Technische Universität München

TUM School of Computation, Information and Technology

Scalability in Ill-posed Machine Learning Problems: Bridging Least Squares Methods with (Non-)Convex Algorithms

Claudio Mayrink Verdun

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:

Prof. Donna Ankerst, Ph. D.

Prüfer*innen der Dissertation:

1. Prof. Dr. Felix Krahmer
2. Prof. Dr. Akram Aldroubi
3. Jun.-Prof. Dr. Max Pfeffer

Die Dissertation wurde am 27.06.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 27.10.2023 angenommen.

Abstract

This thesis explores the relationship between the method of least squares and modern (non-)convex optimization techniques for addressing ill-conditioned and ill-posed inverse problems prevalent in machine learning and data science. Our focus lies in leveraging least squares to develop simple, computationally efficient and statistically robust algorithms with provable guarantees for various tasks, including matrix completion, sparse recovery, and noise-blind regression. We provide contributions in four key areas. First, we address the convergence rate of the iteratively reweighted least squares (IRLS) algorithm for sparse recovery, solving an open problem and demonstrating its global linear convergence rate. Second, we introduce an algorithm capable of efficiently completing highly ill-conditioned low-rank matrices using the information-theoretically optimal number of samples. Additionally, we prove that the algorithm achieves a local quadratic convergence rate. Our extensive numerical experiments showcase its superiority over existing methods for statistically hard problems. Third, we extend the IRLS theory to noise-blind regression problems, where accurately estimating the noise level is challenging, and propose a practical algorithm to tackle this scenario. Additionally, we explore the application of ideas from overparametrized neural networks to solve constrained least squares problems in a scalable manner, harnessing the inherent bias of gradient descent.

Zusammenfassung

Diese Arbeit untersucht die Beziehung zwischen der Methode der kleinsten Quadrate und modernen (nicht-)konvexen optimierungstechniken zur Bewältigung von schlecht konditionierten und schlecht gestellten inversen Problemen, die im maschinellen Lernen und der Datenwissenschaft weit verbreitet sind. Der Fokus liegt darauf, die Methode der kleinsten Quadrate zu nutzen, um einfache, recheneffiziente und statistisch robuste Algorithmen mit nachweisbaren Garantien für verschiedene Aufgaben zu entwickeln, einschliesslich Matrixvervollständigung, dünnbesetzter Wiederherstellung und rauschblinder Regression. In diesem Sinne werden die folgenden vier Bereiche betrachtet. Als erstes wird das offene Problem der Konvergenzgeschwindigkeit des iterativ gewichteten kleinste Quadrate (IRLS) Algorithmus für die dünnbesetzte Wiederherstellung behandelt. Im Zuge dessen wird zusätzlich die globale lineare Konvergenzrate des Algorithmus' demonstriert. Als nächstes wird ein weiterer Algorithmus mit lokal quadratischer Konvergenzrate vorgestellt, der schlecht konditionierte Niedrigrangmatrizen effizient mit der informationstheoretisch optimalen Anzahl von Stichproben vervollständigen kann. Umfangreiche numerische Experimente illustrieren seine Performancevorteil gegenüber bestehenden Methoden für statistisch schwierige Probleme. Im dritten Teil wird die IRLS-Theorie auf rauschblinde Regressionsprobleme, bei denen eine genaue Schätzung des Rauschniveaus herausfordernd ist, erweitert. Zuletzt werden bekannte Ideen aus dem Bereich der überparametrisierten neuronalen Netzwerken erforscht, um Problem der beschränkten kleinsten Quadrate auf skalierbare Weise zu lösen, indem der inhärenten Bias des Gradientenverfahrens ausgenutzt wird.

Acknowledgment

I want to express my deepest gratitude to my supervisor, Felix Krahmer, who is not only a great mind and a great supervisor but also an amazing human being. On the professional side, I will never forget all your dedication, feedback, opportunities that you offered me, collaboration and insights, the grants that we wrote together, the projects that we had done, the students that we supervised, the late (and funny) calls and everything that you did to support me and my career. And also for doing all of that with joy and fun. The jokes and the daily smile on your face made me understand that this academic journey should be enjoyed daily. On a personal level, thank you for just being you. You've always been so attentive and a great listener (not just to me but to everyone in the group), and you've been there for me during tough times when I felt like giving up. You showed me it was worth staying and that the outcome would be positive. Working with you has been an absolute privilege, and I'm truly happy to have had you as my supervisor.

A very special thanks to my mentor, Prof. Marion Menzel. You have played a major role in getting me to where I am today. Thanks for all the amazing conversations about life, the universe, and everything else. Your career advice has been invaluable, and you've helped me navigate the ups and downs of my Ph.D. journey. Thanks for showing me how to handle those quirky individuals we encounter along the way. And let's not forget about all the fun and laughter we've shared and how much I've learned about MRI from you. You've celebrated every little success of mine, and I'm truly lucky to have the opportunity to work with you.

A huge thanks to Prof. Akram Aldroubi and Prof. Max Pfeffer for agreeing to be part of my thesis committee without hesitation. And a special thanks to Prof. Akram Aldroubi for the incredible support from our first interaction. You've brought amazing opportunities into my life, like the Focus Program on Data Science, Approximation Theory, and Harmonic Analysis at the Fields Institute in 2022. That experience definitely opened doors for me professionally. I'm also extremely grateful to Prof. Donna Ankerst for chairing my thesis committee and for her enthusiasm and support in discussing my future career steps.

I want to express my gratitude to Prof. Frank Filbir for being an incredible friend during my Ph.D. adventure. Our engaging conversations about math, history, science, and politics were always enjoyable, and I will always remember those moments. Your support meant a lot to me, especially during these challenging times. By the way, don't forget, you promised to visit me! Boston eagerly awaits your arrival, my dear friend! I want to thank Prof. Holger Rauhut for the book that deeply influenced me, for telling me that I should come to Germany, and for all the fascinating mathematical discussions. Thanks for believing in the work that Frederik and I have been developing. I also want to thank Prof. Massimo Fornasier for a few conversations about life and research, the enthusiasm when we wrote a grant proposal together, and the support with my career's next steps!

I want to give a huge shoutout to my collaborators and friends who have taught me countless things and with whom I have an absolute blast discussing math, our collaborative research, and everything in between. We have many more exciting joint projects ahead! A special thanks to Frederik Hoppe, Johannes Maly, Hung-Hsu Chou, Heudson Mirandola, Peter Jung, Stefan Kolek, Niclas Popp, Manos Theodosis, and Alexander Lin for all the delightful conversations. I also want to express my sincere gratitude to Christian Kümmerle, Tim Fuchs, Carlos Amendola, Alihan Kaplan, Marco Mondelli, and Dominik Stöger for their insightful scientific discussions and for sharing their extensive math knowledge with me.

I want to thank all my amazing colleagues and friends from M15. Not only do you guys produce some seriously interesting research, but our research group is truly one of a kind. You've turned this small, cloistered aisle into a vibrant and fun academic environment. Thank you for the incredible friendships, the shared lunches, the engaging discussions, and all the memorable moments we've had together. Thank you Tim Fuchs, Oleh Melnyk, Carlos Amendola, Hannah Laus, Alessandro Lupoli, Stefan Bamberger, Konstantin Riedl, Pascal Heid, Marco Morandotti, Marco Rauscher, Alessandro Scagliotti, Stefano Almi, Olga Graf, Nada Sissouno, Anna Veselovska, Patricia Römer, Cristina Cipriani, Sara Krause-Solberg, Johannes Maly, Ricardo Acevedo Cabra, Sandro Belz, Felipe Pagginelli Patricio, and Dominik Stöger. I'd also like to extend my gratitude to my friends Haris Ceribasic, Alihan Kaplan, Ezra Tampubolon, and Sajad Saeedinaeeni for all the coffee-fueled scientific discussions and for making my second academic home, the LTI at the Electrical and Computer Engineering Department, even more fun and exciting!

Thanks to Martina Beck for your unwavering support, encouraging words, and constant availability. You've played a crucial role in making all this possible and my personal growth. A special thanks to Nada Sissouno for all the incredible and cheerful support. Your presence and friendship throughout my journey at the university have meant the world to me. I also want to express my gratitude to Isabella Wiegand for your support. Your impact on my PhD journey cannot be overlooked. And I can't forget to thank Valeria Soave for the entertaining and great lessons and, more importantly, for the wonderful friendship we've developed over the years, filled with shared interests. Our morning meetings have been a source of joy amid pure stress. I want to thank Bruno Braga, Guilherme Sales, Pedro Maia, Luis Felipe Velloso, Roberto Velho, Adriano Cortes, Filipe Goulart, and Rafael Ribeiro for everything you've done for me. You are the best friends anyone could ever ask for.

While at TUM, I got the chance to teach and mentor some brilliant Master's students. Interacting with them was an incredible experience that truly helped me grow as a teacher and a researcher. I want to express my gratitude to Haris Ceribasic, Akshay Pimpalkar, Niclas Popp, Julia Wenke, Wien Boukhzar, Tamas Hakkel, Thomas Weinberger, Maximilian Fürst, Christina Strunz, Raul Moreno, Sebastian Kaiser, and Yiwei Zhu. You all made my time as a mentor truly memorable.

I want to express my sincere gratitude to the staff members at TUM who have provided invaluable assistance in various bureaucratic and personal matters. Your professionalism and friendliness have been instrumental in resolving complex issues and navigating the Kafkaesque administrative processes. Special thanks to Silvia Toth-Pinter, Annemarie Meinel, Isabella Wiegand, Kerstin Weinberger, Claudia Koch, and Elizabeth Söder for their continuous support and dedication to the university. Your contributions are highly appreciated and have greatly contributed to everyone's academic journey in math and the electrical engineering department.

I am very grateful to Amir Beck, Ivan Markovsky, and Demba Ba for generously inviting me to visit their research groups. Their warm hospitality and kind support made my experience truly memorable. I am also thankful for our numerous enlightening scientific discussions during my visit. Through your papers and unique approaches to problem-solving, I have gained valuable insights and expanded my understanding of mathematics and electrical engineering.

I want to thank all the amazing friends that I met during my time in Munich. This thesis would not have been possible without all the nice moments and support you provided. Thank you, Carlos, Caio, Tom, Amanda, David, Pedro, Zoé, Haris, Felipe, Stefan, Siena, Alihan, Mariella, Azada, Janosch, Lloyd, Franco, Brenda, Robson, Roxana, Thiago, Maria Eugênia, Daniel, André, Camila, William, Barbara, Lara, Benedetta, Gaia, Oleh, Tim, Hung-Hsu (aka Edward), Frederik, Hector, Cornelia, Gabriel, Marianna, Menna, Luca, Sebastian, Carolin, Akshay, Ezra, Sajad and Milica.
I want to thank my former professors/masters Luiz Wagner Biscainho, Fabio Ramos, Bernardo da Costa, Cesar Niche, Eduardo Silva, Amit Bhaya, Felipe Acker, and Roberto Imbuzeiro. You've been an incredible source of guidance, knowledge, and motivation these past few years. Our conversations have always been enlightening and inspiring. I truly appreciate the support you provided during my studies and when I moved to Germany. I am truly grateful for your friendship and everything you've done!

I want to express my profound gratitude to a group of special individuals who have been there for me throughout this incredible journey, cheering me on and providing unwavering support from various corners of the world. I am truly thankful to Ivani Ivanova, Heudson Mirandola, Bruno Braga, Filipe Goulart, Luis Felipe Velloso, Gabriela Lewenfus, Rafael Ribeiro, Guilherme Sales, João Cordeiro, Lloyd Hughes, Hugo Carvalho, Carlos Lechner, Tiago Domingues, Daniel Soares, Douglas Picciani, Carolina Casals, Leonardo Ribeiro, Arnesh Sujanani. I also want to thank my dear friends from the *Khalifado aplicado*. Your constant support, shared laughter, and amusing conspiracy theories brought so much joy and light to my days. While they may have occasionally diverted my attention from my thesis, they made this journey more enjoyable. I would have had a better thesis without you! Thank you Pedro Maia, Rogerio Lourenço, Roberto Velho, Yuri Saporito, and Rodrigo Targino for your unwavering friendship and camaraderie.

I feel incredibly lucky to have found a second family in Munich. These amazing people have shown me so much love, care, and support that words cannot fully express my gratitude. A big thanks to Isabel von Blume, Eliana Lopes, Ronald von Blume, Andre Bechtel, and Carlos Améndola. Their love and constant support have kept me motivated throughout my journey. It's truly incredible to consider them as part of my family now.

Finally, I would like to express my heartfelt gratitude to my family for their support and encouragement. First, my family from this side of the ocean supported me not only with amazing food but with love, affection, and joy. I am sorry for having spent less time with you than I would have liked and for all the late nights that I was always awake working, and you were waiting for me. Thank you Franco, Maria, Jle, Narciso (in memoriam), and Verena, for being the best on this side of the ocean. From that side of the ocean, I want to thank my family, who were always there for everything I needed. For their unconditional love, support, understanding and sacrifice. I want to thank Tia Marilene and Tia Marisa for being a source of comfort and strength and always being there for me. An immense thanks to my grandma Maria da Gloria (in memoriam) for all the love and all the life lessons. In particular, for teaching me that life has, on average, 3800 weeks and we should enjoy every single one of them. Until her last day, her constant questions about when I would fly back home made me realize how we should truly like what we do and the purpose of all of that.
Last but not least, I want to dedicate a heartfelt acknowledgment to my parents, Marilia and Luiz Claudio, who have been my pillars of strength and unwavering source of love and support. You have been my guiding light, source of inspiration, and unwavering cheerleaders throughout every step of my journey, including the times of triumph and the moments of despair. Thanks, Mom, for the roots, your presence, and daily support, and thanks, Dad, for the courage to move to another country and pursue my dreams and for the enthusiasm for science.
I feel incredibly lucky for the serendipitous encounters in life that led me to meet some people. Above all, I am eternally grateful for the immense love, support, and patience that my life and adventure partner, Selene, showed during this time. Thank you for being the amazing person you are, your understanding and affection, and the countless memorable moments we've shared since our paths crossed. This is only the beginning, and I am excited about what lies ahead and what we will construct together.

Thanks to each and every one of you for your valuable contributions to this thesis and my academic life and, more importantly, for being a part of my life journey!

This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the Collaborative Research Project SparseMRI3D+, by the German Research Foundation (DFG) through the Grant PO 1347/3-1 and by the German Science Foundation (DFG) in the context of the project KR 4512/1-1. I also thank the Munich Center for Machine Learning for their academic support.

Del rigor en la ciencia

En aquel Imperio, el Arte de la Cartografía logró tal Perfección que el Mapa de una sola Provincia ocupaba toda una Ciudad, y el Mapa del Imperio, toda una Provincia. Con el tiempo, estos Mapas Desmesurados no satisficieron y los Colegios de Cartógrafos levantaron un Mapa del Imperio, que tenía el Tamaño del Imperio y coincidía puntualmente con él. Menos Adictas al Estudio de la Cartografía, las Generaciones Siguientes entendieron que ese dilatado Mapa era Inútil y no sin Impiedad lo entregaron a las Inclemencias del Sol y los Inviernos. En los Desiertos del Oeste perduran despedazadas Ruinas del Mapa, habitadas por Animales y por Mendigos; en todo el País no hay otra reliquia de las Disciplinas Geográficas.[1]

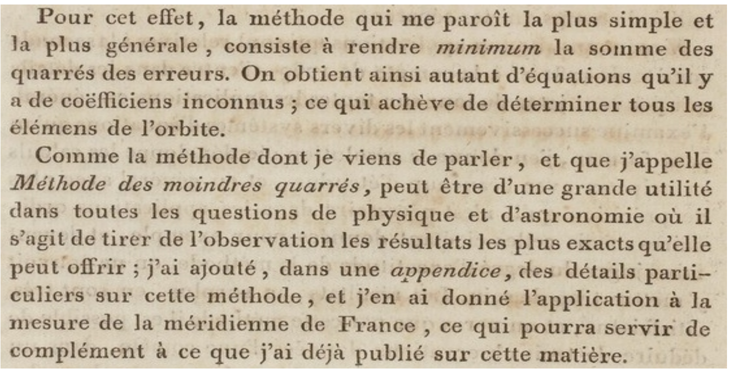
Suárez Miranda: Viajes de varones prudentes
Libro Cuarto, cap. XLV, Lérida, 1658. In [Bor13].

[1] In a free translation: *On the Exactitude in Science - In that Empire, the Art of Cartography attained such Perfection that the Map of a single Province occupied an entire City, and the Map of the Empire, an entire Province. Over time, these Excessive Maps proved unsatisfactory, and the Colleges of Cartographers erected a Map of the Empire that had the Size of the Empire and coincided point for point with it. Less inclined to the Study of Cartography, the subsequent Generations understood that this vast Map was Useless and not without Impiety they delivered it to the Inclemencies of the Sun and the Winters. In the Deserts of the West, still today, there are scattered Ruins of the Map, inhabited by Animals and Beggars; in all the Country there is no other relic of the Disciplines of Geography.*

Contents

Chapter 1

Introduction



Pour cet effet, la méthode qui me paroît la plus simple et la plus générale, consiste à rendre *minimum* la somme des quarrés des erreurs. On obtient ainsi autant d'équations qu'il y a de coëfficiens inconnus ; ce qui achève de déterminer tous les élémens de l'orbite.

Comme la méthode dont je viens de parler, et que j'appelle *Méthode des moindres quarrés*, peut être d'une grande utilité dans toutes les questions de physique et d'astronomie où il s'agit de tirer de l'observation les résultats les plus exacts qu'elle peut offrir ; j'ai ajouté, dans une *appendice*, des détails particuliers sur cette méthode, et j'en ai donné l'application à la mesure de la méridienne de France, ce qui pourra servir de complément à ce que j'ai déjà publié sur cette matière.

Adrien Marie Legendre, *Nouvelles methodes pour la determination des orbites des cometes.* Page VIII. Paris, 1805 [Leg06].[1].

Since the beginning of the new millennium, we have witnessed a significant increase in data collection, processing and analysis, accompanied by the emergence of data science as a scientific discipline. To illustrate this massive growth, consider the staggering statistics for a single minute in the year 2022: Google users conducted 5.9 million searches, YouTube users uploaded 500 hours of video, Twitter users shared approximately 347,000 tweets, Instagram users shared 66,000 photos, and people collectively spent around 104,000 hours in Zoom meetings. To put this into perspective, a decade ago, there were roughly 2 million Google queries, 48 hours of video uploaded to YouTube, and 100,000 tweets and 3,600 photos shared on Instagram per minute [dat23]. Furthermore, the acquisition and processing of vast amounts of data have profoundly impacted numerous scientific disciplines, including genomics, image and audio processing, economics, neuroscience, environmental sciences, robotics, and computer vision, among many oth-

[1] In a free translation *"for this purpose, the method which seems to me the simplest and the most general, consists in minimizing the sum of the squares of the errors. We thus obtain as many equations as there are unknown coefficients; which completes the determination of all the orbit elements. Like the method of which I have just spoken, and which I call the* ***method of least squares****, can be of great use in all questions of physics and astronomy where it is a question of drawing from observation the more exact than it can offer; I have added, in an appendix, specific details on this method, and I have given its application to the measurement of the meridian of France, which could serve as a complement to what I have already published on this matter."*

ers. The expectation is that this impact will continue to grow in the coming years. Indeed, as David Donoho said in the very influential article "50 years of Data Science" [Don17], *"Because all of science itself will soon become data that can be mined, the imminent revolution in Data Science is not about mere 'scaling up', but instead the emergence of scientific studies of data analysis science-wide.* In various scientific disciplines, there has been a notable shift in perspective. Previously, the focus was primarily on acquiring all possible data, aligning with the notion of *"measure what can be measured"*. This quote is often erroneously attributed to Galileo Galilei (cf. [Kle09]). However, a new philosophy has emerged, as articulated by Thomas Strohmer in the survey [Str12], advocating for the concept of *"measure what should be measured"*.

Ever since Johannes Kepler's groundbreaking analysis of astronomical data and his profound insights into planetary motions, which established him as the first renowned data scientist in history [Ost20], science has experienced a profound transformation. This shift has led to the emergence of a fourth paradigm centered around data-driven approaches that, complementing the existing paradigms of empirical evidence, scientific theory, and computational science, is driving a deep transformation in how science has been developed. Moreover, this transition and progress have been compared to the revolutionary impact of the invention of the printing press, highlighting their profound influence on recent scientific advancement [HTT$^+$09]. Indeed, in the words of Jim Gray, *"The world of science has changed, and there is no question about this. The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration"* [HTT$^+$09, Page xix].

However, in this era of data deluge, one premise stands out as particularly significant: the realization that the information encompassed within datasets is considerably smaller than the sheer volume of data. This observation underscores the notion that real-world datasets possess inherent low complexity despite their high dimensionality. This *principle of parsimony* serves as a foundational concept that permeates scientific domains employing data-driven methodologies. In essence, it suggests that a mere small subset of features from the dataset suffices for conducting inference, constructing models, addressing scientific inquiries, and developing practical applications. This principle has many facets and can be translated into several concepts, e.g., sparsity, low-rankness, and positivity. These concepts are also connected to the advent of High-Dimensional Statistics [Wai19] and Compressive Sensing [FR13]. By no means does this thesis intend to provide a comprehensive overview of these vast and fastly developing fields. There are several wonderful books and surveys about them [Wai19, FR13, Vid19, WM22]. We intend, however, to discuss simple but powerful algorithms that can be developed and analyzed to work with very large (and potentially ill-conditioned) datasets.

By harnessing the power of this principle of parsimony, it is possible to retrieve and analyze high-dimensional data from a limited number of measurements; this idea lies at the heart of modern machine learning. As the storage, analysis, inference, and downstream tasks associated with large-scale high-dimensional data become increasingly important, the development of fast and scalable algorithms to process such data emerges as a key objective. This thesis aims to revisit an old yet important idea that originated in a captivating chapter of scientific history during the early 19th century. It was a period marked by the development of statistics and probability theory and by advancements in mechanics, leading to remarkable achievements in solving diverse problems in fields such as cartography and astronomy.

Specifically, this thesis focuses on the principle of least squares —- a technique that traces its roots back to the pioneering work of Lagrange and Gauss and holds a central position in mathematics and statistics. The importance of this idea can be assessed from various perspectives. To illustrate its significance, as of the time of writing, a search on Google Scholar yields 4,470,000 papers with "least squares" in their titles, of which 63,400 have been published since the beginning of 2022. The essence of this thesis lies in harnessing the power of least squares and developing scalable algorithms inspired by it. The objective is to provide solutions with provable convergence guarantees, allowing for the establishment of optimal results under minimal assumptions for some machine learning tasks.

1.1 The method of least squares

The method of least squares, first introduced in 1805 by Adrien-Marie Legendre in an extensive paper of eighty pages accompanied by a fifty-five-page appendix written in French [Leg06][2], who coined its name, and, independently, by Carl Friedrich Gauss in the early 19th century [Gau77][3], marks a significant milestone in the history of numerical analysis and statistical estimation[4]. See also [AW02] for a detailed explanation of Gauss' least-squares calculations.

In particular, as discussed on [Sti86, Page 16], there were three important problems to be solved that are associated with the development of the method of least squares and that were addressed by both of them:

- to determine the shape of the Earth;
- to determine and represent the motions of the moon mathematically;
- and to account for an apparently secular (that is, nonperiodic) inequality that had been observed in the motions of the planets Jupiter and Saturn.

[2] As noted by Stigler in [Sti86, Page 13] *"For stark clarity of exposition the presentation is unsurpassed; it must be counted as one of the clearest and most elegant introductions of a new statistical method in the history of statistics"*.

[3] As indicated by Stigler in [Sti86, Page 145], there was a credit dispute between Legendre and Gauss. In fact, Gauss referred to the method of least squares as *principium nostrum* ("our principle") and claimed that he had been using the method since I795.

[4] The first English translate of Legendre's work on least squares was published in 1822 and appeared in [Har22]. As for Gauss' work, the first English translation is from 1857 [Gau57].

To address these challenges, they stood on the shoulders of other giants such as Mayer, Simpson, Cotes, Lambert, De Moivre, Euler, Boscovich, Laplace, and numerous other prominent astronomers, mathematicians, and cartographers of that era. They built upon the foundational work laid by these scientists, incorporating their ideas into their own research. Generally speaking, Legendre and Gauss independently developed this powerful technique to address the problem of fitting a mathematical model to a set of observed data points. Legendre's work in 1805 focused on determining the best-fit line for a set of data points by minimizing the sum of the squares of the vertical deviations between the data and the line. Shortly afterward, Gauss published his own treatise in 1809. As Kahaner, Moler, and Nash point out in the book [KMN89], "*On January 1, 1801, Giuseppe Piazzi discovered the asteroid Ceres. Ceres was only visible for forty days before being lost to view behind the sun. Using three observations, extensive analysis, and the method of least squares, Gauss was able to determine the orbit with such accuracy that Ceres was easily found when it reappeared in late 1801.*". Indeed, on December 7, 1801, the astronomers Franz von Zach and Heinrich Olbers were able to observe it after using the predictions made by Gauss with the method of least squares [Gol12]. Unfortunately, a comprehensive history of the least squares method and all of its scientific consequences is beyond the scope of this thesis. Nevertheless, there are excellent sources that can be consulted and that highly inspired this chapter, e.g., [Gla72, Mer77b, Mer77a, Pla49, Eis61, She73, Har74a, Har74b, Har75a, Har75b, Har75c, Sti81, Sti86, Wat90, She93, Ald98, Nie01, GW11]. In particular, the publication [Mer77a] provides a comprehensive analysis of the notable developments in the method of least squares spanning the years 1805 to 1864. For readers interested in delving further into this captivating chapter of scientific history, we recommend consulting the books [Sti86, Far99, Gol12, Tod14] and, especially, the excellent book [Gor16]. These sources offer valuable insights and in-depth exploration of the subject matter.

Throughout the 19th and 20th centuries, further advancements were made in numerical analysis, expanding the applications and capabilities of the method of least squares. Notably, the development of matrix algebra and linear regression techniques was crucial in extending the method's reach. In particular, the advent of digital computers in the mid-20th century further accelerated the use of least squares in various fields that relied on least squares to handle large datasets and estimate parameters with high precision.

The field of statistics also embraced the method of least squares as a fundamental tool for regression analysis, hypothesis testing, and model selection. Interestingly, the statistical notion of linear regression and its interpretation date from 1877-85 and were developed much later than the seminar seminal works by Legendre and Gauss. The name *regression* itself, which means *regression towards the mean*, was suggested by Francis Galton and comes from the study of human intelligence and talent [Sti97, Gal89]. Later, the influential works of researchers such as Ronald A. Fisher, Jerzy Neyman, Udny Yule, Francis Edgeworth, and Karl Pearson at the beginning of the 20th century provided a statistical framework for the method and established its theoretical underpinnings. See the books [Leh11, Gor16] for more information on their contribution to statistics. Moreover, the works on numerical mathematics by Gene Golub, William Kahan, Alston Householder, Pete G. W. Stewart, James Wilkinson, Åke Björck, Charles Lawson, Richard Hanson, among others, in the late sixties, were fundamental to establishing the modern computational theory of least squares, see [Hig02, Section 20.11] and many historical comments through the book [LH95].

More precisely, the least squares method is employed to fit a linear mathematical model to a set of given observations. To mitigate the impact of errors in the observed data, it is desirable to have a larger number of measurements compared to the number of unknown parameters in the model. Formally speaking, given a dataset of observations denoted as $b \in \mathbb{R}^m$, obtained through a (possibly noisy) measurement process described by a linear operator $A \in \mathbb{R}^{m\times N}$, the goal is to find a solution $x \in \mathbb{R}^N$ to the optimization problem

$$\min_{x\in\mathbb{R}^N} \|Ax - b\|_2, \tag{1.1}$$

where $\|.\|_2$ denotes the Euclidean norm. The solution $x \in \mathbb{R}^N$ is referred to as the *linear least squares solution* to the linear system $Ax = b$. In the case that we have a larger number of measurements compared to the number of unknown parameters, i.e., $m > N$, which leads to an overdetermined linear system, the solution above is the one where the data b is *best* approximated by Ax. It is called the least squares solution since it minimizes the residual vector $r = Ax - b$. This solution may not be unique when $\text{rank}(A) < N$. However, there is a unique solution that minimizes the 2-norm $\|x\|_2$. An important property of such a solution is described by the following theorem.

Theorem 1.1.1. *[Bjö96, Theorem 1.1.2] A vector $\hat{x} \in \mathbb{R}^N$ is a solution of the problem* (1.1) *if and only if the orthogonal condition $A^T(b - A\hat{x}) = 0$ holds.*

Proof. Assume that $\hat{x}$ satisfies $A^T\hat{r} = 0$, where $\hat{r} = b - A\hat{x}$. Then for any $x \in \mathbb{R}^N$, we have $r = b - Ax = \hat{r} + A(\hat{x} - x)$. By calculating the square norm $\|r\|_2^2$ and using the Pythagorean Theorem, we obtain

$$\|r\|_2^2 = \langle r, r\rangle = \langle \hat{r} + A(\hat{x} - x), \hat{r} + A(\hat{x} - x)\rangle = \langle \hat{r}, \hat{r}\rangle + \|A(\hat{x} - x)\|_2^2.$$

The minimum in the expression above is attained precisely when $x = \hat{x}$. Conversely, suppose that $z =: A^T\hat{r} \neq 0$ and consider $x = \hat{x} + \varepsilon z$, for a certain small $\varepsilon > 0$. Then, $r = \hat{r} - \varepsilon Az$, and

$$\langle r, r\rangle = \langle \hat{r}, \hat{r}\rangle - 2\varepsilon\langle \hat{r}, Az\rangle + \varepsilon^2\langle Az, Az\rangle = \langle \hat{r}, \hat{r}\rangle - 2\varepsilon\langle z, z\rangle + \varepsilon^2\langle Az, Az\rangle < \langle \hat{r}, \hat{r}\rangle,$$

for a sufficiently small ε. This implies that $\hat{x}$ is not the least squares solution. □

The theorem above shows that the residual vector $r = Ax - b$ lies in the kernel of A^T. In particular, by denoting by $R(A)$ the range space of the matrix $A \in \mathbb{R}^{m\times N}$, the geometric interpretation is that any least squares solution x can decompose the data $b \in \mathbb{R}^m$ in a unique

way into two orthogonal components, i.e., $b = Ax + r$, where $r \in \ker(A^T)$ and $Ax \in R(A)$. Consequently, Ax is the orthogonal projection of b onto the range (column space) of A.
From the discussion above, it is possible to see that the least squares solution can be found by solving the so-called *normal equations*, $A^TAx = A^Tb$, a term coined by Gauss. In particular, for $A \in \mathbb{R}^{m\times N}$ with $m \geq N$ of full rank N, the matrix A^TA is positive definite, and the unique least squares solution is given by

$$x_{LS} = (A^TA)^{-1}A^Tb.$$

In fact, the least squares solution can be characterized via the singular value decomposition that exists for every matrix $A \in \mathbb{C}^{m\times N}$.

Proposition 1.1.2. *[FR13, Proposition A.13] For $A \in \mathbb{C}^{m\times N}$, there exist unitary matrices $U \in \mathbb{C}^{m\times m}$, $V \in \mathbb{C}^{N\times N}$, and uniquely defined nonnegative numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{m,N\}}$, called singular values of A, such that*

$$A = U\Sigma V^* \qquad \Sigma = \operatorname{diag}[\sigma_1, \dots, \sigma_{\min\{m,N\}}]$$

In particular, if the matrix $A \in \mathbb{C}^{m\times N}$ has rank r, we can use the following notation:

$$A = U \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} V^*, \qquad U = \begin{pmatrix} U_r & U_{m-r} \end{pmatrix}, \qquad V = \begin{pmatrix} V_r & V_{N-r} \end{pmatrix},$$

We can finally characterize any solution to the least squares problem with the following proposition, which can be found in several books about least squares, e.g., [FR13, Proposition A.20].

Proposition 1.1.3. *Let $A \in \mathbb{R}^{m\times N}$ and $b \in \mathbb{R}^m$. Then, for all $z \in \mathbb{R}^{N-r}$, the vector*

$$\hat{x} = V_r\Sigma_r^{-1}U_r^*b + V_{N-r}z,$$

is a solution of the least squares problem (1.1)*. Moreover, these are all the least squares solutions.*

Proof. Let $x \in \mathbb{R}^N$. We start by partitioning it into two pieces,

$$V^*x = \begin{pmatrix} V_r^*x \\ V_{m-r}^*x \end{pmatrix} =: \begin{pmatrix} w \\ z \end{pmatrix}.$$

Now, we substitute the SVD of A into the residual $Ax - b$, which yields

$$Ax - b = U \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} V^*x - UU^*b = U \begin{pmatrix} \Sigma_r w - U_r^*b \\ -U_{m-r}^*b \end{pmatrix}.$$

Thus,

$$\|Ax - b\|_2^2 = \|U^*(Ax - b)\|_2^2 = \|\Sigma_r w - U_r^*b\|_2^2 + \|U_{m-r}^*b\|_2^2,$$

since the ℓ_2-norm is invariant by unitary matrices. Note that the second summand is independent of w and z. Therefore, the norm is minimized if and only if the first summand is zero, i.e., when $\hat{w} = \Sigma_r^{-1} U_r^* b$. Hence, the solutions of (1.1) are given by

$$\hat{x} = V \begin{pmatrix} \hat{w} \\ z \end{pmatrix} = V_r \hat{w} + V_{N-r} z = V_r \Sigma_r^{-1} U_r^* b + V_{N-r} z.$$

For any $z \in \mathbb{R}^{N-r}$, $V_{N-r} z \in \ker(A)$, since A has rank r. Therefore, z does not influence the function $\|Ax - b\|_2$ and $z \in \mathbb{R}^{N-r}$ can be chosen arbitrarily. □

We have that if A is not full rank, then $\ker(A)$ is non-trivial and, since the columns of V_{N-r} are orthogonal, as a consequence of the theorem above, we can deduce that the least squares problem (1.1) has infinitely many solutions. It also shows that any solution contains one term given the *pseudo-inverse*, also known as the Moore–Penrose inverse, of the matrix A, $A^\dagger = V_r \Sigma_r^{-1} U_r^*$.

In the case that $N \geq m$ and the matrix $A \in \mathbb{R}^{m \times N}$ is of full rank m, then the matrix AA^T is invertible and the minimum norm solution to the underdetermined system $Ax = b$ has a solution given by the *normal equation of the second kind* [FR13, Corollary A.22],

$$x_{LS} = A^T (AA^T)^{-1} b.$$

There are a few ways to numerically find the least squares solution, i.e., to solve (1.1). They are usually divided into *iterative methods* and *direct methods.* As examples of direct methods, we can cite methods based on the *Cholesky factorization* ($mN^2 + \frac{1}{3}N^3$ flops), those based on the *QR factorization* ($2mN^2 - \frac{2}{3}N^3$ flops) and those that utilize the *Singular Value Decomposition* ($2mN^2 + 11N^3$ flops), see, e.g., [TB22, Chapter 11]. Each method has its own advantages and disadvantages, which depend on factors such as problem size, conditioning, and how close to being singular the problem is. However, when dealing with large matrices or structured matrices, it is beneficial to utilize iterative methods that exploit the matrix-vector multiplication Ax. In such cases, *Krylov methods*[5], which include the conjugate gradient method [HS+52a] and the minimum residual method (MINRES) [PS75] as specific instances, have been developed [Sog23]. In particular, we can cite the very relevant and widely used LSQR [PS82], LSMR [FS11], and LSLQ [EOS19] methods for the iterative solution of linear systems.

The least squares method is a rich and multifaceted topic, extensively covered in the classic book by Bjöerck [Bjö96], which remains the most comprehensive reference on the subject. Additional sources worth consulting include [TB22, LH95, Far18]. In this thesis, we will not delve deeply into the statistical or numerical properties of least squares. Instead, our focus will be on demonstrating how this technique can be efficiently employed to solve high-dimensional inverse problems that arise in machine learning and data science in a simple yet effective manner.

[5] Krylov methods are considered one of the top 10 Algorithms of the 20th century, see [Cip00, DS00, VDV00].

1.2 High-dimensional inverse problems

For nearly two centuries, the focus of the discussion surrounding the solution of linear systems revolved around cases where there were more measurements than unknown parameters. Specifically, as mentioned in the previous section, a central point was how to best fit the model to a large set of measurements. However, this scenario changed with the emergence of scientific problems where data acquisition is costly and only a limited number of high-dimensional signal measurements are available – “measure what should be measured” [Str12]. In such cases, the linear system has infinitely many solutions, making it impossible to fully recover the underlying signal $x \in \mathbb{R}^N$ from the data $b \in \mathbb{R}^m$ unless some structural assumption about the solution is made. In many applications, a fundamental premise is that the object to be retrieved has a parsimonious representation, i.e., it can be represented using minimal information. These high-dimensional inverse problems have become ubiquitous in fields like machine learning, signal processing, and data science. Under these assumptions, numerous works over the past three decades have analyzed the conditions under which it becomes possible to unambiguously identify the true underlying signal [FR13, HTW15, WM22]. In such problems, one designs a regularizer that controls the variable selection and allows for the identification of a solution having a certain structure. One notable instance is when the data is assumed to be sparse, i.e., when only a small subset of the variable has nonzero coefficients or when many extracted features among the predictors are irrelevant. This concept corresponds to the idea that for a given phenomenon, usually there are only a few relevant causes, and it forms the basis of the fields of compressive sensing and sparse recovery.

In 2006, groundbreaking research by Donoho, Candès, Tao, and Romberg established a significant link between structure and randomness. They demonstrated how the combination of concepts from convex optimization, stochastic processes, harmonic analysis, and approximation theory could successfully address the challenge of finding sparse solutions of underdetermined systems of equations [Don06, CT05, CT06, CRT06b, CRT06a]. These influential works, currently cited over 80,900 times according to Google Scholar, also highlighted the immense potential of these ideas in the fields of signal processing and statistics. Mathematically speaking, the high-dimensional inverse problem can be formulated as

$$y = Ax + \varepsilon,$$

where $A \in \mathbb{R}^{m \times N}$ is the measurement/design matrix, $y \in \mathbb{R}^m$ is the data, $\varepsilon \in \mathbb{R}^m$ is the noise which is, often, modeled as a random variable following a certain probability distribution, and $x \in \mathbb{R}^N$ is the underlying sparse signal/regression coefficient with at most s coefficients different from zero, i.e., with $||x||_0 \leq s$. And the main question is how to retrieve the signal $x \in \mathbb{R}^N$ from the given data $y \in \mathbb{R}^m$, in a setting where $N \gg m$. Compressive sensing, also known as compressive sampling, sparse recovery, or compressed sensing, has profoundly impacted various fields of science and technology. These innovative techniques have revolutionized how we think about data acquisition, processing, and analysis. Since its first successful application in the field of magnetic resonance imaging (MRI) [LDSP08, JFL15, LDP07, VMA$^+$11] the range of applications has been extended to many fields such as imaging processing in general, telecommunications, image and video compression, biomedical engineering, and remote sensing, where data acquisition costs, storage, and transmission bandwidth are critical factors. There are several books on the topic that discuss its mathematical foundations [FR13, Ela10, Vid19, WM22] as well as books about specific applications of compressive sensing to MRI [Maj15], optics [Ste16], image processing [AH21], radar signal processing [DMEH19] as well as hardware implementation of compressive sensing methodologies [MPC$^+$18].

Several algorithms were devised to find a sparse solution to an underdetermined linear system. We can cite *greedy algorithms*, such as the orthogonal matching pursuit, *thresholding-based algorithms*, such as the hard thresholding pursuit, and *optimization methods*, such as ℓ_1-minimization that aim to minimize a function that promotes sparsity, see [FR13, Chapter 3] or [LW21] for more details on algorithms for sparse recovery. Unfortunately, a least squares solution, or ℓ_2-minimization, does not promote sparsity in the solution. Instead, it produces solutions with smaller but non-zero coefficients for all variables, effectively shrinking the magnitudes of the coefficients (cf. discussion in [HTFF09, Section 3.4.3]. However, in this thesis, we will explore how least squares can be leveraged to develop highly efficient algorithms for efficient reconstruction or estimation of the underlying sparse structure. The algorithm proposed is remarkably simple to implement and versatile since it can be extended to various machine learning problems beyond sparse recovery. Moreover, it is accompanied by robust theoretical guarantees, further enhancing its reliability.

1.3 Least squares in modern machine learning

In recent decades, the rise of machine learning and data-driven approaches has propelled the method of least squares into new realms of application. With the advent of massive datasets and complex models, least squares techniques found widespread use beyond forming the basis of regression models and maximum likelihood estimators in statistics, in problems such as Linear Discriminant Analysis [Ye07], Canonical Correlation Analysis [SJY09], Support Vector Machines [SV99], Importance Estimation [KHS09], Optimal Transport [PBtTB$^+$15], showing that this old technique can be remarkably useful in modern applications.

Furthermore, the concept of overparametrization has gained prominence in modern machine learning [Bel21]. Overparametrized models, characterized by a higher number of parameters than necessary for fitting the data, have shown remarkable capabilities such as generalization to unseen data. Least squares methods and linear models in general, coupled with overparametrization, play a prominent role in understanding large models such as deep neural networks from several different perspectives. As discussed in [HMRT22, Section 1.2], let us consider the scenario where we aim to train a neural network with parameters $\theta \in \mathbb{R}^p$, $f(\cdot;\theta) : \mathbb{R}^d \to \mathbb{R}$, $z \mapsto f(z;\theta)$

from a given i.i.d. dataset $\{(y_i, z_i)\}_{i=1}^n$, $y_i \in \mathbb{R}$, $z_i \in \mathbb{R}^d$.

Assuming the number of parameters is very large, the training process results in minimal changes to θ, when initiated with random parameters $\theta_0 \in \mathbb{R}^p$. Consequently, one tries to understand the linearized model around θ_0. Assuming that the initialization satisfies $f(z;\theta_0) \approx 0$ and denoting $\theta = \theta_0 + \varepsilon$, for a given small ε, we can approximate the statistical model $z \mapsto f(z;\theta)$ using the following equation:

$$z \mapsto \nabla_\theta f(z;\theta_0)^T \varepsilon. \tag{1.2}$$

Although this model remains nonlinear in the input z, it is linear in the parameters ε. Since it is assumed that the initialization θ_0 is random, the high dimensional linear regression problem (with p much greater than n) has random features given by $x_i = \nabla_\theta f(z_i;\theta_0)$, $i = 1,\ldots,n$. Furthermore, due to the dimensionality of the problem ($p > n$), there exist multiple vectors β that result in a model that perfectly interpolates the data. Thus, it is crucial to understand least squares in this context, and a significant body of research is currently being undertaken to fulfill this objective [AZLS19, COB19, JGH18]. In particular, the notion of *neural tangent kernel* plays a major role in this analysis [JGH18, ADH$^+$19].

To summarize, the method of least squares has undergone a transformative journey, from its inception by Legendre and Gauss to its modern applications in machine learning and overparametrized models. The method's versatility, robustness, and rich theoretical foundations have made it a cornerstone of numerical analysis and statistical inference, empowering researchers and practitioners to extract valuable insights from data and build accurate predictive models as well as an understanding of very large models. Here, in this monograph, we delve into the enduring significance of the method of least squares as a valuable source of inspiration for algorithm development in data science and the analysis of complex machine learning scenarios. We explore the reasons why this method remains relevant and is expected to continue shaping advancements in these fields for the foreseeable future.

Contribution: In this thesis, we present a comprehensive study of iterative reweighted least squares (IRLS) algorithms in three different problems, namely, sparse recovery, low-rank matrix completion, and high-dimensional noise-blind regression. We show their theoretical analysis as well as empirical performance. We also present results about constrained least squares and its connection with overparametrization and the implicit bias of the gradient descent. This thesis is divided into four main chapters, each containing an introduction and an extensive literature review. Here, we tried to ensure historical accuracy by including references that initiated specific lines of research or presented results for the first time. Although this task can be challenging, we aimed to provide a comprehensive and precise account of the scientific advancements. Moreover, in each chapter, we left a few open directions and problems that we believe are interesting to pursue. In the concluding chapter, we highlighted two significant areas of research that we consider highly relevant for further exploration and extension of the topics discussed in this work. The contributions of each chapter are as follows:

- **Chapter 2: Global Convergence Rate of IRLS.** This chapter focuses on the problems of sparse recovery from few measurements. Our focus is on how to use least squares to devise an algorithm that provably solves a compressive sensing problem. In particular, we solve an open conjecture and propose an Iteratively Reweighted Least Squares (IRLS) that converges with a global linear convergence rate to the underlying sparse vector. The convergence analysis is performed under sharp, minimal assumptions. This chapter is a joint work with C. Kümmerle and D. Stöger and it is an edited version of the paper *C. Kümmerle, C. Mayrink Verdun, D. Stöger. Iteratively Reweighted Least Squares for Basis Pursuit with Global Linear Convergence Rate.* In Advances in Neural Information Processing Systems 34, 2873-2886, 2021.

- **Chapter 3: Completion of Ill-Conditioned Matrices from Few Measurements.** This chapter focuses on the completion of ill-conditioned low-rank matrices using only a limited number of measurements. We show that algorithms that usually perform well on the recovery of low-rank matrices from few measurements may not do so if the underlying object is an *ill-conditioned* low-rank matrix. We explore novel techniques based on least squares that leverage highly non-convex objective functions to recover missing entries in such matrices in a fast and scalable way. The proposed algorithm and analysis provide insights into solving challenging matrix completion problems with few measurements in highly ill-conditioned scenarios. We establish a local quadratic convergence rate for our method. For the first time, we also provide an extensive numerical comparison of different state-of-the-art methods from the matrix completion literature. This chapter is a joint work with C. Kümmerle, and it is based on the paper *C Kümmerle, C. Mayrink Verdun. A Scalable Second Order Method for Ill-Conditioned Matrix Completion from Few Samples.* In International Conference on Machine Learning, 5872-5883, 2021.

- **Chapter 4: IRLS Algorithm for Noise-Blind High-Dimensional Statistical Problems.**

  This chapter introduces an IRLS algorithm specifically designed for noise-blind high-dimensional statistical problems, focusing on the sqrt-LASSO problem. We address the challenges of handling noisy and high-dimensional datasets by leveraging the power of IRLS. We provide a detailed convergence analysis under general assumptions as well as under assumptions that are commonly used in the sparse recovery literature. We establish a global linear converge rate for the algorithm in the context of high-dimensional statistical inference when assessing the noise level in the measurements is impossible. The results contribute to the development of efficient and accurate solutions for noise-blind estimation problems. This chapter is based on a joint collaboration with O. Melnyk, F. Krahmer, and P. Jung, and the results have not yet been published. The final paper containing these results as well as a numerical evaluation (unfortunately not presented in this thesis) is in preparation.

- **Chapter 5: Overparametrization in Machine Learning for Least Squares Problems.**

  In this chapter, we explore the concept of overparametrization in machine learning models and how this can be used to solve constraint least squares problems. By intentionally introducing additional parameters into the least squares model, we investigate how overparametrization, combined with gradient descent, leads to a biased solution and how this can be harnessed to provide efficient and accurate solutions to the non-negative least squares problem. We analyze the implications of model complexity and generalization, shedding light on the benefits of overparametrization in the context of solving constrained least squares problems. The findings provide valuable insights into leveraging overparametrization as a tool for efficient optimization in machine learning. This chapter is based on a joint collaboration with H.-H. Chou, J. Maly, and H. Mirandola. The preprint *H.-H. Chou, J. Maly, C. Mayrink Verdun. Non-negative Least Squares via Overparametrization. arXiv preprint arXiv:2207.08437* is available on arXiv, and some corrections are being made in order to submit it to a journal.

By investigating these four interconnected topics, we aim to contribute to the understanding and advancement of some machine learning and data science problems. In particular, we advocate for the use of least squares to tackle diverse applications. With that, we hope to pave the way for further research and practical implementations of simple and scalable algorithms for challenging problems.

According to Legendre himself, *"Of all the principles that can be proposed for this purpose, I think there is none more general, more exact, or easier to apply, than that which we have used in this work; it consists of making the sum of the squares of the errors a minimum."* [Leg06, Page 72].

In addition to the contributions discussed in the subsequent chapters, it is worth mentioning that a few other works were written during the period when the research presented in this thesis was conducted. The list of authors in each of these papers is alphabetical, except for the second and he last papers.

- C Kümmerle, C. Mayrink Verdun. "Completion of structured low-rank matrices via iteratively reweighted least squares". 2019 International Conference on Sampling Theory and Applications (SampTA 2019) [KV19].

- C. Mayrink Verdun, et al.. "Group testing for SARS-CoV-2 allows for up to 10-fold efficiency increase across realistic scenarios and testing strategies". Front. Public Health 9, 1205, 2021 [VFH+21]. Featured by David Donoho in his Distinguished Lecture on Mathematics of Data Science: `https://www.youtube.com/watch?v=VOzl-RC4IIs&t=1674s`.

- F. Hoppe, F. Krahmer, C. Mayrink Verdun, M. Menzel, H. Rauhut. "Uncertainty quantification for sparse Fourier recovery". arXiv preprint ArXiv:2212.14864, 2022 [HKV+22].

- F. Hoppe, F. Krahmer, C. Mayrink Verdun, M. Menzel, H. Rauhut. "High-dimensional confidence regions in sparse MRI". IEEE International Conference on Acoustics, Speech and Signal Processing 2023 (ICASSP 2023) [HKV$^+$23a]. This work won the *Best Student Paper Award.*

- F. Hoppe, F. Krahmer, C. Mayrink Verdun, M. Menzel, H. Rauhut. "Sampling Strategies for Compressive Imaging Under Statistical Noise". 2023 International Conference on Sampling Theory and Applications (SampTA 2023) [HKV$^+$23b].

- F. Hoppe, F. Krahmer, H. Laus, C. Mayrink Verdun, H. Rauhut. "Uncertainty Quantification for Learned ISTA". 2023 IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2023) [HVL$^+$23].

- S. Endt, M. Engel, E. Naldi, R. Assereto, M. Molendowska, L. Müller, C. Mayrink Verdun, C. M. Pirkl, M. Palombo, D. K. Jones, M. I. Menzel. "In-vivo myelin water quantification using diffusion-relaxation correlation MRI: a comparison of 1D and 2D methods. Appl. Magn. Reson. 54, 1571–1588 (2023) [EEN$^+$23].

1.4 Notation

In this section, we state standard notational conventions that will be used in the remainder of the thesis. We denote by $\|.\|_p$, for $p > 0$, the ℓ_p-quasinorm of a vector. For $p = 0$, we denote by $\|.\|_0$ the pseudo-norm that counts the number of non-zero entries of a vector. We denote the cardinality of a set I by $|I|$. The support of a vector $x \in \mathbb{R}^N$, i.e., the index set of its nonzero entries, is denoted by $\operatorname{supp}(x) = \{j \in [N] : x_j \neq 0\}$. We call a vector s-sparse if at most s of its entries are nonzero, i.e., if $\|x\|_0 \leq s$. We denote by x_I the restriction of x onto the coordinates indexed by I, and use the notation $I^c := \mathbb{R}^N \setminus I$ to denote the complement of a set I. Furthermore, for $p > 0$, the *ℓ_p-error of best s-term approximation* to a vector $x \in \mathbb{R}^N$ is defined by $\sigma_s(x)_p = \inf\{||x - z||_p, z \in \mathbb{R}^N \text{ is } s\text{-sparse}\}$.

We use $\odot$ to denote the Hadamard product, i.e., the vectors $\mathbf{x}\odot\mathbf{y}$ and $\mathbf{x}^{\odot p}$ have entries $(\mathbf{x}\odot\mathbf{y})_n = x_n y_n$ and $(\mathbf{x}^{\odot p})_n = x_n^p$, respectively. We abbreviate $\tilde{\mathbf{x}} := \bigodot_{k\in[L]} \mathbf{x}^{(k)} = \mathbf{x}^{(1)} \odot \cdots \odot \mathbf{x}^{(L)}$. The logarithm is applied entry-wise to positive vectors, i.e., $\log(\mathbf{x}) \in \mathbb{R}^N$ with $\log(\mathbf{x})_n = \log(x_n)$. For convenience, we denote by $\mathbf{x} \geq \mathbf{y}$ the entry-wise bound $x_n \geq y_n$, for all n, and define $\mathbb{R}_+^N = \{\mathbf{x} \in \mathbb{R}^N : \mathbf{x} \geq \mathbf{0}\}$. The all-zero and all-ones vectors are denoted by $\mathbf{0}$ and $\mathbf{1}$, where the dimension is always clear from the context. For $x_+ \in S_+ := \arg\min_{\mathbf{z}\geq\mathbf{0}} \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2^2.$, we furthermore define

$$\mathbf{y}_+ := \mathbf{A}\mathbf{x}_+, \tag{1.3}$$

which is the unique Euclidean projection of $\mathbf{y}$ onto the convex and closed set

$$C_+ := \{\mathbf{A}\mathbf{z} \colon \mathbf{z} \in \mathbb{R}_{\geq 0}\}. \tag{1.4}$$

We denote the rank and the kernel of a matrix A by $\mathrm{rank}(A)$ and $\ker(A)$. For $0 \leq p \leq \infty$, $\|A\|_{S_p}$ denotes the Schatten-p quasi-norm $\|A\|_{S_p} = (\sum_{i=1}^{d} \sigma_i(A)^p)^{\frac{1}{p}}$ of a matrix $A \in \mathbb{R}^{d_1 \times d_2}$. In particular, for $p = 0$, the Schatten-0 represents the rank of a matrix. For $p = 1$, it is called the nuclear norm, also denoted by $\|A\|_*$. For $p = 2$, it is also called the Frobenius norm, a norm induced by the Frobenius scalar product of two matrices $\langle X, Y \rangle = \mathrm{tr}(X^* Y)$. For $p = \infty$, the norm is also called the spectral norm, which coincides with the largest singular value of the matrix $\sigma_1(A)$. The condition number of a rank-r matrix is denoted by $\kappa(A) = \sigma_1(A)/\sigma_r(A)$, where σ_r denotes the smallest singular value different from zero.

Chapter 2

Iteratively Reweighted Least Squares for Sparse Recovery

"Unfortunately, there is little guidance available on choosing a good technique for a given parameter regime."
J. Tropp and A. Gilbert [TG07].

In this chapter, we present an algorithm for sparse signal recovery. We propose a majorization-minimization approach, drawing inspiration from robust regression techniques. After providing an overview of the existing literature, we delve into the computation and theoretical properties of our proposed method. The work presented in this chapter was written in collaboration with Dr. Christian Kümmerle and Dr. Dominik Stöger and it was published as a spotlight in the Conference on Neural Information Processing Systems (NeurIPS) 2021 under the title *Iteratively Reweighted Least Squares for Basis Pursuit with Global Linear Convergence Rate* [KMVS21]. We highlight our main contribution:

> We establish a new iteratively reweighted least squares method for sparse recovery and we solve an open question from the literature by showing that it converges with a global linear rate under minimal assumptions.

2.1 Introduction

As discussed in Chapter 1, the goal in high-dimensional statistics and in compressive sensing is to retrieve a mathematical structure that has a parsimonious representation, e.g., an (approximately) sparse vector x, from few measurements. Due to the nature of several applications, e.g., magnetic resonance imaging [LL00], it is reasonable to assume that the measurement process is described by a linear operator. In many of these applications, there is a need for saving time or resources, which mathematically translates into acquiring as few measurements as possible [LDP07]. Therefore, signal retrieval, usually a very high-dimensional problem, is then modeled via an underdetermined system of linear equations of the form $y = Ax$ where the number of columns will usually be much larger than the number of rows. In its simplest version, when one assumes that there is no noise in the system, the problem is formulated in the following way:

$$\min_{x \in \mathbb{R}^N} \|x\|_0 \qquad \text{subject to } Ax = y, \tag{P_0}$$

where $\|x\|_0$ denotes the number of nonzero entries of the vector $x \in \mathbb{R}^N$. The study of the optimal solution of (P_0) and its uniqueness was studied in several works, e.g., [GR97b, DE03b, DH$^+$01].

Theorem 2.1.1. *[GR97b, Theorem 1] Suppose that $y = Ax_0$ for a s-sparse vector $x_0 \in \mathbb{R}^N$. If $\ker(A)$ contains no $2s$-sparse vectors, x_0 is the unique optimal solution of (P_0).*

It was proven that the problem above belongs to the class of NP-hard problems, see [Nat95, Theorem 1], by reducing it to the *exact cover by 3-sets problem (X3C)* which, in turn, is an NP-complete problem [Kar72]. This means that solving the problem (P_0) is as hard as the X3C problem. In fact, it was proven that a more general version of the problem, here denoted by $(P_{0,\eta})$, where one assumes not an equality constraint $Ax = y$ but rather the inequality (noisy) constraint $\|Ax - y\|_2 \leq \eta$, is NP-hard. See, e.g., [Nat95, DMA97].

Theorem 2.1.2. *[Nat95, Theorem 1] For any $\eta \geq 0$, the ℓ_0-minimization problem $(P_{0,\eta})$ for general $A \in \mathbb{C}^{m \times N}$ and $y \in \mathbb{C}^m$ is NP-hard.*

2.1.1 The convex relaxation

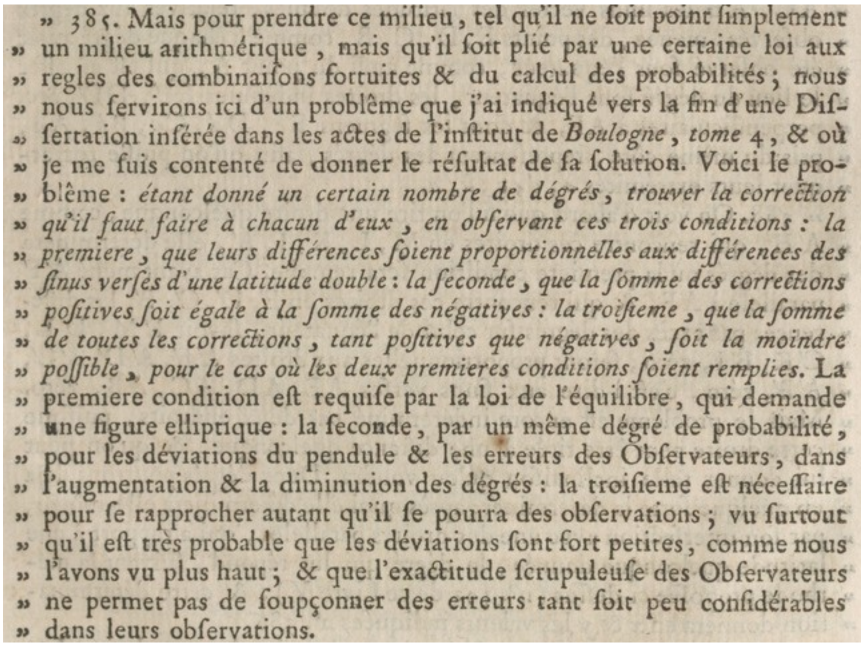
One interesting aspect of this type of result is that it concerns the tractability of the problem in the general case for all possible measurement matrices A and all possible vectors y. In principle, this kind of worst-case analysis tells us nothing about specific choices of A and y and tractable algorithms for these cases.

It is also important to notice that the NP-hardness of this problem lies in the fact that we don't know in principle where the non-zero entries are located. If the sparsity pattern is fixed and equal to s, this problem turns into a convex problem. In the worse case, one needs to solve all $\binom{N}{s}$ sparsity patterns.

One way of tackling this issue is by substituting the ℓ_0-norm by the ℓ_1-norm, which is the convex hull of the intersection of ℓ_0-norm ball with the ℓ_∞-norm ball [WM22, Theorem 2.11] (see also [CRPW12]). This tractable convex formulation, in the context of sparse recovery, was developed and explored in the Ph.D. thesis of S. S. Chen [CD94a] and became known as *Basis Pursuit* [CD94b, CDS01a]. It reads as follows:

$$\min_{x\in\mathbb{R}^N} \|x\|_1 \qquad \text{subject to } Ax = y, \tag{P_1}$$

The minimization of the ℓ_1-*norm* has a long history, before the advent of compressive sensing or high-dimensional statistics and of all the theoretical and numerical achievements of the last 20 years. Such a minimization procedure can date back to the 1750s and the work of Boscovitch [Bos50, She73], which studied the *least absolute deviation* problem for the first time. Boscovitch proposed a method to address errors in the measurements of meridian arcs. He was likely the first scientist to propose a ℓ_1-minimization method. In particular, the following passage can be found in the book by Boscovitch and Maire [MB70, Page 501]:



» 385. Mais pour prendre ce milieu, tel qu'il ne ſoit point ſimplement
» un milieu arithmétique, mais qu'il ſoit plié par une certaine loi aux
» regles des combinaiſons fortuites & du calcul des probabilités; nous
» nous ſervirons ici d'un problême que j'ai indiqué vers la fin d'une Diſ-
» ſertation inſérée dans les actes de l'inſtitut de *Boulogne*, *tome* 4, & où
» je me ſuis contenté de donner le réſultat de ſa ſolution. Voici le pro-
» blême : *étant donné un certain nombre de dégrés, trouver la correction*
» *qu'il faut faire à chacun d'eux, en obſervant ces trois conditions : la*
» *premiere, que leurs différences ſoient proportionnelles aux différences des*
» *ſinus verſés d'une latitude double : la ſeconde, que la ſomme des corrections*
» *poſitives ſoit égale à la ſomme des négatives : la troiſieme, que la ſomme*
» *de toutes les corrections, tant poſitives que négatives, ſoit la moindre*
» *poſſible, pour le cas où les deux premieres conditions ſoient remplies.* La
» premiere condition eſt requiſe par la loi de l'équilibre, qui demande
» une figure elliptique : la ſeconde, par un même dégré de probabilité,
» pour les déviations du pendule & les erreurs des Obſervateurs, dans
» l'augmentation & la diminution des dégrés : la troiſieme eſt néceſſaire
» pour ſe rapprocher autant qu'il ſe pourra des obſervations; vu ſurtout
» qu'il eſt très probable que les déviations ſont fort petites, comme nous
» l'avons vu plus haut; & que l'exactitude ſcrupuleuſe des Obſervateurs
» ne permet pas de ſoupçonner des erreurs tant ſoit peu conſidérables
» dans leurs obſervations.

Figure 2.1: An excerpt from page 501 of the book [MB70]. Available at `https://gallica.bnf.fr/ark:/12148/bpt6k9629131h.texteImage`.

We can highlight the section that states: *"Being given a certain number of degrees, find the correction that must be made to each of them, supposing these three conditions are complied with: …the third, that the sum of all the corrections, positive as well as negative, shall be the least possible …"*

Boscovich initially proposed the method in a geometric and verbal formulation without providing any analytic formulation. It was Laplace who later developed an algebraic treatment of the least absolute error method, building upon Boscovitch's ideas and naming it the *"méthode de situation"* [mdL25]. Interestingly, in an annotated translation of Laplace's book on celestial mechanics to English, Bowditch wrote a footnote saying that Boscovich's method should be preferred over the least squares method because it gave less weight to defective observations. In his own words, *"This method, proposed by Boscovitch, and peculiarly well adapted to the present problem, is not now so much used as it ought to be; instead of it, the principle of making the sum of the squares of the errors a minimum is generally adopted. This method of the least squares...is extremely well adapted to a set of observations, in which all the measured arcs are of nearly the same length and subject to the same degree of uncertainty...But if the measure of one of these degrees should differ very much from the rest, the method of the least squares, applied in the usual manner, would give by far too great an influence to this defective observation, in the determination of the figure of the earth... We shall hereafter find, in several instances, that the method of the least squares, when applied to a system of observations, in which one of the extreme errors is very great, does not generally give so correct a result as the methods proposed by Boscovitch"* [LB32, Page 434]. Bowditch could not have foreseen the profound impact his statement, rediscovered by many researchers, would have on signal processing, robust statistics, and machine learning, even 150 years later. See also [She77], [Sti86, Page 50], [BS83, Section 1.4] and [Far99, Section 9.7] for more details on the (pre-)history of ℓ_1-minimization. Later, several developments in the geophysics community [LF81, CB83, TBM79, WU83, CM73, SS86] led to a better numerical understanding of the properties of this convex but non-differentiable function. In particular, [SS86, DL92] showed that it is possible to retrieve the full wideband seismic signal from incomplete measurements where no low-frequency were acquired due to the nature of the seismic measurements. From the theoretical point of view, the work [Log65] was probably the first to study guarantees for sparse reconstruction via ℓ_1(-norm) minimization.

This problem is equivalent, in the case of real measurements – $A \in \mathbb{R}^{m\times N}$ – to a linear program, a fact that has been known since the 1950s, e.g., [BS83, Chapter 6]. To show this equivalence, we introduce the variables $x^+, x^- \in \mathbb{R}^N$. For $x \in \mathbb{R}^N$, let

$$x_j^+ = \begin{cases} x_j & \text{if } x_j > 0 \\ 0 & \text{if } x_j \leq 0 \end{cases} \qquad \text{and} \qquad x_j^- = \begin{cases} 0 & \text{if } x_j > 0 \\ -x_j & \text{if } x_j \leq 0. \end{cases} \tag{2.1}$$

By doing so, the problem (P_1) is equivalent to the following linear optimization problem for the variables $x^+, x^- \in \mathbb{R}^N$:

$$\min_{x^+, x^- \in \mathbb{R}^N} \sum_{i=1}^{N} (x^+ + x^-) \qquad \text{subject to } [A| - A] \begin{bmatrix} x^+ \\ x^- \end{bmatrix} = y, \quad \begin{bmatrix} x^+ \\ x^- \end{bmatrix} \geq 0. \tag{P_1'}$$

Once the solution of this problem, here denoted by $(x^+)^\#$, $(x^-)^\#$, is obtained, the solution of the original Basis Pursuit (P_1) will be given by $x^{\#} = (x^+)^{\#} - (x^-)^{\#}$. See [Til15, Theorem 2.2] for more details on this equivalence.

Remark 2.1.3. *This equivalence establishes a new interesting research direction for general linear programming (LP). It would be, in theory, possible to obtain new bounds for solving LPs by reducing them to Basis Pursuit and using methods tailored for the latter. See [Til15].*

In the complex case where $x \in \mathbb{C}^N$, this problem is equivalent to a second-order cone program [FR13]. Because of this equivalence, general-purpose methods for linear programming such as the Simplex Method or Interior-Point Methods could, in theory, be used to solve this problem [NN94, BT97]. For example, it is worth mentioning that [CD94a, Section 4.2] starts by emphasizing that *"Basis Pursuit is only thinkable because of recent advances in linear programming via interior point methods"*. This was the starting point for numerical methods for ℓ_1-minimization in high-dimensional statistics and signal processing.

The term *Basis Pursuit* originates from the field of *atomic decomposition* [CDS01b]. It derives its name from the simplex algorithm [BT97, Chapter 3] employed to solve it. The algorithm initially identifies a set of m linearly independent vectors and subsequently iterates by replacing one vector in the set with another not yet included. This process ensures an iteration-wise improvement of the objective function unless the optimal solution has already been attained. This strategy of iteratively improving a basis until a solution is reached is the reason behind the meaning of the name of the method.

Besides being computationally tractable in general, it is also possible to analyze the minimizers of this problem and to understand under which assumptions the minimizer of (P_1) is also the minimizer of (P_0). To do so, various tools were developed, such as the notion of coherence [DE03a] or restricted isometry property (RIP) [BCT11, CT05]. For an analysis of Basis Pursuit based on coherence, see [GN03, DE03b, Fuc04]. For an analysis based on the RIP, see [CT05, CRT06b]. For an analysis of the phase transition phenomenon for ℓ_1-minimization and the probability of success of ℓ_1-minimization, see [DT09a, DT10b, ALMT14, DT09b]. However, the most important theoretical tool for the analysis of ℓ_1-minimization is the *null space property* that will be discussed further below in 2.2.2. This property was distilled in [FN03] and further analyzed in the seminal paper [CDD09], from which it finally got its name.

An unconstrained variant of (P_1), more popular in statistics and machine learning, often called the *Least Absolute Shrinkage and Selection Operator (LASSO)*[1], amounts to

[1] The original LASSO formulation minimizes $||Ax - b||_2$ subject to $||x|| \leq \tau$ [Tib96b]. Later, the unconstrained formulation, originally called *Lagragian LASSO* started to be called LASSO.

the most well-studied tractable estimator for variable selection in high-dimensional inference [Tib96b, HTW19, MB$^+$06]. Besides the conditions cited in the last paragraph, many sufficient conditions were developed to theoretically understand (un)constrained ℓ_1-minimization problems; see [VDGB09] for an overview. This problem has many applications, and it has become ubiquitous in science and engineering; nowadays, it is even called the *modern least squares* [CWB08]. Still, the Basis Pursuit formulation is an optimization program and not an algorithm that finds the sparse vector. So far, nothing has been said on how to actually minimize this convex but non-differentiable function.

2.1.2 Which algorithm should one choose?

As many problems of interest in applications are very high-dimensional, it is usually not a good idea to use general-purpose solvers for minimizing the ℓ_1-norm that do not take the structure of the problem into account. Therefore, a huge amount of research in high-dimensional sparse models was devoted to designing specialized solvers tailored to solve (P_1). Among the most important ones, one can cite the Homotopy Method [DT08], forward-backward algorithms [CW05], Alternating Direction Method of Multipliers [BPC$^+$11], Bregman iterative regularization [YOGD08] and Semismooth Newton Augmented Lagrangian Methods [LST18]. Given the variety of methods available to solve this problem, it is very hard to develop a study that reliably says which method is the best one for a certain task or for a certain parameter regime, as the quote from [TG07] opening this chapter indicates. In particular, a common problem in reproducible research is that it is usually hard to benchmark numerically every single new method that claims to be state-of-the-art for a given numerical problem [2].

The goal of this chapter will be to develop algorithms that have strong provable guarantees (under minimal assumptions) and that are simple to implement and use. Most often, these are the ones that pass the test of time. In Chapter 3, we will show how this kind of idea, when applied to non-convex problems, can lead to excellent numerical performance.

The focus and contributions of this chapter are on another well-established solver for the ℓ_1-minimization problem (P_1), namely, the so-called Iteratively Reweighted Least Squares (IRLS). It corresponds to a family of algorithms that elegantly minimizes non-smooth functions by solving several least squares problems in an iterative way, which motivates the title of this thesis. The idea of this method can be traced back to a method proposed by Weiszfeld[3] for the *Fermat-Weber problem* (that could also be called the Fermat-Torricelli-

[2] The project *State of the Art* from the website `https://paperswithcode.com/sota` is an initiative that tries to present benchmarks and to compare algorithms according to a certain metric for several different machine learning and data science tasks in a systematic way.

[3] Endre Weiszfeld, a Hungarian Jew, fled Europe in 1930 and, upon arriving in the USA, changed

Simpson-Steiner-Weber problem) [Wei37, BS15]. See also [EM11] for historical details about this problem. This problem involves locating a point on the plane that minimizes the total transportation costs from that point to n destination points, i.e., finding a point minimizing the sum of weighted distances from give.

It has numerous uses that go beyond sparse recovery and many problems, such as robust regression in statistics [HW77, MGJK20], total variation regularization in image processing [GR92, NN05, AIG06], joint learning of neural networks [ZHH$^+$19], robust subspace recovery [LM18], numerical methods for elliptic PDEs [DFRW20], design of FIR filters [BBS94], learning sparse and low-rank priors for image problems [LK23] and the recovery of low-rank matrices [MF12a, FRW11a, KS18, KMV21] can be solved efficiently in practice by IRLS. The method relies on efficient and simple linear algebra since one only needs to solve the linear systems arising from the quadratic problems at each iteration without the need for careful initialization or intricate parameter tuning. Despite its successful application in several problems, it is very challenging to analyze it theoretically since the quadratic problem that is solved changes at each iteration.

The main contribution of this chapter is a deeper understanding of this family of methods that led to new formulations and new ways of looking at their theoretical guarantees. In particular, we will establish fast global convergence rates for IRLS for sparse recovery under the most general assumption, namely, the NSP. In Section 2.1.3, we discuss the progress in establishing theoretical guarantees for it. Before we develop our contributions to the theory of IRLS for sparse recovery, in the next section, we will go one step back and discuss the history of IRLS for related problems.

2.1.3 Related work

As discussed in the previous section, IRLS has a long history that dates back to the 1930s. It has appeared under different names within different communities, e.g., similar algorithms are usually called *half-quadratic algorithms* in image processing [Idi01, AIG06] and the *Kacanov method* in numerical PDEs [DFRW20]. It appeared in approximation theory in the study of Chebyshev polynomials, where it became known as *Lawson algorithm* [Cli72, Law61], c.f. the survey [Bur12]. The most common application of IRLS has probably been in robust regression and maximum likelihood estimation problems [HW77, Gre84, Zha21]. For p-norm regression, [APS19] proposed a version of IRLS for which convergence results for $p \in [2, \infty)$ were established, solving a problem that was open for over thirty years. Also, for robust regression, by using an ℓ_1-objective on the

his name to Andrew Vázsonyi. His work, which was originally written in French, was published in the Japanese journal “Tohoku Mathematical Journal” remained unknown for several decades. For an English translation, see [WP09]. And for more details about his life, see [Vaz02].

residual, [MGJK19] recently showed global convergence of IRLS with a linear rate, with high probability for sub-Gaussian data.

The work [BDMS09] discussed the IRLS algorithm for constrained problems with general convex functions and convex constraints that appear in robust regression problems and establishes some convergence results for a general IRLS-type procedure. In [ODBP15], the authors provided a general framework for formulating IRLS algorithms for the optimization of a quite general class of non-convex and non-smooth functions; however, without the smoothing update step that is crucial for the numerical success of IRLS, c.f. Section 2.2. They used techniques developed in [ABS13] to show convergence of the sequence of iterates to a critical point under the Kurdyka-Lojasiewicz property [BDL07]. However, no results about convergence rates were presented.

For the sparse recovery problem, the topic of this chapter, the references [LYW13, FPRW16a, VD17] analyzed IRLS for an unconstrained version of (P_1), which is usually a preferable formulation if the measurements are corrupted by noise. The work [WN10] discussed the connection between IRLS and Sparse Bayesian Learning. Additionally, the work [FPRW16a] addressed the question of how to solve successive quadratic optimization problems. The authors developed a theory that shows, under the null space property, how accurately the quadratic subproblems need to be solved via the conjugate gradient method to preserve the convergence results established in [DDFG10].

A connection between IRLS and bilevel optimization was recently established in [PP21]. This work describes a simple reparametrization of the IRLS formulation for ℓ_p-minimization (with $p \in (2/3, 1)$) that leads to a smooth bilevel optimization problem without any spurious minima, i.e., the stationary points of this new formulation are either global minima or strict saddles.

IRLS was also successfully employed for the so-called subspace prototype problems when one needs to find the median of a dataset [MKPK22], to system identification [BMDFT22] and to point cloud alignment problems [AH15]. For the related problems of low-rank matrix recovery and completion, the topic of Chapter 3, IRLS strategies have emerged as one of the most successful methods in terms of data efficiency and scalability [FRW11a, MF12a, KS18, KMV21]. But before describing the method and our contributions, we present a short overview of how IRLS was developed and used in compressive sensing and high-dimensional statistical problems.

2.1.4 IRLS for sparse recovery

Despite its long history in other fields, as mentioned above, in the sparse recovery context, to the best of the author's knowledge, the first variants of IRLS were introduced in

[RKD99, GR97a], for the ℓ_p-quasinorm minimization problem (P_p) with $0 < p \leq 1$ that is similar to (P_1), but with $\|x\|_p$ instead of $\|x\|_1$ as an objective. It is important to note that unlike most of the methods cited above, IRLS is one of the few methods (ADMM being the other one, e.g., [BPC$^+$11]) that provides a framework to solve both constrained and unconstrained formulations of ℓ_p-minimization problems.

$$\min_{x \in \mathbb{R}^N} \|x\|_p^p \qquad \text{subject to } Ax = y, \tag{P_p}$$

In [CY08], modifications of the method of [RKD99, GR97a] using specific smoothing parameter update rules, c.f. Section 2.2, were observed to exhibit excellent numerical performance for solving (P_p), retrieving the underlying sparse vector with very few measurements when most of the methods fail. This excellent numerical performance, where saddle points can be avoided, combined with its simplicity to tackle highly non-convex problems such as the minimization of ℓ_p-quasinorms for $p \ll 1$, is one of the main reasons for the popularity of IRLS [CY08, DDFG10]. A measure for the popularity of IRLS, only considering the use of the method for sparse recovery, can be highlighted by the number of Google Scholar citations of the four key papers about it [GR97a, CY08, DDFG10, LYW13], which surpassed 5500 as of the writing of this thesis. While we do not study these non-convex variants here in this chapter, in Chapter 3 we will develop the theory and practice of IRLS for non-convex objective functions that appear in the matrix completion problem. For now, this chapter aims to understand the convergence of IRLS in the case of ℓ_1-norm as our objective function. Since the seminal paper [DDFG10], that fundamentally changed the way this type of algorithm is analyzed, despite several extensions and analyses of the IRLS algorithm [ABH19, FPRW16a], the following fundamental algorithmic question has remained unanswered:

What is the global convergence rate of the IRLS algorithm for ℓ_1-minimization?

Contribution of this chapter:

We resolve this question, formally stated in [SV21], and present a new IRLS algorithm that *converges linearly* to a sparse ground truth, *starting from any initialization*, as stated in Theorem 2.3.3. Our algorithm returns a feasible solution with δ-accuracy, i.e., $\|x_* - x^k\|_1 \leq \delta$, where x_* is the underlying s-sparse vector, in $k = O(N\sqrt{(\log N)/m}\log(1/\delta))$ iterations. Analogous to [DDFG10], it is assumed that the measurement matrix A satisfies the so-called null space property [CDD09], which is the minimum possible assumption required for sparse recovery. We also provide a similar result for approximately sparse vectors. Our proof relies on a novel quantification of the descent of a carefully chosen objective function in the direction of the ground truth given by a simple quadratic polynomial equation. Additionally, we support the theoretical claims with numerical simulations indicating that we capture the correct dimension dependence. We also believe that the new analysis techniques in this paper are of independent interest and will pave the way for establishing global convergence rates for other variants of IRLS, such as in low-rank matrix recovery [FRW11a] or in the analysis of certain numerical methods for PDEs [DFRW20].

Remark 2.1.4. *The work [DDFG10] states that the algorithm converges* exponentially fast *in the iteration number. Here we will adopt the standard nomenclature from optimization theory and refer to it as* linear convergence*.*

2.2 IRLS for ℓ_1-minimization

We start by deriving the Iteratively Reweighted Least Squares (IRLS) algorithm for ℓ_1-minimization. The idea behind the method is that the ℓ_1-norm can be approximated by a weighted ℓ_2-norm $\|x\|_{2,w} = \sum_{i=1}^N w_i x_i^2$ and, therefore, a non-smooth convex problem can be turned into a quadratic problem, i.e., a least squares one. In fact, one can write $|x_i| = \frac{x_i^2}{|x_i|}$. The problem with such a calculation is that the ground-truth vector to be retrieved via ℓ_1-minimization will typically be sparse, which makes most of the coefficients $x_i = 0$.

In this case, a smoothing parameter ε needs to be introduced to guarantee that the denominator will not blow up. There are several ways to do so. The most common way found in the literature is given by

$$|x_i| = \frac{x_i^2}{|x_i|} = \frac{x_i^2}{\sqrt{|x_i|^2}} \approx \frac{x_i^2}{\sqrt{|x_i|^2 + \varepsilon}}, \tag{2.2}$$

for a very small $\varepsilon \neq 0$. Another way to do so is by using that $x_i \approx \max(x_i, \varepsilon)$, i.e.,

$$|x_i| = \frac{x_i^2}{|x_i|} \approx \frac{x_i^2}{\max(x_i, \varepsilon)} := w_i x_i^2. \tag{2.3}$$

While the first approach was quite well studied in the literature of smooth methods for the ℓ_1-norm, in this thesis, we will depart from this view and explore the second (sharper) approach. In particular, we will show that such an approach leads to better convergence guarantees as well as better numerical properties.

Remark 2.2.1. *The smoothing parameter ε was introduced in order to deal with sparse vectors in high-dimensional problems. The IRLS algorithm is well-defined and applicable even in the absence of such a parameter in its description, as observed in the original*

Weiszfeld method. However, it is important to note that in such instances, there is a possibility that the iterates $\{x_k\}$ may, eventually, coincide with one point in the set of destination points, the so-called anchor set. This scenario leads to a division by zero in the calculation for subsequent iterates. For precise conditions and a thorough convergence analysis of the method, see [BS15].

One of the key points in our work is the interpretation of IRLS algorithms as a variant of a Majorize-Minimize (MM) algorithm [SBP17, Lan16], as we will lay out in the following paragraphs. This point of view differs from [DDFG10], which considered the IRLS method as a consequence of a variational principle. Our derivation, on the other hand, is highly inspired by robust statistics techniques [Hub64, Mey21].

Remark 2.2.2. *Another approach for sparse recovery that was quite popular for some time in the literature was the* iteratively reweighted ℓ_1-norm *which, as the name suggested, solves a sequence of weighted ℓ_1-minimization problems [CWB08]. However, the convergence analysis for this method is very intricate [WZW22]; IRLS is computationally simpler due to the least squares minimization, and it was shown that it also leads to better numerical performance. See [CY08] and [DDFG10, Section 8.2].*

It mitigates the non-smoothness of the $\|\cdot\|_1$-norm by introducing a smoothed objective function $\mathcal{J}_\varepsilon : \mathbb{R}^N \to \mathbb{R}$, which is defined, for a given $\varepsilon > 0$, by

$$\mathcal{J}_\varepsilon(x) := \sum_{i=1}^{N} j_\varepsilon(x_i) \quad \text{with} \quad j_\varepsilon(x) := \begin{cases} |x|, & \text{if } |x| > \varepsilon, \\ \frac{1}{2}\left(\frac{x^2}{\varepsilon} + \varepsilon\right), & \text{if } |x| \leq \varepsilon. \end{cases} \tag{2.4}$$

We note that the relationship between the ℓ_1-norm and its smoothed version $\mathcal{J}_\varepsilon$ of Equation 2.4 is very similar to the smoothing achieved by using Huber M-estimators instead of ℓ_1-residuals in robust regression [LS98]. Moreover, the function $\mathcal{J}_\varepsilon$ is continuously differentiable and fulfills $|x| \leq j_\varepsilon(x) \leq |x| + \varepsilon$ for each $x \in \mathbb{R}$.

The reason for choosing such function $\mathcal{J}_\varepsilon$, which can be considered as a scaled Huber loss function, can be explained from several points of view. This function is by no means the only approximation to the ℓ_1-norm. See, e.g. [BBZ04, Appendix A], [Kal18] and [Bar19]. For a general discussion on this problem and the importance of this technique in optimization, see [BT12, Nes05]. Here, we follow the presentation in [Bec17] and start by defining the notion of a smooth function and a smoothable function.

Definition 2.2.3. *Let $L \geq 0$. A function $f : \mathbb{R}^N \to (-\infty, \infty]$ is said to be L-smooth over a set $D \subset \mathbb{R}^N$ if its gradient is a Lipschitz function, i.e., if it is differentiable over D and satisfies*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \textit{for all} x, y \in D.$$

The constant L is called the smoothness parameter.

An important step in the analysis of non-smooth functions and the design of algorithms to minimize them is the concept of a *smoothable function.*

Definition 2.2.4. *A convex function $h : \mathbb{R}^n \to \mathbb{R}$ is called (α, β)-smoothable $\alpha, \beta > 0$ if for any $\mu > 0$ there exists a convex differentiable function $h_\mu : \mathbb{R}^n \to \mathbb{R}$ such that the following holds*

I. $h_\mu(x) \leq h(x) \leq h_\mu + \beta\mu$ for all $x \in \mathbb{R}^N$.

ii. h_μ is $\frac{\alpha}{\mu}$-smooth.

The function h_μ is called a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β).

We also need the notion of the Moreau envelope of a function h [Mor65]. Given a proper closed convex function $h : \mathbb{R}^n \to (-\infty, \infty]$ and $\mu > 0$, the Moreau envelope of f is given by the function

$$M_h^\mu(x) = \min_{u \in \mathbb{R}^n} \{h(x) + \frac{1}{2\mu}\|x - u\|^2\} \tag{2.5}$$

With this tool at our disposal, it is possible to establish that the Moreau envelope of a Lipschitz convex function is a $\frac{1}{\mu}$-smooth approximation.

Theorem 2.2.5. *[Bec17, Theorem 10.51] Let $h : \mathbb{R}^n \to \mathbb{R}$ be a convex function satisfying*

$$|h(x) - h(y)| \leq \ell_h \|x - y\| \quad \textit{for all } x, y \in \mathbb{R}^N.$$

Then for any $\mu > 0$, the Moreau envelope M_h^μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(1, \frac{\ell_h^2}{2})$.

As a consequence of this theorem, every convex and Lipschitz with constant ℓ_h is $(1, \frac{\ell_h^2}{2})$-smoothable. Since the Moreau envelope of the absolute value function $|.|$ is given by the Huber function [Bec17, Example 6.54],

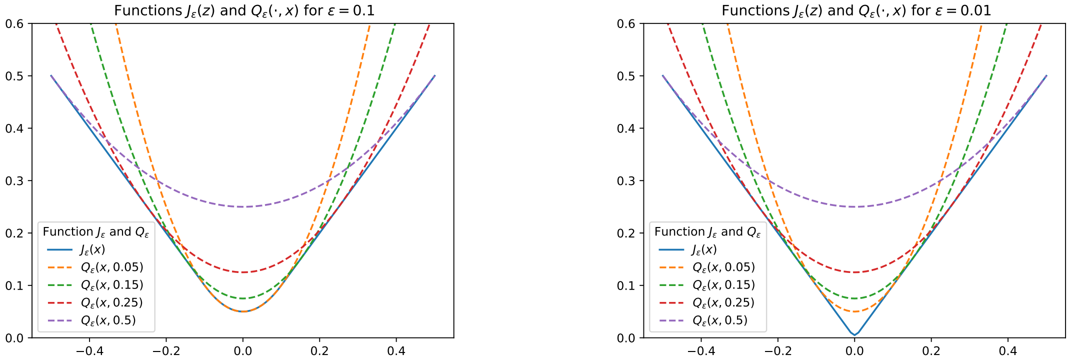
$$H_\mu(x) = \begin{cases} \frac{1}{2\mu}|x|^2, & |x|^2 \leq \mu \\ |x| - \frac{\mu}{2}, & |x|^2 > \mu, \end{cases} \tag{2.6}$$

and the Moreau envelope of separable functions is given by the sum of the Moreau envelopes for each coordinate [Bec17, Theorem 6.58], we have a natural candidate for smoothing the ℓ_1-norm. To simplify calculations, our IRLS method is based on a slightly modified version of the Huber loss, Equation 2.4.

The central idea of IRLS is to use several iterations of the least squares method for a smooth approximation of the objective function. A different quadratic function will be created and minimized in each iteration with standard numerical methods for least squares [Bjö96]. Instead of minimizing the function $\mathcal{J}_\varepsilon$ directly, the idea of IRLS is to minimize instead a suitable chosen quadratic function $Q_\varepsilon(\cdot, x)$, which majorizes $\mathcal{J}_\varepsilon$ such that $Q_\varepsilon(z, x) \geq \mathcal{J}_\varepsilon(z)$ for all $z \in \mathbb{R}^N$. This function is furthermore chosen such that $Q_\varepsilon(x, x) = \mathcal{J}_\varepsilon(x)$ holds, which implies that $\min\limits_{z \in \mathbb{R}^n} Q_\varepsilon(z, x) \leq \mathcal{J}_\varepsilon(x)$. The latter inequality implies that by minimizing $Q_\varepsilon(\cdot, x)$, IRLS actually achieves an improvement in the value of $\mathcal{J}_\varepsilon$ as well. More specifically, $Q_\varepsilon(\cdot, x)$ is defined by

$$\begin{aligned} Q_\varepsilon(z, x) &:= \mathcal{J}_\varepsilon(x) + \langle \nabla \mathcal{J}_\varepsilon(x), z - x \rangle + \frac{1}{2}\langle (z - x), \operatorname{diag}(w_\varepsilon(x))(z - x) \rangle \\ &= \mathcal{J}_\varepsilon(x) + \frac{1}{2}\langle z, \operatorname{diag}(w_\varepsilon(x))z \rangle - \frac{1}{2}\langle x, \operatorname{diag}(w_\varepsilon(x))x \rangle, \end{aligned} \tag{2.7}$$

where $\nabla \mathcal{J}_\varepsilon(x) = \left(\begin{cases} \frac{x_i}{|x_i|}, \text{if} & |x_i| > \varepsilon \\ \frac{x_i}{\varepsilon}, \text{if} & |x_i| \leq \varepsilon \end{cases} \right)_{i=1}^N$ is the gradient of $\mathcal{J}_\varepsilon$ at x and the weight vector $w_\varepsilon(x) \in \mathbb{R}^N$ is a vector of *weights* such that $w_\varepsilon(x)_i := [\max(|x_i|, \varepsilon)]^{-1}$ for $i \in [N]$. Figure 2.2 illustrates the function $J_\varepsilon(x)$ and its quadratic majorizer $Q_\varepsilon(x, z)$ for different values of z.



(a) J_ε and Q_ε with $\varepsilon = 0.1$.

(b) J_ε and Q_ε with $\varepsilon = 0.01$.

Figure 2.2: J_ε and its quadratic majorizer $Q_\varepsilon(x, z)$ for $z = 0.05, 0.15, 0.25, 0.5$.

The following lemma shows that $Q_\varepsilon(\cdot, \cdot)$ has indeed the above-mentioned properties.

Lemma 2.2.6. *Let $\varepsilon > 0$, let $\mathcal{J}_\varepsilon : \mathbb{R}^N \to \mathbb{R}$ be defined as in (2.4) and $Q_\varepsilon : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ as defined in (4.14). Then, for any $z, x \in \mathbb{R}^N$, the following affirmations hold:*

i. $\operatorname{diag}(w_\varepsilon(x))x = \nabla \mathcal{J}_\varepsilon(x)$, *ii.* $Q_\varepsilon(x, x) = \mathcal{J}_\varepsilon(x)$, *iii.* $Q_\varepsilon(z, x) \geq \mathcal{J}_\varepsilon(z)$.

Proof. We prove each of the three statements separately.

1. Let $x \in \mathbb{R}^N$. Then the i-th coordinate of $\operatorname{diag}(w_\varepsilon(x))x$ is given by

$$(\operatorname{diag}(w_\varepsilon(x))x)_i = \begin{cases} \frac{x_i}{|x_i|} = \operatorname{sgn}(x_i), & \text{if } |x_i| > \varepsilon, \\ \frac{x_i}{\varepsilon}, & \text{if } |x_i| \le \varepsilon. \end{cases} = j_\varepsilon'(x_i) = \left(\nabla \mathcal{J}_\varepsilon(x)\right)_i,$$

where $\mathcal{J}_\varepsilon(x)$ is the gradient of $\mathcal{J}_\varepsilon$ at x.

2. This follows directly from the definition of $Q_\varepsilon(x,z)$ and by setting $x = z$.

3. We define $I := \{i \in [N] : |x_i| > \varepsilon\}$ and write the difference $Q_\varepsilon(z,x) - \mathcal{J}_\varepsilon(z)$ as

$$\begin{aligned} Q_\varepsilon(z,x) - \mathcal{J}_\varepsilon(z) &= \frac{1}{2}\left(\langle z, \operatorname{diag}(w_\varepsilon(x))z\rangle - \langle x, \operatorname{diag}(w_\varepsilon(x))x\rangle\right) \\ &= \sum_{i\in I}\left(\frac{1}{2}|x_i| + \frac{1}{2}\frac{z_i^2}{|x_i|} - j_\varepsilon(z_i)\right) + \sum_{i\in I^c}\left(\frac{1}{2}\varepsilon + \frac{1}{2}\frac{z_i^2}{\varepsilon} - j_\varepsilon(z_i)\right) \end{aligned}$$

and show that each summand of the two sums is non-negative. In particular, if $i \in I$, then assume first that $|z_i| > \varepsilon$. Then

$$\frac{1}{2}|x_i| + \frac{1}{2}\frac{z_i^2}{|x_i|} - j_\varepsilon(z_i) = \frac{1}{2}\left(|x_i| + \frac{z_i^2}{|x_i|}\right) - |z_i| \ge |z_i| - |z_i| = 0$$

due to inequality $a \le \frac{1}{2}(a^2/b + b)$, which holds for any $b > 0$.

On the other hand, if $|z_i| \le \varepsilon$, then

$$\begin{aligned} \frac{1}{2}|x_i| + \frac{1}{2}\frac{z_i^2}{|x_i|} - j_\varepsilon(z_i) &= \frac{1}{2}|x_i| + \frac{1}{2}\frac{z_i^2}{|x_i|} - \frac{1}{2}\left(\frac{z_i^2}{\varepsilon} + \varepsilon\right) \\ &= \frac{1}{2}\big(|x_i| - \varepsilon\big) + \frac{1}{2}z_i^2\left(\frac{1}{|x_i|} - \frac{1}{\varepsilon}\right) \\ &\ge \frac{1}{2}\big(|x_i| - \varepsilon\big) + \frac{1}{2}\varepsilon^2\left(\frac{1}{|x_i|} - \frac{1}{\varepsilon}\right) \\ &= \frac{1}{2}\left(|x_i| + \frac{\varepsilon^2}{|x_i|}\right) - \varepsilon \ge \varepsilon - \varepsilon = 0, \end{aligned} \tag{2.8}$$

where we used that $\frac{1}{|x_i|} - \frac{1}{\varepsilon} < 0$ in the first inequality. In the second inequality, we again used $a \le \frac{1}{2}(a^2/b + b)$ for any $b > 0$. Now let $i \in I^c$. We again consider the two cases, $|z_i| \le \varepsilon$ and $|z_i| > \varepsilon$. In the first case we have that $\frac{1}{2}\varepsilon + \frac{1}{2}\frac{z_i^2}{\varepsilon} - j_\varepsilon(z_i) = 0$, and in the second case we have that

$$\frac{1}{2}\varepsilon + \frac{1}{2}\frac{z_i^2}{\varepsilon} - j_\varepsilon(z_i) = \frac{1}{2}\varepsilon + \frac{1}{2}\frac{z_i^2}{\varepsilon} - |z_i| \ge |z_i| - |z_i| = 0,$$

which concludes the proof.

□

As can be seen from the equality in Equation (2.7), minimizing of $Q_\varepsilon(\cdot, x)$ corresponds to a minimizing *(re-)weighted least squares objective* $\langle \cdot, \operatorname{diag}(w_\varepsilon(x)) \cdot \rangle$, which lends its name to the method. A pivotal point in the design of the algorithm, and one of our major contributions to the theory of IRLS, is the observation that the choice of an objective function that approximates the ℓ_1-norm (or any other non-smooth function for which an IRLS algorithm needs to be designed) should match the choice of weights. In particular, one can see from Equation (4.14) that due to the approximation Equation (2.3), the term $\langle \nabla \mathcal{J}_\varepsilon(x), z - x \rangle$ is canceled out due to our choice of weights, which leads to a "pure" quadratic problem without any linear terms. However, unlike a classical MM approach, IRLS comes with an *update* step *of the smoothing parameter* ε at each iteration. This update is crucial for its excellent numerical performance but makes the analysis of the method much harder. We outline the method in Algorithm 1.

Algorithm 1 Iteratively Reweighted Least Squares for ℓ_1-minimization

```
**Input:** Measurement matrix $A \in \mathbb{R}^{m \times N}$, data vector $y \in \mathbb{R}^m$,
initial weight vector $w_0 \in \mathbb{R}^N$ (default: $w_0 = (1, 1, \ldots, 1)$).
Set $\varepsilon_0 = \infty$.
for $k = 0, 1, 2, \ldots$ do
    $x^{k+1} := \arg\min_{z \in \mathbb{R}^N} \langle z, \operatorname{diag}(w_k) z \rangle$ subject to $Az = y$,   (2.9)
    $\varepsilon_{k+1} := \min\left(\varepsilon_k, \frac{\sigma_s(x^{k+1})_{\ell_1}}{N}\right)$,   (2.10)
    $(w_{k+1})_i := \frac{1}{\max\left(|x_i^{k+1}|, \varepsilon_{k+1}\right)}$ for each $i \in [N]$,   (2.11)
end for
return Sequence $(x^k)_{k \geq 1}$.
```

A consequence of Lemma 2.2.6, step Equation 2.10, the fact that $\varepsilon \to \mathcal{J}_\varepsilon(z)$ is monotonously non-decreasing, and that $k \mapsto \varepsilon_k$ is non-increasing so $k \mapsto \mathcal{J}_{\varepsilon_k}(z)$ is non-increasing in k. This implies that the iterates x^k, x^{k+1} of Algorithm 1 fulfill

$$\mathcal{J}_{\varepsilon_{k+1}}(x^{k+1}) \leq \mathcal{J}_{\varepsilon_k}(x^{k+1}) \leq Q_{\varepsilon_k}(x^{k+1}, x^k) \leq Q_{\varepsilon_k}(x^k, x^k) = \mathcal{J}_{\varepsilon_k}(x^k). \tag{2.12}$$

This shows in particular that the sequence $\left\{\mathcal{J}_{\varepsilon_k}\left(x^k\right)\right\}_{k=0}^{\infty}$ is non-increasing. For this reason, it can be shown that each accumulation point of the sequence of iterates $(x^k)_{k \geq 0}$ is a (first-order) stationary point of the smoothed ℓ_1-objective $J_{\overline{\varepsilon}}(\cdot)$ subject to the measurement constraint imposed by A and y, where $\overline{\varepsilon} = \lim_{k\to\infty} \varepsilon_k$. See [DDFG10, Theorem 5.3] for the proof for a related IRLS algorithm.

2.2.1 Computational considerations

While the crucial steps for the success of IRLS are the weighting choice and the update of the smoothing parameter ε, the core of the method is the solution of the least squares problem at each iteration. Roughly speaking, the idea is to smooth the original non-smooth function and, then, to design a sequence of parabolas that will get closer and closer to the original function as we move towards the minimum as Figure 2.3 shows.

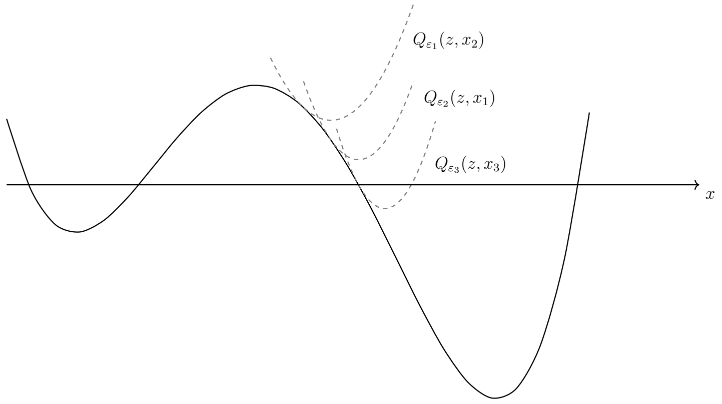

$Q_{\varepsilon_1}(z, x_2)$
$Q_{\varepsilon_2}(z, x_1)$
$Q_{\varepsilon_3}(z, x_3)$
x

Figure 2.3: Majorization-minimization scheme like Algorithm 1.

When $A \in \mathbb{R}^{m\times N}$ $(m \leq N)$ is of full rank, its pseudo-inverse will be given by $A^\dagger = A^*(AA^*)^{-1}$. Hence, by making the substitution $x = W_k^{1/2} z$, where $W_k = \operatorname{diag}(w_k)$, the weighted quadratic problem

$$\min_{z\in\mathbb{R}^N} \langle z, \operatorname{diag}(w_k)\, z\rangle = \|z\|^2_{2,w} = (\sum_{i=1}^{N} |z_j|^2 w_j) \quad \text{subject to } Az = y \tag{2.13}$$

is equivalent to a standard least squares problem with the constraint given by $AW^{-\frac{1}{2}}x = y$. Therefore, the constrained weighted least squares update (2.9) can be computed such that $x^{k+1} = W_k^{-1}A^*(AA^*)^{-1}(y) = W_k^{-1}A^*(AW_k^{-1}A^*)^{-1}(y)$ with $W_k = \operatorname{diag}(w_k)$, with the solution of the $(m \times m)$ linear system $(AW_k^{-1}A^*)z = y$ as a main computational step. This linear system is positive-definite and suitable for the use of iterative solvers such as Krylov methods [Saa03, HS52b, Meu06]. This is especially advantageous if the measurement matrix A is sparse or allows for fast matrix-vector multiplications [Vor12, FPRW16a].

The convergence theory for iterative methods, such as CG, for IRLS, was finally established by [FPRW16a]. However, their analysis was indirect since the shape of the spectrum of the system matrix $AW_k^{-1}A^*$, which plays a major role in the convergence of CG iterations, was not calculated, and the issue of bad conditioning of the IRLS system matrices was not solved. Before this work, it is interesting to note that a few works already used conjugate gradient (CG) in the context of IRLS for sparse recovery, e.g., [Vor12, VD17] for ℓ_1-minimization and [VO98] for TV-minimization. However, none of these works analyzed the convergence of iterative methods to solve the linear system behind IRLS.

The authors of [FPRW16a] provide conditions on the accuracy of conjugate gradient solvers of the successive linear systems for an IRLS algorithm for basis pursuit using weights that are given by

$$(w_k)_i = 1/\sqrt{|x_i^k|^2 + \varepsilon_k^2} \tag{2.14}$$

for each $i \in [n]$ (coinciding with the choice of the weights of [DDFG10, ABH19]).

Independently of how we perform the inner iteration of the method, i.e., how we solve the linear system at each iteration, one caveat of IRLS methods is the fact that the linear system $(AW_k^{-1}A^*)z = y$ becomes highly ill-conditioned once the algorithm gets closer to the ground truth – a sparse vector – since the weights, as well as the smoothing parameter ε_k, will get close to zero [FPRW16a, Section 5.2]. It has been observed that the blow-up of the condition number of $AW_k^{-1}A^*$ can be a problem for an inexact solver of the weighted least squares system [Vor12]. In particular, this numerical issue is shared by many of the IRLS methods developed in the literature so far. For example, it was observed that the IRLS proposed in [DDFG10] usually does not converge when applied to ℓ_p-minimization for $p < 0.5$ due to ill-conditioned linear systems [DDFG10, Section 8]

Another interesting aspect of our contribution is that unlike previous formulations of IRLS [DDFG10, FPRW16a, ABH19], our formulation tackles this issue, previously open in the literature for IRLS methods, and we can establish that our choice of weights leads to a well-conditioned linear system.

The starting point of our algorithm is the observation that for our choice of weight (2.11), i.e., for

$$(w_k)_i := \frac{1}{\max\left(|x_i^k|, \varepsilon_k\right)}$$

for each $i \in [N]$, unlike for (2.14), it is possible to write the inverse weight matrix $W_k^{-1} \in \mathbb{R}^{N\times N}$ such that

$$W_k^{-1} = W_{I_k}^{-1} + \varepsilon_k\left(\mathrm{Id}_N - P_{I_k}\right) = \left(W_{I_k}^{-1} - \varepsilon_k P_{I_k}\right) + \varepsilon_k \mathrm{Id}, \tag{2.15}$$

where $I_k := \{i \in [N] : |x_i^k| > \varepsilon_k\}$, $W_{I_k}^{-1} \in \mathbb{R}^{N\times N}$ is the diagonal matrix with entries $|x_i^k|$ if $i \in I_k$ and 0 otherwise, and P_{I_k} denotes the projection matrix such that $P_{I_k}x = x_{I_k}$. In particular, we observe that $(W_k)^{-1}$ is the sum of a scaled identity matrix Id and a (diagonal) matrix with only $s_k := |I_k|$ non-zero entries. Furthermore, due to the update rule Equation 2.10 of the smoothing parameter ε_k, it can be seen that s_k is small and of the order of s if ε_k approaches 0, i.e. if the k-th iterate x^k of Algorithm 1 has only small coordinates outside a set of s large coordinates. We show this by reformulating the main computational step, i.e., the weighted least squares problem such that x^{k+1} can be computed by solving a positive definite linear system of size $(s_k \times s_k)$, which is well-conditioned. In this way, we avoid solving the ill-conditioned system $\left(AW_k^{-1}A^*\right) z = y$. In order to do so, we need the Sherman-Morrison-Woodbury matrix inversion formula [Woo50]:

Lemma 2.2.7. *[Woo50] Let $B \in \mathbb{R}^{n\times n}, C \in \mathbb{R}^{k\times k}, E \in \mathbb{R}^{n\times k}$ and $F \in \mathbb{R}^{k\times n}$. Then, $(ECF^* + B)$ is invertible if and only if C and $F^*B^{-1}E$ are invertibles and it holds that*

$$(ECF^* + B)^{-1} = B^{-1} - B^{-1}E(C^{-1} + F^*B^{-1}E)^{-1}F^*B^{-1}$$

Our implementation uses the matrix $V \in \mathbb{R}^{m\times N}$ with orthonormal columns which denotes the projection onto the range space of the measurement matrix $A \in \mathbb{R}^{m\times N}$, as well as the left singular vector matrix $U \in \mathbb{R}^{m\times m}$ of A and the diagonal matrix $\Sigma_A \in \mathbb{R}^{m\times m}$ containing the singular values of A. These can be pre-computed before using IRLS, for example, via a singular value decomposition of A. Likewise, the vector

$$\widetilde{y} = V\Sigma_A^{-1}(U^*y) \tag{2.16}$$

can be pre-computed and can be re-used at each outer iteration of IRLS.

In the following, for $I \subset [N]$, we denote by $M_I \in \mathbb{R}^{m\times|I|}$ the restriction of a matrix $M \in \mathbb{R}^{m\times N}$ to the columns indexed by I, and by $Q_{I_k} \in \mathbb{R}^{N\times I_k}$ the projector matrix such that $P_{I_k} = Q_{I_k}Q_{I_k}^*$. Furthermore, let $D_{I_k}^{-1} \in \mathbb{R}^{I_k\times I_k}$ be a diagonal matrix such that $(D_{I_k}^{-1})_{ii} = |x_i^k|$ for each $i \in I_k$. Now, we verify that x^{k+1} as computed by Algorithm 2 is a solution of the weighted least squares problem Equation 2.9.

Lemma 2.2.8. *If $x^{k+1} \in \mathbb{R}^N$ is the output of Algorithm 2, then x^{k+1} coincides with the solution of the weighted least squares problem Equation 2.9.*

Proof of Lemma 2.2.8. We recall from above that if $x_*^{k+1} \in \mathbb{R}^N$ is the solution of Equation 2.9, it holds that $x_*^{k+1} = W_k^{-1}A^*z$ where z is as in $\left(AW_k^{-1}A^*\right) z = y$. Using Equation 2.15, we see that

$$AW_k^{-1}A^* = A\left(\left(W_{I_k}^{-1} - \varepsilon_k P_{I_k}\right) + \varepsilon_k \mathrm{Id}\right)A^* = A_{I_k}\left(D_{I_k}^{-1} - \varepsilon_k \mathrm{Id}\right)A_{I_k}^* + \varepsilon_k AA^*.$$

By identifying $B := \varepsilon_k AA^*$, $C := \left(D_{I_k}^{-1} - \varepsilon_k \mathrm{Id}\right)$ and $E = F := A_{I_k}$, and by noting that

Algorithm 2 Practical implementation of weighted LS step of IRLS for small ε_k

```
Input: Matrix $V \in \mathbb{R}^{m \times N}$ projecting onto range space of measurement matrix $A$, $\widetilde{y} \in \mathbb{R}^N$ from (2.16), smoothing parameter $\varepsilon_k$, projection $\gamma_k^{(0)} = Q_{I_k}^* Q_{I_{k-1}}(\gamma_{k-1}) \in \mathbb{R}^{I_k}$ of solution $\gamma_{k-1} \in \mathbb{R}^{I_{k-1}}$ of linear system (2.17) for previous iteration $k-1$.
Compute $h_k^0 = Q_{I_k}^* \widetilde{y} - \left(\varepsilon_k \left(D_{I_k}^{-1} - \varepsilon_k \mathrm{Id}\right)^{-1} + (V^*)_{I_k}^* (V^*)_{I_k}\right) \gamma_k^{(0)} \in \mathbb{R}^{I_k}$.
Solve
    $\left(\varepsilon_k \left(D_{I_k}^{-1} - \varepsilon_k \mathrm{Id}\right)^{-1} + (V^*)_{I_k}^* (V^*)_{I_k}\right) \Delta\gamma_k = h_k^0 \qquad (2.17)$
  for $\Delta\gamma_k \in \mathbb{R}^{I_k}$ by the *conjugate gradient* method [HS52b, Meu06].
Compute $\gamma_k = \gamma_k^{(0)} + \Delta\gamma_k \in \mathbb{R}^{I_k}$.
Compute residual $r_{k+1} := \widetilde{y} - V(V^*)_{I_k}(\gamma_k) \in \mathbb{R}^N$.
Set $x^{k+1} = r_{k+1}$.
Set $(x^{k+1})_{I_k} = (x^{k+1})_{I_k} + \gamma_k$.
Output: $x^{k+1} \in \mathbb{R}^N$ and $\gamma_k \in I_k$.
```

the matrix C is invertible, since, on the set I_k, we have $|x_i^k| > \varepsilon_k$, we obtain by using Lemma 2.2.7 that

$$\left(A(W_k)^{-1}A^*\right)^{-1} = \varepsilon_k^{-1} Z - \varepsilon_k^{-1} Z A_{I_k} G^{-1} A_{I_k}^* Z,$$

where we used the notation $Z := (AA^*)^{-1}$ and $G := \varepsilon_k C^{-1} + A_{I_k}^* Z A_{I_k}$. Therefore, we have

$$z = \left(A(W_k)^{-1}A^*\right)^{-1} y = \varepsilon_k^{-1} Z(y) - \varepsilon_k^{-1} Z A_{I_k} G^{-1} A_{I_k}^* Z(y) = \varepsilon_k^{-1} Z(y - A_{I_k} G^{-1} A_{I_k}^* Z(y)) \tag{2.18}$$

and hence

$$\begin{aligned} A_{I_k}^*(z) &= \varepsilon_k^{-1}(A_{I_k}^* Z(y) - A_{I_k}^* Z A_{I_k} G^{-1} A_{I_k}^* Z(y)) \\ &= \varepsilon_k^{-1}(A_{I_k}^* Z(y) - A_{I_k}^* Z A_{I_k} \left(\varepsilon_k C^{-1} + A_{I_k}^* Z A_{I_k}\right)^{-1} A_{I_k}^* Z(y)) \\ &= \varepsilon_k^{-1} \left(A_{I_k}^* Z(y) - \left(A_{I_k}^* Z A_{I_k} \pm \varepsilon_k C^{-1}\right) \left(\varepsilon_k C^{-1} + A_{I_k}^* Z A_{I_k}\right)^{-1} A_{I_k}^* Z(y)\right) \\ &= C^{-1} \left(\varepsilon_k C^{-1} + A_{I_k}^* Z A_{I_k}\right)^{-1} A_{I_k}^* Z(y). \end{aligned} \tag{2.19}$$

Thus, for the $(k+1)$-th iteration of IRLS, we obtain for the solution x_*^{k+1} of (2.9) the representation

$$\begin{aligned} x_*^{k+1} &= W_k^{-1} A^*(z) = \left(\varepsilon_k \mathrm{Id} + Q_{I_k} \left(D_{I_k}^{-1} - \varepsilon_k \mathrm{Id}\right) Q_{I_k}^*\right) A^*(z) \\ &= \left[\varepsilon_k \mathrm{Id} + Q_{I_k} C Q_{I_k}^*\right] A^* z = \varepsilon_k A^* z + Q_{I_k} C A_{I_k}^*(z) \\ &= \varepsilon_k A^* z + Q_{I_k} C C^{-1} \left(\varepsilon_k C^{-1} + A_{I_k}^* Z A_{I_k}\right)^{-1} A_{I_k}^* Z(y) \\ &= \varepsilon_k A^* z + Q_{I_k} G^{-1} A_{I_k}^* Z(y) \end{aligned} \tag{2.20}$$

using (2.19) in the third equality. Next, if V is as in the input of Algorithm 2, since

$$A_{I_k}^* Z A_{I_k} = Q_{I_k}^* A^*(AA^*)^{-1} A Q_{I_k} = Q_{I_k}^* V V^* Q_{I_k} = (V^*)_{I_k}^* (V^*)_{I_k},$$

we observe that the matrix G from above actually coincides with the linear system matrix of (2.17).

Furthermore, if γ_k is as in step 3 of Algorithm 2, it satisfies

$$\gamma_k = \gamma_k^{(0)} + \Delta\gamma_k = \gamma_k^{(0)} + G^{-1}\mathbf{h}_k^0 = \gamma_k^{(0)} + G^{-1}\left(Q_{I_k}^* \widetilde{y} - G\gamma_k^{(0)}\right) = G^{-1} Q_{I_k}^* \widetilde{y},$$

where we also observe that

$$Q_{I_k}^* \widetilde{y} = Q_{I_k}^* V \Sigma_A^{-1} U^*(y) = Q_{I_k}^* A^* \left(AA^*\right)^{-1} y = A_{I_k}^* Z y.$$

Using this last equation, we can identify z of (2.18) such that

$$z = \varepsilon_k^{-1} Z(y - A_{I_k} G^{-1} A_{I_k}^* Z y) = \varepsilon_k^{-1} Z\left(y - A_{I_k} G^{-1} Q_{I_k}^* \widetilde{y}\right) = \varepsilon_k^{-1} Z\left(y - A_{I_k}\gamma_k\right).$$

Inserting this into (2.20), we obtain

$$\begin{aligned}
x_*^{k+1} &= \varepsilon_k A^* z + Q_{I_k} G^{-1} A_{I_k}^* Z y = \varepsilon_k A^* z + Q_{I_k} G^{-1} Q_{I_k}^* \widetilde{y} = \varepsilon_k A^* z + Q_{I_k}\gamma_k \\
&= A^* Z\left(y - A_{I_k}\gamma_k\right) + Q_{I_k}\gamma_k \\
&= V\Sigma_A^{-1} U^*\left(y - U\Sigma_A V^* Q_{I_k}\gamma_k\right) + Q_{I_k}\gamma_k \\
&= \widetilde{y} - V(V^*)_{I_k}\gamma_k + Q_{I_k}\gamma_k \\
&= r_{k+1} + Q_{I_k}\gamma_k,
\end{aligned}$$

where the residual r_{k+1} is as in step 4 of Equation 2. The last equation tells us that the solution of the linear system can be represented by a residual. The first, $Z(y)$, can be precomputed from the data, and the second, $ZAP_{I_k}G^{-1}P_{I_k}^* A^* Z(y)$, is updated iteratively according to the weights.

Comparing the equation $x_*^{k+1} = r_{k+1} + Q_{I_k}\gamma_k$ with the steps 5 and 6 of Algorithm 2, we observe that the output x^{k+1} of Algorithm 2 coincides with x_*^{k+1}, which finishes the proof. □

If a linear system is solved inexactly with an iterative solver (and a limited number of iterations), a sufficient condition for quantifying its accuracy can be achieved by bounding the condition number of the system matrix (see, e.g., [NW06, Section 5.1]). In particular, it can be shown that the convergence rate of the conjugate gradient is given by

Theorem 2.2.9 ([QSS10, Theorem 4.12]). *Let the matrix A be Hermitian and positive definite. The conjugate gradient algorithm converges to the solution of the system $Ax = y$ after at most N steps. Moreover, the error $x^i - x$ is such that*

$$\|A^{\frac{1}{2}}(x^i - x)\|_2 \leq \frac{2c_A^i}{1 + c_A^{2i}}\|A^{\frac{1}{2}}(x^0 - x)\|_2, \quad \text{with } c_A = \frac{\sqrt{\kappa_A} - 1}{\sqrt{\kappa_A} + 1} < 1,$$

where $\kappa_A = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$ is the condition number of the matrix A and $\sigma_{\max}(A)$ (resp. $\sigma_{\min}(A)$) is the largest (resp. smallest) singular value of A.

In fact, using a few iterations of a CG method to solve the system (2.17) typically leads to quite accurate solutions, particularly if ε_k is small. This can be seen by analyzing the condition number of the matrix G

$$G = \varepsilon_k \left(D_{I_k}^{-1} - \varepsilon_k \mathrm{Id}_{I_k}\right)^{-1} + A_{I_k}^* Z A_{I_k} =: M_{1,k} + M_{2,k} \tag{2.21}$$

of Equation 2.17 in Algorithm 2. To do so, we assume that the matrix A_{I_k} is well-conditioned in the sense that it has the restricted isometry property.

Definition 2.2.10. *A matrix $A \in \mathbb{R}^{m\times N}$ satisfies the restricted isometry property (RIP) of order $1 \leq s \leq N$ if there is a constant $\delta_s \in (0,1)$ such that*

$$(1 - \delta_s)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_s)\|x\|_2^2$$

for all s-sparse vectors $x \in \mathbb{R}^N$.

We start by bounding the condition number of $M_{2,k}$. For that, let x such that $\|x\|_2 = 1$. We calculate that

$$\|x^T A_{I_k}^* (AA^*)^{-1} A_{I_k} x\| \leq \|A_{I_k} x\|^2 \|(AA^*)^{-1}\| \leq (1 + \delta)^2 \|(AA^*)^{-1}\|,$$

where in the last inequality, we assumed that A satisfies the restricted isometry property. From that, as long as the cardinality of the set $I_k := \{i \in [N] : |x_i^k| > \varepsilon_k\}$ is smaller than s, the inequality holds. This implies that

$$\|M_{2,k}\| = \|A_k^*(AA^*)^{-1}A_k\| \leq (1 + \delta)^2 \sigma_{\min}(A)^{-2}. \tag{2.22}$$

Similarly, we can derive that

$$\sigma_{\min}\left(A_k^*(AA^*)^{-1}A_k\right) \geq (1 - \delta)^2 \|A\|^{-2}. \tag{2.23}$$

We note that $M_{1,k}$ is a diagonal matrix with entries given by $\varepsilon_k/(x_i^k - \varepsilon_k)$. Now, by additionally assuming that the iteration x_k is already close to the ground truth and, without loss of generality, by assuming that k is such that $||x^k - x_*||_\infty \leq c \min_{i \in S}(x_*)_i$, for $c > 0$, we have

$$||M_{1,k}||_2 = \varepsilon_k/(x_i^k - \varepsilon_k) \leq \frac{\varepsilon_k}{(1-c)\min_{i\in S}|(x_*)_i| - \varepsilon_k}. \tag{2.24}$$

Note that this term becomes arbitrarily small when ε_k becomes arbitrarily small. Hence, for ε_k small enough it follows that

$$\kappa(G) = \frac{\|G\|_2}{\sigma_{\min}(G)} \overset{(a)}{\leq} \frac{\|M_{1,k}\|_2 + \|M_{2,k}\|_2}{\sigma_{\min}(M_{2,k}) - \|M_{1,k}\|_2} \overset{(b)}{\leq} \frac{2(1+\delta)^2\|A\|_2^2}{(1-\delta)^2\sigma_{\min}(A)^2} = \frac{2(1+\delta)^2}{(1-\delta)^2}\kappa(A)^2.$$

Here, in (a), we used Weyl's inequality [HHJ94] and (b) holds as soon as ε_k is small enough due to inequalities (2.22), (2.23), and (2.24). Hence, we have shown that if A is well-conditioned, the matrix G will also be well-conditioned, and the CG method will yield very accurate solutions. This fact, which forms the computational basis of our IRLS implementation, solves the problem of ill-conditioned systems that appear when such methods are used. The method of choice for the solution of the linear system in the inner iteration will dictate how scalable the method is. Note that in this thesis, we have not explored modern randomized techniques for such tasks, and we leave this remark for future work. A first step in this direction was given in [Mel21], but we believe that its full potential has not yet been explored. See also the discussion in [DY23].

Open Problem: How can randomized methods, e.g., [GR15], when applied to the inner iteration of IRLS, improve the algorithm's scalability? What are the limitations of such an approach?

Now that we have discussed the computational side of IRLS, we turn to the development of the convergence theory, which is our main result in this chapter. To do so, we start by introducing a central property in compressive sensing and sparse signal recovery, the *null space property.*

2.2.2 Null space property

As mentioned in Section 2.1.1, several sufficient conditions exist for the sparse recovery via ℓ_1-minimization. The remarkable fact about this problem is that we also know a precise characterization of a *necessary condition*, at least in the noiseless case. In fact, the seminal work [CDD09] defined the null space property (NSP), a condition on the kernel of measurement matrix A, that is necessary and sufficient for sparse recovery via Basis Pursuit. This is a very rare example of a provably hard combinatorial problem in optimization where there exists a property that dictates when the solution to this problem can be found via solving a convex problem. They also defined a more general stable version of the NSP that accounts for compressible vectors, i.e., not exactly but rather approximately sparse. This is precisely the content of the next definition.

Definition 2.2.11. *A matrix $A \in \mathbb{R}^{m\times N}$ is said to satisfy the* ℓ_1-null space property (ℓ_1-NSP) *of order $s \in N$ if for any set $S \subset [N]$ of cardinality $|S| \leq s$, it holds that $\|v_S\|_1 < \|v_{S^c}\|_1$, for all $v \in \ker(A)\backslash\{0\}$. We say that it satisfies the* stable null space property *if there exists a constant $0 < \rho_s < 1$ such that $\|v_S\|_1 \leq \rho_s \|v_{S^c}\|_1$, for all $v \in \ker(A)\backslash\{0\}$.*

The precise theorem can be found in [FR13, Theorem 4.5].

Theorem 2.2.12. (Essentially [CDD09, Theorem 3.2]): *Given a matrix $A \in \mathbb{R}^{m\times N}$, every s-sparse vector $x \in \mathbb{R}^N$ is the unique solution of (P_1) with $Ax = y$ if and only if A satisfies the null space of order s.*

This property is not a vacuous one since many examples of matrices satisfy it with high probability, as the next result shows.

Theorem 2.2.13. *[FR13, Theorem 9.29] Let $A \in \mathbb{R}^{m\times N}$ be a random drawing of a Gaussian matrix. Assume that*

$$\frac{m^2}{m+1} \geq 2s\ln(eN/s)\left(1 + \rho^{-1} + D(s/N) + \sqrt{\frac{\ln(\varepsilon^{-1})}{s\ln(eN/s)}}\right)^2,$$

where D is a function that satisfies $D(\alpha) \leq 0.92$ for all $\alpha \in (0, 1]$ and $\lim_{\alpha\to 0} D(\alpha) = 0$. Then, with probability at least $1 - \varepsilon$ the matrix A satisfies the stable null space property of order s with constant ρ.

The proof of Theorem 2.2.13, which appeared for the first time in [FR13, Theorem 9.29] relies on *Gordon's Escape Through the Mesh Theorem* [Gor88] and was highly inspired by [RV08a, Theorem 4.1] and the discussion and propositions presented in [Sto10]. Later, the book [FR13] introduced a robustness constant to account for robustness with respect to additive noise. See [FR13, Chapter 4] or [PJ22] for an overview. Furthermore, the ℓ_1-NSP is known to hold for random matrices with i.i.d. entries whose distribution has a logarithmic number of finite moments and fulfills a small-ball condition, which includes a number of more heavy-tailed random matrices as long as the number of rows and columns satisfy a certain relationship [ML17, DLR18].

The ℓ_1-NSP is implied by the restricted isometry property despite a fundamental theoretical difference between them (see, e.g., [CCW16, DLR16] for a discussion), which is fulfilled by a large class of random matrices with high probability. For example, this includes matrices with (sub-)Gaussian entries and random partial Fourier matrices [RV08b, BDDW08].

As in [DDFG10], the analysis we present in this chapter is based on the assumption that the measurement matrix A satisfies the NSP since this is the weakest possible assumption we could require for the analysis of sparse recovery. Before we state the main result, Theorem 2.3.3, we discuss some existing theoretical results for IRLS.

2.2.3 Existing theory

A major step forward in the theoretical understanding of IRLS was achieved in the seminal paper [DDFG10], where the authors showed that a variant of IRLS for (P_1) converges globally to the ℓ_1-minimizer if the measurement operator A fulfills the NSP of sufficient order, which essentially ensures that an ℓ_1-minimizer is actually sparse. However, since their proof relies on a compactness argument, their proof is non-constructive and does not reveal any rate for *global* convergence. Furthermore, the analysis of [DDFG10] provides a *locally* linear convergence rate, but this local linear rate has the drawback that it only applies *if the support of the true signal has been discovered*, which was not emphasized in their contribution and which is arguably the difficult part of ℓ_0-minimization—cf. Theorem 2.3.1 below and Section 2.4.1. Later, [Bec15a] established a nonasymptotic sublinear rate of convergence for the LASSO as a particular case of a general result for alternating minimization schemes. However, the result does not apply to an IRLS scheme with a smoothing parameter that changes at each iteration, as is the case here. We will discuss this general result [Bec15a] in Chapter 4.

A predecessor of IRLS for the sparse recovery problem (P_1), and more generally, for ℓ_p-quasinorm minimization with $0 < p \leq 1$, is the *FOCal Underdetermined System Solver* (FOCUSS) proposed by Gorodnitsky, Rao and Kreutz-Delgado [GR97a, RKD99]. Asymptotic convergence of FOCUSS to a stationary point from any initialization was claimed in [RKD99], but the proof was not correct, as pointed out by [CY16]. One limitation of FOCUSS is that unlike in IRLS as presented in Algorithm 1, no smoothing parameter ε is used, which leads to ill-conditioned linear systems and to solutions that could be non-sparse in the case of $p < 1$.

To mitigate this, [CY08] proposed an IRLS method that uses smoothing parameters ε (such as used in Q_ε defined above) that are updated iteratively. It was observed that this leads to a better condition number for the linear systems to be solved in each step of

IRLS and, furthermore, that this smoothing strategy has the advantage of finding sparser vectors if the weights of IRLS are chosen to minimize a non-convex ℓ_p-quasinorm for $p < 1$.

Further progress for IRLS designed to minimize an ℓ_1-norm was achieved in the seminal paper [DDFG10]. In [DDFG10], it was shown that if the measurement operator fulfills the ℓ_1-null space property as in Definition 2.2.11, an IRLS method with iteratively updated smoothing converges to the ℓ_1-minimizer, coinciding with the s-sparse solution, if there exists one that is compatible with the measurements. This method uses not exactly the update rule of Equation 2.10, but rather updates the smoothing parameter such that $\varepsilon_{k+1} = \min(\varepsilon_k, R(x^{k+1})_{s+1}/N)$, where $R(x^{k+1})_{s+1}$ is the $(s+1)^{\text{st}}$-largest element of the set $\{|x_j^{k+1}|, j \in [N]\}$. Furthermore, a *local linear convergence rate* of IRLS was established [DDFG10, Theorem 6.1] under the same conditions.

However, the analysis of [DDFG10] has its limitations: first, there is a gap in the assumption of their convergence results between the sparsity s of a vector to be recovered and the order $\widehat{s}$ of the NSP of the measurement operator. Recently, this gap was circumvented in [ABH19] with an IRLS algorithm that uses a smoothing update rule based on an ℓ_1-norm, namely, $\varepsilon_{k+1} = \min(\varepsilon_k, \eta(1-\rho_s)\sigma_s(x^{k+1})_{\ell_1}/N)$, where $\eta \in (0,1)$, and ρ_s is the NSP constant of the order s of the NSP fulfilled by the measurement matrix A – this rule is quite similar to the rule Equation 2.10 that we use in Algorithm 1. In particular, [ABH19, Theorem III.6] establishes convergence with a local linear rate similar to [DDFG10] but without the gap mentioned above. The main limitation, however, of the theory of [DDFG10], which is also shared by more recent paper [ABH19], is that the linear convergence rate only holds *locally*, i.e., in a situation where the support of the sparse vector has already been identified, see also Section 2.3 and Section 2.4.1 for a discussion.

We finally mention three relevant papers for the theoretical understanding of IRLS. The work [BBPB13] established a correspondence between the IRLS algorithm and a class of Expectation-Maximization (EM) algorithms for constrained maximum likelihood estimation under a Gaussian scale mixture distribution. Without requiring any connection between sparse recovery and ℓ_1-minimization, [EV19] shows that an IRLS-like algorithm for Equation P_1, requires $O(N^{1/3}\log(\log(N)) + N^{1/3}\log(1/\delta)/\delta^{2/3} + \log(N)/\delta^2)$ iterations to obtain an multiplicative error of $1+\delta$ on the minimizer $||x||_1$. Unlike the result established in this chapter, Theorem 2.3.3, this corresponds not to a linear but to a sublinear convergence rate. Finally, the recent [SV21] explored the curious relationship of IRLS for ℓ_1-minimization and slime mold dynamics, interpreting both as an instance of the same meta-algorithm. We expect that this kind of insight could be explored in the future for new theoretical guarantees as well as for the design of new algorithms for data science and machine learning problems.

2.3 IRLS for Basis Pursuit with Global Linear Rate

The main theoretical result for (a modern implementation of) IRLS for the sparse recovery problem was established in the work [DDFG10]. In particular, they established the following local convergence-rate theorem[4]:

Proposition 2.3.1. *[DDFG10, Theorem 6.1] Assume that $A \in \mathbb{R}^{m\times N}$ satisfies the NSP of order $\widehat{s} > s$ with constant $\rho_{\widehat{s}}$ such that $0 < \rho_{\widehat{s}} < 1 - \frac{2}{\widehat{s}+2}$ and $\hat{s} > s + \frac{2\rho_{\widehat{s}}}{1-\rho_{\widehat{s}}}$ hold. Let $x_* \in \mathbb{R}^N$ be an s-sparse vector and set $y = Ax_*$. Assume that there exists an integer $k_0 \geq 1$ and a positive number $\xi > 0$ such that*

$$\xi := \frac{\|x^{k_0} - x_*\|_1}{\min_{i\in S} |(x_*)_i|} < 1. \tag{2.25}$$

Then the iterates $\{x^{k_0}, x^{k_0+1}, x^{k_0+2}, \ldots\}$ of the IRLS method in [DDFG10] converge linearly *to x_*, i.e., for all $k \geq k_0$, the kth iteration of IRLS satisfies*

$$\|x^{k+1} - x_*\|_1 \leq \frac{\rho_{\widehat{s}}(1+\rho_{\widehat{s}})}{1-\xi}\left(1 + \frac{1}{\widehat{s}-1-s}\right)\|x^k - x_*\|_1. \tag{2.26}$$

A closer look at the locality condition (2.25) reveals that its *basin of attraction* is very restrictive: this condition means that the *support identification* problem underlying the sparse recovery *has already been solved*, i.e., if x^k, the kth IRLS iteration, is already close enough to the ground truth, then we would observe that the support would have been already identified and, consequently, the hardest part of the sparse recovery problem would have been solved. Only the identification of the correct magnitudes would be missing, which can be done via least squares applied to an overdetermined system. Since this was not discussed in [DDFG10], we establish this fact in the following proposition as part of our contribution.

Proposition 2.3.2. *Let $x^k, x_* \in \mathbb{R}^N$, let $S \subset [N]$ be the support set of x_* of size $|S| = s$. If Equation 2.25 holds, i.e., if $\|x^k - x_*\|_1 < \min_{i\in S}|(x_*)_i|$, then the set $S_k \subset [N]$ of the s largest coordinates of x^k coincides with S.*

Proof. Let $j \in S^c$, where S is the support set of x_*. Then

$$|x_j^k| \leq \sum_{i\in S^c} |x_i^k| < \min_{i\in S} |(x_*)_i| - \sum_{i\in S} |x_i^k - (x_*)_i|,$$

using the assumption $\sum_{i\in S^c} |(x^k)_i| + \sum_{i\in S} |x_i^k - (x_*)_i| = \|x^k - x_*\|_1 < \min_{i\in S}|(x_*)_i|$.

[4] In [DDFG10, Section 7.3], the authors also established a local superlinear convergence rate with rate $2-p$ for the case of ℓ_p-minimization, where $0 < p < 1$

On the other hand, for $j \in S$, we can estimate that

$$|x_j^k| = |x_j^k - (x_*)_j + (x_*)_j| \geq |(x_*)_j| - |x_j^k - (x_*)_j| \geq \min_{i \in S} |(x_*)_i| - \sum_{i \in S} |x_i^k - (x_*)_i|.$$

Taking the previous two inequalities together, we conclude that

$$\max_{j \in S^c} |x_j^k| < \min_{j \in S} |x_j^k|,$$

which finishes the proof. □

Now, we will overcome the local assumption such as the one presented in Equation 2.25 and solve a problem stated in [SV21]. We show that IRLS as defined in this chapter by Algorithm 1 *exhibits a global linear convergence rate*, i.e., there is a linear convergence rate starting from any initialization, as early as in the first iteration. This shows that a very simple strategy for sparse recovery, based on solving a least squares problem multiple times, can be fruitful and have strong theoretical guarantees when the weighting and smoothing parameters are suitably chosen. Moreover, our proof strategy is quite simple and involves bounding two terms together with a solution of a quadratic inequality.
Our first main result, Theorem 2.3.3, deals with the scenario that the ground truth vector x_* is exactly s-sparse. Our second result, Theorem 2.3.9, generalizes the first one to the more realistic situation of approximately sparse vectors.

Theorem 2.3.3. *Consider the problem of recovering an unknown s-sparse vector $x_* \in \mathbb{R}^N$ from known measurements of the form $y = Ax_*$. Assume that the measurement matrix $A \in \mathbb{R}^{m \times N}$ fulfills the ℓ_1-NSP of order s with constant $\rho_s < 1/2$. Let the IRLS iterates $\left\{x^k\right\}_k$ and $\{\varepsilon_k\}_k$ be defined by the IRLS algorithm* (2.9) *and* (2.10) *with initialization x^0. Then, for all $k \in \mathbb{N}$, it holds that*

$$\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 \leq \left(1 - \frac{c}{\rho_1 N}\right)^k \left(\mathcal{J}_{\varepsilon_0}(x^0) - \|x_*\|_1\right) \tag{2.27}$$

as well as

$$\|x^k - x_*\|_1 \leq 9 \left(1 - \frac{c}{\rho_1 N}\right)^k \|x^0 - x_*\|_1. \tag{2.28}$$

Here $c = 1/768$ is an absolute constant and $\rho_1 < 1/2$ denotes the ℓ_1-NSP constant of order 1.

Inequality 2.27 says that the difference $\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1$ converges linearly with a uniform upper bound of $1 - \frac{c}{\rho_1 N}$ on the linear convergence factor. As our proof shows, this implies inequality 2.28, which implies that also $\|x_* - x^k\|_1$ exhibits linear convergence in the number of iterations k. In particular, this means that for some error tolerance $\delta > 0$, we obtain $\|x_* - x^k\|_1 \leq \delta$ after $O\left(\rho_1 N \log\left(\frac{\|x_*-x^0\|_1}{\delta}\right)\right)$ iterations.

Remark 2.3.4. *Note that it follows directly from Definition 4.1.4 that the constant ρ_1 of the ℓ_1-NSP of order 1 satisfies $\rho_1 \leq \rho_s \leq 1$, which implies that δ-accuracy is obtained after $O\left(N \log\left(\frac{\|x_*-x^0\|_1}{\delta}\right)\right)$ iterations. This bound can be improved in many scenarios where one can obtain more explicit bounds on ρ_1, for example, when A is a Gaussian matrix. Namely, inspecting [FR13, p. 142 and Thm. 9.2], we observe in this scenario that $\rho_1 \lesssim \sqrt{(\log N)/m}$ with high probability. Hence, in this scenario, at most $O\left(N\sqrt{\frac{\log N}{m}} \log\left(\frac{\|x_*-x^0\|_1}{\delta}\right)\right)$ iterations are needed to achieve δ-accuracy.*

The key idea in our proof is to use the fact that the quadratic functional $Q_{\varepsilon_k}(\cdot, x^k)$ approximates the ℓ_1-norm in a neighborhood of the current iterate x^k. For this reason, we also expect that for $t > 0$ sufficiently small, we have that $Q_{\varepsilon_k}(x^k + tv^k, x^k) < Q_{\varepsilon_k}(x^k, x^k)$ if $v^k = x_* - x^k$ is the vector between x^k and the ground truth x_*. Then, by choosing t properly, we can guarantee a sufficient decrease of the functional $\mathcal{J}_{\varepsilon^k}\left(x^k\right)$ in each iteration.

Before we start with the proof of Proposition 2.3.7, we first state and prove the following technical lemma, which gives an upper and lower bound for the quantity for which we are going to show linear convergence, namely, $\mathcal{J}_{\varepsilon}(x) - \|x_*\|_1$.

Lemma 2.3.5. *Let $x_*, x \in \mathbb{R}^N$. Assume that A fulfills the ℓ_1-NSP of order s with constant $\rho_s < 1$. Furthermore, suppose $Ax_* = Ax$ and that $\varepsilon \leq \frac{1}{N}\sigma_s(x)_{\ell_1}$. Then, it holds that*

$$\frac{1-\rho_s}{1+\rho_s}\|x - x_*\|_1 - 2\sigma_s(x_*)_{\ell_1} \leq \mathcal{J}_{\varepsilon}(x) - \|x_*\|_1 \leq 3\sigma_s(x)_{\ell_1}. \tag{2.29}$$

To prove Lemma 2.3.5, we need the following technical lemma.

Lemma 2.3.6. *[DDFG10, Lemma 4.3] Assume that the matrix $A \in \mathbb{R}^{m\times N}$ has the ℓ_1-NSP holds for some s and $\rho_s < 1$. Then for all $z, x_* \in \mathbb{R}^N$ such that $Az = Ax_*$ it holds that*

$$\|z - x_*\|_1 \leqslant \frac{1+\rho_s}{1-\rho_s}\left(\|x_*\|_1 - \|z\|_1 + 2\sigma_s(z)_{\ell_1}\right).$$

Proof of Lemma 2.3.5. We observe that $\mathcal{J}_{\varepsilon}(x) \geq \|x\|_1$, which follows directly from the definition of $\mathcal{J}_\varepsilon(x)$, see Equation 2.4. Hence, we obtain that

$$\begin{aligned}\mathcal{J}_\varepsilon(x) - \|x_*\|_1 &\geq \|x\|_1 - \|x_*\|_1 \\ &= \|x_{S^c}\|_1 + \|x_S\|_1 - \|x_*\|_1 \\ &\geq \|x_{S^c}\|_1 - \| (x - x_*)_S \|_1 - \| (x_*)_{S^c} \|_1 \\ &\geq \| (x - x_*)_{S^c} \|_1 - \| (x - x_*)_S \|_1 - 2\| (x_*)_{S^c} \|_1,\end{aligned}$$

where in each of the last two inequalities, we have applied the reverse triangle inequality. Since $x - x_*$ is contained in the null space of A, it follows from the nullspace property that $\| (x - x_*)_S \|_1 \leq \rho_s \| (x - x_*)_{S^c} \|_1$. Hence, we have shown that

$$\mathcal{J}_\varepsilon(x) - \|x_*\|_1 \geq (1 - \rho_s) \| (x - x_*)_{S^c} \|_1 - 2\| (x_*)_{S^c} \|_1.$$

Since it follows from the null space property that $\| (x - x_*)_{S^c} \|_1 \geq \frac{\|x - x_*\|_1}{1+\rho_s}$, this shows the first inequality in 2.29.

Next, we will prove the reverse inequality in 2.29. For that, set $I := \{i \in [N] : |x_i^k| > \varepsilon_k\}$ and denote by S the set, which contains the s largest entries of x in absolute value. Then we observe that

$$\begin{aligned}\mathcal{J}_\varepsilon(x) - \|x_*\|_1 &= \|x_I\|_1 + \frac{1}{2}\sum_{i \in I^c} \left(\frac{x_i^2}{\varepsilon} + \varepsilon \right) - \|x_*\|_1 \\ &\leq \|x_I\|_1 + |I^c|\varepsilon - \|x_*\|_1 \qquad (2.30) \\ &\leq \|x_I\|_1 + \sigma_s(x)_{\ell_1} - \|x_*\|_1 \\ &\leq \|x\|_1 + \sigma_s(x)_{\ell_1} - \|x_*\|_1.\end{aligned}$$

In the third line we used the assumption $\varepsilon \leq \frac{1}{N}\sigma_s(x)_{\ell_1}$. To proceed, we first derive an appropriate upper bound for $\|x\|_1 - \|x_*\|_1$. For that, we note

$$\begin{aligned}\left(\frac{1-\rho_s}{1+\rho_s} + 1 \right) (\|x\|_1 - \|x_*\|_1) &\leq \frac{1-\rho_s}{1+\rho_s}\|x - x_*\|_1 - (\|x_*\|_1 - \|x\|_1) \\ &\leq \left(\|x_*\|_1 - \|x\|_1 + 2\sigma_s(x)_{\ell_1}\right) - (\|x_*\|_1 - \|x\|_1) \\ &\leq 2\sigma_s(x)_{\ell_1},\end{aligned}$$

where in the second line, we have used Lemma 2.3.6. This shows that $\|x\|_1 - \|x_*\|_1 \leq \frac{2\sigma_s(x)_{\ell_1}}{1+\frac{1-\rho_s}{1+\rho_s}}$. Combining this with Equation 2.30, we obtain

$$\mathcal{J}_\varepsilon(x) - \|x_*\|_1 \leq 3\sigma_s(x)_{\ell_1},$$

which finishes the proof of inequality 2.29. □

The next key proposition states that the quantity $\mathcal{J}_{\varepsilon^k}\left(x^k\right) - \|x_*\|_1$ decays linearly under appropriate conditions. But before that, we define the ℓ_1-error of the best s-term approximation as $\sigma_s(x_*)_{\ell_1} = \inf\{\|x_* - z\|_1 : \ z \in \mathbb{R}^N \text{ is } s\text{-sparse}\}$.

Proposition 2.3.7. *Let $x_* \in \mathbb{R}^N$ be an approximately s-sparse vector with support S. Let $A \in \mathbb{R}^{m\times N}$ and $y = Ax_*$. Assume that A fulfills the ℓ_1-NSP of order s with constant $\rho_s < 3/4$, if $\sigma_s\left(x_*\right)_{\ell_1} = 0$, and $\rho_s < 1/4$ otherwise. Moreover, assume that A has the ℓ_1-NSP of order 1 with constant $\rho_1 < 1$.*
Let the IRLS iterates $\left\{x^k\right\}_k$ and $\{\varepsilon_k\}_k$ be defined by Equation 2.9 and Equation 2.10 with initialization x^0. Then, for all $k \in \mathbb{N}$, such that $\|\left(x_*\right)_{S^c}\|_1 \leq \frac{2}{9}\|\left(x_*\right)_{S^c} - x^\ell_{S^c}\|_1$ for all $\ell < k$, the following holds

$$\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 \leq \left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^k \left(\mathcal{J}_{\varepsilon_0}(x^0) - \|x_*\|_1\right). \tag{2.31}$$

where the constant c_{ρ_s} is defined by

$$c_{\rho_s} := \begin{cases} \frac{(3/4-\rho_s)^2}{48} & \text{if } \sigma_s\left(x_*\right)_{\ell_1} = 0 \\ \frac{(1/4-\rho_s)^2}{48} & \text{else} \end{cases}$$

Before proving this statement, let us describe the main ideas of our proof. Recall that x_* has minimal $\|\cdot\|_1$-norm among all vectors x that satisfy the constraint $Ax = y$, since we have assumed that the NSP holds. Hence, setting $v^k = x_* - x^k$ due to convexity of the ℓ_1-norm we that $\|x^k + tv^k\|_1 < \|x^k\|_1$ for all $0 < t < 1$. Since that the quadratic functional $Q(\cdot, x^k)$ approximates the objective function $\mathcal{J}_\varepsilon$, which is a surrogate for the ℓ_1-norm, in a neighborhood of the current iterate x^k, we also expect that for $t > 0$ sufficiently small we have that $Q(x^k + tv^k, x^k) < Q(x^k, x^k)$. To show that the decrease is sufficiently large, we also need to show that t can be chosen large enough. This will guarantee a sufficient decrease of $\mathcal{J}_{\varepsilon^k}\left(x^k\right)$ in each iteration.

Proof of Theorem 2.3.7. In order to show inequality Equation 2.31 we will prove by induction that for each k, such that $\|\left(x_*\right)_{S^c}\|_1 \leq \frac{2}{9}\|\left(x_*\right)_{S^c} - x^\ell_{S^c}\|_1$ for all $\ell \leq k$, it holds that

$$0 \leq \mathcal{J}_{\varepsilon_{k+1}}(x^{k+1}) - \|x_*\|_1 \leq \left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right) \left(\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1\right).$$

Now choose such a $k \geq 1$ and assume that the statement has been shown for all $k' < k$.

Set $v^k = x_* - x^k$. For $t \in \mathbb{R}$, we have, by the optimality of x^{k+1} in Equation2.9, that

$$\mathcal{J}_{\varepsilon_{k+1}}(x^{k+1}) \leq Q_{\varepsilon_k}(x^{k+1}, x^k) \leq Q_{\varepsilon_k}(x^k + tv^k, x^k). \tag{2.32}$$

Moreover, by the definition of the quadratic objective $Q_{\varepsilon_k}(\cdot, x^k)$ (see Equation 4.14), it holds that

$$Q_{\varepsilon_k}(x^k + tv^k, x^k) - \mathcal{J}_{\varepsilon_k}(x^k) = t \langle \nabla \mathcal{J}_{\varepsilon_k}(x^k), v^k \rangle + \frac{t^2}{2} \langle v^k, \operatorname{diag}(w(x^k, \varepsilon_k)) v^k \rangle. \tag{2.33}$$

Our goal is to show that by picking t large enough, we can make $Q_{\varepsilon_k}(x^k + tv^k, x^k) - \mathcal{J}_{\varepsilon_k}(x^k) < 0$ sufficiently small. For that, we now control the terms $\langle \nabla \mathcal{J}_{\varepsilon_k}(x^k), v^k \rangle$ and $\langle v^k, \operatorname{diag}(w_{\varepsilon_k}(x^k)) v^k \rangle$ separately.

Part I: Bounding the linear term $\langle \nabla \mathcal{J}_{\varepsilon_k}(x^k), v^k \rangle$:
Let $I := \{i \in [N] : |x_i^k| > \varepsilon_k\}$ and denote by S the set which contains the s largest entries of x_* in absolute value. In the case that x_* is sparse, S is given by the support of x_*, i.e. $S = \operatorname{supp}(x_*)$. Consider

$$\langle \nabla \mathcal{J}_{\varepsilon_k}(x^k), v^k \rangle = \sum_{i=1}^{N} \frac{x_i^k}{\max(|x_i^k|, \varepsilon_k)} v_i^k = \sum_{i \in S} \frac{x_i^k}{\max(|x_i^k|, \varepsilon_k)} v_i^k + \sum_{i \in S^c} \frac{x_i^k}{\max(|x_i^k|, \varepsilon_k)} v_i^k.$$

The first summand can be bounded by

$$\begin{aligned}
\sum_{i \in S} \frac{x_i^k}{\max(|x_i^k|, \varepsilon_k)} v_i^k &= \sum_{i \in S \cap I} \operatorname{sgn}(x_i^{(k)}) v_i^k + \sum_{i \in S \cap I^c} \frac{x_i^k}{\varepsilon_k} v_i^k \\
&\leq \|v_{S \cap I}^k\|_1 + \|v_{S \cap I^c}^k\|_1 \\
&= \|v_S^k\|_1 \\
&\leq \rho_s \|v_{S^c}^k\|_1.
\end{aligned}$$

Note that this is precisely the first place where the definition of ε_k proposed in Algorithm

1 played a role in the proof of our main result. For the second summand, we have that

$$\begin{aligned}
&\sum_{i\in S^c} \frac{x_i^k}{\max(|x_i^k|,\varepsilon_k)} v_i^k \\
=& \sum_{i\in S^c\cap I} \operatorname{sgn}\left(x_i^k\right) v_i^k + \sum_{i\in S^c\cap I^c} \frac{x_i^k v_i^k}{\varepsilon_k} \\
=& \sum_{i\in S^c\cap I} \operatorname{sgn}\left(x_i^k\right) (x_*)_i - \sum_{i\in S^c\cap I} \operatorname{sgn}\left(x_i^k\right) x_i^k + \sum_{i\in S^c\cap I^c} \frac{x_i^k\,(x_*)_i}{\varepsilon_k} - \sum_{i\in S^c\cap I^c} \frac{(x_i^k)^2}{\varepsilon_k} \\
\leq& \|\,(x_*)_{S^c\cap I}\,\|_1 - \|x^k_{S^c\cap I}\|_1 + \|\,(x_*)_{S^c\cap I^c}\,\|_1 - \frac{\|x^k_{S^c\cap I^c}\|_2^2}{\varepsilon_k} \\
=& -\|x^k_{S^c\cap I}\|_1 + \|\,(x_*)_{S^c}\,\|_1 - \frac{\|x^k_{S^c\cap I^c}\|_2^2}{\varepsilon_k} \\
=& \|\,(x_*)_{S^c}\,\|_1 - \|x^k_{S^c}\|_1 + \|x^k_{S^c\cap I^c}\|_1 - \frac{\|x^k_{S^c\cap I^c}\|_2^2}{\varepsilon_k} \\
\leq& 2\|\,(x_*)_{S^c}\,\|_1 - \|v^k_{S^c}\|_1 + \|x^k_{S^c\cap I^c}\|_1 - \frac{\|x^k_{S^c\cap I^c}\|_2^2}{\varepsilon_k}.
\end{aligned}$$

To proceed, we note that from the elementary inequality $ab \leq \frac{1}{2}(a^2+b^2)$ and from $\|x^k_{S^c\cap I^c}\|_1 \leq \sqrt{N}\|x^k_{S^c\cap I^c}\|_2$, it follows that

$$\|x^k_{S^c\cap I^c}\|_1 \leq \frac{1}{2}\left(\frac{\varepsilon_k\|x^k_{S^c\cap I^c}\|_1^2}{2\|x^k_{S^c\cap I^c}\|_2^2} + 2\frac{\|x^k_{S^c\cap I^c}\|_2^2}{\varepsilon_k}.\right) \leq \frac{\varepsilon_k N}{4} + \frac{\|x^k_{S^c\cap I^c}\|_2^2}{\varepsilon_k}.$$

Hence, using that $\varepsilon_k \leq \sigma_s(x^k)_{\ell_1}/N$, we have shown that

$$\begin{aligned}
\sum_{i\in S^c} \frac{x_i^k}{\max(|x_i^k|,\varepsilon_k)} v_i^k &\leq 2\|\,(x_*)_{S^c}\,\|_1 - \|v^k_{S^c}\|_1 + \frac{\varepsilon_k N}{4} \\
&= 2\|\,(x_*)_{S^c}\,\|_1 - \|v^k_{S^c}\|_1 + \frac{\sigma_s(x^k)_{\ell_1}}{4} \\
&\leq 2\|\,(x_*)_{S^c}\,\|_1 - \|v^k_{S^c}\|_1 + \frac{\|x^k_{S^c}\|_1}{4} \\
&\leq 2\|\,(x_*)_{S^c}\,\|_1 - \|v^k_{S^c}\|_1 + \frac{\|v^k_{S^c}\|_1}{4} + \frac{\|(x_*)_{S^c}\|_1}{4} \\
&= \frac{9}{4}\|\,(x_*)_{S^c}\,\|_1 - \frac{3}{4}\|v^k_{S^c}\|_1,
\end{aligned}$$

where we used the triangular inequality for the vector $v^k = x^k - x_*$ on the set S^c and the fact that $\sigma_s(x^k)_{\ell_1} \leq \|x^k_{S^c}\|_1$. Hence, by adding up terms, we obtain that

$$\langle \nabla \mathcal{J}_{\varepsilon_k}(x^k), v^k\rangle \leq \frac{9}{4}\|\,(x_*)_{S^c}\,\|_1 - \left(\frac{3}{4} - \rho_s\right)\|v^k_{S^c}\|_1 \leq -(\beta-\rho_s)\|v^k_{S^c}\|_1.$$

Here, we have set $\beta = 3/4$ in the case that $\sigma_s\left(x_*\right)_{\ell_1} = 0$ and $\beta = 1/4$ else. Moreover, we used the assumption $\| \left(x_*\right)_{S^c} \|_1 \leq \frac{2}{9} \|v^k_{S^c}\|_1$.

Part II: Bounding the quadratic term $\langle v^k, \operatorname{diag}(w_{\varepsilon_k}(x^k))v^k \rangle$

In order bound the quadratic term in eq. (2.33) we first decompose it into two parts

$$\langle v^k, \operatorname{diag}(w_{\varepsilon_k}(x^k))v^k \rangle = \sum_{i=1}^{N} \frac{(v_i^k)^2}{\max(|x_i^k|, \varepsilon_k)} = \sum_{i\in S} \frac{(v_i^k)^2}{\max(|x_i^k|, \varepsilon_k)} + \sum_{i\in S^c} \frac{(v_i^k)^2}{\max(|x_i^k|, \varepsilon_k)}. \tag{2.34}$$

For the first summand, we note that

$$\sum_{i\in S} \frac{(v_i^k)^2}{\max(|x_i^k|, \varepsilon_k)} \leq \frac{\|v_S^k\|_1 \|v_S^k\|_\infty}{\varepsilon_k} \leq \rho_s \frac{\|v_{S^c}^k\|_1 \|v^k\|_\infty}{\varepsilon_k} \leq \frac{\|v_{S^c}^k\|_1 \|v^k\|_\infty}{\varepsilon_k}. \tag{2.35}$$

For the second summand, it holds that

$$\sum_{i\in S^c} \frac{\left(v_i^k\right)^2}{\max(|x_i^k|, \varepsilon_k)} \leq \frac{\|v_{S^c}^k\|_\infty \|v_{S^c}^k\|_1}{\varepsilon_k} \leq \frac{\|v_{S^c}^k\|_1 \|v^k\|_\infty}{\varepsilon_k}, \tag{2.36}$$

Hence, by adding eq. (2.35) and eq. (2.36) up, it follows that

$$\langle v^k, \operatorname{diag}(w_{\varepsilon_k}(x^k))v^k \rangle \leq 2 \frac{\|v_{S^c}^k\|_1 \|v^k\|_\infty}{\varepsilon_k}.$$

Next, we note that

$$\|v^k\|_\infty \leq \rho_1 \|v^k\|_1 \leq \rho_1 \left(1+\rho_s\right) \|v_{S^c}^k\|_1 \leq 2\rho_1 \|v_{S^c}^k\|_1.$$

Hence, we have shown that

$$\langle v^k, \operatorname{diag}(w_{\varepsilon_k}(x^k))v^k \rangle \leq 4\rho_1 \frac{\|v_{S^c}^k\|_1^2}{\varepsilon_k}.$$

Part III: Combining the bounds to obtain a decrease in k-th step:

Inserting the bounds of Part I and Part II into eq. (2.33) we obtain

$$Q_{\varepsilon_k}(x^k + tv^k, x^k) - \mathcal{J}_{\varepsilon_k}(x^k) \leq -tb + t^2 a =: h(t), \tag{2.37}$$

where the function $h : \mathbb{R} \to \mathbb{R}$ is a quadratic polynomial with coefficients $b = \left(\beta - \rho_s\right) \|v_{S^c}^k\|_1$ and $a = 4\rho_1 \frac{\|v_{S^c}^k\|_1^2}{\varepsilon_k}$. We observe that the minimizer of h is given by $t = \frac{b}{2a}$. Inserting this into h, we obtain that

$$h\left(\frac{b}{2a}\right) = -\frac{b^2}{4a} = -\frac{(\beta-\rho_s)^2\|v^k_{S^c}\|_1^2\varepsilon_k}{16\rho_1\|v^k_{S^c}\|_1^2} = -\frac{(\beta-\rho_s)^2}{16\rho_1}\varepsilon_k. \tag{2.38}$$

Combining this with Equation 2.12, we obtain, for $t = \frac{b}{2a}$,

$$\begin{aligned}\mathcal{J}_{\varepsilon_{k+1}}(x^{k+1}) - \mathcal{J}_{\varepsilon_k}(x^k) &\le Q_{\varepsilon_k}(x^{k+1}, x^k) - \mathcal{J}_{\varepsilon_k}(x^k)\\ &\le Q_{\varepsilon_k}(x^k + tv^k, x^k) - \mathcal{J}_{\varepsilon_k}(x^k) \le -\frac{(\beta-\rho_s)^2}{16\rho_1}\varepsilon_k.\end{aligned}$$

Hence, by rearranging terms, it follows that

$$\mathcal{J}_{\varepsilon_{k+1}}(x^{k+1}) - \|x_*\|_1 \le \mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 - \frac{(\beta-\rho_s)^2}{16\rho_1}\varepsilon_k. \tag{2.39}$$

In order to proceed, we need to bound ε_k from below. Here, for the second time, we need to use the definition of ε_k proposed in Algorithm 1. For that, we note that

$$\varepsilon_k = \min\left(\varepsilon_{k-1}, \frac{\sigma_s(x^k)_{\ell_1}}{N}\right) = \frac{\sigma_s\left(x^\ell\right)_{\ell_1}}{N}$$

for some $\ell \le k$. By Lemma 2.3.5, we have the following inequality chain

$$N\varepsilon_k = \sigma_s\left(x^\ell\right)_{\ell_1} \ge \frac{1}{3}\left(\mathcal{J}_{\varepsilon^\ell}\left(x^\ell\right) - \|x_*\|_1\right) \ge \frac{1}{3}\left(\mathcal{J}_{\varepsilon^k}\left(x^k\right) - \|x_*\|_1\right),$$

where in the second inequality, we have used that, by induction, $\mathcal{J}_{\varepsilon_k}\left(x^k\right) \le \mathcal{J}_{\varepsilon^\ell}\left(x^\ell\right)$. Plugging this into Equation 2.39 leads to

$$\mathcal{J}_{\varepsilon_{k+1}}(x^{k+1}) - \|x_*\|_1 \le \left(1 - \frac{(\beta-\rho_s)^2}{48\rho_1 N}\right)\left(\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1\right).$$

This finishes the induction step and concludes the proof of Theorem 2.3.7.

□

From Theorem 2.3.7 we can deduce Theorem 2.3.3, the first main result of this manuscript.

Proof of Theorem 2.3.3 . Recall that by Theorem 2.3.7 we have for all $k \in \mathbb{N}$ that

$$\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 \le \left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^k \left(\mathcal{J}_{\varepsilon_0}(x^0) - \|x_*\|_1\right)$$

with a constant $c_{\rho_s} = \frac{(3/4-\rho_s)^2}{48}$ and where S denotes the set which contains the s largest entries of x_* in absolute value. By our assumption $\rho_s < 1/2$ it follows that $c_{\rho_s} \geq 1/768$, which implies that inequality Equation 2.27 holds.

By Lemma 2.3.5 we have that

$$\mathcal{J}_{\varepsilon_0}(x^0) - \|x_*\|_1 \leq 3\sigma_s\left(x^0\right)_{\ell_1} \leq 3\|x^0 - x_*\|_1.$$

Next, we note that, again by Lemma 2.3.5, it holds that

$$\frac{1-\rho_s}{1+\rho_s}\|x - x_*\|_1 \leq \mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1.$$

Combining the three inequalities in this proof together with the assumption $\rho_s \leq 1/2$ yields inequality 2.28, which finishes the proof. □

We now generalize Theorem 2.3.3 to the scenario where the ground truth x_* is only approximately sparse. By that, we mean that the vector x_* can be well-approximated by an s-sparse vector in the sense that the ℓ_1-error of the best s-term approximation $\sigma_s(x_*)_{\ell_1} = \inf\{\|x_* - z\|_1 : \; z \in \mathbb{R}^N \text{ is } s\text{-sparse}\}$ is small, which is a commonly used quantity to measure the model misfit to a sparse vector [FR13, Section 2.1]. If x_* is approximately sparse in this sense, we can only hope to *approximately* recover x_* by the ℓ_1-minimization program (P_1) as the next theorem shows.

Theorem 2.3.8. *Suppose that a matrix $A \in \mathbb{R}^{m\times N}$ satisfies the stable null space property of order s with constant $0 < \rho < 1$. Then, for any $x \in \mathbb{R}^N$, a solution x_* of (P_1) with $y = Ax$ approximates the vector x with ℓ_1-error*

$$||x - x_*||_1 \leq \frac{2(1+\rho)}{(1-\rho)}\sigma_s(x)_1. \tag{2.40}$$

Indeed, [DDFG10, Theorem 5.3 (iv)] showed that under a suitable null space property, IRLS for (P_1) finds a vector x, such that $\|x - x_*\|_1$ is at most a constant multiple of the optimal best s-term approximation error $\sigma_s(x_*)_{\ell_1}$. However, as for exactly sparse vectors x_*, only a local but no global convergence rate was provided in previous literature [DDFG10, Theorem 6.4].

The following result shows that we can also obtain global linear convergence of Algorithm 1 in this case. More precisely, Theorem 2.3.9 implies that $\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1$ decays exponentially fast until a certain accuracy is reached, which $\sigma_s\left(x_*\right)_{\ell_1}$ up to a constant multiple.

Theorem 2.3.9. *Consider the problem of recovering an unknown vector $x_* \in \mathbb{R}^N$ from known measurements of the form $y = Ax_*$. Assume that the measurement matrix $A \in$*

$\mathbb{R}^{m\times N}$ fulfills the ℓ_1-NSP of order s with constant $\rho_s < 1/8$. Let the IRLS iterates $\left\{x^k\right\}_k$ and $\{\varepsilon_k\}_k$ be defined by (2.9) and (2.10) with initialization x^0. Then, the following three statements hold.

1. *For $k \le \hat{k} := \min\left\{k \in \mathbb{N} : \sigma_s\left(x_*\right)_{\ell_1} > \frac{2}{9}\left\|\left(x_* - x^k\right)_{S^c}\right\|_1\right\}$ it holds that*

$$\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 \le \left(1 - \frac{c}{\rho_1 N}\right)^k \left(\mathcal{J}_{\varepsilon_0}(x^0) - \|x_*\|_1\right), \tag{2.41}$$

where S denotes the support of the s largest entries of x_*.

2. *For all $1 \le k \le \hat{k}$ it holds that*

$$\|x^k - x_*\|_1 \le 6\left(1 - \frac{c}{\rho_1 N}\right)^k \|x^0 - x_*\|_1 + 10\sigma_s\left(x_*\right)_{\ell_1}. \tag{2.42}$$

3. *Moreover, for all integers $k \gtrsim \rho_1 N \log\left(\frac{\|x^0 - x_*\|_1}{\sigma_s(x_*)_{\ell_1}}\right)$ we have that*

$$\|x^k - x_*\|_1 \le 20\sigma_s\left(x_*\right)_{\ell_1}. \tag{2.43}$$

Here $c = 1/3072$ and $\rho_1 < 1/8$ denotes the constant for the ℓ_1-NSP of order 1.

Remark 2.3.10. *Applying Theorem 2.3.9 to the special case $\sigma_s\left(x_*\right)_{\ell_1} = 0$, we observe that inequality 2.28 yields a seemingly sharper result than inequality 2.42 in Theorem 2.3.3, which may seem somewhat counterintuitive. However, note that in Theorem 2.3.3 we require $\rho_s < 1/2$, whereas in Theorem 2.3.9 we have the stronger assumption $\rho_s < 1/8$. Indeed, a closer inspection of the proofs reveals that both the factors 3 and 6 in the inequalities 2.28 and 2.42 can be replaced by the factor $\frac{3(1+\rho_s)}{1-\rho_s}$, reconciling those two results.*

Now, we prove the second main result in this manuscript, Theorem 2.3.9, which deals with the approximately sparse case.

Proof of Theorem 2.3.9. Recall that

$$\hat{k} := \min\left\{k \in \mathbb{N} :\ \sigma_s\left(x_*\right)_{\ell_1} > \frac{2}{9}\|\left(x_*\right)_{S^c} - x^k_{S^c}\|_1\right\}.$$

Moreover, we note that by Theorem 2.3.7 we have for $k \le \hat{k}$ that

$$\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 \le \left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^k \left(\mathcal{J}_{\varepsilon_0}(x^0) - \|x_*\|_1\right) \tag{2.44}$$

with a constant $c_{\rho_s} = \frac{(1/4-\rho_s)^2}{48}$. Hence, by our assumption $\rho_s < 1/8$ we obtain $c_{\rho_s} \geq 1/3072$ and inequality Equation 2.41 follows, which proves the first statement. To prove the second statement, let $\tilde{k}$ and k be natural numbers, such that $\tilde{k} \leq \hat{k}$ and $k \geq \tilde{k}$ holds. Then, we obtain that

$$\begin{aligned}
\frac{1-\rho_s}{1+\rho_s}\|x^k - x_*\|_1 - 2\sigma_s\left(x_*\right)_{\ell_1} &\leq \mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 \\
&\leq \mathcal{J}_{\varepsilon_{\tilde{k}}}(x^{\tilde{k}}) - \|x_*\|_1 \\
&\leq \left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^{\tilde{k}} \left(\mathcal{J}_{\varepsilon_0}(x^0) - \|x_*\|_1\right) \\
&\leq 3\left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^{\tilde{k}} \sigma_s\left(x^0\right)_{\ell_1},
\end{aligned}$$

where in the first inequality, we applied Lemma 2.3.5. In the second inequality, we used that the sequence $\left\{\mathcal{J}_{\varepsilon^\ell}\left(x^\ell\right)\right\}_\ell$ is monotonically decreasing and in the third inequality we used inequality 2.44. In the fourth inequality, we again used Lemma 2.3.5. By rearranging terms and using the assumption $\rho_s < 1/8$ it follows for all integers $\tilde{k}$ and k such that $\tilde{k} \leq \hat{k}$ and $k \geq \tilde{k}$

$$\|x^k - x_*\|_1 \leq 6\left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^{\tilde{k}} \sigma_s\left(x^0\right)_{\ell_1} + 4\sigma_s\left(x_*\right)_{\ell_1}. \tag{2.45}$$

To proceed, recall that S denotes the support of the s largest entries of x_*. Then we note that

$$\sigma_s\left(x^0\right)_{\ell_1} \leq \|x^0_{S^c}\|_1 \leq \|\left(x^0 - x_*\right)_{S^c}\|_1 + \|\left(x_*\right)_{S^c}\|_1 \leq \|x^0 - x_*\|_1 + \sigma_s\left(x_*\right)_{\ell_1}. \tag{2.46}$$

Hence, we have shown that for all integers $\tilde{k}$ and k such that $\tilde{k} \leq \hat{k}$ and $k \geq \tilde{k}$ it holds that

$$\|x^k - x_*\|_1 \leq 6\left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^{\tilde{k}} \|x^0 - x_*\|_1 + 10\sigma_s\left(x_*\right)_{\ell_1}. \tag{2.47}$$

By setting $k = \tilde{k}$, we observe that this implies inequality 2.42, which shows the second statement. To prove the third statement, we will distinguish two cases. For the first case, assume that $\hat{k} \geq \left\lceil \frac{\rho_1 N}{c_{\rho_s}} \log\left(\frac{\|x^0 - x_*\|_1}{\sigma_s(x_*)_{\ell_1}}\right) \right\rceil$. Then for $k \geq \tilde{k} := \left\lceil \frac{\rho_1 N}{c_{\rho_s}} \log\left(\frac{\|x^0 - x_*\|_1}{\sigma_s(x_*)_{\ell_1}}\right) \right\rceil$ it follows from inequality eq. (2.47) that

$$\|x^k - x_*\|_1 \leq 6\left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^{\frac{\rho_1 N}{c_{\rho_s}} \log\left(\frac{\|x^0 - x_*\|_1}{\sigma_s(x_*)_{\ell_1}}\right)} \|x^0 - x_*\|_1 + 10\sigma_s\left(x_*\right)_{\ell_1} \leq 20\sigma_s\left(x_*\right)_{\ell_1}.$$

where in the second inequality, we have used the elementary inequality $\log(1+t) \leq t$ for $t > -1$. This shows the third statement in the first case. To prove the second case, assume that $\hat{k} < \left\lceil \frac{\rho_1 N}{c_{\rho_s}} \log\left(\frac{\|x^0 - x_*\|_1}{\sigma_s(x_*)_{\ell_1}} \right) \right\rceil$. Then, we can compute that

$$\begin{aligned}
\frac{1-\rho_s}{1+\rho_s}\|x^k - x_*\|_1 - 2\sigma_s(x_*)_{\ell_1} &\leq \mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 \\
&\leq \mathcal{J}_{\varepsilon_{\tilde{k}}}(x^{\hat{k}}) - \|x_*\|_1 \\
&\leq 3\sigma_s\left(x^{\hat{k}}\right)_{\ell_1} \\
&\leq 3\|x^{\hat{k}} - x_*\|_1 + 3\sigma_s(x_*)_{\ell_1} \\
&\leq 3(1+\rho_s)\left\|\left(x^{\hat{k}} - x_*\right)_{S^c}\right\|_1 + 3\sigma_s(x_*)_{\ell_1} \\
&\leq 20\sigma_s(x_*)_{\ell_1}.
\end{aligned}$$

In the first and third inequality, we have used Lemma 2.3.5. In the second inequality, we have used the monotonicity of the sequence $\left\{\mathcal{J}_{\varepsilon_k}\left(x^k\right)\right\}_k$. In the fourth inequality, we have argued as in inequality 2.46, and in the fifth inequality, we have used the null space property. In the last inequality, we have used that by definition of $\hat{k}$ it holds that $\sigma_s(x_*)_{\ell_1} > \frac{2}{9}\|(x_*)_{S^c} - x^{\hat{k}}_{S^c}\|_1$. This shows that the third statement holds in the second case, which finishes the proof. □

2.4 Numerical experiments

In this section, we first examine whether IRLS indeed exhibits two distinct convergence phases, a *global* one, as described in this paper, and a *local* one, as described in [DDFG10, ABH19], corresponding to different linear convergence rate factors. This question is very important in the optimization literature, and, indeed, many algorithms for sparse recovery and low-rank matrix recovery exhibit this type of behavior, e.g., [SLCX23, LFP17]. This phenomenon is also known as *manifold identification* [SJNS19, BIM22], a problem that dates back to [Dun87]. See also [LW11, BM88] for a discussion of this problem in the case of constrained optimization, a question that will arise again in Chapter 5. Second, we explore to which extent the dimension dependence in the convergence rates (2.27) and if we can expect a dimension-free linear convergence rate factor or if rather the (2.28) indicated by Theorem 2.3.3 is necessary.

2.4.1 Two-phase convergence phase

As discussed in the beginning of Section 2.3 the local convergence result of [DDFG10, Theorem 6.1] depends on the locality condition $\xi(k) := \frac{\|x^k - x_*\|_1}{\min_{i\in S}|(x_*)_i|} < 1$, cf. Equation 2.25. Under this condition and a null space property of order $\widehat{s} > s$ with constant $\rho_{\widehat{s}}$ such that $0 < \rho_{\widehat{s}} < 1 - \frac{2}{\widehat{s}+2}$ and $\hat{s} > s + \frac{2\rho_{\widehat{s}}}{1-\rho_{\widehat{s}}}$ hold, Theorem 2.3.1 stated above implies that

$$\|x^{k+1} - x_*\|_1 \leq \mu\|x^k - x_*\|_1$$

with an absolute constant $\mu < 1$ which, in particular, *does not depend on the dimension* $N, m,$ *and* s. This corresponds to a local linear rate for IRLS. A very similar condition to (2.25) is required by the comparable and more recent local convergence statement for the IRLS variant considered in [ABH19].

Theorem 2.4.1. *[ABH19, Theorem III.6] Suppose that the measurement matrix* $A \in \mathbb{R}^{m\times N}$ *satisfies the null space property of order* s *for a certain* $0 < \rho < 1$, *and suppose that* $A^{-1}(y)$, *the set of solutions to the equation* $Ax = y$, *contains a* s*-sparse vector* x^*. *Set* $T := \{i | x_i^* \neq 0, 1 \leq i \leq N\}$ *and choose* $\tilde{\rho} \in (0, 1 - \rho(1 + \eta(1-\rho)))$, *where* ρ *is the NSP constant and* $\eta \in (0,1)$ *is initialized in the IRLS. Then there is a smallest* $k_0 \in \mathbb{N}$ *such that*

$$\|(x^{k_0} - x_*)_{T^c}\|_1 \leq \tilde{\rho} \min_{i\in T} |(x_*)_i| \ .$$

Moreover, for all $k \geq k_0$,

$$\|(x^{k+1} - x_*)_{S^c}\|_1 \leq \mu\|(x^k - x_*)_{S^c}\|_1, \tag{2.48}$$

$$\|x^k - x_*\|_1 \leq (1+\rho)\mu^{k-k_0}\|x^{k_0} - x_*\|_1, \tag{2.49}$$

where $\mu := \frac{\rho(1+\eta(1-\rho))}{1-\rho} < 1$.

We now explore the behavior of the IRLS algorithm for ℓ_1-minimization proposed here, with weights given by Equation (2.11), Algorithm 1, and the sharpness of Theorem 2.3.2. In order to do so, we design experiments that build on those of [DDFG10, Section 8.1]. Therefore, we consider vectors lying in a vector space of dimension $N = 8000$. We also sample independently a 200-sparse vector $x_* \in \mathbb{R}^N$ with random support $S \subset [N]$, $s = 200 = |S|$, chosen uniformly at random such that $(x_*)_S$ is chosen according to the Haar measure on the sphere of a 200-dimensional unit ℓ_2-ball. We choose a measurement matrix $A \in \mathbb{R}^{m\times N}$ with i.i.d. Gaussian entries such that $A_{ij} \sim \mathcal{N}(0, 1/m)$ while setting $m = \lfloor 2s\log(N/s)\rfloor$, which is given by the theory of compressive sensing [FR13, Chapter 9]. As described above, such a matrix is known to fulfill with high probability the ℓ_1-null space property of order s with constant $\rho_s < 1$ [FR13, Theorem 9.29].

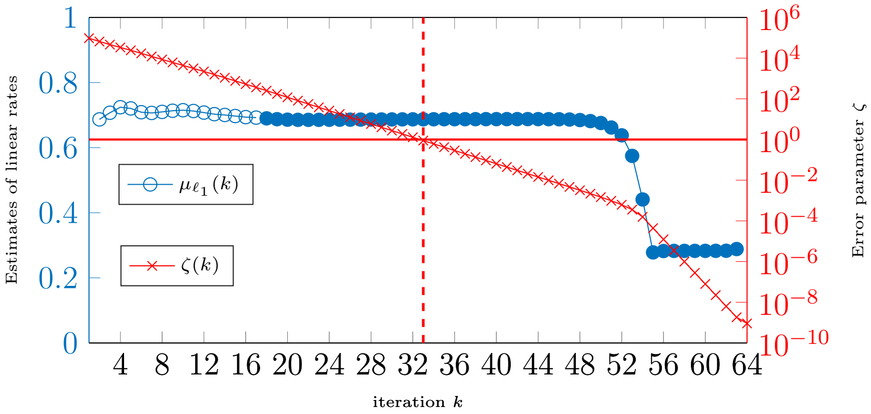


Figure 2.4: Instantaneous linear convergence rates of IRLS for ℓ_1-minimization ($N = 8000$): Linear convergence factors $\mu_{\ell_1}(k) := \|x^k - x_*\|_1 / \|x^{k-1} - x_*\|_1$ (in blue), filled blue circle if $S_k = S$ with S_k of Theorem 2.3.2 (support identification), and error parameter $\zeta(k) := \|x^k - x_*\|_1 / \min_{i \in S} |(x_*)_i|$ (in red). Horizontal (red) line: Threshold $\zeta = 1$. Vertical (red) line: First iterate k with $\zeta(k) < 1$.

In Figure 2.4, we track the decay of the ℓ_1-error $\|x^k - x_*\|_1$ of the iterates x^k rof Algorithm 1 via the values of $\zeta(k) := \|x^k - x_*\|_1 / \min_{i \in S} |(x_*)_i|$, depicted in red, and the behavior of the factor $\mu_{\ell_1}(k) := \|x^k - x_*\|_1 / \|x^{k-1} - x_*\|_1$, depicted in blue. We observe that the condition 2.25 for local convergence with the fast, dimension-less linear rate Equation 2.26 is satisfied after $k = 33$ iterations, as indicated by the vertical dashed red line.

In the first few iterations, $\zeta(k)$ is larger than 1 by several orders of magnitudes, suggesting that the local convergence rate results of [DDFG10, ABH19] do *not* apply until the later stages of the simulation: In fact, we observe that the support S of x_* is already perfectly identified via the s largest coordinates of x^k as soon as $k \geq 18$. For iterations $18 \leq k \leq 50$, the linear rate $\mu_{\ell_1}(k)$ remains very stable around ≈ 0.7, after which an accelerated linear rate can be observed. The latter phenomenon cannot be observed for the IRLS algorithm of [DDFG10] as it uses a slightly different objective function and weights than Algorithm 1. We believe it would also not be observed for the Algorithm from [ABH19]. In the example presented here, before $k = 18$, the rate $\mu(k)$ floats around the value 0.7. For all iterations k, $\mu(k)$ is smaller than 1, in line with the global linear convergence rate implied by Theorem 2.3.3.

In a similar experiment for a larger ambient space dimension $N = 16000$ and a smaller measurement-to-sparsity ratio such that $m = \lfloor 1.75 s \log(N/s) \rfloor$ results in a qualitatively similar situation, as seen in Figure 2.5: In Figure 2.5, we add also a plot of the linear convergence factor $\mu(k) := \frac{\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1}{\mathcal{J}_{\varepsilon_{k-1}}(x^{k-1}) - \|x_*\|_1}$ that tracks the behavior of the linear convergences in the smoothed ℓ_1-norm objective $\mathcal{J}$, cf. (2.31). In addition to what has been observed in Figure 2.4, we see that $\mu(k)$ and $\mu_{\ell_1}(k)$ exhibit very similar behavior for this example.

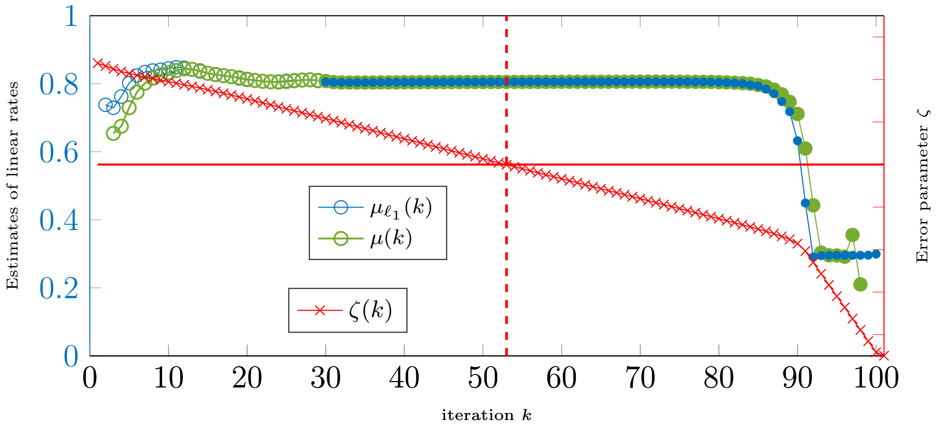


Figure 2.5: Standard initialization (uniform weights $(w_0)_i = 1$ for all i). Instantaneous linear convergence rates of IRLS for ℓ_1-minimization for $N = 16000$. Linear convergence factors $\mu_{\ell_1}(k) := \frac{\|x^k - x_*\|_1}{\|x^{k-1} - x_*\|_1}$ (in blue) and $\mu(k) := \frac{\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1}{\mathcal{J}_{\varepsilon_{k-1}}(x^{k-1}) - \|x_*\|_1}$ (in green), filled circles if $S_k = S$ (perfect support identification), and error parameter $\zeta(k) := \|x^k - x_*\|_1 / \min_{i \in S} |(x_*)_i|$ (in red), horizontal and vertical red lines as in Figure 2.4.

Hence, these experiments indicate that we can distinguish two phases. A few works discuss why such behavior is typical for non-smooth objective functions [SLCX23] and why methods like the (sub)gradient descent method, for example, would require decaying stepsizes in the first phase of the convergence. In the first phase of IRLS, the global one, we observe a linear convergence where the instantaneous linear convergence rate has not yet stabilized. In the second one, the instantaneous linear convergence stabilizes when the support identification problem has been solved.

As discussed in the introduction of this chapter, other methods, such as proximal algorithms, exhibit similar behavior. In particular, there are some algorithms for which convergence results with a two-phase behavior have already been established. For example, [LFP17] showed that a forward-backward method applied to the Lasso problem exhibits local linear convergence and that after a finite number of iterations, the region of fast convergence is reached. In particular, [LFP17, Proposition 3.6(ii)] provides a bound on this number of iterations, which scales proportionally with $||x_* - x^0||_2^2$. On the other hand, what is remarkable is that our result for IRLS, (2.3.3), provides a bound on the number of iterations until the fast linear convergence rate is reached that scales proportionally with $\log(||x_* - x^0||_2)$, but also proportionally with the dimension N. Moreover, most of these results require stronger assumptions than the NSP, such as the restricted isometry property, restricted condition number [AS21], or a restricted strong convexity/smoothness property. At the time of this thesis, some results for linear convergence (with high probability) were recently established, followed by a study of the two-phase behavior for iterative hard thresholding (a projected sub-gradient descent-type algorithm). However, they rely on the strong assumption that the measurement matrix is given by a Gaussian matrix with independent rows [SLCX23].

2.4.2 Global convergence rate and its dimension dependence

In this section, we explore to what extent the dependence on N in the convergence rates given by Equations 2.28 and 2.43 is sharp or if we can rather expect a dimension-free linear convergence rate factor. Such kind of *complexity discussion* appeared already in the literature for, for example, interior-point algorithms such as the barrier method [NN94]. In particular, in [BV04, Section 11.5.3], one can find a discussion about the number of necessary iterations for the Newton step in the barrier method, which also exhibits a two-phase convergence phenomenon. In Chapter 3, we will discuss the connection between IRLS and Newton's method. Therefore, there are parallels in the complexity analysis for both methods.

To analyze the dimension dependency, we design a *hard* experiment where a variant of IRLS is initialized with the weight vector $w_0 \in \mathbb{R}^N$ not uniformly as in Algorithm 1, but based on an *adversary initialization*, here denoted by z^{adv}. Since the hardest step in sparse recovery is to retrieve the support, we start with a vector that contains all the information off-support. More specifically, we first compute a minimizer

$$z^{\text{adv}} \in \underset{z \in \mathbb{R}^{S^c}: A_{S^c} z = y}{\arg\min} \|z\|_1$$

of the ℓ_1-minimization problem restricted to the off-support coordinates of x_* indexed by S^c and set then $x^0 \in \mathbb{R}^N$ such that $x^0_{S^c} := z^{\text{adv}}$ and $x^0_S = 0$. Based on this *initialization* x^0, we compute $\varepsilon_0 := \frac{\sigma_s(x^0)_{\ell_1}}{N}$ and set the first weight vector such that for all $i \in [N]$,

$$(w_0)_i := \frac{1}{\max\left(|x^0_i|, \varepsilon_0\right)}, \tag{2.50}$$

before proceeding with the IRLS steps Equation 2.9, 2.10 and 2.11 until convergence.

We observe in Figure 2.6 that this initialization, which is *adversary* as it sets very large initial weights on the coordinates of S that correspond to the true support of x_*, eventually results in the same behavior of Algorithm 1 as for the standard initialization by uniform weights, identifying the true support at iteration $k = 39$ compared to $k = 30$. However, in the first few iterations, we see that the instantaneous linear convergence factor $\mu(k)$ is close to 1 with $\mu(1) = 0.980$, decreasing only slowly before stabilizing around 0.79 after around $k = 30$.

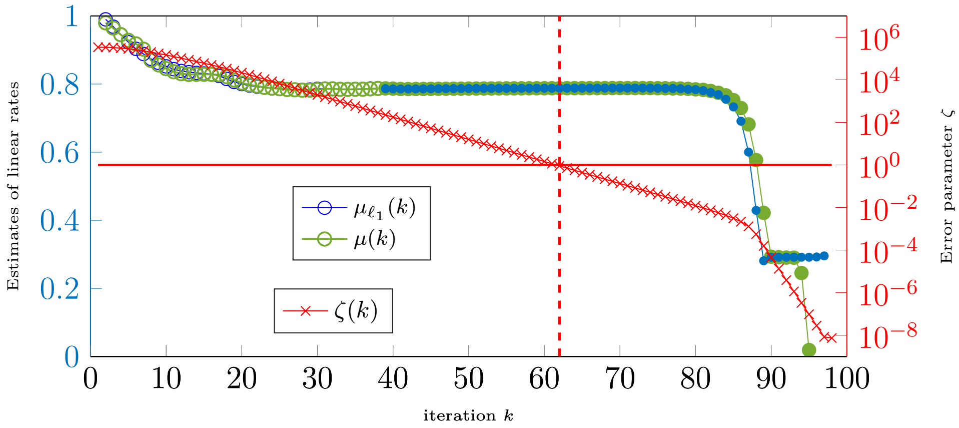


Figure 2.6: Adversary initialization (weights $(w_0)_i$ as in (2.50)). Instantaneous linear convergence rates of IRLS for ℓ_1-minimization for $N = 16000$: Linear convergence factors $\mu_{\ell_1}(k) := \frac{\|x^k - x_*\|_1}{\|x^{k-1}-x_*\|_1}$ (in blue) and $\mu(k) := \frac{\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1}{\mathcal{J}_{\varepsilon_{k-1}}(x^{k-1}) - \|x_*\|_1}$ (in green), filled circles if $S_k = S$ (perfect support identification), and error parameter $\zeta(k) := \|x^k - x_*\|_1 / \min_{i \in S} |(x_*)_i|$ (in red), horizontal and vertical red lines as in Figure 2.4.

While this is just one example, this already indicates that a linear rate such as Equation 2.26, i.e., without dependence on the dimension N (which has been proven locally in [DDFG10, Theorem 6.1] and [ABH19, Theorem III.6]) might not hold in general.

In our next experiment, we further investigate numerically the dimension dependence of the worst-case linear convergence factor $\mu(k) := \frac{\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1}{\mathcal{J}_{\varepsilon_{k-1}}(x^{k-1}) - \|x_*\|_1}$, which is upper bounded by the result of Theorem 2.3.3. We saw that in the experiment using the adversary initialization mentioned above and depicted in Figure 2.6, the maximal value was attained in the first iteration, i.e., for $\mu(1)$, as the effect of the adversary initialization is most eminent for $k = 1$ and it decreases with the iteration of the algorithm,.

In the third experiment, IRLS is executed for different ambient dimensions $N = 125 \cdot 2^{\ell/2}$ for $\ell = 0, 1, \ldots, 14$ and it is started from the adversary initialization. For each of the values of N, vectors $x_* \in \mathbb{R}^N$ of sparsity $s = 40$ are sampled from the same random model as above. The number of i.i.d. Gaussian measurements is given by $m = \lfloor 2s \log(N/s) \rfloor$. We average the resulting values for $\mu(1)$ across 500 independent realizations of the experiment. In Figure 2.7, we see that dependence on N of linear convergence factor $\mu(1)$ as observed in this experiment is quite well described by the upper bound (2.27) provided by our main result Theorem 2.3.3. Roughly speaking, our result predicts that the rate $\mu \approx 1 - \frac{1}{N}$. We observe that $\frac{1}{1-\mu(1)}$ scales almost linearly with N. As 2.3.4 after Theorem 2.3 indicates, the constant ρ_1 of the null space property of order 1 scales with $\sqrt{\frac{\log N}{m}}$, and therefore a precise dependence on all the parameters such as m and s might be more complicated than what can be observed in this experiment.

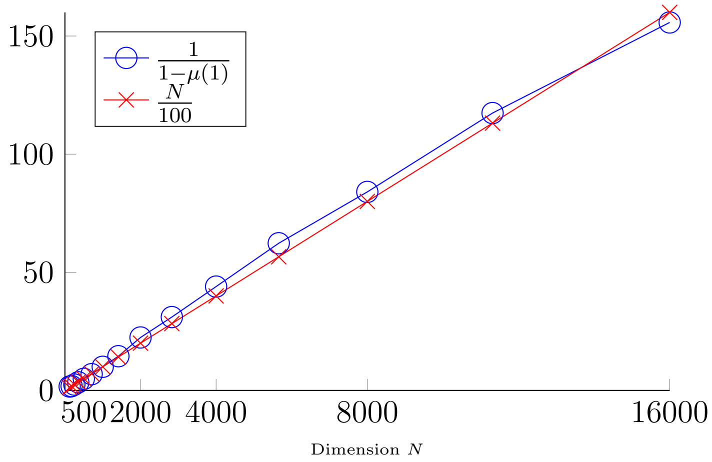


Figure 2.7: Comparison of $\frac{N}{100}$ and $\frac{1}{1-\mu(1)}$ (for which Proposition 2.3.7 provides an upper bound of $\frac{\rho_1 N}{c}$) for different dimension parameters N, where $\mu(1) = \frac{\mathcal{J}_{\varepsilon_1}(x^1)-\|x_*\|_1}{\mathcal{J}_{\varepsilon_0}(x^0)-\|x_*\|_1}$ is the linear convergence factor, for IRLS initialized from adversary initialization.

Nevertheless, Figure 2.7 provides a piece of strong evidence that one should not expect the linear convergence rate of Theorem 2.3.7 to be dimension-free. That being said, it is an interesting open problem to investigate the precise parameter dependence of μ in greater detail. As mentioned at the beginning of this section, a similar question was analyzed for other methods such as interior-point methods, cf. [BV04, Section 11.5] and [N$^+$18, Section 5.2] and all the known results in the literature, to the best of the author's knowledge, are dimension-dependent. Our convergence proof is quite conservative, and it is possible that with different weighting and smoothing schemes for IRLS, one could obtain a sharper convergence rate. We leave this as an open problem.

Open Problem: Establish the optimal dependence on N for the convergence rate of IRLS methods.

In view of the experiments described above and our (potentially pessimistic) analysis, it is interesting to investigate whether a global convergence rate with better dimension dependency is possible, for example, via a *smoothed analysis* [ST03, DH18], which measures the expected performance of algorithms under slight random perturbations of worst-case inputs.

Recently, some works developed the notion of *Trimmed LASSO* that possesses excellent numerical properties [BCM17] and for which a majorization-minimization strategy was also developed [ABN21]. We think it is an interesting problem to leverage the current theoretical knowledge about algorithms for the trimmed LASSO with the techniques described in this chapter.

Open Problem: Is it possible to extend the analysis provided in this chapter to algorithms designed for the Trimmed LASSO objective function?

2.5 Chapter Conclusion

In this chapter, we presented the first idea related to least squares in this thesis, and we discussed how a specific choice of objective function and weights leads to a powerful, scalable, and fast algorithm for ℓ_1-minimization. After covering a part of the vast history of IRLS and its developments for several different problems, we showed how new ideas can lead to improved numerical stability and stronger theoretical results as compared to previous methods. In particular, we solved an open problem in the algorithmic theory for sparse recovery. Namely, we established a global linear convergence rate for IRLS under minimal assumptions. We have corroborated our theory with numerical experiments that, first, discussed the difference between the local and global convergence phase and, second, elucidated the optimality of the dimension dependence of convergence rate given by our main theorem, Theorem 2.3.3. As a potential future work direction, we note that there are currently no convergence rates available for IRLS optimizing a convex objective function that resembles the ℓ_1-norm, such as the nuclear norm. This objective function is also discussed in the literature for, for example, matrix completion [MF12a, FRW11a, KS18] and tensor completion problems [YZ16]. It would be interesting to generalize the theory presented in this chapter to these problems. We conclude by noting that we left a few open problems and future research directions that we believe are interesting to pursue.

Chapter 3

Ill-conditioned low-rank matrix completion

> *"Some movies are notoriously bad, and people who did not like them always give them as negative examples, indicating what they do not want to watch. However, for the other part of the population, who liked those movies, they are not going to be remembered long as salient positive examples. Thus, when rating in bulk, long after watching the movie, only those who disliked the movie will rate it."*
> Yehuda Koren describing the progress on the Netflix prize [Kor09]

This chapter proposes an IRLS-type algorithm to complete highly ill-conditioned low-rank matrices. This algorithm, which can be interpreted as a saddle-escaping smoothing quasi-Newton method or a variable metric proximal gradient method, combines the favorable data efficiency of IRLS approaches with improved scalability by several orders of magnitude. The work presented in this chapter was written in collaboration with Dr. Christian Kümmerle, and it was published at the Workshop on "Beyond first-order methods in ML systems" at the 37th International Conference on Machine Learning, 2020 under the title *Escaping Saddle Points in Ill-Conditioned Matrix Completion with a Scalable Second Order Method* [KMV20]. And also at the main International Conference on Machine Learning 2021 under the title *A Scalable Second Order Method for Ill-Conditioned Matrix Completion from Few Samples* [KV21]. The results of this chapter also appeared in the Ph.D. thesis of Dr. Christian Kümmerle. We highlight our main contribution:

We establish a minimization scheme for the low-rank matrix completion problem that is able to retrieve highly ill-conditioned matrices in the optimal information theoretical regime, where the number of sufficient samples for the completion *does not depend on the condition number* and scales logarithmically with the dimension. We show that the method attains a *local quadratic convergence rate* and provide extensive experiments that indicate its advantages against other state-of-the-art methods.

3.1 Introduction

In the preceding chapter, we extensively explored the concept of sparsity as the primary principle of parsimony, and we devised an efficient and scalable algorithm based on the least squares framework to recover sparse vectors. In the present chapter, our focus will shift towards tabular data in the form of matrices and will generalize the previously developed concepts. Dealing with matrices introduces additional complexities in employing the least squares approach, but we will nonetheless proceed to design an algorithm capable of completing matrices from missing entries. Moreover, we will establish that this algorithm, which is based on the sequential minimization of majorizers of a highly non-convex function, attains local quadratic convergence. This represents a notable departure from the previous chapter, where such accelerated convergence was not established.

In the era of machine learning and data-driven models, low-rank matrices became ubiquitous in science and engineering due to their ability to capture simplified representations of complex information since they provide an efficient and parsimonious presentation of tabular data. In particular, the crucial information contained in an (approximately) low-rank dataset can be expressed by using a few vectors since the number of degrees of freedom of a low-rank matrix is much lower than the ambient dimension of the matrix, see Theorem 3.1.6 below. Thanks to that, low-rank structures usually allow for fast computations and efficient data storage and processing, which made their use well spread in large-scale or real-time applications. See [HMT11, HNWL21] and references therein.

Since the introduction of the *Singular Value Decomposition* (SVD) [Ste93] and the method of *Principal Component Analysis* (PCA) [Pea01, Hot33] (for more historical details, see the notes in the book [SS90]), where one aims for a low-rank matrix that explains the data in the sense that it minimizes the approximation error in the least squares sense, several methods having low-rank matrices at their core were proposed for dealing with large-scale data and for making sense of it. One of the earliest examples of using such matrices was developed in [Hot36]. This work employed the use of eigenvectors of a covariance matrix between two sets of variables to understand the relation between them, and the notion of low-rankness played a fundamental role in the analysis.

The development of tools from linear algebra that leverage low-rank models for compression [HCMTH15], sketching [W$^+$14], or streaming [TYUC19] of data, among other techniques for large-scale data, is a very active topic. Also, the numerous applications of such methods in machine linear and data science, such as in cardiac MRI [ZHBL10], statistics [LR19], recommender systems [RYL$^+$18], large language models [HSW$^+$21], hyperspectral imaging [PSL$^+$21], environmental sensing [KXL$^+$13], blind source separation [KMM$^+$18] geophysics [YMO13], phase retrieval [CESV15], molecular biology [WWZG15], system

identification [LV10], video processing [JLSX10], MIMO channel estimation in telecommunications [XGJ16], and in quantum information theory [GLF$^+$10], to name a few.
Low-rank matrices are more omnipresent than what was thought, as the result below shows. In particular, we can interpret it as *large datasets have a low-rank structure.*

Theorem 3.1.1. *[UT19, Theorem 1] Let $X \in \mathbb{R}^{m\times n}$ with $m \geq n$ and $0 < \epsilon < 1$. Then, with $r = \lceil 72 \log(2n+1)/\epsilon^2 \rceil$ we have*

$$\inf_{\operatorname{rank}(Y)\leq r} \|X - Y\|_{\max} \leq \epsilon \|X\|_2 \tag{3.1}$$

Remark 3.1.2. *As the authors of [UT19] discussed, it is important to assume $r < n$, which requires n to be extremely large; otherwise, the theorem is meaningless. In this case, the theorem says that any large matrix with a small spectral norm is well approximated by a low-rank matrix.*

In many of the applications mentioned above, when only one can access partial or incomplete data, an important sub-problem is to infer the matrix model from just a few samples. The general mathematical question is whether a partially specified rectangular array of data could be completed into a matrix satisfying certain properties. This problem of retrieving a matrix from partial observations is a highly ill-posed one since there are infinitely many matrices that could be used to complete a given subset of tabular data. Therefore, one needs to impose a certain structure on the underlying data to have a well-defined matrix completion (MC) problem. There are various MC problems [Joh90], such as positive definite completions[1], maximum entropy completions, and low-rank completions – the latter, known as *low-rank matrix completion* (LRMC) [NKS19, DR16, CLC19], being the subject of investigation in this chapter.
Among the LRMC problems, one of the most famous examples occurs in the field of recommender systems, where a rating matrix that represents how users rate certain products needs to be inferred from a few samples [CW22]. This lack of information comes from the fact that many users usually haven't rated most of the products but rather just a few of them. The problem is predicting the rates a certain user would give a certain item.

3.1.1 The Netflix problem

The problem described above gained a lot of attention from the media (`https://www.nytimes.com/2006/10/02/technology/02netflix.html`) after the streaming company

[1] This is still a very active field of research. See [CNX22] and `https://www.alignment.org/blog/prize-for-matrix-completion-problems/` for more details.

Netflix[2] created a public competition on 2 October 2006, where they offered the prize of $US\$1,000,000$ for developing an algorithm that could outperform their own method, called Cinematch. A team would be considered a winner if they could improve over Cinematch by 10% on the root mean squared error (RMSE) of the algorithm's prediction against the actual rating that a subscriber provides. The company provided a training data set of $100,480,507$ ratings, scores from 1 to 5, that $480,189$ users (nowadays the company has over 232 million users worldwide) gave to $17,770$ movies and a qualifying set of $2,817,131$ ratings that were used to test the proposed algorithms. The dataset was formed by randomly selecting a subset of all users who provided at least 20 ratings between October 1998 and December 2005. Interestingly, it took six days until a team had already found a better algorithm than the one by Netflix. But it was only in July 2009 that Bellkor's Pragmatic Chaos team was declared the winner after an improvement of 10.06% on the RMSE.

Mathematically speaking, the rating process can be described by matrix $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ whose (i,j) entry contains the rate that the user i's attributed to the item j. If we define the set $\Omega = \{(i,j)|$ rate that user i has given to movie $j\}$, the problem translates into retrieving $\mathbf{X}_0$ from the observed entries $Y = P(\mathbf{X}_0)$, where P is the projection operator onto the subset Ω. As described in Section 3.1, without any additional assumption, the problem is ill-posed since any completion is feasible. It is widely accepted in the recommender system community that only a few latent factors describe the rate that a user gives to a movie since their choices tend to be correlated by, for example, movie style or director [CW22]. Therefore, the underlying matrix would be (approximately) low rank. In Figure 3.1, we can observe this fact for the MovieLens dataset [HK15].

The low-rankness assumption was crucial in many of the solutions proposed for the Netflix challenge. In particular, *matrix factorization methods* that relied on this assumption were successfully employed. Interestingly, another important technique that was used by the winners was the method of *alternating least squares minimization*, which also related to the main theme of this thesis [KBV09]. Despite the great accuracy of such matrix completion methods for recommender systems, it is worth noting that they were not sufficient for practical settings[3]. Indeed, to make matrix completion methods useful for recommender systems, one needs to consider several other aspects of human-computer interaction, machine learning, and quantitative marketing. Nevertheless, this competition

[2] When the problem was announced, Netflix was still a DVD rental company. They started their streaming service in January 2007.

[3] Indeed, the company Netflix did not implement the winning solution. In their own words *"the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment"*. See `https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429`.

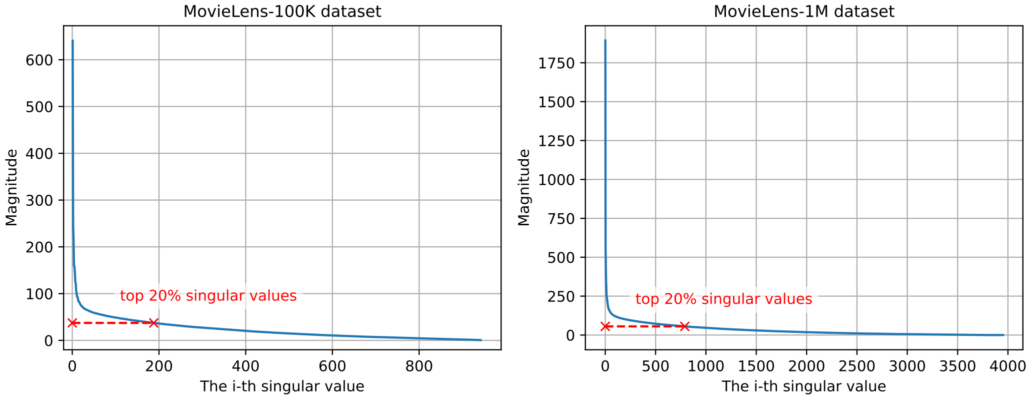


Figure 3.1: Singular values of rating matrices generated from two widely used datasets for recommender systems: (a) MovieLens-100K (1682 Movies and 943 Users) and (b) MovieLens-1M (3952 Movies and 6040 Users), where top 20% singular values account for 51.10% and 55.97% of the sum of all singular values, respectively.

stimulated a lot of research in the field of matrix completion.

3.1.2 The mathematical formulation

The low-rank matrix completion problem can be formulated as follows. Given a matrix $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ of rank-r and an index set $\Omega \subset [d_1] \times [d_2]$, the task is to reconstruct $\mathbf{X}_0$ just from the knowledge of Ω and $P_\Omega(\mathbf{X}_0)$, where $P_\Omega : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ (with $m \ll d_1 d_2$) is the subsampling operator that maps a matrix to the set of entries indexed by Ω.

$$\min_{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}} \operatorname{rank}(\mathbf{X}) \quad \text{subject to } P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{X}_0). \tag{3.2}$$

A connection exists between Equation (3.2) and the sparse recovery problem discussed in Chapter 2. If we describe the underlying matrix $\mathbf{X}_0$ by its SVD decomposition, in the case that U and V are nonsingular, we have $\operatorname{rank}(\mathbf{X}_0) = \operatorname{rank}(\Sigma)$. And $\operatorname{rank}(\Sigma)$ will be given by the number of non-zero elements of the vector of singular values $\vec{\sigma}(\mathbf{X}_0) = (\sigma_1, \ldots, \sigma_{\min\{d_1, d_2\}})$. Therefore, problem 3.2 is simply the problem of minimizing the ℓ_0 norm of the vector of singular values, subject to the constraints given by the data. From this connection, one can convert the sparse recovery problem Equation (P_0) into a rank minimization problem by considering the vector to be retrieved as a diagonal matrix. This shows that the rank minimization problem is as hard as the sparse recovery problem and, hence, it is an NP-hard problem.

Remark 3.1.3. *The low-rank matrix completion can actually be seen as a particular instance of a more general problem, the so-called* low-rank matrix recovery *[Mar18], where the linear operator $P_\Omega(\mathbf{X})$ will not be given by an orthogonal projection onto the set of* *entries but rather will be of the form $P_\Omega(\mathbf{X})_i = tr(A_i^T\mathbf{X})$ for general matrices $A_1, \ldots, A_m \in \mathbb{R}^{d_1\times d_2}$. This general problem is also known as the* affine rank minimization *problem in the control theory community [FHB03, MP97]. LRMC can be considered a (hard) particular example when $\{A_i\}_{kj} = 1$ if $(k,l) \in \Omega$, where Ω is the set of known entries and 0 elsewhere. For the recovery problem, under certain assumptions on the matrices $\{A_i\}_{i=1}^m$, several works established uniform and non-uniform recovery guarantees[4] (see [FR13, Section 9.2] for a discussion) with algebraic geometric [RWX21, CRWX18] or probabilistic [ENP12] techniques. See also [BCMN14, CEHV15] for a discussion about the injectivity of the measurement operator P_Ω in the particular case of phase retrieval and [KK17] for a discussion in the context of bilinear inverse problems.*

The focus of this chapter will be on the LRMC problem, which is considered a hard instance of the general matrix recovery problem since most of the techniques used for the general problem assume that the operator P preserves the geometry of certain structure sets, which is known as the *restricted isometry property* [RFP10], and this does not hold in the case of matrix completion. It is beyond the scope of this thesis to analyze the general recovery problem fully, and we will rather focus on the use of least squares for the LRMC problem.

From an optimization point of view, (3.2) is particularly difficult to handle due to two properties: its *non-convexity* and its *non-smoothness*. A widely studied approach in the literature replaces the rank($\mathbf{X}$) by the (convex) nuclear norm $\|\mathbf{X}\|_* = \sum_{i=1}^d \sigma_i(\mathbf{X})$ [FHB03], which is the tightest convex envelope of the rank, as the following theorem shows:

Theorem 3.1.4. *[FHB03, Theorem 1] The nuclear norm $\|\mathbf{X}\|_*$ is the convex envelope of the function* $\operatorname{rank}(\mathbf{X})$ *on the set $C = \{\mathbf{X} \in \mathbb{R}^{d_1\times d_2} \big| \|\mathbf{X}\|_{2\to 2} \leq 1\}$.*

The theorem above is the matrix version of what was described in Section 2.1.1, i.e., that the ℓ_1-norm ball is the convex hull of the intersection of ℓ_0-norm ball with the ℓ_∞-norm ball. With this in mind, we can state the most well-studied problem for matrix completion, namely, nuclear norm minimization (NNM):

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \|\mathbf{X}\|_* \quad \text{subject to } P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{X}_0). \tag{3.3}$$

Analogously to what was established in the case of sparse recovery, a mature theory has been developed from this fruitful convex approach, e.g., a necessary and sufficient condition – a matrix version of the NSP – for the low-rank matrix recovery problem via nuclear norm minimization [KKRT16, RXH08, RXH11]. Nevertheless, despite this analogy, the

[4] They are also known as strong and weak recovery. See [TBD11, Section 2.D.].

theory for matrix completion is much more involved, and the matrix completion framework is still a very active field of research, cf. the very recent papers that became available on ArXiv at the time of writing of this thesis, e.g., [BCZ23]. The theory for matrix completion includes, for example, performance guarantees for a near-optimal sample complexity [CT10, Che15] and robustness to (adversarial) noise [CP10a, CCF+20, FGJ+22].
However, from a practical point of view, using such a convex relaxation to find a low-rank completion is *computationally very demanding*, as it is equivalent to a semidefinite program as the next theorem shows:

Theorem 3.1.5. *[FHB03, Lemma 1] The problem Equation* (3.3) *is equivalent to the following semidefinite programming problem:*

$$\min_{\substack{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}, Y\in\mathbb{R}^{d_1\times d_1}\\ Z\in\mathbb{R}^{d_2\times d_2}}} \operatorname{tr} Y + \operatorname{tr} Z \qquad \textit{subject to } P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{X}_0) \ \ \textit{and} \ \ \begin{bmatrix} Y & \mathbf{X} \\ \mathbf{X}^T & Z \end{bmatrix} \succcurlyeq 0. \tag{3.4}$$

In principle, this problem could be tackled with generic semidefinite solvers based on interior point methods such as MOSEK [ApS22] or SeDuMi[Stu99]. However, the problem's computational complexity is at least cubic in the dimensions of $\mathbf{X}_0$. See [ZL18, CLC19] and [GM12, Chapter 2] for more details. Even first-order solvers have the same bad arithmetic complexity. Thus, convex relaxations are of little use in large-scale applications of the model, such as in recommender systems [KBV09], where even storing the dense matrix $\mathbf{X}_0 \in \mathbb{R}^{d_1\times d_2}$ is prohibitive.
From a statistical point of view, the most important question is how many measurements are necessary for the recovery. In this case, the important quantity is the *degrees of freedom* of a matrix, described in the theorem below.

Theorem 3.1.6. *A matrix $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$ of rank r has $deg_{\mathbf{X}} := r(d_1 + d_2 - r)$ degrees of freedom.*

Proof. Let the $\mathbf{X} = U\Sigma V^T$ be the singular value decomposition of the rank-r matrix $\mathbf{X}$. Since $\mathbf{X}$ is of rank k, the singular value matrix Σ can be written as a $k \times k$ diagonal matrix containing the k non-zero singular values of $\mathbf{X}$ and the singular vector matrices U and V will be of size $d_1 \times r$ and $d_2 \times r$, respectively. Each of the r singular values contributes as a free parameter. Then, the first column of the matrix U has $d_1 - 1$ degrees of freedom since the vector has d_1 and it has a unit norm. The next columns of U must be orthogonal to the ones before, which gives $(d_1 - 1) + (d_1 - 2) + \cdots + (d_1 - r) = d_1 r - r(r + 1)/2$ free parameters. In a similar way, V has $d_2 r - r(r + 1)/2$ free parameters. All together, the matrix $\mathbf{X}$ has a total of $r + d_1 r + d_2 r - r(r + 1) = r(d_1 + d_2 - r)$ degrees of freedom. □

Another important but less well-known issue is that a convex relaxation is *typically not as data efficient* as certain other algorithms [TW13, BNZ21], i.e., nuclear norm minimization typically necessitates a larger amount of samples m than other methods, measured by the quotient $\rho := m/(d_1+d_2-r)$ (oversampling ratio) between m and the number of degrees of freedom of $\mathbf{X}_0$, to identify $\mathbf{X}_0$ correctly [ALMT14]. The main contribution of this chapter is to develop a method that is able to retrieve matrices using as few measurements as possible, i.e., close to the information-theoretical regime $m = r(d_1 + d_2 - r)$, and to prove a quadratic convergence to the ground truth under very weak assumptions. This type of convergence means, as mentioned in [BV04, Section 9.5], that *"roughly speaking, ... after a sufficiently large number of iterations, the number of correct digits doubles at each iteration"*.

Before we develop our algorithm, which will also be based on a sequential quadratic programming idea where, at each iteration, a parabola is chosen to avoid local minima, we will discuss a few contributions to the topic of matrix completion.

3.1.3 Related work

Even before the Netflix challenge, the LRMC problem was already important in fields like control theory and graph theory. In the former, the problem of reconstructing a discrete linear time-invariant dynamical system from the first n time samples of its impulse response can be modeled as a structured matrix completion problem [LV10, FHB03]. In the latter, the problem of Euclidean distance matrix estimation, where a configuration of points needs to be reconstructed from its pairwise distances, can also be recast as a completion of a rank-deficient matrix [MMS11].

But it was not before the seminal papers by Fazel, Hindi, and Boyd [FHB03, Faz02] and, a few years later, by Candes, Recht, and Tao [CR09, Rec11, CT10], that theoretical progress for this problem was achieved. In particular, these papers introduced the first tools from optimization and non-asymptotic probability that helped to shape the field. The works [FHB03, Faz02] formulated the *trace* heuristic (see also [Sha82] for a pioneering paper on this idea) that coincides with the nuclear norm when the underlying matrix is a positive semidefinite one, and the *log-det* one that very much inspired the work in this chapter. They employed it to complete low-rank Hankel matrices and Euclidean distance matrices [Faz02, Chapter 6]. Later, the work [CR09] analyzed the probability of success of the nuclear norm minimization, introduced in [Faz02], for the retrieval of low-rank matrices. In particular, they proved that $\Omega(D^{5/4} r \log D)$ entries are sufficient for the semidefinite program Equation (3.4), where $D = \max(d_1, d_2)$, to recover all entries of $\mathbf{X}_0$ with high probability. Several works made progress on this approach and established very general recovery guarantees even when the measurements are corrupted by noise [Gro11, CT10, KS21b, CP10a, Rec11, Klo14, Che15, TKL11, CCF$^+$20]. See also the survey [FGJ$^+$22]. Notably, the paper [Che15] showed sharp guarantees for matrix completion via nuclear norm minimization. It proved that $\Omega(\deg_{\mathbf{X}_0} \log^2 D)$ measurements suffice for retrieving low-rank matrices while it was already known that $\Omega(\deg_{\mathbf{X}_0} \log D)$ are necessary for matrix completion, cf. [CT10, Theorem 1.7]. Despite all the theoretical progress, the lack of scalability of NNM and the non-optimality in the information-theoretical sense led to the development of various alternative methods.

Among the most popular ones are *non-convex* algorithms based on a variational formulation of the nuclear norm and matrix factorization ideas [BM03, MMBS13]. Indeed, the three formulas below are equivalent to the nuclear norm of a matrix $\|\mathbf{X}_0\|_*$, see [WM22, Proposition 4.6].

- $\|\mathbf{X}_0\|_* = \min_{U,V} \sum_k \|u_k\|_2 \|v_k\|_2$ s.t. $\mathbf{X}_0 = UV^* = \sum_k u_k v_k^*$.
- $\|\mathbf{X}_0\|_* = \min_{U,V} \|U\|_F \|V\|_f$ s.t. $\mathbf{X}_0 = UV^*$.
- $\|\mathbf{X}_0\|_* = \min_{U,V} \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2)$ s.t. $\mathbf{X}_0 = UV^*$.

Therefore, the NNM problem can be factorized and written, for example, as

$$\min_{\substack{U\in\mathbb{R}^{d_1\times r}, V\in\mathbb{R}^{d_2\times r}\\ \text{s.t.} \mathbf{X}_0 = UV^*}} \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2) \qquad \text{subject to } P_\Omega(UV^*) = P_\Omega(\mathbf{X}_0). \tag{3.5}$$

Remark 3.1.7. *Such an approach has been generalized to any Schatten-p quasi-norm for* $0 < p < 1$. *In particular, the work [SLS$^+$20] showed that any Schatten-p quasi-norm can be written as a product of the Schatten-r (quasi-)norm and Schatten-q (quasi-)norm of its two-factor matrices provided that* $1/p = 1/q + 1/r$. *This result was also generalized to the case of three or more factor matrices.*

It is common in the literature where factorization methods are employed to see the unconstrained version of the problem above, e.g., [RS05], namely,

$$J(\mathbf{U},\mathbf{V}) := \|P_\Omega(\mathbf{U}\mathbf{V}^*) - P_\Omega(\mathbf{X}_0)\|_F^2 + \frac{\lambda}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right) \tag{3.6}$$

for $\lambda \geq 0$, which use (projected) gradient descent on the two-factor matrices [SL16, ZL16, MWCC18], or related methods. These methods are much more scalable than those optimizing a convex rank surrogate while also allowing for theoretical analysis; see [CLC19] for a recent survey. Still, even though a huge amount of literature was devoted to understanding how well these methods perform for matrix recovery problems [BL21, ZBL21], in the case of matrix completion, recent results indicate that they also have limitations when the amount of available information is close to the information-theoretic limit [YZLS22]. Furthermore, among the most data-efficient methods for low-rank completion are those that minimize a smooth objective over the Riemannian manifold of fixed rank matrices [Van13, WCCL20, BA15, BNZ21, ZN22, CA16, SK22, DGHG22]. These approaches are likewise scalable and often able to reconstruct the low-rank matrix from fewer samples m than a convex formulation. The idea is that the optimization is performed over the smooth submanifold of matrices of fixed rank r with explicit knowledge of the tangent spaces and an efficient representation of tangent vectors. However, strong performance guarantees have remained elusive so far. To the best of the author's knowledge, this is the best theoretical paper available [WCCL20].

In many instances of the LRMC problem, such as in the discretization of PDE-based inverse problems [BSS21, CCBB14] or in spectral estimation problems modeled by structured low-rank matrices [Fas95, UC16], it is an additional difficulty that the matrix of interest $\mathbf{X}_0$ is severely *ill-conditioned*, i.e., $\kappa = \sigma_1(\mathbf{X}_0)/\sigma_r(\mathbf{X}_0)$ might be very large. For example, in spectral estimation, the matrices can reach up to $\kappa = 10^{15}$ [FL12]. Therefore, the main question was how to design an algorithm that is *scalable*, from the optimization point of view, *data-efficient*, from the statistical point of view, and is also able to retrieve highly ill-conditioned matrices.

A scaled (preconditioned) version of gradient descent specially designed for this purpose appeared in a few papers [TW13, NS12] and was finally analyzed in [TMC21]. This work claimed that they rigorously established a method that can complete highly ill-conditioned matrices from few measurements. However, a closer inspection shows that the sample complexity for their algorithm scales as $\Omega\left(\kappa^2 r \mathrm{deg}_{X_0} \max(\log(D), \mu_0\kappa^2)\right)$ [TMC21, Theorem 8]. Moreover, the numerical experiments conducted in the paper only completed matrices with condition number $\kappa = 20$, which would not be considered an ill-conditioned matrix by anyone working, for example, with numerical analysis! With that in mind, we address the following question:

How to develop an algorithm for matrix completion that is data-efficient and retrieves the matrix from only $m \approx \mathrm{deg}_{X_0} = r(d_1 + d_2 - r)$ entries, but that is also scalable, provable, can retrieve highly ill-conditioned matrices and is robust to noise errors?

Contribution of this chapter:

We propose a solution based on least squares where, at each iteration, the designed quadratic problem is tailored to the curvature of a highly non-convex function that can be seen as a sharp *non-convex "relaxations"* of the rank function and, thanks to that, is able to complete matrices in the information-theoretical regime. We also analyze the algorithm, called *Matrix Iteratively Reweighted Least Squares* (`MatrixIRLS`) and show that it is designed to find low-rank completions that are potentially very ill-conditioned, allowing for a scalable implementation. While being severely non-convex, we note that our method fundamentally differs from a typical non-convex approach with an objective such as (3.6).

Let $D = \max(d_1, d_2)$ and $d = \min(d_1, d_2)$. From a theoretical angle, we establish that if the m sampled entries are distributed uniformly at random and if $m = \Omega(\mu_0 r D \log D)$, with high probability, `MatrixIRLS` *exhibits local convergence* to $\mathbf{X}_0$ *with a local quadratic convergence rate*, where μ_0 is the incoherence factor [CCF$^+$21, Chapter 3] that will be discussed below in more details. This sample complexity does not depend on the condition number κ, is *optimal* under the sampling model and improves, to the best of our knowledge, on the state-of-the-art of any algorithmic sample complexity result for low-rank matrix completion—albeit, with the caveat that, unlike many other results, our guarantee is inherently *local.* Table 3.1 shows our result with respect to other recent ones:

Table 3.1: Comparison of sample complexity for different state-of-the-art algorithm

Name of the algorithm	**Suff. condition on m for convergence**
Nuclear Norm Min. [Rec11, Che15]	$\Omega(\mu_0 \text{deg}_{X_0} \log^2 D)$
OptSpace [KMO10]	$\Omega(\mu_0 \kappa^2 \text{deg}_{X_0} \max(\log D, \kappa^4 r))$
AltMin [HW14]	$\Omega(\mu_0^2 \log(\kappa) r^8 \text{deg}_{X_0} \log^2 D)$
GD on matrix factorization [CLL20]	$\Omega(\mu_0^2 \kappa^{14} r^2 \log D)$
ScaledGD [TMC21]	$\Omega(\mu_0 \kappa^2 r \text{deg}_{X_0} \max(\log D, \mu_0 \kappa^2))$
MatrixIRLS (our result)	$\Omega(\mu_0 \text{deg}_{X_0} \log D)$ (only local convergence)
Necessary condition [CT10]	$\Omega(\mu_0 \text{deg}_{X_0} \log D)$

As one can see in Table 3.1, essentially all algorithms have a strong dependency on the condition number. The only exception is the alternating minimization scheme of [HW14]. This algorithm's trade-off is that its sample complexity strongly depends on the rank of the matrix. The algorithm developed in this chapter, again based on an interesting application of least squares but this time applied to non-convex functions attains the optimal necessary condition. The caveat is that the theory provided here is of a local nature, and global convergence results for our method still remain elusive.

Nevertheless, we obtain very competitive numerical results, and our algorithm can be implemented in a sub-quadratic per-iteration cost in D, without the need to store dense $d_1 \times d_2$ matrices. Also, by assuming a random sampling model, the linear systems to be solved in the main computational step of `MatrixIRLS` are well-conditioned, even close to the ground truth, unlike the systems of comparable IRLS algorithms in the literature

[DDFG10, FRW11b, MF12a, KS18]. Our method's data efficiency and scalability compared to several state-of-the-art methods are finally explored in numerical experiments involving simulated data.

3.2 MatrixIRLS for log-det rank surrogate

The starting point of the derivation of our method is the observation that minimizing a *non-convex surrogate* objective F with more regularity than $\text{rank}(\mathbf{X})$ can lead to effective methods for solving (3.2) that may combine some of the aforementioned properties, e.g., if F is chosen as a *log-determinant* [Faz02, CESV15], Schatten-p quasi-norm (with $0 < p < 1$) [GVRH20] or a smoothed clipped absolute deviation (SCAD) of the singular values [MSW20]. In particular, it has been observed in several works [Faz02, CESV15] that optimizing the smoothed log-det objective $\sum_{i=1}^{d} \log(\sigma_i(\mathbf{X} + \epsilon \mathbf{I}))$ for some $\epsilon > 0$ can lead to less biased solutions than a nuclear norm minimizer—very generally, it can be shown that a minimizer of non-convex spectral functions, i.e., functions of matrices that depend only on their singular values or on their eigenvalues [Bec17, Chapter 7], such as the smoothed log-det objective coincides as least as often with the rank minimizer as the convex nuclear norm minimizer [Fou18, Corollary 3]. The rationale behind this is that, on the one hand, the function rank can be seen as a limit of the Schatten-p norms, but, on the other hand, with the correct scaling, this limit converges to the log-det function. Indeed, for a matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ of rank-r with singular values $\{\sigma_i\}_{i=1}^r$, we have

$$\begin{aligned}\log\det(\mathbf{X}) &= \log\left(\prod_{i=1}^{r} \sigma_i(\mathbf{X})\right) = \sum_{i=1}^{r} \log \sigma_i(\mathbf{X}) = \sum_{i=1}^{r} \lim_{p\to 0} \frac{(\sigma_i(\mathbf{X}))^p - 1}{p} \\ &= \lim_{p\to 0} \sum_{i=1}^{r} \frac{(\sigma_i(\mathbf{X}))^p - 1}{p} = \lim_{p\to 0} \frac{\|\mathbf{X}\|_{S_p}^p}{p} - \frac{r}{p}\end{aligned} \tag{3.7}$$

On the other hand, we have $\lim_{p\to 0} \|\mathbf{X}\|_{S_p}^p = \text{rank}(\mathbf{X})$. Therefore, since the minimizer $\mathbf{X}$ of the function $\log\det(\mathbf{X})$ does not depend on p and r, the calculation above indicates that the minimizes of both functions are connected.

Relevant algorithmic approaches to minimize non-convex rank surrogates include iterative thresholding methods [MSW20], iteratively reweighted least squares [FRW11b, MF12a, KS18] and iteratively reweighted nuclear norm [LTYL15] algorithms. However, finding the global minimizer of a non-convex and non-smooth rank surrogate can be very challenging, as the existence of sub-optimal local minima and saddle points might deter the success of many local optimization approaches. Furthermore, applications such as in recommender systems [KBV09] require solving very high-dimensional problem instances so that it is impossible to store full matrices, let alone to calculate many singular values of these matrices, ruling out the applicability of many of the existing methods for non-convex surrogates. A major shortcoming is that the available convergence theory for such algorithms is still very immature—a convergence theory quantifying the sample complexity or convergence rates is, to the best of our knowledge, not available for any method of this class.

To derive our method, let now $\epsilon > 0$ and $F_\epsilon : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ be the *smoothed log-det objective* – see Figure 3.2 – defined as $F_\epsilon(\mathbf{X}) := \sum_{i=1}^{d} f_\epsilon(\sigma_i(\mathbf{X}))$ with $d = \min(d_1, d_2)$ and

$$f_\epsilon(\sigma) = \begin{cases} \log|\sigma|, & \text{if } \sigma \geq \epsilon, \\ \log(\epsilon) + \frac{1}{2}\Big(\frac{\sigma^2}{\epsilon^2} - 1\Big), & \text{if } \sigma < \epsilon. \end{cases} \tag{3.8}$$

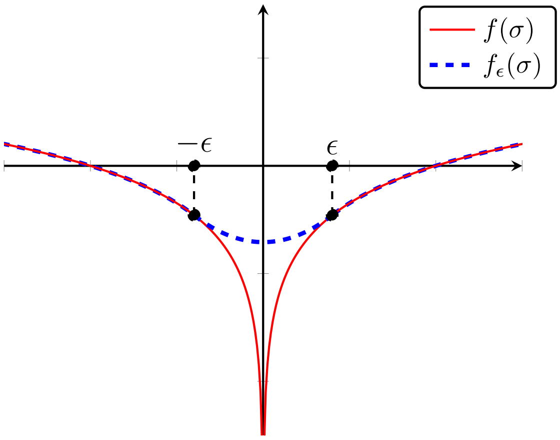


Figure 3.2: Smooth approximation $f_\epsilon(\sigma)$ for $f(\sigma) = \log(|\sigma|)$

It can be shown – see [LS05, Section 7] – that that F_ϵ is continuously differentiable with ϵ^{-2}-Lipschitz gradient given by ,

$$\nabla F_{\epsilon_k}(\mathbf{X}) = \mathbf{U}\,\mathrm{diag}\left(\frac{\sigma_i(\mathbf{X})}{\max(\sigma_i(\mathbf{X}), \epsilon_k)^2}\right)_{i=1}^{d} \mathbf{V}^*,$$

where $\mathbf{X}$ has a singular value decomposition $\mathbf{X} = \mathbf{U}\,\mathrm{diag}\big(\sigma(\mathbf{X})\big)\mathbf{V}^* = \mathbf{U}\,\mathrm{diag}\big(\sigma\big)\mathbf{V}^*$. It is clear that the optimization landscape of F_ϵ crucially depends on the smoothing parameter ϵ. Here, we proceed exactly as in Chapter 2 and will create a sequence of quadratic models to be minimized. The difference, however, lies in the complexity of the quadratic model since its construction, in the case of non-convex functions for matrix completion, is much more involved. Instead of minimizing F_{ϵ_k} directly, our method minimizes, for $k \in N$,

$\epsilon_k > 0$ and $\mathbf{X}^{(k)}$ a *quadratic model*

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

under the data constraint $P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{X}_0)$, where $W^{(k)}$ is the following operator, which describes the geometry of the non-convex function and is one of the main contributions of this work.

Definition 3.2.1. *Let $\epsilon_k > 0$ and $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ be a matrix with singular value decomposition $\mathbf{X}^{(k)} = \mathbf{U}_k \operatorname{diag}(\sigma^{(k)}) \mathbf{V}_k^*$, i.e., $\mathbf{U}_k \in \mathbb{R}^{d_1 \times d_1}$ and $\mathbf{V}_k \in \mathbb{R}^{d_2 \times d_2}$ are orthonormal matrices. Then we call the linear operator $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ the* weight operator *of the ϵ_k-smoothed log-det objective F_{ϵ_k} of* (3.8) *at $\mathbf{X}^{(k)}$ if for $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$,*

$$W^{(k)}(\mathbf{Z}) = \mathbf{U}_k \left[\mathbf{H}_k \circ (\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k) \right] \mathbf{V}_k^*, \tag{3.9}$$

where $\mathbf{H}_k \circ (\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k)$ denotes the entrywise product of $\mathbf{H}_k$ and $\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k$ and $\mathbf{H}_k \in \mathbb{R}^{d_1 \times d_2}$ is a matrix with positive entries such that $(\mathbf{H}_k)_{ij} := \Big(\max(\sigma_i^{(k)}, \epsilon_k) \max(\sigma_j^{(k)}, \epsilon_k) \Big)^{-1}$.

We note that *Iteratively Reweighted Least Squares (IRLS)* methods with certain similarities to Algorithm 3 had been proposed [FRW11b, MF12a, KS18] for the minimization of Schatten-p quasi-norms for $0 < p \leq 1$. The difference, however, lies in the definition of the matrix $(\mathbf{H}_k)_{ij}$. For example, the work [FRW11b] chose $\mathbf{H}_k := \operatorname{diag}(1/[(\sigma_i^{(k)})^2 + \epsilon_k^2])$. By doing so, they showed that the sequence generated by their IRLS-type algorithm converges to stationary points of the underlying smoothed functional. However, the authors were not able to show convergence rate results for the non-convex case of the Schatten-p norm, with $p < 1$.

One of the crucial properties for the design of the correct sequence of least squares problems, i.e., the design of the parabolas that imitate the geometry of the non-convex function, is to have $W^{(k)}(\mathbf{Z}) = \nabla F_{\epsilon_k}(\mathbf{X})$. This property ensures quadratic bounds that mimic the landscape of the function that is being majorized.

Comparing the gradients of smoothed Schatten-p quasi-norms and of (3.8), minimizing a smoothed log-det objective can be considered as a limit case for $p \to 0$ as shown in Equation (3.7). Most importantly, however, our algorithm has two distinct conceptual differences compared to these methods: Firstly, the weight operator of Definition 3.2.1 has the crucial property for the design of the correct sequence of least squares problems, namely, it is able to capture the *second-order information* of F_{ϵ_k}, allowing for an interpretation of `MatrixIRLS` as a saddle-escaping smoothing Newton method, cf. Section 3.3.2, unlike the methods of [FRW11b, MF12a, KS18] due to the different structure

Algorithm 3 `MatrixIRLS` for low-rank matrix completion

```
Input: Set $\Omega$, observations $\mathbf{y} \in \mathbb{R}^m$, rank estimate $\widetilde{r}$.
Initialize $k = 0$, $\epsilon^{(0)} = \infty$ and $W^{(0)} = \mathrm{Id}$.
for $k = 1$ to $K$ do
    Solve weighted least squares: Use a conjugate gradient method to solve
        $\mathbf{X}^{(k)} = \underset{\mathbf{X}:P_\Omega(\mathbf{X})=\mathbf{y}}{\arg\min} \langle \mathbf{X}, W^{(k-1)}(\mathbf{X})\rangle.$    (3.10)
    Update smoothing: Compute $\widetilde{r}+1$-th singular value of $\mathbf{X}^{(k)}$ to update
        $\epsilon_k = \min\left(\epsilon_{k-1}, \sigma_{\widetilde{r}+1}(\mathbf{X}^{(k)})\right).$    (3.11)
    Update weight operator: For $r_k := |\{i \in [d] : \sigma_i(\mathbf{X}^{(k)}) > \epsilon_k\}|$, compute the first $r_k$
    singular values $\sigma_i^{(k)} := \sigma_i(\mathbf{X}^{(k)})$ and matrices $\mathbf{U}^{(k)} \in \mathbb{R}^{d_1 \times r_k}$ and $\mathbf{V}^{(k)} \in \mathbb{R}^{d_2 \times r_k}$ with
    leading $r_k$ left/ right singular vectors of $\mathbf{X}^{(k)}$ to update $W^{(k)}$ defined in Equation (3.9).
end for
Output: $\mathbf{X}^{(K)}$.
```

of their weight operators. Secondly, the interplay of F_{ϵ_k} and the weight operator $W^{(k)}$ in Algorithm 3 is designed to allow for efficient numerical implementations, cf. Section 3.6.

The weight operator $W^{(k)}$ is a positive, self-adjoint operator with strictly positive eigenvalues that coincide with the entries of the matrix $\mathbf{H}_k \in \mathbb{R}^{d_1 \times d_2}$, and it also verifies that $W^{(k)}(\mathbf{X}^{(k)}) = \nabla F_{\epsilon_k}(\mathbf{X}^{(k)})$. Based on this, it follows that the minimization of the quadratic model $Q_{\epsilon_k}(\mathbf{X}|\mathbf{X}^{(k)})$ boils down to a minimization of a quadratic form weighted by $W^{(k)}$. This enables us to design the iterative method *Matrix Iteratively Reweighted Least Squares (MatrixIRLS)*, which we describe in Algorithm 3.

Apart from the weighted least squares step (3.10), which minimizes the quadratic model $Q_{\epsilon_{k-1}}(\cdot|\mathbf{X}^{(k-1)})$ of $F_{\epsilon_{k-1}}$ for fixed ϵ_{k-1}, an indispensable ingredient of our scheme is the *update of the smoothing parameter* ϵ_k, which is performed in the spirit of smoothing methods for non-smooth objectives [Che12]. In particular, the update rule (3.11), which is similar to the update rule of [KS18], makes sure that if the rank estimate $\widetilde{r}$ is chosen such that $\widetilde{r} \geq r$, the smoothing parameter ϵ_k converges to 0 as the iterates approach a rank-r solution.

Finally, we note that it is non-trivial to show that the quadratic model $Q_{\epsilon_k}(\cdot|\mathbf{X}^{(k)})$ induced by $W^{(k)}$ from Definition 3.2.1 is actually a *majorant* of $F_{\epsilon_k}(\cdot)$ such that $F_{\epsilon_k}(\mathbf{X}) \leq Q_{\epsilon_k}(\mathbf{X}|\mathbf{X}^{(k)})$ for all $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$. A proof of this fact, as well as a proof of the optimality of the majorant, will be the subject of an upcoming paper. It is already available in Dr. Kümmerle thesis [Küm19, Theorem 2.4].

3.3 How to interpret MatrixIRLS?

Despite its conceptual simplicity, `MatrixIRLS` has in its core an intricate construction of the weight operator. A brief comparison with the algorithm developed in Chapter 2 shows that the formula $W^{(k)}(\mathbf{Z}) = \mathbf{U}_k \left[\mathbf{H}_k \circ (\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k)\right] \mathbf{V}_k^*$ deserves a deeper explanation.

A common trace of all modern versions of IRLS is that the weights of quadratic form match the first-order derivative of the non-convex objective function to be minimized, i.e., $W^{(k)}(\mathbf{X}^{(k)}) = \nabla F_{\epsilon_k}(\mathbf{X}^{(k)})$. This holds for our algorithm, `MatrixIRLS`, as well as for previous versions of IRLS applied to non-convex spectral functions [FRW11b, MF12a, KS18], see also a general discussion in [ODBP15, Section 5]. However, this local property, together with the local quadratic convergence rate that will be established in Section 3.4, does not explain the numerically observed global convergence behavior, see Section 3.7, which is remarkable due to the non-convexity of the objective function.

These different versions of the IRLS algorithm applied to matrix problems can be interpreted as a variable metric forward-backward method, as we discuss in Section 3.3.1. However, unlike previous contributions, our method can also be interpreted in a different way. In the next two sections, we will show how `MatrixIRLS` can be explained either as a *variable metric forward-backward method* or as a *saddle-escaping smoothing Newton method.* This will help to shed some light on its numerical performance.

3.3.1 MatrixIRLS as variable metric forward-backward method

An instructive angle to understand our method comes from the framework of *variable metric forward-backward methods* [BGLS95, CPR14, FGP15], which can be seen as a combination of a gradient descent method and a proximal point algorithm [CP11] that can be used to minimize the sum of a non-smooth function and a function with Lipschitz continuous gradients. In particular, if F is a proper, lower semi-continuous function, G is differentiable with Lipschitz gradient ∇G and $(\alpha_k)_k$ a sequence of step sizes, the iterations of the forward-backward algorithm [ABS13] are such that $\mathbf{X}^{(k+1)} \in \operatorname{prox}_{\alpha_k F}\left(\mathbf{X}^{(k)} - \alpha_k \nabla G(\mathbf{X}^{(k)})\right)$, where $\operatorname{prox}_{\alpha_k F}(\cdot)$ is the proximity operator of $\alpha_k F$. Typically, in such an algorithm, F would be chosen as the structure-promoting objective (such as the smoothed log-det objective F_ϵ above) and G as a data-fit term such as $G(\mathbf{X}) = \|P_\Omega(\mathbf{X}) - \mathbf{y}\|_2^2/\lambda$, leading to thresholding-type algorithms such as the FISTA [BT09]. Algorithm 3, fits into this framework if we transform the constrained problem into an unconstrained one with the help of an indicator function. More specifically, if we choose, for $\epsilon_k > 0$, the non-smooth part F as the indicator function $F := \chi_{P_\Omega^{-1}(\mathbf{y})} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ of the constraint set $P_\Omega^{-1}(\mathbf{y}) := \{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2} : P_\Omega(\mathbf{X}) = \mathbf{y}\}$ and the smooth part G such that $G := F_{\epsilon_k} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ as in (3.8), while offsetting the distortion induced by the non-Euclidean nature of the level sets of F_{ϵ_k} via an appropriate choice of a *variable metric* $d_{A_k}(\mathbf{X},\mathbf{Z}) = \sqrt{\langle \mathbf{X}-\mathbf{Z}, A_k(\mathbf{X}-\mathbf{Z})\rangle_F}$ for a positive definite linear operator $A_k : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$, such that

$$\mathbf{X}^{(k+1)} \in \operatorname{prox}_{\alpha_k F}^{A_k}\left(\mathbf{X}^{(k)} - \alpha_k A_k^{-1}(\nabla G(\mathbf{X}^{(k)}))\right),$$

where $\operatorname{prox}_F^{A_k}(\mathbf{X}) := \arg\min_{\mathbf{Z}\in\mathbb{R}^{d_1\times d_2}} F(\mathbf{Z}) + \frac{1}{2}d_{A_k}(\mathbf{X},\mathbf{Z})^2$ is the proximity operator of F *scaled in the metric* d_{A_k} at $\mathbf{X}$ [CPR14]. Specifically, if we choose the metric induced by the weight operator of (3.9) such that $A_k := W^{(k)}$ and unit step sizes $\alpha_k = 1$, we obtain

$$\begin{aligned}
&\operatorname{prox}_{\alpha_k F}^{A_k}\left(\mathbf{X}^{(k)} - \alpha_k A_k^{-1}(\nabla G(\mathbf{X}^{(k)}))\right) \\
&= \operatorname{prox}_{\chi_{P_\Omega^{-1}}}^{W^{(k)}}\left(\mathbf{X}^{(k)} - W_k^{-1}(\nabla F_{\epsilon_k}(\mathbf{X}^{(k)}))\right) \\
&= \operatorname{prox}_{\chi_{P_\Omega^{-1}}}^{W^{(k)}}\left(\mathbf{X}^{(k)} - W_k^{-1}W_k(\mathbf{X}^{(k)}))\right) = \operatorname{prox}_{\chi_{P_\Omega^{-1}}}^{W^{(k)}}(\mathbf{0}) \\
&= \underset{\mathbf{X}:P_\Omega(\mathbf{X})=\mathbf{y}}{\arg\min}\ \frac{1}{2}d_{A_k}(\mathbf{X},\mathbf{0})^2 = \underset{\mathbf{X}:P_\Omega(\mathbf{X})=\mathbf{y}}{\arg\min}\ \langle \mathbf{X}, W^{(k)}(\mathbf{X})\rangle,
\end{aligned}$$

where we used that $W_k(\mathbf{X}^{(k)}) = \nabla F_{\epsilon_k}(\mathbf{X}^{(k)})$ in the third line. This shows that this update rule for $\mathbf{X}^{(k+1)}$ coincides with (3.10). Thus, `MatrixIRLS` can be considered as a forward-backward method with a variable metric induced by the weight operator $W^{(k)}$, using a unit step size $\alpha_k = 1$ for each k. One advantage of our method is that, unlike many methods in this family, there is no step size to be tuned. A crucial difference, which makes the existing theory for splitting methods, e.g., [FGP15], not directly applicable for the convergence analysis of `MatrixIRLS`, is that the smooth function $G = F_{\epsilon_k}$ *is changing* at each iteration due to the smoothing parameter update (3.11). On the other hand, the results of [FGP15] already imply the finite sequence length of $(\mathbf{X}^{(k)})_k$ in the case that the smoothing parameter ϵ_k stagnates for $k \geq k_0$, using a Kurdyka-Lojasiewicz property [BDL07] of $F_{\epsilon_k} + \chi_{P_\Omega^{-1}(\mathbf{y})}$, see [FGP15, Theorem 4.1]. In any case, the theory of convergence for variable-metric algorithms is still in its infancy [BPR21], and we do expect to see new developments in the near future that could be used to further analyze the global convergence of IRLS-type methods.

Finally, we note that previous IRLS methods [FRW11b, MF12a, KS18] would also fit in the presented splitting framework, however, without fully capturing the underlying geometry as their weight operator has no strong connection to the Hessian $\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})$ of F_{ϵ_k}, as explained in the next section.

3.3.2 MatrixIRLS as saddle-escaping smoothing Newton method

Another interesting way to interpret `MatrixIRLS` is as a *saddle-escaping smoothing Newton* method. Smoothing Newton methods minimize a non-smooth and possibly non-convex function F by using derivatives of certain smoothing proxies of F [CQS98, Che12]. Interpreting the optimization problem $\min_{\mathbf{X}:P_\Omega(X)=\mathbf{y}} F_{\epsilon_k}(\mathbf{X})$ as an unconstrained optimization problem over the null space of P_Ω, we can write

$$\begin{aligned}\mathbf{X}^{(k+1)} &= \mathbf{X}^{(k)} - P_{\Omega^c}^* \left(P_{\Omega^c} W^{(k)} P_{\Omega^c}^*\right)^{-1} P_{\Omega^c} W^{(k)}(\mathbf{X}^{(k)}) \\ &= \mathbf{X}^{(k)} - P_{\Omega^c}^* \left(P_{\Omega^c} \overline{\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})} P_{\Omega^c}^*\right)^{-1} P_{\Omega^c} \nabla F_\epsilon(\mathbf{X}^{(k)}),\end{aligned}$$

if $\Omega^c = [d_1] \times [d_2] \setminus \Omega$ corresponds to the unobserved indices, where $\overline{\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ is a *modified* Hessian of F_{ϵ_k} at $\mathbf{X}^{(k)}$ that replaces negative eigenvalues of the Hessian $\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})$ by positive ones and slightly increases small eigenvalues. This line of research, which was initiated [Gre67], see also a discussion about modifications of Newton's Methods [NW06, Chapter 3.4] and about Trust-Region Newton-CG methods in [NW06, Chapter 7], witnessed important recent achievements. For example, in [PMR19a], it has been proved that for a fixed smooth function F_{ϵ_k}, similar modified Newton-type steps are able to escape the first-order saddle points at a rate that is independent of the problem's condition number.

Theorem 3.3.1. *[PMR19b, Theorem 2.2] Let $f(x)$ be a function that is twice continuously differentiable with gradient and Hessian that are Lipschitz continuous. Moreover, suppose that there exists a positive constant B such that $\|x^\dagger\| \leq B$ for all $x^\dagger$ such that $\|\nabla f(x^\dagger)\| = 0$ and $\nabla^2 f(x^\dagger) \succ 0$ and assume that the local minima and saddles are nondegenerate, i.e., all the eigenvalue of the Hessian of f at a local minimum or a saddle point is strictly positive. Let $\varepsilon > 0$ be the desired accuracy of the saddle-escaping smoothing Newton method and $\alpha \in (0,1)$ be one of its inputs. If $m < \xi/2$ and*

$$\|\nabla f_-(x_0)\| \geq \max\{(5L/2m^2)\left\|\nabla f(x_0)\right\|^2, \varepsilon\} \tag{3.12}$$

and $\|\nabla f(x_0)\| \leq \delta/2$, we have that $\|\nabla f(x_{K_1})\| \geq \delta/2$, with $K_1 \leq 1 + \log_{3/2}\left(\frac{\delta}{2\varepsilon}\right)$.

As explained in [PMR19b], the result shows that if the projection of the gradient (at a certain iteration x_0 onto the orthogonal subspace associated with the negative eigenvalues of the Hessian $\nabla^2 f(x^\ddagger)$ satisfies a certain condition, then there is an upper bound for the number of iterations that it takes for the modified Newton-type method to escape the saddle point. In particular, this escaping is of the order $O(log(1/\varepsilon))$. Moreover, they showed that even when the condition Equation (3.12) is not satisfied, adding noise to the iterations ensures that it will be fulfilled with high probability.

Now, regarding `MatrixIRLS`, as it was discussed in the beginning of Section 3.2, if $\epsilon_k > 0$, then $F_{\epsilon_k} : \mathbb{R}^{d_1\times d_2} \to R$ given by the ϵ_k-smoothed log-det objective Equation (3.8) is continuously differentiable with ϵ_k^{-2}-Lipschitz gradient $\nabla F_{\epsilon_k}(\mathbf{X}) = \mathbf{U}\,\mathrm{dg}\left(\frac{\sigma_i(\mathbf{X})}{\max(\sigma_i(\mathbf{X}),\epsilon_k)^2}\right)_{i=1}^{d}\mathbf{V}^*$ for any matrix $\mathbf{X}$ with singular value decomposition $\mathbf{X} = \mathbf{U}\,\mathrm{dg}\big(\sigma(\mathbf{X})\big)\mathbf{V}^* = \mathbf{U}\,\mathrm{dg}\big(\sigma\big)\mathbf{V}^*$. Moreover, it holds that ∇F_{ϵ_k} is differentiable at $\mathbf{X}$ if and only if the second derivative $f''_{\epsilon_k} : \mathbb{R} \to \mathbb{R}$ of f_{ϵ_k} from (3.8) exists at all $\sigma = \sigma_i(\mathbf{X})$, $i \in [d]$, which is the case if $\mathbf{X} \in \mathcal{D}_{\epsilon_k} := \big\{\mathbf{X} : \sigma_i(\mathbf{X}) \neq \epsilon_k \text{ for all } i \in [d]\big\}$. The latter statement follows from the theory of *non-Hermitian Löwner functions* [Yan09, DSST18][5], as $\mathbf{X} \mapsto \nabla F_{\epsilon_k}(\mathbf{X})$ is such a function.

Let $\mathbf{X}^{(k)} \in \mathcal{D}_{\epsilon_k} := \big\{\mathbf{X} : \sigma_i(\mathbf{X}) \neq \epsilon_k \text{ for all } i \in [d]\big\}$ with singular value decomposition given by

$$\mathbf{X}^{(k)} = \mathbf{U}_k \,\mathrm{diag}(\sigma^{(k)})\mathbf{V}_k^* = \begin{bmatrix}\mathbf{U}^{(k)} & \mathbf{U}_\perp^{(k)}\end{bmatrix}\begin{bmatrix}\mathbf{\Sigma}^{(k)} & 0\\ 0 & \mathbf{\Sigma}_\perp^{(k)}\end{bmatrix}\begin{bmatrix}\mathbf{V}^{(k)*}\\ \mathbf{V}_\perp^{(k)*}\end{bmatrix}, \tag{3.13}$$

where $\mathbf{U}_k \in \mathbb{R}^{d_1\times d_1}$ and $\mathbf{V}_k \in \mathbb{R}^{d_2\times d_2}$, and corresponding submatrices $\mathbf{U}^{(k)} \in \mathbb{R}^{d_1\times r_k}$, $\mathbf{U}_\perp^{(k)} \in \mathbb{R}^{d_1\times(d_1-r_k)}$, $\mathbf{V}^{(k)} \in \mathbb{R}^{d_2\times r_k}$, $\mathbf{V}_\perp^{(k)} \in \mathbb{R}^{d_2\times(d_2-r_k)}$, $\mathbf{\Sigma}^{(k)} := \mathrm{diag}(\sigma_1^{(k)}, \ldots \sigma_{r_k}^{(k)})$ and $\mathbf{\Sigma}_\perp^{(k)} := \mathrm{dg}(\sigma_{r_k+1}^{(k)}, \ldots \sigma_d^{(k)})$, and $r_k := |\{i \in [d] : \sigma_i(\mathbf{X}^{(k)}) > \epsilon_k\}| = |\{i \in [d] : \sigma_i^{(k)} > \epsilon_k\}|$. In this case, it can be calculated that the Hessian $\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})$ at $\mathbf{X}^{(k)}$, which is a function that maps $\mathbb{R}^{d_1\times d_2}$ to $\mathbb{R}^{d_1\times d_2}$ matrices, satisfies, in the case of $d_1 = d_2$,

$$\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})(\mathbf{Z}) = \mathbf{U}_k\left[\mathbf{M}^{\mathrm{S}} \circ S(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k) + \mathbf{M}^{\mathrm{T}} \circ T(\mathbf{U}_k^*\mathbf{Z}\mathbf{V}_k)\right]\mathbf{V}_k^*, \tag{3.14}$$

for any $\mathbf{Z} \in \mathbb{R}^{d_1\times d_2}$, where $S : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ and $T : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ are the *symmetrization operator* and *antisymmetrization operator*, respectively, that map any $\mathbf{X} \in \mathbb{R}^{d\times d}$ to

$$S(\mathbf{X}) = \frac{1}{2}(\mathbf{X} + \mathbf{X}^*), \quad \text{and} \quad T(\mathbf{X}) = \frac{1}{2}(\mathbf{X} - \mathbf{X}^*)$$

for any $\mathbf{X} \in \mathbb{R}^{d\times d}$, and $\mathbf{M}^{\mathrm{S}}, \mathbf{M}^{\mathrm{T}} \in \mathbb{R}^{d_1\times d_2}$ fulfill

$$\mathbf{M}^{\mathrm{S}} = \left[\begin{array}{c|c} -\mathbf{H}^{(k)} & \mathbf{M}_{1,2}^{-} \\ \hline \mathbf{M}_{2,1}^{-} & \epsilon_k^{-2}\mathbf{1}\end{array}\right] \qquad \mathbf{M}^{\mathrm{T}} = \left[\begin{array}{c|c} -\mathbf{H}^{(k)} & \mathbf{M}_{1,2}^{+} \\ \hline \mathbf{M}_{2,1}^{+} & \epsilon_k^{-2}\mathbf{1}\end{array}\right],$$

the matrix $\mathbf{H}^{(k)} \in \mathbb{R}^{r_k\times r_k}$ satisfies

$$\mathbf{H}_{ij}^{(k)} = \left(\sigma_i^{(k)}\sigma_j^{(k)}\right)^{-1} \text{ for all } i, j \in [r_k], \tag{3.15}$$

[5] These functions can also be called *generalized matrix functions* in the literature [Nof17].

and the $(d_1 - r_k) \times (d_2 - r_k)$ $\mathbf{1}$ matrix where all entries are equal to one. Furthermore, the matrices $\mathbf{M}_{1,2}^{-}, \mathbf{M}_{1,2}^{+} \in (d_1 - r_k) \times r_k$ are such that

$$\left(\mathbf{M}_{1,2}^{\pm}\right)_{ij} = \frac{(\sigma_i^{(k)})^{-1} \pm \sigma_{j+r_k}^{(k)} \epsilon_k^{-2}}{\sigma_i^{(k)} \pm \sigma_{j+r_k}^{(k)}} \quad \text{for} i \in [r_k], j \in [d_2 - r_k] \text{ and}$$

$$\left(\mathbf{M}_{2,1}^{\pm}\right)_{ij} = \frac{(\sigma_j^{(k)})^{-1} \pm \sigma_{i+r_k}^{(k)} \epsilon_k^{-2}}{\sigma_j^{(k)} \pm \sigma_{i+r_k}^{(k)}} \quad \text{for} j \in [r_k], i \in [d_1 - r_k].$$

The formula (3.14) for $\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})$ follows by inserting the operator ∇F_{ϵ_k} into Theorem 2.2.6 of [Yan09], Corollary 3.10 [Nof17] or Theorem 4 of [DSST18]. By realizing that $0 \leq \sigma_\ell^{(k)} \leq \epsilon_k$ for all $\ell > r_k$, we see that

$$\frac{1}{(\sigma_i^{(k)})^2} \leq \left(\mathbf{M}_{1,2}^{+}\right)_{ij} = \left(\mathbf{M}_{2,1}^{+}\right)_{ji} \leq \frac{1}{\sigma_i^{(k)} \epsilon_k} \text{ and } -\frac{1}{\sigma_i^{(k)} \epsilon_k} \leq \left(\mathbf{M}_{1,2}^{-}\right)_{ij} = \left(\mathbf{M}_{2,1}^{-}\right)_{ji} \leq \frac{1}{(\sigma_i^{(k)})^2}$$

for all i and j.

To explain the connection to $\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})(\mathbf{Z})$ and to explain the formula for the weighting operator, i.e., $W^{(k)}(\mathbf{Z}) = \mathbf{U}_k \left[\mathbf{H}_k \circ (\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k)\right] \mathbf{V}_k^*$, we need to understand the role of $\mathbf{H}_k \in \mathbb{R}^{d_1 \times d_2}$. Recall that the entries of this matrix were defined such that, for all i and j, we have $(\mathbf{H}_k)_{ij} := \Big(\max(\sigma_i^{(k)}, \epsilon_k) \max(\sigma_j^{(k)}, \epsilon_k)\Big)^{-1}$. Then, we write

$$\mathbf{H}_k = \begin{bmatrix} \mathbf{H}^{(k)} & \mathbf{H}_{1,2}^{(k)} \\ \mathbf{H}_{2,1}^{(k)} & \epsilon_k^{-2}\mathbf{1}, \end{bmatrix} \tag{3.16}$$

where the matrices $\mathbf{H}^{(k)} \in \mathbb{R}^{r_k \times r_k}$ was defined above in Equation (3.15) and $\mathbf{H}_{1,2}^{(k)} \in \mathbb{R}^{r_k \times (d_2 - r_k)}$ and $\mathbf{H}_{2,1}^{(k)} \in \mathbb{R}^{(d_1 - r_k) \times r_k}$ are such that $\left(\mathbf{H}_{1,2}^{(k)}\right)_{ij} = \left(\sigma_i^{(k)} \epsilon_k\right)^{-1}$ for all $i \in [r_k]$ and $j \in [d_2 - r_k]$ and $\left(\mathbf{H}_{2,1}^{(k)}\right)_{ij} = \left(\epsilon_k \sigma_j^{(k)}\right)^{-1}$ for all $i \in [d_1 - r_k]$ and $j \in [r_k]$.

Now, comparing M^{S} and M^{T} with $\mathbf{H}_k$, the core of the weight operator $W^{(k)}$, we see that the upper left blocks of M^{S} and M^{T} are just the *negative* of the upper left block $\mathbf{H}^{(k)}$ of $\mathbf{H}_k$, while the lower right blocks coincide. Furthermore, the lower left and the upper right blocks are related such that

$$\left|\left(\mathbf{M}_{1,2}^{\pm}\right)_{ij}\right| \leq \frac{1}{\sigma_i^{(k)} \epsilon_k} = (\mathbf{H}_{1,2}^{(k)})_{ij} \text{ for all } i \in [r_k], j \in [d_2 - r_k], \tag{3.17}$$

and

$$\left|\left(\mathbf{M}_{2,1}^{\pm}\right)_{ij}\right| \leq \frac{1}{\sigma_j^{(k)} \epsilon_k} = (\mathbf{H}_{2,1}^{(k)})_{ij} \text{ for all} i \in [d_2 - r_k], j \in [r_k]. \tag{3.18}$$

We can finally shed some light on the relationship between the above considerations and the analysis performed in [PMR19a]. In their paper, the authors assumed to have a smooth function F_{ϵ_k} that needs to be minimized in the framework of *unconstrained minimization.* Under these assumptions, [PMR19a] considers using *modified Newton steps*

$$\mathbf{X}^{(k+1)} := \mathbf{X}^{(k)} - \eta_k \left|\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})\right|_c^{-1} \nabla F_{\epsilon_k}(\mathbf{X}^{(k)})$$

where the Hessian $\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})$ is replaced by a positive definite truncated eigenvalue matrix $\left|\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})\right|_c$, which replaces the large negative eigenvalues of $\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})$ by their modulus for eigenvalues that have large modulus and eigenvalues of small modulus by an appropriate constant c. Then, the paper [PMR19a] shows – see Theorem 3.3.1 above – that such steps are, unlike conventional Newton steps (which are often *attracted* by saddle points), able to *escape* saddle points with an exponential rate that does *not* depend on the conditioning of the problem. Experimental observations of such behavior have also been reported in other works [DPG$^+$14]. See also the discussion in [PDGB14, Section 4].

In view of this, we observe that the weight operator $W^{(k)}$ is nothing but a refined variant of $\left|\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})\right|_c$, as the eigenvalues of $\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})$ from (3.14) are simply $\{(\mathbf{M}^{\mathrm{S}}_{ij}, i \leq j\} \cup \{(\mathbf{M}^{\mathrm{T}}_{ij}, i < j\}$, c.f., [Nof17, Theorem 4.5]. In particular, the refinement is such that the small eigenvalues of $\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})$, which can be found in the entries of $\mathbf{M}^{\pm}_{1,2}$ and $\mathbf{M}^{\pm}_{2,1}$, are replaced not by a uniform constant, but by *different* upper bounds $(\sigma_i^{(k)}\epsilon_k)^{-1}$ and $(\sigma_j^{(k)}\epsilon_k)^{-1}$ that depend either on the row index i or the column index j.

Besides this connection, important differences exist between our algorithm and the algorithm analyzed in [PMR19a]. While that paper considers the minimization of a fixed smooth function, we update the smoothing parameter ϵ_k and thus the function F_{ϵ_k} at each iteration. Furthermore, Algorithm 1 of [PMR19a] uses backtracking for each modified Newton step, which would be prohibitive to perform as evaluations of F_{ϵ_k} are very expensive for our smoothed log-det objectives, as they would require the calculation of all singular values. On the other hand, `MatrixIRLS` uses fully modified Newton steps, and we can assure that these are always a descent direction in our case, as explained in [Küm19]. Lastly, we do not add noise to the iterates as [PMR19b] in order to make the condition Equation (3.12) hold with high probability. We believe it is an interesting problem to make this connection rigorous since this could boost the use of methods containing information about the Hessian in large-scale machine learning problems.

Open Problem: Develop a rigorous analysis, along the same lines of [PMR19b], for constrained non-smooth problems with a variable metric formulation.

As mentioned in Section 3.2, `MatrixIRLS` is by no means the *first* algorithm for low-rank matrix recovery that can be considered as an iteratively reweighted least squares algorithm. However, the IRLS algorithms [FRW11b, MF12b, LXY13, KS18] are different from `MatrixIRLS` not only in their computational aspects and efficiency but also since they do *not* allow for a close relationship between their weight operator $W^{(k)}$ and the Hessian $\nabla^2 F_{\epsilon_k}(\mathbf{X}^{(k)})$ at $\mathbf{X}^{(k)}$ as described above.

3.4 Local Convergence with Quadratic Rate

In this section, we will finally establish the quadratic rate of for `MatrixIRLS`. Before formally stating our result, we need to state a few facts about the geometry of the set of matrices of rank-r

$$\mathbb{R}_r^{d_1 \times d_2} = \{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2} \,|\, \mathrm{rank}(\mathbf{X}) = r\}. \tag{3.19}$$

A nice introduction to the topic of Riemannian Optimization can be found in, e.g., [UV20, Bou23]. Indeed, for every $0 \leq r \leq \min(d_1, d_2)$, this set is an embedded submanifold of $\mathbb{R}_r^{d_1 \times d_2}$ of dimension $r(d_1 + d_2 - r)$, see [Bou23, Chapter 7.5]. Every vector tangent that is tangent to this space belongs to the direct sum of the row and column spaces, i.e., the tangent space is the linear space spanned by elements of the form $u_k x^*$ and $y v_k^*$, $1 \leq k \leq r$, where u_k and v_k are singular vectors of $\mathbf{X}$ and x and y are arbitrary. The tangent space of $\mathbb{R}_r^{d_1 \times d_2}$ at a point $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^*$ can, alternatively, be represented by

$$\begin{aligned} T_X(\mathbb{R}_r^{d_1 \times d_2}) &:= \left\{ \begin{bmatrix} \mathbf{U}\mathbf{U}_\perp \end{bmatrix} \begin{bmatrix} \mathbb{R}^{r_k \times r_k} & \mathbb{R}^{r_k(d_2 - r_k)} \\ \mathbb{R}^{(d_1 - r_k) r_k} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}\mathbf{V}_\perp \end{bmatrix}^* \right\} \\ &= \left\{ \begin{bmatrix} \mathbf{U}\mathbf{U}_\perp \end{bmatrix} \begin{bmatrix} \mathbf{M}_1 \mathbf{M}_2 \\ \mathbf{M}_3 \;\mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}\mathbf{V}_\perp \end{bmatrix}^* : \mathbf{M}_1 \in \mathbb{R}^{r_k \times r_k}, \mathbf{M}_2 \in \mathbb{R}^{r_k \times (d_2 - r_k)}, \mathbf{M}_3 \in \mathbb{R}^{(d_1 - r_k) \times r_k} \right\} \\ &= \Big\{ \mathbf{U}\Gamma_1\mathbf{V}^* + \mathbf{U}\Gamma_2\big(\mathbf{I} - \mathbf{V}\mathbf{V}^*\big) + \big(\mathbf{I} - \mathbf{U}\mathbf{U}^*\big)\Gamma_3\mathbf{V}^* : \Gamma_1 \in \mathbb{R}^{r \times r}, \Gamma_2 \in \mathbb{R}^{r \times d_2}, \Gamma_3 \in \mathbb{R}^{d_1 \times r} \Big\}, \end{aligned} \tag{3.20}$$

where $\Gamma_1 \in \mathbb{R}^{r \times r}, \Gamma_2 \in \mathbb{R}^{r \times d_2}, \Gamma_3 \in \mathbb{R}^{d_1 \times r}$ and it holds that $\Gamma_2 \mathbf{V} = 0$ and $\mathbf{U}^*\Gamma_3 = 0$. Here we have represented the tangent space by a decomposition into three mutually orthogonal subspaces represented by the three matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$. Alternatively, we can also represent it with smaller matrices Γ_1, Γ_2, and Γ_3 that are more suitable to be used for calculations. The orthogonal projection of a matrix $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$ onto $T_X(\mathbb{R}_r^{d_1 \times d_2})$, now denote by T, can be obtained by projecting separately onto each of these subspaces. The formula for $\mathcal{P}_T : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ is given by [AO15],

$$\mathcal{P}_T(\mathbf{Z}) = \mathcal{P}_U \mathbf{Z} \mathcal{P}_V + (\mathbf{Z} - \mathcal{P}_U \mathbf{Z})\mathcal{P}_V + \mathcal{P}_U(\mathbf{Z}^* - \mathcal{P}_V \mathbf{Z}^*)^* = \mathbf{U}\mathbf{U}^*\mathbf{Z} + \mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^*.$$

Also, the projection onto the perpendicular space is given by

$$\mathcal{P}_{T^\perp} = (\mathbf{I} - \mathcal{P}_T)(\mathbf{Z}) = (\mathbf{I} - \mathbf{U}\mathbf{U}^*)\mathbf{Z}(\mathbf{I} - \mathbf{V}\mathbf{V}^*).$$

If $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$ are the left and right singular matrices corresponding to the r non-zero singular values of $\mathbf{X}$. With the tangent space of the rank-r manifold in mind, we can formulate the weight operator $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ of Definition 3.2.1 in a different way. Recalling that $\mathbf{H}_k \in \mathbb{R}^{d_1 \times d_2}$ is such that $(\mathbf{H}_k)_{ij} := \Big(\max(\sigma_i^{(k)}, \epsilon_k)\max(\sigma_j^{(k)}, \epsilon_k)\Big)^{-1}$ for all i and j, we write for each $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$

$$\begin{aligned}
W^{(k)}(\mathbf{Z}) &= \mathbf{U}_k \left[\mathbf{H}_k \circ (\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k)\right] \mathbf{V}_k^* \\
&= \begin{bmatrix} \mathbf{U}^{(k)} & \mathbf{U}_\perp^{(k)} \end{bmatrix} \left(\mathbf{H}_k \circ \begin{bmatrix} \mathbf{U}^{(k)*}\mathbf{Z}\mathbf{V}^{(k)} & \mathbf{U}^{(k)*}\mathbf{Z}\mathbf{V}_\perp^{(k)} \\ \mathbf{U}_\perp^{(k)*}\mathbf{Z}\mathbf{V}^{(k)} & \mathbf{U}_\perp^{(k)*}\mathbf{Z}\mathbf{V}_\perp^{(k)} \end{bmatrix} \right) \begin{bmatrix} \mathbf{V}^{(k)*} \\ \mathbf{V}_\perp^{(k)*} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{U}^{(k)} & \mathbf{U}_\perp^{(k)} \end{bmatrix} \left(\begin{bmatrix} \mathbf{H}^{(k)} & \mathbf{H}_{1,2}^{(k)} \\ \mathbf{H}_{2,1}^{(k)} & \epsilon_k^{-2}\mathbf{1} \end{bmatrix} \circ \begin{bmatrix} \mathbf{U}^{(k)*}\mathbf{Z}\mathbf{V}^{(k)} & \mathbf{U}^{(k)*}\mathbf{Z}\mathbf{V}_\perp^{(k)} \\ \mathbf{U}_\perp^{(k)*}\mathbf{Z}\mathbf{V}^{(k)} & \mathbf{U}_\perp^{(k)*}\mathbf{Z}\mathbf{V}_\perp^{(k)} \end{bmatrix} \right) \begin{bmatrix} \mathbf{V}^{(k)*} \\ \mathbf{V}_\perp^{(k)*} \end{bmatrix} \\
&= \left(P_{T_k} \mathbf{D}_{S_k} P_{T_k}^* + \epsilon_k^{-2} \left(\mathbf{I} - P_{T_k} P_{T_k}^*\right) \right) \mathbf{Z},
\end{aligned} \tag{3.21}$$

where the matrices $\mathbf{H}^{(k)} \in \mathbb{R}^{r_k \times r_k}$, $\mathbf{H}_{1,2}^{(k)} \in \mathbb{R}^{r_k \times (d_2 - r_k)}$ and $\mathbf{H}_{2,1}^{(k)} \in \mathbb{R}^{(d_1 - r_k) \times r_k}$ are defined in Equation (3.54). Furthermore, $\mathbf{1}$ in the third line is the $((d_1 - r_k) \times (d_2 - r_k))$-matrix of ones $\mathbf{1}$, and $\mathbf{I}$ is the identity operator. Also, the operator $\mathbf{D}_{S_k} : S_k \to S_k$ is defined implicitly through the last equality. We observe that $\mathbf{D}_{S_k}$ is a diagonal matrix with the entries of $\mathbf{H}^{(k)}$, $\mathbf{H}_{1,2}^{(k)}$ and $\mathbf{H}_{2,1}^{(k)}$ enumerated on its diagonal.

In this thesis, we consider the canonical uniform random sampling model studied in [CR09, Rec11, Che15] where the sampling set $\Omega = (i_\ell, j_\ell)_{\ell=1}^m \subset [d_1] \times [d_2]$ consists of m double indices that are drawn uniformly at random without replacement. Not *each* rank-r matrix $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ is expected to be identifiable from a small number of samples m under this sampling model.

The model above is not the only sampling model studied in the literature. In particular, some works studied deterministic patterns and the problem of unique completability of a given sampling pattern, see [PABN16, SXZ18, SC10, Tsa23, LMSH23]. Another interesting recent line of research is to obtain results for when the sampling mask is dependent on the underlying matrix to be retrieved. One can think about a situation where sensors are monitoring the environment, and the sensors have a saturation value. After a range, the sensors return a truncated value that can be treated as missing data. In this case, the sampling pattern depends on the underlying matrix itself. This problem is known as *truncated matrix completion* [NTTB22].

In particular, we need to quantify how scattered the information of the non-zero entries is since large matrices with all non-zero entries concentrated in a certain region will hardly be retrieved unless all the measurements (observations) are performed in this region. Another mathematical way to describe this fact is by measuring the alignment of a matrix with the standard basis of $\mathbb{R}^{d_1 \times d_2}$ or, in other words, how concentrated the singular vectors are. The following notion of *incoherence* quantifies this phenomenon:

Definition 3.4.1. *We say that a rank-r matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ with singular value decomposition $\mathbf{X} = \mathbf{U}\operatorname{diag}(\sigma)\mathbf{V}^*$, $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$, $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$, is μ_0-*incoherent* if there exists a constant $\mu_0 \geq 1$ such that*

$$\max_{1 \leq i \leq d_1, 1 \leq j \leq d_2} \|\mathcal{P}_T(e_i e_j^*)\|_F \leq \sqrt{\mu_0 r \frac{d_1 + d_2}{d_1 d_2}}, \tag{3.22}$$

where $T = T_{\mathbf{X}} = \{\mathbf{U}\mathbf{M}^* + \widetilde{\mathbf{M}}\mathbf{V}^* : \ \mathbf{M} \in \mathbb{R}^{d_2 \times r}, \ \widetilde{\mathbf{M}} \in \mathbb{R}^{d_1 \times r}\}$ is the tangent space onto the rank-r matrix manifold at $\mathbf{X}$ and $\mathcal{P}_T$ is the projection operator onto T .

This definition is slightly weaker than related conditions of [Rec11, Che15].

Another interesting line of research deals with the problem of non-uniform random sampling patterns [SSS$^+$16]. In particular, results in the literature show how to complete *any* low-rank matrix, even without the coherence assumption, using a biased sampling strategy [CBSW15]. In particular, those methods proposed as a strategy for the matrix completion a *weighted nuclear norm minimization* scheme. An interesting question is to extend those results to IRLS, which, as already discussed here, is computationally simpler. We leave this as an open problem.

Open Problem: How to develop an IRLS scheme that provably works to complete low-rank matrices from non-uniform sampling distributions?

Now, by assuming a uniform sampling model, we can finally state the main result of this chapter. As defined in the Section 1.4, by denoting the *spectral norm* (or *Schatten-∞ norm*) of a matrix $\mathbf{X}$ by $\|\mathbf{X}\|_{S_\infty} = \sigma_1(\mathbf{X})$, we obtain the following local convergence result.

Theorem 3.4.2 (Local convergence of `MatrixIRLS` with Quadratic Rate)**.** *Let $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ be a matrix of rank r that is μ_0-incoherent, and let $P_\Omega : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ be the subsampling operator corresponding to an index set $\Omega = (i_\ell, j_\ell)_{\ell=1}^m \subset [d_1] \times [d_2]$ that is drawn uniformly without replacement. If the sample complexity fulfills $m \gtrsim \mu_0 r(d_1 +$*

$d_2)\log(d_1+d_2)$, then with high probability, the following holds: If the output matrix $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1\times d_2}$ of the k-th iteration of `MatrixIRLS` *with inputs P_Ω, $\mathbf{y} = P_\Omega(\mathbf{X}_0)$ and $\widetilde{r} = r$ updates the smoothing parameter in* (3.11) *such that $\epsilon_k = \sigma_{r+1}(\mathbf{X}^{(k)})$ and fulfills*

$$\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \lesssim \min\left(\sqrt{\frac{\mu_0 r}{d}}, \frac{\mu_0}{d\log(D)\kappa}\right)\sigma_r(\mathbf{X}_0), \tag{3.23}$$

where $\kappa = \sigma_1(\mathbf{X}_0)/\sigma_r(\mathbf{X}_0)$, then the local convergence rate is quadratic *in the sense that $\|\mathbf{X}^{(k+1)} - \mathbf{X}_0\|_{S_\infty} \leq \min(\mu\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty}^2, \|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty})$ with $\mu \leq \frac{d\log(D)}{\mu_0\sigma_r(\mathbf{X}_0)}\kappa$, and furthermore $\mathbf{X}^{(k+\ell)} \xrightarrow{\ell\to\infty} \mathbf{X}_0$ if additionally $\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \lesssim \min\left(\sqrt{\frac{\mu_0 r}{d}}, \frac{\mu_0^{3/2} r^{1/2}}{d^2\log(D)^{3/2}\kappa}\right)\sigma_r(\mathbf{X}_0)$.*

While a comparable local convergence result had been obtained for an IRLS algorithm for (non-convex) Schatten-p minimization [KS18], that result is *not* applicable for matrix completion, as the proof relied on a *null space property* [Rec11] of the measurement operator, which is not fulfilled by P_Ω since there are always rank-ones matrices in the null space of the entry-wise operator P_Ω.

As shown in Table 3.1, unlike the theory of other algorithms, the sample complexity assumption of Theorem 3.4.2 is *optimal* as it matches a well-known lower bound for this sampling model [CT10] that is necessary for unique identifiability. Among the weakest sufficient conditions for existing algorithms are $m \gtrsim \mu_0 r(d_1+d_2)\log^2(d_1+d_2)$ for nuclear norm minimization [Che15], $m \gtrsim \mu_0\kappa^{14}r^2(d_1+d_2)\log^2(d_1+d_2)$ for gradient descent [CLL20] on a variant of Equation (3.6) and $m \gtrsim \kappa^6(d_1+d_2)r^2\log(d_1+d_2)$ required random samples for the Riemannian gradient descent algorithm of [WCCL20]. In particular, our result does not depend on the condition number, indicating that our method will potentially retrieve highly ill-conditioned matrices while other methods fail, cf. Section 3.7. On the other hand, in contrast to other results, Theorem 3.4.2 only quantifies *local* convergence.

3.5 Proof of the Local Convergence Rate

In this section, we prove under a random sampling model on the location of the provided entries, `MatrixIRLS` converges locally to a low-rank completion of the data with high probability with a quadratic convergence rate, as described in Theorem 3.4.2.

First, we shortly elaborate on our notion of *incoherence*, see Definition 3.4.1, which quantifies the alignment of the standard basis $(e_i e_j^*)_{i=1,j=1}^{d_1,d_2}$ of $\mathbb{R}^{d_1\times d_2}$ with the *tangent space* onto the manifold of low-rank matrices at a specific rank-r matrix. The definition of incoherence dates back to the works [DE03a, CR07] in the context of sparse recovery, although the definition meant something slightly different in that context. This, in turn, appeared already implicitly in the paper [MZ93].

As mentioned in the previous section, a small incoherence parameter is a way to guarantee that the information of the column and row spaces is not very concentrated in just a few columns or rows. In the case of *structure low-rank matrix completion* such as low-rank Hankel matrix completion, which is a technique used to recover spectrally sparse signals, there is an interesting connection between the coherence parameter and the minimum separation condition between frequency pairs, the so-called Rayleigh resolution [CFG14]. In that case, the coherence is given by the reciprocal of the smallest singular value of a certain 2D Dirichlet kernel matrix that encapsulates the Rayleigh condition, see [CC14]. Here, we require the basis elements $(e_i e_j^*)_{i=1,j=1}^{d_1,d_2}$ of $\mathbb{R}^{d_1 \times d_2}$ of the space of $d_1 \times d_2$ matrices to have a small projection component on the tangent space of the manifold of rank-r matrices.

Remark 3.5.1. *We note that the assumption that a rank-r matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ is μ_0-incoherent according to Definition 3.4.1 is* weaker *than similar assumptions described in Definition 1.2, A0 and A1 of [CR09] and Definition 1 and Theorem 2 of [Rec11], and even than the assumption (2) of [Che15], which is the weakest available incoherence condition in the literature that is used for showing successful completion by nuclear norm minimization. More precisely, [Che15] calls a matrix $\mathbf{X}$ μ_0-incoherent if*

$$\max_{1 \le i \le d_1} \|\mathbf{U}^* e_i\|_2 \le \sqrt{\frac{\mu_0 r}{d_1}} \quad \textit{and} \quad \max_{1 \le j \le d_2} \|\mathbf{V}^* e_j\|_F \le \sqrt{\frac{\mu_0 r}{d_2}}. \tag{3.24}$$

In fact, condition (3.24) *is* stronger *than* (3.22)*. If $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$ are the left and right singular matrices corresponding to the r non-zero singular values of $\mathbf{X}$, we can write the projection operator $\mathcal{P}_T : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ that projects onto the tangent space T such $\mathcal{P}_T(\mathbf{Z}) = \mathbf{U}\mathbf{U}^*\mathbf{Z} + \mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^*$. Therefore, it can be seen that*

$$\begin{aligned}
\|\mathcal{P}_T(e_i e_j^*)\|_F^2 &= \|\mathbf{U}\mathbf{U}^* e_i e_j^* + e_i e_j^* \mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^* e_i e_j^* \mathbf{V}\mathbf{V}^*\|_F^2 = \|\mathbf{U}\mathbf{U}^* e_i e_j^* (\mathbf{I} - \mathbf{V}\mathbf{V}^*) + e_i e_j^* \mathbf{V}\mathbf{V}^*\|_F^2 \\
&= \|\mathbf{U}\mathbf{U}^* e_i e_j^* (\mathbf{I} - \mathbf{V}\mathbf{V}^*)\|_F^2 + \|e_i e_j^* \mathbf{V}\mathbf{V}^*\|_F^2 \le \|\mathbf{U}\mathbf{U}^* e_i e_j^*\|_F^2 \|\mathbf{I} - \mathbf{V}\mathbf{V}^*\|^2 + \|e_i e_j^* \mathbf{V}\mathbf{V}^*\|_F^2 \\
&\le \|\mathbf{U}^* e_i e_j^*\|_F^2 + \|e_i e_j^* \mathbf{V}\|_F^2 = \|\mathbf{U}^* e_i\|_2^2 + \|\mathbf{V}^* e_j\|_2^2 \le \frac{\mu_0 r}{d_1} + \frac{\mu_0 r}{d_2} \le \frac{\mu_0 r (d_1 + d_2)}{d_1 d_2}
\end{aligned}$$

for any $i \in [d_1]$, $j \in [d_2]$, if (3.24) *is fulfilled, which holds since*

$$\|\mathbf{U}^* e_i e_j^*\|_F^2 = \operatorname{tr}(e_j e_i^* \mathbf{U}\mathbf{U}^* e_i e_j^*) = \operatorname{tr}(e_i^* \mathbf{U}\mathbf{U}^* e_i) = e_i^* \mathbf{U}\mathbf{U}^* e_i = \|\mathbf{U}^* e_i\|_2^2$$

and similarly $\|e_i e_j^* \mathbf{V}\|_F^2 = \|\mathbf{V}^* e_j\|_2^2$.

3.5.1 Proof Roadmap

Here, we will briefly detail a roadmap for the quadratic convergence result, Theorem 3.4.2 since it consists of several steps. In the proof, we will denote by $\mathcal{T}_{r_k}(\mathbf{X}^{(k)})$ the *best rank-r_k approximation* of $\mathbf{X}^{(k)}$, that is given by the Eckardt-Young-Mirsky theorem [Mir60].

Theorem 3.5.2. *Given a matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, the best rank-$r$ approximation $\mathcal{T}_{r_k}(\mathbf{X}^{(k)})$ to $\mathbf{X}$ is given by its truncated singular value decomposition, i.e.,*

$$\mathcal{T}_{r_k}(\mathbf{X}^{(k)}) := \underset{\mathbf{Z}:\mathrm{rank}(\mathbf{Z}) \leq r_k}{\arg\min} \|\mathbf{Z} - \mathbf{X}^{(k)}\| = \mathbf{U}^{(k)}\mathbf{\Sigma}^{(k)}\mathbf{V}^{(k)*}. \tag{3.25}$$

Eckart and Young proved the theorem above in 1936 [EY36] for the Frobenius norm and generalized by Mirsky for any unitarily invariant norm. An interesting generalization of this theorem was established in [GHS87].

The first step in the convergence proof is understanding the interplay between the sampling operator, where we assume that the knowing entries are revealed uniformly at random without replacement and the tangent space of the manifold of rank-r matrices. In particular, by using the definition of coherence, it was shown that the operator representing the sampling model is well-conditioned on the tangent space of the rank-r manifold. This is presented in Lemma 3.5.4.

The second step is to show that this well-conditioning property from the first step, which can be interpreted as a *restricted isometry property* on the tangent space, can "be transferred" from a tangent space at a point $\mathbf{X}_0$ to a tangent space at a point $\mathbf{X}^{(k+1)}$ provided that both point $\mathbf{X}_0$ and $\mathbf{X}^{(k+1)}$ are close enough. Consequently, we can quantify "how relevant" a vector $\eta \in \ker P_\Omega$ is. In particular, we prove that most of its energy is concentrated in the space orthogonal to the tangent space, i.e., we will show that if $\|\eta\|_F$ is large, then the projection $\|\mathcal{P}_{T_k^\perp}(\eta)\|_F$ must also be large. This is presented in Lemma 3.5.6.

After that, in the third step, we introduce a classical matrix perturbation argument that quantifies how close two subspaces are. See Lemma 3.5.9.

By using this, in the fourth step, we observe that $\mathbf{X}^{(k+1)} - \mathbf{X}_0$, where $\mathbf{X}^{(k+1)}$ is an iteration of `MatrixIRLS`, is an element of $\ker P_\Omega$ and we use the previous result to show that this distance its distance can be bounded by the nuclear norm of the weight operator $\|W^{(k)}(\mathbf{X}_0)\|_{S_1}$. This is the content of Lemma 3.5.11

Finally, by taking the structure of the weight operator into account, we break it into four pieces and estimate them by the norm of the vector $\mathbf{X}^{(k+1)} - \mathbf{X}_0$. In this step, the definition of the weighted matrix developed for `MatrixIRLS` will play an important role. In particular, the smoothing parameter ε_k will appear in the bounds, and this will be finally estimated in the last step.

In the end, we also connect the decay in the smoothing parameter ε_k with the norm of $\mathbf{X}^{(k+1)} - \mathbf{X}_0$. To conclude, we show that if the current iteration of the algorithm lies in a local basin of attraction, we can ensure that the $(r+1)$-st singular value $\sigma_{r+1}(\mathbf{X}^{(k)})$ of the current iterate is strictly decreasing. Then, we wrap up the proof by combining all the steps described above. But we start by discussing the interaction between the projection onto the tangent space and the random sampling operator.

3.5.2 Interplay between sampling operator and tangent space

In the statement of Theorem 3.4.2, we assume that the index set Ω is *drawn uniformly at random without replacement.* In our proof below, however, we use a sampling model on the locations $\Omega = (i_\ell, j_\ell)_{\ell=1}^m$ corresponding to *independent sampling with replacement.* It is well-known (see, e.g., Proposition 3 of [Rec11]) that the statement then carries over to the above sampling model without replacement.

As a preparation for our proof, we recall a result from [Rec11] that bounds the number of repetitions of each location in Ω under the random sampling model with replacement.

Lemma 3.5.3. *[Rec11, Proposition 5] Let $D = \max(d_1, d_2)$ and $\beta > 1$, let $\Omega = (i_\ell, j_\ell)_{\ell=1}^m$ be a multiset of double indices from $[d_1] \times [d_2]$ fulfilling $m < d_1 d_2$ that are sampled independently with replacement. Then with probability at least $1 - D^{2-2\beta}$, the maximal number of repetitions of any entry in Ω is less than $\frac{8}{3}\beta \log(D)$ for $D \geq 9$ and $\beta > 1$. Consequently, we have that with probability of at least $1 - D^{2-2\beta}$, the operator $\mathcal{R}_\Omega : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ defined such that*

$$\mathcal{R}_\Omega(\mathbf{X}) := P_\Omega^*(P_\Omega(\mathbf{X})) = \sum_{\ell=1}^m \langle e_{i_\ell} e_{j_\ell}^*, \mathbf{X} \rangle e_{i_\ell} e_{j_\ell}^* \tag{3.26}$$

fulfills

$$\|\mathcal{R}_\Omega\|_{S_\infty} \leq \frac{8}{3}\beta \log(D).$$

Now, we state a lemma from [Rec11] that shows that the operator $\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0}$ is well-conditioned. As discussed in Remark 3.5.5 below, this result can be seen as a *restricted isometry property* on the space of matrices. We provide the proof for completeness since we use the weaker incoherence definition of Definition 3.4.1 instead of the incoherence notions of [Rec11, Che15]. As discussed in [Rec11], for a sampling without replacement model given by a Bernoulli distribution where each entry is revealed independently with probability equal to p, the next result is highly non-trivial and uses several results of probability in Banach spaces. See [CR09, Theorem 4.1]. Here, on the other hand, the proof can be simplified due to the assumption on the sampling strategy.

Lemma 3.5.4. *[Rec11, Theorem 6] Let $0 < \epsilon \leq \frac{1}{2}$, let $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ be a μ_0-incoherent* *matrix whose tangent space* $T_0 = T_{\mathbf{X}_0}$ *onto the rank-r manifold* $T_0 = T_{\mathbf{X}_0}\mathcal{M}_r$ *(see* (3.20)*) fulfills* (3.22) *and* $\mathcal{R}_\Omega : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ *be defined as in* (3.26) *from* m *independent uniformly sampled locations. Let* $\mathcal{P}_{T_0} : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ *be the projection operator associated to* T_0*. Then*

$$\left\| \frac{d_1 d_2}{m} \mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0} - \mathcal{P}_{T_0} \right\|_{S_\infty} \leq \varepsilon \tag{3.27}$$

holds with probability at least $1-(d_1+d_2)^{-2}$ *provided that*

$$m \geq \frac{7}{\varepsilon^2}\mu_0 r(d_1+d_2)\log(d_1+d_2). \tag{3.28}$$

Proof of Lemma 3.5.4. First we define the family of operators $\mathcal{Z}_\ell, \widetilde{\mathcal{Z}}_\ell : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ such that for $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$,

$$\mathcal{Z}_\ell(\mathbf{X}) := \frac{d_1 d_2}{m}\langle e_{i_\ell}e_{j_\ell}^*, \mathcal{P}_{T_0}(\mathbf{X})\rangle \mathcal{P}_{T_0}(e_{i_\ell}e_{j_\ell}^*) - \frac{1}{m}\mathcal{P}_{T_0}(\mathbf{X}) := \frac{d_1 d_2}{m}\widetilde{\mathcal{Z}}_\ell(\mathbf{X}) - \frac{1}{m}\mathcal{P}_{T_0}(\mathbf{X})$$

for any $\ell \in [m]$. Then

$$\mathbb{E}[\mathcal{Z}_\ell] = \frac{1}{d_1 d_2}\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\frac{d_1 d_2}{m}\langle e_i e_j^*, \mathcal{P}_{T_0}(\cdot)\rangle\mathcal{P}_{T_0}(e_i e_j^*) - \frac{1}{m}\mathcal{P}_{T_0} = \frac{1}{d_1 d_2}\frac{d_1 d_2}{m}\mathcal{P}_{T_0}\mathbf{I}\mathcal{P}_{T_0} - \frac{1}{m}\mathcal{P}_{T_0} = 0. \tag{3.29}$$

Since for $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$

$$\langle e_{i_\ell}e_{j_\ell}^*, \mathcal{P}_{T_0}(\mathbf{X})\rangle\mathcal{P}_{T_0}(e_{i_\ell}e_{j_\ell}^*) = \langle \mathcal{P}_{T_0}(e_{i_\ell}e_{j_\ell}^*), \mathbf{X}\rangle\mathcal{P}_{T_0}(e_{i_\ell}e_{j_\ell}^*),$$

we obtain

$$\|\langle e_{i_\ell}e_{j_\ell}^*, \mathcal{P}_{T_0}(\mathbf{X})\rangle\mathcal{P}_{T_0}(e_{i_\ell}e_{j_\ell}^*)\|_F \leq \left|\langle \mathcal{P}_{T_0}(e_{i_\ell}e_{j_\ell}^*), \mathbf{X}\rangle\right| \|\mathcal{P}_{T_0}(e_{i_\ell}e_{j_\ell}^*)\|_F \leq \|\mathcal{P}_{T_0}(e_{i_\ell}e_{j_\ell}^*)\|_F^2 \|X\|_F$$

by Cauchy-Schwartz, and thus the norm bound

$$\begin{aligned} \frac{d_1 d_2}{m}\left\|\widetilde{\mathcal{Z}}_\ell\right\|_{S_\infty} &\leq \frac{d_1 d_2}{m}\|\mathcal{P}_{T_0}(e_{i_\ell}e_{j_\ell}^*)\|_F^2 \leq \frac{d_1 d_2}{m}\max_{i\in[d_1],j\in[d_2]}\|\mathcal{P}_{T_0}(e_i e_j^*)\|_F^2 \\ &\leq \frac{d_1 d_2}{m}\frac{\mu_0 r(d_1+d_2)}{d_1 d_2} = \frac{\mu_0 r(d_1+d_2)}{m} \end{aligned} \tag{3.30}$$

using the incoherence assumption (3.22) in the last inequality. Similarly,

$$\left\|\frac{1}{m}\mathcal{P}_{T_0}\right\|_{S_\infty} = \left\|\frac{1}{m}\mathcal{P}_{T_0}\mathbf{I}\mathcal{P}_{T_0}\right\|_{S_\infty} \leq \frac{1}{m}\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\left\|\langle\mathcal{P}_{T_0}(e_i e_j^*), (\cdot)\rangle\mathcal{P}_{T_0}(e_i e_j^*)\right\|_{S_\infty} \leq \frac{\mu_0 r(d_1+d_2)}{m}. \tag{3.31}$$

We note that if operators $\mathcal{A}$ and $\mathcal{B}$ are positive semidefinite, then $\|\mathcal{A}-\mathcal{B}\|_{S_\infty} \leq \max(\|\mathcal{A}\|_{S_\infty}, \|\mathcal{B}\|_{S_\infty})$, and as both $\widetilde{\mathcal{Z}}_\ell$ and $\mathcal{P}_{T_0}$ are positive semidefinite, we have

$$\|\mathcal{Z}_\ell\|_{S_\infty} \leq \max\left(\frac{d_1 d_2}{m}\left\|\widetilde{\mathcal{Z}}_\ell\right\|_{S_\infty}, \frac{1}{m}\|\mathcal{P}_{T_0}\|_{S_\infty}\right) = \frac{\mu_0 r(d_1+d_2)}{m}$$

for all $\ell \in [m]$. By taking the expectation of the squares of $\mathcal{Z}_\ell$, we obtain

$$\begin{aligned}\mathbb{E}\,\mathcal{Z}_\ell\mathcal{Z}_\ell^* &= \frac{(d_1 d_2)^2}{m^2}\,\mathbb{E}\left[(\widetilde{\mathcal{Z}}_\ell)^*\widetilde{\mathcal{Z}}_\ell\right] - \frac{d_1 d_2}{m^2}\,\mathbb{E}\left[\widetilde{\mathcal{Z}}_\ell\right]\mathcal{P}_{T_0} - \frac{d_1 d_2}{m^2}\mathcal{P}_{T_0}\,\mathbb{E}\left[\widetilde{\mathcal{Z}}_\ell\right] + \frac{1}{m^2}\mathcal{P}_{T_0}\\ &= \frac{(d_1 d_2)^2}{m^2}\,\mathbb{E}\left[(\widetilde{\mathcal{Z}}_\ell)^*\widetilde{\mathcal{Z}}_\ell\right] + (1-2)\frac{1}{m^2}\mathcal{P}_{T_0},\end{aligned}$$

as $\mathcal{P}_{T_0}^2 = \mathcal{P}_{T_0}$ and $\mathbb{E}[\widetilde{\mathcal{Z}}_\ell] = \frac{1}{d_1 d_2}\mathcal{P}_{T_0}$. Hence,

$$\begin{aligned}\left\|\sum_{\ell=1}^m \mathbb{E}\,\mathcal{Z}_\ell\mathcal{Z}_\ell^*\right\|_{S_\infty} &\leq \sum_{\ell=1}^m \|\mathbb{E}\,\mathcal{Z}_\ell\mathcal{Z}_\ell^*\|_{S_\infty} = \sum_{\ell=1}^m \left\|\frac{(d_1 d_2)^2}{m^2}\,\mathbb{E}\left[(\widetilde{\mathcal{Z}}_\ell)^2\right] - \frac{1}{m^2}\mathcal{P}_{T_0}\right\|_{S_\infty}\\ &\leq \sum_{\ell=1}^m \max\left(\frac{(d_1 d_2)^2}{m^2}\left\|\mathbb{E}\left[(\widetilde{\mathcal{Z}}_\ell)^2\right]\right\|_{S_\infty}, \frac{1}{m^2}\|\mathcal{P}_{T_0}\|_{S_\infty}\right)\\ &\leq \sum_{\ell=1}^m \max\left(\frac{(d_1 d_2)^2}{m^2}\left\|\mathbb{E}\left[\|\mathcal{P}_{T_0}(e_{i_\ell}e_{j_\ell}^*)\|_F^2\widetilde{\mathcal{Z}}_\ell\right]\right\|_{S_\infty}, \frac{1}{m^2}\right)\\ &\leq \sum_{\ell=1}^m \max\left(\frac{(d_1 d_2)(d_1+d_2)\mu_0 r}{m^2}\left\|\mathbb{E}\,\widetilde{\mathcal{Z}}_\ell\right\|_{S_\infty}, \frac{1}{m^2}\right)\\ &\leq \sum_{\ell=1}^m \max\left(\frac{(d_1+d_2)\mu_0 r}{m^2}, \frac{1}{m^2}\right) = \frac{\mu_0 r(d_1+d_2)}{m},\end{aligned}$$

where we used that $\|\mathcal{P}_{T_0}\|_2 \leq 1$ since $\mathcal{P}_{T_0}$ is a projection in the third inequality, the definition of μ_0 in the fourth and the fact that $\mathbb{E}\,\widetilde{\mathcal{Z}}_\ell = \frac{1}{d_1 d_2}\mathcal{P}_{T_0}$, since $\mathbb{E}\,\mathcal{Z}_\ell = 0$ due to Equation (3.29), in the fifth. As all the operators $\mathcal{Z}_\ell$ are Hermitian, it follows by the matrix Bernstein inequality [Ver18, Theorem 5.4.1] that

$$\begin{aligned}P\left(\left\|\frac{d_1 d_2}{m}\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0} - \mathcal{P}_{T_0}\right\|_{S_\infty} \geq \varepsilon\right) &\leq (d_1+d_2)\exp\left(-\frac{m\varepsilon^2/2}{\mu_0 r(d_1+d_2) + \mu_0 r(d_1+d_2)\epsilon/3}\right)\\ &\leq (d_1+d_2)\exp\left(-\frac{m\varepsilon^2}{2\mu_0 r(d_1+d_2) + \mu_0 r(d_1+d_2)/3}\right),\end{aligned}$$

using that $\varepsilon \leq \frac{1}{2}$ in the last inequality. Furthermore, if (3.28) is fulfilled, then

$$(d_1+d_2)\exp\left(-\frac{m\varepsilon^2}{\frac{7}{3}\mu_0 r(d_1+d_2)}\right) \leq (d_1+d_2)^{-2},$$

which shows that (3.27) holds with a probability of at least $1 - (d_1 + d_2)^{-2}$. □

Remark 3.5.5. *The result above can be seen (and it is usually referred to) as a restricted isometry property (RIP) on the tangent space. Indeed, the inequality* $\left\|\frac{d_1 d_2}{m}\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0} - \mathcal{P}_{T_0}\right\|_{S_\infty} \leq \varepsilon$ *implies that exist* $C_1, C_2 > 0$ *such that*

$$C_1\|\mathcal{P}_{T_0}(\mathbf{X}_0)\|_F \leq \|\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0}(\mathbf{X}_0)\|_F \leq C_2\|\mathcal{P}_{T_0}(\mathbf{X}_0)\|_F \tag{3.32}$$

This, in particular, shows that the operator $\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0}$ *mapping the tangent space of the rank-r matrices onto itself is well-conditioned and hence invertible. This remarkable fact is one of the crucial ingredients in essentially any rigorous analysis of algorithms for matrix completion problems.*

To prove our convergence rate theorem, we will use the local restricted isometry statement of (3.27) for tangent spaces $T_{\mathbf{X}}$ corresponding to matrices $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ *that are close* to $\mathbf{X}_0$. We show the following auxiliary result, which is a refinement of Lemma 4.2 [WCCL20] as we obtain a bound in the S_∞-norm in (c) instead of in the Frobenius norm.

Lemma 3.5.6. *Let* $\mathbf{X}_0, \mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ *be matrices and assume that* $0 < \varepsilon < 1$ *and that the following three conditions hold:*

(a) For $\mathcal{R}_\Omega : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ *as in* (3.26),

$$\|\mathcal{R}_\Omega\|_{S_\infty} \leq \frac{16}{3}\log(D).$$

(b) The tangent space $T_0 = T_{\mathbf{X}_0}$ *onto the rank-r manifold* $\mathcal{M}_r$ *at* $\mathbf{X}_0$ *fulfills*

$$\left\|\frac{d_1 d_2}{m}\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0} - \mathcal{P}_{T_0}\right\|_{S_\infty} \leq \varepsilon.$$

(c) The spectral norm distance between $\mathbf{X}$ *and* $\mathbf{X}_0$ *fulfills*

$$\|\mathbf{X} - \mathbf{X}_0\|_{S_\infty} \leq \frac{\sqrt{3}}{32\sqrt{\log(D)}\sqrt{(1+\varepsilon)}}\varepsilon\sqrt{\frac{m}{d_1 d_2}}\sigma_r(\mathbf{X}_0).$$

Then the tangent space $T = T_{\mathbf{X}}$ *onto the rank-r manifold at* $\mathbf{X}$ *fulfills*

$$\left\|\frac{d_1 d_2}{m}\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_T - \mathcal{P}_T\right\|_{S_\infty} \leq 4\varepsilon. \tag{3.33}$$

Proof. For any $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$, we have

$$\begin{aligned}
&\|\mathcal{R}_\Omega \mathcal{P}_{T_0}(\mathbf{Z})\|_F^2 = \langle \mathcal{R}_\Omega \mathcal{P}_T(\mathbf{Z}), \mathcal{R}_\Omega \mathcal{P}_T(\mathbf{Z})\rangle \leq \frac{16}{3}\log(D)\,\langle \mathcal{P}_{T_0}(\mathbf{Z}), \mathcal{R}_\Omega \mathcal{P}_{T_0}(\mathbf{Z})\rangle \\
&= \frac{16}{3}\log(D)\,\langle \mathcal{P}_{T_0}(\mathbf{Z}), \mathcal{P}_{T_0}\mathcal{R}_\Omega \mathcal{P}_{T_0}(\mathbf{Z})\rangle \\
&= \frac{16}{3}\log(D)\left(\left\langle \mathcal{P}_{T_0}(\mathbf{Z}), \frac{m}{d_1 d_2}\mathcal{P}_{T_0}(\mathbf{Z})\right\rangle + \left\langle \mathcal{P}_{T_0}(\mathbf{Z}), \left(\mathcal{P}_{T_0}\mathcal{R}_\Omega \mathcal{P}_{T_0}(\mathbf{Z}) - \frac{m}{d_1 d_2}\mathcal{P}_{T_0}(\mathbf{Z})\right)\right\rangle\right) \\
&\leq \frac{16}{3}\log(D)\left(\frac{m}{d_1 d_2} + \varepsilon\frac{m}{d_1 d_2}\right)\|\mathcal{P}_{T_0}(\mathbf{Z})\|_F^2 \leq \frac{16}{3}\log(D)(1+\varepsilon)\frac{m}{d_1 d_2}\|\mathbf{Z}\|_F^2,
\end{aligned}$$

where the first inequality follows from condition (a) and the second one from condition (b). It follows that

$$\|\mathcal{R}_\Omega \mathcal{P}_{T_0}\| \leq \sqrt{\frac{16}{3}\log(D)(1+\varepsilon)\frac{m}{d_1 d_2}}. \tag{3.34}$$

Furthermore, if $\mathbf{U}, \mathbf{U}_0 \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V}, \mathbf{V}_0 \in \mathbb{R}^{d_2 \times r}$ are the matrices of first r left and right singular vectors of $\mathbf{X}$ and $\mathbf{X}_0$, respectively, it holds that for any $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$,

$$\begin{aligned}
(\mathcal{P}_T - \mathcal{P}_{T_0})(\mathbf{Z}) &= \mathbf{U}\mathbf{U}^*\mathbf{Z} + \mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}_0\mathbf{U}_0^*\mathbf{Z} - \mathbf{Z}\mathbf{V}_0\mathbf{V}_0^* + \mathbf{U}_0\mathbf{U}_0^*\mathbf{Z}\mathbf{V}_0\mathbf{V}_0^* \\
&= (\mathbf{U}\mathbf{U}^* - \mathbf{U}_0\mathbf{U}_0^*)\,\mathbf{Z}(\mathbf{I} - \mathbf{V}_0\mathbf{V}_0^*) + (\mathbf{I} - \mathbf{U}\mathbf{U}^*)\mathbf{Z}(\mathbf{V}\mathbf{V}^* - \mathbf{V}_0\mathbf{V}_0^*),
\end{aligned}$$

which we use to estimate

$$\begin{aligned}
\|(\mathcal{P}_T - \mathcal{P}_{T_0})(\mathbf{Z})\|_F &\leq \|\mathbf{U}\mathbf{U}^* - \mathbf{U}_0\mathbf{U}_0^*\|_{S_\infty}\|\mathbf{Z}\|_F\|\mathbf{I} - \mathbf{V}_0\mathbf{V}_0^*\|_{S_\infty} + \|\mathbf{I} - \mathbf{U}\mathbf{U}^*\|_{S_\infty}\|\mathbf{Z}\|_F\|\mathbf{V}\mathbf{V}^* - \mathbf{V}_0\mathbf{V}_0^*\|_{S_\infty} \\
&\leq \frac{\|\mathcal{T}_r(\mathbf{X}) - \mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)}\|\mathbf{Z}\|_F \cdot 1 + 1 \cdot \|\mathbf{Z}\|_F \frac{\|\mathcal{T}_r(\mathbf{X}) - \mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)} \\
&\leq 2\frac{\|\mathcal{T}_r(\mathbf{X}) - \mathbf{X}\|_{S_\infty} + \|\mathbf{X} - \mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)}\|\mathbf{Z}\|_F,
\end{aligned}$$

where $\mathcal{T}_r(\mathbf{X})$ is the best rank-r approximation (3.25). Here, we used the results

$$\|\mathbf{U}\mathbf{U}^* - \mathbf{U}_0\mathbf{U}_0^*\|_{S_\infty} \leq \frac{\|\mathcal{T}_r(\mathbf{X}) - \mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)}$$

and

$$\|\mathbf{V}\mathbf{V}^* - \mathbf{V}_0\mathbf{V}_0^*\|_{S_\infty} \leq \frac{\|\mathcal{T}_r(\mathbf{X}) - \mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)}$$

of Lemma 4.2, inequality (4.3) of [WCCL16], which bound the distance between the projections onto the left and right singular subspaces of $\mathbf{X}$ and $\mathbf{X}_0$.

From the Eckardt-Young-Mirsky theorem (3.25), it then follows that

$$\|(\mathcal{P}_T - \mathcal{P}_{T_0})\|_{S_\infty} \leq \frac{4\|\mathbf{X} - \mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)}. \tag{3.35}$$

With this, we further bound

$$\begin{aligned}
\|\mathcal{R}_\Omega \mathcal{P}_T\|_{S_\infty} &\leq \|\mathcal{R}_\Omega(\mathcal{P}_T - \mathcal{P}_{T_0})\|_{S_\infty} + \|\mathcal{R}_\Omega \mathcal{P}_{T_0}\|_{S_\infty} \\
&\leq \frac{16}{3}\log(D)\frac{4\|\mathbf{X}-\mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)} + \|\mathcal{R}_\Omega \mathcal{P}_{T_0}\|_{S_\infty} \\
&\leq \frac{16}{3}\log(D)\frac{\sqrt{3}}{8\sqrt{\log(D)}\sqrt{(1+\varepsilon)}}\varepsilon\sqrt{\frac{m}{d_1 d_2}} + \sqrt{\frac{16}{3}\log(D)(1+\varepsilon)\frac{m}{d_1 d_2}} \qquad (3.36) \\
&= \frac{2}{\sqrt{3}}\sqrt{\log(D)}\frac{1}{\sqrt{(1+\varepsilon)}}\varepsilon\sqrt{\frac{m}{d_1 d_2}} + \sqrt{\frac{16}{3}\log(D)(1+\varepsilon)\frac{m}{d_1 d_2}} \\
&\leq 2\sqrt{3}\sqrt{\log(D)}\sqrt{1+\varepsilon}\sqrt{\frac{m}{d_1 d_2}},
\end{aligned}$$

where the second inequality follows from (3.35) and the third from condition (c). To prove the statement (3.33), we calculate

$$\begin{aligned}
\left\|\frac{d_1 d_2}{m}\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_T - \mathcal{P}_T\right\|_{S_\infty} &\leq \|\mathcal{P}_T - \mathcal{P}_{T_0}\|_{S_\infty} + \frac{d_1 d_2}{m}\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_T - \mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_{T_0}\|_{S_\infty} \\
&\quad + \frac{d_1 d_2}{m}\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_{T_0} - \mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0}\|_{S_\infty} + \left\|\mathcal{P}_{T_0} - \frac{d_1 d_2}{m}\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0}\right\|_{S_\infty} \\
&\leq \|\mathcal{P}_T - \mathcal{P}_{T_0}\|_{S_\infty} + \frac{d_1 d_2}{m}\|\mathcal{R}_\Omega\mathcal{P}_T\|_{S_\infty}\|\mathcal{P}_T - \mathcal{P}_{T_0}\|_{S_\infty} \\
&\quad + \frac{d_1 d_2}{m}\|\mathcal{R}_\Omega\mathcal{P}_{T_0}\|_{S_\infty}\|\mathcal{P}_T - \mathcal{P}_{T_0}\|_{S_\infty} + \left\|\mathcal{P}_{T_0} - \frac{d_1 d_2}{m}\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0}\right\|_{S_\infty} \\
&\leq \frac{4\|\mathbf{X}-\mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)} + \frac{d_1 d_2}{m}\|\mathcal{R}_\Omega\mathcal{P}_T\|_{S_\infty}\frac{4\|\mathbf{X}-\mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)} \\
&\quad + \frac{d_1 d_2}{m}\|\mathcal{R}_\Omega\mathcal{P}_{T_0}\|_{S_\infty}\frac{4\|\mathbf{X}-\mathbf{X}_0\|_{S_\infty}}{\sigma_r(\mathbf{X}_0)} + \left\|\mathcal{P}_{T_0} - \frac{d_1 d_2}{m}\mathcal{P}_{T_0}\mathcal{R}_\Omega\mathcal{P}_{T_0}\right\|_{S_\infty} \\
&\leq 4\varepsilon
\end{aligned}$$

where in the second inequality, we utilized the fact $\mathcal{R}_\Omega^* = \mathcal{R}_\Omega$ so that $\|\mathcal{P}_T\mathcal{R}_\Omega\|_{S_\infty} = \|\mathcal{R}_\Omega\mathcal{P}_T\|_{S_\infty}$. The very last estimate follows from conditions (b) and (c) and the bounds (3.34) and (3.36) for $\|\mathcal{R}_\Omega\mathcal{P}_T\|_{S_\infty}$ and $\|\mathcal{R}_\Omega\mathcal{P}_{T_0}\|_{S_\infty}$. □

In the following lemma, we combine the previous results to show that under our sampling model, with high probability, a local restricted isometry property holds with respect to tangent spaces T_k that are in some sense close to $\mathbf{X}_0$.

Lemma 3.5.7. *Let $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ be a matrix of rank r that is μ_0-incoherent, and let $\Omega = (i_\ell, j_\ell)_{\ell=1}^m$ be a random index set of cardinality $|\Omega| = m$ that is sampled uniformly without replacement, or, alternatively, sampled independently with replacement. There* *exists constants* $C, \widetilde{C}, C_1$ *such that if*

$$m \geq C\mu_0 r(d_1+d_2)\log(d_1+d_2), \tag{3.37}$$

then, with probability at least $1-2D^{-2}$*, the following holds: For each matrix* $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1\times d_2}$ *fulfilling*

$$\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \leq C_1\sqrt{\frac{\mu_0 r}{d}}\sigma_r(\mathbf{X}_0), \tag{3.38}$$

it follows that the projection $\mathcal{P}_{T_k} : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ *onto the tangent space* $T_k := T_{\mathcal{T}_r(\mathbf{X}^{(k)})}\mathcal{M}_r$ *satisfies*

$$\left\| \frac{d_1 d_2}{m}\mathcal{P}_{T_k}P_\Omega^* P_\Omega \mathcal{P}_{T_k} - \mathcal{P}_{T_k}\right\|_{S_\infty} \leq \frac{2}{5},$$

and furthermore,

$$\|\eta\|_F \leq \sqrt{\frac{\widetilde{C}d\log(D)}{\mu_0 r}}\|\mathcal{P}_{T_k^\perp}(\eta)\|_F$$

for each matrix $\eta \in \ker P_\Omega$ *in the null space of the subsampling operator* $P_\Omega : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$.

Proof of Lemma 3.5.7. Assume that there are m locations $\Omega = (i_\ell, j_\ell)_{\ell=1}^m$ in $[d_1]\times[d_2]$ sampled independently uniformly *with replacement*, where m fulfills (3.37) with $C := 7/\varepsilon^2$ and $\varepsilon = 0.1$. By Lemma 3.5.3, it follows that the corresponding operator $\mathcal{R}_\Omega : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ from (3.26) fulfills

$$\|\mathcal{R}_\Omega\|_{S_\infty} \leq \frac{16}{3}\log(D) \tag{3.39}$$

on an event called E_Ω, which occurs with a probability of at least $1-D^{-2}$, and by Lemma 3.5.4, the tangent space $T_0 = T_{\mathbf{X}_0}\mathcal{M}_r$ corresponding to the μ_0-incoherent rank-r matrix $\mathbf{X}_0$ fulfills

$$\left\| \frac{d_1 d_2}{m}\mathcal{P}_{T_0}P_\Omega^* P_\Omega \mathcal{P}_{T_0} - \mathcal{P}_{T_0}\right\|_{S_\infty} \leq \varepsilon$$

on an event called E_{Ω,T_0}, which occurs with a probability of at least $1-D^{-2}$. Let $\tilde{\epsilon} = \frac{1}{10}$. If $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1\times d_2}$ is such that $\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \leq \widetilde{\xi}\sigma_r(\mathbf{X}_0)$ with

$$\widetilde{\xi} = \frac{\sqrt{3}}{32}\frac{\epsilon}{\sqrt{\log(D)(1+\epsilon)}}\sqrt{\frac{m}{d_1 d_2}} = \frac{\sqrt{3}}{32}\frac{1}{10\sqrt{\log(D)(11/10)}}\sqrt{\frac{m}{d_1 d_2}}, \tag{3.40}$$

it follows by Lemma 3.5.6 that on the event $E_\Omega \cap E_{\Omega,T_0}$, the tangent space $T_k := \mathbf{X}^{(k)}$ onto the rank-r manifold at $\mathbf{X}^{(k)}$ fulfills

$$\left\| \frac{d_1 d_2}{m}\mathcal{P}_{T_k}\mathcal{R}_\Omega \mathcal{P}_{T_k} - \mathcal{P}_{T_k}\right\|_{S_\infty} \leq 4\tilde{\epsilon} = \frac{2}{5}. \tag{3.41}$$

Next, we claim that on the event $E_\Omega \cap E_{\Omega,T_0}$,

$$\|\eta\|_F \leq \sqrt{\frac{\widetilde{C} d \log(D)}{\mu_0 r}} \|\mathcal{P}_{T_k^\perp}(\eta)\|_F. \tag{3.42}$$

for any for each matrix $\eta \in \ker P_\Omega$ in the null space of the subsampling operator $P_\Omega : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$. Indeed, to show this claim, we first note that $\eta \in \ker P_\Omega$ if and only if $\eta \in \ker \mathcal{R}_\Omega : P_\Omega^* P_\Omega$. Let $\eta \in \ker \mathcal{R}_\Omega$. Then

$$\begin{aligned}
\|\mathcal{P}_{T_k}(\eta)\|_F^2 &= \langle \mathcal{P}_{T_k}(\eta), \mathcal{P}_{T_k}(\eta)\rangle \\
&= \left\langle \mathcal{P}_{T_k}(\eta), \frac{d_1 d_2}{m} \mathcal{P}_{T_k} \mathcal{R}_\Omega \mathcal{P}_{T_k}(\eta)\right\rangle + \left\langle \mathcal{P}_{T_k}(\eta), \mathcal{P}_{T_k}(\eta) - \frac{d_1 d_2}{m} \mathcal{P}_{T_k} \mathcal{R}_\Omega \mathcal{P}_{T_k}(\eta)\right\rangle \\
&\leq \left\langle \mathcal{P}_{T_k}(\eta), \frac{d_1 d_2}{m} \mathcal{P}_{T_k} \mathcal{R}_\Omega \mathcal{P}_{T_k}(\eta)\right\rangle + \|\mathcal{P}_{T_k}(\eta)\|_F \left\|\mathcal{P}_{T_k} - \frac{d_1 d_2}{m} \mathcal{P}_{T_k} \mathcal{R}_\Omega \mathcal{P}_{T_k}\right\|_{S_\infty} \|\mathcal{P}_{T_k}(\eta)\|_F \\
&\leq \left\langle \mathcal{P}_{T_k}(\eta), \frac{d_1 d_2}{m} \mathcal{P}_{T_k} \mathcal{R}_\Omega \mathcal{P}_{T_k}(\eta)\right\rangle + 4\epsilon \|\mathcal{P}_{T_k}(\eta)\|_F^2,
\end{aligned}$$

using (3.41) in the last inequality, which implies that

$$\begin{aligned}
\|\mathcal{P}_{T_k}(\eta)\|_F^2 &\leq \frac{1}{1-4\epsilon} \frac{d_1 d_2}{m} \langle \mathcal{P}_{T_k}(\eta), \mathcal{P}_{T_k} \mathcal{R}_\Omega^2 \mathcal{P}_{T_k}(\eta)\rangle = \frac{1}{1-4\epsilon} \frac{d_1 d_2}{m} \|\mathcal{R}_\Omega \mathcal{P}_{T_k}(\eta)\|_F^2 \\
&\leq \frac{2 d_1 d_2}{m} \|\mathcal{R}_\Omega \mathcal{P}_{T_k}(\eta)\|_F^2
\end{aligned}$$

using the fact that $\mathcal{R}_\Omega : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ is positive semidefinite and has eigenvalues that are 0 or larger or equal than 1 only. Furthermore, we used that $\epsilon \leq \frac{1}{10}$ in the last inequality.

Since $\eta \in \ker \mathcal{R}_\Omega$, it holds that

$$0 = \|\mathcal{R}_\Omega(\eta)\|_F = \left\|\mathcal{R}_\Omega \left(\mathcal{P}_{T_k}(\eta) + \mathcal{P}_{T_k^\perp}(\eta)\right)\right\|_F \geq \|\mathcal{R}_\Omega \mathcal{P}_{T_k}(\eta)\|_F - \|\mathcal{R}_\Omega \mathcal{P}_{T_k^\perp}(\eta)\|_F$$

so that

$$\|\mathcal{R}_\Omega \mathcal{P}_{T_k}(\eta)\|_F \leq \|\mathcal{R}_\Omega \mathcal{P}_{T_k^\perp}(\eta)\|_F \leq \frac{16}{3} \log(D) \|\mathcal{P}_{T_k^\perp}(\eta)\|_F,$$

where we used (3.39), i.e., $\|\mathcal{R}_\Omega\|_{S_\infty} \leq \frac{16}{3} \log(D)$, in the last inequality. Inserting this above, we obtain

$$\begin{aligned}\|\eta\|_F^2 = \|\mathcal{P}_{T_k}(\eta)\|_F^2 + \|\mathcal{P}_{T_k^\perp}(\eta)\|_F^2 &\leq \left(\frac{2d_1d_2}{m}\frac{16^2}{3^2}\log(D)^2+1\right)\|\mathcal{P}_{T_k^\perp}(\eta)\|_F^2 \\ &\leq \left(\frac{2d_1d_2}{C\mu_0 r(d_1+d_2)\log(d_1+d_2)}\frac{16^2}{3^2}\log(D)^2+1\right)\|\mathcal{P}_{T_k^\perp}(\eta)\|_F^2 \\ &\leq \frac{\widetilde{C}d\log(D)}{\mu_0 r}\|\mathcal{P}_{T_k^\perp}(\eta)\|_F^2,\end{aligned}$$

where we used the sample complexity condition (3.37) in the second inequality and the definition

$$\widetilde{C} := \frac{4\cdot 16^2}{C\cdot 3^2}$$

for the constant $\widetilde{C}$.

Moreover, we observe that for $C_1 := \frac{\sqrt{C}}{320}\sqrt{\frac{30}{11}}$ where C is the constant of (3.37), it holds that

$$C_1\sqrt{\frac{\mu_0 r}{d}} \leq \frac{\sqrt{3}}{32}\frac{1}{10\sqrt{\log(D)(11/10)}}\sqrt{\frac{C\mu_0 r(d_1+d_2)\log(d_1+d_2)}{d_1d_2}} \leq \widetilde{\xi},$$

implying that the two statements of Lemma 3.5.7 are satisfied on the event $E_\Omega \cap E_{\Omega,T_0}$ if (3.23) holds. By the above-mentioned probability bounds and a union bound, $E_\Omega \cap E_{\Omega,T_0}$ occurs with a probability of at least $1-2D^{-2}$, finishing the proof for the sampling with replacement model. By the argument of [Rec11, Proposition 4], the result extends to the model of sampling locations drawn uniformly at random without replacement, with the same probability bound. This concludes the proof of Lemma 3.5.7. □

The following lemma will also play a role in the proof of Theorem 3.4.2.

Lemma 3.5.8. *Let $C, \widetilde{C}, C_1$ be the constants of Lemma 3.5.7 and μ_0 be the incoherence factor of a rank-r matrix $\mathbf{X}_0$. If*

$$m \geq C\mu_0 r(d_1+d_2)\log(d_1+d_2)$$

and if $\eta^{(k)} = \mathbf{X}^{(k)} - \mathbf{X}_0$ fulfills

$$\|\eta^{(k)}\|_{S_\infty} \leq \xi\sigma_r(\mathbf{X}_0),$$

with

$$\xi := \min\left(C_1\sqrt{\frac{\mu_0 r}{d}}, \frac{\mu_0}{4(1+6\kappa)d\log(D)\widetilde{C}}\right)$$

then, on the event of Lemma 3.5.7, it holds that

$$\|\eta^{(k)}\|_{S_\infty} < \sqrt{\frac{4\widetilde{C}d(d-r)\log(D)}{\mu_0 r}}\sigma_{r+1}(\mathbf{X}^{(k)}). \tag{3.43}$$

Proof. First, we compute that

$$\begin{aligned}\|\mathcal{P}_{T_k^\perp}(\eta^{(k)})\|_F &\leq \|\mathcal{P}_{T_k^\perp}(\mathbf{X}^{(k)})\|_F + \|\mathcal{P}_{T_k^\perp}(\mathbf{X}^0)\|_F \leq \sqrt{\sum_{i=r+1}^{d}\sigma_i^2(\mathbf{X}^{(k)})} + \left\|\mathbf{U}_\perp^{(k)}\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\mathbf{V}_\perp^{(k)*}\right\|_F \\ &\leq \sqrt{d-r}\sigma_{r+1}(\mathbf{X}^{(k)}) + \|\mathbf{U}_\perp^{(k)*}\mathbf{U}_0\|_{S_\infty}\|\mathbf{\Sigma}_0\|_F\|\mathbf{V}_0^*\mathbf{V}_\perp^{(k)}\|_{S_\infty} \\ &\leq \sqrt{d-r}\sigma_{r+1}(\mathbf{X}^{(k)}) + \frac{2\|\eta^{(k)}\|_{S_\infty}^2}{(1-\zeta)^2\sigma_r^2(\mathbf{X}_0)}\sqrt{r}\sigma_1(\mathbf{X}_0) \\ &= \sqrt{d-r}\sigma_{r+1}(\mathbf{X}^{(k)}) + \frac{2\|\eta^{(k)}\|_{S_\infty}^2}{(1-\zeta)^2\sigma_r(\mathbf{X}_0)}\sqrt{r}\kappa,\end{aligned}$$

where $0 < \zeta < 1$ such that $\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \leq \zeta\sigma_r(\mathbf{X}_0)$, using Lemma 3.5.9 twice in the fourth inequality and $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_{S_\infty}\|\mathbf{B}\|_F$, which holds for all matrices $\mathbf{A}$ and $\mathbf{B}$.

Using Lemma 3.5.7 for $\eta^{(k)} = \mathbf{X}^{(k)} - \mathbf{X}^0$, we obtain on the event on which the statement of Lemma 3.5.7 holds that

$$\begin{aligned}\|\eta^{(k)}\|_{S_\infty} \leq \|\eta^{(k)}\|_F &\leq \sqrt{\frac{\widetilde{C}d\log(D)}{\mu_0 r}}\|\mathcal{P}_{T_k^\perp}(\eta^{(k)})\|_F \\ &\leq \sqrt{\frac{\widetilde{C}d\log(D)}{\mu_0 r}}\left(\sqrt{d-r}\sigma_{r+1}(\mathbf{X}^{(k)}) + \frac{8\sqrt{r}\kappa\|\eta^{(k)}\|_{S_\infty}^2}{\sigma_r(\mathbf{X}_0)}\right) \\ &\leq \sqrt{\frac{\widetilde{C}d\log(D)}{\mu_0 r}}\left(\sqrt{d-r}\sigma_{r+1}(\mathbf{X}^{(k)}) + \frac{8\sqrt{r}\kappa\mu_0\sigma_r(\mathbf{X}_0)}{4(1+6\kappa)d\log(D)\widetilde{C}\sigma_r(\mathbf{X}_0)}\|\eta^{(k)}\|_{S_\infty}\right) \\ &= \sqrt{\frac{\widetilde{C}d(d-r)\log(D)}{\mu_0 r}}\sigma_{r+1}(\mathbf{X}^{(k)}) + \frac{1}{3}\sqrt{\frac{\mu_0}{\widetilde{C}d\log(D)}}\|\eta^{(k)}\|_{S_\infty}.\end{aligned}$$

Since $\mu_0 \leq \frac{d}{r}$, we have that $\frac{1}{3}\sqrt{\frac{\mu_0}{\widetilde{C}d\log(D)}} < \frac{1}{2}$, and therefore we obtain, after rearranging,

$$\left(1-\frac{1}{2}\right)\|\eta^{(k)}\|_{S_\infty} < \left(1-\frac{1}{3}\sqrt{\frac{\mu_0}{\widetilde{C}d\log(D)}}\right)\|\eta^{(k)}\|_{S_\infty} \leq \sqrt{\frac{\widetilde{C}d(d-r)\log(D)}{\mu_0 r}}\sigma_{r+1}(\mathbf{X}^{(k)}),$$

which implies the statement of this lemma. □

3.5.3 Weight operator and matrix perturbation

In the following, we use a well-known bound on perturbations of the singular value decomposition, which originally appeared in the work [Wed72], as an extension of a theorem established in the paper [DK70]. The result bounds the alignment of the subspaces spanned by the singular vectors of two matrices by their norm distance, given a gap between the first singular values of one matrix and the last singular values of the other matrix that is sufficiently pronounced.

Lemma 3.5.9 (Wedin's bound [Ste06])**.** *Let* $\mathbf{X}$ *and* $\widehat{\mathbf{X}}$ *be two matrices of the same size and their singular value decompositions*

$$\mathbf{X} = \begin{pmatrix} \mathbf{U} & \mathbf{U}_\perp \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma} & 0 \\ 0 & \mathbf{\Sigma}_\perp \end{pmatrix} \begin{pmatrix} \mathbf{V}^* \\ \mathbf{V}_\perp^* \end{pmatrix} \quad \textit{and} \quad \widehat{\mathbf{X}} = \begin{pmatrix} \widehat{\mathbf{U}} & \widehat{\mathbf{U}}_\perp \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{\Sigma}} & 0 \\ 0 & \widehat{\mathbf{\Sigma}}_\perp \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{V}}^* \\ \widehat{\mathbf{V}}_\perp^* \end{pmatrix},$$

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

$$\|\widehat{\mathbf{U}}_\perp^* \mathbf{U}\|_{S_\infty} \leq \sqrt{2} \frac{\|\mathbf{X} - \widehat{\mathbf{X}}\|_{S_\infty}}{\delta} \text{ and } \|\widehat{\mathbf{V}}_\perp^* \mathbf{V}\|_{S_\infty} \leq \sqrt{2} \frac{\|\mathbf{X} - \widehat{\mathbf{X}}\|_{S_\infty}}{\delta}. \tag{3.44}$$

We also use a lemma, which provides an explicit formula for calculating the new iterate $\mathbf{X}^{(k)}$ of `MatrixIRLS` and its characterization by optimality conditions. It is well-known in the IRLS literature, see, e.g., [DDFG10, Equation 1.9 and Lemma 5.2] or [FRW11b, Lemma 5.1], and is very general as it holds for any positive definite weight operator.

Lemma 3.5.10. *Let* $P_\Omega : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ *be the sampling operator, let* $\mathbf{y} \in \mathbb{R}^m$*. Let* $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ *be the weight operator of Definition 3.2.1 defined based on* $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$*. Then the solution of the weighted least squares step* (3.10) *of Algorithm 3 is unique and*

$$\mathbf{X}^{(k+1)} = \underset{P_\Omega(\mathbf{X}) = \mathbf{y}}{\arg\min} \langle \mathbf{X}, W^{(k)}(\mathbf{X}) \rangle = (W^{(k)})^{-1} P_\Omega^* \left(P_\Omega (W^{(k)})^{-1} P_\Omega^* \right)^{-1} (\mathbf{y}), \tag{3.45}$$

where $(W^{(k)})^{-1} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ *is the inverse matrix operator of* $W^{(k)}$*.*
Moreover, a matrix $\mathbf{X}^{(k+1)} \in \mathbb{R}^{d_1 \times d_2}$ *coincides with the one of* (3.45) *if and only if*

$$\langle W^{(k)}(\mathbf{X}^{(k+1)}), \eta \rangle = 0 \ \textit{ for all } \ \eta \in \ker P_\Omega \quad \textit{and} \quad P_\Omega(\mathbf{X}^{(k+1)}) = \mathbf{y}. \tag{3.46}$$

We show the following lemma. Wherever it appears, $\|\mathbf{X}\|_{S_1}$ denotes the nuclear norm $\|\mathbf{X}\|_{S_1} = \sum_{i=1}^d \sigma_i(\mathbf{X})$ of a matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$.

Lemma 3.5.11. *Let $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ be a matrix of rank r, let $\mathbf{X}^{(k)}$ be the k-th iterate of Algorithm 3 for input parameters Ω, $\mathbf{y} = P_\Omega(\mathbf{X}_0)$ and $\widetilde{r} = r$. Assume that $\epsilon_k = \sigma_{r+1}(\mathbf{X}^{(k)})$ and that*

$$\|\eta\|_F \leq c(\mu_0, r, d_1, d_2)\|\mathcal{P}_{T_k^\perp}\eta\|_F \qquad \text{for all } \eta \in \ker P_\Omega \tag{3.47}$$

for some constant $c(\mu_0, r, d_1, d_2)$ that may depend on μ_0, r, d_1, d_2, where $T_k = T_{\mathcal{T}_r(\mathbf{X}^{(k)})}\mathcal{M}_r$ is tangent space onto the manifold of rank-r matrices at $\mathcal{T}_r(\mathbf{X}^{(k)})$. Then

$$\|\mathbf{X}^{(k+1)} - \mathbf{X}_0\|_{S_\infty} \leq c(\mu_0, r, d_1, d_2)^2 \epsilon_k^2 \|W^{(k)}(\mathbf{X}_0)\|_{S_1}, \tag{3.48}$$

if $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ is the weight operator of Definition 3.2.1 corresponding to $\mathbf{X}^{(k)}$.

Proof of Lemma 3.5.11. Let $\eta^{(k+1)} := \mathbf{X}^{(k+1)} - \mathbf{X}_0$. Since $\eta^{(k+1)}$ is in the nullspace $\ker P_\Omega$, it follows from (3.47) that

$$\|\eta^{(k+1)}\|_{S_\infty}^2 \leq \|\eta^{(k+1)}\|_F^2 \leq c(\mu_0, r, d_1, d_2)^2 \|\mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\|_F^2. \tag{3.49}$$

Recalling the definition of the weight operator $W^{(k)} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ from Definition 3.2.1, i.e., $W^{(k)}(\mathbf{Z}) = \mathbf{U}_k \left[\mathbf{H}_k \circ (\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k)\right] \mathbf{V}_k^*$, we see that, if

$$\mathbf{X}^{(k)} = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^* = \begin{bmatrix} \mathbf{U}^{(k)} & \mathbf{U}_\perp^{(k)} \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}^{(k)} & 0 \\ 0 & \mathbf{\Sigma}_\perp^{(k)} \end{bmatrix} \begin{bmatrix} \mathbf{V}^{(k)*} \\ \mathbf{V}_\perp^{(k)*} \end{bmatrix} \tag{3.50}$$

is a singular value decomposition with $\mathbf{U}^{(k)} \in \mathbb{R}^{d_1 \times r}$, $\mathbf{U}_\perp^{(k)} \in \mathbb{R}^{d_1 \times (d_1 - r)}$, $\mathbf{V}^{(k)} \in \mathbb{R}^{d_2 \times r}$, $\mathbf{V}_\perp^{(k)} \in \mathbb{R}^{d_2 \times (d_2 - r)}$, we have that

$$\langle \mathbf{Z}, W^{(k)}(\mathbf{Z}) \rangle = \langle \mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k, \mathbf{H}_k \circ (\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k) \rangle \tag{3.51}$$

where $\mathbf{H}_k \in \mathbb{R}^{d_1 \times d_2}$ is as in Definition 3.2.1.

If $\mathbf{Z} = \mathcal{P}_{T_k^\perp}(\eta^{(k+1)}) \in T_k^\perp$, we know that $\mathbf{U}^{(k)*}\mathbf{Z} = 0$ and $\mathbf{Z}\mathbf{V}^{(k)} = 0$, and therefore

$$\mathbf{U}_k^* \mathbf{Z} \mathbf{V}_k = \begin{bmatrix} \mathbf{U}^{(k)*} \\ \mathbf{U}_\perp^{(k)*} \end{bmatrix} \mathbf{Z} \begin{bmatrix} \mathbf{V}^{(k)} & \mathbf{V}_\perp^{(k)} \end{bmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{U}_\perp^{(k)*} \mathbf{Z} \mathbf{V}_\perp^{(k)} \end{pmatrix}$$

with $\mathbf{U}_\perp^{(k)*} \mathbf{Z} \mathbf{V}_\perp^{(k)} \in \mathbb{R}^{(d_1 - r) \times (d_2 - r)}$.

By assumption of Lemma 3.5.11, we know that $\epsilon_k = \sigma_{r+1}(\mathbf{X}^{(k)})$, which means that $r_k := |\{i \in [d] : \sigma_i(\mathbf{X}^{(k)}) > \epsilon_k\}| = r$, and therefore $(\mathbf{H}_k)_{ij} = \epsilon_k^{-2}$ for all $i, j > r$. This entails with (3.51) that

$$\begin{aligned}\langle \mathcal{P}_{T_k^\perp}(\eta^{(k+1)}), W^{(k)}(\mathcal{P}_{T_k^\perp}(\eta^{(k+1)}))\rangle &= \epsilon_k^{-2}\langle \mathbf{U}_k^*\mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\mathbf{V}_k, \mathbf{U}_k^*\mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\mathbf{V}_k\rangle \\ &= \epsilon_k^{-2}\langle \mathcal{P}_{T_k^\perp}(\eta^{(k+1)}), \mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\rangle = \epsilon_k^{-2}\|\mathcal{P}_{T_k^\perp}(\eta^{(k+1)})\|_F^2,\end{aligned}$$

using the cyclicity of the trace and the fact that $\mathbf{U}_k$ and $\mathbf{V}_k$ are orthonormal matrices. Inserting this into (3.49), we obtain

$$\begin{aligned}\|\eta^{(k+1)}\|_{S_\infty}^2 &\leq c(\mu_0, r, d_1, d_2)^2\epsilon_k^2 \left\langle \mathcal{P}_{T_k^\perp}(\eta^{(k+1)}), W^{(k)}(\mathcal{P}_{T_k^\perp}(\eta^{(k+1)}))\right\rangle \\ &\leq c(\mu_0, r, d_1, d_2)^2\epsilon_k^2 \left\langle \eta^{(k+1)}, W^{(k)}(\eta^{(k+1)})\right\rangle,\end{aligned} \tag{3.52}$$

where the last inequality holds since $W^{(k)}$ is positive definite and since

$$\left\langle \mathcal{P}_{T_k^\perp}(\eta^{(k+1)}), W^{(k)}(\mathcal{P}_{T_k}(\eta^{(k+1)}))\right\rangle = 0$$

due to the orthogonality of T_k and $T_k^\perp$. Due to Lemma 3.5.10, we know that the new iterate $\mathbf{X}^{(k+1)}$ fulfills

$$0 = \langle W^{(k)}(\mathbf{X}^{(k+1)}), \eta^{(k+1)}\rangle = \langle W^{(k)}(\eta^{(k+1)} + \mathbf{X}_0), \eta^{(k+1)}\rangle,$$

and therefore

$$\left\langle \eta^{(k+1)}, W^{(k)}(\eta^{(k+1)})\right\rangle = -\left\langle W^{(k)}(\mathbf{X}_0), \eta^{(k+1)}\right\rangle \leq \|W^{(k)}(\mathbf{X}_0)\|_{S_1}\|\eta^{(k+1)}\|_{S_\infty},$$

using Hölder's inequality for Schatten-p (quasi-)norms [GGK12, Theorem 11.2]. Dividing (3.52) by $\|\eta^{(k+1)}\|_{S_\infty}$ concludes the proof of Lemma 3.5.11. □

To obtain a fast local convergence rate, it is crucial to bound $\|W^{(k)}(\mathbf{X}_0)\|_{S_1}$. For this, we split $\|W^{(k)}(\mathbf{X}_0)\|_{S_1}$ into three parts and estimate the parts separately by using the classical singular subspace perturbation result of Lemma 3.5.9.

Lemma 3.5.12. *Let $W^{(k)} : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ be the weight operator* (3.9) *of Definition 3.2.1 corresponding to $\mathbf{X}^{(k)}$, let $\epsilon_k = \sigma_{r+1}(\mathbf{X}^{(k)}) = \sigma_r^{(k)}$ and $\mathbf{X}_0 \in \mathbb{R}^{d_1\times d_2}$ be a rank-r matrix. Assume that there exists $0 < \zeta < 1$ such that*

$$\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \leq \zeta\sigma_r(\mathbf{X}_0). \tag{3.53}$$

Then

$$\left\|W^{(k)}(\mathbf{X}_0)\right\|_{S_1} \leq r(1-\zeta)^{-2}\sigma_r(\mathbf{X}_0)^{-1}\left(1+4\frac{\|\eta^{(k)}\|_{S_\infty}}{\epsilon_k}\frac{\sigma_1(\mathbf{X}_0)}{\sigma_r(\mathbf{X}_0)}+2\frac{\|\eta^{(k)}\|_{S_\infty}^2}{\epsilon_k^2}\frac{\sigma_1(\mathbf{X}_0)}{\sigma_r(\mathbf{X}_0)}\right).$$

Proof. Recalling the notation $\sigma_\ell^{(k)} = \sigma_\ell(\mathbf{X}^{(k)})$ for the ℓ-th singular value of $\mathbf{X}^{(k)}$ and the decomposition

$$\mathbf{H}_k = \begin{bmatrix} \mathbf{H}^{(k)} & \mathbf{H}_{1,2}^{(k)} \\ \mathbf{H}_{2,1}^{(k)} & \epsilon_k^{-2}\mathbf{1} \end{bmatrix} \tag{3.54}$$

of (3.21), we bound the entries of the different blocks $\mathbf{H}^{(k)}$, $\mathbf{H}_{1,2}^{(k)}$ and $\mathbf{H}_{2,1}^{(k)}$ separately. Since $(\mathbf{H}_k)_{ij} = \left(\max(\sigma_i^{(k)},\epsilon_k)\max(\sigma_j^{(k)},\epsilon_k)\right)^{-1}$ for each $i \in [d_1]$ and $j \in [d_2]$ due to definition of $\mathbf{H}_k$, we observe that

$$\max_{i\in[r],j\in[r]}(\mathbf{H}^{(k)})_{ij} \leq (\sigma_r^{(k)})^{-2}, \tag{3.55}$$

and

$$\max\left(\max_{i,j}((\mathbf{H}_{1,2}^{(k)})_{ij}), \max_{i,j}((\mathbf{H}_{2,1}^{(k)})_{ij})\right) = \max_{i\in[r],r+1\leq j\leq d_2}(\mathbf{H}_{1,2}^{(k)})_{ij} \leq (\sigma_r^{(k)})^{-1}\epsilon_k^{-1}, \tag{3.56}$$

In view of these entrywise bounds on the submatrices of $\mathbf{H}_k$ and $\mathbf{H}_2^{(k)}$, we compute, using (3.21), that

$$\begin{aligned}
\|W^{(k)}(\mathbf{X}_0)\|_{S_1} &= \left\|\begin{bmatrix}\mathbf{U}^{(k)} & \mathbf{U}_\perp^{(k)}\end{bmatrix}\left(\begin{bmatrix} \mathbf{H}^{(k)} & \mathbf{H}_{1,2}^{(k)} \\ \mathbf{H}_{2,1}^{(k)} & \epsilon_k^{-2}\mathbf{1} \end{bmatrix} \circ \begin{bmatrix}\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)} & \mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)} \\ \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)} & \mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\end{bmatrix}\right)\begin{bmatrix}\mathbf{V}^{(k)*} \\ \mathbf{V}_\perp^{(k)*}\end{bmatrix}\right\|_{S_1} \\
&\leq \left\|\mathbf{U}^{(k)}[\mathbf{H}^{(k)} \circ (\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)})]\mathbf{V}^{(k)*}\right\|_{S_1} + \left\|\mathbf{U}_k\begin{bmatrix} 0 & \mathbf{H}_{1,2}^{(k)} \circ (\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}) \\ \mathbf{H}_{2,1}^{(k)} \circ (\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)}) & 0\end{bmatrix}\mathbf{V}_k^*\right\|_{S_1} \\
&+ \epsilon_k^{-2}\left\|\mathbf{U}_\perp^{(k)}\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\mathbf{V}_\perp^{(k)*}\right\|_{S_1} =: \text{(I)} + \text{(II)} + \text{(III)}.
\end{aligned}$$

We now bound the terms (I), (II) and (III) separately.
First, we see that

$$\begin{aligned}
\text{(I)} &= \left\|\mathbf{H}^{(k)} \circ (\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)})\right\|_{S_1} \leq \sqrt{r}\left\|\mathbf{H}^{(k)} \circ (\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)})\right\|_F \\
&\leq \sqrt{r}\left\|\mathbf{H}^{(k)} \circ (\mathbf{U}^{(k)*}\mathbf{X}^{(k)}\mathbf{V}^{(k)})\right\|_F + \sqrt{r}\left\|\mathbf{H}^{(k)} \circ (\mathbf{U}^{(k)*}\eta^{(k)}\mathbf{V}^{(k)})\right\|_F \\
&\leq \sqrt{r}\left\|\mathbf{H}^{(k)} \circ \mathbf{\Sigma}^{(k)}\right\|_F + \sqrt{r}(\sigma_r^{(k)})^{-2}\|\mathbf{U}^{(k)*}\eta^{(k)}\mathbf{V}^{(k)}\|_F,
\end{aligned}$$

where we used the Cauchy-Schwarz inequality in the first inequality, the notation $\eta^{(k)} =$

$\mathbf{X}^{(k)} - \mathbf{X}_0$ and the triangle inequality in the second inequality, and finally, (3.55) in the third inequality. $\Sigma^{(k)}$ is here as in (3.13).

Since

$$\left\|\mathbf{H}^{(k)} \circ \mathbf{\Sigma}^{(k)}\right\|_F = \left(\sum_{i=1}^{r} (\sigma_i^{(k)})^{-2}\right)^{1/2} \leq \sqrt{r}(\sigma_r^{(k)})^{-1}$$

and

$$\|\mathbf{U}^{(k)*}\eta^{(k)}\mathbf{V}^{(k)}\|_F \leq \sqrt{r}\|\mathbf{U}^{(k)*}\eta^{(k)}\mathbf{V}^{(k)}\|_{S_\infty} \leq \sqrt{r}\|\eta^{(k)}\|_{S_\infty} \leq \sqrt{r}\zeta\sigma_r(\mathbf{X}_0)$$

from assumption (3.53), it follows then that

$$\text{(I)} \leq r(\sigma_r^{(k)})^{-2}\left(\sigma_r^{(k)} + \zeta\sigma_r(\mathbf{X}_0)\right).$$

We can use the proximity assumption (3.53) further to get rid of the dependence on k in the bound, as

$$\sigma_r(\mathbf{X}_0) = \sigma_r(\mathbf{X}^{(k)} - \eta^{(k)}) \leq \sigma_r^{(k)} + \sigma_1(\eta^{(k)}) = \sigma_r^{(k)} + \|\eta^{(k)}\|_{S_\infty} \leq \sigma_r^{(k)} + \zeta\sigma_r(\mathbf{X}_0),$$

using $\sigma_{i+j-1}(\mathbf{A}) \leq \sigma_i(\mathbf{A}+\mathbf{B}) + \sigma_j(\mathbf{B})$ for any i, j (cf. Theorem 3.3.16 of [HHJ94]) with $\mathbf{A}+\mathbf{B} = X^{(k)} - \eta^{(k)}$ and $\mathbf{B} = \eta^k$ so that

$$\sigma_r^{(k)} \geq (1-\zeta)\sigma_r(\mathbf{X}_0), \tag{3.57}$$

and hence

$$\text{(I)} \leq r\sigma_r(\mathbf{X}_0)^{-2}(1-\zeta)^{-2}\left(\sigma_r(\mathbf{X}_0)(1-\zeta) + \zeta\sigma_r(\mathbf{X}_0)\right) = r(1-\zeta)^{-2}\sigma_r(\mathbf{X}_0)^{-1}. \tag{3.58}$$

For the term (II), we compute that

$$\begin{aligned}\text{(II)} &\leq \sqrt{2r}\left\|\begin{bmatrix} 0 & \mathbf{H}_{1,2}^{(k)} \circ (\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}) \\ \mathbf{H}_{2,1}^{(k)} \circ (\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)}) & 0 \end{bmatrix}\right\|_F \\ &\leq \sqrt{2r}(\sigma_r^{(k)})^{-1}\epsilon_k^{-1}\left(\left\|\mathbf{U}^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\right\|_F + \left\|\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}^{(k)}\right\|_F\right) \\ &\leq \sqrt{2r}(\sigma_r^{(k)})^{-1}\epsilon_k^{-1}\left(\|\mathbf{U}^{(k)*}\mathbf{U}_0\mathbf{\Sigma}_0\|_F\|\mathbf{V}_0^*\mathbf{V}_\perp^{(k)}\|_{S_\infty} + \|\mathbf{U}_\perp^{(k)*}\mathbf{U}_0\|_{S_\infty}\|\mathbf{\Sigma}_0\mathbf{V}_0^*\mathbf{V}^{(k)}\|_F\right),\end{aligned}$$

using the singular value decomposition $\mathbf{X}_0 = \mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^*$ of the rank-r matrix $\mathbf{X}_0$ with $\mathbf{U}_0 \in \mathbb{R}^{d_1 \times r}$, $\mathbf{V}_0 \in \mathbb{R}^{d_2 \times r}$. This allows us to use the singular subspace perturbation result of Lemma 3.5.9, so that $\|\mathbf{V}_0^*\mathbf{V}_\perp^{(k)}\|_{S_\infty}$ and $\|\mathbf{U}_\perp^{(k)*}\mathbf{U}_0\|_{S_\infty}$ can compensate for the negative power of the ϵ_k, avoiding a blow-up of the term (II): Indeed, using Lemma 3.5.9 with

$\mathbf{X} = \mathbf{X}_0$, $\widehat{\mathbf{X}} = \mathbf{X}^{(k)}$, $\alpha = \sigma_r(\mathbf{X}_0)$ and $\delta = (1-\zeta)\sigma_r(\mathbf{X}_0)$ results in

$$\max(\|\mathbf{V}_0^*\mathbf{V}_\perp^{(k)}\|_{S_\infty}, \|\mathbf{U}_\perp^{(k)*}\mathbf{U}_0\|_{S_\infty}) \leq \frac{\sqrt{2}\|\eta^{(k)}\|_{S_\infty}}{(1-\zeta)\sigma_r(\mathbf{X}_0)},$$

and since $\|\mathbf{U}^{(k)*}\mathbf{U}_0\mathbf{\Sigma}_0\|_F \leq \|\mathbf{\Sigma}_0\|_F \leq \sqrt{r}\sigma_1(\mathbf{X}_0)$, $\|\mathbf{\Sigma}_0\mathbf{V}_0^*\mathbf{V}^{(k)}\|_F \leq \sqrt{r}\sigma_1(\mathbf{X}_0)$, we obtain with (3.57) that

$$(\mathrm{II}) \leq 4r(1-\zeta)^{-2}\sigma_r(\mathbf{X}_0)^{-1}\frac{\|\eta^{(k)}\|_{S_\infty}}{\epsilon_k}\frac{\sigma_1(\mathbf{X}_0)}{\sigma_r(\mathbf{X}_0)}. \tag{3.59}$$

It remains to bound the last term (III). For (III), we can use the subspace perturbation lemma *twice* in the same summand such that

$$\begin{aligned}(\mathrm{III}) &= \epsilon_k^{-2}\left\|\mathbf{U}_\perp^{(k)}\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\mathbf{V}_\perp^{(k)*}\right\|_{S_1} = \epsilon_k^{-2}\|\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\|_{S_1} \leq \sqrt{r}\epsilon_k^{-2}\|\mathbf{U}_\perp^{(k)*}\mathbf{X}_0\mathbf{V}_\perp^{(k)}\|_F \\ &\leq \sqrt{r}\epsilon_k^{-2}\|\mathbf{U}_\perp^{(k)*}\mathbf{U}_0\|_{S_\infty}\|\mathbf{\Sigma}_0\|_F\|\mathbf{V}_0^*\mathbf{V}_\perp^{(k)}\|_{S_\infty} \\ &\leq \sqrt{r}\epsilon_k^{-2}\frac{\sqrt{2}\|\eta^{(k)}\|_{S_\infty}}{(1-\zeta)\sigma_r(\mathbf{X}_0)}\sqrt{r}\sigma_1(\mathbf{X}_0)\frac{\sqrt{2}\|\eta^{(k)}\|_{S_\infty}}{(1-\zeta)\sigma_r(\mathbf{X}_0)} = 2r(1-\zeta)^{-2}\sigma_r(\mathbf{X}_0)^{-1}\frac{\|\eta^{(k)}\|_{S_\infty}^2}{\epsilon_k^2}\frac{\sigma_1(\mathbf{X}_0)}{\sigma_r(\mathbf{X}_0)}.\end{aligned} \tag{3.60}$$

Combining Equation (3.58), Equation (3.59) and Equation (3.60) finally yields the statement of Lemma 3.5.12. □

The bound on perturbations of the singular value decomposition, Lemma 3.5.9, used above are not optimal since it is uniform for both the left and right singular spaces. In particular, the bounds used here lead to a quite small basin of attraction for our algorithm. Recently, some optimal perturbation bounds were developed, see [CZ18] and the discussion in [FWZ18].

Open Problem: Is it possible to develop unbalanced weights to complete matrices such that the local basin of attraction has a larger radius? Can the results from [CZ18] be used to establish such results?

3.5.4 Connecting the dots and finishing the proof

We can now put Lemma 3.5.7, Lemma 3.5.11 and Lemma 3.5.12 together to prove the local convergence statement of Theorem 3.4.2, also showing that we attain locally quadratic convergence.

Proof of Theorem 3.4.2. Let $k = k_0$ and $\mathbf{X}^{(k)}$ be the k-th iterate of `MatrixIRLS` with the parameters stated in Theorem 3.4.2. Under the sampling model of Theorem 3.4.2, if the number of samples m fulfills $m \geq C\mu_0 r(d_1+d_2)\log(d_1+d_2)$, where C is the constant of Lemma 3.5.7, we know from Lemma 3.5.7 that with a probability of at least $1-2D^{-2}$, inequality (3.47) is satisfied with $c(\mu_0, r, d_1, d_2) = \sqrt{\frac{\widetilde{C}d\log(D)}{\mu_0 r}}$, if furthermore $\eta^{(k)} := \mathbf{X}^{(k)} - \mathbf{X}_0$ fulfills

$$\|\eta^{(k)}\|_{S_\infty} \leq \xi \sigma_r(\mathbf{X}_0) \tag{3.61}$$

with

$$\xi \leq C_1 \sqrt{\frac{\mu_0 r}{d}}, \tag{3.62}$$

and thus, by Lemma 3.5.11,

$$\|\mathbf{X}^{(k+1)} - \mathbf{X}_0\|_{S_\infty} \leq \frac{\widetilde{C}d\log(D)}{\mu_0 r} \epsilon_k^2 \|W^{(k)}(\mathbf{X}_0)\|_{S_1}. \tag{3.63}$$

We denote the event that this is fulfilled by E. Furthermore, on this event, if $\xi \leq 1/2$ in (3.61) and denoting the condition number by $\kappa = \sigma_1(\mathbf{X}_0)/\sigma_r(\mathbf{X}_0)$, it follows from Lemma 3.5.12 that

$$\|\mathbf{X}^{(k+1)} - \mathbf{X}_0\|_{S_\infty} \leq \frac{\widetilde{C}d\log(D)}{\mu_0} 4\sigma_r(\mathbf{X}_0)^{-1} \left(\epsilon_k^2 + 4\epsilon_k \|\eta^{(k)}\|_{S_\infty}\kappa + 2\|\eta^{(k)}\|_{S_\infty}^2 \kappa\right)$$

Furthermore, if $\mathbf{X}_r^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ denotes the best rank-r approximation of $\mathbf{X}^{(k)}$ in any unitarily invariant norm, we estimate that

$$\epsilon_k \leq \sigma_{r+1}(\mathbf{X}^{(k)}) = \|\mathbf{X}^{(k)} - \mathbf{X}_r^{(k)}\|_{S_\infty} \leq \|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} = \|\eta^{(k)}\|_{S_\infty},$$

Inserting these two bounds into (3.63), we obtain

$$\|\eta^{(k+1)}\|_{S_\infty} = \|\mathbf{X}^{(k+1)} - \mathbf{X}_0\|_{S_\infty} \leq \frac{\widetilde{C}d\log(D)}{\mu_0} 4\sigma_r(\mathbf{X}_0)^{-1} (1+6\kappa) \|\eta^{(k)}\|_{S_\infty}^2.$$

Finally, if, additionally, (3.61) is satisfied for

$$\xi \leq \frac{\mu_0}{4(1+6\kappa)d\log(D)\widetilde{C}}, \tag{3.64}$$

we conclude that

$$\|\eta^{(k+1)}\|_{S_\infty} < \|\eta^{(k)}\|_{S_\infty}$$

and also, we observe a quadratic decay in the spectral error such that

$$\|\eta^{(k+1)}\|_{S_\infty} \leq \mu \|\eta^{(k)}\|_{S_\infty}^2$$

with a constant $\mu = \frac{4\widetilde{C}d\log(D)(1+6\kappa)}{\mu_0 \sigma_r(\mathbf{X}_0)}$. This shows inequality (3.23) of Theorem 3.4.2.

To show the remaining statement, we can use Lemma 3.5.8 to show that if $\mathbf{X}^{(k)}$ is close enough to $\mathbf{X}_0$, we can ensure that the $(r+1)$-st singular value $\sigma_{r+1}(\mathbf{X}^{(k)})$ of the current iterate is strictly decreasing. More precisely, assume now the stricter assumption of

$$\|\eta^{(k)}\|_{S_\infty} \leq \sqrt{\frac{\mu_0 r}{4\widetilde{C}d(d-r)\log(D)}}\xi\sigma_r(\mathbf{X}_0). \tag{3.65}$$

In fact, if ξ fulfills (3.62) and (3.64), we can conclude that on the event E,

$$\begin{aligned}\sigma_{r+1}(\mathbf{X}^{(k+1)}) \leq \|\eta^{(k+1)}\|_{S_\infty} &\leq \frac{\widetilde{C}d\log(D)}{\mu_0}4\sigma_r(\mathbf{X}_0)^{-1}\left(1+6\kappa\right)\|\eta^{(k)}\|_{S_\infty}\cdot\|\eta^{(k)}\|_{S_\infty}\\ &< \frac{\widetilde{C}d\log(D)}{\mu_0}4\sigma_r(\mathbf{X}_0)^{-1}\left(1+6\kappa\right)\sqrt{\frac{\mu_0 r}{4\widetilde{C}d(d-r)\log(D)}}\xi\sigma_r(\mathbf{X}_0)\\ &\cdot\sqrt{\frac{4\widetilde{C}d(d-r)\log(D)}{\mu_0 r}}\sigma_{r+1}(\mathbf{X}^{(k)}) \leq \sigma_{r+1}(\mathbf{X}^{(k)})\end{aligned}$$

using Lemma 3.5.8 for one factor $\|\eta^{(k)}\|_{S_\infty}$ and (3.65) for the other factor $\|\eta^{(k)}\|_{S_\infty}$ in the third inequality, and (3.64) in the last inequality. Taking the update rule (3.11) for the smoothing parameter into account, this implies that $\epsilon_{k+1} = \sigma_{r+1}(\mathbf{X}^{(k+1)})$, which ensures that the first statement of Theorem 3.4.2 is fulfilled likewise for iteration $k+1$. By induction, this implies that $\mathbf{X}^{(k+\ell)} \xrightarrow{\ell\to\infty} \mathbf{X}_0$, which finishes the proof of Theorem 3.4.2. □

Some of the ingredients used in the proof of Theorem 3.4.2 are similar to the *local superlinear convergence* of [KS18, Theorem 11] when the objective function is the Schatten-p and also have similarities with techniques developed for other matrix completion algorithms [CWW19]. The proof presented in [KS18], however, relies on the null space property, which does not hold in the matrix completion setting, and cannot be extended to the limit of Schatten-p quasi-norm, namely, the log-det function, as we did here. Moreover, the previous versions of the IRLS algorithm developed in [FRW11b, MF12b] cannot have superlinear convergence, which is also not observed in practice. One way to explain that is that with their choice of weights for the weighted least square problem, the estimations developed in Lemma 3.5.12 would be too large.

3.6 Computational Aspects

The main contributions of this chapter are on the theoretical quadratic convergence and on the numerical comparison of state-of-the-art methods for matrix completion. However, for the sake of completeness, we also add some computational aspects of `MatrixIRLS`.

These aspects were developed in [KMV21, Appendix A and Appendix C] and [Küm19]. In a similar way to what was established in Section 2.2.1, under the uniform sampling model, it is possible to prove that, with high probability, the underlying linear system from `MatrixIRLS`, described in Equation (3.10), is well-conditioned. This shows that a very accurate solution of this system can be obtained with an iterative solver like the conjugate gradient one in just a few iterations. In particular, the following theorem can be established:

Theorem 3.6.1 (Well-conditioning of system matrices of `MatrixIRLS`). *In the setup and sampling model of Theorem 3.6.2, if* $m \gtrsim \mu_0 r(d_1+d_2)\log(d_1+d_2)$*, the following holds with high probability: If* $\epsilon_k = \sigma_{r+1}(\mathbf{X}^{(k)}) < \sigma_r(\mathbf{X}^{(k)})$ *and if* $\|\mathbf{X}^{(k)} - \mathbf{X}_0\|_{S_\infty} \lesssim \min\left(\sqrt{\frac{\mu_0 r}{d}}, \frac{1}{4}\right)\sigma_r(\mathbf{X}_0)$*, the matrix* $\mathbf{A}_k \in \mathbb{R}^{r(d_1+d_2-r)\times r(d_1+d_2-r)}$ *in the linear system matrix of the weighted least squares step* (3.10) *has a spectrum* $\lambda(\mathbf{A}_k)$ *that satisfies* $\lambda(\mathbf{A}_k) \subset \frac{m}{d_1 d_2}\left[\frac{6}{10}; \frac{24}{10}\right]$*, and thus, the condition number of* $\mathbf{A}_k$ *fulfills* $\kappa(\mathbf{A}_k) \leq 4$.

Similarly to what happened in our IRLS scheme designed for sparse recovery in Chapter 2, this result shows that `MatrixIRLS` is able to successfully overcome a prevalent issue encountered by numerous IRLS algorithms. Indeed, our choice of objective function and weights induce the construction of a linear system that is not ill-conditioned like the linear systems presented in [DDFG10, FPRW16b, MF12a, FRW11b, KS18]. Still, the proof of this theorem relies on probabilistic arguments and the uniform sampling model. Therefore, we propose the following question.

Open Problem: Is it possible to obtain proof for the well-conditioning of the linear system of `MatrixIRLS` when a deterministic sampling pattern is assumed?

A crucial property of Algorithm 3 is that due to the structure of the weight operator (3.9) and the smoothing update rule (3.11), in fact, the weighted least squares step (3.10) can be computed by solving a positive definite linear system of size $(r_k(d_1+d_2-r_k)) \times (r_k(d_1+d_2-r_k))$, where r_k is the number of singular values of $\mathbf{X}^{(k)}$ that are larger than ϵ_k, which is typically equal or very close to $\widetilde{r}$. Conceptually, this corresponds to a linear system in the tangent space T_k of the rank-r_k matrix manifold at the best rank-r_k approximation of $\mathbf{X}^{(k)}$, $T_k = \left\{ \begin{bmatrix} \mathbf{U}^{(k)} \mathbf{U}_\perp^{(k)} \end{bmatrix} \begin{bmatrix} \mathbb{R}^{r_k \times r_k} & \mathbb{R}^{r_k(d_2-r_k)} \\ \mathbb{R}^{(d_1-r_k)r_k} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}^{(k)} \mathbf{V}_\perp^{(k)} \end{bmatrix}^* \right\}$.

In our implementation, it is noteworthy that the computation of more than r_k singular vector pairs and singular values of $\mathbf{X}^{(k)}$ is never required. Moreover, the matrix $\mathbf{X}^{(k)}$ can be represented as a sum of a sparse matrix and a matrix in T_k. Thus, when using an iterative solver such as *conjugate gradients* to solve the linear system, as discussed above, we obtain an implementation of `MatrixIRLS` with a time and space complexity of the same order as for state-of-the-art first-order algorithms based on matrix factorization (i.e., of Burer-Monteiro type) [CC18]. We refer to [KMV21, Appendix A] for a proof and an extensive discussion.

Theorem 3.6.2. *Let $\mathbf{X}^{(k)} \in \mathbb{R}^{d_1 \times d_2}$ be the k-th iterate of* `MatrixIRLS` *for an observation vector $\mathbf{y} \in \mathbb{R}^m$ and $\widetilde{r} = r$. Assume that $\sigma_i^{(k)} \leq \epsilon_k$ for all $i > r$ and $\sigma_r^{(k)} > \epsilon_k$. Then an implicit representation of the new iterate $\mathbf{X}^{(k+1)} \in \mathbb{R}^{d_1 \times d_2}$ can be calculated in a* time complexity *of*

$$O\left((mr + r^2 D) \cdot N_{CG\_inner}\right),$$

where N_{CG_inner} is the number of inner iterations used in the conjugate gradient method and $D = \max(d_1, d_2)$. More precisely, $\mathbf{X}^{(k+1)}$ can be represented as

$$\mathbf{X}^{(k+1)} = P_\Omega^*(\mathbf{r}_{k+1}) + \mathbf{U}^{(k)} \mathbf{M}_1^{(k+1)*} + \mathbf{M}_2^{(k+1)} \mathbf{V}^{(k)*},$$

where $\mathbf{r}_{k+1} \in \mathbb{R}^m$, $\mathbf{M}_1^{(k+1)} \in \mathbb{R}^{d_2 \times r}$ and $\mathbf{M}_2^{(k+1)} \in \mathbb{R}^{d_1 \times r}$, i.e., with a space complexity *of $O(m + rD)$.*

Theorem 3.6.2 showcases the computational superiority of `MatrixIRLS` over previous iteratively reweighted least squares algorithms when it comes to solving low-rank matrix recovery problems [FRW11b, MF12a, KS18], which all require the storage and updates of full $(d_1 \times d_2)$-matrices and the calculation of singular value decompositions of these.

According to Theorem 3.6.2, since $P_\Omega^*(\mathbf{r}_{k+1}) \in \mathbb{R}^{d_1 \times d_2}$ is m-sparse, $\mathbf{X}^{(k+1)}$ can be seen a sum of a sparse and two rank-r matrices. Intuitively, this representation is possible as the weight operator $W^{(k)}$ of Definition 3.2.1 can be written as "identity + diagonal on T_k", and due to the Sherman-Morrison-Woodbury formula applied to the inverse in $\mathbf{X}^{(k+1)} = (W^{(k)})^{-1} P_\Omega^* \left(P_\Omega (W^{(k)})^{-1} P_\Omega^*\right)^{-1}(\mathbf{y})$, which is an explicit representation of the solution of Equation (3.10).

As a result, fast matrix-vector multiplications can be used to explore the full potential of methods such as Lanczos bidiagonalization [SZ00] or randomized Block Krylov [MM15] to compute r_{k+1} singular values and vectors of $\mathbf{X}^{(k+1)}$ in step 3 of Algorithm 3.

3.7 Numerical Experiments

In this section, we finally explore the performance of `MatrixIRLS` for completing synthetic low-rank matrices in terms of statistical and computational efficiency compared to state-of-the-art algorithms in the literature. Our selection criteria are guided by the aim of capturing a comprehensive overview of state-of-the-art algorithms for matrix completion.

This includes algorithms demonstrating scalability for high-dimensional problems, those accompanied by the best theoretical guarantees, and those specifically designed to perform well in completing *ill-conditioned* matrices. All the methods are provided with the true rank r of $\mathbf{X}_0$ as an input parameter. If possible, we use the MATLAB implementation provided by the authors of the respective papers, as described in Section 3.7.1.

We opted for a diverse selection of algorithms to test against `MatrixIRLS`. These methods can be grouped into three main categories: the non-convex matrix factorization ones which include `LMaFit` [WYZ12], `ScaledASD` [TW16] and `ScaledGD` [TMC21], the Riemannian optimization on the manifold of fixed rank matrices ones which include `LRGeomCG` [Van13], `RTRMC` [BA15] and `R3MC` [MS14], one alternating projection method on the manifold of fixed rank matrices, `NIHT` [TW13] (see [WCCL20] for a connection between NIHT and Riemannian methods), and the recent `R2RILS` [BNZ21] which can be seen as a factorization based method but also contains ideas from the Riemannian optimization family of algorithms.

As discussed at the beginning of this chapter, we are interested in completing matrices in the information-theoretical regime. This means that we are interested in finding low-rank completions from a sampling set Ω of sample size $|\Omega| =: m = \lfloor \rho r(d_1 + d_2 - r) \rfloor$, where ρ is an oversampling ratio since $r(d_1 + d_2 - r)$ is just the number of degrees of freedom of an $(d_1 \times d_2)$-dimensional rank-r matrix. For a given Ω, the solution of Equation (3.2) might *not* coincide with $\mathbf{X}_0$, or the solution might not be unique, even if the sample set Ω is chosen uniformly at random. In particular, this will be the case if Ω is such that there is a row or a column with *fewer than* r revealed entries, which is a necessary condition for the uniqueness of the Equation (3.2), see the discussion in [PABN16]. To mitigate this problem, which is rather related to the structure of the sampling set rather than to the performance of a certain algorithm, we, in fact, adapt the sampling model of uniform sampling without replacement. For a given factor $\rho \geq 1$, we sample a set $\Omega \subset [d_1] \times [d_2]$ of size $m = \lfloor \rho r(d_1 + d_2 - r) \rfloor$ indices randomly without replacement. Then we check whether the condition such that each row and each column in Ω has at least r observed entries, and resample Ω if this condition is not fulfilled. This procedure is repeated up to a maximum of 1000 resamplings.

We consider the following setup: we sample a pair of random matrices $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$ with r orthonormal columns and define the diagonal matrix $\Sigma \in \mathbb{R}^{r \times r}$ such that $\Sigma_{ii} = \kappa \exp(-\log(\kappa)\frac{i-1}{r-1})$ for $i \in [r]$. With this definition, we define a ground truth matrix $\mathbf{X}_0 = \mathbf{U}\Sigma\mathbf{V}^*$ of rank r that has exponentially decaying singular values between κ and 1.

3.7.1 Algorithmic Parameter Choice

This section provides a comprehensive description of the numerical experiments, including the corresponding download links for the implementations. While explicit parallelization is not utilized in any of the methods, many of them incorporate compiled C subroutines to efficiently execute sparse evaluations of matrix factorizations. We set the maximum number of outer iterations for the second-order methods to $N_0 = 400$. The second-order type algorithms considered for this paper, including their parameter choices, are:

- `MatrixIRLS`, as described in Algorithm 3. As a stopping criterion, we choose a threshold of 10^{-9} for the relative change of the Frobenius norm $\frac{\|\mathbf{X}^{(k+1)}-\mathbf{X}^{(k)}\|_F}{\|\mathbf{X}^{(k)}\|_F}$. We use the conjugate gradient (CG) method for solving the linear system (2.17) without any preconditioning. We terminate the CG method if a maximum number of $N_{\text{CG\_inner}} = 500$ inner iterations is reached or if a relative residual of $\text{tol}_{\text{inner}} = 10^{-9}$ is reached, whichever happens first.[6] For the weight operator update step, we use a variant of the randomized Block Krylov method [MM15] based on the implementation provided by the authors[7], setting the parameter for the maximal number of iterations to 20.

- `R2RILS` [BNZ21] or *rank* $2r$ *iterative least squares*, a method that optimizes a least squares data fit objective $\|P_\Omega(\mathbf{X}_0) - P_\Omega(\mathbf{X})\|_F$ over $\mathbf{X} \in T_{\mathbf{Z}^{(k)}}\mathcal{M}_r$, where $T_{\mathbf{Z}^{(k)}}\mathcal{M}_r$ is a tangent space onto the manifold of rank-r matrices, while iteratively updating this tangent space. As above, we stop the outer iterations a threshold of 10^{-9} is reached for the relative change of the Frobenius norm $\frac{\|\mathbf{X}^{(k+1)}-\mathbf{X}^{(k)}\|_F}{\|\mathbf{X}^{(k)}\|_F}$. At each outer iteration, `R2RILS` solves an overdetermined least squares problem of size $(m \times r(d_1+d_2))$ via the iterative solver LSQR, for which we choose the maximal number of inner iterations as $N_{\text{LSQR\_inner}} = 500$ and a termination criterion based on a relative residual of 10^{-10}. We use the implementation based on the author's code but adapted for these stopping criteria.[8]

- `RTRMC`, the preconditioned Riemannian trust-region method called RTRMC 2p of [BA15], which was reported to achieve the best performance among a variety of matrix completion algorithms for the task of completing matrices of a condition number of up to $\kappa = 150$. We use the implementation provided by the authors[9] with default

[6] While this stopping condition uses the condition number κ, which will probably be unknown in practice, it can be generally chosen independently of κ without any convergence issues.

[7] `https://github.com/cpmusco/bksvd`

[8] `https://github.com/Jonathan-WIS/R2RILS`

[9] RTRMC v3.2 from `http://web.math.princeton.edu/~nboumal/RTRMC/index.html`, together with the toolbox Manopt 6.0 (`https://www.manopt.org/`) [BMAS14].

options except from setting the maximal number of inner iterations to $N_{\text{inner}} = 500$ and setting the parameter for the tolerance on the gradient norm to 10^{-15}. Furthermore, as the algorithm otherwise would often run into certain submatrices that are not positive definite for ρ between 1 and 1.5, we set the regularization parameter $\lambda = 10^{-8}$, which is small enough not to deter high precision approximations of $\mathbf{X}_0$ if enough samples are provided.

Furthermore, we consider the following first-order algorithms, setting the maximal number of outer iterations to $N_0 = 4000$:

- `LRGeomCG` [Van13], a local optimization method for a quadratic data fit term based on gradients with respect to the Riemannian manifold of fixed rank matrices. We use the author's implementation[10] while setting the parameters related to the stopping conditions `abs_grad_tol`, `rel_grad_tol`, `abs_f_tol`, `rel_f_tol`, `rel_tol_change_x` and `rel_tol_change_res` each to 10^{-9}. The rank-adaptive variant of `LRGeomCG`, called `LRGeomCG Pursuit` [UV15, TTW+14], is used with the same algorithmic parameters as `LRGeomCG` for the inner iterations, and with a rank increase of 1 each outer iteration.

- `LMaFit` or low-rank matrix fitting [WYZ12], a nonlinear successive over-relaxation algorithm based on matrix factorization. We use the implementation provided by the authors[11], setting the tolerance threshold for the stopping condition (which is based on a relative data fit error $\|P_\Omega(\mathbf{X}^{(k)}) - \mathbf{y}\|_2/\|\mathbf{y}\|_2$) to $5 \cdot 10^{-10}$.

- `ScaledASD` or scaled alternating steepest descent [TW16], a gradient descent method based on matrix factorization which scales the gradients in a quasi-Newton fashion. We use the implementation provided by the authors[12] with the stopping condition given by $\|P_\Omega(\mathbf{X}^{(k)}) - \mathbf{y}\|_2/\|\mathbf{y}\|_2 \leq 10^{-9}$.

- `ScaledGD` or scaled gradient descent [TMC21], a method that is very similar to `ScaledASD`, but for which a non-asymptotic local convergence analysis has been established for matrix completion. We use an adapted version of the author's implementation[13]: We choose a step size of $\eta = 0.5$, but increase the normalization parameter p by a factor of 1.5 in case the unmodified algorithm `ScaledGD` leads to divergent algorithmic iterates, using the same stopping condition as for `ScaledASD`.

[10] `http://www.unige.ch/math/vandereycken/matrix_completion.html`

[11] `http://lmafit.blogs.rice.edu`

[12] `http://www.sdspeople.fudan.edu.cn/weike/code/mc20140528.tar`

[13] `https://github.com/Titan-Tong/ScaledGD`

- `NIHT` or normalized iterative hard thresholding [TW13], which performs iterative hard thresholding steps with adaptive step sizes. We use the implementation provided by the authors [12] with a stopping threshold of 10^{-9} for the relative data fit error $\|P_\Omega(\mathbf{X}^{(k)}) - \mathbf{y}\|_2/\|\mathbf{y}\|_2$ and for the convergence rate threshold parameter.

- `R3MC` [MS14], a Riemannian nonlinear conjugate-gradient that also optimizes a least squares data fit objective $\|P_\Omega(\mathbf{X}_0) - P_\Omega(\mathbf{X})\|_F$ by exploiting a three-factor matrix factorization similar to the SVD. The algorithm performs a search on a quotient manifold defined from the manifold of rank r matrices and uses the factorization and symmetries from the action of the orthogonal group. We use the author's implementation[14] while choosing the Polyak-Ribier rule for the nonlinear CG and the Armijo line search with a maximum of 50 line searches allowed at each iteration. Also, we set the tolerance parameter for the stopping criterion to 10^{-9}. `R3MC w/ Rank Upd` corresponds to the method described in the section on *rank updating* of [MS14].

3.7.2 Data-efficient recovery of ill-conditioned matrices

First, we run `MatrixIRLS` and the algorithms `R2RILS` , `RTRMC`, `LRGeomCG`, `LMaFit`, `ScaledASD`, `ScaledGD`, `NIHT` and `R3MC` to complete $\mathbf{X}_0$ from $P_\Omega(\mathbf{X}_0)$ where Ω corresponds to different oversampling factors ρ between 1 and 4, and where the condition number of $\mathbf{X}_0$ is $\kappa = \sigma_1(\mathbf{X}_0)/\sigma_r(\mathbf{X}_0) = 10$. In Figure 3.3, we report the median Frobenius errors $\|\mathbf{X}^{(K)} - \mathbf{X}_0\|_F/\|\mathbf{X}_0\|_F$ of the respective algorithmic outputs $\mathbf{X}^{(K)}$ across 100 independent realizations.

We see that `MatrixIRLS` and `R2RILS` are the only algorithms that are able to complete $\mathbf{X}_0$ already for $\rho = 1.5$. In our experiment, `R3MC` completes $\mathbf{X}_0$ in a majority of instances starting from $\rho = 2.0$, whereas the other algorithms, except `NIHT`, are able to reconstruct the matrix most of the times if ρ is at least between 2.4 and 3.0. This confirms the findings of [BNZ21], which show that even for quite well-conditioned matrices, fewer samples are required if second-order methods such as `R2RILS` or `MatrixIRLS` are used.

We repeat this experiment for ill-conditioned matrices $\mathbf{X}_0$ with $\kappa = 10^5$. In Figure 3.4, we see that current state-of-the-art methods are *not able* to achieve exact recovery of $\mathbf{X}_0$. This is, in particular, true as given the exponential decay of the singular values, to recover the subspace corresponding to the smallest singular value of $\mathbf{X}_0$, a relative Frobenius error of 10^{-5} or even several orders of magnitude smaller needs to be achieved. We observe

[14] `https://bamdevmishra.in/codes/r3mc/`. We used the version from Sep. 2020 which already includes the rank updating strategy.

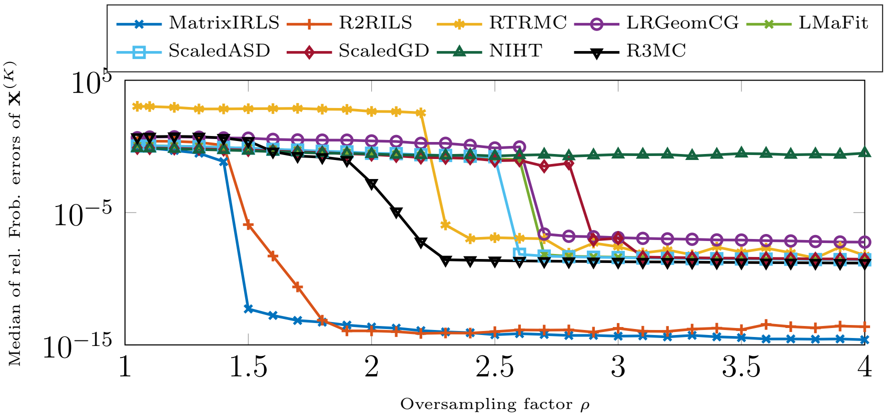


Figure 3.3: Performance of matrix completion algorithms for 1000 × 1000 matrices of rank $r = 5$ with condition number $\kappa = 10$, given $m = \lfloor \rho r(d_1 + d_2 - r) \rfloor$ random samples. Median of Frobenius errors $\|\mathbf{X}^{(K)} - \mathbf{X}_0\|_F / \|\mathbf{X}_0\|_F$ of 100 independent realizations.

that MatrixIRLS is the only method that is able to complete $\mathbf{X}_0$ for any of the considered oversampling factors.

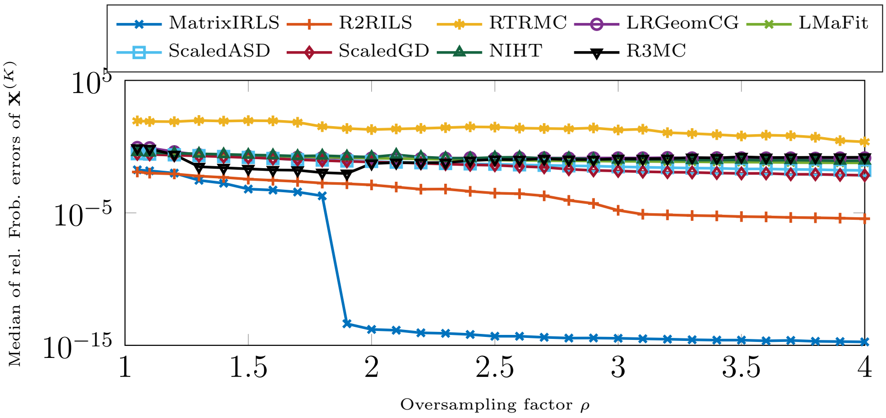


Figure 3.4: Performance of matrix completion algorithms as in Figure 3.3, but with $\kappa = 10^5$. Median of 50 realizations.

3.7.3 Running time for ill-conditioned problems

In Figure 3.5, for an oversampling ratio of $\rho = 4$, we illustrate the completion of one single extremely ill-conditioned 1000 × 1000 matrix with rank = 10 and $\kappa = 10^{10}$ and exponentially interpolated singular values as described above. We again can see that only second-order methods such as R2RILS or MatrixIRLS can achieve a relative Frobenius error $\approx 10^{-5}$ or smaller. MatrixIRLS goes beyond that and attains a relative Frobenius error of the order of the machine precision and, remarkably, *exactly recovers all the singular values up to 15 digits*. This also shows that the conjugated gradient and the randomized block Krylov method used at the inner core of our implementation can be extremely precise when properly adjusted. R2RILS can obtain relatively low Frobenius error, but unlike our method, it is not able to retrieve all the singular values with high accuracy. For the ill-conditioned matrix under consideration, other methods were found to result in a negligible reduction in relative error, rendering them ineffective for the task.

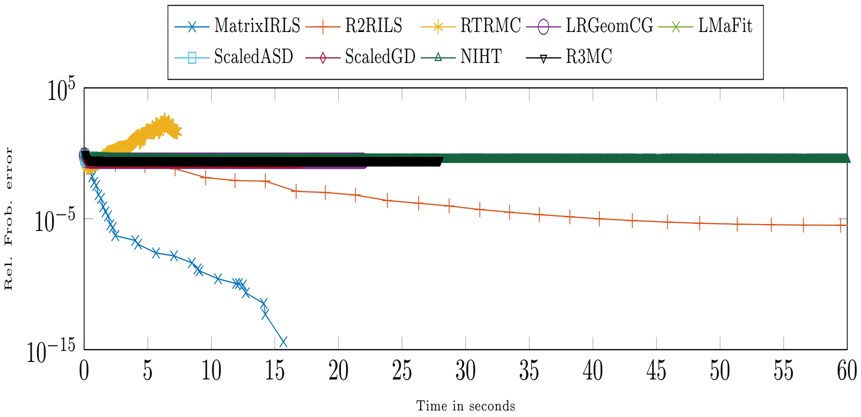


Figure 3.5: Completion task for a highly ill-conditioned 1000×1000 matrix of rank $r = 10$ with $\kappa = 10^{10}$ $(\rho = 4)$.

In Figure 3.6, we thoroughly investigate the execution time of R2RILS and MatrixIRLS for a range of ground truth matrices with increasing dimension for an oversampling ratio of $\rho = 2.5$, whose singular values are linearly interpolated between κ and 1. We observe that the larger the dimensions are, the larger the discrepancy in the running time of the two algorithms. Other algorithms are not considered in this experiment because they typically do not reach a relative error below 10^{-4} for $\kappa \gg 10^{2}$.

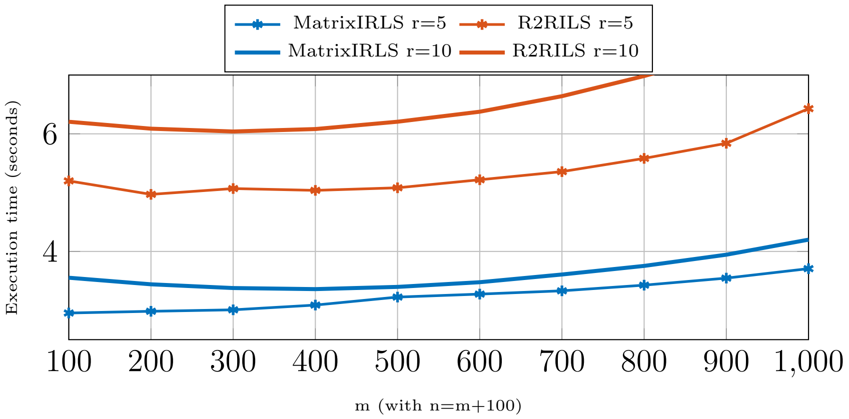


Figure 3.6: Execution time of R2RILS and MatrixIRLS for completion of rank $r \in \{5, 10\}$ matrices of size $m \times (m + 100)$ and condition number $\kappa = 10^{2}$, averaged across 50 independent realizations.

3.7.4 MatrixIRLS versus rank-adaptive strategies

In Section 3.7.2, all methods were provided with the correct rank r of the ground truth, which was used to determine the size of the matrix factors or the rank of the fixed rank manifold. Even in this case, we illustrated numerically that most of the methods are not able to recover highly ill-conditioned matrices. To handle such ill-conditioned completion problems, [MS14, UV15, TTW$^+$14] proposed very promising rank-adaptive variants of the methods `R3MC` and `LRGeomCG`. These variants, which we call `LRGeomCG Pursuit`[15] [UV15, TTW$^+$14] and `R3MC w/ Rank Update` [MS14], respectively, combine fixed-rank optimization with outer iterations that increase $\widetilde{r}$ from 1 to a target rank r, while warm starting each outer iteration with the output of the previous iteration. The main idea of these is that the rank-r manifold is not a closed set since a sequence of matrices in this set may converge to a matrix that has a strictly smaller rank and, therefore, it is not easy to develop a convergence theory for Riemannian method on this set. Moreover, in this manifold, it is usually the case that second-order approximation to the function to be minimized contains terms that are of the order of $1/\sigma_r(\mathbf{X})$, which means that these methods can perform really badly for retrieving ill-conditioned with very singular values, see [UV15] and [Van13, Section A.2]. Then, the problem is formulated on the (closed) set of matrices of rank *at most* r. To compare the data efficiency of `MatrixIRLS` with one of these three algorithms, we repeat the experiments of Section 3.7.2 for these methods and report the median Frobenius errors for the completion of 1000×1000 matrices of rank $r = 5$ with condition numbers $\kappa = 10$ and $\kappa = 10^5$, respectively, with those of `MatrixIRLS` in Figures 3.7 and 3.8.

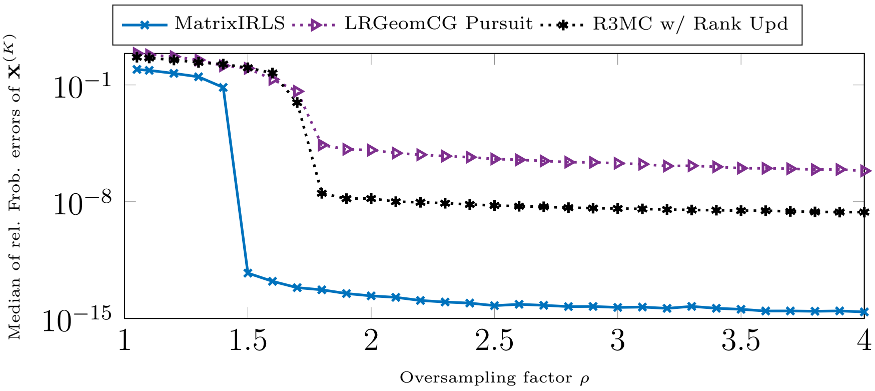


Figure 3.7: Completion of 1000×1000 matrices of rank $r = 5$ with condition number $\kappa = 10$, experiment as in Figure 3.3.

Figure 3.7 demonstrates that MatrixIRLS exhibits superior data efficiency compared to

[15] The MATLAB code containing the rank update was provided by B. Vandereycken in private communication.

the two rank-adaptive methods when considering a relatively small condition number of $\kappa = 10$. The phase transition for the rank-adaptive methods occurs at a larger oversampling factor,$\rho = 1.8$, in contrast to MatrixIRLS, where it happened at $\rho = 1.5$.

Contrarily, 3.8 highlights the effectiveness of the rank-adaptive techniques LRGeomCG Pursuit and R3MC w/ Rank Update when dealing with matrices possessing a large condition number like $\kappa = 10^5$. These methods exhibit a phase transition at approximately $\rho = 1.8$ and $\rho = 1.7$, respectively, whereas MatrixIRLS exhibits its phase transition at $\rho = 1.9$. This shows that for large condition numbers, rank adaptive strategies can outperform the data efficiency of MatrixIRLS, and in both experiments, the phase transitions are considerably better than for their fixed rank versions LRGeomCG and R3MC, cf. Figures 3.3 and Figure 3.4.

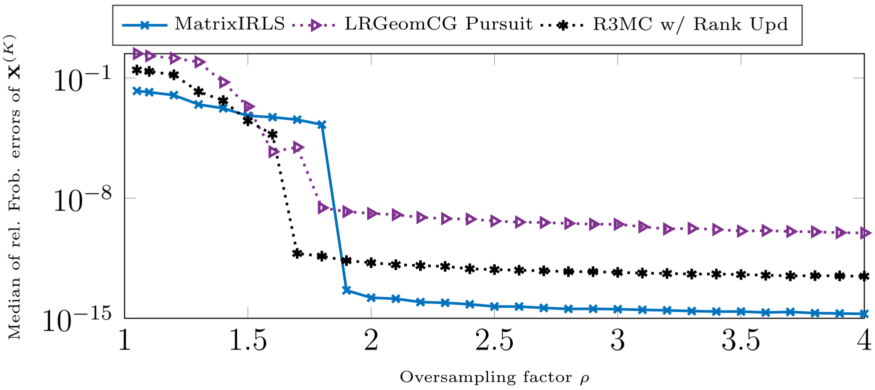


Figure 3.8: Completion of 1000×1000 matrices of rank $r = 5$ with condition number $\kappa = 10^5$, experiment as in Figure 3.4.

In all experiments so far, we have considered low-rank matrices with r singular values that exponentially decrease from κ to 1, as described at the beginning of this section. This might be a setting that is particularly suitable for rank-adaptive strategies that increase the rank parameter $\widetilde{r}$ one-by-one, as the singular subspaces are all one-dimensional and well-separated. For this reason, in a last experiment, we change this to a very challenging setup and consider ground truth matrices $\mathbf{X}_0$ that have a *plateau* in the set of singular values, potentially presenting a larger challenge for completion methods due to a higher dimensional subspace spanned by a set of multiple singular vectors. In particular, we consider the completion of a 1000×1000 matrix $\mathbf{X}_0$ with 10 singular values equal to $10^{10} \cdot \exp(-10 \cdot \log(10) \frac{14}{29})$, and with 10 singular values linearly interpolated on a logarithmic scale between this value and 10^{10} and, and another 10 between this value and 1. For a random instance of such a matrix, we report the relative Frobenius error vs. execution time for the methods MatrixIRLS against the rank-adaptive variants of LRGeomCG and R3MC, here denoted by LRGeomCG Pursuit and R3MC w/ Rank Update in Figure 3.9, from random samples with a small oversampling factor of $\rho = 1.5$.

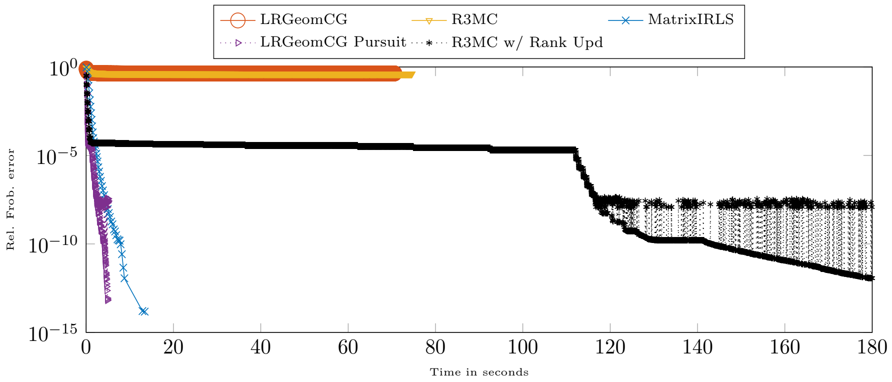


Figure 3.9: Comparison of matrix completion algorithms for 1000×1000 matrices of rank $r = 30$ with condition number $\kappa = 10^{10}$ and 10 equal singular values, oversampling factor of $\rho = 1.5$.

We observe that the fixed-rank variants `LRGeomCG` and `R3MC` are not able to complete the matrix, which is in line with the experiment of Section 3.7.3. `R3MC w/ Rank Update` exhibits a quick error decrease to a range around $6 \cdot 10^{-5}$, after which it just decreases very slowly for around 110 seconds before converging to $\mathbf{X}_0$ up to an error of around 10^{-12} within another 70 seconds. The stagnation phase presumably corresponds to the learning of the 10-dimensional singular space of $\mathbf{X}_0$ in the central part of its spectrum. `LRGeomCG Pursuit`, on the other hand, reaches an error of around 10^{-12} already after 5 seconds, albeit without monotonicity with a fluctuation phase between errors of 10^{-8} and 10^{-12} from seconds 3 to 5. For `MatrixIRLS`, we use a tolerance parameter for the relative residual in the conjugate gradient method of $\text{tol}_{\text{inner}} = 10^{-3}$ and a maximal number of 3 iterations for the randomized Block Krylov method (cf. 3.7.1 for the default parameters) and observe that the method successfully converges to $\mathbf{X}_0$ slightly slower with a convergence within 13 seconds, but, remarkably, unlike `LRGeomCG Pursuit`, with a monotonous error decrease.[16]

Nevertheless, relying solely on tracking the relative Frobenius error to assess the performance of methods in recovering highly ill-conditioned matrices without taking into account the condition number κ, may not provide a comprehensive understanding. It is important to note that the recovery of singular spaces associated with the smallest singular values is typically achieved only when the relative error becomes smaller than $1/\kappa$.

To address this, we present the singular values of the recovered matrices $\mathbf{X}^{(K)}$ in Fig-

[16] For the default choice of algorithmic parameters as described in Section 3.7.1, we obtain a qualitatively similar behavior for `MatrixIRLS`, but with a small runtime multiple due to the higher required precision at each iteration.

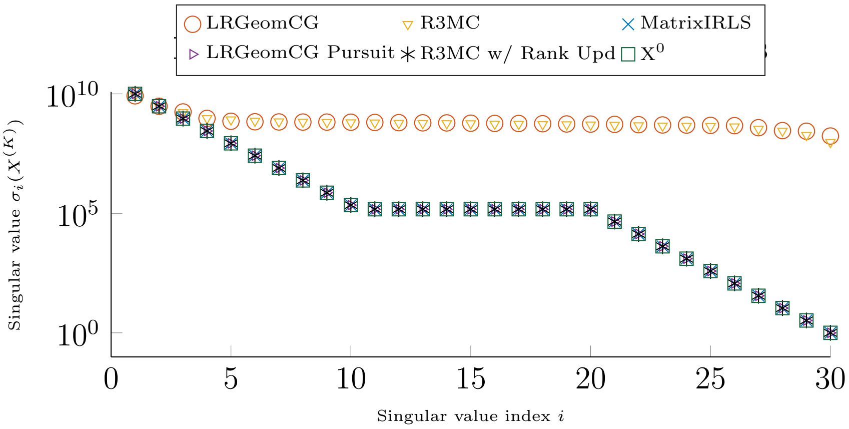


Figure 3.10: Spectrum of output matrices $\mathbf{X}^{(K)}$ and $\mathbf{X}^0$

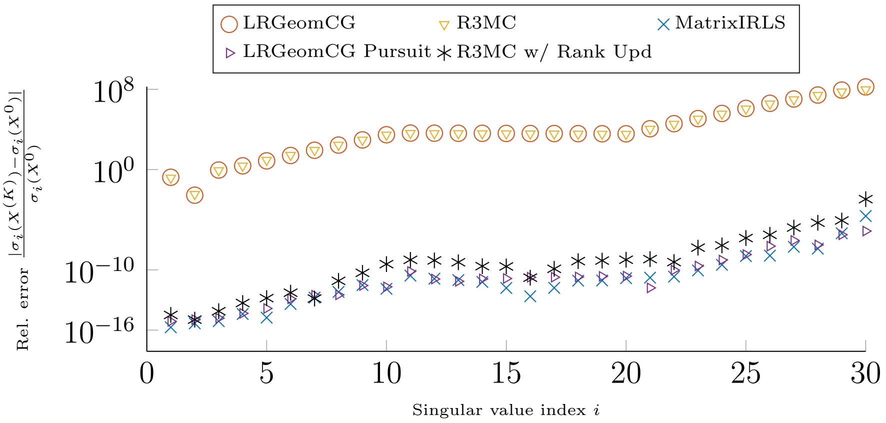


Figure 3.11: Relative errors $\frac{|\sigma_i(\mathbf{X}^{(K)})-\sigma_i(\mathbf{X}^0)|}{\sigma_i(\mathbf{X}^0)}$

Figure 3.12: Spectrum information of algorithmic output $\mathbf{X}^{(K)}$ after convergence, experiment of Figure 3.9 (1000×1000 matrix, $r = 30$, $\kappa = 10^{10}$, $\rho = 1.5$)

ure 3.10, along with the corresponding relative error for each individual singular value in Figure 3.11. These results pertain to the experiment conducted in Section 3.7.4, as illustrated in Figure 3.9. Notably, `MatrixIRLS`, `LRGeomCG Pursuit` and `R3MC w/ Rank Upd` exhibit remarkable capability in accurately recovering even the smallest singular values, specifically those with indices $i = 28, 29, 30$. The relative error for these singular values ranges from 10^{-7} to 10^{-3}, indicating a high level of precision.

This shows that despite a not too restrictive choice of the tolerance on of the inner conjugate gradient iterations (such as $\text{tol}_{\text{inner}} = 10^{-3}$), `MatrixIRLS` is successful in recovering the complete spectrum of $\mathbf{X}_0$, indicating that an implementation of `MatrixIRLS` that solves Equation (2.17) via conjugate gradient method together with weight updates based on a randomized block Krylov method can be very precise even without requiring a very high precision on the iterative solver.

These observations suggest that `MatrixIRLS` and Riemannian optimization methods with adaptive rank updates such as `LRGeomCG Pursuit` and `R3MC w/ Rank Upd` are good alternatives to solve hard matrix recovery problems in a numerically efficient way, warranting further investigations for a better theoretical understanding. In particular, looking into the possible connections between rank-adaptive Riemannian methods and `MatrixIRLS` would be interesting.

3.8 Chapter Conclusion

In this chapter, we presented the second idea related to the methods of least squares method of this thesis, and we discussed how to develop a second-order method, here called `MatrixIRLS`, that is able to efficiently complete large, extremely ill-conditioned matrices from few samples, a problem for which most state-of-the-art methods fail. We designed weighted least squares problems that can tackle highly non-convex objective functions such as the log-det by applying a suitable smoothing strategy combined with saddle-escaping Newton-type steps.

We established a local quadratic convergence rate for our algorithm under very general coherence assumptions and we have corroborated our theory with numerical experiments that show how efficient the method is for retrieving very ill-conditioned low-rank matrices. While some Riemannian optimization methods are guaranteed to have an asymptotic quadratic rate of convergence [MMBS13, BA15], their theory does not specify the number of samples needed to perform the completion. Also, it is not clear what the basin of attraction for such methods is. See also [ABG07] for a discussion about the convergence rate.

In addition, we examined an efficient implementation focused on the matrix completion problem. It is important to note that our analysis can be extended to situations where the measurements are corrupted by noise. In such cases, the constraint $P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{X}_0)$ can be integrated into the objective function as $\lambda\|P_\Omega(\mathbf{X}) - P_\Omega(\mathbf{X}_0)\|_2^2$, where λ is a parameter that balances the data-fidelity and the non-convex proxy function that enforces the low-rank model. Since the data-fidelity term in this scenario is represented by a quadratic function, incorporating it into the reweighted least squares analysis can be done easily.

Another very relevant issue relates to the more realistic scenario of online optimization. In recommender systems, for instance, the algorithms can operate in an online and interactive fashion, continuously adapting based on user feedback. While the theory for online matrix completion is not as well-established as its offline counterpart, cf. [PJ23], recent advancements have been made in completing ill-conditioned matrices within an online framework [ZCZ22]. We believe that an extension of `MatrixIRLS` to this case is extremely relevant.

Open Problem: How to develop a reweighted least squares algorithm for online matrix completion? Is it possible to use techniques developed for modified least squares problems, e.g., [Bjö96, Chapter 3], for this task?

Another interesting issue is related to the local convergence rate. Usually, is it possible to establish a two-phase convergence regime for Newton's method, the so-called damped Newton phase and pure Newton phase [BV04, Chapter 9.5]. Under certain assumptions, it is possible to prove that this type of method achieves a linear convergence in the beginning, and after a finite number of steps, it reaches the basin of attraction for which an accelerated superlinear convergence is guaranteed. Given the resemblance of our algorithm with a non-convex Newton's method [PMR19b], it would be interesting to establish results of this nature to `MatrixIRLS`.

Open Problem: How to rigorously establish a two-phase global convergence guarantees for `MatrixIRLS`?

Another interesting line of research deals with tensors and tensor completion methods: a higher-order generalization of the matrix completion ones. However, everything is more involved in this scenario. Even the singular value decomposition [DOT18, VNVM14]. In this case, developing an algorithm that is fast, scalable, and with optimal memory storage is even more crucial. Extending the theory presented in this chapter to address the tensor completion problem would be fascinating.

Chapter 4

IRLS for Sparse Noise-Blind Optimization

> "Finding patterns is easy in any kind of data-rich environment; that's what mediocre gamblers do. The key is in determining whether the patterns represent noise or signal." [Sil12]
>
> *The Signal and the Noise: Why So Many Predictions Fail-but Some Don't, Nate Silver*

In this chapter, we study the noise-blind sparse recovery problem, a problem that is also connected to robust optimization. We then propose an efficient algorithm based on the method of least squares to minimize the so-called square-root LASSO, an optimization problem for which the optimal solution provably does not depend on the noise added to the measurements. The work presented in this chapter, still in progress, is in collaboration with Dr. Oleh Melnyk, Dr. Peter Jung and Prof. Felix Krahmer and the final version of the manuscript is currently in preparation.

We establish a minimization scheme for the square-root LASSO, a convex but non-smooth problem that appears in high-dimensional statistics and robust optimization for which the optimal regularization parameter does not depend on the noise level. By extending the techniques for the first chapter of this thesis, we show that the method attains a *global convergence rate*.

4.1 Introduction

In Chapter 2, we discussed and analyzed a constrained optimization program for sparse recovery, the so-called *Basis Pursuit* Equation (P_1). However, in several applications, the data is corrupted with noise, i.e., the measurements are subject to errors. Given the presence of noise in the measurements, typically modeled via

$$y = Ax + \epsilon, \text{ where } \epsilon \text{ is a random variable,} \tag{4.1}$$

it is crucial to adopt a formulation that incorporates and withstands its impact. Taking this into consideration, the Quadratically Constrained Basis Pursuit (QCBP), introduced in [CD94b], substitutes the equality constraint $Ax = y$ with an inequality constraint $\frac{1}{m}\|Ax - y\|_2^2 \leq \eta$. Consequently, the objective now shifts towards solving

$$\min_{z\in\mathbb{R}^N} \|z\|_1 \qquad \text{subject to} \qquad \frac{1}{m}\|Az - y\|_2^2 \leq \eta, \tag{4.2}$$

Remark 4.1.1. *An alternative approach to address this challenge involves solving the* Equality Constrained Basis Pursuit *program blindly, without any prior knowledge or estimation of the noise level. The aim is to identify conditions under which this approach guarantees successful recovery, meaning that it provides an accurate solution up to the noise level. The seminal work by [Woj10] (see also [Fou14, DPW09]) introduced the quotient property, which is discussed in Chapter 5. The authors demonstrated that if the measurement matrix satisfies this property, it is possible to achieve noise-blind recovery guarantees for the* Equality Constrained Basis Pursuit*. However, while this property can be established for broad matrices ensembles, the supporting arguments rely on concentration inequalities and random polytopes. To the best of the author's knowledge, there is currently no theory to establish this property for structured matrices commonly used in applications. Therefore, this chapter will focus on methods widely used in signal processing, statistics, and machine learning.*

At around the same time, in the statistics literature, the work [Tib96a] proposed the LASSO[1], which stands for *Least Absolute Shrinkage and Selection Operator*, which consists in solving, for some parameter $\tau \geq 0$,

$$\min_{z\in\mathbb{R}^N} \frac{1}{m}\|Az - y\|_2^2 \qquad \text{subject to} \qquad \|z\|_1 \leq \tau. \tag{4.3}$$

Since both problems are constrained ones, the development of techniques that are designed to solve them is harder than the correspondent techniques for the equivalent unconstrained problem, which was originally called *Basis Pursuit Denoising* (also called Lagrangian LASSO, see [Wai19, Chapter 7]). This problem consists in solving, for some parameter $\lambda \geq 0$,

$$\min_{z\in\mathbb{R}^N} \lambda\|z\|_1 + \frac{1}{m}\|Az - y\|_2^2. \tag{4.4}$$

The following theorem demonstrates the interrelation among these three problems, highlighting their connection. For a comprehensive analysis encompassing all three problems, refer to [HTW15].

[1] This, in turn, was inspired by the so-called empirical atomic decomposition [CDS01b].

Theorem 4.1.2. *[FR13, Proposition 3.2] There exists an equivalence between the LASSO, the Basis Pursuit Denoising and the Quadratically Constrained Basis Pursuit as the following three statements show.*

i. If x is a minimizer of the Basis Pursuit Denoising with $\lambda > 0$, then there exists $\eta = \eta(x)$ such that x is a minimizer of the Quadratically Constrained Basis Pursuit.

ii. If x is a unique minimizer of the Quadratically Constrained Basis Pursuit with $\eta \geq 0$, then there exists $\tau = \tau(x) \geq 0$ such that x is a unique minimizer of the LASSO.

iii. If x is a unique minimizer of the LASSO with $\tau > 0$, then there exists $\lambda = \lambda(x) \geq 0$ such that x is a minimizer of the Basis Pursuit Denoising.

The aforementioned theoretical equivalence implies that the unconstrained problem represented by Equation (4.4) could be a favorable option for designing optimization algorithms. In this scenario, there is no need to consider if the feasible points lie within the constraint set. The Lagrangian LASSO problem described by Equation (4.4) has recently been referred to as the LASSO, a terminology that we will maintain throughout this thesis. Due to this equivalence, this penalized estimator also achieves, under certain assumptions, variable selection by promoting sparsity among the estimated regression coefficients.

From the computational point of view, several efficient and scalable algorithms were proposed to minimize this function. See, e.g. [LST18, WYL^{+}22] and references therein. From the theoretical point of view, this estimator attains optimal minimax rates for the prediction error [BRT09, Section 6] and in the ℓ_∞-norm [Lou08, BZ22]. Results for the size of the support of the LASSO minimizer were also derived in [FTZ22]. However, such optimal results depend on a regularization choice that relies on oracle knowledge about the noise variance, which is usually unavailable in many applications [Gir15, Chapter 5]. If no prior information about the noise is available, the sub-optimally tuned LASSO yields suboptimal recovery guarantees. Moreover, the estimation of the error variance for LASSO-type problems is a non-trivial problem that still attracts significant interest [RTF16, YB19, GHV12].

Several (almost optimal) approaches have been developed to address the selection of λ in a provable manner without assuming any prior knowledge about the noise distribution. These include techniques such as cross-validation [CLC21] and adaptive calibration schemes [CLW16]. However, they lack scalability or, in the second case, are just tailored to ℓ_∞-norm results, and it is not clear how to establish other minimax bounds for such approaches. Besides that, the LASSO estimator lacks some important properties such as scale invariance, see e.g., [Gir15, Section 5.1], or asymptotic normality, see e.g., [JM18, Section 1] and references therein for a discussion. Both properties are fundamental, for example, to characterize the uncertainty associated with a parameter through the computation of confidence intervals and p-values.

To address the aforementioned challenges, the groundbreaking paper [BCW11] introduced the *square-root LASSO* (sqrt-LASSO)[2]. In the authors' own words, this novel estimator effectively *"handles the unknown scale, heteroscedasticity, and (drastic) non-Gaussianity of the noise"*. Notably, the key characteristic of this estimator is that the tuning parameter λ, which ensures minimax oracle inequalities, remains independent of the noise level [vdG16]. Mathematically, it is described as follows:

$$\hat{x}_\lambda \in \operatorname*{arg\,min}_{x\in\mathbb{R}^N} \frac{1}{\sqrt{m}}\|Ax-y\|_2 + \lambda\|x\|_1 =: \operatorname*{arg\,min}_{x\in\mathbb{R}^N} f_0(x). \tag{4.5}$$

Subsequent advancements in the field yielded sharp oracle inequalities for the sqrt-LASSO, building upon the findings of [BCW11], as demonstrated in some works such as [Der18, SvdG17]. Estimates for the support size of the obtained solution [Fou23] and recovery guarantees against adversarial noise [PJ21] have also been established. In particular, the work [Fou23] shows that if the measurement matrix A satisfies the RIP, then the output of any LASSO-type solution, including the sqrt-LASSO, is Cs-sparse for a certain constant C.

Theorem 4.1.3. *[Fou23, Theorem 5] Let $p \in [1,2]$, $q \geq 1$, and $r \geq 1$. Consider a vector $x \in \mathbb{R}^N$ such that $B^{-1}x$ is s-sparse for some invertible matrix $B \in \mathbb{R}^{N\times N}$ with condition number $\kappa_B := \|B\|_{2\to2}\|B^{-1}\|_{2\to2}$. Suppose that the vector x is measured via $y = Ax + \epsilon \in \mathbb{R}^m$ for some error vector $e \in \mathbb{R}^m$ with $\|\epsilon\|_p \leq (1/3)\|y\|_p$ and some matrix $A \in \mathbb{R}^{m\times N}$ satisfying the nonstantard restricted isometry property of order $t := \lfloor(6\gamma\kappa_B)^2 s\rfloor + 1$ with ratio $\gamma = \beta/\alpha$, i.e.,*

$$\alpha\|z\|_2 \leq \|Az\|_p \leq \beta\|z\|_2 \qquad \textit{whenever } \|B^{-1}z\|_0 \leq t.$$

Then, for any $\lambda \geq \lambda^* := 2^{q-1}\beta^r\|B\|_{2\to2}^r\|\epsilon\|_p^{q-r}$, the solution x^λ of the LASSO-type procedure

$$x_\lambda \in \operatorname*{arg\,min}_{z\in\mathbb{R}^N} \frac{1}{q}\|y-Az\|_p^q + \lambda\frac{1}{r}\|B^{-1}z\|_1^r$$

has sparsity at most proportional to s, namely

$$\|B^{-1}x^\lambda\|_0 \leq \lfloor\chi^2 s\rfloor, \qquad \chi := 6\gamma\kappa_B.$$

Applying the theorem above to the case where $q = r = 1$ and $B = \mathbf{I}$, the threshold λ^*

[2] It is also called ℓ_2-lasso in the signal processing literature [OTH13].

reduces to a quantity that is independent of the magnitude of noise, and it yields the corresponding result concerning the sparsity of the sqrt-LASSO solution. In particular, for $\lambda \geq \beta$, it holds that $\|x_\lambda\|_0 \leq \lfloor 36\frac{\beta}{\alpha}s \rfloor$.

The most celebrated result for the sqrt-LASSO estimator is in the form of sharp recovery guarantees. Namely, it is that it is possible to establish that for a certain choice of regularization parameter λ that is *independent* from the noise level, the sqrt-LASSO attains optimal error. For example, by assuming that ϵ in Equation (4.1) is a random noise, with values in $\mathbb{R}^m$, distributed as $N(0, \sigma^2\mathbf{I})$, it was proven in [BCW11, Theorem 1] and [Der18, Theorem 3.1] that, for a certain λ that does not depend on σ, the solution of the sqrt-LASSO x^λ attains an error of the order of $s/m \log(N/s)$, that is known to be sharp, see [BLT18] and [RWY11]. Later, this was generalized to an adversarial error, more commonly discussed in the signal processing literature, by assuming that the measurement matrix satisfies the *robust null space property*, a generalization of the null space property discussed in Chapter 2.

Definition 4.1.4 ([FR13, Definition 4.17]). *A matrix $A \in \mathbb{R}^{m \times N}$ is said to satisfy the* robust null space property (NSP) *of order $s \in [N]$ with constants $0 < \rho < 1$ and $\tau > 0$ if for any set $S \subset [N]$ of cardinality $|S| \leq s$, it holds that*

$$\|v_S\|_1 \leq \rho\|v_{S^c}\|_1 + \tau\|Av\|_2, \text{ for all } v \in \mathbb{R}^N. \tag{4.6}$$

Theorem 4.1.5. *[PJ21, Theorem 3.1] Let $A \in \mathbb{R}^{m \times N}$ have ℓ_q-RNSP of order S wrt $\|.\|$ with constants ρ and τ. Let $\lambda \geq \frac{2}{1+\rho}\tau$. Then for all $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^m$, any minimizer $\mathbf{x}^\lambda$ of*

$$\min_{z \in \mathbb{R}^N} \|Az - y\|_2 + \lambda\|z\|_1$$

obeys

$$\|x^\lambda - x\|_1 \leq 2\frac{1+\rho}{1-\rho}\sigma_s(x)_1 + \left(\frac{2}{1-\rho}\tau + \frac{1+\rho}{1-\rho}\lambda\right)\|y - Ax\|_2. \tag{4.7}$$

Also, if $\lambda \geq \frac{3+\rho}{(1+\rho)^2}\tau\sqrt{S}$, any minimizer x^λ of the sqrt-LASSO

$$\|x^\lambda - x\|_2 \leq 2\frac{(1+\rho)^2}{1-\rho}\sqrt{S}\sigma_s(x)_1 + \left(\frac{3+\rho}{1-\rho}\tau + \frac{(1+\rho)^2}{1-\rho}\sqrt{S}\lambda\right)\|y - Ax\|_2. \tag{4.8}$$

Another remarkable property of the sqrt-LASSO is that it is equivalent to a robust regression problem, usually formulated as the following min-max problem,

$$\min_{\mathbf{x} \in \mathbb{R}^m} \left\{ \max_{\Delta A \in \mathcal{U}} \|b - (A + \Delta A)x\|_2 \right\}, \tag{4.9}$$

where $\mathcal{U}$ denotes the so-called uncertainty set representing all possible perturbations of the measurement matrix A. One typical class of uncertainty sets is given by

$$\mathcal{U} =: \left\{(\delta_1, \cdots, \delta_m)\Big| \|\delta_i\|_2 \leq c_i, \;\; i = 1, \cdots, m\right\}, \tag{4.10}$$

for given $c_i \geq 0$.

Theorem 4.1.6. *[XCM10, Theorem 1] The robust regression problem (4.9) with uncertainty set of the form (4.10) is equivalent to the following sqrt-LASSO regression problem:*

$$\min_{x\in\mathbb{R}^m} \left\{\|b - Ax\|_2 + \sum_{i=1}^{m} c_i|x_i|\right\}. \tag{4.11}$$

Although the sqrt-LASSO exhibits comparable theoretical characteristics to the original LASSO without requiring tuning efforts and many interesting theoretical properties, as illustrated above, its computational aspect poses greater challenges due to the *non-differentiability* of the data fidelity term. The original paper demonstrated that the sqrt-LASSO can be reformulated as a second-order conic program [BCW11, Section 4], allowing for the utilization of second-order conic solvers like MOSEK [ApS22]. However, these solvers do not make use of the intrinsic problem structure and are not scalable [SCGS16], necessitating the development of specialized solvers tailored to noise-blind sparse recovery problems in high-dimensional settings. While several algorithms have been proposed - see Section 4.3 - essentially all lack rigorous analysis or scalability. Despite significant progress in understanding the theoretical aspects of this estimator and recognizing its vast potential in applications where noise estimation is a challenging task, such as magnetic resonance imaging [VS16], there still remains a critical question regarding the design of an efficient computational method for minimizing this function. Therefore, the main question to be solved in this chapter is

> What approaches can be utilized to devise a scalable algorithm for the sqrt-LASSO, ensuring provable global convergence with a linear rate?

Contribution of this chapter:

The goal of this chapter is to affirmatively answer this question by leveraging the techniques from Chapter 2 and by establishing an Iteratively Reweighted Least Squares method for the sqrt-LASSO problem that has a global convergence rate. Unlike the previous chapters, we discuss a general convergence theory of the algorithm even in scenarios where structural assumptions for sparse recovery, such as the null space property of the incoherence property, are not made.

4.2 Minimizing the sqrt-LASSO via IRLS

The objective function in (4.5) is convex, but both of its terms are non-smooth, and an efficient way to minimize it is to design a majorization-minimization strategy, i.e., to introduce a smoothed objective $f_\varepsilon(x)$ that mitigates the non-smoothness of the $\|\cdot\|_1$-norm as well as the non-smoothness of the $\|\cdot\|_2$-norm but that, at the same time, majorizes the sqrt-LASSP objective function to be minimized. After that, we establish quadratic upper bounds for $f_\varepsilon(x)$ that can be optimized efficiently by employing least square ideas.
Note that the objective function (4.5) is non-smooth at the points where either the data fidelity term or the ℓ_1-norm vanishes. Thus, smoothing the objective around these points will be sufficient. By following the steps from Chapter 2 we use a scaled Huber loss function, Equation (2.4), on both terms of our objective function. Our new objective function will be then given by

$$f_\varepsilon(x) = j_{\varepsilon_0}(\|Ax-b\|_2) + \sum_{i=1}^{N} j_\varepsilon(x_i), \tag{4.12}$$

with some smoothing parameters $\varepsilon_0, \varepsilon > 0$ that consider how far the proxy is from the original function. Clearly, f_ε and f_0 coincide if $\varepsilon_0 = \varepsilon = 0$. Furthermore, the distance in the function values between smooth and non-smooth objectives can be quantified.

Lemma 4.2.1. *Let $\varepsilon_0, \varepsilon > 0$. Then, we have*

$$f_0(x) \le f_\varepsilon(x) \le f_0(x) + \varepsilon_0 + \lambda N\varepsilon \quad \textit{and} \quad f_{\varepsilon_1}(x) \le f_{\varepsilon_2}(x), \quad 0 \le \varepsilon_1 \le \varepsilon_2.$$

To tackle the objective function above, we perform a change of variables

$$Ax - b = \begin{bmatrix} -b & A \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \tilde{A}\tilde{x}.$$

The embedding $\tilde{v} := (1, v^T)^T$ for vectors $v \in \mathbb{R}^N$ will be used throughout this chapter with the added dimension indexed by zero, i.e., $\tilde{v}_0 = 1$. With this change of variables, our objective function becomes

$$f_\varepsilon(x) = j_\varepsilon(\|Ax-b\|_2) + \sum_{i=1}^{N} j_\varepsilon(x_i) = j_\varepsilon(\|\tilde{A}\tilde{x}\|_2) + \sum_{i=1}^{N} j_\varepsilon(x_i) =: \tilde{f}_\varepsilon(\tilde{x})$$

and the unconstrained optimization of the proposed smoothed objective becomes

$$\min f_\varepsilon(x) = \min_{\tilde{x}\in\mathbb{R}^{N+1},\ \tilde{x}_0=1} \tilde{f}_\varepsilon(\tilde{x}). \tag{4.13}$$

The main idea of the IRLS strategy developed in this paper is to solve Equation (4.13) by defining a sequence of iterates x^k, $k \geq 0$, which performs the following steps. As previously discussed, a quadratic majorant of the form

$$\begin{aligned} Q_\varepsilon(\tilde{z},\tilde{x}) :&= \tilde{f}_\varepsilon(\tilde{x}) + \langle \nabla \tilde{f}_\varepsilon(\tilde{x}), \tilde{z} - \tilde{x} \rangle + \frac{1}{2}\langle \tilde{z} - \tilde{x}, W_\varepsilon(\tilde{x})(\tilde{z} - \tilde{x})\rangle \\ &= \tilde{f}_\varepsilon(\tilde{x}) + \frac{1}{2}\langle \tilde{z}, W_\varepsilon(\tilde{x})\tilde{z}\rangle - \frac{1}{2}\langle \tilde{x}, W_\varepsilon(\tilde{x})\tilde{x}\rangle, \end{aligned} \tag{4.14}$$

is constructed with the symmetric and positive semidefinite matrix W_ε given by

$$W_\varepsilon(x) = \frac{\tilde{A}^T\tilde{A}}{\max\{\|\tilde{A}\tilde{x}\|_2, \varepsilon_0\}} + \lambda \begin{bmatrix} 0 & 0 \\ 0 & \operatorname{diag}(\{\max^{-1}\{|x_i|, \varepsilon\}\}_{i=1,\dots,N}) \end{bmatrix}. \tag{4.15}$$

This choice of $W_\varepsilon(x)$ ensures that $Q_\varepsilon(\tilde{z},\tilde{x})$ is a majorizer, as the following lemma shows. Moreover, as explained in Section 2.2, the weight matrix is constructed in a way that captures the information given by the first-order derivative. By doing so, we are able to derive a "pure" quadratic problem, i.e., without first-order terms, to be minimized at each iteration.

Lemma 4.2.2. *The function Equation* (4.14) *with* W_ε *as in* (4.15) *admits*

i. $W_\varepsilon(x)\tilde{x} = \nabla \tilde{f}_\varepsilon(\tilde{x})$, *ii.* $Q_\varepsilon(\tilde{x},\tilde{x}) = \tilde{f}_\varepsilon(\tilde{x})$, *iii.* $Q_\varepsilon(\tilde{z},\tilde{x}) \geq \tilde{f}_\varepsilon(\tilde{z})$.

Proof. The first derivative of the smoothed objective function is given by

$$\nabla \tilde{f}_\varepsilon(\tilde{x}) = \frac{\tilde{A}^T\tilde{A}\tilde{x}}{\max\{\|\tilde{A}\tilde{x}\|_2, \varepsilon_0\}} + \lambda \sum_{j=1}^{N} \frac{x_j e_j}{\max\{|x_j|, \varepsilon\}}, \tag{4.16}$$

with $\{e_i\}_{i=0,\dots,N}$ being the standard basis vectors in $\mathbb{R}^{1+N}$. In view of the first condition, it is natural to define $W_\varepsilon(x)$ as in (4.15).

The second condition follows directly from Equation (4.14). To show that $Q_\varepsilon(\tilde{z},\tilde{x})$ majorizes $f_\varepsilon(\tilde{z})$ for all $z \in \mathbb{R}^N$, we first rewrite Equation (4.14) as

$$\begin{aligned} Q_\varepsilon(\tilde{z},\tilde{x}) &:= \tilde{f}_\varepsilon(\tilde{x}) + \langle \nabla \tilde{f}_\varepsilon(\tilde{x}), \tilde{z}\rangle - \langle \nabla \tilde{f}_\varepsilon(\tilde{x}), \tilde{x}\rangle + \frac{1}{2}\langle \tilde{z}, W_\varepsilon(\tilde{x})\tilde{z}\rangle + \frac{1}{2}\langle \tilde{x}, W_\varepsilon(\tilde{x})\tilde{x}\rangle - \langle \tilde{z}, W_\varepsilon(\tilde{x})\tilde{x}\rangle \\ &= \tilde{f}_\varepsilon(\tilde{x}) + \frac{1}{2}\langle \tilde{z}, W_\varepsilon(\tilde{x})\tilde{z}\rangle - \frac{1}{2}\langle \nabla \tilde{f}_\varepsilon(\tilde{x}), \tilde{x}\rangle, \end{aligned} \tag{4.17}$$

where we used the second condition. Thus, to establish the majorization property iii., we need to prove the inequality

$$0 \leq Q_\varepsilon(\tilde{z},\tilde{x}) - \tilde{f}_\varepsilon(\tilde{z}) = \tilde{f}_\varepsilon(\tilde{x}) - \tilde{f}_\varepsilon(\tilde{z}) - \frac{1}{2}\langle \nabla \tilde{f}_\varepsilon(\tilde{x}), \tilde{x}\rangle + \frac{1}{2}\langle \tilde{z}, W_\varepsilon(\tilde{x})\tilde{z}\rangle$$

$$= j_\varepsilon(\|\tilde{A}\tilde{x}\|_2) - j_\varepsilon(\|\tilde{A}\tilde{z}\|_2) - \frac{\|\tilde{A}\tilde{x}\|_2^2 - \|\tilde{A}\tilde{z}\|_2^2}{2\max\{\|\tilde{A}\tilde{x}\|_2, \varepsilon_0\}} + \lambda \sum_{i=1}^{N}\left[j_\varepsilon(x_i) - j_\varepsilon(z_i) - \frac{|x_i|^2 - |z_i|^2}{2\max\{|x_i|,\varepsilon\}}\right].$$

As all summands have a similar structure, we can prove that for $M \in \mathbb{R}^{p\times q}$, $\gamma > 0$ and for all $v, u \in \mathbb{R}^q$, it holds that

$$j_\gamma(\|Mv\|_2) - j_\gamma(\|Mu\|_2) - \frac{\|Mv\|_2^2 - \|Mu\|_2^2}{2\max\{\|Mv\|_2, \gamma\}} \geq 0. \tag{4.18}$$

Then, note that the theorem will be established by applying Equation (4.18), first with $M = \tilde{A}$ and $\gamma = \varepsilon_0$ and finally with $M = E^{i,i}$, $i = 1, \dots, N$ and $\gamma = \varepsilon$, where $E^{i,i}$ is a matrix with a single non-zero element $E^{i,i}_{i,i} = 1$.

To prove Equation (4.18), we start by noting that the left-hand side may take four different values depending on $\|Mv\|_2$, $\|Mu\|_2$ and γ. Let us consider each of them separately.
Case 1: $\|Mv\|_2 < \gamma$, $\|Mu\|_2 < \gamma$. Then, the left-hand side is given by

$$\frac{\|Mv\|_2^2}{2\gamma} + \frac{\gamma}{2} - \frac{\|Mu\|_2^2}{2\gamma} - \frac{\gamma}{2} - \frac{\|Mv\|_2^2 - \|Mu\|_2^2}{2\gamma} = 0.$$

Case 2: $\|Mv\|_2 < \gamma$, $\|Mu\|_2 \geq \gamma$. Using the arithmetic-geometric mean inequality, we get

$$\frac{\|Mv\|_2^2}{2\gamma} + \frac{\gamma}{2} - \|Mu\|_2 - \frac{\|Mv\|_2^2 - \|Mu\|_2^2}{2\gamma} = \frac{\|Mu\|_2^2}{2\gamma} + \frac{\gamma}{2} - \|Mu\|_2 \geq 0.$$

Case 3: $\|Mv\|_2 \geq \gamma$, $\|Mu\|_2 \geq \gamma$. Likewise, as in the previous case,

$$\|Mv\|_2 - \|Mu\|_2 - \frac{\|Mv\|_2^2 - \|Mu\|_2^2}{2\|Mv\|_2} = \frac{\|Mv\|_2}{2} + \frac{\|Mu\|_2^2}{2\|Mv\|_2} - \|Mu\|_2 \geq 0.$$

Case 4: $\|Mv\|_2 \geq \gamma$, $\|Mu\|_2 < \gamma$. We have

$$\|Mv\|_2 - \frac{\|Mu\|_2^2}{2\gamma} - \frac{\gamma}{2} - \frac{\|Mv\|_2^2 - \|Mu\|_2^2}{2\|Mv\|_2} = \frac{1}{2}\left[\|Mv\|_2 - \gamma + \|Mu\|_2^2\left[\frac{1}{\|Mv\|_2} - \frac{1}{\gamma}\right]\right].$$

Since $\|Mv\|_2 \geq \gamma$, the second term is negative, and we can further decrease it by applying

$\|Mu\|_2 < \gamma$, which gives

$$\begin{aligned}\|Mv\|_2 - \frac{\|Mu\|_2^2}{2\gamma} - \frac{\gamma}{2} - \frac{\|Mv\|_2^2 - \|Mu\|_2^2}{2\|Mv\|_2} &\geq \frac{1}{2}\left[\|Mv\|_2 - \gamma + \gamma^2\left[\frac{1}{\|Mv\|_2} - \frac{1}{\gamma}\right]\right] \\ &= \frac{1}{2}\left[\|Mv\|_2 + \frac{\gamma}{\|Mv\|_2} - 2\gamma\right] \geq 0,\end{aligned}$$

where the last inequality is the arithmetic-geometric mean inequality again. □

The change of variables introduced above is motivated by a more natural construction of the weight matrix. Note that if instead of $\tilde{f}_\varepsilon$, we try to work with f_ε directly, its derivative

$$\nabla f_\varepsilon(x) = \frac{A^T(Ax-b)}{\max\{\|Ax-b\|_2, \varepsilon_0\}} + \lambda\sum_{j=1}^N \frac{x_j e_j}{\max\{|x_j|, \varepsilon\}}$$

would be affine in x and not linear, making the construction of W_ε more involved. With this transformation in mind, given a previous iterate x^k of the algorithm, we obtain x^{k+1} by solving

$$\tilde{x}^{k+1} := \underset{\tilde{z}\in\mathbb{R}^{N+1},\ \tilde{z}_0=1}{\arg\min}\ Q_\varepsilon(\tilde{z}, \tilde{x}^k). \tag{4.19}$$

Despite (4.19) being a constrained optimization, it can be seen as a classical least squares problem, as the following lemma shows

Lemma 4.2.3. *The iterate x^{k+1} defined in* (4.19) *is the minimizer of the unconstrained least squares problem*

$$\min_{z\in\mathbb{R}^N} \frac{\|Az-b\|_2^2}{\max\{\|Ax^k-b\|_2, \varepsilon_0\}} + \lambda\sum_{j=1}^N \frac{|z_j|^2}{\max\{|x_j^k|, \varepsilon\}}.$$

Proof. In view of (4.17), the minimizer of $Q_\varepsilon(\tilde{z}, \tilde{x}^k)$ is also the minimizer of

$$\begin{aligned}\frac{1}{2}\langle \tilde{z}, W_\varepsilon(\tilde{x})\tilde{z}\rangle &= \frac{\|\tilde{A}\tilde{z}\|_2^2}{\max\{\|\tilde{A}\tilde{x}^k\|_2, \varepsilon_0\}} + \lambda\sum_{j=1}^N \frac{|\tilde{z}_j|^2}{\max\{|\tilde{x}_j^k|, \varepsilon\}} \\ &= \frac{\|Az - \tilde{z}_0 b\|_2^2}{\max\{\|Ax^k - \tilde{x}_0^k b\|_2, \varepsilon_0\}} + \lambda\sum_{j=1}^N \frac{|z_j|^2}{\max\{|x_j^k|, \varepsilon\}}.\end{aligned}$$

Reversing the change of variables from $\tilde{z}$ to z with the equalities $\tilde{z}_0 = \tilde{x}_0^k = 1$ gives the desired unconstrained least squares problem,

$$\begin{aligned}\underset{\tilde{z}\in\mathbb{R}^{N+1},\ \tilde{z}_0=1}{\arg\min}\ Q_\varepsilon(\tilde{z},\tilde{x}^k) &= \underset{\tilde{z}\in\mathbb{R}^{N+1},\ \tilde{z}_0=1}{\arg\min} \frac{\|Az-\tilde{z}_0 b\|_2^2}{\max\{\|Ax^k-\tilde{x}_0^k b\|_2,\varepsilon_0\}} + \lambda\sum_{j=1}^N \frac{|z_j|^2}{\max\{|x_j^k|,\varepsilon\}} \\ &= \underset{z\in\mathbb{R}^N}{\arg\min} \frac{\|Az-b\|_2^2}{\max\{\|Ax^k-b\|_2,\varepsilon_0\}} + \lambda\sum_{j=1}^N \frac{|z_j|^2}{\max\{|x_j^k|,\varepsilon\}}.\end{aligned}$$

□

Hence, to determine the new iterate x^{k+1}, we solve a weighted least squares problem, where the weights $\max\{\|Ax^k-b\|_2,\varepsilon_0\}$ and $\max\{|x_j^k|,\varepsilon\}$, $j=1,\dots,N$, are updated after each iteration according to a certain decaying rule. By Lemma 4.2.2, for a fixed regularization parameter ε, the new iterate satisfies

$$0 \le f_\varepsilon(x^{k+1}) = \tilde{f}_\varepsilon(\tilde{x}^{k+1}) \le Q_\varepsilon(\tilde{x}^{k+1},\tilde{x}^k) \le Q_\varepsilon(\tilde{x}^k,\tilde{x}^k) = \tilde{f}_\varepsilon(\tilde{x}^k) = f_\varepsilon(x^k)$$

and, thus, the sequence $\{f_\varepsilon(x^k)\}_{k\ge 0}$ converges. However, as discussed previously in Chapter 2, the key idea is to gradually decrease ε at each iteration to obtain the minimizer of the underlying nonsmooth function (4.24), which constitutes Algorithm 4.

Algorithm 4 Quadratic minimization for Sqrt-LASSO

```
**Input:** Measurement matrix $A\in\mathbb{R}^{m\times N}$, data vector $y\in\mathbb{R}^m$,
initial weight vector $w_0\in\mathbb{R}^N$ (default: $w_0=(1,1,\dots,1)$).
Set $\varepsilon_0=\infty$.
for $k=0,1,2,\dots$ do
    $$x^{k+1} := \underset{z\in\mathbb{R}^N}{\arg\min} \frac{\|Az-b\|_2^2}{\max\{\|Ax^k-b\|_2,\varepsilon_{k,0}\}} + \lambda\sum_{j=1}^N \frac{|z_j|^2}{\max\{|x_j^k|,\varepsilon_k\}}.  \quad (4.20)$$
    Update $\varepsilon_{k+1},\varepsilon_{k+1,0}$ such that $0<\varepsilon_{k+1}\le\varepsilon_k$ and $0<\varepsilon_{k+1,0}\le\varepsilon_{k,0}$   (4.21)
end for
return Sequence $(x^k)_{k\ge 1}$.
```

Note that due to Lemma 4.2.1, our convergence argument above is still applicable, namely,

$$0 \le f_{\varepsilon_{k+1}}(x^{k+1}) \le f_{\varepsilon_k}(x^{k+1}) \le f_{\varepsilon_k}(x^k). \tag{4.22}$$

While this guarantees that Algorithm 4 eventually stops, from this qualitative argument, it is neither possible to infer the properties of the limit point nor quantify the convergence speed. Thus, a more involved analysis is needed. In this chapter, we establish three types of results: convergence speed of IRLS with fixed regularization parameter $\varepsilon_k = \varepsilon$, their counterparts for decaying ε_k, namely, a $O(1/k)$ rate, and global linear convergence under the assumption that A satisfies robust nullspace property. This property is a crucial assumption when dealing with sparse recovery. Moreover, as discussed in Chapter 2, the NSP is a necessary and sufficient condition for sparse recovery via Basis Pursuit [FR13, Chapter 4]. See also [PJ21, Section 3.3] for an extensive discussion.

4.3 Related works

Before the proposal of the sqrt-LASSO (4.5), there were other attempts to define a scale-invariant and pivotal estimator with respect to the noise level. For example, the work [SBvdG10b] proposed an estimator that is scaling invariant and simultaneously estimates the noise level to the sparse vector. However, this estimator still relies on tuning the regularization parameter. This work suggested performing the parameter tuning via a cross-validation procedure or a via Bayesian information criteria (BIC), which is usually not scalable [SBvdG10b, Section 3.4]. This estimator was later called *Scaled LASSO*, and it was claimed that it could be efficiently minimized using coordinate descent [SBvdG10a]. However, no further optimization discussion was provided. The short note [Ant10] proposed to address the concomitant scale invariance and the noise estimation problems with a slightly different estimator inspired by some classical robust regression techniques [Hub81, Chapter 7]. In particular, this note proposed to solve, instead of the LASSO,

$$(\hat{x}_\lambda, \hat{\sigma}_\lambda) \in \operatorname*{arg\,min}_{x\in\mathbb{R}^N, \sigma>0} \frac{1}{2m\sigma}\|y - Ax\|_2^2 + \frac{\sigma}{2} + \lambda\|x\|_1 \tag{4.23}$$

Specifically, [Ant10] raised the question of whether using the estimator (4.23) instead of the original Scaled LASSO proposed by [SBvdG10b] could potentially result in a more efficient optimization algorithm. In his own words *"Do the authors think that such a parametrization could lead to a more efficient optimization algorithm?"*

Subsequently, the paper [SZ12] referred to the estimator proposed by [Ant10], Equation (4.23), as the Scaled LASSO, which led to some confusion in the literature as highlighted in [vdG16, Section 3.1]. This estimator, which jointly estimates the noise and sparse vector, was also implicitly mentioned in [Owe07][3]. It is worth noting that some subsequent works have referred to it as the Concomitant LASSO [NFG$^+$17] or as SPICE (SParse Iterative Covariance-based Estimation) in the signal processing literature [BS14]. Moreover, since its introduction, the sqrt-LASSO estimator has been extended to en-

[3] Indeed, [Owe07] exhibited the Sqrt-LASSO as a member of a more general family of penalized estimators defined. See equations (8) and (9) in [Owe07].

compass various notions of parsimony beyond its initial formulation. For instance, it has been adapted to address group sparsity [BLS14] and matrix completion [Klo14], and more broadly, to accommodate any regularization defined by a norm that fulfills the weak decomposability condition [SvdG17]. It was also studied from the distributionally robust optimization point of view [BKM19], and results about the out-of-sample prediction error are also available [ORVW22].

From the computational point of view, the initial approach presented in [BCW11] employed interior-point and first-order methods for conic programs. Specifically, they utilized the SDT3 implementation of an interior-point method [TTT12] and the TFOCS implementation of first-order methods [BCG11] to minimize Equation (4.5). Subsequently, an ADMM-based solver was implemented in [LZYL15]. Since the minimization sub-steps of ADMM are very costly, a Primal-Dual Hybrid Gradient was developed with simpler sub-steps [?]. The equivalent Scaled LASSO was addressed using a combination of gradient descent and alternating minimization [SZ12], and a coordinate descent strategy was proposed for a smoothed version of the Scaled LASSO, coupled with a pathwise optimization for the tuning parameter to enhance convergence empirically [NFG$^+$17]. The work [?] introduced an overparametrized variational formulation to solve this type of objective function, but they did not manage to obtain a convergence rate for their method. Proximal gradient descent and proximal Newton methods were introduced by [LJH$^+$20] to handle the square-root LASSO objective function, establishing local linear and local quadratic convergence guarantees, respectively. Additionally, both [NFG$^+$17] and [LJH$^+$20] improved convergence performance through the inclusion of a pathwise optimization strategy. More recently, [TWST20] presented a proximal majorization-minimization method for the square-root LASSO. The majority of the papers mentioned above lack theoretical guarantees for their algorithms, except for [TWST20], which demonstrated the convergence of their method to a d-stationary point (cf. [CP21, Chapter 6]), [LJH$^+$20], which established a local convergence theory but required the assumption of a locally restricted strongly smooth condition for its validity, and [?], which developed a primal-dual method with a $O(1/k)$ convergence rate.

In this chapter, we will establish a global convergence rate of IRLS applied to the sqrt-LASSO under minimal assumptions, namely, the robust null space property. However, before that, we start by discussing the convergence theory in the case of a fixed smoothing parameter ε and without any assumption on the measurement matrix A.

4.4 Global convergence with fixed smoothing

We start our convergence analysis by showing a sublinear convergence rate in the case we have a fixed regularization parameter ε, i.e., we do not update ε as in Equation (4.21), but it is rather chosen as a very small positive number. This analysis does not need any structural assumption from the measurement matrix A and can be established only with tools from convex analysis. In our analysis, we will denote the sqrt-LASSO objective function by

$$\|Ax - b\|_2 + \lambda\|x\|_1 =: f_0(x). \tag{4.24}$$

Moreover, we will denote the minimizer of the regularized function f_ε by x_ε^*. The first step is to quantify the decay in the function value for a single iteration. This idea is inspired by the analysis performed for basis pursuit in Chapter 2.

Lemma 4.4.1 (General function value decay rate)**.** *Fix $\varepsilon, \varepsilon_0 > 0$ and let $x \in \mathbb{R}^N$. If the iterate x^k of IRLS, defined by* (2.9)*, satisfies $\langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), \tilde{x} - \tilde{x}^k\rangle \leq 0$ and $x \neq x^k$, then we have*

$$f_\varepsilon(x^{k+1}) - f_\varepsilon(x^k) \leq Q_\varepsilon(\tilde{x}^{k+1}, \tilde{x}^k) - f_\varepsilon(x^k) \leq -\frac{|\langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), \tilde{x} - \tilde{x}^k\rangle|^2}{2\langle W_\varepsilon(x^k)(\tilde{x} - \tilde{x}^k), \tilde{x} - \tilde{x}^k\rangle}.$$

Proof. By construction, $f_\varepsilon(x^{k+1}) = Q_\varepsilon(\tilde{x}^{k+1}, \tilde{x}^k) \leq Q_\varepsilon(\tilde{z}, \tilde{x}^{k+1})$ for any $\tilde{z} \in \mathbb{R}^{1+N}$ such that $\tilde{z}_0 = 1$. Consider $v^k = \tilde{x} - \tilde{x}^k$ and let us evaluate the difference $Q_\varepsilon(\tilde{x}^k + tv^k, \tilde{x}^k) - \tilde{f}_\varepsilon(\tilde{x}^k)$ for some $t > 0$. The idea is that for a properly chosen $t > 0$, the difference $Q_\varepsilon(\tilde{x}^k + tv^k, \tilde{x}^k) - \tilde{f}_\varepsilon(\tilde{x}^k)$ will be negative, which will imply that $\tilde{f}_\varepsilon(\tilde{x}^{k+1}) < \tilde{f}_\varepsilon(\tilde{x}^k)$. Expanding $Q_\varepsilon(\tilde{x}^k + tv^k, \tilde{x}^k)$ yields

$$Q_\varepsilon(\tilde{x}^k + tv^k, \tilde{x}^k) - \tilde{f}_\varepsilon(\tilde{x}^k) = t\langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle + \frac{t^2}{2}\langle W_\varepsilon(x^k)v^k, v^k\rangle = bt + at^2. \tag{4.25}$$

This is a quadratic polynomial with a positive leading coefficient. Indeed,

$$a = \frac{1}{2}\langle W_\varepsilon(x^k)v^k, v^k\rangle = \frac{1}{2}\frac{\|\tilde{A}v^k\|_2^2}{\max\{\|\tilde{A}\tilde{x}^k\|_2, \varepsilon_0\}} + \frac{\lambda}{2}\sum_{j=1}^N \frac{|v_j^k|^2}{\max\{|\tilde{x}_j^k|, \varepsilon\}} \geq 0.$$

The equality is possible if and only if all summands are equal to zero. This means that $v^k = 0$, i.e., $x = x^k$ which contradicts our assumption.

Hence, the quadratic polynomial $Q_\varepsilon(\tilde{x}^k + tv^k, \tilde{x}^k) - \tilde{f}_\varepsilon(x^k)$ attains its minimum at $t =$

$-b/2a$. Therefore, we have that,

$$\begin{aligned} f_\varepsilon(x^{k+1}) - f_\varepsilon(x^k) &\leq Q_\varepsilon(\tilde{x}^{k+1}, \tilde{x}^k) - f_\varepsilon(x^k) \leq \min_{z \ s.t. \ z_0=1} Q_\varepsilon(\tilde{z}, \tilde{x}^k) - f_\varepsilon(x^k) \\ &\leq Q_\varepsilon(\tilde{x}^k + tv^k, \tilde{x}^k) - f_\varepsilon(x^k) \leq -\frac{b^2}{4a} = -\frac{|\langle \nabla \tilde{f}_\varepsilon(x^k), v^k\rangle|^2}{2\langle W_\varepsilon(x^k)v^k, v^k\rangle}. \end{aligned}$$

□

Remark 4.4.2. *The essence of Section 4.4 shares similarities with the sufficient decrease lemma [Bec17, Lemma 10.4], which is commonly employed to establish convergence properties of the proximal gradient descent method. However, an important distinction arises in the approach taken by the author. In the case of Section 4.4, the author introduces the gradient mapping G_L as a generalization of the traditional gradient concept. This definition enables the establishment of convergence guarantees tailored specifically for the proximal gradient descent algorithm.*

With the help of Lemma 4.4.1, we can establish the first result regarding the sublinear convergence of IRLS. In particular, we will show that the sequence $\{x^k\}$ converges sublinearly to the minimizer of the regularized problem, here denoted by x_ε^*.

Theorem 4.4.3 (Sublinear convergence to the minimizer of the regularized problem)**.** *Let $\varepsilon, \varepsilon_0 > 0$. Let x_ε^* be the minimizer of f_ε. Whenever $k \geq 2$, for the iterates x^k of IRLS the inequality*

$$f_\varepsilon(x^k) - f_\varepsilon(x_\varepsilon^*) \leq \max\left\{ \left(\frac{1}{2}\right)^{(k-1)/2} (f_\varepsilon(x^0) - f_\varepsilon(x_\varepsilon^*)), \frac{8(\varepsilon_0^{-1}\|A\|^2 + \lambda\varepsilon^{-1})(f_\varepsilon(x^0) + f_0(x_\varepsilon^*))^2}{\lambda^2(k-1)} \right\}$$

holds.

Proof. We apply Lemma 4.4.1 with $x = x_\varepsilon^*$. Let $v^k = \tilde{x}_\varepsilon^* - \tilde{x}^k$. Note that convexity of $\tilde{f}_\varepsilon$ gives

$$f_\varepsilon(x_\varepsilon^*) = \tilde{f}_\varepsilon(\tilde{x}_\varepsilon^*) \geq \tilde{f}_\varepsilon(\tilde{x}^k) + \langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle = f_\varepsilon(x^k) + \langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle,$$

or, equivalently,

$$\langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle \leq f_\varepsilon(x_\varepsilon^*) - f_\varepsilon(x^k) \leq 0 \text{ and } |\langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle| \geq f_\varepsilon(x^k) - f_\varepsilon(x_\varepsilon^*). \tag{4.26}$$

Hence, Lemma 4.4.1 yields

$$f_\varepsilon(x^{k+1}) - f_\varepsilon(x^k) \leq -\frac{|\langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle|^2}{2\langle W_\varepsilon(x^k)v^k, v^k\rangle} \leq -\frac{(f_\varepsilon(x^k) - f_\varepsilon(x_\varepsilon^*))^2}{2\langle W_\varepsilon(x^k)v^k, v^k\rangle}. \tag{4.27}$$

Let us show that the denominator is bounded from above. By construction we have

$$\begin{aligned}\langle W(x^k)v^k, v^k\rangle &= \frac{\|\tilde{A}v^k\|_2^2}{\max\{\|\tilde{A}\tilde{x}^k\|_2, \varepsilon_0\}} + \lambda\sum_{j=1}^N \frac{|v_j^k|^2}{\max\{|\tilde{x}_j^k|^2, \varepsilon\}}\\ &\leq \frac{\|A(x_\varepsilon^* - x^k)\|_2^2}{\varepsilon_0} + \frac{\lambda}{\varepsilon}\sum_{j=1}^N |(x_\varepsilon^* - x^k)_j|^2 \leq (\varepsilon_0^{-1}\|A\|^2 + \lambda\varepsilon^{-1})\|x_\varepsilon^* - x^k\|_2^2.\end{aligned}$$

Consequently, we only need to bound $\|x_\varepsilon^* - x^k\|_2$. This can be done by contradiction. Assume that $\|x_\varepsilon^* - x^k\|_2 > \lambda^{-1}(f_0(x^k) + f_0(x_\varepsilon^*))$. Then, by the reverse triangle inequality, we have

$$\begin{aligned}f_0(x^k) &= \|Ax^k - b\|_2 + \lambda\|x^k\|_1\\ &\geq \|A(x_\varepsilon^* - x^k)\|_2 - \|Ax_\varepsilon^* + b\|_2 + \lambda\|x_\varepsilon^* - x^k\|_1 - \lambda\|x_\varepsilon^*\|_1.\end{aligned}$$

The first term is nonnegative. The second and the fourth terms together are equal to $-f_0(x_\varepsilon^*)$. The third term can be bounded from below by $\lambda\|x_\varepsilon^* - x^k\|_2$ by monotonicity of the ℓ_p-norms. Hence, using the assumption $\|x_\varepsilon^* - x^k\|_2 > \lambda^{-1}(f_0(x^k) + f_0(x_\varepsilon^*))$, we get

$$f_0(x^k) \geq 0 + \lambda\|x_\varepsilon^* - x^k\|_2 - f_0(x_\varepsilon^*) > \lambda\lambda^{-1}(f_0(x^k) + f_0(x_\varepsilon^*)) - f_0(x_\varepsilon^*) = f_0(x^k),$$

which is a contradiction. Therefore,

$$\|x_\varepsilon^* - x^k\|_2 \leq \lambda^{-1}(f_0(x^k) + f_0(x_\varepsilon^*)) \leq \lambda^{-1}(f_\varepsilon(x^k) + f_0(x_\varepsilon^*)) \leq \lambda^{-1}(f_\varepsilon(x^0) + f_0(x_\varepsilon^*)),$$

where we used that $f_\varepsilon(x^k) \geq f_\varepsilon(x^{k+1}) \geq f_0(x^{k+1})$ for all $k \geq 0$. Now, we substitute the obtained bound in (4.27), which leads to

$$\begin{aligned}[f_\varepsilon(x^{k+1}) - f_\varepsilon(x_\varepsilon^*)] - [f_\varepsilon(x^k) - f_\varepsilon(x_\varepsilon^*)] &\leq -\frac{\lambda^2(f_\varepsilon(x^k) - f_\varepsilon(x_\varepsilon^*))^2}{2(\varepsilon_0^{-1}\|A\|^2 + \lambda\varepsilon^{-1})(f_\varepsilon(x^0) + f_0(x_\varepsilon^*))^2}\\ &\leq -\frac{\lambda^2(f_\varepsilon(x^{k+1}) - f_\varepsilon(x_\varepsilon^*))^2}{2(\varepsilon_0^{-1}\|A\|^2 + \lambda\varepsilon^{-1})(f_\varepsilon(x^0) + f_0(x_\varepsilon^*))^2} \qquad (4.28)\end{aligned}$$

The result of the theorem follows by applying [Bec15b, Lemma 3.8] for the sequence $\{f_\varepsilon(x^k) - f_\varepsilon(x^*)\}_{k\geq 0}$. □

Remark 4.4.4. *Theorem 4.4.3 is similar in its nature to [Bec15b, Theorem 4.2]. The difference is that in [Bec15b, Theorem 4.2], the proof is given for the alternating minimization strategy. This would correspond to the regularized Scaled LASSO objective function, while here, we establish it directly for the regularized sqrt-LASSO formulation without an alternating procedure.*

It is essential to highlight that the convergence rate presented in Theorem Theorem 4.4.3 can be divided into two distinct components. The first case demonstrates the potential for achieving a linear convergence rate, while the second sublinear case becomes dominant in the asymptotic regime. Additionally, it is worth noting that, in principle, it is impossible to theoretically ascertain which of the two cases will apply. Thus, in our next theorem, we derive an alternative result, which provides a guaranteed linear convergence to a neighborhood of the solution x_ε^*.

Theorem 4.4.5 (Linear convergence to the neighborhood of the minimizer of the smoothed problem)**.** *Let* $\varepsilon, \varepsilon_0 > 0$. *Let* x_ε^* *be the minimizer of* f_ε. *If* $f_\varepsilon(x^k) - f_\varepsilon(x_\varepsilon^*) \geq \gamma$ *for some constant* $\gamma > 0$, *then*

$$f_\varepsilon(x^{k+1}) - f_\varepsilon(x_\varepsilon^*) \leq \big(f_\varepsilon(x^k) - f_\varepsilon(x_\varepsilon^*)\big) \left[1 - \frac{\gamma}{2\max\{\frac{1}{\varepsilon_0}, \frac{1}{\lambda\varepsilon}\} f_0^2(x_\varepsilon^*) + 4\gamma}\right].$$

Proof. We start by noting that it would be possible to obtain a linear decay rate by substituting the assumption $f_\varepsilon(x^k) - f_\varepsilon(x_\varepsilon^*) \geq \gamma$ into (4.28). However, such a rate would depend on $f_0(x^0)$, which can potentially be large. Thus, we take a step back to (4.27) and bound the denominator $\langle W_\varepsilon(x^k)v^k, v^k\rangle$ with $v^k = \tilde{x}_\varepsilon^* - \tilde{x}^k$ differently. More precisely, we first decompose it into two parts and connect it with the first derivative $\langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle$ by using that $W_\varepsilon(x)\tilde{x} = \nabla \tilde{f}_\varepsilon(\tilde{x})$ and that $W_\varepsilon(x)$ is a self-adjoint and positive semidefinite matrix,

$$\begin{aligned}
\langle W_\varepsilon(x^k)v^k, v^k\rangle &= \langle W_\varepsilon(x^k)v^k, \tilde{x}_\varepsilon^*\rangle - \langle W_\varepsilon(x^k)v^k, \tilde{x}^k\rangle \\
&= \langle W_\varepsilon(x^k)(\tilde{x}_\varepsilon^* - \tilde{x}^k), \tilde{x}_\varepsilon^*\rangle - \langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle \\
&= \langle W_\varepsilon(x^k)\tilde{x}_\varepsilon^*, \tilde{x}_\varepsilon^*\rangle - \langle W_\varepsilon(x^k)\tilde{x}^k, \tilde{x}_\varepsilon^* \pm \tilde{x}^k\rangle - \langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle \qquad (4.29)\\
&= \langle W_\varepsilon(x^k)\tilde{x}_\varepsilon^*, \tilde{x}_\varepsilon^*\rangle - \langle W_\varepsilon(x^k)\tilde{x}^k, \tilde{x}^k\rangle - 2\langle \nabla \tilde{f}_\varepsilon(\tilde{x}_\varepsilon^k), v^k\rangle \\
&\leq \langle W_\varepsilon(x^k)\tilde{x}_\varepsilon^*, \tilde{x}_\varepsilon^*\rangle + 2|\langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle|.
\end{aligned}$$

Next, the term $\langle W_\varepsilon(x^k)\tilde{x}_\varepsilon^*, \tilde{x}_\varepsilon^*\rangle$ is bounded from above as

$$\begin{aligned}
\langle W_\varepsilon(x^k)\tilde{x}_\varepsilon^*, \tilde{x}_\varepsilon^*\rangle &= \frac{\|\tilde{A}\tilde{x}_\varepsilon^*\|_2^2}{\max\{\|\tilde{A}\tilde{x}^k\|_2, \varepsilon_0\}} + \lambda \sum_{j=1}^{N} \frac{|(x_\varepsilon^*)_j|^2}{\max\{|x_j^k|, \varepsilon\}} \\
&\leq \frac{\|\tilde{A}\tilde{x}_\varepsilon^*\|_2^2}{\varepsilon_0} + \frac{\lambda^2}{\lambda} \sum_{j=1}^{N} \frac{|(x_\varepsilon^*)_j|^2}{\varepsilon} \\
&\leq \max\left\{\frac{1}{\varepsilon_0}, \frac{1}{\lambda\varepsilon}\right\} \left[\|Ax_\varepsilon^* - b\|_2^2 + \lambda^2 \sum_{j=1}^{N} |(x_\varepsilon^*)_j|^2\right] \\
&\leq \max\left\{\frac{1}{\varepsilon_0}, \frac{1}{\lambda\varepsilon}\right\} \left[\|Ax_\varepsilon^* - b\|_2 + \lambda \sum_{j=1}^{N} |(x_\varepsilon^*)_j|\right]^2 = \max\left\{\frac{1}{\varepsilon_0}, \frac{1}{\lambda\varepsilon}\right\} f_0^2(x_\varepsilon^*)
\end{aligned} \tag{4.30}$$

Now, turning to the nominator in Equation (4.27), the bound in (4.26) gives

$$|\langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle|^2 \geq |\langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle| [f_\varepsilon(x^k) - f_0(x_\varepsilon^*)]. \tag{4.31}$$

By plugging Equation (4.29), Equation (4.30) and Equation (4.31) into Equation (4.27), we obtain

$$f_\varepsilon(x^{k+1}) - f_\varepsilon(x_\varepsilon^*) \leq \left[f_\varepsilon(x^k) - f_\varepsilon(x_\varepsilon^*)\right] \left[1 - \frac{|\langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle|}{2\max\{\frac{1}{\varepsilon_0}, \frac{1}{\lambda\varepsilon}\} f_0^2(x_\varepsilon^*) + 4|\langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle|}\right],$$

The second term of the right-hand side has the form $1 - \frac{t}{a+4t} = \frac{3}{4} + \frac{a}{4a+16t}$, where $t = |\langle \nabla \tilde{f}_\varepsilon(\tilde{x}^k), v^k\rangle| \geq \gamma$. The function $\frac{a}{4a+16t}$ is decreasing and, thus, attains its maximum at $t = \gamma$. This gives

$$f_\varepsilon(x^{k+1}) - f_\varepsilon(x^*) \leq \left[f_\varepsilon(x^{k+1}) - f_\varepsilon(x^*)\right] \left[1 - \frac{\gamma}{2\max\{\frac{1}{\varepsilon_0}, \frac{1}{\lambda\varepsilon}\} f_0^2(x^*) + 4\gamma}\right], \tag{4.32}$$

□

Consequently, by Theorem 4.4.5, IRLS converges with a linear rate to an arbitrary neighborhood of the solution of the smoothed problem at a rate determined by the neighborhood's diameter. Once the neighborhood is reached, i.e., once $f_\varepsilon(x^k) - f_\varepsilon(x_\varepsilon^*) < \gamma$, only a sublinear convergence can be guaranteed by Theorem 4.4.3.

The preceding analysis assumed a fixed regularization parameter and derived convergence results towards the solution x_ε^* of the regularized problem. However, the true potential of

IRLS-type algorithms lies in constructing a sequence of objective functions with a varying regularization parameter ε_k, allowing the algorithm to converge towards the solution of the original non-smooth function, as demonstrated in Chapter 2. In the subsequent section, we will establish the possibility of obtaining convergence rate results where the sequence x_ε^k converges to the minimizer of the original sqrt-LASSO problem (4.5), here denoted by x_0^*.

4.5 Global Convergence with decaying ε_k

Now, we turn to the analysis of Algorithm 4 with decaying ε_k. Since $\varepsilon_k, \varepsilon_{k,0} \to 0$ as $k \to \infty$, Algorithm 4 is expected to converge to the solution of (4.5). Just as for constant ε_k, we would like to quantify the convergence of $f_{\varepsilon_k}(x^k)$ to $f_0(x_0^*)$. In contrast to the previous theorem, this situation necessitates a more intricate analysis as the condition $\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{x}^k), \tilde{x}_0^* - \tilde{x}^k \rangle \leq 0$ used in Lemma 4.4.1 may no longer hold. Consequently, an alternative to Theorem 4.4.3 is required.

Theorem 4.5.1 (Sublinear convergence of iterates). *Let $\{\varepsilon_k\}_{k\geq 0}$ and $\{\varepsilon_{k,0}\}_{k\geq 0}$ be two non-increasing sequences. Then, the iterates x^k generated by Algorithm 4 satisfy*

$$\lim_{k\to\infty} \|x^{k+1} - x^k\|_2 = 0 \text{ and } \min_{k=0,\ldots,K-1} \|x^{k+1} - x^k\|_2^2 \leq \frac{2\max\{\lambda\varepsilon_0, f_{\varepsilon_0}(x^0)\}}{\lambda^2 K}[f_{\varepsilon_0}(x^0) - f_0(x_0^*)].$$

Proof. We start by quantifying the difference between $f_{\varepsilon_{k+1}}(x^{k+1})$ and $f_{\varepsilon_k}(x^k)$, i.e., by proving that

$$f_{\varepsilon_{k+1}}(x^{k+1}) - f_{\varepsilon_k}(x^k) \leq Q_{\varepsilon_k}(\tilde{x}^{k+1}, \tilde{x}^k) - f_{\varepsilon_k}(x^k) = -\tfrac{1}{2}\langle W_{\varepsilon_k}(x^k)(\tilde{x}^{k+1} - \tilde{x}^k), \tilde{x}^{k+1} - \tilde{x}^k\rangle.$$

The first inequality is trivial if $x^{k+1} = x^k$. Otherwise, we apply Lemma 4.4.1 with $x = x^{K+1}$ instead of $x = x_0^*$ as it was done in the previous proofs. Note that by the convexity of $\tilde{f}_{\varepsilon_k}$ and definition of x^{k+1}, the assumption of Lemma 4.4.1 is satisfied,

$$\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{x}^k), x^{k+1} - x^k \rangle \leq f_{\varepsilon_k}(x^{k+1}) - f_{\varepsilon_k}(x^k) \leq 0. \tag{4.33}$$

Thus, we get

$$Q_{\varepsilon_k}(\tilde{x}^{k+1}, \tilde{x}^k) - f_{\varepsilon_k}(x^k) \leq -\frac{|\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{x}^k), \tilde{x}^{k+1} - \tilde{x}^k\rangle|^2}{2\langle W_{\varepsilon_k}(x^k)(\tilde{x}^{k+1} - \tilde{x}^k), \tilde{x}^{k+1} - \tilde{x}^k\rangle}.$$

Substituting the definition of $Q_{\varepsilon_k}(\tilde{x}^{k+1}, \tilde{x}^k)$ and combining it with (4.33) gives

$$-|\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{x}^k), \tilde{x}^{k+1} - \tilde{x}^k \rangle| + \frac{1}{2}\langle W_{\varepsilon_k}(x^k)(\tilde{x}^{k+1} - \tilde{x}^k), \tilde{x}^{k+1} - \tilde{x}^k \rangle$$
$$\leq -\frac{|\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{x}^k), \tilde{x}^{k+1} - \tilde{x}^k \rangle|^2}{2\langle W_{\varepsilon_k}(x^k)(\tilde{x}^{k+1} - \tilde{x}^k), \tilde{x}^{k+1} - \tilde{x}^k \rangle}.$$

Let us denote

$$a := |\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{x}^k), \tilde{x}^{k+1} - \tilde{x}^k \rangle| \geq 0 \quad \text{and} \quad b := \langle W_{\varepsilon_k}(x^k)(\tilde{x}^{k+1} - \tilde{x}^k), \tilde{x}^{k+1} - \tilde{x}^k \rangle > 0.$$

Then, the inequality above is equivalent to $-2ab + b^2 \leq -a^2$ and $(a-b)^2 \leq 0$, which is only possible if $a = b$. Another way of seeing that this holds is to look at the KKT conditions of the problem $\min_{z \in \mathbb{R}^{N+1},\ z_0 = 1} Q_{\varepsilon_k}(\tilde{z}, \tilde{x}^k)$. In fact, it holds that $\langle W_{\varepsilon_k}(x^k)(\tilde{x}^{k+1} - \tilde{x}^k), v \rangle = -\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{x}^k), v \rangle$ for all $v \in \mathbb{R}^{N+1}$ such that $v_0 = 0$. Substituting the obtained equality into $Q_{\varepsilon_k}(\tilde{x}^{k+1}.\tilde{x}^k)$ gives

$$Q_{\varepsilon_k}(\tilde{x}^{k+1}, \tilde{x}^k) = \tilde{f}_{\varepsilon_k}(\tilde{x}^k) - a + \tfrac{b}{2} = \tilde{f}_{\varepsilon_k}(\tilde{x}^k) - \tfrac{1}{2}\langle W_{\varepsilon_k}(x^k)(\tilde{x}^{k+1} - \tilde{x}^k), \tilde{x}^{k+1} - \tilde{x}^k \rangle.$$

Together with the majorization property from Lemma 4.2.2, it yields the inequality stated at the beginning of the proof, namely,

$$\tilde{f}_{\varepsilon_{k+1}}(\tilde{x}^{k+1}) \leq \tilde{f}_{\varepsilon_k}(\tilde{x}^{k+1}) \leq Q_{\varepsilon_k}(\tilde{x}^{k+1}, \tilde{x}^k) = \tilde{f}_{\varepsilon_k}(\tilde{x}^k) - \tfrac{1}{2}\langle W_{\varepsilon_k}(x^k)(\tilde{x}^{k+1} - \tilde{x}^k), \tilde{x}^{k+1} - \tilde{x}^k \rangle. \tag{4.34}$$

Now, we bound the quadratic term from below in terms of the squared distance $\|x^{k+1} - x^k\|_2^2$. Using the definition of $W_{\varepsilon_k}(x^k)$, we get

$$\langle W_{\varepsilon_k}(x^k)(\tilde{x}^{k+1} - \tilde{x}^k), \tilde{x}^{k+1} - \tilde{x}^k \rangle = \frac{\|\tilde{A}(\tilde{x}^{k+1} - \tilde{x}^k)\|_2^2}{\max\{\|\tilde{A}\tilde{x}^k\|_2^2, \varepsilon_{k,0}\}} + \lambda \sum_{j=1}^{N} \frac{|(\tilde{x}^{k+1} - \tilde{x}^k)_j|^2}{\max\{|\tilde{x}_j^k|, \varepsilon_k\}}$$
$$\geq 0 + \lambda \sum_{j=1}^{N} \frac{|(x^{k+1} - x^k)_j|^2}{\max\{|x_j^k|, \varepsilon_k\}} \geq \frac{\lambda \|x^{k+1} - x^k\|_2^2}{\max\{\|x^k\|_\infty, \varepsilon_k\}}.$$

Furthermore, by construction, we have $\varepsilon_k \leq \varepsilon_0$ and

$$\lambda \|x^k\|_\infty \leq \lambda \|x^k\|_1 \leq \|Ax^k - b\|_2 + \lambda \|x^k\|_1 = f_0(x^k) \leq f_{\varepsilon_k}(x^k) \leq f_{\varepsilon_0}(x^0).$$

Consequently, the quadratic term satisfies

$$\langle W_{\varepsilon_k}(x^k)(\tilde{x}^{k+1} - \tilde{x}^k), \tilde{x}^{k+1} - \tilde{x}^k \rangle \geq \frac{\lambda \|x^{k+1} - x^k\|_2^2}{\max\{\lambda^{-1} f_{\varepsilon_0}(x^0), \varepsilon_0\}} = \frac{\lambda^2 \|x^{k+1} - x^k\|_2^2}{\max\{f_{\varepsilon_0}(x^0), \lambda\varepsilon_0\}}.$$

Returning to Equation (4.34), we obtain

$$\frac{\lambda^2\|x^{k+1}-x^k\|_2^2}{2\max\{f_{\varepsilon_0}(x^0),\lambda\varepsilon_0\}} \leq \tilde{f}_{\varepsilon_k}(\tilde{x}^k) - \tilde{f}_{\varepsilon_{k+1}}(\tilde{x}^{k+1}) = f_{\varepsilon_k}(x^k) - f_{\varepsilon_{k+1}}(x^{k+1})$$

Summing up for $k = 0, \dots, K-1$, for some $K \in \mathbb{N}$, leads to

$$\sum_{k=0}^{K-1} \frac{\lambda^2\|x^{k+1}-x^k\|_2^2}{2\max\{f_{\varepsilon_0}(x^0),\lambda\varepsilon_0\}} \leq \sum_{k=0}^{K-1}[f_{\varepsilon_k}(x^k) - f_{\varepsilon_{k+1}}(x^{k+1})] = f_{\varepsilon_0}(x^0) - f_{\varepsilon_k}(x^K).$$

By taking $\frac{\lambda^2}{2\max\{f_{\varepsilon_0}(x^0),\lambda\varepsilon_0\}}$ to the right-hand side, we observe that the partial sum of the series is bounded from above by

$$\sum_{k=0}^{K-1}\|x^{k+1}-x^k\|_2^2 \leq \frac{2\max\{f_{\varepsilon_0}(x^0),\lambda\varepsilon_0\}}{\lambda^2}[f_{\varepsilon_0}(x^0)-f_{\varepsilon_k}(x^K)] \leq \frac{2\max\{f_{\varepsilon_0}(x^0),\lambda\varepsilon_0\}}{\lambda^2}[f_{\varepsilon_0}(x^0)-f_0(x_0^*)].$$

This bound is independent of K and, thus, the series $\sum_{k=0}^{\infty}\|x^{k+1}-x^k\|_2^2$ is convergent. As a result its summands $\|x^{k+1}-x^k\|_2^2$ converge to zero as $k \to \infty$. Finally, we bound the minimum of the first K summands by their mean,

$$\min_{k=0,\dots,K-1}\|x^{k+1}-x^k\|_2^2 \leq \frac{1}{K}\sum_{k=0}^{K-1}\|x^{k+1}-x^k\|_2^2 \leq \frac{2\max\{f_{\varepsilon_0}(x^0),\lambda\varepsilon_0\}}{\lambda^2 K}[f_{\varepsilon_0}(x^0)-f_0(x_0^*)].$$

□

Unlike previous results, Theorem 4.5.1 states convergence in terms of distance between two consequence iterates instead of function values. Furthermore, it is, in principle, not clear if $f_{\varepsilon_k}(x^k)$ converges to the optimal value $f_0(x_0^*)$, which is the aim of the next part of this subsection. First, we derive a linear convergence to the neighborhood of the minimizer x_0^*, as it was analogously done to the case when ε_k is fixed.

Theorem 4.5.2 (Linear convergence to the neighborhood of the minimizer of the original problem). *Let $\varepsilon_k, \varepsilon_{k,0} > 0$ and set x_0^* as the minimizer of f_0. If $f_{\varepsilon_k}(x^k) - f_0(x_0^*) \geq \gamma(\lambda N\varepsilon_k + \varepsilon_{k,0})$ for some constant $\gamma > 1$, then*

$$f_{\varepsilon_k}(x^{k+1}) - f_0(x_0^*) \leq \left(f_{\varepsilon_k}(x^k) - f_0(x_0^*)\right)\left[1 - \frac{(\gamma-1)^2(\lambda N\varepsilon_k + \varepsilon_{k,0})}{2\gamma\max\{\frac{1}{\varepsilon_{k,0}},\frac{1}{\lambda\varepsilon_k}\}f_0^2(x_0^*) + 4(\gamma-1)^2(\lambda N\varepsilon_k+\varepsilon_{k,0})}\right].$$

Proof. The proof follows a similar structure to that of Theorems 4.4.3 and 4.4.5 discussed earlier. Specifically, we employ Lemma 4.4.1 and establish an upper bound for the denominator, analogous to the approach utilized in Theorem 4.4.5. However, Lemma 4.4.1

relies on (4.26), which is no longer true for x_0^*. Consequently, we derive an alternative bound to address this issue. Similarly to the previous case, the convexity of $\tilde{f}_\varepsilon$ yields

$$f_{\varepsilon_k}(x_0^*) = \tilde{f}_{\varepsilon_k}(\tilde{x}_0^*) \geq \tilde{f}_{\varepsilon_k}(\tilde{x}^k) + \langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{x}^k), \tilde{x}_0^* - \tilde{x}^k \rangle = f_{\varepsilon_k}(x^k) + \langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{x}^k), \tilde{x}_0^* - \tilde{x}^k \rangle$$

Then, by Lemma 4.2.1 and the assumption $f_{\varepsilon_k}(x^k) - f_0(x_0^*) \geq \gamma(\lambda N \varepsilon_k + \varepsilon_{k,0})$, we obtain

$$\begin{aligned} -\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{x}^k), \tilde{x}_0^* - \tilde{x}^k \rangle &\geq f_{\varepsilon_k}(x^k) - f_{\varepsilon_k}(x_0^*) \geq f_{\varepsilon_k}(x^k) - f_0(x_0^*) - (\lambda N \varepsilon_k + \varepsilon_{k,0}) \\ &\geq (\gamma - 1)(\lambda N \varepsilon_k + \varepsilon_{k,0}) \geq 0, \end{aligned} \tag{4.35}$$

and

$$-\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{x}^k), \tilde{x}_0^* - \tilde{x}^k \rangle \geq f_{\varepsilon_k}(x^k) - f_0(x_0^*) - (\lambda N \varepsilon_k + \varepsilon_{k,0}) \geq (1 - 1/\gamma)[f_{\varepsilon_k}(x^k) - f_0(x_0^*)].$$

Using these bounds instead of (4.26) leads to the desired result. □

4.6 Discussion on ε decay

An important element of Theorem 4.5.2 is the condition $f_{\varepsilon_k}(x^k) - f_0(x_0^*) \geq \gamma(\lambda N \varepsilon_k + \varepsilon_{k,0})$, which plays a crucial role in establishing convergence. While Theorem 4.4.5 and Theorem 4.5.2 share similarities, the latter exhibits a convergence rate that approaches one as ε_k and $\varepsilon_{k,0}$ tend to zero. However, depending on the rate at which the smoothing parameters approach zero, the bound presented in Theorem 4.5.2 may not always yield a meaningful convergence result. In this section, we establish a connection between the decay of smoothing parameters and the convergence rate, providing further insights into the analysis.

Lemma 4.6.1. *Let $K \geq 1$, $\gamma > 1$ and $0 < \nu < 1$. Assume that $\varepsilon_{k,0} = \lambda \varepsilon_k$ and define*

$$c := \gamma \lambda (N+1) \quad \textit{and} \quad d := \frac{2\gamma f_0^2(x_0^*)}{\lambda^2(\gamma-1)^2(N+1)} \tag{4.36}$$

Then, the iterate x^{K+1} of Algorithm 4 admits

$$f_{\varepsilon_{K+1}}(x^{K+1}) - f_0(x_0^*) \leq \max \left\{ c\varepsilon_{\lfloor K^\nu \rfloor}, (f_{\varepsilon_0}(x^0) - f_0(x_0^*)) \prod_{k=\lfloor K^\nu \rfloor + 1}^{K} \left[1 - \frac{1}{d\varepsilon_k^{-2} + 4} \right] \right\},$$

where $\varepsilon_{\lfloor K^\nu \rfloor}$ denotes the regularization parameter ε_k at the iteration $k = \lfloor K^\nu \rfloor$ for a certain $0 < \nu < 1$.

Proof. The proof differentiates between two possible cases depending on how many times the inequality $f_{\varepsilon_k}(x^k) - f_0(x_0^*) < c\varepsilon_k$ is satisfied. Firstly, assume that it is satisfied for at least $\lfloor K^\nu \rfloor$ indices k and denote the largest one of them by k_0, i.e., $k_0 \geq \lfloor K^\nu \rfloor$. Then, by the inequality (4.22), we have

$$f_{\varepsilon_{K+1}}(x^{K+1}) \leq f_{\varepsilon_k}(x^K) \leq \ldots \leq f_{\varepsilon_{k_0}}(x^{k_0}) \leq f_{\varepsilon_{\lfloor K^\nu \rfloor}}(x^{\lfloor K^\nu \rfloor}) < f_0(x_0^*) + c\varepsilon_{\lfloor K^\nu \rfloor}.$$

Now, let us assume that the opposite holds, i.e., that there are less than $\lfloor K^\nu \rfloor$ indices k for which $f_{\varepsilon_k}(x^k) - f_0(x_0^*) < c\varepsilon_k$. In this case, this inequality implies that there are at least $K - \lfloor K^\nu \rfloor + 1$ indices k such that the opposite holds. Let us denote all these indices by a set $\mathcal{K}$. If $k \notin \mathcal{K}$, we can use the bound $f_{\varepsilon_{k+1}}(x^{k+1}) - f_0(x_0^*) \leq f_{\varepsilon_k}(x^k) - f_0(x_0^*)$ that holds due to the monotonicity of ε. Otherwise, by Theorem 4.5.2, we have

$$f_{\varepsilon_{k+1}}(x^{k+1}) - f_0(x_0^*) \leq f_{\varepsilon_k}(x^k) - f_0(x_0^*) \left[1 - \frac{1}{d\varepsilon_k^{-2} + 4}\right].$$

Combining these two bounds together yields

$$f_{\varepsilon_{k+1}}(x^{K+1}) - f_0(x_0^*) \leq (f_{\varepsilon^0}(x^0) - f_0(x_0^*)) \prod_{k \in \mathcal{K}} \left[1 - \frac{1}{d\varepsilon_k^{-2} + 4}\right]$$

By construction, $\varepsilon_{k+1} \leq \varepsilon_k$. Thus, the product on the right-hand side is the largest, when the set $\mathcal{K}$ is $\{\lfloor K^\nu \rfloor, \ldots, K\}$, which gives

$$f_{\varepsilon_{k+1}}(x^{K+1}) - f_0(x_0^*) \leq (f_{\varepsilon^0}(x^0) - f_0(x_0^*)) \prod_{k=\lfloor K^\nu \rfloor}^{K} \left[1 - \frac{1}{d\varepsilon_k^{-2} + 4}\right].$$

□

We first note that assumption $\varepsilon_0^k = \lambda\varepsilon_k$ is only used to simplify the formulas. Lemma 4.6.1 highlights the impact of ε_k. If ε_k decays slowly, the product quickly becomes small and the first term dominates. On the other hand, if ε_k decays fast, the product may converge to a nonzero value. By [LTVB22, Theorem 2.2.2], the infinite product $\prod_{k\geq 0}\left[1 - \frac{1}{d\varepsilon_k^{-2}+4}\right]$ diverges to zero [4] if and only if the series $\sum_{k\geq 0}[d\varepsilon_k^{-2} + 4]^{-1}$ diverges. The latter, in turn, is equivalent to the divergence of the series $\sum_{k\geq 0} \varepsilon_k^2$.

For instance, consider a sequence $\varepsilon_k = \varepsilon_0(1+k)^{-\theta}$ with starting value $\varepsilon_0 > 0$ and decay parameter $\theta > 0$. Consequently, if $\theta > 1/2$, the product does not diverge to zero. Yet, Lemma 4.6.1 only provides an upper bound for the rate and does not imply that $f_0(x_0^*)$

[4] We follow the standard denomination from the theory of infinite products that treats zero as a special case since the product *diverges to zero* if and only if the series $\sum_{n=1}^{\infty} \log(a_n)$ diverges to $-\infty$.

is not the limit of $f_{\varepsilon_k}(x^k)$. When $\theta \le 1/2$, the product diverges to zero. However, even if the product vanishes quickly, the right-hand side is proportional to ε_k, which decays sublinearly. In general, we are able to establish the following sublinear convergence rate.

Theorem 4.6.2. *Consider the sequence $\varepsilon_k = \varepsilon_0(1+k)^{-\theta}$ with starting value $\varepsilon_0 > 0$ and decay parameter and $0 < \theta \le 1/2$. For $K \ge 1$ and $\gamma > 1$, the sequence x^k generated by Algorithm 4 admits the following decay*

$$f_{\varepsilon_{K+1}}(x^{K+1}) - f_0(x_0^*) \le \max\left\{c\varepsilon_0, e^{\varepsilon_0^2/d}(6 + 4d^{-1}\varepsilon_0^2)^{\varepsilon_0^2/d}(f_{\varepsilon_0}(x^0) - f_0(x_0^*))\right\} K^{-\frac{\theta\varepsilon_0^2}{d\theta+\varepsilon_0^2}}.$$

where constants c and d are defined in (4.36).

Proof. Let $q = d^{-1}\varepsilon_0^2$ and $0 < \nu < 1$, whose precise value of which will be determined later. Then, an application of Lemma 4.6.1 gives

$$f_{\varepsilon_{K+1}}(x^{K+1}) - f_0(x_0^*) \le \max\left\{c\varepsilon_0(\lfloor K^\nu \rfloor + 1)^{-\theta}, (f_{\varepsilon_0}(x^0) - f_0(x_0^*)) \prod_{k=\lfloor K^\nu \rfloor}^{K}\left[1 - \frac{1}{q^{-1}(k+1)^{2\theta} + 4}\right]\right\}.$$

The first term is already of the desired form since

$$c\varepsilon_0(\lfloor K^\nu \rfloor + 1)^{-\theta} \le c\varepsilon_0(K^\nu - 1 + 1)^{-\theta} = c\varepsilon_0 K^{-\nu\theta}.$$

Hence, we look at the second term and estimate the product. Note that this product is increasing in θ, and we can bound it by

$$\begin{aligned}\prod_{k=\lfloor K^\nu \rfloor}^{K}\left[1 - \frac{1}{q^{-1}(k+1)^{2\theta} + 4}\right] &\le \prod_{k=\lfloor K^\nu \rfloor}^{K}\left[1 - \frac{1}{q^{-1}(k+1) + 4}\right] \\ &= \prod_{k=\lfloor K^\nu \rfloor}^{K}\left[1 - \frac{q}{k+1+4q}\right] \le \prod_{k=\lfloor K^\nu \rfloor}^{K}\left[1 - \frac{q}{k+1+4\lceil q\rceil}\right] = \prod_{k=\lfloor K^\nu \rfloor+1+4\lceil q\rceil}^{K+1+4\lceil q\rceil}\left[1 - \frac{q}{k}\right].\end{aligned}$$

Let $n = K + 4\lceil q\rceil$, $p = \lfloor K^\nu \rfloor + 4\lceil q\rceil$, $s = \lfloor q \rfloor$ and $r = q - s$. Then, the product can be expressed in terms of gamma function Γ,

$$\begin{aligned}\prod_{k=p+1}^{n+1}\left[1 - \frac{q}{k}\right] = \prod_{k=p+1}^{n+1}\frac{k-q}{k} &= \frac{(n+1-q)\cdot\ldots\cdot(p+1-q)}{(n+1)\cdot\ldots\cdot(p+1)}\cdot\frac{\Gamma(p+1-q)}{\Gamma(p+1-q)}\cdot\frac{p!}{p!} \\ &= \frac{\Gamma(n+2-q)\,p!}{(n+1)!\,\Gamma(p+1-q)} = \frac{\Gamma(n-s+1+1-r)\,p!}{(n+1)!\,\Gamma(p-s+1-r)}.\end{aligned}$$

Note that all factors here are strictly positive since

$$(n+1-q) \geq \ldots \geq (p+1-q) = \lfloor K^\nu \rfloor + 4\lceil q \rceil + 1 - q \geq \lfloor K^\nu \rfloor + 3\lceil q \rceil + 1 \geq 2 > 0.$$

The next step is to bound the Gamma functions using Gautschi's double inequality [Gau59, Qi10]. This leads to

$$\Gamma(j+1)(j+1)^{-(1-\alpha)} \leq \Gamma(j+\alpha) \leq \Gamma(j+1)j^{-(1-\alpha)}, \quad j \in \mathbb{N},\ 0 \leq \alpha \leq 1.$$

Its application for $j = n-s+1$ and $j = p-s$ with $\alpha = 1-r$ gives

$$\begin{aligned}\frac{\Gamma(n-s+1+1-r)\ p!}{(n+1)!\ \Gamma(p-s+1-r)} &\leq \frac{\Gamma(n-s+2)\ (n-s+1)^{-r}\ p!}{(n+1)!\Gamma(p-s+1)\ (p-s+1)^{-r}} \\ &= \frac{(n-s+1)!\ p!\ (p-s+1)^r}{(n+1)!\ (p-s)!\ (n-s+1)^r}.\end{aligned}$$

Recall the definition of the binomial coefficient $\binom{j}{k}$, it holds the bound $\frac{j^k}{k^k} \leq \binom{j}{k} \leq \frac{e^k j^k}{k^k}$. In our case, this leads to

$$\frac{(n-s+1)!\ p!\ (p-s+1)^r}{(n+1)!\ (p-s)!\ (n-s+1)^r} = \frac{\binom{p}{s}\ (p-s+1)^r}{\binom{n+1}{s}\ (n-s+1)^r} \leq \frac{e^s p^s (p-s+1)^r}{(n+1)^s (n-s+1)^r}.$$

To simplify the resulting fraction, we note that $e^s \leq e^{r+s}$, $p^s \leq (p+1)^s$, $(p-s+1)^r \leq (p+1)^r$ and that

$$n+1 \geq n+1-s = K + 4\lceil q \rceil + 1 - \lfloor q \rfloor \geq K + 3q + 1 \geq K.$$

Consequently, these estimates yield

$$\begin{aligned}\frac{e^s p^s (p-s+1)^r}{(n+1)^s (n-s+1)^r} &\leq \frac{e^{s+r}(p+1)^{s+r}}{K^{s+r}} = \frac{e^q (p+1)^q}{K^q} = \frac{e^q(\lfloor K^\nu \rfloor + 4\lceil q \rceil + 1)^q}{K^q} \\ &\leq e^q(2+4\lceil q \rceil)^q K^{\nu q} K^{-q} \leq e^q (6+4q)^q K^{-(1-\nu)q}.\end{aligned}$$

Combining everything together, we arrive at

$$f_{\varepsilon_{k+1}}(x^{K+1}) - f_0(x_0^*) \leq \max\left\{c\varepsilon_0 K^{-\nu\theta}, e^q(6+4q)^q (f_{\varepsilon_0}(x^0) - f_0(x_0^*))K^{-(1-\nu)q}\right\}.$$

The last step is to select ν such that the powers $\nu\theta$ and $(1-\nu)q$ coincide. This gives $\nu = q/(\theta+q)$, which implies that $0 < \nu < 1$ and

$$K^{-(1-\nu)q} = K^{-\nu\theta} = K^{-\theta q/(\theta+q)}.$$

Substituting $q = d^{-1}\varepsilon_0^2$ concludes the proof. □

Theorem 4.6.2 derives a sublinear convergence rate, which is slightly worse than $K^{-\theta}$ the decay rate of ε_k. As the power $\theta\varepsilon_0^2/(d\theta + \varepsilon_0^2)$ is increasing as a function of ε_0, $K^{-\theta q/(\theta+q)}$ will converge asymptotically to $K^{-\theta}$ as $\varepsilon_0 \to +\infty$. However, in this case, the constant $e^{\varepsilon_0^2/d}(6 + 4d^{-1}\varepsilon_0^2)^{\varepsilon_0^2/d}$ blows up at a much faster pace. In any case, it is important to understand the rate given by the previous theorem. Indeed, the maximum of the exponent $-\theta q/(\theta + q)$ is given by

$$\max_{q,\theta} \frac{\theta q}{\theta + q} = \max_{q,\theta} \frac{(\theta + q)q}{\theta + q} - \frac{q^2}{\theta + q} = \max_{q,\theta} q - \frac{q^2}{\theta + q}. \tag{4.37}$$

Since $\theta \in (0, 1/2]$. the maximum above, as a function of θ, is attained when $\theta = 1/2$. Hence, this yields

$$\max_{q} q - \frac{q^2}{1/2 + q} = \max_{q} \frac{1}{2}\frac{q}{q + 1/2}. \tag{4.38}$$

When $q \to \infty$, the maximum of the expression above is 1/2. There $K^{-1/2}$ is, in principle, the best possible rate that one can obtain by using this technique. As discussed previously, Theorem 4.5.1 also reveals that the convergence of the iterates exhibits a rate of $K^{-1/2}$ for any θ. Remarkably, in the recent work [Chi21], it was demonstrated that for a related algorithm, namely the Bregman gradient method (but operating in an infinite-dimensional space of measures), the rate $K^{-1/2}$ is proven to be optimal and cannot be improved without imposing additional assumptions.

It is important to put the two main results established so far in this chapter into perspective. Theorem 4.4.3 established a K^{-1} for a fixed smoothing parameter ε, whereas the aforementioned result, Theorem 4.6.2, exhibits a slower convergence rate given by $K^{-1/2}$. At first glance, it may appear that we are obtaining a less favorable result for a more desirable problem, as the decay of ε_k brings us closer to the true function f_0. However, as ε_k decreases, the problem gets closer to a non-differentiable one, which poses greater challenges for optimization due to a deteriorating Lipschitz constant. Also, the limit point in both cases is different. While in Theorem 4.4.3, the algorithm convergence to x_ε^*, in Theorem 4.6.2, the limit point is x_0^*.

At the end of this section, we consider a construction of ε_k that, to a certain extent, optimizes the constant involved. Now, we choose a sequence ε_k that makes the rate established in Theorem 4.6.2 independently from $d := \frac{2\gamma f_0^2(x_0^*)}{\lambda^2(\gamma-1)^2(N+1)}$.

Corollary 4.6.3. *Consider the sequence*

$$\varepsilon_k = \frac{f_{\varepsilon_{k-1}}(x^k)}{2\lambda\sqrt{N+1}(1+k)^\theta}, \quad k \geq 0,$$

where, for convenience, $\varepsilon_{-1} > 0$ and $0 < \theta \leq 1/2$. For $K \geq 1$, the sequence x^k generated by Algorithm 4 admits

$$f_{\varepsilon_{k+1}}(x^{K+1}) - f_0(x_0^*) \leq \max\left\{\sqrt{N+1} f_{\varepsilon_0}(x^0), 10e(f_{\varepsilon_0}(x^0) - f_0(x_0^*))\right\} K^{-\frac{\theta}{\theta+1}}.$$

Proof. The proof is similar to the proof of Theorem 4.6.2 with only minor changes. An application of Lemma 4.6.1 with $\gamma = 2$ and any $0 < \nu < 1$ gives

$$\begin{aligned} f_{\varepsilon_{k+1}}(x^{K+1}) - f_0(x_0^*) \leq \max &\left\{ \frac{c f_{\varepsilon_{\lfloor K^\nu \rfloor - 1}}(x^{\lfloor K^\nu \rfloor})}{\lambda\sqrt{2N+2}} (\lfloor K^\nu \rfloor + 1)^{-\theta}, \right. \\ &\left. (f_{\varepsilon_0}(x^0) - f_0(x_0^*)) \prod_{k=\lfloor K^\nu \rfloor}^{K} \left[1 - \frac{1}{d f_{\varepsilon_{k-1}}^{-2}(x^k) \lambda^2 (2N+2)(k+1)^{2\theta} + 4} \right] \right\}, \end{aligned}$$

where c and d are defined in (4.36). The first term is again bounded by

$$\frac{c f_{\varepsilon_{\lfloor K^\nu \rfloor - 1}}(x^{\lfloor K^\nu \rfloor})}{2\lambda\sqrt{N+1}} (\lfloor K^\nu \rfloor + 1)^{-\theta} \leq \sqrt{N+1} f_{\varepsilon_0}(x^0)(K^\nu - 1 + 1)^{-\theta} = \sqrt{N+1} f_{\varepsilon_0}(x^0) K^{-\nu\theta}.$$

For the second term, we observe that

$$d f_{\varepsilon_{k-1}}^{-2}(x^k) \lambda^2 (2N+2) = \frac{f_0^2(x_0^*)}{f_{\varepsilon_{k-1}}^2(x^k)} \leq 1, \quad \text{for all } k \geq 0.$$

Using this bound for the product, the rest of the proof repeats the steps from Theorem 4.6.2 with $q = 1$. In particular, $\nu = 1/(1+\theta)$. □

Remark 4.6.4. *In the proof above, we could also have used the monotone sequence $\varepsilon_k = \frac{\min_{s=0,\ldots,k} f_0(x^s)}{2\lambda\sqrt{N+1}(1+k)^\theta}$ instead of $\varepsilon_k = \frac{f_{\varepsilon_{k-1}}(x^k)}{2\lambda\sqrt{N+1}(1+k)^\theta}$.*

Without any additional assumptions, deriving a global linear convergence rate result for IRLS, aimed at minimizing a scaled Huber loss function $j_\varepsilon(x)$, for the sqrt-LASSO problem, appears to be unattainable, both in cases with a fixed ε and with a decaying ε, as we expect that similar lower bounds as the one from [Chi21] can be obtained. However, it is possible to establish such convergence rate for a fixed ε by choosing, instead of the Huber function, the smoothing function $\sqrt{|x|^2 + \varepsilon^2}$, which leads to the objective

$$\hat{f}_\varepsilon(x) := \sqrt{\|Ax - y\|_2^2 + \varepsilon_0^2} + \lambda \sum_{j=1}^{N} \sqrt{|x_j|^2 + \varepsilon^2}.$$

In this context, we outline a concise explanation for establishing the linear convergence rate by utilizing the KL property, an essential analytical tool for studying nonconvex nonsmooth problems [ABRS10, YLP22].

Definition 4.6.5 (Kurdyka-Lojasiewicz property). *We say that a proper closed function $h : \mathbb{X} \to \mathbb{R} \cup \{\infty\}$ satisfies the Kurdyka-Lojasiewicz (KL) property at $\hat{x} \in \operatorname{dom} \partial h$ if there are $a \in (0, \infty]$, a neighborhood V of $\hat{x}$ and a continuous concave function $\varphi : [0, a) \to [0, \infty)$ with $\varphi(0) = 0$ such that*

(i) *φ is continuously differentiable on $(0, a)$ with $\varphi' > 0$ on $(0, a)$;*

(ii) *For any $x \in V$ with $h(\hat{x}) < h(x) < h(\hat{x}) + a$, it holds that*

$$\varphi'(h(x) - h(\hat{x}))\operatorname{dist}(0, \partial h(x)) \geq 1. \tag{4.39}$$

If h satisfies the KL property at $\hat{x} \in \operatorname{dom} \partial h$ and the $\varphi(s)$ in (4.39) can be chosen as $\bar{c}\, s^{1-\alpha}$ for some $\bar{c} > 0$ and $\alpha \in [0, 1)$, then we say that h satisfies the KL property at $\hat{x}$ with exponent α.
A proper closed function h satisfying the KL property at every point in $\operatorname{dom} \partial h$ is said to be a KL function, and a proper closed function h satisfying the KL property with exponent $\alpha \in [0, 1)$ at every point in $\operatorname{dom} \partial h$ is said to be a KL function with exponent α.

First note that the results of Theorem 4.4.5 remain true for IRLS applied to $\hat{f}_\varepsilon$. Next, we establish a local linear convergence rate of IRLS based on [BP16, Proposition 4]. For that, we need three conditions to be satisfied:

1. $\hat{f}_\varepsilon(x^{k+1}) - \hat{f}_\varepsilon(x^k) \leq C_1 \|\nabla \hat{f}_\varepsilon(x^k)\|_2$ for some constant $C_1 > 0$;
2. $\|\nabla \hat{f}_\varepsilon(x^k)\|_2 \leq C_2 \|x^{k+1} - x^k\|_2$ for some constant $C_2 > 0$;
3. $\hat{f}_\varepsilon$ satisfies Kurdyka-Lojasiewicz (KL) property with exponent 1/2 or less (see Lemma 2.2 of [YLP22]).

The first property follows similarly to Theorem 4.5.1. The second property follows from the $1/\varepsilon$-smoothness of $\sqrt{|x|^2 + \varepsilon^2}$ (see [Bec17, Example 10.44]) combined with Karush-Kuhn-Tucker conditions for the constrained problem described in Equation (4.19). Lastly, unlike the scaled Huber function $j_\varepsilon(x)$, the function $\sqrt{|x|^2 + \varepsilon^2}$ can be represented via linear matrix inequalities [YLP22], and, by [YLP22, Theorem 4.3], the loss $\hat{f}_\varepsilon$ has KL exponent 1/2. Thus, by [BS17, Lemma 2.5], the KL constant $\bar{c} > 0$ can be estimated and, therefore, by following an argument similar to the one developed in [BNPS17], it is possible to show that there exists a neighborhood of x_ε^*, estimated via $\bar{c} > 0$, in which IRLS admits a linear convergence rate. Then, by selecting γ in Theorem 4.4.5 appropriately, the IRLS algorithm designed to minimize the function $\sqrt{|x|^2 + \varepsilon^2}$ admits a linear convergence rate to this neighborhood, which implies that IRLS has a global linear convergence rate to a solution of the smoothed SQRT-LASSO problem. However, this solution potentially differs from the solution of the true sqrt-LASSO objective function. Moreover, this proof will not remain true for a decaying sequence ε_k as the smoothness of $\hat{f}_\varepsilon$ deteriorates as ε_k vanishes and, therefore, the second property no longer holds. We leave it as an open problem to characterize the set of smoothed versions of the sqrt-LASSO for which the technique above can be applied.

Open Problem: For which other functions, beyond $\sqrt{|x|^2 + \varepsilon^2}$, is it possible to apply the technique above? Is there a simple characterization of this set? Moreover, is it possible to extend the analysis from [BP16] to the ε_k decaying case, i.e., for a sequence of smoothing functions indexed by ε_k that smooth the sqrt-LASSO?

So far we provided a general convergence analysis of the IRLS method applied to the objective function

$$f_\varepsilon(x) = j_{\varepsilon_0}(\|Ax - b\|_2) + \sum_{i=1}^{N} j_\varepsilon(x_i),$$

that is designed to tackle the non-smoothness of the sqrt-LASSO objective function.
As previously emphasized in this chapter, the sqrt-LASSO finds its highest significance for noise-blind high-dimensional statistics or compressive sensing tasks. These applications aim to recover a sparse vector and perform variable selection without requiring prior noise level knowledge. Hence, in the next section, our goal is to establish the principal theorem of this chapter, which is the linear convergence rate of IRLS for the sqrt-LASSO under a natural assumption prevalent in compressive sensing.

4.7 Convergence under the Null Space Property

In the field of sparse regression, it is common to assume that the design matrix $A \in \mathbb{R}^{n\times p}$ is, in some sense, well-conditioned on the set of sparse vectors or that the kernel of such matrices has a nice geometry. A general way to do so is via the compatibility condition [VDGB09], which is the sharpest condition to obtain oracle inequalities for estimation and prediction.

Definition 4.7.1. *A matrix $A \in \mathbb{R}^{n\times p}$ is said to satisfy the* (L,S)-compatibility condition *if there exists $L \in (1,\infty)$ such that for the set*

$$\Delta_{L,S} := \left\{v \in \mathbb{R}^N : \|v_{S^c}\|_1 \le L\|v_S\|_1 \text{ and } \|v_S\|_1 \neq 0\right\},$$

the condition $\inf_{v \in \Delta_{L,S}} \frac{S\|Av\|_2^2}{\|v_S\|_1^2} > 0$ *holds true.*

It turns out that the compatibility condition is equivalent to the so-called robust null space property more commonly used in the signal processing literature, e.g., [PJ21][Theorem XX]. This property extends the necessary and sufficient property used for the analysis of equality constrained ℓ_1-minimization to include robustness with respect to noise. Before we start the convergence analysis, we state a few facts connected to the NSP. The first one is an equivalent formulation, which is more suitable for our analysis.

Lemma 4.7.2. *[FR13, Lemma 4.20] The matrix* $A \in \mathbb{C}^{m \times N}$ *satisfies the robust null space property with constants* $0 < \rho < 1$ *and* $\tau > 0$ *if and only if for any set* $S \subset [N]$ *of cardinality* $|S| \leq s$ *we have*

$$\|z - x\|_1 \leq \frac{1+\rho}{1-\rho}(\|z\|_1 - \|x\|_1 + 2\|x_{S^c}\|_1) + \frac{2\tau}{1-\rho}\|A(z-x)\|_2 \tag{4.40}$$

for all vectors $x, z \in \mathbb{C}^N$.

The first benefit of NSP, when used in the convergence analysis, is that it is possible to track the distance to the ground truth signal x in terms of the function value gap.

Lemma 4.7.3 (Error bound in terms of function value gap)**.** *Let* $A \in \mathbb{C}^{m \times N}$ *admit the robust null space property with constants* $0 < \rho < 1$ *and* $\tau > 0$ *of order* s. *If* $\lambda \leq \frac{(1+\rho)}{2\tau}$, *then for all* $z \in \mathbb{R}^n$ *and all* $\varepsilon, \varepsilon_0 \geq 0$ *we have*

$$\|z - x_*\|_1 \leq \frac{2(1+\rho)}{(1-\rho)\lambda}\left[\lambda\sigma_s(x_*)_{\ell_1} + \|\epsilon\|_2 + \frac{1}{2}(f_0(z) - f_0(x_*))\right],$$

where x_* *is the ground truth signal that gives origin to the data* $y = Ax_* + \epsilon$.

Proof. By Lemma 4.7.2 and the choice of λ, we get

$$\begin{aligned}\|z - x_*\|_1 &\leq \frac{1+\rho}{1-\rho}(\|z\|_1 - \|x\|_1 + 2\sigma_s(x)_{\ell_1}) + \frac{2\tau}{1-\rho}\|A(z-x_*)\|_2 \\ &\leq \frac{1+\rho}{(1-\rho)\lambda}\left[\lambda(\|z\|_1 - \|x_*\|_1 + 2\sigma_s(x)_{\ell_1}) + \|A(z-x_*)\|_2\right]\end{aligned}$$

Since x_* is the true signal, we have $Ax_* = y - \epsilon$ and $\|\epsilon\|_2 = \|Ax_* - y\|_2$. Hence, triangle inequality gives

$$\begin{aligned}\|z - x_*\|_1 &\leq \frac{1+\rho}{(1-\rho)\lambda}\left[2\lambda\sigma_s(x_*)_{\ell_1} + \lambda\|z\|_1 + \|Az - y\|_2 + \|\epsilon\|_2 - \lambda\|x_*\|_1\right] \\ &= \frac{1+\rho}{(1-\rho)\lambda}\left[2\lambda\sigma_s(x_*)_{\ell_1} + 2\|\epsilon\|_2 + f_0(z) - f_0(x_*)\right].\end{aligned}$$

□

For the iterates of IRLS, a consequence of Lemma 4.7.3 is the following statement:

Corollary 4.7.4. *Under assumptions of Lemma 4.7.3, the iterates x^k generated by Algorithm 4 admit*

$$\|x^k - x_*\|_1 \leq \frac{2(1+\rho)}{(1-\rho)\lambda}\left[\lambda\sigma_s(x_*)_{\ell_1} + \|\epsilon\|_2 + \frac{1}{2}(f_{\varepsilon_k}(x^k) - f_0(x_0^*))\right].$$

Proof. It follows Lemma 4.7.3 combined with inequalities $f_0(_*) \geq f_0(x_0^*)$ and $f_0(x^k) \leq f_{\varepsilon_k}(x^k)$. □

Consequently, the minimization of the objective function (4.5) implies that the distance between the iterates and the ground truth is being minimized. Furthermore, the bounds established in Section 4.4 and Section 4.5 can be combined with Corollary 4.7.4 to derive the convergence results in terms of the distance to the ground truth. A similar inequality can be derived for the minimizer of the sqrt-LASSO, Equation (4.5).

Corollary 4.7.5. *Under assumptions of Lemma 4.7.3 solution x_0^* of Sqrt-LASSO admits*

$$\|x_0^* - x_*\|_1 \leq \frac{2(1+\rho)}{(1-\rho)\lambda}\left[\lambda\sigma_s(x_*)_{\ell_1} + \|\epsilon\|_2\right]$$

and

$$\|Ax_0^* - b\|_2 \leq \left[1 + \frac{2(1+\rho)}{(1-\rho)}\right]\|\epsilon\|_2 + \frac{2(1+\rho)\lambda}{(1-\rho)}\sigma_s(x_*)_{\ell_1}.$$

Proof. The first inequality follows from Lemma 4.7.3 with $z = x_0^*$ and $\varepsilon = \varepsilon_0 = 0$. For the second inequality, we use the optimality of x_0^*,

$$\|Ax_0^* - b\|_2 + \lambda\|x_0^*\|_1 \leq \|Ax_* - b\|_2 + \lambda\|x_*\|_1 = \|\epsilon\|_2 + \lambda\|x\|_1. \tag{4.41}$$

By bringing $\|x_0^*\|_1$ to the right-hand side and applying the first inequality, we get

$$\begin{aligned}\|Ax_0^* - b\|_2 &\leq \|\epsilon\|_2 + \lambda(\|x_*\|_1 - \|x_0^*\|_1) \leq \|\epsilon\|_2 + \lambda\|x_* - x_0^*\|_1\\ &\leq \|\epsilon\|_2 + \frac{2(1+\rho)}{1-\rho}\left[\lambda\sigma_s(x_*)_{\ell_1} + \|\epsilon\|_2\right]\\ &= \left[1 + \frac{2(1+\rho)}{1-\rho}\right]\|\epsilon\|_2 + \frac{2(1+\rho)\lambda}{1-\rho}\sigma_s(x_*)_{\ell_1}.\end{aligned}$$

□

The first bound is the bound Theorem 4.1.5 [PJ21, Theorem 3.1] that was proven here for the sake of completeness[5]. It implies that if the noise is absent and x is sparse, Sqrt-LASSO recovers x uniquely. The second bound is rather a technical result, which will be useful later in this section. We can now proceed to formally state the principal theorem.

Theorem 4.7.6. *Let A satisfies NSP with constants $0 < \rho < \frac{1}{4}$ and $\tau > 0$ and assume that $y = Ax_* + \epsilon$. Consider the sequence $\varepsilon_k = \min\left\{\varepsilon_{k-1}, \frac{\|Ax^k - b\|_2 + \lambda\sigma_s(x^k)_{\ell_1}}{\lambda(N+1)}\right\}$ and $\varepsilon_{k,0} = \lambda\varepsilon_k$ for $k \geq 0$ and $\varepsilon_{-1} = +\infty$. Then, for $k \leq \hat{k} := \min\left\{k \in \mathbb{N} : f_{\varepsilon_k}(x^k) - f_0(x_0^*) > 3\lambda(N+1)\varepsilon_k/4\right\}$ it holds that the following is true for the iterates x^k of Algorithm 4 with $\lambda \leq \rho/\tau$:*

$$f_{\varepsilon_{k+1}}(x^{k+1}) - f_0(x_0^*) \leq \left[1 - \frac{(1-\rho)^4}{96(1+\rho)^2(2+\rho)^2(N+1)}\right]\left[f_{\varepsilon_k}(x^k) - f_0(x_0^*)\right].$$

Moreover, for $0 \leq f_{\varepsilon_k}(x^k) - f_0(x_0^*) \leq 3\lambda(N+1)\varepsilon_k/4$, it holds that

$$\|x^k - x_*\|_1 \leq \frac{2(1+\rho)}{(1-\rho)\lambda}\left[1 + \frac{3(1+\rho)}{2-8\rho}\right]\left[\lambda\sigma_s(x_*)_{\ell_1} + \|\epsilon\|_2\right].$$

The proof of Theorem 4.7.6 is based on Lemma 4.4.1 with $x = x_0^*$, $v^k = x_0^* - x^k$ and $\tilde{v}^k = (0, v^k) = \tilde{x}_0^* - \tilde{x}^k$ and follows the same lines of the proof of Proposition 2.3.7. Let us denote by S the support of the k largest entries of x^k in absolute value. It consists of two complementary parts, similar to what happens for Basis Pursuit, see Theorem 2.3.9. In the first part, we will establish that outside a certain region, i.e., when $f_{\varepsilon_k}(x^k) - f_0(x_0^*) > C\varepsilon_k$ for a certain $C > 0$, we obtain a linear decay on the function value. Then, in the second part, we will prove that when we reach the basin of attraction, we can also conclude that the iterates x^k are already close enough to the ground truth.
The first part consists of three main steps:

i Bound the first order term $-|\langle\nabla\tilde{f}_{\varepsilon_k}(\tilde{x}^k), \tilde{x}_0^* - \tilde{x}^k\rangle|$ from below.

ii Bound the second order term $\langle W_\varepsilon(x^k)\tilde{v}^k, \tilde{v}^k\rangle$ from above.

iii Bound the function value gap $f_{\varepsilon_k}(x^{k+1}) - f_{\varepsilon_k}(x^k)$ is bounded by ε_k.

iv To finish, the convergence rate will finally be obtained by using a certain choice of ε together with the three bounds above.

Proof. The proof will be divided into two parts. First, assume that $f_{\varepsilon_k}(x^k) - f_0(x_0^*) > 3\lambda(N+1)\varepsilon_k/4$ holds. In this case, the idea of the proof is based on Theorem 4.4.5 and Theorem 4.5.2.

[5] At the beginning of the chapter, we stated this theorem with the condition $\lambda \geq \frac{2}{1+\rho}\tau$ as it is originally stated. Here, for our convenience, we performed the change of variable $\lambda \mapsto \frac{1}{\lambda}$ and stated it under the assumption $\lambda \leq \frac{(1+\rho)}{2\tau}$.

Part I: Bounding the linear term: By assuming that $\varepsilon_{k,0} = \lambda\varepsilon_k$, the first-order term can be rewritten as

$$\langle \nabla \tilde{f}_{\varepsilon_k}(x^k), \tilde{v}^k \rangle = \frac{\langle \tilde{v}^k, \tilde{A}^T \tilde{A}\tilde{x}^k \rangle}{\max\{\|\tilde{A}\tilde{x}^k\|_2, \lambda\varepsilon_k\}} + \lambda \sum_{i=1}^{N} \frac{v_i^k x_i^k}{\max\{|x_i^k|, \varepsilon_k\}} = \lambda \sum_{i=0}^{N} \frac{\langle M_i\tilde{v}^k, M_i\tilde{x}^k \rangle}{\max\{\|M_i\tilde{x}^k\|_2, \varepsilon_k\}},$$

with $M_0 = \lambda^{-1}\tilde{A}$ and $M_i = E^{i,i}$ for $i = 1, \dots, N$, where $E^{i,i}$ is a matrix with a single non-zero entry $E_{i,i}^{i,i} = 1$. For a single summand, we have

$$\begin{aligned}
\frac{\langle M_i\tilde{v}^k, M_i\tilde{x}^k \rangle}{\max\{\|M_i\tilde{x}^k\|_2, \varepsilon_k\}} &= \frac{\langle M_i\tilde{x}_0^*, M_i\tilde{x}^k \rangle}{\max\{\|M_i\tilde{x}^k\|_2, \varepsilon_k\}} - \frac{\|M_i\tilde{x}^k\|_2^2}{\max\{\|M_i\tilde{x}^k\|_2, \varepsilon_k\}} \\
&\leq \frac{\|M_i\tilde{x}_0^*\|_2 \|M_i\tilde{x}^k\|_2}{\max\{\|M_i\tilde{x}^k\|_2, \varepsilon_k\}} - \frac{\|M_i\tilde{x}^k\|_2^2}{\max\{\|M_i\tilde{x}^k\|_2, \varepsilon_k\}} \\
&\leq \|M_i\tilde{x}_0^*\|_2 - \frac{\|M_i\tilde{x}^k\|_2^2}{\max\{\|M_i\tilde{x}^k\|_2, \varepsilon_k\}}.
\end{aligned}$$

If $\|M_i\tilde{x}^k\|_2 \geq \varepsilon_k$, then

$$\frac{\langle M_i\tilde{v}^k, M_i\tilde{x}^k \rangle}{\max\{\|M_i\tilde{x}^k\|_2, \varepsilon_k\}} \leq \|M_i\tilde{x}_0^*\|_2 - \|M_i\tilde{x}^k\|_2 = \|M_i\tilde{x}_0^*\|_2 - j_{\varepsilon_k}(\|M_i\tilde{x}^k\|_2).$$

Otherwise, if $\|M_i\tilde{x}^k\|_2 < \varepsilon_k$, we have

$$\begin{aligned}
\frac{\langle M_i\tilde{v}^k, M_i\tilde{x}^k \rangle}{\max\{\|M_i\tilde{x}^k\|_2, \varepsilon_k\}} &\leq \|M_i\tilde{\tilde{x}}_0^*\|_2 - \frac{\|M_i\tilde{x}^k\|_2^2}{\varepsilon_k} \\
&= \|M_i\tilde{x}_0^*\|_2 - \frac{\|M_i\tilde{x}^k\|_2^2}{2\varepsilon_k} - \frac{\varepsilon_k}{2} + \frac{\varepsilon_k}{2} = \|M_i\tilde{x}_0^*\|_2 - j_{\varepsilon_k}(\|M_i\tilde{x}^k\|_2) + \frac{\varepsilon_k}{2}.
\end{aligned}$$

Thus, in any case, the latter bound applies since $\frac{\varepsilon_k}{2} > 0$. Hence, the first-order term can be bounded by

$$\begin{aligned}
\langle \nabla \tilde{f}_{\varepsilon_k}(x^k), \tilde{v}^k \rangle &\leq \lambda \sum_{i=0}^{N} \left[\|M_i\tilde{x}_0^*\|_2 - j_{\varepsilon_k}(\|M_i\tilde{x}^k\|_2) + \tfrac{\varepsilon_k}{2}\right] \\
&= \|\tilde{A}\tilde{x}_0^*\|_2 + \lambda\|x_0^*\|_1 - j_{\varepsilon_{k,0}}(\|\tilde{A}\tilde{x}^k\|_2) + \lambda \sum_{i=1}^{N} j_{\varepsilon_k}(x_j^k) + \tfrac{1}{2}\lambda(N+1)\varepsilon_k \\
&= f_0(x_0^*) - f_{\varepsilon_k}(x^k) + \tfrac{1}{2}\lambda(N+1)\varepsilon_k.
\end{aligned}$$

Now, by using the hypothesis $f_{\varepsilon_k}(x^k) - f_0(x_0^*) > \frac{3}{4}\lambda(N+1)\varepsilon_k$, it follows that

$$\langle \nabla \tilde{f}_{\varepsilon_k}(x^k), \tilde{v}^k \rangle \leq f_0(x_0^*) - f_{\varepsilon_k}(x^k) + \tfrac{1}{2}\lambda(N+1)\varepsilon_k \leq -\tfrac{1}{4}\lambda(N+1)\varepsilon_k \leq 0,$$

and

$$\begin{aligned}\langle \nabla \tilde{f}_{\varepsilon_k}(x^k), \tilde{v}^k \rangle &\leq f_0(x_0^*) - f_{\varepsilon_k}(x^k) + \frac{1}{2}\lambda(N+1)\varepsilon_k \\ &\leq f_0(x_0^*) - f_{\varepsilon_k}(x^k) + \frac{2}{3}[f_{\varepsilon_k}(x^k) - f_0(x_0^*)] \leq \tfrac{1}{3}[f_0(x_0^*) - f_{\varepsilon_k}(x^k)] \leq 0.\end{aligned}$$

Hence, for $x^k \neq x_0^*$ we apply Lemma 4.4.1, which gives

$$f_{\varepsilon_k}(x^{k+1}) - f_{\varepsilon_k}(x^k) \leq -\frac{|\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{x}^k), \tilde{v}^k \rangle|^2}{2\langle W_{\varepsilon_k}(x^k)\tilde{v}^k, \tilde{v}^k \rangle} \leq -\frac{\lambda(N+1)\varepsilon_k(f_{\varepsilon_k}(x^k) - f_0(x_0^*))}{24\langle W_{\varepsilon_k}(x^k)v^k, v^k \rangle}. \tag{4.42}$$

Now, we must bound the denominator $\langle W_{\varepsilon_k}(x^k)v^k, v^k \rangle$.

Part II: Bounding the quadratic term: From the definition of W_ε, we obtain

$$\begin{aligned}\langle W_\varepsilon(x^k)\tilde{v}^k, \tilde{v}^k \rangle &= \frac{\|\tilde{A}\tilde{v}^k\|_2^2}{\max\{\|\tilde{A}\tilde{x}^k\|_2, \lambda\varepsilon_k\}} + \lambda \sum_{j=1}^N \frac{|v_j^k|^2}{\max\{|x_j^k|, \varepsilon_k\}} \\ &\leq \frac{\|Av^k\|_2^2}{\lambda\varepsilon_k} + \frac{\lambda^2}{\lambda\varepsilon_k}\|v^k\|_2^2 = \frac{1}{\lambda\varepsilon_k}\left[\|Av^k\|_2^2 + \lambda^2\|v^k\|_2^2\right] \\ &\leq \frac{1}{\lambda\varepsilon_k}\left[\|Av^k\|_2 + \lambda\|v^k\|_2\right]^2 \leq \frac{1}{\lambda\varepsilon_k}\left[\|A(x_0^* - x^k)\|_2 + \lambda\|x_0^* - x^k\|_1\right]^2.\end{aligned}$$

Next, we bound the term $\|A(x_0^* - x^k)\|_2 + \lambda\|x_0^* - x^k\|_1$ by using the NSP. Lemma 4.7.2 yields

$$\begin{aligned}&\|A(x_0^* - x^k)\|_2 + \lambda\|x_0^* - x^k\|_1 \\ &\leq \frac{\lambda(1+\rho)}{1-\rho}\left[\|x_0^*\|_1 - \|x^k\|_1 + 2\sigma_s(x^k)_{\ell_1}\right] + \left[1 + \frac{2\tau\lambda}{1-\rho}\right]\|A(x_0^* - x^k)\|_2 \\ &\leq \frac{\lambda(1+\rho)}{1-\rho}\left[\|x_0^*\|_1 - \|x^k\|_1 + 2\sigma_s(x^k)_{\ell_1}\right] + \left[1 + \frac{2\tau\lambda}{1-\rho}\right](\|Ax_0^* - b\|_2 + \|Ax^k - b\|_2)\end{aligned}$$

Since x_0^* is the minimizer of f_0, we have

$$\|Ax_0^* - b\|_2 + \lambda\|x_0^*\|_1 \leq \|Ax^k - b\|_2 + \lambda\|x^k\|_1,$$

which is equivalent to

$$\lambda(\|x_0^*\|_1 - \|x^k\|_1) \leq \|Ax^k - b\|_2 - \|Ax_0^* - b\|_2.$$

Thus, we get

$$\|A(x_0^*-x^k)\|_2+\lambda\|x_0^*-x^k\|_1 \le \frac{2\lambda(1+\rho)}{1-\rho}\sigma_s(x^k)_{\ell_1} \\ +\left[1+\frac{2\tau\lambda}{1-\rho}+\frac{1+\rho}{1-\rho}\right]\|Ax^k-b\|_2+\left[1+\frac{2\tau\lambda}{1-\rho}-\frac{1+\rho}{1-\rho}\right]\|Ax_0^*-b\|_2.$$

By assumption $\lambda \le \rho/\tau$, which implies that

$$\left[1+\frac{2\tau\lambda}{1-\rho}-\frac{1+\rho}{1-\rho}\right]\le 0 \quad \text{and} \quad \left[1+\frac{2\tau\lambda}{1-\rho}+\frac{1+\rho}{1-\rho}\right]\le \frac{2(1+\rho)}{1-\rho}.$$

Therefore, we obtain

$$\|A(x_0^*-x^k)\|_2+\lambda\|x_0^*-x^k\|_1 \le \frac{2(1+\rho)}{1-\rho}\left[\|Ax^k-b\|_2+\lambda\sigma_s(x^k)_{\ell_1}\right],$$

and

$$\langle W_\varepsilon(x^k)\tilde{v}^k,\tilde{v}^k\rangle \le \frac{4(1+\rho)^2}{\lambda\varepsilon_k(1-\rho)^2}\left[\|Ax^k-b\|_2+\lambda\sigma_s(x^k)_{\ell_1}\right]^2. \tag{4.43}$$

The remaining major step of this part is to bound $\|Ax^k-b\|_2+\lambda\sigma_s(x^k)_{\ell_1}$ in terms of ε_k. Recall that, by hypothesis, $\varepsilon_k=\min\left\{\varepsilon_{k-1},\frac{\|Ax^k-b\|_2+\lambda\sigma_s(x^k)_{\ell_1}}{\lambda(N+1)}\right\}$. If the minimum is attained by the second term, the bound is trivial since

$$\|Ax^k-b\|_2+\lambda\sigma_s(x^k)_{\ell_1}=\lambda(N+1)\varepsilon_k \le \frac{2+\rho}{(1-\rho)}\lambda(N+1)\varepsilon_k.$$

Otherwise, there exists index $j<k$, such that

$$\varepsilon_k=\varepsilon_j=\frac{\|Ax^j-b\|_2+\lambda\sigma_s(x^j)_{\ell_1}}{\lambda(N+1)}.$$

By Lemma 4.2.1 and by the construction of the iterates, we have

$$f_0(x^k)\le f_{\varepsilon_k}(x^k)\le f_{\varepsilon_j}(x^j)\le f_0(x^j)+\lambda(N+1)\varepsilon_j.$$

Expanding both the right and left parts leads to

$$\|Ax^k-b\|_2+\lambda\|x^k\|_1 \le 2\|Ax^j-b\|_2+\lambda\|x^j\|_1+\lambda\sigma_s(x^j)_{\ell_1}$$

Let us denote by S_j the set of indices corresponding to the best-s term approximation of x^j. That is, we have $\|x^j_{S_j}\|_1=\sigma_s(x^j)_{\ell_1}$ and $\|x^k_{S_j}\|_1\ge\sigma_s(x^k)_{\ell_1}$. Thus, by splitting the norms $\|x^t\|_1 = \|x^t_{S_j}\|_1 + \|x^t_{S_j^c}\|_1$, for $t = j, k$, we arrive at

$$\|Ax^k - b\|_2 + \lambda\|x^k_{S_j}\|_1 + \lambda\|x^k_{S_j^c}\|_1 \leq (1+c)\|Ax^j - b\|_2 + \lambda\|x^j_{S_j}\|_1 + (1+c)\lambda\sigma_s(x^j)_{\ell_1}.$$

Hence, bringing $\lambda\|x^k_{S_j}\|_1$ to the right-hand side, yields

$$\begin{aligned}\|Ax^k - b\|_2 + \lambda\|x^k_{S_j^c}\|_1 &\leq 2\left[\|Ax^j - b\|_2 + \lambda\sigma_s(x^j)_{\ell_1}\right] + \lambda[\|x^j_{S_j}\|_1 - \|x^k_{S_j}\|_1] \\ &\leq 2\left[\|Ax^j - b\|_2 + \lambda\sigma_s(x^j)_{\ell_1}\right] + \lambda\|(x^j - x^k)_{S_j}\|_1. \qquad (4.44)\end{aligned}$$

Moreover, the definition of NSP gives

$$\begin{aligned}\|(x^j - x^k)_{S_j}\|_1 &\leq \rho\|(x^j - x^k)_{S_j^c}\|_1 + \tau\|A(x^j - x^k)\|_2 \\ &\leq \rho\|x^j_{S_j^c}\|_1 + \rho\|x^k_{S_j^c}\|_1 + \tau\|Ax^j - b\|_2 + \tau\|Ax^k - b\|_2 \\ &\leq \rho\sigma_s(x^j)_{\ell_1} + \rho\|x^k_{S_j^c}\|_1 + \tfrac{\rho}{\lambda}\|Ax^j - b\|_2 + \tfrac{\rho}{\lambda}\|Ax^k - b\|_2.\end{aligned}$$

Incorporating this bound in Equation (4.44) leads to

$$\|Ax^k - b\|_2 + \lambda\|x^k_{S_j^c}\|_1 \leq (2+\rho)\left[\|Ax^j - b\|_2 + \lambda\sigma_s(x^j)_{\ell_1}\right] + \rho\|Ax^k - b\|_2 + \lambda\|x^k_{S_j^c}\|_1,$$

which, in turn, is equivalent to

$$\|Ax^k - b\|_2 + \lambda\|x^k_{S_j^c}\|_1 \leq \frac{2+\rho}{1-\rho}\left[\|Ax^j - b\|_2 + \lambda\sigma_s(x^j)_{\ell_1}\right].$$

Since $\|x^k_{S_j}\|_1 \geq \sigma_s(x^k)_{\ell_1}$, we obtain

$$\|Ax^k - b\|_2 + \lambda\sigma_s(x^k)_{\ell_1} \leq \frac{2+\rho}{1-\rho}\left[\|Ax^j - b\|_2 + \lambda\sigma_s(x^j)_{\ell_1}\right] = \frac{2+\rho}{(1-\rho)}\lambda(N+1)\varepsilon_k.$$

Returning to the bound for the quadratic term (4.43), this gives

$$\langle W_\varepsilon(x^k)\tilde{v}^k, \tilde{v}^k\rangle \leq \frac{4(1+\rho)^2(2+\rho)^2\lambda(N+1)^2\varepsilon_k}{(1-\rho)^4}. \qquad (4.45)$$

Adding the pieces together: Now, the bound for the quadratic term can be combined with the inequality (4.42). This finally gives

$$f_{\varepsilon_k}(x^{k+1}) - f_{\varepsilon_k}(x^k) \leq -\frac{\lambda(N+1)\varepsilon_k(1-\rho)^4}{96(1+\rho)^2(2+\rho)^2\lambda(N+1)^2\varepsilon_k}\left[f_{\varepsilon_k}(x^k) - f_0(x^*_0)\right].$$

As for the last step, we add and subtract $f_0(x_0^*)$ and rearrange all the terms, which gives

$$\begin{aligned} f_{\varepsilon_{k+1}}(x^{k+1}) - f_0(x_0^*) &\le f_{\varepsilon_k}(x^{k+1}) - f_0(x_0^*) \\ &\le \left[1 - \frac{(1-\rho)^4}{96(1+\rho)^2(2+\rho)^2(N+1)}\right] \left[f_{\varepsilon_k}(x^k) - f_0(x_0^*)\right]. \end{aligned}$$

Now, we assume the opposite inequality, i.e., we assume that $0 \le f_{\varepsilon_k}(x^k) - f_0(x_0^*) \le 3\lambda(N+1)\varepsilon_k/4$ holds. Since $\lambda \le \rho/\tau = 2\rho/2\tau \le (1+\rho)/2\tau$, by Corollary 4.7.4, we have

$$\begin{aligned} \|x^k - x\|_1 &\le \frac{2(1+\rho)}{(1-\rho)\lambda} \left[\lambda\sigma_s(x)_{\ell_1} + \|\epsilon\|_2 + \frac{1}{2}(f_{\varepsilon_k}(x^k) - f_0(x_0^*))\right] \\ &\le \frac{2(1+\rho)}{(1-\rho)\lambda} \left[\lambda\sigma_s(x)_{\ell_1} + \|\epsilon\|_2 + \tfrac{3}{8}\lambda(N+1)\varepsilon_k\right]. \end{aligned} \tag{4.46}$$

To finish the proof, we will establish a bound of the form

$$\|Ax^k - b\|_2 + \lambda\sigma_s(x^k)_{\ell_1} \le c_1[\|\epsilon\|_2 + \lambda\sigma_s(x)_{\ell_1}],$$

for a given $c_1 \ge 0$. Since x_0^* is the minimizer of f_0, we have

$$f_0(x^k) \le f_{\varepsilon_k}(x^k) \le f_0(x_0^*) + \tfrac{3}{4}\lambda(N+1)\varepsilon_k \le f_0(x) + \tfrac{3}{4}[\|Ax^k - b\|_2 + \lambda\sigma_s(x^k)_{\ell_1}].$$

Expanding f_0 and rearranging terms gives

$$(1 - \tfrac{3c}{4})\|Ax^k - b\|_2 + \lambda\|x^k\|_1 - \tfrac{3\lambda}{4}\sigma_s(x^k)_{\ell_1} \le \|Ax - b\|_2 + \lambda\|x\|_1 = \|\epsilon\|_2 + \lambda\|x\|_1.$$

In a similar way to what was done above, let us denote by S the set of indices corresponding to the best-s term approximation of x. That is, we have $\|x_S\|_1 = \sigma_s(x)_{\ell_1}$ and $\|x_S^k\|_1 \ge \sigma_s(x^k)_{\ell_1}$. Thus, by splitting the norms $\|v\|_1 = \|v_S\|_1 + \|v_{S^c}\|_1$ for $v = x^k$ and $v = x$, we arrive at

$$(1 - \tfrac{3}{4})\|Ax^k - b\|_2 + \lambda\|x_S^k\|_1 + \lambda\|x_{S^c}^k\|_1 - \tfrac{3\lambda}{4}\sigma_s(x^k)_{\ell_1} \le \|\epsilon\|_2 + \lambda\|x_{S_*}\|_1 + \lambda\sigma_s(x)_{\ell_1}.$$

This, together with reverse triangle inequality, yields

$$\begin{aligned} (1 - \tfrac{3}{4})\|Ax^k - b\|_2 + \lambda\|x_{S^c}^k\|_1 - \tfrac{3\lambda}{4}\sigma_s(x^k)_{\ell_1} &\le \|\epsilon\|_2 + \lambda[\|x_S\|_1 - \|x_S^k\|_1] + \lambda\sigma_s(x)_{\ell_1} \\ &\le \|\epsilon\|_2 + \lambda[\|(x - x^k)_S\|_1] + \lambda\sigma_s(x)_{\ell_1} \end{aligned}$$

By the definition of NSP, we have

$$\begin{aligned}\|(x-x^k)_S\|_1 &\leq \rho\|(x-x^k)_{S^c}\|_1 + \tau\|A(x-x^k)\|_2\\ &\leq \rho\|x_{S^c}\|_1 + \rho\|x_{S^c}\|_1 + \tau\|Ax-b\|_2 + \tau\|Ax^k-b\|_2\\ &\leq \rho\sigma_s(x)_{\ell_1} + \rho\|x^k_{S^c}\|_1 + \tfrac{\rho}{\lambda}\|\epsilon\|_2 + \tfrac{\rho}{\lambda}\|Ax^k-b\|_2.\end{aligned}$$

Thus,

$$(1-\tfrac{3}{4}-\rho)\|Ax^k-b\|_2 + (1-\rho)\lambda\|x^k_{S^c}\|_1 - \tfrac{3\lambda}{4}\sigma_s(x^k)_{\ell_1} \leq (1+\rho)[\|\epsilon\|_2 + \lambda\sigma_s(x)_{\ell_1}]$$

Since, by assumption, $1-\frac{3}{4}-\rho>0$ and $\|x^k_S\|_1 \geq \sigma_s(x^k)_{\ell_1}$, we get

$$\begin{aligned}\|Ax^k-b\|_2 + \lambda\sigma_s(x^k)_{\ell_1} &\leq \|Ax^k-b\|_2 + \frac{1-\rho}{1-\frac{3}{4}-\rho}\lambda\|x^k_{S^c}\|_1 - \frac{\frac{3\lambda}{4}}{1-\frac{3}{4}-\rho}\sigma_s(x^k)_{\ell_1}\\ &\leq \frac{1+\rho}{1-\frac{3}{4}-\rho}[\|\epsilon\|_2 + \lambda\sigma_s(x)_{\ell_1}].\end{aligned}$$

Thus, $\lambda(N+1)\varepsilon_k$ is bounded from above as

$$\lambda(N+1)\varepsilon_k \leq [\|Ax^k-b\|_2 + \lambda\sigma_s(x^k)_{\ell_1}] \leq \frac{(1+\rho)}{1-\frac{3}{4}-\rho}[\|\epsilon\|_2 + \lambda\sigma_s(x)_{\ell_1}].$$

By applying this bound to (4.46), we finally conclude that

$$\|x^k-x\|_1 \leq \frac{2(1+\rho)}{(1-\rho)\lambda}\left[1+\frac{3(1+\rho)}{2-8\rho}\right][\lambda\sigma_s(x)_{\ell_1} + \|\epsilon\|_2].$$

□

Remark 4.7.7. *Inspired null space constant $\rho < 1/8$ and the constant $\frac{3}{4}$ in $f_{\varepsilon_k}(x^k) - f_0(x_0^*) > 3\lambda(N+1)\varepsilon_k/4$ are not optimized. In fact, inspired by [ABH19], one could choose the smoothing parameter ε_k as $\varepsilon_k = \min\left\{\varepsilon_{k-1}, c\frac{\|Ax^k-b\|_2+\lambda\sigma_s(x^k)_{\ell_1}}{\lambda(N+1)}\right\}$ for a certain constant $0<c<2$. This constant would appear in the definition of the null space constant in the form $0<\rho<1-\frac{3c}{4}$ and in the hypothesis for the "basis of attraction" that would be given by $f_{\varepsilon_k}(x^k) - f_0(x_0^*) > \min\{1, c^{-1}\}\lambda(N+1)\varepsilon_k$. Hence, in this case, the convergence results read as*

$$f_{\varepsilon_{k+1}}(x^{k+1}) - f_0(x_0^*) \leq \left[1 - \frac{c^2(1-\rho)^4}{96(1+\rho)^2(1+c+\rho)^2(N+1)}\right]\left[f_{\varepsilon_k}(x^k) - f_0(x_0^*)\right],$$

and

$$||x^k-x||_1 \leq \frac{2(1+\rho)}{(1-\rho)\lambda}\left[1+\frac{3c(1+\rho)}{8-6c-8\rho}\right][\lambda\sigma_s(x)_{\ell_1} + ||\eta||_2].$$

4.8 Chapter Conclusion

In this chapter, which represents an ongoing work, we have introduced the third use of the method of least squares method in this thesis. Specifically, we have delved into developing an algorithm to address non-smooth noise-blind problems. Furthermore, we have extensively discussed various types of convergence results. Our next step involves comparing the IRLS approach devised in this chapter with alternative strategies, such as proximal gradient methods and the semismooth Newton method.

Furthermore, as discussed in Section 4.1, it is worth noting that this particular objective function can also be viewed from the perspective of robust optimization, as highlighted in recent works [BKZ+21, CTZ22]. We believe that exploring the implications of this IRLS strategy in the realm of robust optimization holds promising potential for future research.

Chapter 5

Overparametrization and the solution of convex problems

> But in high dimension, there is no such thing as interpolation.
> In high dimension, everything is extrapolation.
> *Yann LeCun* [twi21]

In this chapter, we discuss how certain very recent ideas related to *benign overfitting* and *overparametrization* in machine learning can be used to develop an algorithm that solves constrained least squares problems in an unconstrained way. The algorithm is highly scalable as it is based purely on gradient descent. We present its convergence theory based on continuous arguments for a certain gradient flow and illustrate its power with numerical experiments. The work presented in this chapter was developed in collaboration with Dr. Hung-Hsu Chou, Dr. Johannes Maly and Dr. Heudson Mirandola. The first version of our manuscript, *"Non-negative Least Squares via Overparametrization"*, is currently in preparation for journal submission [CMV22]. Although this chapter closely follows the article, several parts were edited to improve the text's clarity and the math proofs.

5.1 Introduction

Until now, this thesis has primarily focused on developing a simple least squares-based algorithm for tackling various tasks, along with establishing the corresponding convergence theory. However, we now embark on a different path and delve into analyzing a *constrained* least squares problem known as the non-negative least squares problem (NNLS). In doing so, we delve into one of the central questions in theoretical machine learning: why do models with a large number of parameters, such as deep neural networks, exhibit remarkable performance beyond the training data? In connection with this question, it is crucial to understand another pertinent and related question in this field: why does (stochastic) gradient descent converge towards a desirable overparameterized model with nice generalization properties when employed to train such models?

5.2 The importance of overparametrization

> Why don't heavily parameterized neural networks overfit the data?
> *Leo Breiman* [Bre18]

One of the central points of learning theory is how a model with a certain number of parameters can explain a given dataset, i.e., in some sense, how rich the model is. The standard (and now classical) learning theory says that there should be a balance between the abundance of features and details to capture the underlying structure of the data and parsimony to avoid overfitting [Vap99]. The well-studied *bias-variance tradeoff* phenomenon, see [HTFF09, Chapter 7], quantifies this phenomenon and advocates for a balance that does not highly underfit the training data but, at the same time, does not interpolate it, causing overfitting.

However, the recent success of large-scale machine learning models [GPAM$^+$14] in many tasks, including challenging humans, e.g., playing the game Go [SHM$^+$16] or predicting protein structures [SEJ$^+$20], has raised questions regarding such trade-off explanation and traditional statistical learning theory. The training of deep neural networks, in particular, has shown the ability to find parameter configurations that lead to models with highly effective generalization properties on unseen data despite the potential for overfitting. This challenges the conventional notion from learning theory that models should not excessively fit the training data to avoid poor performance on future data. In fact, modern machine learning methods have demonstrated the opposite phenomenon, where remarkable prediction rules can be achieved alongside a strong fit to the training data. This contradicts the belief expressed in the following quote from one of the most important books in the field: *"This means that we cannot use residual sum-of-squares on the training data to determine these parameters as well, since we would always pick those that gave interpolating fits and hence zero residuals. Such a model is unlikely to predict future data well at all."*[HTFF09, Chapter 2].

It was observed that very large models could go beyond the interpolation threshold and perform well on unseen data, i.e., if we progressively increase the size of the model, the so-called test risk, which is the error on data that does not belong to the training set, will also go to zero if the number of parameters is large enough [BHMM19].

The development of state-of-the-art models in various machine learning domains, such as natural language processing and image classification, has witnessed a significant increase in model size over time, as evidenced for benchmark problems [pap23a, pap23b]. This trend highlights the advantages associated with training larger models. Consequently, investigating the phenomenon of the test error exhibiting the so-called *double descent* curve has emerged as a paramount theoretical question in the field of machine learning.

Based on this observation, a number of research papers have attempted to provide an explanation for this phenomenon. See the survey [Bel21] and references therein. These studies have focused on determining the conditions under which *benign overfitting* can occur and have also sought to understand why, within the overparametrized regime, gradient descent tends to converge towards parameter configurations that yield good generalization performance, a phenomenon known as *implicit bias*.

The objective functions in such tasks typically have infinitely many global minimizers – usually, there are infinitely many networks fitting the training samples exactly in the overparameterized regime [ZBH$^+$17] –, which means that the choice of the algorithm strongly influences which minimum is picked in this highly non-convex optimization problem.

Through rigorous numerical simulations, starting with the seminal work of [SHN$^+$18] on classification and support vector machines, several works [JT19, NTSS17, NTS15, ZBH$^+$17, ZBH$^+$21, WGL$^+$20] have undertaken a systematic exploration of the implicit bias of gradient descent in the training of deep neural networks.

Whereas it is not even clear by now how to measure the implicit bias — one could quantify it, e.g., in low complexity [NTS15] or in high generalizability [HS97] —, the empirical studies observed that the factorized structure of such networks is crucial for successful training. However, due to the complexity of the model class, there is still very little corresponding thorough theoretical analysis/understanding available. To close the gap between theory and practice for such a complex phenomenon, some simplified "training" models have been proposed and analyzed in many works. In particular, several works in this line of research performed an analysis of overparametrized linear regression models, i.e., overparametrized *least squares* objective functions, [BHX20, HMRT22, BLLT20, MZS23], on vector factorization [Hof17], [ZYH19], [VKR19], [WGL$^+$20], [LNHW21], [CMR23], [LZQY22], [YZQM20] or matrix factorization [ACH18], [ACHL19], [CGMR20], [GKK20], [GBLJ19], [GSD20], [GLSS18], [SS21], [WCZT22], [GWB$^+$17], [NTSS17], [NTS15], [RC20], [SHN$^+$18], [WR21].

Essentially, all of these works agree on the point that, if initialized close to zero and applied to a plain least-squares overparametrized formulation, Equation (5.2) vanilla gradient descent/flow exhibits an implicit bias towards global minimizers that are sparse, in the vector case, or of low-rank, in the matrix case. This is remarkable since it shows that in overparametrized regimes, gradient descent has an implicit tendency toward *simple* solutions.

Such simplified analysis and results can even be connected to general neural networks in the infinite width regime via the neural tangent kernel [JGH18]. However, a comprehensive understanding of the implicit bias of gradient descent in the training of finite-width networks remains a challenge and one of the main problems in theoretical machine learning, especially for large models present in natural language processing such as transformers [MRG$^+$20]. Whereas the obtained insights on vector and matrix factorization might not yet resolve the mystery of deep learning, they provide valuable tools for more classical problems. One such example is sparse recovery, which lies at the interface of high-dimensional statistics and signal processing. Over the past two decades, several methods have been developed to recover intrinsically low-dimensional data, such as sparse vectors or low-rank matrices, which have been the subject of the previous chapters. For example, one can ask if obtaining new algorithms for such problems is possible by using overparametrized models.

While previous contributions tried to explain the implicit bias phenomenon and its connection to the double descent behavior, most of these works were not interested in how to use this bias to leverage current approaches to solve a certain optimization problem. For example, is it possible to use an overparametrized objective function together with gradient descent to solve a problem with a certain structure, e.g., a problem with convex constraints? In this chapter, we will give the first step in this direction. The goal here is to link the implicit bias of gradient descent/flow to a ubiquitous problem of numerical mathematics, namely (NNLS), the topic of the next section.

5.3 Non-negative least squares

The task of finding (approximate) solutions to a linear system holds significant importance and occurs frequently in the field of numerical mathematics as well as in various applications. Extensive research in scientific computing and related fields has focused on this problem, as it serves as a fundamental component for numerous methods employed in data science. In many scenarios, the physical quantities of interest are inherently non-negative, such as in deconvolution and demixing problems like source separation. Consequently, seeking a solution that minimizes the least-squared error while adhering to additional non-negativity constraints is common. This leads to the formulation of the *non-negative least-squares* (NNLS) problem. Given a linear operator $\mathbf{A} \in \mathbb{R}^{M \times N}$ and data $\mathbf{y} \in \mathbb{R}^M$, NNLS is formally defined as finding

$$\mathbf{x}_+ \in S_+ := \underset{\mathbf{z} \geq \mathbf{0}}{\arg\min} \, \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2^2. \tag{NNLS}$$

Large scale applications in which (NNLS) appears include NMR relaxometry [SK21], imaging deblurring [BZZB09], biometric templates [KKKT19], transcriptomics [KBP+11], magnetic microscopy [MLL+19], sparse hyperspectral unmixing [ELX13], and system identification [CRBH11]. The formulation in (NNLS) is also closely related to non-negative matrix factorization [CSEP21] and supervised learning methods such as support vector machines [Vap99]. We refer the interested reader to [CP10b] for a survey about the development of algorithms that enforce non-negativity.

Compared to the standard least squares approach, including the non-negativity constraint introduces additional challenges. By now, three primary algorithmic approaches have been developed for solving this problem: (i) interior point methods, (ii) active set methods, and (iii) projected gradient methods. In the first approach, the non-negative least squares (NNLS) problem can be recast as the quadratic problem

$$\operatorname*{arg\,min}_{\mathbf{x}\geq\mathbf{0}} \frac{1}{2}\langle \mathbf{x}, \mathbf{Q}\mathbf{x}\rangle + \langle \mathbf{c}, \mathbf{x}\rangle, \tag{5.1}$$

where $\mathbf{Q} = \mathbf{A}^\top\mathbf{A}$ and $\mathbf{c} = -\mathbf{A}^\top\mathbf{y}$, which is then solved via interior point methods. For $M = N$, these are guaranteed to converge to a ε-solution in $O(N^3 \ln \varepsilon^{-1})$ time [BMM06]. The (ii) active set methods [LH95] represent the most commonly used solution to (NNLS). They exploit the fact that the solution of (NNLS) can be found by solving an unconstrained problem with *inactive variables* that do not contribute to the constraints. Both (i) and (ii) require solving a linear system at each iteration, limiting their scalability. In contrast, (iii) projected gradient methods like projected gradient descent (PGD) only require matrix-vector multiplications and the projection to the positive orthant can be trivially performed, e.g., [Lin07, KSD13, BTT91].

However, selecting an appropriate step size poses a challenge for these methods. Despite the guaranteed convergence of projected gradient descent (PGD) with a step size determined by the inverse of the Lipschitz constant, this approach exhibits slow convergence for ill-conditioned problems. Employing step size acceleration methods, such as the Barzilai-Borwein step-size [BB88], in such scenarios can lead to cyclic behavior in PGD, causing it to fail to converge even when the chosen step-size is proven to be effective for unconstrained gradient descent [DF05]. In light of recent insights into the implicit bias of (vanilla) gradient descent [Hof17, ZYH19, VKR19, WGL+20, LNHW21, CMR23], we ask the following question.

Is it possible to design solvers for (NNLS) that are based on unconstrained optimization by exploiting the implicit bias of first-order methods?

Contribution of this chapter:

this chapter aims to affirmatively answer this question by leveraging the implicit bias of the gradient descent for large models. First, we formulate a nonconvex unconstrained but overparametrized ℓ_2-functional capable of capturing the geometry imposed by the convex constraints in (NNLS).

Furthermore, when applied to this functional, we demonstrate that the vanilla gradient flow method exhibits implicit bias towards non-negative solutions. From a conceptual standpoint, we introduce a novel approach for trading-off side constraints in optimization problems with the complexity of the optimization landscape without relying on the concept of Lagrangian multipliers. This trade-off becomes particularly advantageous in ill-conditioned problems, where the geometry of the constrained set, although convex, presents challenges in selecting an appropriate step size. Advanced techniques for stochastic and accelerated step-size tuning can be explored by employing an unconstrained substitute.

As mentioned in [VKR19, CMR23], the implicit sparsity regularization bridges the recent studies on gradient descent and compressive sensing. In fact, the overparametrized gradient descent provides a tuning-free alternative to established algorithms like LASSO and Basis Pursuit. A nice by-product of our approach is that, by choosing the initialization of gradient flow close to zero, one can add an additional (weighted-)ℓ_1-bias on top of non-negativity. This latter feature is inherited from previous works like [VKR19, CMR23] and is only of interest in the special case of applying NNLS to sparse recovery.

Before detailing our results, we however need to remember some notation already introduced in Chapter 1. We use $\odot$ to denote the Hadamard product, i.e., the vectors $\mathbf{x} \odot \mathbf{y}$ and $\mathbf{x}^{\odot p}$ have entries $(\mathbf{x} \odot \mathbf{y})_n = x_n y_n$ and $(\mathbf{x}^{\odot p})_n = x_n^p$, respectively. We abbreviate $\tilde{\mathbf{x}} := \bigodot_{k\in[L]} \mathbf{x}^{(k)} = \mathbf{x}^{(1)} \odot \cdots \odot \mathbf{x}^{(L)}$. The logarithm is applied entry-wise to positive vectors, i.e., $\log(\mathbf{x}) \in \mathbb{R}^N$ with $\log(\mathbf{x})_n = \log(x_n)$. For convenience, we denote by $\mathbf{x} \geq \mathbf{y}$ the entry-wise bound $x_n \geq y_n$, for all n, and define $\mathbb{R}_+^N = \{\mathbf{x} \in \mathbb{R}^N : \mathbf{x} \geq \mathbf{0}\}$. The all-zeros and all-ones vectors are denoted by $\mathbf{0}$ and $\mathbf{1}$, where the dimension is always clear from the context. For $\mathbf{x}_+ \in S_+$, we furthermore define

$$\mathbf{y}_+ := \mathbf{A}\mathbf{x}_+,$$

which is the unique Euclidean projection of $\mathbf{y}$ onto the convex and closed set

$$C_+ := \{\mathbf{A}\mathbf{z} \colon \mathbf{z} \in \mathbb{R}_{\geq 0}\}.$$

5.4 Solving NNLS via vanilla gradient descent

In this section, we demonstrate that the implicit regularization exhibited by the (vanilla) gradient flow/descent method can effectively solve the (NNLS) problem through unconstrained least-square optimization. The approach involves artificially introducing excessive parameters – the overparametrization technique – to the objective function and minimizing it using a scalable first-order method like gradient descent. Subsequently, we establish that the algorithm's trajectory in the parameter space tends to favor a non-negative solution, a characteristic that is not inherently implied by the loss function.

Although the formulation proposed here in this chapter is inherently discrete since it uses gradient descent (see [Lem12] for a historical perspective on the method), which, for a given step-size $\eta > 0$, is given by

$$x_{k+1} = x_k - \eta \nabla f(x_k),$$

the analysis developed in this chapter will adopt a continuous approach through an infinitesimal stepsize limit, i.e., a gradient flow argument.

More precisely, this is obtained by considering that our iterates x_k are sampled at each multiple of η, from a function $x : \mathbb{R}_+ \to \mathbb{R}^N$ such that $x_k = x(k\eta)$. We extend the definition of this function to each point in the domain by interpolating them with an affine interpolation. This gives, for $t = k\eta$, $x(t+\eta) = x_{k+1} = x_k - \eta\nabla f(x_k) = x(t) - \eta\nabla f(x(t))$. Therefore, the gradient flow arises when the step size becomes arbitrarily small, i.e., when $\eta \to 0$. Indeed, this leads to

$$\frac{d}{dt}x(t) = \lim_{\eta \to 0} \frac{x(t+\eta) - x(t)}{\eta} = -\nabla f(x(t)).$$

From this observation, gradient descent can be seen as an Euler discretization for the gradient flow equation.

This analysis simplifies various challenges encountered in analyzing discrete algorithms, eliminating the need to devise rules for the step size. Additionally, techniques developed for ordinary differential equations (ODEs) and continuous flows, a quite developed field, can be applied in such analyses [San17]. However, a significant challenge arises when attempting to translate continuous results derived from gradient flow arguments into discrete ones, and a comprehensive theory addressing this issue is still elusive [GBR21, EC21]. Nonetheless, the continuous analysis provides strong evidence for the existence of the implicit bias phenomenon and suggests the possibility of extending it to a discrete but potentially more complicated proof. It is worth noting that the majority of works on implicit bias focus on continuous scenarios, see [YKM21, AMN$^+$21, LWLA22] and references therein, and we will follow a similar approach in this chapter.

To be more precise, we consider gradient flow on the factorized loss-function

$$\mathcal{L}_{\text{over}}\big(\mathbf{x}^{(1)},\dots,\mathbf{x}^{(L)}\big) := \frac{1}{2}\left\|\mathbf{A}\big(\mathbf{x}^{(1)}\odot\cdots\odot\mathbf{x}^{(L)}\big)-\mathbf{y}\right\|_2^2, \tag{5.2}$$

which appeared before in the context of implicit ℓ_1-regularization [VKR19, LNHW21, CMR23]. In these works, it was shown that if there exists a non-negative solution $\mathbf{x} \geq \mathbf{0}$ with $\mathbf{A}\mathbf{x} = \mathbf{y}$ and all factors $\mathbf{x}^{(k)}$ are initialized with $\alpha\mathbf{1}$ at $t = 0$, for $\alpha > 0$ sufficiently small, then the product $\mathbf{x}^{(1)}(t)\odot\cdots\odot\mathbf{x}^{(L)}(t)$ of the gradient flow approximates an ℓ_1-minimizer among all positive solutions, for $t \to \infty$. While previous works such as [VKR19, LNHW21, CMR23] found the positivity of the limit to be a technical challenge to be circumvented. This required an adaptation of (5.2), the present work utilizes this positivity to address the problem of solving (NNLS).

Our contribution is twofold. First, by robustifying the argument in [CMR23] we show, for any $\mathbf{A}$, $\mathbf{y}$, and positive (identical) initialization $\mathbf{x}^{(k)}(0) = \mathbf{x}_0 > \mathbf{0}$, that the product $\mathbf{x}^{(1)}(t)\odot\cdots\odot\mathbf{x}^{(L)}(t)$ converges to a solution $\mathbf{x}_+$ of (NNLS). *A crucial point here is that, in contrast to [CMR23], the existence of a solution* $\mathbf{A}\mathbf{x} = \mathbf{y}$ *is not required anymore.* As part of this, we characterize the convergence rate of the trajectory as $\mathcal{O}(\frac{1}{t})$ for the gradient flow formulation. Second, as a nice by-product of relying on the existing theory, we conclude that if $\mathbf{x}_0$ is chosen sufficiently close to zero, the limit of gradient flow is of (approximately) minimizing a weighted ℓ_1-norm among all possible solutions of (NNLS). Note that to guarantee a similar additional regularization in the case of general measurements $\mathbf{A}$, the established techniques for solving (NNLS) — (i)-(iii) discussed above — would require notable adjustments both in methodology and in theory.

The following two theorems formalize these claims. Let us emphasize that a small initialization is only required in Theorem 5.4.1 to obtain additional ℓ_1-regularization. In general, any instance of NNLS can be solved via our method with arbitrary positive initialization:

Theorem 5.4.1. *Let $L \geq 2$, $\mathbf{A} \in \mathbb{R}^{M\times N}$ and $\mathbf{y} \in \mathbb{R}^M$. Define the overparameterized loss function $\mathcal{L}_{over}$ as*

$$\mathcal{L}_{over}\big(\mathbf{x}^{(1)},\dots,\mathbf{x}^{(L)}\big) := \frac{1}{2}\left\|\mathbf{A}\tilde{\mathbf{x}}-\mathbf{y}\right\|_2^2 \tag{5.3}$$

where $\tilde{\mathbf{x}} = \mathbf{x}^{(1)}\odot\cdots\odot\mathbf{x}^{(L)}$. Let $\mathbf{x}_0 > \mathbf{0}$ be fixed and, for any k, let $\mathbf{x}^{(k)}(t)$ follow the flow $\big(\mathbf{x}^{(k)}\big)'(t) = -\nabla_{\mathbf{x}^{(k)}}\mathcal{L}_{over}$ with $\mathbf{x}^{(k)}(0) = \mathbf{x}_0$. Let S_+ be the set defined in (NNLS)*. Then the limit $\tilde{\mathbf{x}}_\infty := \lim_{t\to\infty}\tilde{\mathbf{x}}(t)$ exists and lies in S_+. Also, for $\mathbf{y}_+$ as defined in* (5.3)*, there exists an absolute constant $C > 0$ that only depends on the choice of $\mathbf{A}$, $\mathbf{y}$, and $\mathbf{x}_0$ such* *that*

$$\|\mathbf{A}\tilde{\mathbf{x}}(t) - \mathbf{y}_+\|_2^2 \leq \frac{C}{t},$$

for any $t > 0$. Moreover, let $\epsilon > 0$ and assume in addition that we start the gradient descent with a very small initialization $\mathbf{x}_0 = \alpha\mathbf{1}$. If

$$\alpha \leq h(Q_+, \epsilon) := \begin{cases} \exp\left(-\frac{1}{2} - \frac{Q_+^2 + Ne^{-1}}{2\epsilon}\right) & \text{if } L = 2 \\ \left(\frac{2\epsilon}{L(Q_+ + N + \epsilon)}\right)^{\frac{1}{L-2}} & \text{if } L > 2 \end{cases}, \tag{5.4}$$

where $Q_+ = \min_{\mathbf{z} \in S_+} \|\mathbf{z}\|_1$, then the ℓ_1-norm of $\tilde{\mathbf{x}}_\infty$ satisfies

$$\|\tilde{\mathbf{x}}_\infty\|_1 - \min_{\mathbf{z} \in S_+} \|\mathbf{z}\|_1 \leq \epsilon.$$

Let us highlight two of the most important features of the first theorem right away. First, the limit $\tilde{\mathbf{x}}_\infty$ exists and minimizes $\mathcal{L}_{\text{over}}$ to global optimality for any choice of $\mathbf{A}$ and $\mathbf{y}$ and any initialization magnitude $\alpha > 0$. The non-convex nature of equation (5.2) makes this statement non-trivial. Additionally, it is noteworthy that Theorem 5.4.1 does not impose any specific technical assumptions on matrix $\mathbf{A}$ and vector $\mathbf{y}$, yet it applies to arbitrary problems in the form of non-negative least squares (NNLS). However, it is important to note that we require identical initialization of all factors and consider the continuous gradient flow in our analysis.

The fact that we focus on the case of identical initialization, i.e., $\mathbf{x}^{(1)}(0) = \cdots = \mathbf{x}^{(L)}(0) = \mathbf{x}_0$ is not restrictive when solving (NNLS). Indeed, all solutions of (NNLS) are stationary points of (5.2) and, as such, can be described as the limit of gradient flow on (5.2) under suitably chosen identical initialization. We start by noting that since the problem (NNLS) is convex, the Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient for optimality, see, e.g., [Bjö96].

Theorem 5.4.2. *(Karush-Kuhn-Tucker conditions for NNLS) A point $\mathbf{x}_+ \in \mathbb{R}^N$ is a solution of problem* (NNLS) *if and only if there exists $\mathbf{w}^\star \in \mathbb{R}^n$ and a partition $A \cup P = \{1, \ldots, N\}$ such that*

$$\mathbf{w}^\star = \mathbf{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x}_+), \tag{5.5}$$

$$(x_+)_i = 0, \ i \in A, \quad (x_+)_i > 0, \ i \in P, \tag{5.6}$$

$$w_i^\star \leq 0, \ i \in A, \quad w_i^\star = 0, \ i \in P. \tag{5.7}$$

The equations (5.6-5.7) imply

$$(x_+)_i w_i^\star = 0, \ i = 1, \dots, N, \tag{5.8}$$

which are the KKT complementarity conditions.

From these conditions, one can observe that all solutions of (NNLS) can be represented by stationary points of the functional

$$\mathcal{L}_{\text{over}}\big(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}\big) := \frac{1}{2} \left\| \mathbf{A}\big(\mathbf{x}^{(1)} \odot \dots \odot \mathbf{x}^{(L)}\big) - \mathbf{y} \right\|_2^2$$

in (5.2), i.e., points that satisfy, for all $\ell \in [L]$, the equation

$$\nabla L_{\mathbf{x}^{(\ell)}}\big(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}\big) = \Big[\mathbf{A}^{\mathrm{T}}\Big(\mathbf{A}\big(\mathbf{x}^{(1)} \odot \dots \odot \mathbf{x}^{(L)}\big) - \mathbf{y}\Big)\Big] \odot \Big(\bigodot_{k \neq \ell} \mathbf{x}^{(k)}\Big) = \mathbf{0}. \tag{5.9}$$

Indeed, for any given optimal point $\mathbf{x}_+$ of the (NNLS) problem, it is straightforward to check that the conditions in Theorem 5.4.2 imply that $\mathbf{x}^{(1)} = \dots = \mathbf{x}^{(L)} = \mathbf{x}_+^{\odot \frac{1}{L}}$ is a stationary point of (5.2) with $\mathbf{x}^{(1)} \odot \dots \odot \mathbf{x}^{(L)} = \mathbf{x}_+$. The same argument holds for stationary points of the reduced functional (5.10). In particular, this implies that any solution of (NNLS) can be described as the limit of gradient flow on (5.2) under suitably chosen *identical* initialization.

As we have done in Chapter 3, we develop a roadmap for the proof of Theorem 5.4.1. It consists of three major steps: First, the objective function and the corresponding flow are reduced to a simplified form by using the fact that all factors are initialized identically. Second, we prove that the reduced flow converges and characterize its limits as the minimizer of a specific optimization problem. Finally, we show that if $\mathbf{x}_0 = \alpha \mathbf{1}$, the limit approximately minimizes the ℓ_1-norm among all possible solutions of (NNLS). Whereas the first and the third steps are taken unchanged from [CMR23], the second step, which is the backbone of the argument, requires different reasoning due to the (possible) non-existence of solutions $\mathbf{A}\mathbf{z} = \mathbf{y}$. Let us now start with the proof.

As already mentioned, we use the same reduction as in [CMR23] to analyze the dynamics $\mathbf{x}'(t) = -\nabla \mathcal{L}_{\text{over}}$.

Lemma 5.4.3 (Identical Initialization, [CMR23, Lemma A.2])**.** *Suppose $\mathbf{x}^{(k)}(t)$ follows the negative gradient flow*

$$\big(\mathbf{x}^{(k)}\big)'(t) = -\nabla_{\mathbf{x}^{(k)}} \mathcal{L}_{over}\big(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}\big) .$$

If all initialization vectors are identical, i.e. $\mathbf{x}^{(k)}(0) = \mathbf{x}^{(k')}(0)$ for all $k, k' \in [L]$, then the *vectors remain identical for all $t \geq 0$, i.e.* $\mathbf{x}^{(k)}(t) = \mathbf{x}^{(k')}(t)$. *Moreover, the dynamics will be given by*

$$\mathbf{x}'(t) = -\nabla \mathcal{L}(\mathbf{x}),$$

where $\mathbf{x} = \mathbf{x}^{(1)} = \cdots = \mathbf{x}^{(L)}$ *and* $\mathcal{L}(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}^{\odot L} - \mathbf{y}\|_2^2$.

Due to the identical initialization, all the factors (layers) $\mathbf{x}^{(k)}$ remain identical over time. This idea appeared already in the literature under the name *balancedness condition* [ACHL19, RMC21]. Therefore, based on Lemma 5.4.3, we can restrict ourselves to a simplified loss in the following. Moreover, our present contribution stems from the insight that in most of the above papers [ACH18, ACHL19, CGMR20, CMR23], the signs of components do not change over time when gradient flow is applied. Instead of viewing this feature as an obstacle, cf. [CMR23, Section 2], we use it to naturally link the implicit bias of gradient descent/flow to another ubiquitous problem of numerical mathematics, namely, Equation (NNLS).

Definition 5.4.4 (Reduced Factorized Loss)**.** *Let $L \in \mathbb{N}$, $L \geq 2$. For $\mathbf{A} \in \mathbb{R}^{M \times N}$ and $\mathbf{y} \in \mathbb{R}^M$, the* reduced factorized loss function *is defined as*

$$\mathcal{L} \colon \mathbb{R}^N \to [0, \infty), \qquad \mathcal{L}(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x}^{\odot L} - \mathbf{y}\|_2^2. \tag{5.10}$$

Its derivative is given by $\nabla \mathcal{L}(\mathbf{x}) = \left[\mathbf{A}^T(\mathbf{A}\mathbf{x}^{\odot L} - \mathbf{y})\right] \odot \mathbf{x}^{\odot L-1}$.

One intriguing observation is that the factorization process induces a Riemannian metric, and our method can be regarded as a Riemannian gradient approach. In this framework, the metric distorts the space, ensuring that positive solutions are always obtained. Indeed, originally we have

$$\mathbf{x}'(t) = -L\mathbf{x}^{\odot L-1} \odot [\mathbf{A}^\top(\mathbf{A}\mathbf{x}^{\odot L} - \mathbf{y})].$$

Now, we want to look at the dynamics at the point $\tilde{\mathbf{x}} = \mathbf{x}^{\odot L}$, which will obey the following equation

$$\begin{aligned}
\tilde{\mathbf{x}}'(t) &= L\mathbf{x}^{\odot L-1} \odot \mathbf{x}'(t) \\
&= -L^2\mathbf{x}^{\odot L-1} \odot [\mathbf{x}^{L-1} \odot [\mathbf{A}^\top(\mathbf{A}\mathbf{x}^{\odot L} - \mathbf{y})]] \\
&= -L^2\mathbf{x}^{2L-2} \odot [\mathbf{A}^\top(\mathbf{A}\mathbf{x}^{\odot L} - \mathbf{y})] \\
&= -L^2\tilde{\mathbf{x}}^{2-\frac{2}{L}} \odot [\mathbf{A}^\top(\mathbf{A}\tilde{\mathbf{x}} - \mathbf{y})] \\
&= -L^2\tilde{\mathbf{x}}^{2-\frac{2}{L}} \odot [\nabla \mathcal{L}(\tilde{\mathbf{x}})] \\
&= -\tilde{\mathbf{x}}(t)^{\odot q} \odot \nabla \mathcal{L}(\tilde{\mathbf{x}}).
\end{aligned}$$

Here we used that $q = 2 - \frac{2}{L}$ and $\mathcal{L}(\tilde{\mathbf{x}}) = \frac{1}{2}\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{y}\|_2^2$. This shows that if one considers the matrix $G(x) = L^{-2}\operatorname{diag}(x)^{\frac{2}{L}-2} = L^{-2}\operatorname{diag}(x)^q$, which defines the Riemmanian metric $\langle u, v\rangle_{\mathbf{x}} = \langle L^2\mathbf{x}^{\odot -q}u, v\rangle$, the dynamics of $\tilde{\mathbf{x}}'(t)$ can be recast as

$$\tilde{\mathbf{x}}'(t) = -G(x)^{-1}\nabla\mathcal{L}(\tilde{\mathbf{x}}) = -\nabla_R\mathcal{L}(\tilde{\mathbf{x}}),$$

where ∇_R denotes the Riemannian gradient. See [Bou23, Chapter 3] for more details. To prove that $\mathbf{x}^{\odot L}$ converges to an element in S_+ defined in (NNLS), we use the concept of Bregman divergence, which measures the distance between a function and its first-order approximation. This notion of distance is strongly connected to the mirror descent algorithm [RM15] used in several problems in machine learning and in the analysis of overparametrized models.

Definition 5.4.5 (Bregman Divergence)**.** *Let $F : \Omega \to \mathbb{R}$ be a continuously-differentiable, strictly convex function defined on a closed convex set Ω. The Bregman divergence associated with F for points $p, q \in \Omega$ is defined as*

$$D_F(p, q) = F(p) - F(q) - \langle\nabla F(q), p - q\rangle. \tag{5.11}$$

Although this function is not a metric, since it does not satisfy the triangle inequality, it has several important properties. For example, this function is non-negative, strictly convex in its first argument p, and it is unique up to affine difference, i.e., $D_F(p, q) = D_G(p, q)$ if and only if $F - G$ is an affine function. We can now show convergence of $\mathbf{x}(t)^{\odot L}$ and characterize its limit.

Theorem 5.4.6. *Let $\tilde{\mathbf{x}}(t) = \mathbf{x}(t)^{\odot L}$ and*

$$\mathbf{x}'(t) = -\frac{1}{L^2}\nabla\mathcal{L}(\mathbf{x}(t)) = -\frac{1}{L}\left[\mathbf{A}^T(\mathbf{A}\mathbf{x}^{\odot L}(t) - \mathbf{y})\right] \odot \mathbf{x}^{\odot L-1}(t) \tag{5.12}$$

with $\mathbf{x}(0) \geq 0$. Then $\tilde{\mathbf{x}}_\infty := \lim_{t\to\infty}\tilde{\mathbf{x}}(t)$ exists and

$$\tilde{\mathbf{x}}_\infty = \operatorname*{arg\,min}_{\mathbf{z}\in S_+} g_{\tilde{\mathbf{x}}(0)}(\mathbf{z}) := \operatorname*{arg\,min}_{\mathbf{z}\in S_+} \begin{cases} \langle\mathbf{z}, \log(\mathbf{z}) - \mathbf{1} - \log(\tilde{\mathbf{x}}(0))\rangle & \text{if } L = 2, \\ \langle\mathbf{z}, \tilde{\mathbf{x}}(0)^{\odot\frac{2}{L}-1}\rangle - \frac{L}{2}\|\mathbf{z}\|_{2/L}^{2/L} & \text{if } L > 2 \end{cases} \tag{5.13}$$

where S_+ is defined in (NNLS).

Theorem 5.4.6 resembles [CMR23, Theorem 2.7]. Note, however, that the definition of S_+ is different. The proof unifies and simplifies several of the arguments given there, and that Theorem 5.4.6 does not require the existence of a solution of $\mathbf{A}\mathbf{x} = \mathbf{y}$.

To prove Theorem 5.4.6, we consider a function $F : \mathbb{R}^N_{\geq 0} \to \mathbb{R}$ that is based on an important concept that appears in thermodynamics, namely, the so-called *Tsallis q-logarithm*, introduced in the seminal paper [Tsa88]. It is given by

$$F(\mathbf{x}) = \langle \mathbf{x}, \ln_q \mathbf{x} \rangle, \tag{5.14}$$

where $q = 2 - \frac{2}{L}$ and the q-logarithm is given by

$$\ln_q(u) = \begin{cases} \frac{x^{1-q}-1}{1-q}, \text{ if } q \neq 1; \\ \ln(x), \text{ if } q = 1, \end{cases}$$

where we set $0 \log(0) = \lim_{z \to 0} z \log(z) = 0$. The q-logarithm function $\ln_q(u)$ has some basic properties that will be used in the calculations below.

(i) $\ln_q(1/u) = -u^{q-1} \ln_q u$;

(ii) $\ln_q(uv) = \ln_q u + \ln_q v + (1-q) \ln_q u \ln_q v$

(iii) $\ln_q(u/v) = v^{q-1}(\ln_q u - \ln_q v)$

Furthermore, its derivative is given by $\frac{d}{du}(\ln_q u) = u^{-q}$. Using these properties, it follows that the gradient vector is $\nabla F(\mathbf{x}) = \ln_q \mathbf{x} + \mathbf{x}^{\odot 1-q}$ and the Hessian is described by $\nabla^2 F(\mathbf{x}) = (2-q) \operatorname{diag}(\mathbf{x}^{\odot -q})$. In particular, the function F is convex, since $q = 2 - 2/L < 2$ and, therefore, $\nabla^2 F(\mathbf{x}) \succcurlyeq 0$.

The proof of Theorem 5.4.6 is based on two ingredients, a function for which the Bregman divergence can be nicely calculated and for which its temporal derivative $\partial_t D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t))$ is bounded by the derivative of the objective function and a so-called *Lyapunov functional* (also known as energy functional) that decreases as $t \to \infty$. The convergence rate will naturally appear from the minimization of this functional. This idea, which dates back from the work by Aleksandr Lyapunov [Lya92], became a crucial tool to analyze dynamical systems given by a system of ODEs. We start by calculating the Bregman divergence of the q-logarithm.

Proposition 5.4.7. *The Bregman divergence D_F of the function $F : \mathbb{R}^N_{\geq 0} \to \mathbb{R}$ define above is given by*

$$D_F(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x} \odot \ln_q(\frac{\mathbf{x}}{\mathbf{y}}) - (\mathbf{x} - \mathbf{y}), \mathbf{y}^{\odot 1-q} \rangle.$$

In particular, for $q = 1$, $D_F(\mathbf{x}, \mathbf{y})$ reduces to the Kullback-Leibler divergence for vectors in $\mathbb{R}^N_{\geq 0}$.

Proof. The definition of the Bregman divergence yields

$$\begin{aligned}D_F(\mathbf{x},\mathbf{y}) &= F(\mathbf{x}) - F(\mathbf{y}) - \langle \nabla F(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \\ &= \langle \mathbf{x}, \ln_q \mathbf{x}\rangle - \langle \mathbf{y}, \ln_q \mathbf{y}\rangle - \langle \ln_q \mathbf{y} + \mathbf{y}^{\odot 1-q}, \mathbf{x}-\mathbf{y}\rangle \\ &= \langle \mathbf{x}, \ln_q \mathbf{x}\rangle - \langle \mathbf{y}, \ln_q \mathbf{y}\rangle - \langle \ln_q \mathbf{y}, \mathbf{x}\rangle + \langle \ln_q \mathbf{y}, \mathbf{y}\rangle - \langle \mathbf{y}^{\odot 1-q}, \mathbf{x}\rangle + \langle \mathbf{y}^{\odot 1-q}, \mathbf{y}\rangle \\ &= \langle \mathbf{x}, \ln_q \mathbf{x} - \ln_q \mathbf{y}\rangle - \langle \mathbf{x}-\mathbf{y}, \mathbf{y}^{\odot 1-q}\rangle.\end{aligned}$$

By using that $\ln_q(u-v) = v^{1-q}\ln_q(u/v)$, we have that $\ln_q \mathbf{x} - \ln_q \mathbf{y} = \mathbf{y}^{\odot 1-q} \odot \ln(\frac{\mathbf{x}}{\mathbf{y}})$. Thus,

$$\begin{aligned}D_F(\mathbf{x},\mathbf{y}) &= \langle \mathbf{x}, \mathbf{y}^{1-q} \odot \ln_q(\frac{\mathbf{x}}{\mathbf{y}})\rangle - \langle \mathbf{x}-\mathbf{y}, \mathbf{y}^{\odot 1-q}\rangle \\ &= \langle \mathbf{x} \odot \ln_q(\frac{\mathbf{x}}{\mathbf{y}}), \mathbf{y}^{1-q}\rangle - \langle \mathbf{x}-\mathbf{y}, \mathbf{y}^{\odot 1-q}\rangle.\end{aligned}$$

□

Remark 5.4.8. *This Bregman divergence has a nice interpretation in the particular case $q = 1$. Indeed, for probability vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N_{\geq 0}$ that add up to one, i.e., with $\langle \mathbf{x}, \mathbf{1}\rangle = \langle \mathbf{y}, \mathbf{1}\rangle = 1$, the Bregman divergence D_F reduces to the Kullback-Leibler function $KL[\mathbf{x},\mathbf{y}] = \langle \ln(\frac{\mathbf{x}}{\mathbf{y}}), \mathbf{x}\rangle$. In general, in Information Theory and Information Geometry, such divergence is called Kullback-Leibler divergence if the underlying vectors are positive. See [Ama16, Page 11].*

We will use another Lemma about the boundedness of the Bregman divergence. It is inspired by [CMR23, Lemma 2.5] but we prove it here for the sake of completeness since we have a different Bregman divergence.

Lemma 5.4.9. *Let F be the function defined in Equation (5.14) and $\tilde{\mathbf{x}}(t)\colon \mathbb{R}_+ \to \mathbb{R}^N_+$ be a continuous function with $\tilde{\mathbf{x}}(0) > \mathbf{0}$. Let $\mathbf{z} \geq \mathbf{0}$ be fixed. If $D_F(\mathbf{z}, \tilde{\mathbf{x}}(t))$ is bounded, then $\|\tilde{\mathbf{x}}(t)\|_2$ is bounded.*

Proof. We will prove the statement by using the contrapositive. Let $\|\tilde{\mathbf{x}}(t)\|_2$ be unbounded. Then there exists a sequence $0 < t_1 \leq t_2 \leq \dots$ such that $\|\tilde{\mathbf{x}}(t_k)\|_2 \to \infty$. Hence there exists some $n \in [N]$ and a subsequence $0 < t_{n_1} \leq t_{n_2} \leq \dots$ such that $\tilde{x}_n(t_{n_k}) \to \infty$. Because of that, we only need to analyze a one-dimensional version of the Bregman divergence. From Proposition 5.4.7, we have

$$D_F(\mathbf{z}, \tilde{\mathbf{x}}(t)) = \langle \mathbf{z} \odot \ln_q\left(\frac{\mathbf{z}}{\tilde{\mathbf{x}}(t)}\right) - (\mathbf{z} - \tilde{\mathbf{x}}(t)), \tilde{\mathbf{x}}(t)^{1-q}\rangle.$$

For $z \geq 0$ and $q = 1$, we have

$$\begin{aligned} D_F(\mathbf{z}, \tilde{\mathbf{x}}_n(t)) &= \mathbf{z} \ln\left(\frac{\mathbf{z}}{\tilde{\mathbf{x}}_n(t)}\right) - \mathbf{z} + \tilde{\mathbf{x}}_n(t) \\ &= \tilde{\mathbf{x}}_n(t) - \mathbf{z} \ln\left(\frac{\mathbf{z}}{\tilde{\mathbf{x}}_n(t)}\right) - \mathbf{z} \to \infty, \end{aligned}$$

when $\tilde{\mathbf{x}}_n(t_{n_k}) \to \infty$. For $z \geq 0$ and $q = 2$, we have,

$$\begin{aligned} D_F(\mathbf{z}, \tilde{\mathbf{x}}_n(t)) &= \left(\mathbf{z} \ln_q\left(\frac{\mathbf{z}}{\tilde{\mathbf{x}}_n(t)}\right) - \mathbf{z} + \tilde{\mathbf{x}}_n(t)\right) \tilde{\mathbf{x}}_n(t)^{1-q} \\ &= \mathbf{z}\left[\frac{\left(\frac{\mathbf{z}}{\tilde{\mathbf{x}}_n(t)}\right)^{1-q} - 1}{1-q} - (\mathbf{z} - \tilde{\mathbf{x}}(t))\right] \tilde{\mathbf{x}}_n(t)^{1-q} \\ &= \mathbf{z}\left[\frac{\mathbf{z}^{1-q} - \tilde{\mathbf{x}}_n(t)^{1-q}}{1-q} - \mathbf{z}\mathbf{x}_n(t)^{1-q} + \mathbf{x}_n(t)^{2-q}\right] \to \infty, \end{aligned}$$

when $\tilde{\mathbf{x}}_n(t_{n_k}) \to \infty$, since $1 - q < 0$. Thus $D_F(\mathbf{z}, \tilde{\mathbf{x}}(t_{n_k})) \to \infty$ and consequently $D_F(\mathbf{z}, \tilde{\mathbf{x}}(t))$ is unbounded. □

We will first prove Theorem 5.4.6 and, from that, we will use the balancedness condition from Lemma 5.4.3 to finally establish Theorem 5.4.1.

Proof of Theorem 5.4.6. Let us begin with a brief outline of the proof.

- First, we compute the time derivative of $D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t))$, for $\mathbf{z}_+ \in S_+$, where S_+ was defined in (NNLS).
- Second, with this estimate, we will establish the convergence rate of the objective function, i.e., at which rate $\mathbf{A}\tilde{\mathbf{x}}(t) \to \mathbf{y}_+$. To do so, we will show there exists a certain Lyapunov functional that decreases over time, i.e., such that $\mathcal{E}'(t) < 0$.
- Third, by using Lemma 5.4.9, we deduce the boundedness and convergence of $\|\tilde{\mathbf{x}}(t)\|_2$.
- Fourth, we characterize the limit $\tilde{\mathbf{x}}_\infty = \lim_{t\to\infty} \tilde{\mathbf{x}}(t)$.

Since $C_+ = \{\mathbf{Az} \colon \mathbf{z} \in \mathbb{R}_{\geq 0}\} \subset \mathbb{R}^M$ is a closed convex set, the infimum of the convex function $\inf_{w \in C} \|w - y\|_2$ is attained at y_+, which is a vector that fulfills $Ax_+ = y_+$. This means that the set S_+ is non-empty. Let $\mathbf{z}_+$ be any element of S_+. We start by calculating the derivative

$$\begin{aligned} \partial_t \nabla F(\tilde{\mathbf{x}}(t)) &= (2-q)\,\mathrm{diag}(\tilde{\mathbf{x}}(t)^{\odot -q})\tilde{\mathbf{x}}'(t) = -(2-q)\tilde{\mathbf{x}}(t)^{\odot -q} \odot \tilde{\mathbf{x}}^{\odot q}(t) \odot A^T(A\tilde{\mathbf{x}}(t) - \mathbf{y}) \\ &= -(2-q)A^T(A\tilde{\mathbf{x}}(t) - \mathbf{y}) = -(2-q)\nabla\mathcal{L}(\tilde{\mathbf{x}}). \end{aligned}$$

From that, we obtain the temporal derivative of the Bregman divergence $D_F(\tilde{\mathbf{x}}_\infty, \tilde{\mathbf{x}}(t))$,

$$\begin{aligned}\partial_t D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t)) &= \partial_t \left[F(\mathbf{z}_+) - F(\tilde{\mathbf{x}}(t)) - \langle \nabla F(\tilde{\mathbf{x}}(t)), \mathbf{z}_+ - \tilde{\mathbf{x}}(t)\rangle\right] \\ &= -\langle \nabla F(\tilde{\mathbf{x}}(t)), \tilde{\mathbf{x}}'(t)\rangle + \langle \nabla F(\tilde{\mathbf{x}}(t)), \tilde{\mathbf{x}}'(t)\rangle - \langle \partial_t \nabla F(\tilde{\mathbf{x}}(t)), \mathbf{z}_+ - \tilde{\mathbf{x}}(t)\rangle \\ &= (2-q)\langle \nabla \mathcal{L}(\tilde{\mathbf{x}}(t)), \mathbf{z}_+ - \tilde{\mathbf{x}}(t)\rangle.\end{aligned} \tag{5.15}$$

Now, the crucial step is to exhibit a functional $\mathcal{E}(t)$ that depends on the difference $\mathcal{L}(\tilde{\ }\tilde{(}t\mathbf{x})-$ $\mathcal{L}(\mathbf{z}_+))$ and on the Bregman divergence $D_F(\tilde{\mathbf{x}}_\infty, \tilde{\mathbf{x}}(t))$. The function F is chosen in a way to make the Bregman divergence of $\mathcal{L}$, i.e., $D_{\mathcal{L}}(\mathbf{z}_+, \tilde{\mathbf{x}}(t))$, appear in the derivative of this functional. Therefore, consider the functional

$$\mathcal{E}(t) = t(\mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+)) + \frac{1}{2-q} D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t)).$$

Note that $\mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+) \geq 0$, since

$$\frac{d}{dt}(\mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+)) = \langle \nabla \mathcal{L}(\tilde{\mathbf{x}}(t)), \tilde{\mathbf{x}}'(t)\rangle = -\langle \nabla \mathcal{L}(\tilde{\mathbf{x}}(t)), \tilde{\mathbf{x}}(t)^{\odot q} \nabla \mathcal{L}(\tilde{\mathbf{x}}(t))\rangle \leq 0.$$

Using Equation (5.15), we obtain

$$\begin{aligned}\mathcal{E}'(t) &= \mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+) + t\langle \nabla \mathcal{L}(\tilde{\mathbf{x}}(t)), \tilde{\mathbf{x}}'(t)\rangle + \langle \nabla \mathcal{L}(\tilde{\mathbf{x}}(t)), \mathbf{z}_+ - \tilde{\mathbf{x}}(t)\rangle \\ &= -D_{\mathcal{L}}(\mathbf{z}_+, \tilde{\mathbf{x}}(t)) - t\langle \nabla \mathcal{L}(\tilde{\mathbf{x}}(t)), \tilde{\mathbf{x}}^{\odot q} \nabla \mathcal{L}(\tilde{\mathbf{x}}(t))\rangle \leq 0.\end{aligned}$$

Hence, $\mathcal{E}(t)$ is decreasing, which implies

$$t(\mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+)) \leq \mathcal{E}(t) \leq \mathcal{E}(0) = \frac{1}{2-q} D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(0)).$$

Thus, $\mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+) \leq \mathcal{O}\left(\frac{1}{t}\right)$. This implies that

$$\begin{aligned}\|\mathbf{A}\tilde{\mathbf{x}}(t) - \mathbf{y}_+\|_2^2 &\leq \|\mathbf{A}\tilde{\mathbf{x}}(t) - \mathbf{y}\|_2^2 + \|\mathbf{y}_- \mathbf{y}_+\|_2^2 \\ &= \|\mathbf{A}\tilde{\mathbf{x}}(t) - \mathbf{y}\|_2^2 + \|\mathbf{A}\mathbf{z}_+ - \mathbf{y}\|_2^2 \\ &= \mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+) \leq \mathcal{O}\left(\frac{1}{t}\right).\end{aligned} \tag{5.16}$$

Now, we establish that $\|\tilde{\mathbf{x}}(t)\|_2$ is bounded. In fact, from Equation (5.15), we have

$$\begin{aligned}\partial_t D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t)) &= -(2-q)\langle \mathbf{A}\tilde{\mathbf{x}}(t) - \mathbf{y}, \mathbf{A}\tilde{\mathbf{x}}(t) - \mathbf{A}\mathbf{z}_+\rangle \\ &\leq -(2-q)\|\mathbf{A}\tilde{\mathbf{x}}(t) - \mathbf{y}_+\|_2^2 < 0.\end{aligned} \tag{5.17}$$

Note that $D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t))$ must converge for $t \to \infty$ as it is bounded from below and monotonically decreasing, since $D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t)) \geq 0$ and $\partial_t D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t)) \leq 0$. In particular, $0 \leq D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t)) \leq D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(0))$ for all $t \geq 0$. Moreover, from Lemma 5.4.9, we know that $\|\tilde{\mathbf{x}}(t)\|_2$ is bounded since $0 \leq D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t)) \leq D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(0))$.

Since $x(t)$ is bounded, let us denote by B a sufficiently large compact ball around the origin such that $B \cap S_+ \neq \emptyset$ and $\tilde{\mathbf{x}}(t) \in B$, for all $t \geq 0$. Now assume that there exists no $\mathbf{z}_+ \in S_+ \cap B$ such that $\lim_{t\to\infty} D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t)) = 0$, i.e., by compactness of $S_+ \cap B$ there exists $\varepsilon > 0$ such that $\lim_{t\to\infty} D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t)) > \epsilon$, for all $\mathbf{z}_+ \in S_+ \cap B$. By strict convexity of $D_F(\cdot, \tilde{\mathbf{x}}(t))$, this implies that $\tilde{\mathbf{x}}$ is bounded away from the set S_+ on B and $\|\mathbf{A}\tilde{\mathbf{x}}(t) - \mathbf{y}_+\|_2$ cannot converge to zero contradicting the just obtained convergence $\lim_{t\to\infty} \mathbf{A}\tilde{\mathbf{x}}(t) = \mathbf{y}_+$. Hence, there exists $\mathbf{z}_+ \in S_+$ with $\lim_{t\to\infty} D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t)) = 0$. For any such $\mathbf{z}_+$, let us assume that $\tilde{\mathbf{x}}(t) \nrightarrow \mathbf{z}_+$. Then there exists $\varepsilon > 0$ and a sequence of time steps $t_0, t_1, \ldots$ with $\|\mathbf{z}_+ - \tilde{\mathbf{x}}(t_k)\|_2 \geq \varepsilon$ and $\lim_{k\to\infty} D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t_k)) = 0$. Since $\tilde{\mathbf{x}}(t_k)$ is a bounded sequence, a (not relabeled) subsequence converges to some $\bar{\mathbf{x}}$ with $\|\mathbf{z}_+ - \bar{\mathbf{x}}\|_2 \geq \varepsilon$ and $D_F(\mathbf{z}_+, \bar{\mathbf{x}}) = 0$. Since $D_F(\bar{\mathbf{x}}, \bar{\mathbf{x}}) = 0$ and D_F is non-negative, this is a contradiction to the strict convexity of $D_F(\cdot, \bar{\mathbf{x}})$. Hence, $\tilde{\mathbf{x}}_\infty = \lim_{t\to\infty} \tilde{\mathbf{x}}(t) \in S_+$ exists and is the unique solution satisfying $\lim_{t\to\infty} D_F(\tilde{\mathbf{x}}_\infty, \tilde{\mathbf{x}}(t)) = 0$.

Because $\partial_t D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(t))$ is identical for all $\mathbf{z}_+ \in S_+$ (the second line of (5.17) does not depend on the choice of $\mathbf{z}_+$), the difference

$$\Delta_{\mathbf{z}_+} = D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(0)) - D_F(\mathbf{z}_+, \tilde{\mathbf{x}}_\infty) \tag{5.18}$$

is also identical for all $\mathbf{z}_+ \in S_+$. By non-negativity of D_F,

$$D_F(\mathbf{z}_+, \tilde{\mathbf{x}}(0)) \geq \Delta_{\mathbf{z}_+} = \Delta_{\tilde{\mathbf{x}}_\infty} = D_F(\tilde{\mathbf{x}}_\infty, \tilde{\mathbf{x}}(0)). \tag{5.19}$$

Thus

$$\begin{aligned}
\tilde{\mathbf{x}}_\infty &\in \operatorname*{arg\,min}_{\mathbf{z}\in S_+} D_F(\mathbf{z}, \tilde{\mathbf{x}}(0)) \\
&= \operatorname*{arg\,min}_{\mathbf{z}\in S_+} F(\mathbf{z}) - F(\tilde{\mathbf{x}}(0)) - \langle \nabla F(\tilde{\mathbf{x}}(0)), \mathbf{z} - \tilde{\mathbf{x}}(0)\rangle \\
&= \operatorname*{arg\,min}_{\mathbf{z}\in S_+} F(\mathbf{z}) - \langle \nabla F(\tilde{\mathbf{x}}(0)), \mathbf{z}\rangle \\
&= \operatorname*{arg\,min}_{\mathbf{z}\in S_+} \begin{cases} \sum_{n=1}^N z_n \log(z_n) - z_n - \log(\tilde{x}_n(0)) z_n & \text{if } L = 2 \\ \sum_{n=1}^N -z_n^{\frac{2}{L}} + \frac{2}{L}\tilde{x}_n(0)^{-1+\frac{2}{L}} z_n & \text{if } L > 2 \end{cases} \\
&= \operatorname*{arg\,min}_{\mathbf{z}\in S_+} \begin{cases} \langle \mathbf{z}, \log(\mathbf{z}) - \mathbf{1} - \log(\tilde{\mathbf{x}}(0))\rangle & \text{if } L = 2, \\ \langle \mathbf{z}, \tilde{\mathbf{x}}(0)^{\odot \frac{2}{L} - 1}\rangle - \frac{L}{2}\|\mathbf{z}\|_{2/L}^{2/L} & \text{if } L > 2 \end{cases} \\
&= \operatorname*{arg\,min}_{\mathbf{z}\in S_+} g_{\tilde{\mathbf{x}}(0)}(\mathbf{z}).
\end{aligned}$$

□

We can finally prove the last part of our main result.

Proof of Theorem 5.4.1. By Lemma 5.4.3, the assumption that $\mathbf{x}^{(k)}(0) = \mathbf{x}^{(k')}(0)$, for all $k, k' \in [L]$, implies that $\mathbf{x}^{(k)}(t) = \mathbf{x}^{(k')}(t)$, for all $t \geq 0$. Furthermore, each $\mathbf{x}^{(k)}(t)$ equals $\mathbf{x}(t)$ defined via $\mathbf{x}(0) = \mathbf{x}^{(k)}(0)$ and $\mathbf{x}'(t) = -\frac{1}{L^2}\nabla\mathcal{L}(\mathbf{x})$. By Theorem 5.4.6, the limit $\mathbf{x}_\infty := \lim_{t\to\infty}\mathbf{x}(t)$ exists and $\tilde{\mathbf{x}}_\infty = \mathbf{x}_\infty^{\odot L}$ lies in S_+. Let $\mathbf{z} \in S_+$, i.e., is an admissible solution to (NNLS). The quantitative bound in (5.4) can be deduced by following the steps in [CMR23] since (5.13) and [CMR23, Equation (18)] are identical up to the definition of S_+. □

The Lyapunov analysis developed above can, in principle, be generalized to more sophisticated methods such as stochastic or accelerated gradient flows. See [SBC16, DSKL+22, WRJ21, KS21a, ZWB+21]. In fact, by using slightly different gradient flow dynamics, we can establish an accelerated convergence rate for this dynamical system that solves the problems (NNLS).

5.5 Accelerated gradient flow

In this section, we can generalize the argument given above and prove that an accelerated version of Theorem 5.4.1. Indeed, let $\mathbf{y}(t) = \tilde{\mathbf{x}}(t) + \frac{t}{2}\tilde{\mathbf{x}}'(t)$ and consider the second order ODE:

$$\mathbf{y}'(t) = -\frac{t}{2}\mathbf{y}(t)^{\odot q}\nabla\mathcal{L}(\tilde{\mathbf{x}}(t)).$$
$$\mathbf{y}(0) = \tilde{\mathbf{x}}(0) = \tilde{\mathbf{x}}_0 > \mathbf{0}.$$

First, we observe that the sign of $\mathbf{y}(t)$ does not change. In fact, if $\mathbf{y}(t_0) = 0$, for some $t_0 \geq 0$, then we have $\mathbf{y}(t_0) = \mathbf{y}'(t_0) = 0$. By the Picard-Lindelöf theorem, we have that $\mathbf{y}(t) = 0$, for all t, which contradicts $\mathbf{y}(0) > \mathbf{0}$. Thus, the ODE admits a unique solution $\mathbf{y}(t) > \mathbf{0}$, for all $t \geq 0$.

When $q = 0$, the ODE system above can be used to explain Nesterov's acceleration of gradient descent see, e.g., [SBC16, Section 2], [KBB15, Section 3] and [MJ19, Section 2]. Therefore, Equation (5.20) can be seen as a Riemannian generalization of the accelerated gradient flow.

Proposition 5.5.1. *The flow $\tilde{\mathbf{x}}(t)$ converges with order $O(1/t^2)$. More precisely, it holds that $\mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(z_+) \leq \mathcal{O}(\frac{1}{t^2})$.*

Proof. Consider $\mathbf{z}_+ \in S_+$. We then have that $\mathcal{L}(\mathbf{z}_+) = \frac{1}{2}\|\mathbf{y}_+ - \mathbf{y}\|_2^2$. Now, consider the Lyapunov functional given by

$$\mathcal{E}(t) = t^2(\mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+)) + \frac{4}{2-q}D_F(\mathbf{z}_+, \mathbf{y}(t)).$$

We start by proving that it decreases, i.e., that $\mathcal{E}'(t) \leq 0$.

First we will calculate the derivative $\frac{d}{dt}D_F(\mathbf{z}_+, \mathbf{y}(t))$. Remember from the previous section that the gradient of F is given by $\nabla F(\mathbf{x}) = \ln_q \mathbf{x} + \mathbf{x}^{\odot 1-q}$ and the Hessian is described by $\nabla^2 F(\mathbf{x}) = (2-q)\operatorname{diag}(\mathbf{x}^{\odot -q})$. Using that the Hessian applied at $y(t)$, $\nabla^2 F(\mathbf{y}(t)) = (2-q)\operatorname{diag}(\mathbf{y}(t)^{\odot -q})$, and $\mathbf{y}'(t) = -\frac{t}{2}\mathbf{y}(t)^{\odot q}\nabla L(\tilde{\mathbf{x}})$,

$$\begin{aligned}\partial_t \nabla F(\mathbf{y}(t)) &= \nabla^2 F(\mathbf{y}(t))\mathbf{y}'(t) = (2-q)\operatorname{diag}(\mathbf{y}(t)^{\odot -q})\mathbf{y}'(t)\\ &= -\frac{t}{2}(2-q)\nabla\mathcal{L}(\tilde{\mathbf{x}}(t)),\end{aligned}$$

we obtain

$$\begin{aligned}\frac{d}{dt}D_F(\mathbf{z}_+, \mathbf{y}(t)) &= \frac{d}{dt}\Big(F(\mathbf{z}_+) - F(\mathbf{y}(t)) - \langle\nabla F(\mathbf{y}(t)), \mathbf{z}_+ - \mathbf{y}(t)\rangle\Big)\\ &= -\langle\nabla F(\mathbf{y}(t)), \mathbf{y}'(t)\rangle - \langle\nabla F(\mathbf{y}(t)), -\mathbf{y}'(t)\rangle - \langle\partial_t\nabla F(\mathbf{y}(t)), \mathbf{z}_+ - \mathbf{y}(t)\rangle\\ &= -\langle\partial_t\nabla F(\mathbf{y}(t)), \mathbf{z}_+ - \mathbf{y}(t)\rangle\\ &= \frac{t}{2}(2-q)\langle\nabla\mathcal{L}(\tilde{\mathbf{x}}(t)), \mathbf{z}_+ - \mathbf{y}(t)\rangle. \qquad (5.20)\end{aligned}$$

Thus,

$$\begin{aligned}
\mathcal{E}'(t) &= 2t(\mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+)) + t^2\langle \nabla\mathcal{L}(\tilde{\mathbf{x}}), \tilde{\mathbf{x}}'(t)\rangle + \frac{4}{2-q}\frac{d}{dt}D_F(\mathbf{z}_+, \mathbf{y}(t)).\\
&= 2t(\mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+)) + t^2\langle \nabla\mathcal{L}(\tilde{\mathbf{x}}), \tilde{\mathbf{x}}'(t)\rangle + 2t\langle \nabla\mathcal{L}(\tilde{\mathbf{x}}(t)), \mathbf{z}_+ - \mathbf{y}(t)\rangle\\
&= 2t(\mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+)) + t^2\langle \nabla\mathcal{L}(\tilde{\mathbf{x}}), \tilde{\mathbf{x}}'(t)\rangle + 2t\langle \nabla\mathcal{L}(\tilde{\mathbf{x}}(t)), \mathbf{z}_+ - \tilde{\mathbf{x}}(t) - \frac{t}{2}\tilde{\mathbf{x}}'(t)\rangle\\
&= 2t(\mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+)) + t^2\langle \nabla\mathcal{L}(\tilde{\mathbf{x}}), \tilde{\mathbf{x}}'(t)\rangle + 2t\langle \nabla\mathcal{L}(\tilde{\mathbf{x}}(t)), \mathbf{z}_+ - \tilde{\mathbf{x}}(t) - \frac{t}{2}\tilde{\mathbf{x}}'(t)\rangle\\
&= 2t(\mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+)) + 2t\langle \nabla\mathcal{L}(\tilde{\mathbf{x}}(t)), \mathbf{z}_+ - \tilde{\mathbf{x}}(t)\rangle\\
&= -2tD_{\mathcal{L}}(\mathbf{z}_+, \tilde{\mathbf{x}}(t)) \leq 0.
\end{aligned}$$

We can finally establish the accelerated convergence rate for the dynamics given by Equation (5.20). Indeed, since $\mathcal{E}(t)$ is decreasing. This implies that

$$t^2(\mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(\mathbf{z}_+)) \leq \mathcal{E}(t) \leq \mathcal{E}(0) = \frac{4}{2-q}D_F(\mathbf{z}_+, \tilde{\mathbf{x}}_0).$$

We conclude that $0 \leq \mathcal{L}(\tilde{\mathbf{x}}(t)) - \mathcal{L}(z_+) \leq \mathcal{O}(\frac{1}{t^2})$, as $t \to \infty$.

□

As described above, the discretization of a Riemannian gradient flow can be seen as a mirror descent algorithm [LWLA22, GWS21]. Therefore, the next step, left here as an open problem, is to explore this connection and to develop an extension of the above techniques to the discrete case to establish the respective convergence rates.

Open Problem: How can the gradient flow analysis, which was developed here for the vanilla/accelerated gradient flow applied to overparametrized NNLS, can be extended to discrete methods?

Remark 5.5.2. *The very restrictive form of $\mathbf{x}_0 = \alpha\mathbf{1}$ in Theorem 5.4.1 is only required to get an implicit ℓ_1-bias, which is a classical regularizer for sparse recovery [FR13]. By changing the initialization, one can also achieve other biases like weighted ℓ_1-norms. Below, we show that, for $L = 2$ and any $\mathbf{w} \in (0,1]^N$ with $\|\mathbf{w}\|_\infty = 1$, the initialization vector $\mathbf{x}_0$ defined as*

$$\mathbf{x}_0 = e^{-\frac{1}{2}(\mathbf{1}+\theta\mathbf{w})}$$

will yield an approximate $\ell_{\mathbf{w},1}$-bias in the limit of gradient flow, where $\|\mathbf{z}\|_{\mathbf{w},1} = \|\mathbf{z} \odot \mathbf{w}\|_1$ and $\theta > 0$ has to be chosen sufficiently large depending on the aimed for accuracy.

Weighted ℓ_1-norms have been used in various applications, e.g., polynomial interpolation or sparse polynomial chaos approximation [RW16], [PHD14].

Theorem 5.5.3. *Let $\varepsilon > 0$ and $L = 2$. Under the assumptions of Theorem 5.4.1 with $\mathbf{x}_0 \leq \alpha \leq h(Q_+, \epsilon)$, we get that*

$$\|\tilde{\mathbf{x}}_\infty\|_{\mathbf{w},1} - \min_{\mathbf{z}\in S_+} \|\mathbf{z}\|_{\mathbf{w},1} \leq \epsilon,$$

where $\|\mathbf{z}\|_{\mathbf{w},1} = \|\mathbf{z} \odot \mathbf{w}\|_1$ denotes the weighted ℓ_1-norm for the weights $\mathbf{w} = \delta(-\mathbf{1} - \log(\tilde{\mathbf{x}}_0))$, where $\delta = \frac{1}{\max_{n\in[N]}(-1-\log((\tilde{x}_0)_n)}$.

Proof. For $L = 2$, we obtain from (5.13) that $\tilde{\mathbf{x}}_\infty \in \arg\min_{\mathbf{z}\in S_+} \langle \mathbf{z}, \log(\mathbf{z}) - \mathbf{1} - \log(\tilde{\mathbf{x}}_0)\rangle$. Hence,

$$\langle \tilde{\mathbf{x}}_\infty, \log(\tilde{\mathbf{x}}_\infty) - \mathbf{1} - \log(\tilde{\mathbf{x}}_0)\rangle \leq \langle \mathbf{z}, \log(\mathbf{z}) - \mathbf{1} - \log(\tilde{\mathbf{x}}_0)\rangle,$$

which may be re-stated as

$$\langle \tilde{\mathbf{x}}_\infty - \mathbf{z}, -\mathbf{1} - \log(\tilde{\mathbf{x}}_0)\rangle \leq \langle \mathbf{z}, \log(\mathbf{z})\rangle - \langle \tilde{\mathbf{x}}_\infty, \log(\tilde{\mathbf{x}}_\infty)\rangle.$$

Let $\alpha \in (0, e^{-\frac{1}{2}})$ and assume that $\alpha \geq \mathbf{x}_0 > 0$ (so that $\log(\tilde{\mathbf{x}}_0) < -1$). Note that $1 \geq \mathbf{w} \geq 0$ by assumption. Denote $\|\mathbf{z}\|_{\mathbf{w},1} = \|\mathbf{z} \odot \mathbf{w}\|_1$ to be the weighted ℓ_1-norm. By non-negativity of $\mathbf{w}, \tilde{\mathbf{x}}_\infty, \mathbf{z}$, we get that

$$\|\tilde{\mathbf{x}}_\infty\|_{\mathbf{w},1} - \|\mathbf{z}\|_{\mathbf{w},1} \leq \delta(\langle \mathbf{z}, \log(\mathbf{z})\rangle - \langle \tilde{\mathbf{x}}_\infty, \log(\tilde{\mathbf{x}}_\infty)\rangle).$$

Since $\xi^2 \geq \xi\log(\xi) \geq -e^{-1}$ for $\xi \geq 0$,

$$\|\tilde{\mathbf{x}}_\infty\|_{\mathbf{w},1} - \|\mathbf{z}\|_{\mathbf{w},1} \leq \delta(\|\mathbf{z}\|_2^2 + Ne^{-1}) \leq \delta(\|\mathbf{z}\|_1^2 + Ne^{-1}) \leq \epsilon$$

because $\alpha \leq h(Q_+, \epsilon)$ and $\delta \leq -\frac{1}{1+2\log(\alpha)}$. Take the minimum over all $\mathbf{z} \in S_+$ and we get our conclusion. □

5.6 NNLS and Compressive Sensing

The additional ℓ_1-regularization that is described in Theorem 5.4.1, for small $\alpha > 0$, allows finding NNLS-solutions that are of lower complexity since there is a strong connection between small ℓ_1-norm and (effective) sparsity. In particular, this allows stable reconstruction of (almost) non-negative ground truths if $\mathbf{A}$ is well-behaved, e.g., if $\mathbf{A}$ satisfies standard assumptions for sparse recovery like suitable robust null space and quotient properties [FR13]. The former was defined in Chapter 4, see Definition 4.1.4. Therefore, we start our discussion by defining the latter.

Definition 5.6.1 ([FR13, Definition 11.11])**.** *A measurement matrix* $\mathbf{A} \in \mathbb{R}^{M \times N}$ *is said to possess the* ℓ_1*-quotient property with constant* d *relative to the* ℓ_2*-norm if, for all* $\mathbf{b} \in \mathbb{R}^M$*, there exists* $\mathbf{u} \in \mathbb{R}^N$ *with* $\mathbf{Au} = \mathbf{b}$ *and* $\|\mathbf{u}\|_1 \leq d\sqrt{s_*}\|\mathbf{b}\|_2$*, where* $s_* = M/\ln(eN/M)$*.*

The quotient property was established by [Woj10] in order to study when equality-constrained ℓ_1-minimization would succeed in recovering the underlying ground-truth in a noisy regime, as discussed in Chapter 4. It has a nice geometric interpretation, namely, that the ℓ_2-ball of radius $\frac{\ln(eN/M)}{dM}$ is contained in the image of the ball in the 1-norm, which is the polytope generated by the absolute convex hull of the columns of the measurement operator A. Mathematically speaking, the quotient property says that $\frac{\ln(eN/M)}{dM}B_2^N \subset AB_1^N$.

Note that many types of matrices satisfy the robust NSP of order s and the ℓ_1-quotient property, e.g., Gaussian random matrices, randomly subsampled Fourier-matrices, and randomly subsampled circulant matrices. For instance, a properly scaled matrix with i.i.d. Gaussian entries satisfies both properties with high probability if $M \geq Cs\log(eN/s)$ and $M \leq N/2$, where the constant $C > 0$ only depends on the NSP parameters ρ and τ [FR13]. Combining Theorem 5.4.1 and [CMR23, Theorem 1.4], the following stable recovery result can be derived.

Theorem 5.6.2. *Let* $\mathbf{A} \in \mathbb{R}^{N \times M}$ *be a matrix satisfying the* ℓ_2*-robust null space property with constants* $0 \leq \rho < 1$ *and* $\tau > 0$ *of order* $s := cM/\log(eN/M)$ *and the* ℓ_1*-quotient property with respect to the* ℓ_2*-norm with constant* $d > 0$*.*

For $\mathbf{x}_* \in \mathbb{R}^N$ *and* $\mathbf{y} = \mathbf{Ax}_*$*, recall* $\mathbf{y}_+$ *and* C_+ *from Equations* (5.3) *and* (1.4)*. Decompose* $\mathbf{x}_*$ *into*

$$\mathbf{x}_* = \mathbf{x}_+ - \mathbf{x}_- \tag{5.21}$$

where $\mathbf{Ax}_+ = \mathbf{y}_+$*. For* $\epsilon > 0$ *assume that* $\alpha > 0$ *satisfies*

$$\alpha \leq h(\|\mathbf{x}_+\|_1, \epsilon)$$

for h *defined in* (5.4)*. Then the limit* $\tilde{\mathbf{x}}_\infty$ *defined in Theorem 5.4.1 yields reconstruction error*

$$\|\tilde{\mathbf{x}}_\infty - \mathbf{x}_*\|_2 \leq \frac{C}{\sqrt{s}}(2\sigma_s(\mathbf{x}_+)_{\ell_1} + \epsilon) + \|\mathbf{x}_-\|_2. \tag{5.22}$$

The constants $C, C' > 0$ *only depend on* ρ, τ, c, d*.*

As can be seen from Theorem 5.6.2, our approach to solving (NNLS) is stable with respect to negative entries of the ground truth, i.e., the reconstruction error depends on the sparsity of the positive part $\mathbf{x}_+$ of $\mathbf{x}_*$ and the magnitude of the negative part $\mathbf{x}_-$ of $\mathbf{x}_*$. The experiments we perform in Section 5.8.4 suggest that the established solvers for (NNLS) are less stable under such perturbations.

The observation that under specific assumptions on the measurement operator $\mathbf{A}$, the solution obtained through (NNLS) inherently promotes sparsity without necessitating hyper-parameter tuning can be traced back to seminal works such as [DT05, DT10a], which establish a connection between this subject and the theory of convex polytopes. It is noteworthy that even before the advent of modern sparse recovery theory, as exemplified by compressive sensing, the paper [DJHS92] exploited non-negativity to recover sparse objects, showcasing the early recognition of its significance. Subsequent works studied the uniqueness of positive solutions of underdetermined systems [BEZ08] as well as the extension to low-rank solutions within the positive-definite matrix cone [WXT10]. Furthermore, researchers have established conditions under which NNLS proves effective in recovering sparse vectors even in the presence of noise [KJ17, SJC19, SH11, SH13, Mei13]. Two key concepts in these findings are the null space property, as outlined in Definition 4.1.4, and the $\mathcal{M}_+$ *criterion* [BEZ08].

Definition 5.6.3 ([BEZ08])**.** *Let* $\mathbf{A} \in \mathbb{R}^{M\times N}$. *We say* $\mathbf{A}$ *obeys the* $\mathcal{M}_+$ *criterion with vector* $\mathbf{u}_{\mathcal{M}_+}$ *if there exists* $\mathbf{u}_{\mathcal{M}_+} \in \mathbb{R}^M$ *such that* $\mathbf{A}^\top \mathbf{u}_{\mathcal{M}_+} > 0$*, i.e., if* $\mathbf{A}$ *admits a strictly-positive linear combination of its rows.*

It is important to note that the $\mathcal{M}_+$ criterion, as demonstrated in [WXT10, Theorem 5], serves as a necessary condition for an underdetermined system ($M < N$) to possess a unique non-negative solution. Examples of matrices $\mathbf{A} \in \mathbb{R}^{M\times N}$ satisfying the $\mathcal{M}_+$ criterion include: (i) matrices with independent and identically distributed (i.i.d.) Bernoulli entries, (ii) matrices whose columns can be expressed as independent 1-subgaussian random vectors [SJC19], (iii) matrices whose columns form an outwardly k-neighborly polytope [DT05], and (iv) adjacency matrices of bipartite expander graphs [WXT10]. The following is a recent robustness result for NNLS, relying on the null space property and the $\mathcal{M}_+$ criterion.

Theorem 5.6.4 ([KJ17, Theorem 4])**.** *Suppose that* $\mathbf{A} \in \mathbb{R}^{M\times N}$ *obeys the NSP of order* $s \leq N$ *with constants* $0 < \rho < 1$ *and* $\tau > 0$ *and the* $\mathcal{M}_+$ *criterion with the vector* $\mathbf{u}_{\mathcal{M}_+}$. *Then* $\mathbf{A}$ *allows stable reconstruction of any non-negative s-sparse vector* $\mathbf{x}_*$ *from* $\mathbf{y} = \mathbf{A}\mathbf{x}_* + \boldsymbol{\epsilon}$ *via* (NNLS). *In particular, for any* $1 \leq p \leq q$*, the unique solution* $\mathbf{x}_+$ *of* (NNLS) *is guaranteed to obey*

$$\|\mathbf{x}_+ - \mathbf{x}_*\|_p \leq \frac{C}{s^{1-1/p}}\sigma_s(\mathbf{x})_1 + \frac{D}{s^{1/q-1/p}}(\|\mathbf{u}_{\mathcal{M}_+}\|_2 + \tau)\|\boldsymbol{\epsilon}\|_2, \tag{5.23}$$

where C and D only depend on ρ, the condition number of the diagonal matrix $\operatorname{diag}(\mathbf{A}^\top \mathbf{u}_{\mathcal{M}_+})$*.*

Theorem 5.6.4, which was later generalized to arbitrary ℓ_p-quasinorms [SJC19, Theorem 2], shows that if $\mathbf{A}$ behaves sufficiently well, the solution of (NNLS) stably reconstructs any non-negative s-sparse vector from noisy measurements at least as well as conventional programs for sparse recovery. In particular, neither sparsity regularization nor parameter tuning is required. The program only relies on the geometry imposed by its constraints. The work [SJC19] even showed that (NNLS) outperforms Basis Pursuit Denoising in retrieving a sparse solution from noisy measurements. However, as previously discussed in Section 5.3, the established solvers for (NNLS) unfortunately come with several disadvantages. Let us also mention that measurement operators $\mathbf{A}$ appearing in applications normally do not satisfy the assumptions of Theorem 5.6.4. In such scenarios, an additional sparsity regularization is still needed when solving (NNLS) with these methods. Since our approach does not share the disadvantages of non-scalability or step-size tuning and naturally includes the possibility of adding ℓ_1-regularization, it is well-designed for exploiting the noise robustness of (NNLS) in sparse recovery.

5.7 Related Works on NNLS

The results presented in Section 5.4 establish a connection between two distinct and previously unrelated areas of mathematical research: the long-standing question of efficient methods for solving NNLS and the more recent investigation into the implicit bias of gradient descent. In particular, it was shown in Theorem 5.4.1 that it is possible to use ideas from overparametrization to solve a constrained optimization problem. Prior to delving into a numerical evaluation of our theory, it is appropriate to provide a concise overview of the existing literature about the solution of the non-negative least squares.

The first algorithm proposed to solve (NNLS) appeared in 1974 in the book [LH95, Chapter 23], where its finite convergence was proved, and a Fortran routine was presented.[1] Like the previous papers [Sto71, GS70, GM73] it builds upon the solution of linear systems. Similar to the simplex method, the algorithm is an active-set algorithm that iteratively sets parts of the variables to zero to identify the active constraints and solve the unconstrained least squares sub-problem for this active set of constraints. It is still, arguably, the most famous method for solving (NNLS), and several improvements have been proposed in a series of follow-up papers [BDJ97], [VBK04], [MFLS17], [LD11], [DDM21]. Nevertheless, a limitation of this approach arises from its reliance on the normal equations, rendering it impractical for ill-conditioned or large-scale problems. Furthermore,

[1] This algorithm is the standard one implemented in many languages: *optimize.nnls* in the SciPy package, *nnls* in R, *lsqnonneq* in MATLAB and *nnls.jl* in Julia.

current theoretical guarantees for the algorithm and its modifications have yet to surpass convergence with a finite number of steps [LH95, Chapter 23], [DDM21, Theorem 3], leaving room for improvement in terms of theoretical guarantees.

Another line of research has been developing projected gradient methods for solving (NNLS), which come with linear convergence guarantees [KSD13], [Pol15] [Lin07]. In contrast to active set and interior point methods, projected gradient methods do not require solving a linear system of equations at each step, making them more scalable for high-dimensional problems. However, the effectiveness of these methods hinges on the appropriate choice of step size, and as previously emphasized in Section 5.3, the projection step may render established acceleration techniques for standard gradient descent ineffective.

5.8 Numerical experiments

In the last part of this chapter, we will finally turn to a numerical evaluation of our theoretical insights. We compare the following six methods for solving NNLS here:

- **GD-nL:** Vanilla gradient descent applied to $\mathcal{L}_{\text{over}}$ in (5.2) with n layers, for $n \in \mathbb{N}$. This is the discretized version of the gradient flow we considered in Section 5.4. As initialization we use $\alpha\mathbf{1}$, for $\alpha > 0$.

- **SGD-nL:** Stochastic gradient descent applied to $\mathcal{L}_{\text{over}}$ in (5.2) with n layers, for $n \in \mathbb{N}$. A probabilistic variant of **GD-nL**. In the experiments we use $M/10$ as batch-size for **SGD-nL** and initialize with $\alpha\mathbf{1}$, for $\alpha > 0$.

- **LH-NNLS:** The standard Python NNLS-solver *scipy.optimize.nnls*, which is an active set method and is based on the original Lawson-Hanson method [LH95]. It is not scalable to high dimensions since it requires solving linear systems in each iteration. (An accelerated version of **LH-NNLS** is provided in [BDJ97]. Since both methods produce the same outcome in our experiments, we only provide the results for **LH-NNLS**.)

- **TNT-NN:** An alternative but more recent active set method that heuristically works well and dramatically improves over **LH-NNLS** in performance [MFLS17]. We used the recent Python implementation available at `https://github.com/gdcs92/pytntnn`.

- **CVX-NNLS:** Solving the quadratic formulation of (NNLS) described in (5.1) via ADMM. We use the Python-embedded modeling language CVX [2] which, in turn, uses the solver OSQP [3] for this task.
- **PGD**: Projected gradient descent for NNLS as described in [Pol15].

In particular, we are interested in quantifying the impact of initialization on sparsity regularization, in illustrating the impact of the number of layers on reconstruction performance, in the practicality of state-of-the-art step-size tuning procedures for **GD-nL**, and in the stability of **GD-nL** against negative perturbations in comparison to established methods. At the end of the chapter, we also provide additional large-scale experiments and benchmark tests with PGD.

5.8.1 Initialization

In the first experiment, we validate the ℓ_1-norm regularization that, according to the second part of Theorem 5.4.1, can be induced by using a small initialization for **GD-nL**. For $M = 10$ and $N = 50$, we draw a random Gaussian matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$, create a 3-sparse ground-truth $\mathbf{x} \in \mathbb{R}^N$, and set $\mathbf{y} = \mathbf{Ax}$. Figure 5.1 depicts the ℓ_1-norm of the limits of **GD-2L** and **GD-3L**, for $10^{-3} \leq \alpha \leq 10^{-1}$ and constant step-size $\eta = 10^{-2}$. As a benchmark, the ℓ_1 minimizer among all feasible solutions is computed via basis pursuit (BP). Figure 5.1 shows that **GD-2L** and **GD-3L** converge to ℓ_1-norm minimizers if α is sufficiently small. As predicted by the second part of Theorem 5.4.1, the requirements on α to allow such regularization are milder for the 3-layer case **GD-3L**. Finally, neither **LH-NNLS** nor **CVX-NNLS** reaches ℓ_1-minimality. The matrix $\mathbf{A}$, although sufficiently well-behaved for sparse recovery in general, does not guarantee the uniqueness of the NNLS solution here.

5.8.2 Number of layers

In the second experiment, we take a closer look at the reconstruction behavior of **GD-nL**. In particular, we compare how different ground truth entries are approximated over time for $n = 2$ and $n = 3$ layers. For initialization magnitude and step-size we choose $\alpha = 10^{-2}$ and $\eta = 10^{-2}$. We set $M = 30$ and $N = 50$, draw a random Gaussian matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$, create a 3-sparse non-negative ground-truth $\mathbf{x} \in \mathbb{R}^N$, and set $\mathbf{y} = \mathbf{Ax}$. Note that $\mathbf{A}$ satisfies the assumptions of Theorem 5.6.2. Figure 5.2 depicts the entry-wise error between the three non-zero ground-truth entries and the corresponding entries

[2] `https://www.cvxpy.org/`

[3] `https://osqp.org/docs/solver/index.html`

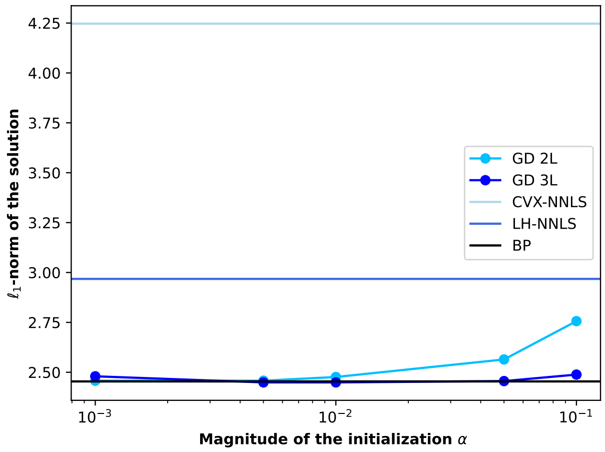


Figure 5.1: Influence of the initialization on the ℓ_1-norm of the solution.

of the iterates of **GD-2L** and **GD-3L**. As already observed in previous related works, we see that a deeper factorization leads to sharper error transitions that occur later and that more dominant entries are recovered faster than the rest. Interestingly, there occurs some kind of overshooting in the dominant entries. The dark blue and purple curves suggest that **GD-nL** does not monotonically decrease the error in all components but rather concentrates heavily on the leading component(s) in the beginning and only starts distributing the error over time. In this way, **GD-nL** could be interpreted as a self-correcting greedy method.

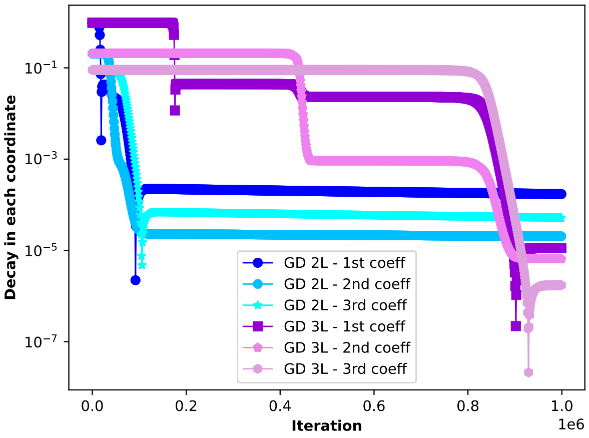


Figure 5.2: Influence of the number of layers on the decay of the solution.

5.8.3 Different stepsizes

In the third experiment, we compare the convergence rates of **GD-nL** and **SGD-nL** for various choices of step size. Apart from using a constant step-size $\eta > 0$, we also consider

Nesterov acceleration [Nes83] and the BB stepsize rules [Fle05, Ray97]. Figure 5.5 shows the decay in training error, i.e., objective value over time in two different settings: the dense case, i.e., we have a quadratic system with $M = N = 50$ and a dense ground-truth, and the sparse case, i.e., we have an underdetermined system with $M = 30$, $N = 50$, and a 3-sparse ground-truth. In both settings, $\mathbf{A}$ has Gaussian entries. As Figures 5.3 and 5.4 show, advanced step-size choices notably improve the gradient methods' convergence rate. Moreover, **GD-nL** seems to profit more from the acceleration than **SGD-nL**.

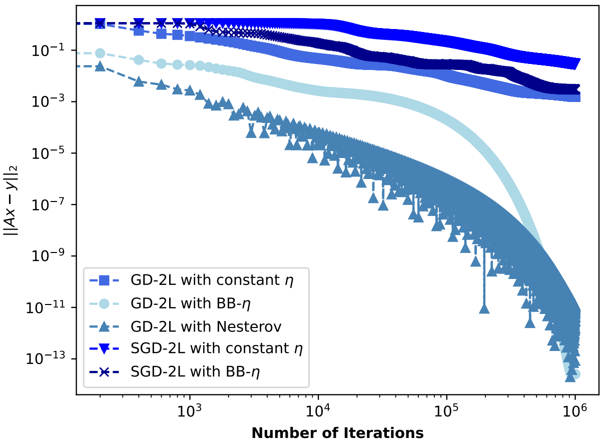


Figure 5.3: Convergence rate for various choices of step-size in the dense case, see Section 5.8.3.

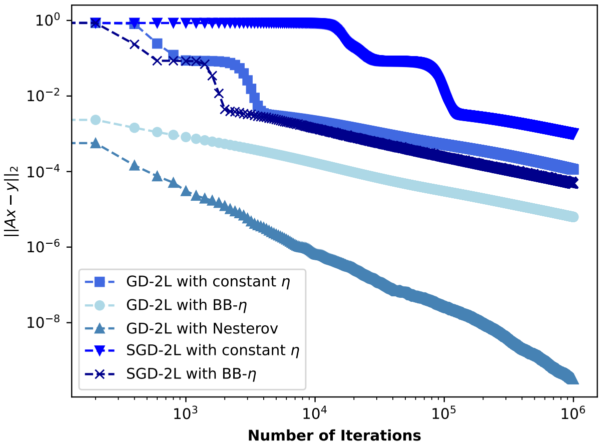
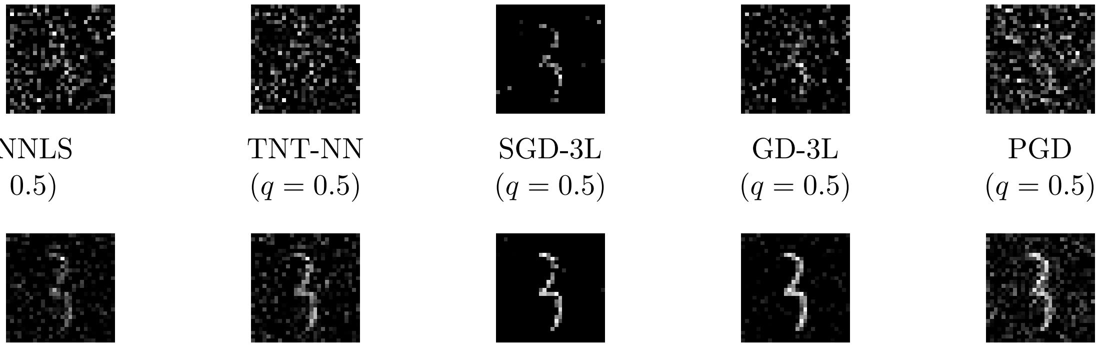


Figure 5.4: Convergence rate for various choices of step-size in the sparse case, see Section 5.8.3.

Figure 5.5: Convergence rate for various choices of step-size, see Section 5.8.3.



HL-NNLS ($q = 0.5$) TNT-NN ($q = 0.5$) SGD-3L ($q = 0.5$) GD-3L ($q = 0.5$) PGD ($q = 0.5$)

HL-NNLS ($q = 0.05$) TNT-NN ($q = 0.05$) SGD-3L ($q = 0.05$) GD-3L ($q = 0.05$) PGD ($q = 0.05$)

Figure 5.6: Illustration of the MNIST reconstruction, see Section 5.8.4.

5.8.4 Stability with Negative Entries

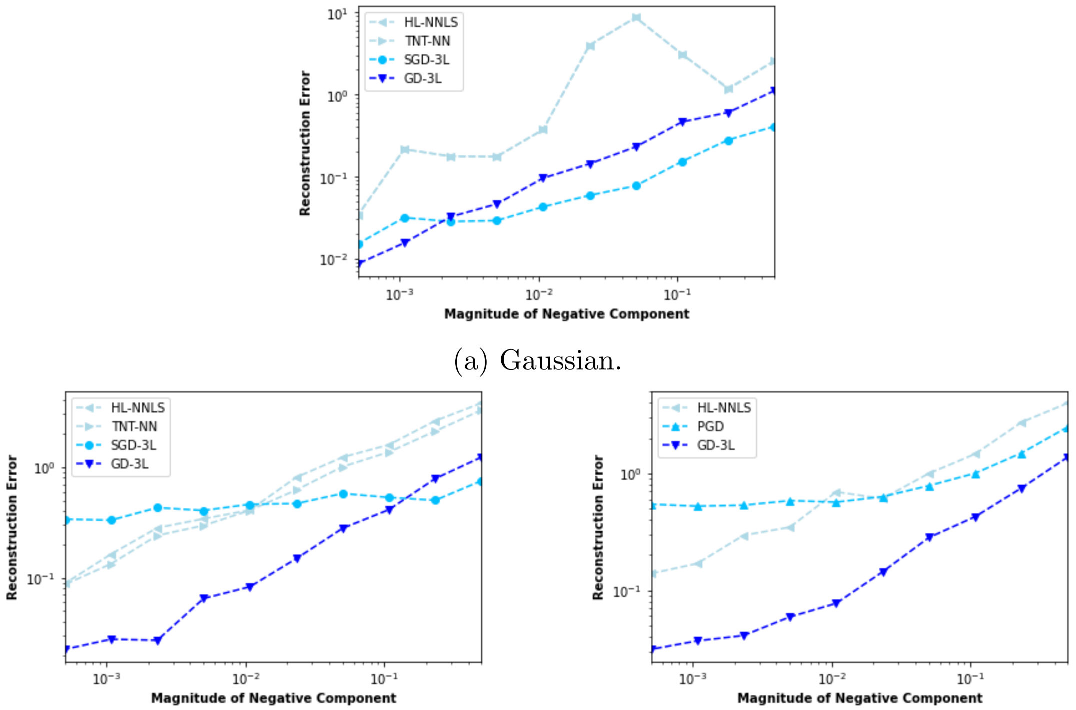
Now, we examine the robustness of the proposed methods in the context of recovering perturbed signals that may not strictly adhere to non-negativity constraints. We explore two distinct scenarios: the first involves generating random signals, while the second utilizes signals sourced from the MNIST data set, which represent more realistic data. In both scenarios we set $\alpha = 10^{-2}$ and $\eta = 10^{-2}$ for **GD-nL** and **SGD-nL**. It is noteworthy that in all tested instances **GD-nL** and **SGD-nL** outperform the established methods in reconstruction quality, cf. Figure 5.6. Whereas **HL-NNLS** and **TNT-NN** are designed to retrieve only non-negative signals, the gradient-based methods can stably deal with vectors with small negative components by not using explicit constraints.

5.8.4.1 Gaussian signals

Let $\mathbf{A} \in \mathbb{R}^{M\times N}$ be a random Gaussian matrix, where $M = 30$ and $N = 50$. We pick a 3-sparse non-negative vector $\mathbf{x}_+ \in \mathbb{R}_+^N$ at random. We, furthermore, define a noise vector $\mathbf{x}_- \in \mathbb{R}_+^N$ that is 0 on the support of $\mathbf{x}_+$ and has positive Gaussian entries everywhere else. For $q \in [0, 1]$, we scale $\mathbf{x}_+, \mathbf{x}_-$ such that

$$\|\mathbf{x}_+\|_2^2 = 1 - q \quad \text{and} \quad \|\mathbf{x}_-\|_2^2 = q. \tag{5.24}$$

The perturbed signal is then given by $\mathbf{x} = \mathbf{x}_+ - \mathbf{x}_-$, and q regulates the negative corruption. We regard the copy of $\mathbf{x}_+$ scaled to ℓ_2-norm with norm equal to $(1 - q)$ as ground truth and, by abuse of notation, also refer to it as $\mathbf{x}_+$. The corresponding measurements are given as $\mathbf{y} = \mathbf{A}\mathbf{x} = \mathbf{A}(\mathbf{x}_+ - \mathbf{x}_-)$.



(a) Gaussian.

(b) MNIST.

(c) MNIST with PGD.

Figure 5.7: Comparison of stability, see Section 5.8.4.

5.8.4.2 MNIST signals

Again $\mathbf{A} \in \mathbb{R}^{M \times N}$ is a random Gaussian matrix, where now $M = 300$ and $N = 28^2 = 784$. We take $\mathbf{x}_+$ to be the original MNIST image (number three) and define a corrupted signal $\mathbf{x}$ including negative Gaussian noise $\mathbf{x}_-$ as in (5.24), for $q \in [0, 1]$. Note that our ground truth is again a re-scaled version of the original image that has ℓ_2-norm equal to $(1 - q)$. Figures 5.7a, 5.7b and 5.7c compare the reconstruction error $\|\hat{\mathbf{x}} - \mathbf{x}_+\|_2$ of **HL-NNLS**, **TNT-NN**, **GD-3L**, and **SGD-3L** over various choices of q. Here we set $\alpha = 10^{-2}$ and $\eta = 10^{-2}$. These figures clearly show that the gradient descent-based methods outperform the established NNLS solvers. Only for small negative noise levels and MNIST data **SGD-3L** yields worse results than **HL-NNLS** and **TNT-NN**. As can be seen from Figure 5.6, this worse error is mainly caused by incorrect values on the support of $\mathbf{x}_+$. Visually, even for small q, the MNIST reconstruction of **SGD-3L** is far better than the one of **HL-NNLS** and **TNT-NN**.

The experiment also reveals two interesting points. First, whereas it has numerically been observed in [PPVF21] that, compared to gradient descent, stochastic gradient descent reduces the generalization (resp. approximation) error if measured in ℓ_2-norm, we observe this (in the case of NNLS) only for large values of q, i.e., strong negative perturbations. For small values of q, Figure 5.5 rather suggests that **GD-3L** outperforms **SGD-3L**. Second, for all values of q, the solutions computed by **SGD-3L** are visually closer to the ground truth than the ones computed by **GD-3L**, cf. Figure 5.6. This suggests that, even in the simple context of sparse recovery, (i) the ℓ_2-norm might not be the appropriate measure for the generalization error and (ii) the stochasticity in **SGD-3L** apparently improves the generalization quality. Formalizing and proving this observation is an appealing topic for future research.

5.8.5 Large-scale NNLS experiments

In this section, we provide additional empirical evidence for our theoretical claims. In particular, (i) we illustrate the performance of our method in an over-determined large-scale NNLS problem and (ii) we compare our method to projected gradient descent (PGD) [Pol15] in terms of the number of iterations, convergence rate, and running time since it is known that PGD converges linearly to the global minimizer [ABS13, Pol15] and it is a numerically efficient method due to the fast calculation of the projection step.

In this first experiment, cf. Figure 5.8, we illustrate the performance of our method for an overdetermined NNLS problem on a larger scale. The matrix $\mathbf{A} \in \mathbb{R}^{2048\times 1024}$ is standard Gaussian and, for a Gaussian random vector $\mathbf{x} \in \mathbb{R}^{1024}$, $\mathbf{y}$ is created as a perturbed version of $\mathbf{Ax}$ such that $\mathbf{y}$ is not in the range of $\mathbf{A}$. Furthermore, we use $\mathbf{x}_0 = \mathbf{1}$ as a generic initialization. Figure 5.8 shows the error $\|\mathbf{A}\mathbf{x}^{\odot L} - \mathbf{y}\|_2$ over the iterations of gradient descent. As predicted by Theorem 5.4.1, our method converges to a solution that solves the NNLS problem (compare the error to the benchmark given by the Lawson-Hanson algorithm).

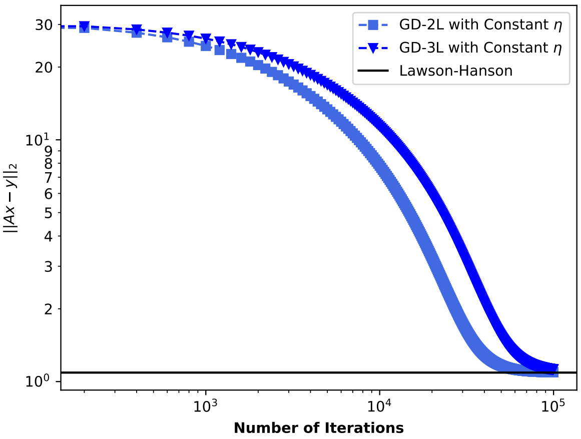


Figure 5.8: Large-scale NNLS problem, see Section 5.8.5.

5.8.5.1 Comparison of the averaged number of iterations

In the second experiment, cf. Figure 5.9, we illustrate the necessary amount of iterations to reach the precision 10^{-3}. The matrix $\mathbf{A} \in \mathbb{R}^{512\times 1024}$ is a normalized 8-sparse random vector $\mathbf{x} \in \mathbb{R}^{1024}$ with Gaussian entries that are made positive by considering only the absolute values of each entry, we generate $\mathbf{y}$ simply by doing $\mathbf{y} = \mathbf{Ax}$, which means that we would expect the methods to converge to zero. We use $\mathbf{x}_0 = \mathbf{1}$ as a generic initialization. As for the stepsize, for the PGD algorithm, it was chosen as $\eta = 1/L = 1/||\mathbf{A}^T\mathbf{A}||_2$, the Lipschitz constant of the gradient. Also, for overparametrized GD-2L, overparametrized GD-3L, and overparametrized SGD-2L, we choose a small but constant stepsize $\eta = 0.02$. Moreover, we compare all of these methods against the GD-2L and GD-3L with the Barzilai-Borwein (BB) stepsize $\eta = \frac{||\mathbf{x}_t - \mathbf{x}_{t-1}||_2^2}{||\mathbf{A}(\mathbf{x}_t - \mathbf{x}_{t-1})||_2^2}$. See [BB88] for more details. We ran it 100 times and took the median of the results in order to avoid outliers.

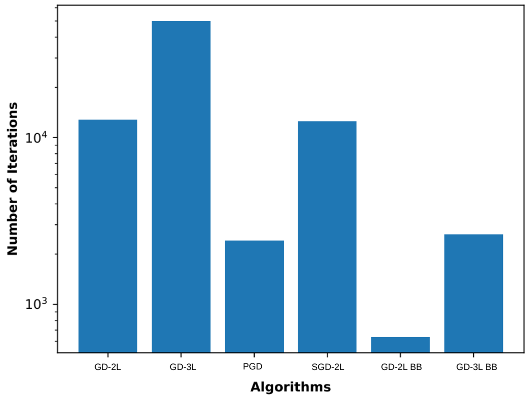


Figure 5.9: Comparison of the number of iterations, cf. Section 5.8.5.1.

As expected, PGD shines as it is extremely fast for the convex formulation of NNLS and reaches good progress in the first 10^3 iterations. Our non-convex formulation, on the other hand, needs more iterations to reach a very good precision. Nevertheless, by combining it with the BB stepsize, we outperform the number of iterations necessary for a certain precision as compared to PGD. This illustrates the potential of further acceleration schemes that could leverage overparametrized formulations.

5.8.5.2 Running time comparison

In this experiment, we show a table with the necessary average running time, in seconds and averaged over 25 realizations, to run 10^6 iteration of the NNLS problem for normalized squared Gaussian matrices. Here, we use the Lipschitz constant of the gradient as the stepsize for PGD, as described above. We also employ the same strategy and use the (iteration-dependent) Lipschitz constant for our method, given by

$$\nabla^2 \mathcal{L}(\mathbf{x}) = L^2 \mathbf{A}^T \mathbf{A} \odot [\mathbf{x}^{\odot L-1} [\mathbf{x}^{\odot L-1}]^T] + L(L-1) \operatorname{diag}\{\mathbf{A}^T (\mathbf{A}\mathbf{x}^{\odot L} - \mathbf{y}) \odot \mathbf{x}^{\odot L-2})\},$$

where $\mathcal{L}$ is defined in (5.10), i.e., $\eta = 1/||\nabla^2 \mathcal{L}(\mathbf{x})||_2$. Since this computation should be done at each iteration, we only precompute the stepsize for the overparametrized algorithm GD-2L and GD-3L at every 1000 iterations to make it more efficient. We use $\mathbf{x}_0 = 0.02\mathbf{1}$ as a generic initialization. For both methods, we precompute $\mathbf{Q} = \mathbf{A}^T \mathbf{A}$ and $\mathbf{p} = \mathbf{A}^T \mathbf{y}$ at the beginning of the simulation. As can be seen from the table, PGD is extremely efficient. Interestingly, our method has a comparable running time, and it even outperforms PGD for large dimensions.

	256 × 256	512 × 512	1024 × 1024	2048 × 2048	4096 × 4096
PGD	4.12	5.05	34.08	144.03	657.87
GD-2L	4.26	6.65	24.16	89.07	403.13
GD-3L	7.04	9.58	24.50	98.81	423.31

Table 5.1: Comparison of running time, cf. Section 5.8.5.2.

5.8.5.3 Convergence rate

In the last experiment, we compare the convergence rates of **GD-2L**, **GD-3L**, and PGD. To this end, we create an NNLS instance with Gaussian $\mathbf{A} \in \mathbb{R}^{256\times256}$, $\mathbf{y} = \mathbf{A}\mathbf{x}$, for $\mathbf{x} \in \mathbb{R}^{256}$ drawn from a Gaussian distribution and entries taken in absolute value, and run all three methods. As Figures 5.10 and 5.11 show, the convergence rates of all three algorithms are the same with the observation that the PGD algorithm has a boosted start provided by the initial projection step.

5.9 Chapter Conclusion

In this chapter, we presented the fourth and last idea related to the methods of least squares method of this thesis. We discussed the phenomenon of implicit bias on overparametrized models, and we have shown that, due to its implicit bias, vanilla gradient descent is a reliable and scalable solver for the decade-old problem of NNLS that, in contrast to some of the established methods, comes with strong theoretical guarantees.

Whereas most works on the implicit bias of gradient descent focus on explaining the still mysterious success of deep learning, we here took a different path and used the implicit bias to solve a constrained optimization problem. We also showed that our

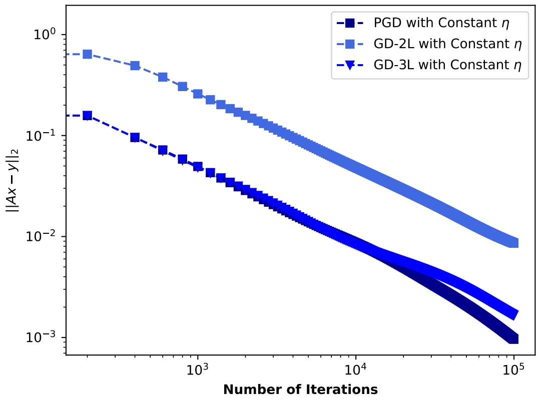


Figure 5.10: Comparison of convergence rate in the dense, non-sparse case, cf. Section 5.8.5.3. With constant step size.

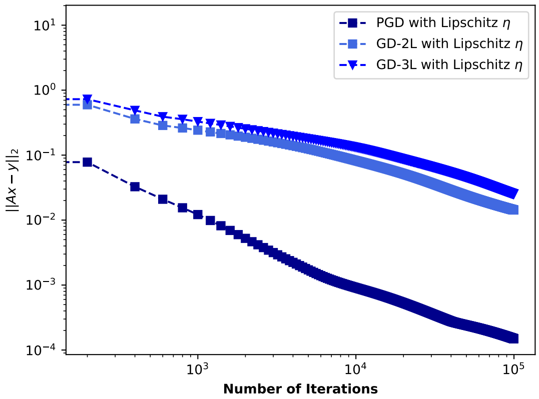


Figure 5.11: Comparison of convergence rate in the dense, non-sparse case, cf. Section 5.8.5.3. With oracle stepsize.

solution has a Riemmanian interpretation, which opens new ways to think about the overparametrized models and the analysis of deep learning architectures. In particular, by using, for example, techniques related to extrapolating iterates, it should be possible to develop Riemannian nonlinear accelerated methods that address more general constraints in an accelerated way [HMJG23]. We leave this as an open problem with potential for future exploration and development.

Open Problem: How to leverage the current overparametrized methods developed in this chapter with accelerated Riemannian methods?

One limitation of the presented analysis resides in its reliance on a gradient flow argument with infinitely small stepsize, whereas the discrete nature of the proposed algorithm calls for the development of a convergence theory specifically tailored to gradient descent. Moreover, a connection between those results and generalization aspects of (linear) neural networks needs a stepsize analysis since the performance of machine learning algorithms is highly sensitive to the choice of stepsize. These results are currently in preparation and will serve as the focus of one of the sections of an upcoming version of our paper before submission.

Another intriguing avenue for exploration involves extending the positivity constraint beyond NNLS to encompass polyhedral or more general convex constraints. This extension can be achieved by devising functions, such as the q-logarithm, that effectively capture the geometry of the constraint set through the Bregman divergence in conjunction with the overparametrized regime. A comprehensive investigation of this topic is also being undertaken in preparation for a forthcoming publication.

A final important aspect is the benefit of stochasticity [PPVF21]. For algorithms like stochastic gradient descent, it has been shown that introducing randomness induces better generalization properties than that of gradient flow. Therefore, given the importance of such a foundational algorithm for machine learning, as a future work, it would be interesting to understand the role of the stepsize and how SGD can be used to retrieve a solution to a constrained problem beyond the numerical simulations presented here.

In any case, we see much potential in exploiting this phenomenon in other contexts and more classical problems as well. With the present chapter, we hope to initiate further research and discussion in this direction.

Chapter 6

Conclusion

What's past is prologue.
The Tempest, William Shakespeare, Act II, Scene I

This thesis has investigated how a simple and old idea, least squares, can be used as the central toolkit for the understanding and development of modern algorithms that solve data science and machine learning problems. By investigating iteratively reweighted least squares, we observed that these algorithms can be seen as a majorization-minimization strategy applied to a surrogate (non-)convex function. Moreover, by analyzing a gradient flow algorithm applied to a non-convex least squares formulation, we developed a scalable method that tackled constrained optimization problems in the case of non-negative constraints.

As the statistician Stephen Stigler wrote in one of the most important books about the history of statistics, *"The method of least squares was the dominant theme – the leitmotif – of nineteenth-century mathematical statistics. In several respects, it was to statistics what the calculus had been to mathematics a century earlier."* [Sti86, Page 11]. Thus, building upon this line of reasoning, we envision this thesis as an incremental step towards advancing simple, scalable, statistically efficient but provable approaches that are grounded in the method of least squares method for tackling large-scale machine learning challenges. Moreover, we advocate for comprehending intricate and complex phenomena in data science and machine learning by initially grasping the linear regression scenario and establishing connections between the complex phenomenon and a simplified least squares framework. Currently, this paradigm is in a state of active development, as demonstrated by some recent papers that, for example, try to understand the *transformer architecture* in deep learning [ASA+22, ZFB23]. We anticipate that this is just the beginning of several theoretical developments, and we can expect numerous comparable advancements to arise in the future.

We now summarize the main results of this thesis and give an overview of some possible future research directions. In this thesis, we developed:

- **A proof:** The first proof that IRLS can converge globally with a linear rate for the sparse recovery problem. In particular, we solved an open problem in the theory of IRLS for sparse recovery.

- **An algorithm:** A scalable second-order algorithm with provable guarantees that is able to retrieve highly ill-conditioned matrices for the provably optimal number of measurements and is highly competitive with state-of-the-art methods.

- **An extension of a theory:** A theory for IRLS with global linear convergence rate in the noise-blind case where the objective function is given by the sum of two non-smooth terms allowing its use for problems when the noise level is not known and when it is hard to perform hyperparameter tuning.

- **An idea:** A connection between overparametrization and the solution of constrained convex problems based on the trade-off between side constraints in optimization problems with the complexity of the optimization landscape.

Furthermore, throughout our discussions, we have conducted numerous numerical experiments and identified several open problems and intriguing directions worth exploring. These aspects have added depth to our research and have laid the groundwork for further investigation.

6.1 Future directions

We conclude this thesis with two future lines of research that we believe to be interesting.

6.1.1 Majorization-minimization

As we discussed in this thesis, the majorization-minimization strategy plays a crucial role in the development of efficient algorithms for ill-conditioned large-scale problems. Recently, this type of strategy was employed to solve geodesically-convex problems in Riemannian optimization [SW22] to get non-asymptotic convergence rates. Can the IRLS-type of majorization-minimization strategy develop new algorithms for geodesically convex functions with better convergence rates?

6.1.2 Overparametrization

In the study of overparametrized least squares, we studied the implicit bias of gradient descent when the overparametrization is given by $\tilde{\mathbf{x}} := \bigodot_{k\in[L]} \mathbf{x}^{(k)} = \mathbf{x}^{(1)} \odot \cdots \odot \mathbf{x}^{(L)}$, which can be seen as a simple linear neural network. It would be interesting to investigate what happens in the case of a wider (and more realistic) set of architectures that involve non-linearities. As an example, the recent study [LJ22] investigated the implicit bias observed in nonhomogeneous feedforward networks. What are the implications of a more general overparametrization when combined with the bias of the gradient descent for the solution of constrained optimization problems?

> *"We have not succeeded in answering all our problems – indeed we sometimes feel we have not completely answered any of them. The answers we have found have only served to raise a whole set of new questions. In some ways, we feel that we are as confused as ever, but we think we are confused on a higher level and about more important things. So this report does not purport to give final answers, or to claim that we now "know how to do it". We see more need for revision than ever. But we are doing better than we did. And this is a progress report, rendered with humility because of the unsolved problems we see now which we could not see before."*
>
> *Earl Kelley*, in [Kel51]

Bibliography

[ABG07] P-A Absil, Christopher G Baker, and Kyle A Gallivan. Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics*, 7:303–330, 2007.

[ABH19] Aleksandr Aravkin, James V Burke, and Daiwei He. IRLS for Sparse Recovery Revisited: Examples of Failure and a Remedy. *arXiv preprint arXiv:1910.07095*, 2019.

[ABN21] Tal Amir, Ronen Basri, and Boaz Nadler. The trimmed lasso: Sparse recovery guarantees and practical optimization by the generalized soft-min penalty. *SIAM journal on mathematics of data science*, 3(3):900–929, 2021.

[ABRS10] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-lojasiewicz inequality. *Mathematics of operations research*, 35(2):438–457, 2010.

[ABS13] Hedy Attouch, Jerome Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods. *Mathematical Programming*, 137:91–129, 2013.

[ACH18] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pages 244–253, 2018.

[ACHL19] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7413–7424, 2019.

[ADH$^+$19] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019.

[AH15] Khurrum Aftab and Richard Hartley. Convergence of iteratively re-weighted least squares to robust m-estimators. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 480–487. IEEE, 2015.

[AH21] Ben Adcock and Anders C Hansen. *Compressive Imaging: Structure, Sampling, Learning.* Cambridge University Press, 2021.

[AIG06] Marc Allain, Jerome Idier, and Yves Goussard. On global and local convergence of half-quadratic algorithms. *IEEE Trans. Image Process.*, 15(5):1130–1142, 2006.

[Ald98] John Aldrich. Doing least squares: perspectives from gauss and yule. *International Statistical Review/Revue Internationale de Statistique*, pages 61–81, 1998.

[ALMT14] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[Ama16] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[AMN$^+$21] Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pages 468–477. PMLR, 2021.

[Ant10] Anestis Antoniadis. Comments on: ℓ_1-Penalization for Mixture Regression Models by N. Stadler, P. Bühlmann and S. van de Geer. *TEST*, 19(2):257–2588, 2010.

[AO15] P-A Absil and Ivan V Oseledets. Low-rank retractions: a survey and new results. *Computational Optimization and Applications*, 62(1):5–29, 2015.

[APS19] Deeksha Adil, Richard Peng, and Sushant Sachdeva. Fast, provably convergent irls algorithm for p-norm linear regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 14189–14200, 2019.

[ApS22] MOSEK ApS. *The MOSEK Optimizer API for Python 10.0.45. Version 10.0.45*, 2022.

[AS21] Kyriakos Axiotis and Maxim Sviridenko. Sparse convex optimization via adaptively regularized hard thresholding. *The Journal of Machine Learning Research*, 22(1):5421–5467, 2021.

[ASA$^+$22] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

[AW02] Assyr Abdulle and Gerhard Wanner. 200 years of least squares method. *Elemente der Mathematik*, 57:45–60, 2002.

[AZLS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[BA15] Nicolas Boumal and P-A Absil. Low-rank matrix completion via preconditioned optimization on the grassmann manifold. *Linear Algebra and its Applications*, 475:200–239, 2015.

[Bar19] Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4331–4339, 2019.

[BB88] Jonathan Barzilai and Jonathan M Borwein. Two-point step size gradient methods. *IMA journal of numerical analysis*, 8(1):141–148, 1988.

[BBPB13] Demba Ba, Behtash Babadi, Patrick L Purdon, and Emery N. Brown. Convergence and stability of iteratively re-weighted least squares algorithms. *IEEE Transactions on Signal Processing*, 62(1):183–195, 2013.

[BBS94] Charles Sidney Burrus, Jose Antonio Barreto, and Ivan W Selesnick. Iterative reweighted least-squares design of fir filters. *IEEE Transactions on Signal Processing*, 42(11):2926–2936, 1994.

[BBZ04] Alexander M Bronstein, Michael M Bronstein, and Michael Zibulevsky. Blind source separation using block-coordinate relative newton method. *Signal processing*, 84(8):1447–1459, 2004.

[BCG11] Stephen Becker, Emmanuel J. Candès, and Michael C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput.*, 3(3):165–218, 2011.

[BCM17] Dimitris Bertsimas, Martin S Copenhaver, and Rahul Mazumder. The trimmed lasso: Sparsity and robustness. *arXiv preprint arXiv:1708.04527*, 2017.

[BCMN14] Afonso S Bandeira, Jameson Cahill, Dustin G Mixon, and Aaron A Nelson. Saving phase: Injectivity and stability for phase retrieval. *Applied and Computational Harmonic Analysis*, 37(1):106–125, 2014.

[BCT11] Jeffrey D Blanchard, Coralia Cartis, and Jared Tanner. Compressed sensing: How sharp is the restricted isometry property? *SIAM review*, 53(1):105–125, 2011.

[BCW11] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 12 2011.

[BCZ23] Fengmiao Bian, Jian-Feng Cai, and Rui Zhang. A preconditioned riemannian gradient descent algorithm for low-rank matrix recovery. *arXiv preprint arXiv:2305.02543*, 2023.

[BDDW08] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28:253–263, 2008.

[BDJ97] Rasmus Bro and Sijmen De Jong. A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 11(5):393–401, 1997.

[BDL07] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.*, 17(4):1205–1223, 2007.

[BDMS09] Nicolai Bissantz, Lutz Dümbgen, Axel Munk, and Bernd Stratmann. Convergence analysis of generalized iteratively reweighted least squares algorithms on convex function spaces. *SIAM Journal on Optimization*, 19(4):1828–1845, 2009.

[Bec15a] Amir Beck. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM Journal on Optimization*, 25(1):185–209, 2015.

[Bec15b] Amir Beck. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM J. Optim.*, 25(1):185–209, 2015.

[Bec17] Amir Beck. *First-order methods in optimization.* SIAM, 2017.

[Bel21] Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

[BEZ08] Alfred M Bruckstein, Michael Elad, and Michael Zibulevsky. On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations. *IEEE Transactions on Information Theory*, 54(11):4813–4820, 2008.

[BGLS95] J Frédéric Bonnans, J Ch Gilbert, Claude Lemaréchal, and Claudia A Sagastizábal. A family of variable metric proximal methods. *Mathematical Programming*, 68:15–47, 1995.

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias-variance tradeoff. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[BHX20] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[BIM22] Gilles Bareilles, Franck Iutzeler, and Jérôme Malick. Newton acceleration on manifolds identified by proximal gradient methods. *Mathematical Programming*, pages 1–34, 2022.

[Bjö96] Åke Björck. *Numerical Methods for Least Squares Problems.* Society for Industrial and Applied Mathematics, 1996.

[BKM19] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[BKZ$^+$21] Jose Blanchet, Yang Kang, Fan Zhang, Fei He, and Zhangyi Hu. Doubly robust data-driven distributionally robust optimization. *Applied Modeling Techniques and Data Analysis 1: Computational Data Analysis Methods and Tools*, 7:75–90, 2021.

[BL21] Yingjie Bi and Javad Lavaei. On the absence of spurious local minima in nonlinear low-rank matrix recovery problems. In *International Conference on Artificial Intelligence and Statistics*, pages 379–387. PMLR, 2021.

[BLLT20] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[BLS14] Florentina Bunea, Johannes Lederer, and Yiyuan She. The group square-root lasso: Theoretical properties and fast algorithms. *IEEE Trans. Inf. Theory*, 60(2):1313–1325, 2014.

[BLT18] Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

[BM88] James V Burke and Jorge J Moré. On the identification of active constraints. *SIAM Journal on Numerical Analysis*, 25(5):1197–1211, 1988.

[BM03] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[BMAS14] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(42):1455–1459, 2014.

[BMDFT22] Jean-Sébastien Brouillon, Keith Moffat, Florian Dörfler, and Giancarlo Ferrari-Trecate. Robust online joint state/input/parameter estimation of linear systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 2153–2158. IEEE, 2022.

[BMM06] Stefania Bellavia, Maria Macconi, and Benedetta Morini. An interior point Newton-like method for non-negative least-squares problems with degenerate solution. *Numerical Linear Algebra with Applications*, 13(10):825–846, 2006.

[BNPS17] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165:471–507, 2017.

[BNZ21] Jonathan Bauch, Boaz Nadler, and Pini Zilber. Rank 2r iterative least squares: efficient recovery of ill-conditioned low rank matrices from few entries. *SIAM Journal on Mathematics of Data Science*, 3(1):439–465, 2021.

[Bor13] Jorge Luis Borges. *El hacedor*. Vintage Español, 2013.

[Bos50] Rogerius Joseph Boscovich. De calculo probabilitatum quae respondent diversis valoribus summae errorum post plures observationes, quarum singulae possint esse erroneae certa quadam quantitate, f.[1]. *Autograf pohranjen u Bancroft Library u sastavu University of California at Berkeley, u zbirci Boscovich Papers, sa signaturom: Carton*, 1:1–6, 1750.

[Bou23] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[BP16] Jerome Bolte and Edouard Pauwels. Majorization-minimization procedures and convergence of sqp methods for semi-algebraic and tame programs. *Mathematics of Operations Research*, 41(2):442–465, 2016.

[BPC$^+$11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3:1–122, 2011.

[BPR21] Silvia Bonettini, Marco Prato, and Simone Rebegoldi. New convergence results for the inexact variable metric forward–backward method. *Applied Mathematics and Computation*, 392:125719, 2021.

[Bre18] Leo Breiman. Reflections after refereeing papers for nips. In *The Mathematics of Generalization*, pages 11–15. CRC Press, 2018.

[BRT09] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705 – 1732, 2009.

[BS83] Peter Bloomfield and William L Steiger. *Least absolute deviations: theory, applications, and algorithms*. Springer, 1983.

[BS14] Prabhu Babu and Petre Stoica. Connection between spice and square-root lasso for sparse parameter estimation. *Signal Processing*, 95:10–14, 2014.

[BS15] Amir Beck and Shoham Sabach. Weiszfeld's method: Old and new results. *J. Optim. Theory Appl.*, 164:1–40, 2015.

[BS17] Amir Beck and Shimrit Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1):1–27, 2017.

[BSS21] Alexandra Bünger, Valeria Simoncini, and Martin Stoll. A low-rank matrix equation method for solving pde-constrained optimization problems. *SIAM Journal on Scientific Computing*, 43(5):S637–S654, 2021.

[BT97] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena scientific Belmont, MA, 1997.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[BT12] Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

[BTT91] Michel Bierlaire, Ph L Toint, and Daniel Tuyttens. On iterative algorithms for linear least squares problems with bound constraints. *Linear Algebra and its Applications*, 143:111–143, 1991.

[Bur12] Charles S. Burrus. Iterative reweighted least squares. *OpenStax CNX: Available online: http://cnx.org/contents/92b90377-2b34-49e4-b26f-7fe572db78a1*, 2012.

[BV04] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[BZ22] Pierre C Bellec and Cun-Hui Zhang. De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli*, 28(2):713–743, 2022.

[BZZB09] Federico Benvenuto, Riccardo Zanella, Luca Zanni, and Mario Bertero. Nonnegative least-squares image deblurring: improved gradient projection approaches. *Inverse Problems*, 26(2):025004, 2009.

[CA16] Léopold Cambier and P-A Absil. Robust low-rank matrix completion by riemannian optimization. *SIAM Journal on Scientific Computing*, 38(5):S440–S460, 2016.

[CB83] NR Chapman and I Barrodale. Deconvolution of marine seismic data using the L1 norm. *Geophysical Journal International*, 72(1):93–100, 1983.

[CBSW15] Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Completing any low-rank matrix, provably. *The Journal of Machine Learning Research*, 16(1):2999–3034, 2015.

[CC14] Yuxin Chen and Yuejie Chi. Robust spectral compressed sensing via structured matrix completion. *IEEE Transactions on Information Theory*, 60(10):6576–6601, 2014.

[CC18] Yudong Chen and Yuejie Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14–31, 2018.

[CCBB14] Alexander Cloninger, Wojciech Czaja, Ruiliang Bai, and Peter J Basser. Solving 2d fredholm integral from incomplete measurements using compressive sensing. *SIAM journal on imaging sciences*, 7(3):1775–1798, 2014.

[CCF$^+$20] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121, 2020.

[CCF$^+$21] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

[CCW16] Jameson Cahill, Xuemei Chen, and Rongrong Wang. The gap between the null space property and the restricted isometry property. *Linear Algebra Appl.*, 501:363–375, 2016.

[CD94a] Shaobing Chen and David Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE, 1994.

[CD94b] Shaobing Scott Chen and David L. Donoho. Basis pursuit. In *Proc. of 1994 28th Asilomar Conference on Signals, Systems and Computers*, pages 41–44, 1994.

[CDD09] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22:211–231, 2009.

[CDS01a] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, 2001.

[CDS01b] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[CEHV15] Aldo Conca, Dan Edidin, Milena Hering, and Cynthia Vinzant. An algebraic characterization of injectivity in phase retrieval. *Applied and Computational Harmonic Analysis*, 38(2):346–356, 2015.

[CESV15] Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

[CFG14] Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on pure and applied Mathematics*, 67(6):906–956, 2014.

[CGMR20] Hung-Hsu Chou, Carsten Gieshoff, Johannes Maly, and Holger Rauhut. Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *arXiv preprint arXiv:2011.13772*, 2020.

[Che12] Xiaojun Chen. Smoothing methods for nonsmooth, nonconvex minimization. *Mathematical programming*, 134:71–99, 2012.

[Che15] Yudong Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

[Chi21] Lénaïc Chizat. Convergence rates of gradient methods for convex optimization in the space of measures. *arXiv preprint arXiv:2105.08368*, 2021.

[Cip00] Barry A Cipra. The best of the 20th century: Editors name top 10 algorithms. *SIAM news*, 33(4):1–2, 2000.

[CLC19] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[CLC21] Denis Chetverikov, Zhipeng Liao, and Victor Chernozhukov. On cross-validated Lasso in high dimensions. *The Annals of Statistics*, 49(3):1300 – 1317, 2021.

[Cli72] AK Cline. Rate of convergence of lawson's algorithm. *Mathematics of Computation*, 26(117):167–176, 1972.

[CLL20] Ji Chen, Dekai Liu, and Xiaodong Li. Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *IEEE Transactions on Information Theory*, 66(9):5806–5841, 2020.

[CLW16] Michael Chichignoud, Johannes Lederer, and Martin J Wainwright. A practical scheme and fast algorithm to tune the lasso with optimality guarantees. *The Journal of Machine Learning Research*, 17(1):8162–8181, 2016.

[CM73] Jon F Claerbout and Francis Muir. Robust modeling with erratic data. *Geophysics*, 38(5):826–844, 1973.

[CMR23] Hung-Hsu Chou, Johannes Maly, and Holger Rauhut. More is less: Inducing sparsity via overparameterization. *Information and Inference: A Journal of the IMA*, 12(3):iaad012, 2023.

[CMV22] Hung-Hsu Chou, Johannes Maly, and Claudio Mayrink Verdun. Non-negative least squares via overparametrization. *arXiv preprint arXiv:2207.08437*, 2022.

[CNX22] Paul Christiano, Eric Neyman, and Mark Xu. Formalizing the presumption of independence. *arXiv preprint arXiv:2211.06738*, 2022.

[COB19] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

[CP10a] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[CP10b] Donghui Chen and Robert J Plemmons. Nonnegativity constraints in numerical analysis. In *The birth of numerical analysis*, pages 109–139. World Scientific, 2010.

[CP11] Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212, 2011.

[CP21] Ying Cui and Jong-Shi Pang. *Modern nonconvex nondifferentiable optimization.* SIAM, 2021.

[CPR14] Emilie Chouzenoux, Jean-Christophe Pesquet, and Audrey Repetti. Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. *Journal of Optimization Theory and Applications*, 162(1):107–132, 2014.

[CQS98] Xiaojun Chen, Liqun Qi, and Defeng Sun. Global and superlinear convergence of the smoothing newton method and its application to general box constrained variational inequalities. *Mathematics of computation*, 67(222):519–540, 1998.

[CR07] Emmanuel Candes and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969, 2007.

[CR09] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[CRBH11] Jie Chen, Cédric Richard, José Carlos M Bermudez, and Paul Honeine. Nonnegative least-mean-square algorithm. *IEEE Transactions on Signal Processing*, 59(11):5225–5235, 2011.

[CRPW12] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12:805–849, 2012.

[CRT06a] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[CRT06b] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

[CRWX18] Jian-Feng Cai, Yi Rong, Yang Wang, and Zhiqiang Xu. Data recovery on a manifold from linear samples: theory and computation. *Annals of Mathematical Sciences and Applications*, 3(1):337–365, 2018.

[CSEP21] Delin Chu, Weya Shi, Srinivas Eswar, and Haesun Park. An alternating rank-k nonnegative least squares framework (ARkNLS) for nonnegative matrix factorization. *SIAM Journal on Matrix Analysis and Applications*, 42(4):1451–1479, 2021.

[CT05] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[CT06] Emmanuel J. Candès and Terence Tao. Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? *IEEE Trans. Inf. Theory*, 52(12):5406–5425, 2006.

[CT10] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[CTZ22] Hong TM Chu, Kim-Chuan Toh, and Yangjing Zhang. On regularized square-root regression problems: Distributionally robust interpretation and fast computations. *Journal of Machine Learning Research*, 23(308):1–39, 2022.

[CW05] Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(2):1168–1200, 2005.

[CW22] Zhaoliang Chen and Shiping Wang. A review on matrix completion for recommender systems. *Knowledge and Information Systems*, pages 1–34, 2022.

[CWB08] Emmanuel J. Candès, Michael B. Wakin, and Stephen P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

[CWW19] Jian-Feng Cai, Tianming Wang, and Ke Wei. Fast and provable algorithms for spectrally sparse signal reconstruction via low-rank hankel matrix completion. *Applied and Computational Harmonic Analysis*, 46(1):94–121, 2019.

[CY08] Rick Chartrand and Wotao Yin. Iteratively reweighted algorithms for compressive sensing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3869–3872, 2008.

[CY16] Rick Chartrand and Wotao Yin. Nonconvex sparse regularization and splitting algorithms. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 237–249. Springer International Publishing, 2016.

[CZ18] T Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018.

[dat23] Data Never Sleeps 10.0. `https://www.domo.com/data-never-sleeps`, 2023. Accessed: 2023-06-19.

[DDFG10] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Comm. Pure Appl. Math.*, 63:1–38, 2010.

[DDM21] Monica Dessole, Marco Dell'Orto, and Fabio Marcuzzi. The lawson-hanson algorithm with deviation maximization: Finite convergence and sparse recovery. *Numerical Linear Algebra with Applications*, page e2490, 2021.

[DE03a] D. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proc. Natl. Acad. Sci. USA*, 100(5):2197–2202, 2003.

[DE03b] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[Der18] Alexis Derumigny. Improved bounds for Square-Root Lasso and Square-Root Slope. *Electronic Journal of Statistics*, 12(1):741 – 766, 2018.

[DF05] Yu-Hong Dai and Roger Fletcher. Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming. *Numerische Mathematik*, 100(1):21–47, 2005.

[DFRW20] Lars Diening, Massimo Fornasier, Tomasi Roland, and Maximilian Wank. A Relaxed Kacanov iteration for the p-poisson problem. *Numer. Math.*, 145(2):1–34, 2020.

[DGHG22] Shuyu Dong, Bin Gao, Wen Huang, and Kyle A Gallivan. On the analysis of optimization with fixed-rank matrices: a quotient geometric view. *arXiv preprint arXiv:2203.06765*, 2022.

[DH$^+$01] David L Donoho, Xiaoming Huo, et al. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

[DH18] Daniel Dadush and Sophie Huiberts. A friendly smoothed analysis of the simplex method. *SIAM J. Comput.*, 49(5):STOC18–449–STOC18–499, 2018.

[DJHS92] David L Donoho, Iain M Johnstone, Jeffrey C Hoch, and Alan S Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):41–67, 1992.

[DK70] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[DL92] David L Donoho and Benjamin F Logan. Signal recovery and the large sieve. *SIAM J. Appl. Math.*, 52(2):577–591, 1992.

[DLR16] Sjoerd Dirksen, Guillaume Lecué, and Holger Rauhut. On the gap between restricted isometry properties and sparse recovery conditions. *IEEE Transactions on Information Theory*, 64(8):5478–5487, 2016.

[DLR18] S. Dirksen, G. Lecué, and H. Rauhut. On the gap between restricted isometry properties and sparse recovery conditions. *IEEE Trans. Inform. Theory*, 64(8):5478–5487, Aug 2018.

[DMA97] Geoffrey Davis, Stephane Mallat, and Marco Avellanedas. Adaptive greedy approximations. *Constr. Approx.*, 13(1):57–98, 1997.

[DMEH19] Antonio De Maio, Yonina C Eldar, and Alexander M Haimovich. *Compressed sensing in radar signal processing.* Cambridge University Press, 2019.

[Don06] David L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theor.*, 52(4):1289–1306, 2006.

[Don17] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[DOT18] Jan Draisma, Giorgio Ottaviani, and Alicia Tocino. Best rank-k approximations for tensors: generalizing eckart–young. *Research in the Mathematical Sciences*, 5(2):27, 2018.

[DPG$^+$14] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27, 2014.

[DPW09] Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Instance-optimality in probability with an ℓ_1-minimization decoder. *Applied and Computational Harmonic Analysis*, 27(3):275–288, 2009.

[DR16] Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[DS00] Jack Dongarra and Francis Sullivan. Guest editors introduction to the top 10 algorithms. *Computing in Science & Engineering*, 2(01):22–23, 2000.

[DSKL$^+$22] Christopher M De Sa, Satyen Kale, Jason D Lee, Ayush Sekhari, and Karthik Sridharan. From gradient flow on population loss to learning with stochastic gradient descent. *Advances in Neural Information Processing Systems*, 35:30963–30976, 2022.

[DSST18] Chao Ding, Defeng Sun, Jie Sun, and Kim-Chuan Toh. Spectral operators of matrices. *Mathematical Programming*, 168:509–531, 2018.

[DT05] David L Donoho and Jared Tanner. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proceedings of the National Academy of Sciences*, 102(27):9446–9451, 2005.

[DT08] David L. Donoho and Yaakov Tsaig. Fast solution of ℓ_1-norm minimization problems when the solution may be sparse. *IEEE Trans. Inform. Theory*, 54:4789–4812, 2008.

[DT09a] David Donoho and Jared Tanner. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22(1):1–53, 2009.

[DT09b] David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

[DT10a] David L Donoho and Jared Tanner. Counting the faces of randomly-projected hypercubes and orthants, with applications. *Discrete & Computational Geometry*, 43(3):522–541, 2010.

[DT10b] David L Donoho and Jared Tanner. Exponential bounds implying construction of compressed sensing matrices, error-correcting codes, and neighborly polytopes by random sampling. *IEEE Transactions on Information Theory*, 56(4):2002–2016, 2010.

[Dun87] John C Dunn. On the convergence of projected gradient processes to singular critical points. *Journal of Optimization Theory and Applications*, 55:203–216, 1987.

[DY23] Michal Derezinski and Jiaming Yang. Solving dense linear systems faster than via preconditioning. *arXiv preprint arXiv:2312.08893*, 2023.

[EC21] Omer Elkabetz and Nadav Cohen. Continuous vs. discrete optimization of deep neural networks. *Advances in Neural Information Processing Systems*, 34:4947–4960, 2021.

[EEN$^+$23] S Endt, M Engel, E Naldi, R Assereto, M Molendowska, L Müller, C Mayrink Verdun, CM Pirkl, M Palombo, DK Jones, et al. In-vivo myelin water quantification using diffusion-relaxation correlation mri: a comparison of 1d and 2d methods. *Appl Magn Reson*, 54:1571–1588, 2023.

[Eis61] Churchill Eisenhart. The background and evolution of the method of least squares. *Bulletin of the British Society for the History of Science*, 2(20):100–102, 1961.

[Ela10] Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*, volume 2. Springer, 2010.

[ELX13] Ernie Esser, Yifei Lou, and Jack Xin. A method for finding structured sparse solutions to nonnegative least squares problems with applications. *SIAM Journal on Imaging Sciences*, 6(4):2010–2046, 2013.

[EM11] Horst A Eiselt and Vladimir Marianov. Pioneering developments in location analysis. *Foundations of location analysis*, pages 3–22, 2011.

[ENP12] Yonina C Eldar, Deanna Needell, and Yaniv Plan. Uniqueness conditions for low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 33(2):309–314, 2012.

[EOS19] Ron Estrin, Dominique Orban, and Michael A Saunders. Lslq: An iterative method for linear least-squares with an error minimization property. *SIAM Journal on Matrix Analysis and Applications*, 40(1):254–275, 2019.

[EV19] Alina Ene and Adrian Vladu. Improved convergence for ℓ_1 and ℓ_∞ regression via iteratively reweighted least squares. In *Proceedings of 2019 International Conference on Machine Learning (ICML'19)*, pages 1794–1801. PMLR, 2019.

[EY36] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[Far99] Richard William Farebrother. *Fitting linear relationships: A history of the calculus of observations 1750-1900.* Springer Science & Business Media, 1999.

[Far18] Richard William Farebrother. *Linear least squares computations.* Routledge, 2018.

[Fas95] Dario Fasino. Spectral properties of hankel matrices and numerical solutions of finite moment problems. *Journal of Computational and Applied Mathematics*, 65(1-3):145–155, 1995.

[Faz02] Maryam Fazel. *Matrix rank minimization with applications.* PhD thesis, PhD thesis, Stanford University, 2002.

[FGJ$^+$22] Tim Fuchs, David Gross, Peter Jung, Felix Krahmer, Richard Kueng, and Dominik Stöger. Proof methods for robust low-rank matrix recovery. In *Compressed Sensing in Information Processing*, pages 37–75. Springer, 2022.

[FGP15] Pierre Frankel, Guillaume Garrigos, and Juan Peypouquet. Splitting methods with variable metric for Kurdyka–Lojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165:874–900, 2015.

[FHB03] Maryam Fazel, Haitham Hindi, and Stephen P Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *Proceedings of the 2003 American Control Conference, 2003.*, volume 3, pages 2156–2162. IEEE, 2003.

[FL12] Albert Fannjiang and Wenjing Liao. Coherence pattern–guided compressive sensing with unresolved grids. *SIAM Journal on Imaging Sciences*, 5(1):179–202, 2012.

[Fle05] Roger Fletcher. On the Barzilai-Borwein method. In *Optimization and control with applications*, pages 235–256. Springer, 2005.

[FN03] Arie Feuer and Arkadi Nemirovski. On sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 49(6):1579–1581, 2003.

[Fou14] Simon Foucart. Stability and robustness of ℓ_1-minimizations with weibull matrices and redundant dictionaries. *Linear Algebra and its Applications*, 441:4–21, 2014.

[Fou18] Simon Foucart. Concave mirsky inequality and low-rank recovery. *SIAM Journal on Matrix Analysis and Applications*, 39(1):99–103, 2018.

[Fou23] Simon Foucart. The sparsity of lasso-type minimizers. *Applied and Computational Harmonic Analysis*, 62:441–452, 2023.

[FPRW16a] Massimo Fornasier, Steffen Peter, Holger Rauhut, and Stephan Worm. Conjugate gradient acceleration of iteratively re-weighted least squares methods. *Comput. Optim. Appl.*, 65(1):205–259, 2016.

[FPRW16b] Massimo Fornasier, Steffen Peter, Holger Rauhut, and Stephan Worm. Conjugate gradient acceleration of iteratively re-weighted least squares methods. *Comput. Optim. Appl.*, 65(1):205–259, 2016.

[FR13] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing.* Applied and Numerical Harmonic Analysis. Springer New York, 2013.

[FRW11a] Massimo Fornasier, Holger Rauhut, and Rachel Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM J. Optim.*, 21:1614–1640, 2011.

[FRW11b] Massimo Fornasier, Holger Rauhut, and Rachel Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM Journal on Optimization*, 21(4):1614–1640, 2011.

[FS11] David Chin-Lung Fong and Michael Saunders. Lsmr: An iterative algorithm for sparse least-squares problems. *SIAM Journal on Scientific Computing*, 33(5):2950–2971, 2011.

[FTZ22] Simon Foucart, Eitan Tadmor, and Ming Zhong. On the sparsity of lasso minimizers in sparse data recovery. *Constructive Approximation*, pages 1–19, 2022.

[Fuc04] J-J Fuchs. On sparse representations in arbitrary redundant bases. *IEEE transactions on Information theory*, 50(6):1341–1344, 2004.

[FWZ18] Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42, 2018.

[Gal89] Francis Galton. *Natural inheritance.* Macmillan and Company, 1889.

[Gau57] Carl Friedrich Gauss. *Theory of the motion of the heavenly bodies moving about the sun in conic sections: a translation of Carl Frdr. Gauss" Theoria motus": With an appendix. By Charles Henry Davis.* Little, Brown and Comp., 1857.

[Gau77] Carl Friedrich Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, volume 7. FA Perthes, 1877.

[Gau59] Walter Gautschi. Some elementary inequalities relating to the gamma and incomplete gamma function. *Journal of Mathematics and Physics*, 38(1-4):77–81, 1959.

[GBLJ19] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, pages 3202–3211, 2019.

[GBR21] Chirag Gupta, Sivaraman Balakrishnan, and Aaditya Ramdas. Path length bounds for gradient descent and flow. *The Journal of Machine Learning Research*, 22(1):3154–3216, 2021.

[GGK12] Israel Gohberg, Seymour Goldberg, and Nahum Krupnik. *Traces and determinants of linear operators*, volume 116. Birkhäuser, 2012.

[GHS87] Gene H Golub, Alan Hoffman, and Gilbert W Stewart. A generalization of the eckart-young-mirsky matrix approximation theorem. *Linear Algebra and its applications*, 88:317–327, 1987.

[GHV12] Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. High-dimensional regression with unknown variance. *Statistical Science*, 27(4):500–518, 2012.

[Gir15] Christophe Giraud. *Introduction to high-dimensional statistics.* Monographs on statistics and applied probability ; 139. CRC Press, 2015.

[GKK20] Kelly Geyer, Anastasios Kyrillidis, and Amir Kalev. Low-rank regularization and solution uniqueness in over-parameterized matrix sensing. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 930–940, 2020.

[Gla72] JWL Glaisher. The method of least squares. *Nature*, 6(138):140–141, 1872.

[GLF$^+$10] David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.

[GLSS18] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018.

[GM73] Philip E Gill and Walter Murray. A numerically stable form of the simplex algorithm. *Linear algebra and its applications*, 7(2):99–138, 1973.

[GM12] Bernd Gärtner and Jiri Matousek. *Approximation algorithms and semidefinite programming.* Springer Science & Business Media, 2012.

[GN03] Rémi Gribonval and Morten Nielsen. Sparse representations in unions of bases. *IEEE transactions on Information theory*, 49(12):3320–3325, 2003.

[Gol12] Herman Heine Goldstine. *A History of Numerical Analysis from the 16th through the 19th Century*, volume 2. Springer Science & Business Media, 2012.

[Gor88] Yehoram Gordon. On milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1986–87*, pages 84–106. Springer, 1988.

[Gor16] Prakash Gorroochurn. *Classic topics on the history of modern mathematical statistics: From Laplace to more recent times*. John Wiley & Sons, 2016.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[GR92] Donald Geman and George Reynolds. Constrained Restoration and the Recovery of Discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(03):367–383, 1992.

[GR97a] Irina Gorodnitsky and Bhaskar Rao. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *IEEE Trans. Signal Process.*, 45(3):600–616, 1997.

[GR97b] Irina F Gorodnitsky and Bhaskar D Rao. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *IEEE Transactions on signal processing*, 45(3):600–616, 1997.

[GR15] Robert M Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[Gre67] John Greenstadt. On the relative efficiencies of gradient methods. *Mathematics of Computation*, 21(99):360–367, 1967.

[Gre84] Peter 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)*, 46(2):149–170, 1984.

[Gro11] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[GS70] Gene H. Golub and Michael A. Saunders. Linear least squares and quadratic programming. In Jean Abadie and Philip Wolfe, editors, *Integer and Non-*

linear Programming, pages 229–256. North Holland Pub. Co., Amsterdam, 1970.

[GSD20] Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. *International Conference on Learning Representations (ICLR).*, 2020.

[GVRH20] Paris Giampouras, René Vidal, Athanasios Rontogiannis, and Benjamin Haeffele. A novel variational form of the schatten-p quasi-norm. *Advances in Neural Information Processing Systems*, 33:21453–21463, 2020.

[GW11] Fred G Gustavson and Jerzy Waśniewski. *Gauss's, Cholesky's and Banachiewicz's Contributions to Least Squares.* DTU Informatics, 2011.

[GWB+17] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

[GWS21] Suriya Gunasekar, Blake Woodworth, and Nathan Srebro. Mirrorless mirror descent: A natural derivation of mirror descent. In *International Conference on Artificial Intelligence and Statistics*, pages 2305–2313. PMLR, 2021.

[Har22] George Harvey. On the method of minimum squares, employed in the reduction of experiments, being a translation of the appendix to an essay of legendre's, entitled nouvelles méthodes pour la détermination des orbites des comètes. *Edinburgh Philosophical Journal*, 7:292–301, 1822.

[Har74a] H Leon Harter. The method of least squares and some alternatives: Part ii. *International Statistical Review/Revue Internationale de Statistique*, 42(3):235–282, 1974.

[Har74b] W Leon Harter. The method of least squares and some alternatives: Part i. *International Statistical Review/Revue Internationale de Statistique*, pages 147–174, 1974.

[Har75a] H Leon Harter. The method of least squares and some alternatives. addendum to part iv. *International Statistical Review/Revue Internationale de Statistique*, 43(3):273–278, 1975.

[Har75b] H Leon Harter. The method of least squares and some alternatives: Part iii. *International Statistical Review/Revue Internationale de Statistique*, 43(1):1–44, 1975.

[Har75c] H Leon Harter. The method of least squares and some alternatives: Part iv. *International Statistical Review/Revue Internationale de Statistique*, 43(2):125–190, 1975.

[HCMTH15] Junhui Hou, Lap-Pui Chau, Nadia Magnenat-Thalmann, and Ying He. Sparse low-rank matrix approximation for data compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(5):1043–1054, 2015.

[HHJ94] Roger A Horn, Roger A Horn, and Charles R Johnson. *Topics in matrix analysis*. Cambridge university press, 1994.

[Hig02] Nicholas J Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2002.

[HK15] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[HKV$^+$22] Frederik Hoppe, Felix Krahmer, Claudio Mayrink Verdun, Marion I Menzel, and Holger Rauhut. Uncertainty quantification for sparse fourier recovery. *arXiv preprint arXiv:2212.14864*, 2022.

[HKV$^+$23a] Frederik Hoppe, Felix Krahmer, Claudio Mayrink Verdun, Marion I Menzel, and Holger Rauhut. High-dimensional confidence regions in sparse mri. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[HKV$^+$23b] Frederik Hoppe, Felix Krahmer, Claudio Mayrink Verdun, Marion I Menzel, and Holger Rauhut. Sampling strategies for compressive imaging under statistical noise. In *2023 International Conference on Sampling Theory and Applications (SampTA)*, pages 1–5. IEEE, 2023.

[HMJG23] Andi Han, Bamdev Mishra, Pratik Jawanpuria, and Junbin Gao. Riemannian accelerated gradient methods via extrapolation. In *International Conference on Artificial Intelligence and Statistics*, pages 1554–1585. PMLR, 2023.

[HMRT22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[HMT11] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[HNWL21] Zhanxuan Hu, Feiping Nie, Rong Wang, and Xuelong Li. Low rank regularization: A review. *Neural Networks*, 136:218–232, 2021.

[Hof17] Peter D Hoff. Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017.

[Hot33] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[Hot36] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[HS$^+$52a] Magnus R Hestenes, Eduard Stiefel, et al. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49(6):409–436, 1952.

[HS52b] Magnus Rudolph Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49(1), 1952.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[HSW$^+$21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[HTFF09] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[HTT$^+$09] Anthony JG Hey, Stewart Tansley, Kristin Michele Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA, 2009.

[HTW15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

[HTW19] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, 2019.

[Hub64] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 1964.

[Hub81] P.J. Huber. *Robust statistics*. Wiley New York, 1981.

[HVL$^+$23] Frederik Hoppe, Claudio Mayrink Verdun, Hannah Laus, Felix Krahmer, and Holger Rauhut. Uncertainty quantification for learned ista. In *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2023.

[HW77] Paul W. Holland and Roy E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, 6:813–827, 1977.

[HW14] Moritz Hardt and Mary Wootters. Fast matrix completion without the condition number. In *Conference on learning theory*, pages 638–678. PMLR, 2014.

[Idi01] Jérôme Idier. Convex half-quadratic criteria and interacting auxiliary variables for image restoration. *IEEE Trans. Image Process.*, 10(7):1001–1009, 2001.

[JFL15] Oren N Jaspan, Roman Fleysher, and Michael L Lipton. Compressed sensing mri: a review of the clinical literature. *The British journal of radiology*, 88(1056):20150487, 2015.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[JLSX10] Hui Ji, Chaoqiang Liu, Zuowei Shen, and Yuhong Xu. Robust video denoising using low rank matrix completion. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1791–1798. IEEE, 2010.

[JM18] Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics*, 46(6A):2593 – 2622, 2018.

[Joh90] Charles R Johnson. *Matrix theory and applications*, volume 40. American Mathematical Soc., 1990.

[JT19] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[Kal18] El Mostafa Kalmoun. An investigation of smooth tv-like regularization in the context of the optical flow problem. *Journal of Imaging*, 4(2):31, 2018.

[Kar72] Richard M Karp. Reducibility among combinatorial problems, complexity of computer computations (re miller and jw thatcher, editors), 1972.

[KBB15] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. *Advances in neural information processing systems*, 28, 2015.

[KBP$^+$11] Hyunsoo Kim, Yingtao Bi, Sharmistha Pal, Ravi Gupta, and Ramana V Davuluri. IsoformEx: isoform level gene expression estimation using weighted non-negative least squares from mRNA-Seq data. *BMC bioinformatics*, 12(1):1–9, 2011.

[KBV09] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[Kel51] Earl C Kelley. The workshop way of learning. 1951.

[KHS09] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

[KJ17] Richard Kueng and Peter Jung. Robust nonnegative sparse recovery and the nullspace property of 0/1 measurements. *IEEE Transactions on Information Theory*, 64(2):689–703, 2017.

[KK17] Michael Kech and Felix Krahmer. Optimal injectivity conditions for bilinear inverse problems with applications to identifiability of deconvolution problems. *SIAM Journal on Applied Algebra and Geometry*, 1(1):20–37, 2017.

[KKKT19] Jun Beom Kho, Jaihie Kim, Ig-Jae Kim, and Andrew BJ Teoh. Cancelable fingerprint template design with randomized non-negative least squares. *Pattern Recognition*, 91:245–260, 2019.

[KKRT16] Maryia Kabanava, Richard Kueng, Holger Rauhut, and Ulrich Terstiege. Stable low-rank matrix recovery via null space properties. *Information and Inference: A Journal of the IMA*, 5(4):405–441, 2016.

[Kle09] Andreas Kleinert. Der messende luchs: Zwei verbreitete fehler in der galilei-literatur. *NTM Zeitschrift für Geschichte der Wissenschaften, Technik und Medizin*, 17:199–206, 2009.

[Klo14] Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

[KMM$^+$18] Daichi Kitamura, Shinichi Mogami, Yoshiki Mitsui, Norihiro Takamune, Hiroshi Saruwatari, Nobutaka Ono, Yu Takahashi, and Kazunobu Kondo. Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation. *EURASIP Journal on Advances in Signal Processing*, 2018(1):1–25, 2018.

[KMN89] David Kahaner, Cleve Moler, and Stephen Nash. *Numerical methods and software.* Prentice-Hall, Inc., 1989.

[KMO10] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

[KMV20] C Kümmerle and C Mayrink Verdun. Escaping saddle points in ill-conditioned matrix completion with a scalable second order method. In *Workshop on "Beyond first-order methods in ML systems" at the 37th International Conference on Machine Learning*, 2020.

[KMV21] Christian Kümmerle and Claudio Mayrink Verdun. A scalable second order method for ill-conditioned matrix completion from few samples. In *Proceedings of 2021 International Conference on Machine Learning (ICML'21)*, 2021.

[KMVS21] Christian Kümmerle, Claudio Mayrink Verdun, and Dominik Stöger. Iteratively reweighted least squares for basis pursuit with global linear convergence rate. *Advances in Neural Information Processing Systems*, 34:2873–2886, 2021.

[Kor09] Yehuda Koren. The bellkor solution to the netflix grand prize. *Netflix prize documentation*, 81(2009):1–10, 2009.

[KS18] Christian Kümmerle and Juliane Sigl. Harmonic mean iteratively reweighted least squares for low-rank matrix recovery. *The Journal of Machine Learning Research*, 19(1):1815–1863, 2018.

[KS21a] Nikola B Kovachki and Andrew M Stuart. Continuous time analysis of momentum methods. *The Journal of Machine Learning Research*, 22(1):760–799, 2021.

[KS21b] Felix Krahmer and Dominik Stöger. On the convex geometry of blind deconvolution and matrix completion. *Communications on Pure and Applied Mathematics*, 74(4):790–832, 2021.

[KSD13] Dongmin Kim, Suvrit Sra, and Inderjit S Dhillon. A non-monotonic method for large-scale non-negative least squares. *Optimization Methods and Software*, 28(5):1012–1039, 2013.

[Küm19] Christian Kümmerle. *Understanding and enhancing data recovery algorithms.* PhD thesis, Technische Universität München, 2019.

[KV19] Christian Kümmerle and Claudio M Verdun. Completion of structured low-rank matrices via iteratively reweighted least squares. In *2019 13th international conference on sampling theory and applications (sampta)*, pages 1–5. IEEE, 2019.

[KV21] Christian Kümmerle and Claudio M Verdun. A scalable second order method for ill-conditioned matrix completion from few samples. In *International Conference on Machine Learning*, pages 5872–5883. PMLR, 2021.

[KXL$^+$13] Linghe Kong, Mingyuan Xia, Xiao-Yang Liu, Min-You Wu, and Xue Liu. Data loss and reconstruction in sensor networks. In *2013 Proceedings IEEE INFOCOM*, pages 1654–1662. IEEE, 2013.

[Lan16] Kenneth Lange. *MM optimization algorithms.* SIAM, 2016.

[Law61] Charles Lawrence Lawson. Contribution to the theory of linear least maximum approximation. *Ph. D. dissertation. Univ. Calif.*, 1961.

[LB32] Marquis de Laplace and Nathaniel Bowditch. *Mécanique céleste, vols.1-4 [extensive annotated translation of Laplace, 1799-1805]*, volume 1-4. Hilliard, Gray, Little and Wilkins, Publishers, 1829-1832.

[LD11] Yuancheng Luo and Ramani Duraiswami. Efficient parallel nonnegative least squares on multicore architectures. *SIAM Journal on Scientific Computing*, 33(5):2848–2863, 2011.

[LDP07] Michael Lustig, David Donoho, and John M Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[LDSP08] Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly. Compressed sensing mri. *IEEE signal processing magazine*, 25(2):72–82, 2008.

[Leg06] Adrien Marie Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes: avec un supplément contenant divers perfectionnemens de ces méthodes et leur application aux deux comètes de 1805.* Courcier, 1806.

[Leh11] Erich L Lehmann. *Fisher, Neyman, and the creation of classical statistics.* Springer Science & Business Media, 2011.

[Lem12] Claude Lemaréchal. Cauchy and the gradient method. *Documenta Mathematica*, pages 251–254, 2012.

[LF81] Shlomo Levy and Peter K Fullagar. Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution. *Geophysics*, 46(9):1235–1243, 1981.

[LFP17] Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Activity identification and local linear convergence of forward–backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437, 2017.

[LH95] Charles L Lawson and Richard J Hanson. *Solving least squares problems.* SIAM, 1995.

[Lin07] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[LJ22] Thien Le and Stefanie Jegelka. Training invariances and the low-rank phenomenon: beyond linear networks. In *The Tenth International Conference on Learning Representations*, 2022.

[LJH$^+$20] Xinguo Li, Haoming Jiang, Jarvis Haupt, Raman Arora, Han Liu, Mingyi Hong, and Tuo Zhao. On fast convergence of proximal algorithms for sqrt-lasso optimization: Don't worry about its nonsmooth loss function. In

Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 49–59, Tel Aviv, Israel, 22–25 Jul 2020. PMLR.

[LK23] Stamatios Lefkimmiatis and Iaroslav Sergeevich Koshelev. Learning sparse and low-rank priors for image recovery via iterative reweighted least squares minimization. In *The Eleventh International Conference on Learning Representations*, 2023.

[LL00] Zhi-Pei Liang and Paul C Lauterbur. *Principles of magnetic resonance imaging.* SPIE Optical Engineering Press Bellingham, 2000.

[LM18] Gilad Lerman and Tyler Maunu. Fast, robust and non-convex subspace recovery. *Information and Inference: A Journal of the IMA*, 7(2):277–336, 2018.

[LMSH23] Hanbyul Lee, Rahul Mazumder, Qifan Song, and Jean Honorio. Matrix completion from general deterministic sampling patterns. *arXiv preprint arXiv:2306.02283*, 2023.

[LNHW21] Jiangyuan Li, Thanh Nguyen, Chinmay Hegde, and Ka Wai Wong. Implicit sparse regularization: The impact of depth and early stopping. In *Advances in Neural Information Processing Systems*, 2021.

[Log65] Benjamin Franklin Logan. *Properties of high-pass signals.* PhD thesis, Columbia University, 1965.

[Lou08] Karim Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2(none):90–102, 2008.

[LR19] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[LS98] W. Li and J. Swetits. The linear l1 estimator and the huber m-estimator. *SIAM J. Optim.*, 8(2):457–475, 1998.

[LS05] Adrian S Lewis and Hristo S Sendov. Nonsmooth analysis of singular values. part i: Theory. *Set-Valued Analysis*, 13:213–241, 2005.

[LST18] Xudong. Li, Defeng Sun, and Kim-Chuan Toh. A highly efficient semismooth newton augmented lagrangian method for solving lasso problems. *SIAM J. Optim.*, 28(1):433–458, 2018.

[LTVB22] Charles HC Little, Kee L Teo, and Bruce Van Brunt. *An Introduction to Infinite Products*. Springer, 2022.

[LTYL15] Canyi Lu, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm. *IEEE Transactions on Image Processing*, 25(2):829–839, 2015.

[LV10] Zhang Liu and Lieven 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.

[LW11] Adrian S Lewis and Stephen J Wright. Identifying activity. *SIAM Journal on Optimization*, 21(2):597–614, 2011.

[LW21] Ming-Jun Lai and Yang Wang. *Sparse solutions of underdetermined linear systems and their applications*. SIAM, 2021.

[LWLA22] Zhiyuan Li, Tianhao Wang, Jason D Lee, and Sanjeev Arora. Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent. *Advances in Neural Information Processing Systems*, 35:34626–34640, 2022.

[LXY13] Ming-Jun Lai, Yangyang Xu, and Wotao Yin. Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *SIAM J. Numer. Anal.*, 51:927–957, 2013.

[Lya92] Aleksandr Mikhailovich Lyapunov. *General problem of the stability of motion*, volume 55. CRC Press, 1992.

[LYW13] Ming-Jun Lai, Xu Yangyang, and Yin Wotao. Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *SIAM J. Numer. Anal.*, 2(51):927–957, 2013.

[LZQY22] Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14153–14172. PMLR, 17–23 Jul 2022.

[LZYL15] Xingguo Li, Tuo Zhao, Xiaoming Yuan, and Han Liu. The flare package for high dimensional linear regression and precision matrix estimation in r. *Journal of Machine Learning Research*, 16(18):553–557, 2015.

[Maj15] Angshul Majumdar. *Compressed Sensing for Magnetic Resonance Image Reconstruction.* Cambridge University Press, 2015.

[Mar18] Ivan Markovsky. *Low-Rank Approximation: Algorithms, Implementation, Applications.* Springer, 2018.

[MB70] Christopher Maire and Ruđer Josip Bo**v**ković. *Voyage astronomique et geographique, dans l'état de l'eglise: entrepris par l'ordre et sous les auspices du pape Benoit XIV, pour mesurer deux dégrés du méridien, & corriger la carte de l'etat ecclésiastique.* Tilliard, 1770.

[MB$^+$06] Nicolai Meinshausen, Peter Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *Annals of statistics*, 34(3):1436–1462, 2006.

[mdL25] Pierre Simon marquis de Laplace. *Traité de mécanique céleste*, volume 1-5. Chez JBM Duprat, libraire pour les mathématiques, quai des Augustins, 1825.

[Mei13] Nicolai Meinshausen. Sign-constrained least squares estimation for high-dimensional regression. *Electronic Journal of Statistics*, 7:1607–1631, 2013.

[Mel21] Kateryna Melnykova. *Theory and algorithms for compressive data acquisition under practical constraints.* PhD thesis, University of British Columbia, 2021.

[Mer77a] Mansfield Merriman. *A List of Writings Relating to the Method of Least Squares: With Historical and Critical Notes*, volume 4. Academy, 1877.

[Mer77b] Mansfield Merriman. On the history of the method of least squares. *The Analyst*, 4(2):33–36, 1877.

[Meu06] Gérard Meurant. *The Lanczos and Conjugate Gradient Algorithms: From Theory to Finite Precision Computations.* Society for Industrial and Applied Mathematics,, 2006.

[Mey21] Gregory P Meyer. An alternative probabilistic interpretation of the huber loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5261–5269, 2021.

[MF12a] Karthik Mohan and Maryam Fazel. Iterative reweighted algorithms for matrix rank minimization. *J. Mach. Learn. Res.*, 13:3441–3473, 2012.

[MF12b] Karthik Mohan and Maryam Fazel. Iterative reweighted algorithms for matrix rank minimization. *J. Mach. Learn. Res.*, 13(1):3441–3473, 2012.

[MFLS17] Joe M Myre, Erich Frahm, David J Lilja, and Martin O Saar. TNT-NN: a fast active set method for solving large non-negative least squares problems. *Procedia Computer Science*, 108:755–764, 2017.

[MGJK19] Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar. Globally-convergent iteratively reweighted least squares for robust regression problems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 313–322. PMLR, 2019.

[MGJK20] Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar. Globally-convergent iteratively reweighted least squares for robust regression problems. *ArXiv e-prints*, 2020.

[Mir60] Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59, 1960.

[MJ19] Michael Muehlebach and Michael Jordan. A dynamical systems perspective on nesterov acceleration. In *International Conference on Machine Learning*, pages 4656–4662. PMLR, 2019.

[MKPK22] Nathan Mankovich, Emily J King, Chris Peterson, and Michael Kirby. The flag median and flagirls. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10339–10347, 2022.

[ML17] Shahar Mendelson and Guillaume Lecué. Sparse recovery under weak moment assumptions. *J. Eur. Math. Soc.*, 19(3):881–904, 2017.

[MLL$^+$19] Joseph M Myre, Ioan Lascu, Eduardo A Lima, Joshua M Feinberg, Martin O Saar, and Benjamin P Weiss. Using TNT-NN to unlock the fast full spatial inversion of large magnetic microscopy data sets. *Earth, Planets and Space*, 71(1):1–26, 2019.

[MM15] Cameron Musco and Christopher Musco. Randomized block Krylov methods for stronger and faster approximate singular value decomposition. *Advances in neural information processing systems*, 28, 2015.

[MMBS13] Bamdev Mishra, Gilles Meyer, Francis Bach, and Rodolphe Sepulchre. Low-rank optimization with trace norm penalty. *SIAM Journal on Optimization*, 23(4):2124–2149, 2013.

[MMS11] Bamdev Mishra, Gilles Meyer, and Rodolphe Sepulchre. Low-rank optimization for distance matrix completion. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 4455–4460. IEEE, 2011.

[Mor65] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[MP97] Mehran Mesbahi and George P Papavassilopoulos. On the rank minimization problem over a positive semidefinite linear matrix inequality. *IEEE Transactions on Automatic Control*, 42(2):239–243, 1997.

[MPC$^+$18] Mauro Mangia, Fabio Pareschi, Valerio Cambareri, Riccardo Rovatti, and Gianluca Setti. *Adapted compressed sensing for effective hardware implementations: A design flow for signal-level optimization of compressed sensing stages*. Springer, 2018.

[MRG$^+$20] William Merrill, Vivek Ramanujan, Yoav Goldberg, Roy Schwartz, and Noah Smith. Effects of parameter norm growth during transformer training: Inductive bias from gradient descent. *arXiv preprint arXiv:2010.09697*, 2020.

[MS14] Bamdev Mishra and Rodolphe Sepulchre. R3MC: A Riemannian three-factor algorithm for low-rank matrix completion. In *53rd IEEE Conference on Decision and Control*, pages 1137–1142. IEEE, 2014.

[MSW20] Rahul Mazumder, Diego Saldana, and Haolei Weng. Matrix completion with nonconvex regularization: Spectral operators and scalable algorithms. *Statistics and Computing*, 30(4):1113–1138, 2020.

[MWCC18] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

[MZ93] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

[MZS23] Per Mattsson, Dave Zachariah, and Petre Stoica. Analysis of the minimum-norm least-squares estimator and its double-descent behavior [lecture notes]. *IEEE Signal Processing Magazine*, 40(3):39–75, 2023.

[N$^+$18] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[Nat95] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[Nes83] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Dokl. akad. nauk SSSR*, volume 269, pages 543–547, 1983.

[Nes05] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103:127–152, 2005.

[NFG$^+$17] Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, Vincent Leclère, and Joseph Salmon. Efficient smoothed concomitant lasso estimation for high dimensional regression. *Journal of Physics: Conference Series*, 904:012006, oct 2017.

[Nie01] Yves Nievergelt. A tutorial history of least squares with applications to astronomy and geodesy. *Numerical Analysis: Historical Developments in the 20th Century*, pages 77–112, 2001.

[NKS19] Luong Trung Nguyen, Junhan Kim, and Byonghyo Shim. Low-rank matrix completion: A contemporary survey. *IEEE Access*, 7:94215–94237, 2019.

[NN94] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming.* SIAM, 1994.

[NN05] Mila Nikolova and Michael K. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Sci. Comput.*, 27(3):937–966, 2005.

[Nof17] Vanni Noferini. A formula for the fréchet derivative of a generalized matrix function. *SIAM Journal on Matrix Analysis and Applications*, 38(2):434–457, 2017.

[NS12] Thanh Ngo and Yousef Saad. Scaled gradients on grassmann manifolds for matrix completion. *Advances in neural information processing systems*, 25, 2012.

[NTS15] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations*, 2015.

[NTSS17] Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

[NTTB22] Rishhabh Naik, Nisarg Trivedi, Davoud Ataee Tarzanagh, and Laura Balzano. Truncated matrix completion-an empirical study. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 847–851. IEEE, 2022.

[NW06] Jorge Nocedal and Stephen Wright. *Numerical optimization.* Springer Science & Business Media, 2006.

[ODBP15] Peter Ochs, Alexey Dosovitskiy, Thomas Brox, and Thomas Pock. On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision. *SIAM J. Imaging Sci.*, 8(1):331–372, 2015.

[ORVW22] José Luis Montiel Olea, Cynthia Rush, Amilcar Velez, and Johannes Wiesel. On the generalization error of norm penalty linear regression models. *arXiv preprint arXiv:2211.07608*, 2022.

[Ost20] Wolfgang Osterhage. *Johannes Kepler: The order of things.* Springer Nature, 2020.

[OTH13] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized LASSO: A precise analysis. In *51st Annual Allerton Conference on Communication, Control, and Computing, Allerton 2013, Allerton Park & Retreat Center, Monticello, IL, USA, October 2-4, 2013*, pages 1002–1009. IEEE, 2013.

[Owe07] A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 19(2):59–72, 2007.

[PABN16] Daniel L Pimentel-Alarcón, Nigel Boston, and Robert D Nowak. A characterization of deterministic sampling patterns for low-rank matrix completion. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):623–636, 2016.

[pap23a] Papers with Code - Image Classification on Imagenet. `https://paperswithcode.com/sota/image-classification-on-imagenet`, 2023. Accessed: 2023-06-01.

[pap23b] Papers with Code - Language Modelling on Wikitext 103. `https://paperswithcode.com/sota/language-modelling-on-wikitext-103`, 2023. Accessed: 2023-06-01.

[PBtTB$^+$15] CR Prins, René Beltman, JHM ten Thije Boonkkamp, Wilbert L IJzerman, and Teus W Tukker. A least-squares method for optimal transport using the monge–ampère equation. *SIAM Journal on Scientific Computing*, 37(6):B937–B961, 2015.

[PDGB14] Razvan Pascanu, Yann N Dauphin, Surya Ganguli, and Yoshua Bengio. On the saddle point problem for non-convex optimization. *arXiv preprint arXiv:1405.4604*, 2014.

[Pea01] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[PHD14] Ji Peng, Jerrad Hampton, and Alireza Doostan. A weighted ℓ_1-minimization approach for sparse polynomial chaos expansions. *Journal of Computational Physics*, 267:92–111, 2014.

[PJ21] Hendrik Bernd Petersen and Peter Jung. Robust instance-optimal recovery of sparse signals at unknown noise levels. *Information and Inference: A Journal of the IMA*, 08 2021.

[PJ22] Hendrik Bernd Petersen and Peter Jung. Robust instance-optimal recovery of sparse signals at unknown noise levels. *Information and Inference: A Journal of the IMA*, 11(3):845–887, 2022.

[PJ23] Soumyabrata Pal and Prateek Jain. Online low rank matrix completion. In *The Eleventh International Conference on Learning Representations*, 2023.

[Pla49] Robin L Plackett. A historical note on the method of least squares. *Biometrika*, 36(3/4):458–460, 1949.

[PMR19a] S. Paternain, A. Mokhtari, and A. Ribeiro. A newton-based method for nonconvex optimization with fast evasion of saddle points. *SIAM J. Optim.*, 29(1):343–368, 2019.

[PMR19b] Santiago Paternain, Aryan Mokhtari, and Alejandro Ribeiro. A newton-based method for nonconvex optimization with fast evasion of saddle points. *SIAM Journal on Optimization*, 29(1):343–368, 2019.

[Pol15] Roman A Polyak. Projected gradient method for non-negative least square. *Contemp Math*, 636:167–179, 2015.

[PP21] Clarice Poon and Gabriel Peyré. Smooth bilevel programming for sparse regularization. *Advances in Neural Information Processing Systems*, 34:1543–1555, 2021.

[PPVF21] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity. *Advances in Neural Information Processing Systems*, 34, 2021.

[PS75] Christopher C Paige and Michael A Saunders. Solution of sparse indefinite systems of linear equations. *SIAM journal on numerical analysis*, 12(4):617–629, 1975.

[PS82] Christopher C Paige and Michael A Saunders. Lsqr: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS)*, 8(1):43–71, 1982.

[PSL+21] Jiangtao Peng, Weiwei Sun, Heng-Chao Li, Wei Li, Xiangchao Meng, Chiru Ge, and Qian Du. Low-rank and sparse representation for hyperspectral image processing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(1):10–43, 2021.

[Qi10] Feng Qi. Bounds for the ratio of two gamma functions. *Journal of Inequalities and Applications*, 2010(1):493058, 2010.

[QSS10] Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical mathematics*, volume 37. Springer Science & Business Media, 2010.

[Ray97] Marcos Raydan. The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. *SIAM Journal on Optimization*, 7(1):26–33, 1997.

[RC20] Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. In *Advances in Neural Information Processing Systems*, pages 21174–21187, 2020.

[Rec11] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

[RFP10] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[RKD99] Bhaskar D. Rao and Kenneth Kreutz-Delgado. An affine scaling methodology for best basis selection. *IEEE Trans. Signal Process.*, 47(1):187–200, 1999.

[RM15] Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

[RMC21] Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in tensor factorization. In *International Conference on Machine Learning*, pages 8913–8924. PMLR, 2021.

[RS05] Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719, 2005.

[RTF16] Stephen Reid, Robert Tibshirani, and Jerome Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67, 2016.

[RV08a] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 61(8):1025–1045, 2008.

[RV08b] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.

[RW16] Holger Rauhut and Rachel Ward. Interpolation via weighted ℓ_1-minimization. *Applied and Computational Harmonic Analysis*, 40(2):321–351, 2016.

[RWX21] Yi Rong, Yang Wang, and Zhiqiang Xu. Almost everywhere injectivity conditions for the matrix recovery problem. *Applied and Computational Harmonic Analysis*, 50:386–400, 2021.

[RWY11] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

[RXH08] Benjamin Recht, Weiyu Xu, and Babak Hassibi. Necessary and sufficient conditions for success of the nuclear norm heuristic for rank minimization. In *2008 47th IEEE Conference on Decision and Control*, pages 3065–3070. IEEE, 2008.

[RXH11] Benjamin Recht, Weiyu Xu, and Babak Hassibi. Null space conditions and thresholds for rank minimization. *Mathematical programming*, 127:175–202, 2011.

[RYL$^+$18] Andy Ramlatchan, Mengyun Yang, Quan Liu, Min Li, Jianxin Wang, and Yaohang Li. A survey of matrix completion methods for recommendation systems. *Big Data Mining and Analytics*, 1(4):308–323, 2018.

[Saa03] Yousef Saad. *Iterative methods for sparse linear systems.* SIAM, 2003.

[San17] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2017.

[SBC16] Weijie Su, Stephen Boyd, and Emmanuel J Candès. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *The Journal of Machine Learning Research*, 17(1):5312–5354, 2016.

[SBP17] Ying Sun, Prabhu Babu, and Daniel P. Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Trans. Signal Process.*, 65(3):794–816, 2017.

[SBvdG10a] Nicolas Städler, Peter Bühlmann, and Sara van de Geer. ℓ_1-penalization for mixture regression models (with discussion). *TEST*, 19(2):280–285, 2010.

[SBvdG10b] Nicolas Städler, Peter Bühlmann, and Sara van de Geer. Rejoinder: ℓ_1-penalization for mixture regression models (with discussion). *TEST*, 19(2):209–285, 2010.

[SC10] Amit Singer and Mihai Cucuringu. Uniqueness of low-rank matrix completion by rigidity theory. *SIAM Journal on Matrix Analysis and Applications*, 31(4):1621–1641, 2010.

[SCGS16] Xinyue Shen, Laming Chen, Yuantao Gu, and Hing-Cheung So. Square-root lasso with nonconvex regularization: An ADMM approach. *IEEE Signal Processing Letters*, 23(7):934–938, 2016.

[SEJ$^+$20] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin **v**Zídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[SH11] Martin Slawski and Matthias Hein. Sparse recovery by thresholded nonnegative least squares. *Advances in Neural Information Processing Systems*, 24, 2011.

[SH13] Martin Slawski and Matthias Hein. Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics*, 7:3004–3056, 2013.

[Sha82] Alexander Shapiro. Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika*, 47:187–199, 1982.

[She73] Oscar B Sheynin. R. j. boscovich's work on probability. *Archive for history of exact sciences*, 9(4/5):306–324, 1973.

[She77] Oscar B Sheynin. Laplace's theory of errors. *Archive for history of exact sciences*, 17(1):1–61, 1977.

[She93] Oscar Sheynin. On the history of the principle of least squares. *Archive for history of exact sciences*, pages 39–54, 1993.

[SHM$^+$16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[SHN$^+$18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[Sil12] Nate Silver. *The signal and the noise: Why so many predictions fail-but some don't.* Penguin, 2012.

[SJC19] Yonatan Shadmi, Peter Jung, and Giuseppe Caire. Sparse non-negative recovery from biased subgaussian measurements using NNLS. *arXiv preprint arXiv:1901.05727*, 2019.

[SJNS19] Yifan Sun, Halyun Jeong, Julie Nutini, and Mark Schmidt. Are we there yet? manifold identification of gradient-related proximal methods. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1110–1119. PMLR, 2019.

[SJY09] Liang Sun, Shuiwang Ji, and Jieping Ye. A least squares formulation for a class of generalized eigenvalue problems in machine learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 977–984, 2009.

[SK21] Grzegorz Stoch and Artur T Krzyżak. Enhanced Resolution Analysis for Water Molecules in MCM-41 and SBA-15 in Low-Field T2 Relaxometric Spectra. *Molecules*, 26(8):2133, 2021.

[SK22] Axel Seguin and Daniel Kressner. Continuation methods for riemannian optimization. *SIAM Journal on Optimization*, 32(2):1069–1093, 2022.

[SL16] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[SLCX23] Yinan Shen, Jingyang Li, Jian-Feng Cai, and Dong Xia. Computationally efficient and statistically optimal robust high-dimensional linear regression. *arXiv preprint arXiv:2305.06199*, 2023.

[SLS$^+$20] Fanhua Shang, Yuanyuan Liu, Fanjie Shang, Hongying Liu, Lin Kong, and Licheng Jiao. A unified scalable equivalent formulation for schatten quasi-norms. *Mathematics*, 8(8):1325, 2020.

[Sog23] Tomohiro Sogabe. *Krylov Subspace Methods for Linear Systems: Principles of Algorithms*, volume 60. Springer Nature, 2023.

[SS86] Fadil Santosa and William W Symes. Linear inversion of band-limited reflection seismograms. *SIAM journal on scientific and statistical computing*, 7(4):1307–1330, 1986.

[SS90] Gilbert W Stewart and Ji-guang Sun. *Matrix perturbation theory*. Academic press, 1990.

[SS21] Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34, 2021.

[SSS$^+$16] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR, 2016.

[ST03] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of termination of linear programming algorithms. *Math. Program., Ser. B*, 97:375–404, 2003.

[Ste93] Gilbert W Stewart. On the early history of the singular value decomposition. *SIAM review*, 35(4):551–566, 1993.

[Ste06] Michael Stewart. Perturbation of the svd in the presence of small singular values. *Linear algebra and its applications*, 419(1):53–77, 2006.

[Ste16] Adrian Stern. *Optical compressive imaging*. CRC Press, 2016.

[Sti81] Stephen M Stigler. Gauss and the invention of least squares. *the Annals of Statistics*, pages 465–474, 1981.

[Sti86] Stephen M Stigler. *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, 1986.

[Sti97] Stephen M Stigler. Regression towards the mean, historically considered. *Statistical methods in medical research*, 6(2):103–114, 1997.

[Sto71] Josef Stoer. On the numerical solution of constrained least-squares problems. *SIAM journal on Numerical Analysis*, 8(2):382–411, 1971.

[Sto10] Mihailo Stojnic. l1 optimization and its various thresholds in compressed sensing. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3910–3913. IEEE, 2010.

[Str12] Thomas Strohmer. Measure what should be measured: progress and challenges in compressive sensing. *IEEE Signal Processing Letters*, 19(12):887–893, 2012.

[Stu99] Jos F Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization methods and software*, 11(1-4):625–653, 1999.

[SV99] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9:293–300, 1999.

[SV21] Damian Straszak and Nisheeth K Vishnoi. Iteratively reweighted least squares and slime mold dynamics: connection and convergence. *Mathematical Programming*, pages 1–33, 2021.

[SvdG17] Benjamin Stucky and Sara van de Geer. Sharp oracle inequalities for square root regularization. *Journal of Machine Learning Research*, 18(67):1–29, 2017.

[SW22] Suvrit Sra and Melanie Weber. On a class of geodesically convex optimization problems solved via euclidean mm methods. *arXiv preprint arXiv:2206.11426*, 2022.

[SXZ18] Alexander Shapiro, Yao Xie, and Rui Zhang. Matrix completion with deterministic pattern: A geometric perspective. *IEEE Transactions on Signal Processing*, 67(4):1088–1103, 2018.

[SZ00] Horst D Simon and Hongyuan Zha. Low-rank matrix approximation using the lanczos bidiagonalization process with applications. *SIAM Journal on Scientific Computing*, 21(6):2257–2274, 2000.

[SZ12] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 09 2012.

[TB22] Lloyd N Trefethen and David Bau. *Numerical linear algebra*, volume 181. SIAM, 2022.

[TBD11] Vincent YF Tan, Laura Balzano, and Stark C Draper. Rank minimization over finite fields: Fundamental limits and coding-theoretic interpretations. *IEEE transactions on information theory*, 58(4):2018–2039, 2011.

[TBM79] Howard L Taylor, Stephen C Banks, and John F McCoy. Deconvolution with the ℓ_1-norm. *Geophysics*, 44(1):39–52, 1979.

[TG07] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

[Tib96a] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

[Tib96b] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[Til15] Andreas M Tillmann. Equivalence of linear programming and basis pursuit. *PAMM*, 15(1):735–738, 2015.

[TKL11] AB Tsybakov, V Koltchinskii, and K Lounici. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011.

[TMC21] Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *The Journal of Machine Learning Research*, 22(1):6639–6701, 2021.

[Tod14] Issac Todhunter. *A History of the Mathematical Theory of Probability: From the Time of Pascal to that of Laplace*. Cambridge University Press, 2014.

[Tsa88] Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52:479–487, 1988.

[Tsa23] Manolis C Tsakiris. Low-rank matrix completion theory via plücker coordinates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[TTT12] Kim-Chuan Toh, Michael J. Todd, and Reha H. Tütüncü. On the Implementation and Usage of SDPT3 - A Matlab Software Package for Semidefinite-Quadratic-Linear Programming, Version 4.0. chapter Chapter 25, pages 715–754. Springer, 2012.

[TTW$^+$14] Mingkui Tan, Ivor W Tsang, Li Wang, Bart Vandereycken, and Sinno Jialin Pan. Riemannian pursuit for big matrix recovery. In *International Conference on Machine Learning*, pages 1539–1547. PMLR, 2014.

[TW13] Jared Tanner and Ke Wei. Normalized iterative hard thresholding for matrix completion. *SIAM Journal on Scientific Computing*, 35(5):S104–S125, 2013.

[TW16] Jared Tanner and Ke Wei. Low rank matrix completion by alternating steepest descent methods. *Applied and Computational Harmonic Analysis*, 40(2):417–429, 2016.

[twi21] Yann LeCun - Twitter from 29.06.2021. `https://twitter.com/ylecun/status/1409940043951742981`, 2021. Accessed: 2023-06-01.

[TWST20] Peipei Tang, Chengjing Wang, Defeng Sun, and Kim-Chuan Toh. A sparse semismooth newton based proximal majorization-minimization algorithm for nonconvex square-root-loss regression problem. *The Journal of Machine Learning Research*, 21(1):9253–9290, 2020.

[TYUC19] Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Streaming low-rank matrix approximation with an application to scientific simulation. *SIAM Journal on Scientific Computing*, 41(4):A2430–A2463, 2019.

[UC16] Konstantin Usevich and Pierre Comon. Hankel low-rank matrix completion: Performance of the nuclear norm relaxation. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):637–646, 2016.

[UT19] Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

[UV15] André Uschmajew and Bart Vandereycken. Greedy rank updates combined with riemannian descent methods for low-rank optimization. In *2015 International Conference on Sampling Theory and Applications (SampTA)*, pages 420–424. IEEE, 2015.

[UV20] André Uschmajew and Bart Vandereycken. Geometric methods on low-rank matrix and tensor manifolds. *Handbook of variational methods for nonlinear geometric data*, pages 261–313, 2020.

[Van13] Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

[Vap99] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[Vaz02] Andrew Vazsonyi. *Which door has the Cadillac: adventures of a real-life mathematician.* iUniverse, 2002.

[VBK04] Mark H Van Benthem and Michael R Keenan. Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(10):441–450, 2004.

[VD17] Sergey Voronin and Ingrid Daubechies. An iteratively reweighted least squares algorithm for sparse regularization. In *Functional Analysis, Harmonic Analysis, and Image Processing: A Collection of Papers in Honor of Björn Jawerth*, pages 391–441. American Mathematical Society, 2017.

[vdG16] Sara van de Geer. *Estimation and Testing Under Sparsity: Cole d't de Probabilits de Saint-Flour XLV - 2015.* Springer Publishing Company, Incorporated, 1st edition, 2016.

[VDGB09] Sara A Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. 2009.

[VDV00] Henk A Van Der Vorst. Krylov subspace iteration. *Computing in science & engineering*, 2(1):32–37, 2000.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[VFH$^+$21] Claudio M Verdun, Tim Fuchs, Pavol Harar, Dennis Elbrächter, David S Fischer, Julius Berner, Philipp Grohs, Fabian J Theis, and Felix Krahmer. Group testing for sars-cov-2 allows for up to 10-fold efficiency increase across realistic scenarios and testing strategies. *Frontiers in Public Health*, 9:583377, 2021.

[Vid19] Mathukumalli Vidyasagar. *An introduction to compressed sensing.* SIAM, 2019.

[VKR19] Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. In *Advances in Neural Information Processing Systems*, pages 2972–2983, 2019.

[VMA$^+$11] SS Vasanawala, MJ Murphy, Marcus T Alley, P Lai, Kurt Keutzer, John M Pauly, and Michael Lustig. Practical parallel imaging compressed sensing mri: Summary of two years of experience in accelerating body mri of pediatric patients. In *2011 ieee international symposium on biomedical imaging: From nano to macro*, pages 1039–1043. IEEE, 2011.

[VNVM14] Nick Vannieuwenhoven, Johannes Nicaise, Raf Vandebril, and Karl Meerbergen. On generic nonexistence of the schmidt–eckart–young decomposition for complex tensors. *SIAM Journal on Matrix Analysis and Applications*, 35(3):886–903, 2014.

[VO98] Curtis R Vogel and Mary E Oman. Fast, robust total variation-based reconstruction of noisy, blurred images. *IEEE transactions on image processing*, 7(6):813–824, 1998.

[Vor12] Sergey Voronin. *Regularization of linear systems with sparsity constraints with applications to large-scale inverse problems.* Ph.D. Thesis, Princeton University, 2012.

[VS16] S Aja-Fernándezand G Vegas-Sánchez. Ferrero, statistical analysis of noise in mri: modeling, filtering and estimation, 2016.

[W$^+$14] David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[Wat90] William C Waterhouse. Gauss's first argument for least squares. *Archive for history of exact sciences*, pages 41–52, 1990.

[WCCL16] Ke Wei, Jian-Feng Cai, Tony F Chan, and Shingyu Leung. Guarantees of riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

[WCCL20] Ke Wei, Jian-Feng Cai, Tony F Chan, and Shingyu Leung. Guarantees of riemannian optimization for low-rank matrix completion. *Inverse Problems and Imaging*, 14(2):233–265, 2020.

[WCZT22] Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large Learning Rate Tames Homogeneity: Convergence and Balancing Effect. In *International Conference on Learning Representations*, 2022.

[Wed72] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

[Wei37] Endre Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43:355–386, 1937.

[WGL$^+$20] Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and Rich Regimes in Overparametrized Models. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 3635–3673, 2020.

[WM22] John Wright and Yi Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.

[WN10] David Wipf and Srikantan 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.

[Woj10] Przemyslaw Wojtaszczyk. Stability and instance optimality for gaussian measurements in compressed sensing. *Foundations of Computational Mathematics*, 10:1–13, 2010.

[Woo50] Max A. Woodbury. Inverting modified matrices. *Memorandum report*, 42(106):336, 1950.

[WP09] Endre Weiszfeld and Frank Plastria. On the point for which the sum of the distances to n given points is minimum. *Annals of Operations Research*, 167:7–41, 2009.

[WR21] Fan Wu and Patrick Rebeschini. Implicit Regularization in Matrix Sensing via Mirror Descent. *Advances in Neural Information Processing Systems*, 34, 2021.

[WRJ21] Ashia C Wilson, Ben Recht, and Michael I Jordan. A lyapunov analysis of accelerated methods in optimization. *The Journal of Machine Learning Research*, 22(1):5040–5073, 2021.

[WU83] Colin Walker and Tad J Ulrych. Autoregressive recovery of the acoustic impedance. *Geophysics*, 48(10):1338–1350, 1983.

[WWZG15] Dingming Wu, Dongfang Wang, Michael Q Zhang, and Jin Gu. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC genomics*, 16(1):1–10, 2015.

[WXT10] Meng Wang, Weiyu Xu, and Ao Tang. A unique "nonnegative" solution to an underdetermined system: From vectors to matrices. *IEEE Transactions on Signal Processing*, 59(3):1007–1016, 2010.

[WYL$^+$22] Guoqiang Wang, Wenjian Yu, Xiubo Liang, Yuanqing Wu, and Bo Yu. An iterative reduction fista algorithm for large-scale lasso. *SIAM Journal on Scientific Computing*, 44(4):A1989–A2017, 2022.

[WYZ12] Zaiwen Wen, Wotao Yin, and Yin 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.

[WZW22] Hao Wang, Hao Zeng, and Jiashan Wang. An extrapolated iteratively reweighted ℓ_1 method with complexity analysis. *Computational Optimization and Applications*, 83(3):967–997, 2022.

[XCM10] Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. *IEEE Transactions on Information Theory*, 7(56):3561–3574, 2010.

[XGJ16] Hongxiang Xie, Feifei Gao, and Shi Jin. An overview of low-rank channel estimation for massive mimo systems. *IEEE Access*, 4:7313–7321, 2016.

[Yan09] Zhe Yang. *A study on nonsymmetric matrix-valued functions*. PhD thesis, Master's thesis, Department of Mathematics, National University of Singapore, 2009.

[YB19] Guo Yu and Jacob Bien. Estimating the error variance in a high-dimensional linear model. *Biometrika*, 106(3):533–546, 2019.

[Ye07] Jieping Ye. Least squares linear discriminant analysis. In *Proceedings of the 24th international conference on Machine learning*, pages 1087–1093, 2007.

[YKM21] Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*, 2021.

[YLP22] Peiran Yu, Guoyin Li, and Ting Kei Pong. Kurdyka-lojasiewicz exponent via inf-projection. *Foundations of Computational Mathematics*, 22(4):1171–1217, 2022.

[YMO13] Yi Yang, Jianwei Ma, and Stanley Osher. Seismic data reconstruction via matrix completion. *Inverse Problems and Imaging*, 7(4):1379–1392, 2013.

[YOGD08] Wotao Yin, Stanley Osher, Donald Goldfarb, and Jerome Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM J. Imaging Sci*, 1(1):143–168, 2008.

[YZ16] Ming Yuan and Cun-Hui Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.

[YZLS22] Baturalp Yalcin, Haixiang Zhang, Javad Lavaei, and Somayeh Sojoudi. Factorization approach for low-complexity matrix completion problems: Exponential number of spurious solutions and failure of gradient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 319–341. PMLR, 2022.

[YZQM20] Chong You, Zhihui Zhu, Qing Qu, and Yi Ma. Robust recovery via implicit bias of discrepant learning rates for double over-parameterization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17733–17744, 2020.

[ZBH$^+$17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[ZBH$^+$21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[ZBL21] Haixiang Zhang, Yingjie Bi, and Javad Lavaei. General low-rank matrix optimization: Geometric analysis and sharper bounds. *Advances in Neural Information Processing Systems*, 34:27369–27380, 2021.

[ZCZ22] Jialun Zhang, Hong-Ming Chiu, and Richard Y Zhang. Accelerating sgd for highly ill-conditioned huge-scale online matrix completion. *Advances in Neural Information Processing Systems*, 35:37549–37562, 2022.

[ZFB23] Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

[Zha21] Tonglin Zhang. Iteratively reweighted least squares with random effects for maximum likelihood in generalized linear mixed effects models. *Journal of Statistical Computation and Simulation*, 91(16):3404–3425, 2021.

[ZHBL10] Bo Zhao, Justin P Haldar, Cornelius Brinegar, and Zhi-Pei Liang. Low-rank matrix recovery for real-time cardiac MRI. In *2010 IEEE international symposium on biomedical imaging: From nano to macro*, pages 996–999. IEEE, 2010.

[ZHH$^+$19] Zaiwei Zhang, Xiangru Huang, Qixing Huang, Xiao Zhang, and Yuan Li. Joint learning of neural networks via iterative reweighted least squares. In *CVPR Workshops*, pages 18–26, 2019.

[ZL16] Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

[ZL18] Richard Y Zhang and Javad Lavaei. Sparse semidefinite programs with near-linear time complexity. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 1624–1631. IEEE, 2018.

[ZN22] Pini Zilber and Boaz Nadler. GNMR: A provable one-line algorithm for low rank matrix recovery. *SIAM Journal on Mathematics of Data Science*, 4(2):909–934, 2022.

[ZWB$^+$21] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on Learning Theory*, pages 4633–4635. PMLR, 2021.

[ZYH19] Peng Zhao, Yun Yang, and Qiao-Chu He. Implicit Regularization via Hadamard Product Over-Parametrization in High-Dimensional Linear Regression. *arXiv preprint: 1903.09367*, 2019.