} \lambda_s p_s (\mathcal{J}^{(0)}_{GIRLS})^{1-\frac{2}{p_s}}\|(X^{(n)} - X^{(n+1)})_{\text{vec}}\|^2_{\ell_2} \end{aligned}$$

$$:= C\|(X^{(n)} - X^{(n+1)})_{\mathrm{vec}}\|_{\ell_2}^2.$$

The monotonicity property stated in Lemma 5.3 and the boundedness of the sequence $\left(\mathcal{J}_{GIRLS}^{(n)}\right)_{n\in\mathbb{N}}$ imply that

$$\lim_{n\to\infty}(\mathcal{J}_{GIRLS}^{(n)} - \mathcal{J}_{GIRLS}^{(n+1)}) = 0,$$

hence also

$$\lim_{n\to\infty}\|(X^{(n)} - X^{(n+1)})_{\mathrm{vec}}\|_{\ell_2}^2 = 0.$$

□

From the monotonicity of the components of $\left(\epsilon_s^{(n)}\right)_{s=1,\dots,t}$ we know that $\epsilon_s := \lim_{n\to\infty}\epsilon_s^{(n)}$ exists and is non-negative. Define $\epsilon := \left(\epsilon_s\right)_{s=1,\dots,t}$. The following functional will play a role in our proof of convergence, especially if all components of ϵ are positive.

Definition 5.6. (ϵ-perturbed objective functional) We define the ϵ-perturbed objective functional to be of the following form

$$F^{\epsilon}(X) := \sum_{s=1}^{t}\lambda_s f_{N_s}^{\epsilon_s}(X)$$

and the corresponding minimization problem

$$\min_{\Phi(X)=Y} F^{\epsilon}(X). \tag{5.21}$$

Notice that, if we knew that $X^{(n)}$ converged to a point $\bar{X}$, then, having in mind (4.25), $F^{\epsilon}(\bar{X})$ would be the limit of $\mathcal{J}_{GIRLS}(X^{(n)}, \left(\epsilon_s^{(n)}\right)_{s=1,\dots,t}, \left(W_s^{(n)}\right)_{s=1,\dots,t})$ for $n \to \infty$. In the case that F^{ϵ} is nonconvex one might practically only be able to find critical points. We denote by $\mathcal{Z}^{\epsilon}(Y)$ its set of global minimizers Z with $\Phi(Z) = Y$. Moreover, let a minimizer in dependence of $\left(\epsilon_s\right)_{s=1,\dots,t}$ be denoted as

$$X^{\epsilon} \in \arg\min_{\Phi(X)=Y} F^{\epsilon}(X) = \mathcal{Z}^{\epsilon}(\tilde{Y}). \tag{5.22}$$

Lemma 5.7. *Let $\epsilon > 0$ and Z with $\Phi(Z) = Y$. Then Z is a critical point of F^{ϵ}, i.e., $\nabla_Z(F^{\epsilon})(Z) = \left(\frac{\partial F^{\epsilon}(Z)}{Z_l}\right)_{l=1}^{d_1\cdot d_2} = 0$ if and only if $\langle Z_{\mathrm{vec}}, \eta_{\mathrm{vec}}\rangle_{W_1(Z,\epsilon)} = 0$ for all $\eta \in \mathcal{N}(\Phi)$, where $W_1(Z,\epsilon) = \sum_{s=1}^{t}\lambda_s p_s W_s(Z,\epsilon_s)$. In the case that F^{ϵ} is convex, i.e., $p_s \geq 1$ for all $s \in [t]$, $\langle Z_{\mathrm{vec}}, \eta_{\mathrm{vec}}\rangle_{W_1(Z,\epsilon)} = 0$ implies that $Z = X^{\epsilon}$ is the unique minimizer.*

Proof. First we prove the "only if" part. Let Z be a critical point of F^ϵ meaning $(F^\epsilon)'(Z) = 0$ and $\eta \in \mathcal{N}(\Phi)$. Consider the function $\widetilde{F^\epsilon}(t) = F^\epsilon(Z + t\eta) - F^\epsilon(Z)$. Note that $(\widetilde{F^\epsilon})'(0) = \langle \nabla_Z(F^\epsilon)(Z)_{\text{vec}}, \eta_{\text{vec}}\rangle$ and that, if Z is a critical point of F^ϵ, 0 is also a critical point of $\tilde{F}^\epsilon$, i.e., $(\widetilde{F^\epsilon})'(0) = 0$. Therefore, if

$$0 = (\tilde{F}^\epsilon)'(0) = \langle \nabla_Z(F^\epsilon)(Z)_{\text{vec}}, \eta_{\text{vec}})\rangle = \left\langle \left(\sum_{s=1}^{t} \lambda_s (\nabla_Z(f^{\epsilon_s}_{N_s})(Z) \right)_{\text{vec}}, \eta_{\text{vec}} \right\rangle$$

where

$$(\nabla_Z(f^{\epsilon_s}_{N_s})(Z)))_{\text{vec}} = \begin{cases} \left(p_s Z_l \left(Z_l^2 + \epsilon_s^2\right)^{\frac{p_s-2}{2}}\right)_{l=1}^{d_1 \cdot d_2} & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{\ell_{p_s}} \\ \left(p_s Z_l \left(\sum_{j=1}^{d_2} |Z_{ij}|^2 + \epsilon_s^2\right)^{\frac{p_s-2}{2}}\right)_{l=1}^{d_1 \cdot d_2} & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{\ell_{2,p_s}}, \\ \left(p_s Z_l \left(\sum_{i=1}^{d_1} |Z_{ij}|^2 + \epsilon_s^2\right)^{\frac{p_s-2}{2}}\right)_{l=1}^{d_1 \cdot d_2} & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{\ell_{p_s,2}}, \\ \left(p_s Z^T \left(ZZ^T + \epsilon_s^2 \cdot I_{d_1}\right)^{\frac{p_s-2}{2}}\right)_{\text{vec}} & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{S_{p_s}}, \\ \left(p_s Z \left(Z^T Z + \epsilon_s^2 \cdot I_{d_2}\right)^{\frac{p_s-2}{2}}\right)_{\text{vec}} & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{S_{p_s}}. \end{cases}$$

$$= \begin{cases} p_s \operatorname{diag}\left(\left(Z_l^2 + \epsilon_s^2\right)^{\frac{p_s-2}{2}}\right)_{l=1}^{d_1 \cdot d_2} Z_{\text{vec}} & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{\ell_{p_s}}, \\ p_s \operatorname{diag}\left(\left(\sum_{j=1}^{d_2} |Z_{ij}|^2 + \epsilon_s^2\right)^{\frac{p_s-2}{2}}\right)_{l=1}^{d_1 \cdot d_2} Z_{\text{vec}} & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{\ell_{2,p_s}}, \\ p_s \operatorname{diag}\left(\left(\sum_{i=1}^{d_1} |Z_{ij}|^2 + \epsilon_s^2\right)^{\frac{p_s-2}{2}}\right)_{l=1}^{d_1 \cdot d_2} Z_{\text{vec}} & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{\ell_{p_s,2}}, \\ p_s \left(\left(\mathbf{I}_{d_2} \otimes \left(ZZ^T + \epsilon_s^2 \cdot I_{d_1}\right)^{\frac{p_s-2}{2}}\right)\right) Z_{\text{vec}} & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{S_{p_s}}, \\ p_s \left(\left(Z^T Z + \epsilon_s^2 \cdot I_{d_2}\right)^{\frac{p_s-2}{2}} \otimes \mathbf{I}_{d_1}\right) Z_{\text{vec}} & \text{for } \|\cdot\|_{N_s} = \|\cdot\|^{p_s}_{S_{p_s}}. \end{cases}$$

It follows that

$$\begin{aligned} 0 &= \langle \sum_{s=1}^{t} \lambda_s p_s (W_s(Z, \epsilon_s) Z)_{\text{vec}}, \eta_{\text{vec}} \rangle \\ &= \langle W_1(Z, \epsilon) Z_{\text{vec}}, \eta_{\text{vec}} \rangle = \langle Z_{\text{vec}}, \eta_{\text{vec}})\rangle_{W_1(Z,\epsilon)}. \end{aligned}$$

Now we treat the "if" part. Let $Z \in M_{d_1 \times d_2}$ be such that $\Phi(Z) = Y$ and assume that for all $\eta \in \mathcal{N}(\Phi)$ holds

$$0 = \langle Z_{\text{vec}}, \eta_{\text{vec}} \rangle_{W_1(Z,\epsilon)}.$$

Following the lines of the previous calculations, we see that

$$0 = \langle Z_{\text{vec}}, \eta_{\text{vec}} \rangle_{W_1(Z,\epsilon)} = \langle W_1(Z, \epsilon) Z_{\text{vec}}, \eta_{\text{vec}} \rangle = \langle (\nabla_Z(F^\epsilon)(Z))_{\text{vec}}, \eta_{\text{vec}} \rangle.$$

This means that $(\nabla_Z(F^\epsilon)(Z))_{\text{vec}}$ is perpendicular to the nullspace of Φ and therefore it holds

$$(\nabla_Z(F^\epsilon)(Z))_{\text{vec}} \in \text{Ran}(\Phi^*) \text{ and } \Phi(Z) = Y.$$

Therefore, Z satisfies the KKT conditions of (5.21) and Z is a critical point of F^ϵ under the linear constraint. This proves the first part of the lemma.
For the convex case, assume that $\Phi(Z) = Y$ and $\langle Z, \eta\rangle_{W_1(Z,\epsilon)} = 0$ for all $\eta \in \mathcal{N}(\Phi)$, where $W_1(Z,\epsilon)$ is defined as above. We shall show that, if $\langle Z_{\text{vec}}, \eta_{\text{vec}}\rangle_{W_1(Z,\epsilon)} = 0$, Z is the minimizer of F_1^ϵ for all $\Phi(Z) = Y$ and, therefore, coincides with X^ϵ. Since F^ϵ is a combination of the functions $f_{N_s}^{\epsilon_s}$, for any point Z_0, we obtain by convexity

$$\begin{aligned}F^\epsilon(\bar{Z}) = \sum_{s=1}^t \lambda_s f_{N_s}^{\epsilon_s}(\bar{Z}) &\geq \sum_{s=1}^t \lambda_s f_{N_s}^{\epsilon_s}(Z_0) + \sum_{s=1}^t \lambda_s \langle (\nabla_Z(f_{N_s}^{\epsilon_s})(Z_0))_{\text{vec}}, (\bar{Z} - Z_0)_{\text{vec}}\rangle \\ &= F^\epsilon(Z_0) + \langle (Z_0)_{\text{vec}}, (\bar{Z} - Z_0)_{\text{vec}}\rangle_{W_1(Z_0,\epsilon)}.\end{aligned}$$

If we take now $\bar{Z}$ such that $\Phi(\bar{Z}) = Y$ and $Z_0 = Z$, we have that $\bar{Z} - Z \in \mathcal{N}(\Phi)$ and obtain

$$F^\epsilon(\bar{Z}) \geq F^\epsilon(Z) + \langle Z_{\text{vec}}, (\bar{Z} - Z)_{\text{vec}}\rangle_{W_1(Z,\epsilon)} = F^\epsilon(Z),$$

which yields the result. □

5.3.3 Convergence results

Finally, we can state the convergence results for Algorithm 5.

Theorem 5.8. *Fix $Y \in \mathbb{R}^m$. Let $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m$ and the functional $\mathcal{J}_{GIRLS}(X^{(n)}, (\epsilon_s^{(n)})_{s=1,\dots,t}, (W_s^{(n)})_{s=1,\dots,t})$ be defined for $(\epsilon_s^{(n)})_{s=1,\dots,t}, (W_s^{(n)})_{s=1,\dots,t}$ as generated by Algorithm 5 for all $n \geq 0$. Assume that the matrix $X_0 \in M_{d_1 \times d_2}$ has property S_s of order k_s for all $s \in [t]$. Let $\bar{\mathcal{Z}}^\epsilon$ be the set of accumulation points of the sequence $(X^{(n)})_{n\in\mathbb{N}}$ generated by Algorithm 5.*

(i) If $\epsilon = (\epsilon_s)_{s=1,\dots,t} = (0)_{s=1,\dots,t}$, and

(a) the matrix Φ fulfills the corresponding NSP for all *structures S_s, $s \in [t]$, of order K_s as defined in Definition 4.10,*

or (b) if $p_s = p$ for all $s \in [t]$ and, if there exists some $s_0 \in [t]$ such that Φ fulfills the NSP corresponding to structure S_{s_0} of order K_{s_0},

then $\bar{\mathcal{Z}}^\epsilon$ consists of a single point $\bar{X}$ that has sparsity structure S_s of order K_s for all $s \in [t]$ and $\bar{X} = X_0$ is the solution to the minimization problem (5.4)*. Moreover, in case (a), we have for $k_s \leq K_s, s \in [t]$ and any Z with $\Phi(Z) = Y$ that*

$$\sum_{s=1}^{t} \lambda_s \|Z - \bar{X}\|_{N_s} \leq \sum_{s=1}^{t} \hat{C}_s \beta_{K_s}(Z)_{N_s}, \text{ where } \hat{C}_s = \frac{2\lambda_s(1+\gamma_s)}{(1-\gamma_s)}. \tag{5.23}$$

(ii) If $\epsilon = (\epsilon_s)_{s=1,\dots,t}$ with $\epsilon_s > 0$ for any s, then each point $\bar{X} \in \bar{\mathcal{Z}}^\epsilon$ is a stationary point of the ϵ-perturbed objective functional F_1^ϵ. In the case that F_1^ϵ is convex, i.e., $p_s \geq 1$ for $s \in [t]$ then $\bar{X} = X^\epsilon$ is the unique global minimizer of $F_1^\epsilon(X)$.

(iii) In the case of (ii), if the matrix Φ fulfills the corresponding NSPs corresponding to the structures s_s of order K_s for all $s \in [t]$ as defined in Definition 4.10 with $\gamma_s < 1 - \frac{2}{K_s+2}$ (or equivalently, if $\frac{2\gamma_s}{1-\gamma_s} < K_s$), then we have, for all $\bar{X} \in \bar{\mathcal{Z}}^\epsilon \cap \mathcal{Z}^\epsilon$ and all Z with $\Phi(Z) = Y$ and any $k_s < K_s - \frac{2\gamma_s}{1-\gamma_s}$ that

$$\sum_{s=1}^{t} \lambda_s \|Z - \bar{X}\|_{N_s} \leq \begin{cases} \sum_{s=1}^{t} \tilde{C}_s \beta_{K_s}(Z)_{N_s} & \text{if } \epsilon_s \geq \tilde{\epsilon} \text{ for all } s \in [t], \\ \sum_{s=1}^{t} \tilde{C}_s \max_s \beta_{K_s}(Z)_{N_s} & \text{if } \epsilon_s < \tilde{\epsilon} \text{ for any } s \in [t] \end{cases}$$

with $\tilde{C}_s := \frac{2\lambda_s(1+\gamma_s)}{1-\gamma_s}\left[\frac{K_s - k_s + \frac{3}{2}}{K_s - k_s - \frac{2\gamma_s}{1-\gamma_s}}\right]$.
As a consequence, this case is excluded if there exists a matrix Z with $\Phi(Z) = Y$ and sparsity structure S_s of order $k_s < K_s - \frac{2\gamma_s}{1-\gamma_s}$.

Proof. (i) Since $\epsilon_s = 0, s \in [t]$ and by monotonicity, there exist $n_s \in \mathbb{N}$ such that $\epsilon_s^{(n_s)} < \overline{\epsilon}$, but $\epsilon_s^{(n_s-1)} \geq \overline{\epsilon}$. By the definitions of $\epsilon_s^{(n)}$, this can only happen if the third terms of the definitions are the minima. This, on the other hand, means that the n_s are equal for all $s \in [t]$. We denote this number as n_0 in the following. It holds that the sequences $(\epsilon_s^{(n)})_{n\in\mathbb{N}}, s \in [t]$ are equal for all $n \geq n_0$. Thus, we can define a sequence $(\epsilon^{(n)})_{n\geq n_0}$ with $\epsilon^{(n)} := \epsilon_s^{(n)}, s \in [t]$.

We first assume that there exists an $\bar{n} \in \mathbb{N}$ such that $\epsilon^{(\bar{n})} = 0$. Then we have $X^{(\bar{n})} = \bar{X}$ and it holds that $\beta_{K_s+1}(\bar{X})_{N_s} = 0$. Otherwise, we have that $\epsilon^{(n)} > 0$ for all $n \in \mathbb{N}$. In this case, there exists a subsequence $(n_l)_{l\in\mathbb{N}}$ of $(n)_{n\geq n_0}$ such that $\epsilon^{(n_l+1)} < \epsilon^{(n_l)}$ for all $l \in \mathbb{N}$. By Lemma 5.4, $(X^{(n_l+1)})_l$ is bounded and we can extract a further subsequence, which we denote again by $(X^{(n_l+1)})_l$, and that converges to some $\bar{X} := \lim_{l\to\infty} X^{(n_l+1)}$. Since $\lim_{l\to\infty} \epsilon^{(n_l+1)} = 0$, it also follows that $\lim_{l\to\infty} \beta_{K_s+1}(X^{(n_l)})_{N_s} = 0$. Moreover, by continuity of the non-increasing rearrangement resp. $K_i + 1$-th singular value, we get that $\beta_{K_s+1}(\bar{X})_{N_s} = 0$ and thus $\bar{X}$ is a solution to $\Phi(X) = Y$ with sparsity structure s_s of order K_s.

We now show that the whole sequence converges to $\bar{X}$. According to (4.25), for $n \geq n_0$, it holds that

$$\mathcal{J}_{GIRLS}(X^{(n)}, (\epsilon_s^{(n)})_{s=1,\dots,t}, (W_s^{(n)})_{s=1,\dots,t}) = \sum_{s=1}^{t} \lambda_s f_{N_s}^{\epsilon_s^{(n)}}(X^{(n)}). \tag{5.24}$$

Since $(X^{(n_l+1)}) \xrightarrow{l\to\infty} \bar{X}$ and $\epsilon^{(n_l+1)} \xrightarrow{l\to\infty} 0$,

$$\mathcal{J}_{GIRLS}(X^{(n)}, (\epsilon_s^{(n)})_{s=1,\dots,t}, (W_s^{(n)})_{s=1,\dots,t}) \xrightarrow{l\to\infty} \sum_{s=1}^{t} \lambda_s \|\bar{X}\|_{N_s}. \tag{5.25}$$

By the non-increasing monotonicity property stated in Lemma 5.3, that the same holds true for the whole sequence $(X^{(n)})_{n\geq\ell_0}$, i.e., $\mathcal{J}_{GIRLS}(X^{(n)}, (\epsilon_s^{(n)})_{s=1,\dots,t}, (W_s^{(n)})_{s=1,\dots,t}) \xrightarrow{n\to\infty} \sum_{s=1}^{t} \lambda_s \|\bar{X}\|_{N_s}$. By introducing the structure-dependent dimension parameter d_s, where

$$d_s = \begin{cases} d_1 \cdot d_2 & \text{for } \|\cdot\|_{\ell_{p_s}}, \\ d_1 & \text{for } \|\cdot\|_{\ell_{2,p_s}}, \\ d_2 & \text{for } \|\cdot\|_{\ell_{p_s,2}}, \\ \min(d_1, d_2) & \text{for } \|\cdot\|_{\ell_{S_{p_s}}}, \end{cases}$$

we see that

$$\mathcal{J}_{GIRLS}^{(n)} - \sum_{s=1}^{t} d_s \lambda_s (\epsilon^{(n)})^{p_s} \leq \sum_{s=1}^{t} \lambda_s \|X^{(n)}\|_{N_s} \leq \mathcal{J}_{GIRLS}^{(n)}, \tag{5.26}$$

for $n \geq n_0$. Since $\lim_{n\to\infty} \sum_{s=1}^{t} d_s \lambda_s (\epsilon^n)^{p_s} = 0$, we conclude that also

$$\sum_{s=1}^{t} \lambda_s \|X^{(n)}\|_{N_s} \xrightarrow{n\to\infty} \sum_{s=1}^{t} \lambda_s \|\bar{X}\|_{N_s}. \tag{5.27}$$

Now, it remains to show that $X^{(n)} \to \bar{X}$.

By Lemma 4.13

$$\|\bar{X} - X^{(n)}\|_{N_s} \leq \frac{1+\gamma_s}{1-\gamma_s} \left(\|\bar{X}\|_{N_s} - \|X^{(n)}\|_{N_s} + 2\beta_{K_s+1}(\bar{X})_{N_s} \right)$$

and therefore,

$$\begin{aligned} \min_{s=1,\ldots,t} \frac{1-\gamma_s}{1+\gamma_s} \sum_{s=1}^{t} \lambda_s \|\bar{X} - X^{(n)}\|_{N_s} &\leq \sum_{s=1}^{t} \lambda_s (\|\bar{X}\|_{N_s} - \|X^{(n)}\|_{N_s}) \\ &= \sum_{s=1}^{t} \lambda_s \|\bar{X}\|_{N_s} - \sum_{s=1}^{t} \lambda_s \|X^{(n)}\|_{N_s}, \end{aligned}$$

as $\beta_{K_s+1}(\overline{X}) = 0$ due to the fact $\mathcal{N}_i(\overline{X}) = 0$. Thus, we can summarize

$$\begin{aligned} \min_{s=1,\cdots,t} \|\bar{X} - X^{(n)}\|_F^{p_s} &\leq \frac{1}{\sum_{s=1}^{t} \lambda_s} \sum_{s=1}^{t} \lambda_s \|\bar{X} - X^{(n)}\|_F^{p_s} \leq \frac{1}{\sum_{s=1}^{t} \lambda_s} \sum_{s=1}^{t} \lambda_s \|\bar{X} - X^{(n)}\|_{N_s} \\ &\leq \frac{\min\limits_{s=1,\cdots,t} \frac{1+\gamma_s}{1-\gamma_s}}{\sum_{s=1}^{t} \lambda_s} \left(\sum_{s=1}^{t} \lambda_s \|\bar{X}\|_{N_s} - \sum_{s=1}^{t} \lambda_s \|X^{(n)}\|_{N_s} \right) \\ &:= C_0 \left(\sum_{s=1}^{t} \lambda_s \|\bar{X}\|_{N_s} - \sum_{s=1}^{t} \lambda_s \|X^{(n)}\|_{N_s} \right). \end{aligned}$$

Taking the limit $n \to \infty$ gives $\lim\limits_{n\to\infty} \|\bar{X} - X^{(n)}\|_F = 0$. Therefore, it follows that $X^{(n)} \to \bar{X}$.

To obtain (5.23), we first consider a variable Z with $\Phi(Z) = Y$ and use Lemma 4.13

$$\begin{aligned} \sum_{s=1}^{t} \lambda_s \|Z - \bar{X}\|_{N_s} &\leq \sum_{s=1}^{t} \frac{\lambda_s(1+\gamma_s)}{(1-\gamma_s)} \left(\|\bar{X}\|_{N_s} - \|Z\|_{N_s} + 2\beta_{K_s}(Z)_{N_s} \right) \\ &\leq \sum_{s=1}^{t} \frac{2\lambda_s(1+\gamma_s)}{(1-\gamma_s)} \beta_{K_s}(Z)_{N_s}, \end{aligned} \tag{5.28}$$

where the second inequality follows from the fact that $\bar{X}$ is the unique minimizer of (5.4).

In the case, where $p_s = p$ for all $s \in [t]$, we only need that the NSP corresponding to only one of the structures s_s holds.

We first observe that for all Z holds that $\frac{\min_s d_s^{\frac{1}{2}-\frac{1}{p}}}{\sum_{s=1}^t \lambda_s} \sum_{s=1}^t \lambda_s \|Z\|_{N_s} \leq \|Z\|_F^p \leq \frac{1}{\sum_{s=1}^t \lambda_s} \sum_{s=1}^t \lambda_s \|Z\|_{N_s}$.
Therefore, we also have $\|X^{(n)}\|_F^p \xrightarrow{n\to\infty} \|\bar{X}\|_F^p$. By the equivalence of (quasi-) norms we can conclude that also $\|X^{(n)}\|_{N_s} \xrightarrow{n\to\infty} \|\bar{X}\|_{N_s}$. Using again Lemma 4.13, we get

$$\|\bar{X} - X^{(n)}\|_{N_s} \leq \frac{1+\gamma_s}{1-\gamma_s} \left(\|\bar{X}\|_{N_s} - \|X^{(n)}\|_{N_s} + 2\beta_{k+1}(\bar{X})_{N_s}\right)$$

and obtain immediately

$$\begin{aligned} \|\bar{X} - X^{(n)}\|_F^p &\leq \frac{1+\gamma_s}{1-\gamma_s} \left(\|\bar{X}\|_{N_s} - \|X^{(n)}\|_{N_s}\right) \\ &:= C_0 \left(\|\bar{X}\|_{N_s} - \|X^{(n)}\|_{N_s}\right). \end{aligned}$$

Taking the limit $n \to \infty$ yields again $\lim_{n\to\infty} \|\bar{X} - X^{(n)}\|_F = 0$ and proves the convergence with less restrictive assumptions.

(ii) We shall first show that $X^{(n)} \to X^\epsilon$ for $n \to \infty$ with X^ϵ being a stationary point of $F^\epsilon(X)$. We already observed that $(X^{(n)})_{n\in\mathbb{N}_0}$ is a bounded sequence and, hence, this sequence has accumulation points. Let $(X^{(n_l)})_{l\in\mathbb{N}_0}$ be any convergent subsequence of $(X^{(n)})_{n\in\mathbb{N}_0}$ and $\bar{X}$ its limit. We want to show that $\bar{X} = X^\epsilon$.
Since $W_1(X, \epsilon)$, as defined in Lemma 5.7, depends continuously on X and ϵ, it follows that $\lim_{l\to\infty} W_1^{(n)} = \lim_{l\to\infty} W_1(X^{(n_l)}, (\epsilon_s^{n_l})_{s=1,\dots,t}) := W_1(\bar{X}, \epsilon) = \overline{W_1}$.
On the other hand, by invoking Lemma 5.5, we obtain also that $X^{(n_l+1)} \to \bar{X}, l \to \infty$ and, therefore, also $\lim_{l\to\infty} W_1^{(n_l+1)} = \overline{W_1}$. We observe that with the minimality property of $X^{(n_l+1)}$ the KKT conditions for the optimization problem in (5.13) are fulfilled, i.e.,

$$W_1^{(n_l)} X^{(n_l+1)} \in \operatorname{Ran}(\Phi^*) \text{ and } \Phi(X^{(n_l+1)}) = Y.$$

This implies that there exists $\theta \in \mathbb{R}^m$ such that $W_1^{(n_l+1)} X^{(n_l)} = \Phi^*(\theta)$. Note that for all $\eta \in \mathcal{N}(\Phi)$ and all $n_l, l > 0$,

$$\langle X_{\text{vec}}^{(n_l+1)}, \eta_{\text{vec}} \rangle_{W_1^{(n_l)}} = \langle W_1^{(n_l)} X_{\text{vec}}^{(n_l+1)}, \eta_{\text{vec}} \rangle = \langle \Phi^*(\theta), \eta_{\text{vec}} \rangle = \langle \theta, \Phi(\eta) \rangle = 0.$$

Consequently, $\langle \bar{X}_{\text{vec}}, \eta_{\text{vec}} \rangle_{\overline{W_1}} = \lim_{l\to\infty} \langle X_{\text{vec}}^{(n_l+1)}, \eta_{\text{vec}} \rangle_{W_1^{(n_l)}} = 0$. By Lemma 5.7, this implies that $\bar{X}$ is a stationary point of F^ϵ and even coincides with the unique minimizer X^ϵ in the convex case.
(iii) To prove the error estimate stated in (iii), we first observe that with the minimizing

property of X^ϵ, for every $\tilde{Z}$ with $\Phi(\tilde{Z}) = Y$, we have

$$\sum_{s=1}^{t} \lambda_s \|X^\epsilon\|_{N_s} \leq F^\epsilon(X^\epsilon) \leq F^\epsilon(\tilde{Z}) \leq \sum_{s=1}^{t} \lambda_s \|\tilde{Z}\|_{N_s} + \sum_{s=1}^{t} d_s \lambda_s \epsilon_s^{p_s}.$$

Hence we obtain

$$\sum_{s=1}^{t} \lambda_s \|X^\epsilon\|_{N_s} - \sum_{s=1}^{t} \lambda_s \|\tilde{Z}\|_{N_s} \leq \sum_{s=1}^{t} d_s \lambda_s \epsilon_s^{p_s}$$

Furthermore, using Lemma 4.13, we have that

$$\sum_{s=1}^{t} \lambda_s \|X^\epsilon - \tilde{Z}\|_{N_s} \leq \sum_{s=1}^{t} \lambda_s \frac{1+\gamma_s}{1-\gamma_s} (d_s \epsilon_s^{p_s} + 2\beta_{k_s}(\tilde{Z})_{N_s}). \tag{5.29}$$

From the definition of ϵ, we obtain

$$\begin{aligned}
\sum_{s=1}^{t} \lambda_s \epsilon_s^{p_s} &= \lim_{n\to\infty} \sum_{s=1}^{t} \lambda_s (\epsilon_s^{(n)})^{p_s} \\
&\leq \begin{cases} \lim\limits_{n\to\infty} \sum_{s=1}^{t} \lambda_s (r_{k_s}(X^{(n)}))^{p_s} & \epsilon_s \geq \tilde{\epsilon} \text{ for all } s \in [t] \\ \lim\limits_{n\to\infty} \sum_{s=1}^{t} \lambda_s \max\limits_s (r_{k_s}(X^{(n)}))^{p_s} & \epsilon_s < \tilde{\epsilon} \text{ for any } s \in [t] \end{cases} \\
&= \begin{cases} \sum_{s=1}^{t} \lambda_s (r_{k_s}(X^{\epsilon}))^{p_s} & \epsilon_s \geq \tilde{\epsilon} \text{ for all } s \in [t] \\ \sum_{s=1}^{t} \lambda_s \max\limits_s (r_{k_s}(X^{\epsilon}))^{p_s} & \epsilon_s < \tilde{\epsilon} \text{ for any } s \in [t]. \end{cases}
\end{aligned}$$

However, it is easy to see that $|\beta_{k_s}(Z)_{N_s} - \beta_{k_s}(Z')_{N_s}| \leq \|Z - Z'\|_{N_s}$. From this observation and Lemma 4.13, we conclude that

$$\begin{aligned}
&(K_s + 1 - k_s) d_s \epsilon_s^{p_s} \\
&\leq \begin{cases} (K_s + 1 - k_s)(r_{k_s}(X^\epsilon))^{p_s} & \epsilon_s \geq \tilde{\epsilon} \text{ for all } s \in [t], \\ (K_s + 1 - k_s) \max\limits_s (r_{k_s}(X^\epsilon))^{p_s} & \epsilon_s < \tilde{\epsilon} \text{ for any } s \in [t] \end{cases} \\
&\leq \begin{cases} (\|X^\epsilon - \tilde{Z}\|_{N_s} + \beta_{k_s}(\tilde{Z})_{N_s}) & \epsilon_s \geq \tilde{\epsilon} \text{ for all } s \in [t], \\ \max\limits_s (\|X^\epsilon - \tilde{Z}\|_{N_s} + \beta_{k_s}(\tilde{Z})_{N_s}) & \epsilon_s < \tilde{\epsilon} \text{ for any } s \in [t] \end{cases} \\
&\leq \begin{cases} (\frac{1+\gamma_s}{1-\gamma_s} [d_s \epsilon_s^{p_s} + 2\beta_{k_s}(\tilde{Z})_{N_s}] + \beta_{k_s}(\tilde{Z})_{N_s}) & \epsilon_s \geq \tilde{\epsilon} \text{ for all } s \in [t], \\ (\frac{1+\gamma_s}{1-\gamma_s} \max\limits_s [d_s \epsilon_s^{p_s} + 2\beta_{k_s}(\tilde{Z})_{N_s} + \beta_{k_s}(\tilde{Z})_{N_s}]) & \epsilon_s < \tilde{\epsilon} \text{ for any } s \in [t]. \end{cases}
\end{aligned}$$

Since we assumed that $K_s - k_s \geq \frac{2\gamma_s}{1-\gamma_s}$,

$$d_s \epsilon_s^{p_s} + 2\beta_{k_s}(\tilde{Z})_{N_s} \leq \frac{2(K_s - k_s) + 3}{(K_s - k_s) - \frac{2\gamma_s}{1-\gamma_s}} \beta_{k_s}(\tilde{Z})_{N_s}$$

in the case $\epsilon_s \geq \tilde{\epsilon}$, for all $s \in [t]$ and

$$d_s \epsilon_s^{p_s} + 2\beta_{k_s}(\tilde{Z})_{N_s} \leq \max_s d_s[(\epsilon_s^{p_s}) + 2(\beta_{k_s}(\tilde{Z})_{N_s})] \leq \frac{2(K_s - k_s) + 3}{(K_s - k_s) - \frac{2\gamma_s}{1-\gamma_s}} \max_s(\beta_{k_s}(\tilde{Z})_{N_s})$$

in the case $\epsilon_s < \tilde{\epsilon}$ for any $s \in [t]$. Plugging this into (5.29) and continuing as in the calculation (5.28) gives

$$\begin{aligned}
&\sum_{s=1}^{t} \lambda_s \|Z - \bar{X}\|_{N_s} \leq \sum_{s=1}^{t} \lambda_s (\|Z - \tilde{Z}\|_{N_s} + \|\tilde{Z} - \bar{X}\|_{N_s}) \\
&\leq \sum_{s=1}^{t} \lambda_s (\sigma_{\min}(\tilde{\Phi}))^{-1} \|\hat{Y} - \tilde{Y}\|_{N_s} + \sum_{s=1}^{t} \frac{\lambda_s(1+\gamma_s)}{(1-\gamma_s)} \left(\|\bar{X}\|_{N_s} - \|\tilde{Z}\|_{N_s} + 2\beta_{k_s}(\tilde{Z})_{N_s} \right) \\
&\leq \begin{cases} \sum_{s=1}^{t} \lambda_s \tilde{C}_s \beta_{k_s}(Z)_{N_s} & \epsilon_s \geq \tilde{\epsilon} \text{ for all } s \in [t] \\ \sum_{s=1}^{t} \lambda_s \tilde{C}_s \max\limits_s \beta_{k_s}(Z)_{N_s} & \epsilon_s < \tilde{\epsilon} \text{ for any } s \in [t] \end{cases}
\end{aligned} \tag{5.30}$$

□

Remark 5.9. The theoretical error bound obtained in (iii) does not outperform the bound obtained by the minimization of one norm, i.e., the exploitation of only one structure. However, numerical experiments in the thesis of Kümmerle [84] show that practically the recovery error is significantly lower for Algorithm 5 than using only one structure (except of cases, where one structure is extremely dominating).

5.4 Algorithm formulation for the unconstrained case

Analogously to the constrained case, we also want to introduce a similar auxiliary functional that will be helpful for the formulation of a second version of IRLS for problems of type (5.5), assuming that $m = m_1 \cdot m_2$ and that $\Phi(X) - Y \in \mathbb{R}^m$ can be reshaped into a $m_1 \times m_2$-matrix:

Definition 5.10. Given $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m, Y \in \mathbb{R}^m$. Consider $\| \cdot \|_{N_s}, \| \cdot \|_{N_r}$ and the corresponding real numbers $\epsilon_s > 0, \epsilon_r > 0$, and weight matrices $W_s \in \mathbb{R}^{d_1 \cdot d_2 \times d_1 \cdot d_2}, W_r \in \mathbb{R}^{m \times m}$ for $s \in [t]$ derived from the norms $\| \cdot \|_{N_s}$ and $\| \cdot \|_{N_r}$. Set $W_1 = \sum_{s=1}^{n_s} \lambda_s p_s W_s$ and $W_2 = \mu p_r W_r$. We define the following auxiliary functional

$$\begin{aligned} &\mathcal{J}_{GIRLS2}(X, (\epsilon_s)_{s=1,\dots,t}, \epsilon_r, (W_s)_{s=1,\dots,t}, W_r) := \\ &\frac{1}{2}\Big(\|X_{\text{vec}}\|^2_{\ell_2(W_1)} + \sum_{s=1}^{t} \|\epsilon_s \cdot \mathbf{1}_{d_1 \cdot d_2}\|^2_{\ell_2(\lambda_s p_s W_s)} + (2 - p_s)\lambda_s \|W_s^{p_s/(p_s-2)}\|^2_{\ell_2} \\ &+ \|\Phi(X) - Y\|^2_{\ell_2(W_2)} + \|\epsilon_r \cdot \mathbf{1}_m\|^2_{\ell_2(W_2)} + (2 - p_r)\mu \|W_r^{p_r/(p_r-2)}\|^2_{\ell_2} \Big) \end{aligned} \tag{5.31}$$

Again, we define an additional auxiliary variable for the formulation of the algorithm

$$\mathcal{N}_r(X) = \begin{cases} r_{K_s+1}\left(\Phi(X) - Y\right)/m^{1/p_r}, & \text{for } \| \cdot \|_{N_r} = \| \cdot \|^{p_r}_{\ell_{p_r}}, \\ r_{K_s+1}\left(\left(\sum\limits_{j=1}^{m_2} (((\Phi(X) - Y)_{ij})^2 \right)^{1/2} \right)/d_1^{1/p_r}, & \text{for } \| \cdot \|_{N_r} = \| \cdot \|^{p_s}_{\ell_{2,p_s}}, \\ r_{K_s+1}\left(\left(\sum\limits_{i=1}^{m_1} ((\Phi(X) - Y)_{ij})^2 \right)^{1/2} \right)/d_2^{1/p_r}, & \text{for } \| \cdot \|_{N_r} = \| \cdot \|^{p_r}_{\ell_{p_r,2}}, \\ \sigma_{K_s+1}(\Phi(X) - Y)/(\min(m_1, m_2))^{1/p_r}, & \text{for } \| \cdot \|_{N_s} = \| \cdot \|^{p_r}_{S_{p_r}}, \end{cases}$$

and

$$\mathcal{M}_r(X) = \begin{cases} r_m\left(\Phi(X) - Y\right)/m^{1/p_r}, & \text{for } \| \cdot \|_{N_r} = \| \cdot \|^{p_r}_{\ell_{p_r}}, \\ r_{m_1}\left(\left(\sum\limits_{j=1}^{m_2} ((\Phi(X) - Y)_{ij})^2 \right)^{1/2} \right)/m_1^{1/p_r}, & \text{for } \| \cdot \|_{N_r} = \| \cdot \|^{p_s}_{\ell_{2,p_s}}, \\ r_{m_2}\left(\left(\sum\limits_{i=1}^{m_1} (((\Phi(X) - Y)_{ij})^2 \right)^{1/2} \right)/m_2^{1/p_r}, & \text{for } \| \cdot \|_{N_r} = \| \cdot \|^{p_s}_{\ell_{p_s,2}}, \\ \sigma_{\min(m_1,m_2)}(\Phi(X) - Y)/(\min(m_1, m_2))^{1/p_r}, & \text{for } \| \cdot \|_{N_r} = \| \cdot \|^{p_r}_{S_{p_r}}. \end{cases}$$

Moreover, let

$$\widetilde{\mathcal{J}}_{GIRLS2}(W)_u^{(n+1)} = \mathcal{J}_{GIRLS}(X^{(n+1)}, (\epsilon_u^{(n+1)})_{u=1,\dots,t}, (W_u^{(n+1)})_{u=1,\dots,u-1}, W, (W_u^{(n)})_{u=u+1,\dots,t+1})$$

for $u \in [t+1]$ corresponding to $s \in [t]$ for $u \in [t]$ or r if $u = t+1$.

GIRLS2 again performs an alternating minimization of the functional $\mathcal{J}_{GIRLS2}$.

Algorithm 6 Generalized IRLS for structured matrices and residuals 2 (GIRLS2)

```
Input: $\Phi : M_{d_1 \cdot d_2} \to \mathbb{R}^m$, $Y = \Phi(X_0) \in \mathbb{R}^m$ for a ground truth matrix $X_0 \in M_{d_1 \times d_2}$,
        nonconvexity parameters $p_s$ for $s \in [t]$, $p_r$.
Output: $X^{(1)}, X^{(2)}, \ldots \in M_{d_1 \times d_2}$
Initialize $\epsilon_s^{(0)} = 1$, set $W_s^{(0)} = \lambda_s p_s \cdot I_{d_1 \cdot d_2}$ for $s \in [t]$ and $\epsilon_r^{(0)} = 1$, set $W_r^{(0)} = \mu p_r \cdot I_{m_1 \cdot m_2}$.
repeat
    $X^{(n+1)} = \underset{\Phi(X)=Y}{\arg\min}\, \mathcal{J}_{GIRLS2}(X, (\epsilon_s^{(n)})_{s=1,\ldots,t}, (W_s^{(n)})_{s=1,\ldots,t}, W_r^{(n)})$   (5.32)
        $= \underset{\Phi(X)=Y}{\arg\min}\, \|X_{\mathrm{vec}}\|_{\ell_2(W_1^{(n)})}^2 + \|\Phi(X) - Y\|_{\ell_2(W_2^{(n)})}^2$
        $= \left[\left(W_1^{(n)} + \Phi^* \circ W_2^{(n)} \circ \Phi\right)^{-1} \circ \Phi^* \circ W_2^{(n)}(Y)\right]_{\mathrm{mat}}$
    for $s \in [t]$   (5.33)
    $\epsilon_s^{(n+1)} = \min\left(\epsilon_s^{(n)}, \max(\mathcal{N}_s(X^{(n+1)}), \tilde{\epsilon}), \max_s(\mathcal{N}_s(X^{(n+1)}))\right)$ and   (5.34)
    $\epsilon_r^{(n+1)} = \min\left(\epsilon_r^{(n)}, \max(\mathcal{N}_r^{(n+1)}, \tilde{\epsilon}), \mathcal{M}_r^{(n+1)}\right)$ with $\tilde{\epsilon} > 0$   (5.35)
    $W_u^{(n+1)} = \begin{cases} \underset{W>0, W \text{ diag}}{\arg\min}\, \widetilde{\mathcal{J}}_{GIRLS}(W)_u^{(n+1)}, & \text{for } \|\cdot\|_{N_u} = \|\cdot\|_{\ell_{p_u}}^{p_u}, \\ \underset{W=\mathbf{I}_{d_2}\otimes\mathbf{W}, \mathbf{W}>0, \mathbf{W}\text{diag}}{\arg\min}\, \widetilde{\mathcal{J}}_{GIRLS2}(W)_u^{(n+1)}, & \text{for } \|\cdot\|_{N_u} = \|\cdot\|_{\ell_{2,p_u}}^{p_u}, \\ \underset{W=\mathbf{W}\otimes\mathbf{I}_{d_1}, \mathbf{W}>0, \mathbf{W}\text{diag}}{\arg\min}\, \widetilde{\mathcal{J}}_{GIRLS2}(W)_u^{(n+1)}, & \text{for } \|\cdot\|_{N_u} = \|\cdot\|_{\ell_{2,p_u}}^{p_u}, \\ \underset{W=\mathbf{I}_{d_2}\otimes\mathbf{W}, \mathbf{W}>0}{\arg\min}\, \widetilde{\mathcal{J}}_{GIRLS2}(W)_u^{(n+1)}, & \text{for } \|\cdot\|_{N_u} = \|\cdot\|_{S_{p_u}}^{p_u}, \\ \underset{W=\mathbf{W}\otimes\mathbf{I}_{d_1}, \mathbf{W}>0}{\arg\min}\, \widetilde{\mathcal{J}}_{GIRLS2}(W)_u^{(n+1)}, & \text{for } \|\cdot\|_{N_u} = \|\cdot\|_{S_{p_u}}^{p_u} \end{cases}$   (5.36)
        $= \begin{cases} W_{s,\epsilon_s^{(n+1)}}(X^{(n+1)}) \text{ as defined in (2.76)} & u \in [t], \\ W_{r,\epsilon_r^{(n+1)}}((\Phi(X^{(n+1)}) - Y) \text{ as defined in (2.76)} & u = t+1. \end{cases}.$
    $W_1^{(n+1)} = \sum_{s=1}^{t} \lambda_s p_s W_s^{(n+1)}$ and $W_2^{(n+1)} = \mu p_r W_r^{(n+1)}$.   (5.37)
    $n = n + 1$.
until stopping criterion is met.;
Set $n_0 = n$.
```

We stop the algorithm if $\epsilon_s = 0$, for $s \in [t]$ and $\epsilon_r = 0$. In this case, we define $X^{(j)} := X^{(n)}$ for $j > n$. However, in general, the algorithm will generate an infinite sequence $(X^{(n)})_{n\in\mathbb{N}}$ of distinct vectors and it is convenient to keep the variables (ϵ_s, W_s)

or (ϵ_r, W_r) fixed, as soon as ϵ_s or ϵ_r, respectively falls below an appropriately chosen threshold and only continue updating the others.

5.5 Theoretical analysis and convergence results for the unconstrained case

The preliminary results obtained for Algorithm 5 can be deduced in an analogous manner, e.g., by using the monotonicity property and the boundedness of the iterates. This is why we omit the details except for the parts that demand for significant adaptions.

5.5.1 Preliminary results

We start with the result that the iterates of Algorithm 6 are coming arbitrarily close for $n \to \infty$.

Lemma 5.11. *Fix $Y \in \mathbb{R}^m$. Let $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m$ with $\sigma_{\min}(\Phi) > 0$. For the iterates of Algorithms 6, it holds that*

$$\lim_{n\to\infty} \|X^{(n)} - X^{(n+1)}\|_F^2 = 0.$$

Proof. For each $n = 1, 2, ...$, by monotonicity we have

$$\begin{aligned}
&2\left[\mathcal{J}_{GIRLS2}^{(n)} - \mathcal{J}_{GIRLS2}^{(n+1)}\right] \\
&\geq 2\left[\mathcal{J}_{GIRLS2}^{(n)} - \mathcal{J}_{GIRLS2}(X^{(n+1)}, \left(\epsilon_s^{(n)}\right)_{s=1,\ldots,t}, \epsilon_r^{(n)}, \left(W_s^{(n)}\right)_{s=1,\ldots,t}, W_r^{(n)})\right] \\
&= \|X_{\text{vec}}^{(n)}\|_{\ell_2(W_1^{(n)})}^2 - \|X_{\text{vec}}^{(n+1)}\|_{\ell_2(W_1^{(n)})}^2 + \|\Phi(X^{(n)}) - Y\|_{\ell_2(W_2^{(n)})}^2 - \|\Phi(X^{(n+1)}) - Y\|_{\ell_2(W_2^{(n)})}^2.
\end{aligned}$$

Moreover, if we exploit the convexity of the functional J_{GIRLS2} in X and the minimality property of $X^{(n+1)}$, we obtain that

$$\begin{aligned}
&\|X_{\text{vec}}^{(n)}\|_{\ell_2(W_1^{(n)})}^2 - \|X_{\text{vec}}^{(n+1)}\|_{\ell_2(W_1^{(n)})}^2 + \|\Phi(X^{(n)}) - Y\|_{\ell_2(W_2^{(n)})}^2 - \|\Phi(X^{(n+1)}) - Y\|_{\ell_2(W_2^{(n)})}^2 \\
&\geq \|X_{\text{vec}}^{(n+1)}\|_{\ell_2(W_1^{(n)})}^2 + 2\langle X_{\text{vec}}^{(n+1)}, (X^{(n)} - X^{(n+1)})_{\text{vec}}\rangle_{W_1^{(n)}} \\
&+ (X^{(n)} - X^{(n+1)})_{\text{vec}}^T W_1^{(n)} (X^{(n)} - X^{(n+1)})_{\text{vec}} - \|X_{\text{vec}}^{(n+1)}\|_{\ell_2(W_1^{(n)})}^2 \\
&+ \|\Phi(X^{(n+1)}) - Y\|_{\ell_2(W_2^{(n)})}^2 + 2\langle \Phi(X^{(n+1)}) - Y, \Phi(X^{(n)} - X^{(n+1)})\rangle_{W_2^{(n)}} \\
&+ [\Phi(X^{(n)} - X^{(n+1)})]^T W_2^{(n)} [\Phi(X^{(n)} - X^{(n+1)})] - \|\Phi(X^{(n+1)}) - Y\|_{\ell_2(W_2^{(n)})}^2
\end{aligned}$$

$$\begin{aligned}&= (X^{(n)} - X^{(n+1)})^T_{\mathrm{vec}} W_1^{(n)} (X^{(n)} - X^{(n+1)})_{\mathrm{vec}} \\ &+ (\Phi(X^{(n)} - X^{(n+1)}))^T W_2^{(n)} \Phi(X^{(n)} - X^{(n+1)}).\end{aligned}$$

Next, we calculate $\sigma_{\min}(W_s^{(n)})$ and $\sigma_{\min}(W_r^{(n)})$ and continue the estimation as follows

$$\begin{aligned}&2\left[\mathcal{J}_{GIRLS2}^{(n)} - \mathcal{J}_{GIRLS2}^{(n+1)}\right] \\ &\geq \sum_{s=1}^{t} \lambda_s p_s \sigma_{\min}(W_s^{(n)}) \|X^{(n)} - X^{(n+1)}\|_F^2 \\ &+ \sum_{s=1}^{t} \mu p_r \sigma_{\min}(W_r^{(n)}) \|\Phi(X^{(n)} - X^{(n+1)})\|_{\ell_2}^2 \\ &\geq \sum_{s=1}^{t} \lambda_s p_s (\widetilde{\mathcal{J}}_{GIRLS2}^{(0)})^{1-\frac{2}{p_s}} \|X^{(n)} - X^{(n+1)}\|_F^2 \\ &+ \sum_{s=1}^{t} \mu p_r (\widetilde{\mathcal{J}}_{GIRLS2}^{(0)})^{1-\frac{2}{p_r}} \sigma_{\min}(\Phi)^4 \|X^{(n)} - X^{(n+1)}\|_F^2 \\ &= \left[\sum_{s=1}^{t} \lambda_s p_s (\widetilde{\mathcal{J}}_{GIRLS2}^{(0)})^{1-\frac{2}{p_s}} + \sum_{s=1}^{t} \mu p_r (\widetilde{\mathcal{J}}_{GIRLS2}^{(0)})^{1-\frac{2}{p_r}} \sigma_{\min}(\Phi)^4\right] \|X^{(n)} - X^{(n+1)}\|_F^2 \\ &:= \tilde{C} \|X^{(n)} - X^{(n+1)}\|_F^2.\end{aligned}$$

Again, by monotonicity and the boundedness of the sequence $\left(\mathcal{J}_{GIRLS2}^{(n)}\right)_{n\in\mathbb{N}}$ we know that

$$\lim_{n\to\infty} \left[\mathcal{J}_{GIRLS2}^{(n)} - \mathcal{J}_{GIRLS2}^{(n+1)}\right] = 0,$$

and, therefore, also

$$\lim_{n\to\infty} \|X^{(n)} - X^{(n+1)}\|_F^2 = 0.$$

□

Remark 5.12. The assumption on the singular values of Φ are very weak and, e.g., fulfilled for random matrices with high probability.

From the monotonicity of $(\epsilon_s^{(n)})_{s=1,\dots,t}$ and $\epsilon_r^{(n)}$, we know that $\epsilon_s := \lim_{n\to\infty} \epsilon_s^{(n)}$ and $\epsilon_r := \lim_{n\to\infty} \epsilon_r^{(n)}$ exist and are non-negative. We define $\epsilon := \left[(\epsilon_s)_{s=1,\dots,t}, \epsilon_r\right]$. The following functional will play a role in our proof of convergence, especially for $\epsilon > 0$.

Definition 5.13. (ϵ-perturbed objective functional for the unconstrained case)
We define the ϵ-perturbed objective functional to be of the following form

$$F_2^{\epsilon}(X) := \sum_{s=1}^{t} \lambda_s f_{N_s}^{\epsilon_s}(X) + f_{N_r}^{\epsilon_r}(\Phi(X) - Y)$$

and the corresponding minimization problem

$$\min_X = F_2^{\epsilon_1,\epsilon_2}(X). \tag{5.38}$$

Notice that, if we knew that $X^{(n)}$ converged to a point $\bar{X}$, then $F_2^{\epsilon}(\bar{X})$ would be the limit of $\mathcal{J}_{GIRLS2}\left(X^{(n)}, (\epsilon_s^{(n)})_{s=1,\dots,t}, (\epsilon_r^{(n)})_{r=1,\dots,n_2}, (W_s^{(n)})_{s=1,\dots,t}, (W_r^{(n)})_{r=1,\dots,n_2}\right)$ for $n \to \infty$. We denote by $\mathcal{Z}^{\epsilon}(\tilde{Y})$ the set of global minimizers Z of the functional $F_2^{\epsilon}(X)$ with $\Phi(Z) = Y$ and in the case that we consider a unique global minimizer in dependence of ϵ with

$$X^{\epsilon} \in \arg\min_X F_2^{\epsilon}(X), \tag{5.39}$$

5.5.2 Convergence results

Finally, we can state the convergence results for Algorithm 6 under the NSP conditions already mentioned above.

Theorem 5.14. *Let $\Phi : M_{d_1 \times d_2} \to \mathbb{R}^m, Y \in \mathbb{R}^m$ and the functionals $\mathcal{J}_{GIRLS2}\left(X^{(n)}, (\epsilon_s^{(n)})_{s=1,\dots,t}, (\epsilon_r^{(n)})_{r=1,\dots,n_2}, (W_s^{(n)})_{s=1,\dots,t}, W_r^{(n)}\right)$ be defined for $(\epsilon_s^{(n)})_{s=1,\dots,t}, \epsilon_r^{(n)}$, $(W_s^{(n)})_{s=1,\dots,t}, W_r^{(n)}$ as generated by Algorithm 6 for all $n \geq 0$. In the following, we want to assume that our desired solution matrix X_0 has property $s_s, s \in [t]$ of order $k_s, s \in [t]$ and $\Phi(X_0) - Y$ has property s_r. Let $\bar{\mathcal{Z}}^{\epsilon}$ be the set of accumulation points of the sequence $(X^{(n)})_{n\in\mathbb{N}}$ generated by Algorithm 6.*

(i) If $\epsilon = [(\epsilon_s)_{s=1,\dots,t}, \epsilon_r] = 0$, and

(a) the map Φ fulfills the corresponding NSP for structure s_s of order K_s, $s \in [t]$ as defined in Definition 4.10,

(b) $p_s = p, s \in [t]$ and the matrix Φ fulfills the NSP corresponding to structure s_s of order K_s

then $\bar{\mathcal{Z}}^{\epsilon}$ consists of one single point $\bar{X} = X_0$ with sparsity structure s_s of order $K_s, s \in [t]$ and $\bar{X}$ is the solution to the minimization problem (5.5).
Moreover, in case (a), we have for $k_s \leq K_s, s \in [t]$ and any Z with $\Phi(Z) = Y$ that

$$\sum_{s=1}^{t} \lambda_s \|Z - \bar{X}\|_{N_s} \leq \sum_{s=1}^{t} \hat{C}_s \beta_{k_s}(Z)_{N_s}, \tag{5.40}$$

where $\hat{C}_s = \frac{2\lambda_s(1+\gamma_s)}{(1-\gamma_s)}$.

(ii) If $\epsilon_1 = (\epsilon_s)_{s=1,\dots,t}$ with $\epsilon_s > 0$ for any s and $\epsilon_r > 0$, then every $\bar{X} \in \bar{\mathcal{Z}}^\epsilon$ is a stationary point of the ϵ-perturbed objective functional F_2^ϵ. In the case that F_2^ϵ is convex, i.e., $p_s, p_r \geq 1$ for $s \in [t]$ we assume for simplicity that X^ϵ is actually the unique global minimizer of F_2^ϵ. Then $\bar{X} = X^\epsilon$.

(iii) If, in addition to the assumptions in case (ii), the map Φ fulfills the corresponding NSPs for the structures s_s of order $K_s, s \in [t]$ as defined in Definition 4.10 with $\gamma_s < 1 - \frac{2}{K_s+2}$ (or equivalently, if $\frac{2\gamma_s}{1-\gamma_s} < K_s$), then we have, for all $\bar{X} \in \bar{\mathcal{Z}}^\epsilon \cap \mathcal{Z}^\epsilon$ and all Z with $\Phi(Z) = Y$ and any $k_s < K_s - \frac{2\gamma_s}{1-\gamma_s}$ that

$$\sum_{s=1}^{t} \lambda_s \|Z - \bar{X}\|_{N_s}$$
$$\leq \begin{cases} \sum_{s=1}^{t} \tilde{C}_s \lambda_s \beta_{k_s}(Z)_{N_s} + C_2 & \text{if } \epsilon_s \geq \tilde{\epsilon}, s \in [t], \\ \sum_{s=1}^{t} \tilde{C}_s \lambda_s \max\limits_{s} \beta_{k_s}(Z)_{N_s} + C_2 & \text{if } \exists s \text{ s.t. } \epsilon_s < \tilde{\epsilon} \end{cases}$$

with $\tilde{C}_s := \frac{2(1+\gamma_s)}{1-\gamma_s} \left[\frac{K_s - k_s + \frac{3}{2}}{K_s - k_s - \frac{2\gamma_s}{1-\gamma_s}} \right]$ and $C_2 = \sum_{s=1}^{t} \frac{1+\gamma_s}{1-\gamma_s} \mu d_r \tilde{\epsilon}$.
As a consequence, this case is excluded if there exists a matrix Z with $\Phi(Z) = Y$ and sparsity structure S_s of order $k_s < K_s - \frac{2\gamma_s}{1-\gamma_s}$.

The proof is in great parts analogous to Theorem 5.8 and we only give additional comments and hints.

Proof. (i) The first part of the proof follows the one of Theorem 5.8. From $\lim_{l\to\infty} \epsilon^{(n_l+1)} = 0$ and the definition of ϵ_r we deduce that the residuals $\Phi(\bar{X}) - Y$ vanish. Thus, $\bar{X}$ is a solution to $\Phi(X) = Y$ with sparsity structure s_s of order K_s.

(iii) To prove the error estimate stated in (iii), we first observe that by the minimizing property of X^ϵ for every Z with $\Phi(Z) = Y$, we have

$$\sum_{s=1}^{t} \lambda_s \|X^\epsilon\|_{N_s} + \mu \|\Phi(X^\epsilon) - Y\|_{N_r} \leq F_2^\epsilon(X^\epsilon) \leq F_2^\epsilon(Z)$$
$$\leq \sum_{s=1}^{t} \lambda_s \|Z\|_{N_s} + \mu \|\Phi(Z) - Y\|_{N_r} + \sum_{s=1}^{t} d_s \lambda_s \epsilon_s^{p_s} + d_r \mu \epsilon_r^{p_r}$$
$$= \sum_{s=1}^{t} \lambda_s \|Z\|_{N_s} + \sum_{s=1}^{t} d_s \lambda_s \epsilon_s^{p_s} + d_r \mu \epsilon_r^{p_r}.$$

From this we obtain that

$$\sum_{s=1}^{t} \lambda_s \|X^\epsilon\|_{N_s} - \sum_{s=1}^{t} \lambda_s \|Z\|_{N_s} \le \sum_{s=1}^{t} d_s \lambda_s \epsilon_s^{p_s} - \mu \|\Phi(X^\epsilon) - Y\|_{N_r} + d_r \mu \epsilon_r^{p_r}.$$

In the case that $\epsilon_r > \tilde{\epsilon}$, we have that

$$-\mu \|\Phi X^\epsilon_{\text{vec}} - Y\|_{N_r} + d_r \mu \epsilon_r^{p_r} < 0.$$

In the other case,

$$-\mu \|\Phi(X^\epsilon) - Y\|_{N_r} + d_r \epsilon_r^{p_r} < d_r \mu \tilde{\epsilon}^{p_r}.$$

By Lemma 4.13, we have that

$$\begin{aligned} \sum_{s=1}^{t} \lambda_s \|X^\epsilon - Z\|_{N_s} &\le \sum_{s=1}^{t} \lambda_s \frac{1+\gamma_s}{1-\gamma_s} (d_s \epsilon_s^{p_s} + \frac{\mu}{\sum_{s=1}^{t} \lambda_s} d_r \tilde{\epsilon}^{p_r} + 2\beta_{k_s}(Z)_{N_s}) \\ &\le \sum_{s=1}^{t} \lambda_s \frac{1+\gamma_s}{1-\gamma_s} (d_s \epsilon_s^{p_s} + 2\beta_{k_s}(Z)_{N_s}) + \sum_{s=1}^{t} \lambda_s \frac{1+\gamma_s}{1-\gamma_s} \frac{\mu}{\sum_{s=1}^{t} \lambda_s} d_r \tilde{\epsilon}_2^{p_r} \\ &\le \sum_{s=1}^{t} \lambda_s \frac{1+\gamma_s}{1-\gamma_s} (d_s \epsilon_s^{p_s} + 2\beta_{k_s}(Z)_{N_s}) + C((\gamma_s)_{s=1,\dots,t}, \mu, d_r, p_r, \tilde{\epsilon}) \\ &= \sum_{s=1}^{t} \lambda_s \frac{1+\gamma_s}{1-\gamma_s} (d_s \epsilon_s^{p_s} + 2\beta_{k_s}(Z)_{N_s}) + C_2. \end{aligned} \tag{5.41}$$

The rest of the proof follows the arguments of Theorem 5.8 (iii) and and this leads directly to the bound in (5.40). □

Bibliography

[1] A. Ahmed and J. Romberg. Compressive multiplexing of correlated signals. *IEEE Transactions on Information Theory*, 61(1):479–498, 2015.

[2] M. Artina, M. Fornasier, and F. Solombrino. Linearly constrained nonsmooth and nonconvex minimization. *SIAM Journal on Optimization*, 23(3):1904–1937, 2013.

[3] K. M. R. Audenaert. A generalisation of Mirsky's singular value inequalities. ArXiv preprint:1410.4941, 2014.

[4] A. Bagirov, N. Karmitsa, and M. M. Mkel. *Introduction to Nonsmooth Optimization: Theory, Practice and Software.* Springer Publishing Company, Incorporated, 2014.

[5] A. E. Beaton and J. W. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974.

[6] R. Bellman. *Dynamic programming.* Courier Corporation, 2013.

[7] James Bennett and Stan Lanning. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35, 2007.

[8] D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas (Second Edition).* Princeton University Press, 2009.

[9] J. D. Blanchard, J. Tanner, and K. Wei. CGIHT: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion. *Information and Inference*, 4(4):289–327, 2015. [using `CGIHT` ('Conjugate Gradient Iterative Hard Thresholding') algorithm, code provided directly by the authors.

[10] P. Bloom. Précis of how children learn the meanings of words. *Behavioral and brain Sciences*, 24(6):1095–1103, 2001.

[11] H. Boche, R. Calderbank, G. Kutyniok, and J. Vybiral. *Compressed sensing and its applications.* Springer, 2015.

[12] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, New York, NY, USA, 2004.

[13] D. A. Braun, C. Mehring, and D. M. Wolpert. Structure learning in action. *Behavioural brain research*, 206(2):157–165, 2010.

[14] C. Brezinski and L. Wuytack. *Numerical Analysis: Historical Developments in the 20th Century.* Elsevier Science, 2012.

[15] D. Butnariu and E. Resmerita. Bregmann distances, totally convex functions and a method for solving operator equations in banach spaces, 2005.

[16] G. Buttazzo. Semicontinuity, relaxation, and integral representation in the calculus of variations, 1989.

[17] V. Cambareri. *Matrix Designs and Methods for Secure and Efficient Compressed Sensing.* PhD thesis, University of Bologna, 2014.

[18] E. J. Candès, Y. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.

[19] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011.

[20] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[21] E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, April 2011.

[22] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[23] E. J. Candès, T. Strohmer, and V. Voroninski. PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[24] E. J. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[25] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[26] A. Chambolle and P.-L. Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188, 1997.

[27] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Sparse and low-rank matrix decompositions. In *Proceedings of the 47th Annual Allerton Conference on Communication, Control, and Computing*, Allerton'09, pages 962–967. IEEE Press, 2009.

[28] R. Chartrand and V. Staneva. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 24(035020):1–14, 2008.

[29] R. Chartrand and W. Yin. Iteratively reweighted algorithms for compressive sensing. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.*, pages 3869–3872, 2008.

[30] J. A. Chavez-Dominguez and D. Kutzarova. Stability of low-rank matrix recovery and its connections to Banach space geometry. *Journal of Mathematical Analysis and Applications*, 427(1):320 – 335, 2015.

[31] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Completing any low-rank matrix, provably. *Journal of Machine Learning Research*, 16:2999–3034, 2015.

[32] A.K. Cline. Rate of Convergence of Lawson's Algorithm. *Mathematics of Computation*, 26(117):167+, 1972.

[33] A. Cohen, W. Dahmen, and R.A. DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 21:211–231, 2009.

[34] B. Dacorogna. *Direct Methods in the Calculus of Variations.* Springer-Verlag New York, Inc., 1989.

[35] I. Daubechies, R. DeVore, M. Fornasier, and C.S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63:1–38, 2010.

[36] I. Daubechies, R. A. DeVore, M. Fornasier, and C. S. Güntürk. Iteratively reweighted least squares minimization: Proof of faster than linear rate for sparse recovery. In *42nd Annual Conference on Information Sciences and Systems, CISS 2008, Princeton, NJ, USA, 19-21 March 2008*, pages 26–29, 2008.

[37] J. Davenport, M. A.; Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10:608–622, 06 2016.

[38] M. Davies and R. Gribonval. On ℓ_p-minimisation, instance optimality, and restricted isometry constants for sparse approximation. In *SAMPTA proc*, 2009.

[39] L. Diening, M. Fornasier, and M. Wank. A relaxed ka\v {c} anov iteration for the p-poisson problem. *arXiv preprint arXiv:1702.03844*, 2017.

[40] D.L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289 –1306, 2006.

[41] J. Drenth. *Principles of Protein X-Ray Crystallography.* Springer, 2007.

[42] R. Dutter. *Robust regression: Different approaches to numerical solution and algorithms.* Fachgruppe für Statistik, Eidgenössische Techn. Hochsch., 1975.

[43] M. Ehler, M. Fornasier, and J. Sigl. Quasi-linear compressed sensing. *Multiscale Modeling and Simulation*, 12(2):725–754, 2014.

[44] Y. C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *Information Theory, IEEE Transactions on*, 55(11):5302–5316, 2009.

[45] H. Fang, S. Sakellaridi, and Y. Saad. Multilevel manifold learning with application to spectral clustering. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 419–428. ACM, 2010.

[46] M. Fazel. *Matrix rank minimization with applications.* PhD thesis, Electrical Engineering Department Stanford University, 2002.

[47] J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, 1982.

[48] M. Fornasier. Numerical methods for sparse recovery. In M. Fornasier, editor, *Theoretical Foundations and Numerical Methods for Sprase Recovery*, volume 9 of *Radon Series on Computational and Applied Mathematics*, pages 93–200. De Gruyter, Berlin, 2010.

[49] M. Fornasier, S. Peter, H. Rauhut, and S. Worm. Conjugate gradient acceleration of iteratively re-weighted least squares methods. *Computational Optimization and Applications*, pages 1–55, 2016.

[50] M. Fornasier and H. Rauhut. Compressive Sensing. In O. Scherzer, editor, *Handbook of Mathematical Methods in Imaging*, pages 187–228. Springer, 2011.

[51] M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM J. on Optimization*, 21(4):1614–1640, 2011.

[52] S. Foucart. Concave Mirsky inequality and low-rank recovery. http://www.math.tamu.edu/ foucart/publi/Mirsky.pdf, preprint, 2016.

[53] S. Foucart and M.-J. Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26(3):395 – 407, 2009.

[54] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing.* Applied and Numerical Harmonic Analysis. Springer New York, 2013.

[55] Y. Gao, J. Peng, S. Yue, and Y. Zhao. On the null space property of ℓ_q -minimization for $0 < q \leq 1$ in compressed sensing. *Journal of Function Spaces*, page 579853, 2015.

[56] E. M. Garau, P. Morin, and C. Zuppa. Convergence of an adaptive Kačanov FEM for quasi-linear problems. *Applied Numerical Mathematics*, 61(4):512–529, 2011.

[57] R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of the phase from image and diffraction plane pictures. *Optik (Jena)*, 35:237+, 1972.

[58] M. Giaquinta and G. Modica. *Mathematical Analysis: Linear and Metric Structures and Continuity.* Mathematical analysis. Birkhäuser Boston, 2007.

[59] B. P. Gibbs. *Linear Least-Squares Estimation: Solution Techniques*, pages 139–192. John Wiley & Sons, Inc., 2011.

[60] J.F. Giovannelli and J. Idier. *Regularization and Bayesian Methods for Inverse Problems in Signal and Image Processing.* ISTE. Wiley, 2015.

[61] P. Godfrey-Smith. *Theory and Reality: An Introduction to the Philosophy of Science.* University Of Chicago Press, first edition edition, 2003.

[62] D. Goel and D. Batra. Predicting user preference for movies using netflix database.

[63] I. Gohberg, S. Goldberg, and N. Krupnik. *Traces and determinants of linear operators*, volume 116. Springer, 2012.

[64] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[65] Z. Gong, Z. Shen, and K.-C. Toh. Image restoration with mixed or unknown noises. *Multiscale Modeling & Simulation*, 12(2):458–487, 2014.

[66] N. D. Goodman, T. D. Ullman, and J. B. Tenenbaum. Learning a theory of causality. *Psychological review*, 118(1):110, 2011.

[67] A. Gopnik, C. Glymour, D. M. Sobel, L. E. Schulz, T. Kushnir, and D. Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004.

[68] I.F. Gorodnitsky and B.D. Rao. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45:600–616, 1997.

[69] P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 149–192, 1984.

[70] R. Gribonval and M. Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Applied and Computational Harmonic Analysis*, 22(3):335–355, 2007.

[71] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[72] D. Gross, F. Krahmer, and R. Kueng. A partial derandomization of phaselift using spherical designs. *Journal of Fourier Analysis and Applications*, 21(2):229–266, 2015.

[73] D. Gross, Y. Liu, S. T. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Physical Review Letters*, 105:150401, 2010.

[74] J. P. Haldar and D. Hernando. Rank-constrained solutions to linear matrix equations using powerfactorization. *IEEE Signal Processing Letters*, 16(7):584–587, July 2009. [using `p_MC_AltMin` (Alternating Minimization) algorithm, partial code available at `http://mr.usc.edu/download/irpf/`].

[75] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[76] G. E. Hinton, P. Dayan, and M. Revow. Modeling the manifolds of images of handwritten digits. *IEEE transactions on Neural Networks*, 8(1):65–74, 1997.

[77] P.J. Huber. *Robust statistics.* Wiley New York, 1981.

[78] P. Jain, R. Meka, and I. S. Dhillon. Guaranteed rank minimization via singular value projection. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 937–945. Curran Associates, Inc., 2010.

[79] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 665–674, New York, NY, USA, 2013. ACM.

[80] A. Jameson. Solution of the equation ax + xb = c by inversion of an m x m or n x n matrix. *SIAM Journal on Applied Mathematics*, 16(5):1020–1023, 1968.

[81] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege. Stable low-rank matrix recovery via null space properties. *Information and Inference: A Journal of the IMA*, 5(4):405, 2016.

[82] S. Koziel and X.S. Yang. *Computational Optimization, Methods and Algorithms.* Studies in Computational Intelligence. Springer Berlin Heidelberg, 2011.

[83] F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904, 2014.

[84] C. Kümmerle. Learning ridge functions by low-rank and sparse optimization. Master's thesis, Technische Universität München, 2015.

[85] A. Kyrillidis and V. Cevher. Matrix recipes for hard thresholding methods. *Journal of Mathematical Imaging and Vision*, 48(2):235–265, 2014. [using `Matrix ALPS II` ('Matrix ALgrebraic PursuitS II') algorithm, code from `http://akyrillidis.github.io/projects/`].

[86] C. Kümmerle and J. Sigl. Harmonic mean iteratively reweighted least squares for low-rank matrix recovery. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 489–493, 2017.

[87] C. Kümmerle and J. Sigl. Harmonic mean iteratively reweighted least squares for low-rank matrix recovery. *ArXiv preprint:1703.05038*, 2017.

[88] Y. Lai, M.and Liu. The null space property for sparse recovery from multiple measurement vectors, 2010.

[89] K. Lange. *MM Optimization Algorithms*. SIAM, 2016.

[90] K. Lange, D. R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20, 2000.

[91] C.L. Lawson. *Contributions to the Theory of Linear Least Maximum Approximation*. University of California, Los Angeles–Mathematics., 1961.

[92] D. Q. Lee. Numerically efficient methods for solving least squares problems. 2012.

[93] J.A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer New York, 2007.

[94] K. Levenberg. A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, 2:164–168, 1944.

[95] Z. Li, W. Shi, X. Shi, and Z. Zhong. A supervised manifold learning method. *Computer Science and Information Systems*, 6(2):205–215, 2009.

[96] Z. Liu, A. Hansson, and L. Vandenberghe. Nuclear norm system identification with missing inputs and outputs. *Systems & Control Letters*, 62(8):605 – 612, 2013.

[97] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2010.

[98] H. Lu, X. Long, and J. Lv. A fast algorithm for recovery of jointly sparse vectors based on the alternating direction methods. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 461–469, 2011.

[99] S. Lu and S. V. Pereverzyev. Multi-parameter regularization and its numerical realization. *Numerische Mathematik*, 118(1):1–31, 2011.

[100] Y. Ma and Y. Fu. *Manifold Learning Theory and Applications*. CRC Press, 2011.

[101] J.R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley Series in Probability and Statistics: Texts and References Section. Wiley, 1999.

[102] S. Maji and J. Malik. Fast and accurate digit classification. Technical report, EECS Department, University of California, Berkeley, 2009.

[103] M. Malek-Mohammadi, M. Babaie-Zadeh, and M. Skoglund. Iterative concave rank approximation for recovering low-rank matrices. *IEEE Transactions on Signal Processing*, 62(20):5213–5226, Oct 2014.

[104] I.B.C. Matheson. A critical comparison of least absolute deviation fitting (robust) and least squares fitting: The importance of error distributions. *Computers & Chemistry*, 14(1):49 – 57, 1990.

[105] P. McCullagh. Quasi-likelihood functions. *The Annals of Statistics*, pages 59–67, 1983.

[106] K. Mohan and M. Fazel. Iterative reweighted least squares for matrix rank minimization. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 653–661, Sept 2010.

[107] B. Moore and B. Natarajan. A general framework for robust compressive sensing based nonlinear regression. In *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2012 IEEE 7th*, pages 225–228. IEEE, 2012.

[108] J.J. More. The levenberg-marquardt algorithm: Implementation and theory. In G.A. Watson, editor, *Numerical Analysis*, volume 630 of *Lecture Notes in Mathematics*, pages 105–116. Springer Berlin Heidelberg, 1978.

[109] K. P Murphy. *Machine learning: a probabilistic perspective*. The MIT press, 2012.

[110] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[111] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 1965.

[112] Inc. Netflix. Netflix prize. http://www.netflixprize.com, 2009.

[113] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

[114] P. Ochs, A. Dosovitskiy, T. Brox, and T. Pock. An iteratively reweighted algorithm for non-smooth non-convex optimization in computer vision. *Technical Report*, 2014.

[115] M.R. Osborne. *Finite Algorithms in Optimization and Data Analysis*. John Wiley & Sons, Inc., New York, NY, USA, 1985.

[116] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. Simultaneously Structured Models with Application to Sparse and Low-rank Matrices, 2014.

[117] S. Oymak, K. Mohan, M. Fazel, and B. Hassibi. A simplified approach to recovery conditions for low rank matrices. In *2011 IEEE International Symposium on Information Theory Proceedings, ISIT 2011, St. Petersburg, Russia, July 31 - August 5, 2011*, pages 2318–2322, 2011.

[118] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Finding Low-rank Solutions to Matrix Problems, Efficiently and Provably. *ArXiv preprints:1606.03168*, 2016. [using `BFGD` ('Bi-Factored Gradient Descent') algorithm, code from `http://akyrillidis.github.io/projects/`].

[119] D.L. Pimentel-Alarcón, N. Boston, and R. D. Nowak. A Characterization of Deterministic Sampling Patterns for Low-Rank Matrix Completion. *ArXiv preprints:1503.02596*, 2015.

[120] X. Pitkow and D. E. Angelaki. Inference in the brain: Statistics flowing in redundant population codes. *Neuron*, 94(5):943–953, 2017.

[121] S. Qaisar, R. M. Bilal, W. Iqbal, M. Naureen, and S. Lee. Compressive sensing: From theory to applications, a survey. *Journal of Communications and networks*, 15(5):443–456, 2013.

[122] N. V Queipo, R. T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, and P. K. Tucker. Surrogate-based analysis and optimization. *Progress in aerospace sciences*, 41(1):1–28, 2005.

[123] H. Rauhut. Compressive sensing and structured random matrices. In M. Fornasier, editor, *Theoretical Foundations and Numerical Methods for Sprase Recovery*, volume 9 of *Radon Series on Computational and Applied Mathematics*, pages 93–200. De Gruyter, 2010.

[124] R.Chartrand. Exact Reconstruction of Sparse Signals via Nonconvex Minimization. *Signal Processing Letters, IEEE*, 14(10):707–710, 2007.

[125] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[126] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[127] P. l. Resnick and H. R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[128] R.Gribonval and M.Nielsen. Sparse representations in unions of bases. *Information Theory, IEEE Transactions on*, 49(12):3320 – 3325, 2003.

[129] J. R. Rice and K. H. Usow. The lawson algorithm and extensions. *Mathematics of Computation*, 22(101):118–127, 1968.

[130] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:1025–1045, 2008.

[131] L.I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Journal of Physics D: Applied Physics*, 60(1-4):259–268, 1992.

[132] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* Pearson Education, 2 edition, 2003.

[133] S.G.Mallat and Z.Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397 –3415, 1993.

[134] H. Shen, L. Peng, L. Yue, Q. Yuan, and L. Zhang. Adaptive norm selection for regularized image restoration and super-resolution. *IEEE transactions on cybernetics*, 46(6):1388–1399, 2016.

[135] J. Sigl. Nonlinear residual minimization by iteratively reweighted least squares. *Computational Optimization and Applications*, 64(3):755–792, 2016.

[136] N. Srebro, J. Rennie, and T.S. Jaakkola. Maximum-margin matrix factorization. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1329–1336. MIT Press, 2005.

[137] D.L. Donoho S.S. Chen and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1998.

[138] J.L. Starck, F. Murtagh, and J. Fadili. *Sparse Image and Signal Processing: Wavelets and Related Geometric Multiscale Analysis.* Cambridge University Press, 2015.

[139] J.L. Starck, F. Murtagh, and J.M. Fadili. *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity.* Cambridge University Press, 2010.

[140] M. Stewart. Perturbation of the SVD in the presence of small singular values. *Linear Algebra and its Applications*, 419(1):53 – 77, 2006.

[141] T. Strutz. *Data fitting and uncertainty: A practical introduction to weighted least squares and beyond.* Vieweg and Teubner, 2010.

[142] J. Sun, Q. Qu, and J. Wright. When are nonconvex problems not scary? ArXiv preprint:1510.06096, 2015.

[143] Q. Sun. Recovery of sparsest signals via ℓ_q-minimization. *Applied and Computational Harmonic Analysis*, 32(3):329 – 341, 2012.

[144] R. Sun and Z. Q. Luo. Guaranteed matrix completion via nonconvex factorization. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 270–289, 2015.

[145] J. Tanner and K. Wei. Normalized Iterative Hard Thresholding for Matrix Completion. *SIAM Journal on Scientific Computing*, 35(5):S104–S125, 2013.

[146] J. Tanner and K. Wei. Low rank matrix completion by alternating steepest descent methods. *Applied and Computational Harmonic Analysis*, 40(2):417 – 429, 2016. [using `ASD` ('Alternating Steepest Descent') algorithm, code from `https://www.math.ucdavis.edu/~kewei/publications.html` resp. `https://www.math.ucdavis.edu/~kewei/code/mc20140528.tar`].

[147] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

[148] V. Torczon. On the convergence of pattern search algorithms. *SIAM Journal on Optimization*, 7(1):1–25, 1997.

[149] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank Solutions of Linear Matrix Equations via Procrustes Flow. *ArXiv preprint: 1507.03566*, 2015.

[150] M. Ulbrich. Semismooth newton methods for operator equations in function spaces. *SIAM Journal on Optimization*, 13(3):805–841, 2002.

[151] L. Van Der Maaten, E. Postma, and J. Van den Herik. Dimensionality reduction: a comparative. *J. Mach. Learn. Res.*, 2009.

[152] C. F. Van Loan. The ubiquitous Kronecker product. *Journal of Computational and Applied Mathematics*, 123(12):85 – 100, 2000.

[153] L. Vandenberghe. Convex optimization techniques in system identification. *IFAC Proceedings Volumes*, 45(16):71–76, 2012.

[154] B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. [using `Riemann_Opt` ('Riemannian Optimization') algorithm, code from `http://www.unige.ch/math/vandereycken/matrix\_completion.html`].

[155] S. V. Vaseghi. *Advanced Digital Signal Processing and Noise Reduction.* John Wiley and Sons, Ltd, 2001.

[156] C.R. Vogel and M.E. Oman. Fast, robust total variation-based reconstruction of noisy, blurred images. *IEEE Transactions on Image Processing*, 7(6):813–824, 1998.

[157] R. Wang, Y. Shen, P. Tino, A. Welchman, and Z. Kourtzi. Learning predictive statistics: strategies and brain mechanisms. *Journal of Neuroscience*, 2017.

[158] Y. Wang and Z. Wang, J.and Xu. On recovery of block-sparse signals via mixed ℓ_2/ℓ_q $(=< q \leq 1)$ norm minimization. *EURASIP Journal on Advances in Signal Processing*, 2013(1), 2013.

[159] R.W.M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61(3):439–447, 1974.

[160] P. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[161] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung. Guarantees of Riemannian Optimization for Low Rank Matrix Completion. *ArXiv preprint:1603.06610*, 2016.

[162] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung. Guarantees of Riemannian Optimization for Low Rank Matrix Recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

[163] Z. Wen, W. Yin, and Y. Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, 2012.

[164] D. Wipf and S. Nagarajan. Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):317–329, 2010.

[165] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

[166] J. Wright. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in Neural Information Processing Systems 22*, 2009.

[167] F. Xu and J. B. Tenenbaum. Word learning as bayesian inference. *Psychological review*, 114(2):245, 2007.

[168] X. Yuan and J. Yang. Sparse and low-rank matrix decomposition via alternating direction methods. Technical report, 2009.

[169] Q. Zheng and J. Lafferty. Convergence Analysis for Rectangular Matrix Completion Using Burer-Monteiro Factorization and Gradient Descent. *ArXiv preprint:1605.07051*, 2016.

[170] T. Zhou and D. Tao. Godec: Randomized low-rank and sparse matrix decomposition in noisy case. In *International Conference on Machine Learning*, 2011.